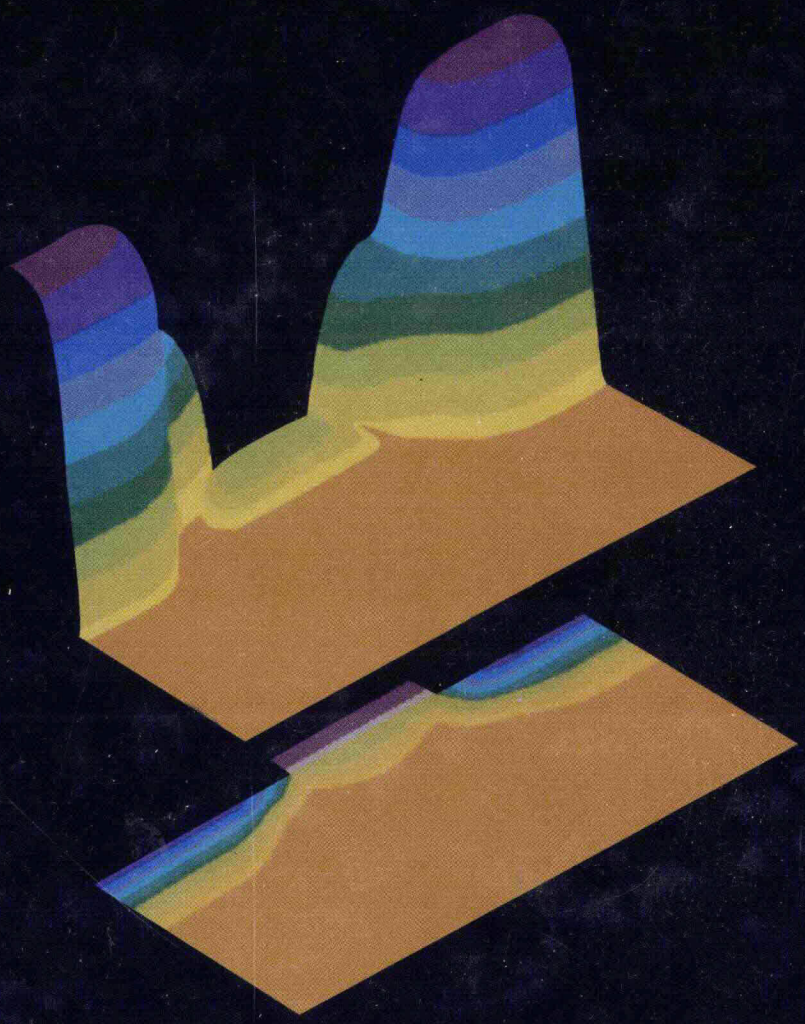
Silicon Processing

for the VLSI Era

Volume 3 - The Submicron MOSFET



S. Wolf

TSMC-1029
U.S. Patent 8,796,779

Silicon Processing for the VLSI Era

Volume 3

The Submicron MOSFET

S. Wolf

LATTICE PRESS

SILICON PROCESSING

FOR

THE VLSI ERA

VOLUME 3:

THE SUBMICRON MOSFET

SILICON PROCESSING

FOR

THE VLSI ERA

VOLUME 3:

THE SUBMICRON MOSFET

STANLEY WOLF Ph.D.
Professor, Department of Electrical Engineering
California State University, Long Beach
Long Beach, California

LATTICE PRESS

Sunset Beach, California

DISCLAIMER

This publication is based on sources and information believed to be reliable, but the authors and Lattice Press disclaim any warranty or liability based on or relating to the contents of this publication.

Published by:

Lattice Press,
Post Office Box 340
Sunset Beach, California 90742, U.S.A.

Cover design by Roy Montibon, Visionary Art Resources, Inc., Los Angeles, CA.

Copyright © 1995 by Lattice Press.
All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system without written permission from the publisher, except for the inclusion of brief quotations in a review.

Library of Congress Cataloging in Publication Data
Wolf, Stanley

Silicon Processing for the VLSI Era
Volume 3 : The Submicron MOSFET

Includes Index
1. Integrated circuits-Very large scale integration. 2. Silicon. I. Title

ISBN 0-961672-5-3

9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

CONTENTS

PREFACE

Chap. 1 - THE ROLE OF PROCESS AND DEVICE MODELS IN MICROELECTRONICS TECHNOLOGY 1

1.1 MODELS IN MICROELECTRONIC TECHNOLOGY 1

1.1.1 Development of Physically Based Models 3
1.1.2 Development of Empirically Based Models 4
1.1.2.1 Empirically Based versus Physically Based Models.
1.1.2.2 Deriving Empirically Based Models.
1.1.2.3 Role of Empirical Models in Semiconductor Simulation.
1.1.2.4 Examples of Quantitative Empirically Based Models.
1.1.2.5 PREDICT-1: A Process Simulator Based on Empirical Models.

1.2 MATHEMATICAL FORMULATION OF THE PHYSICALLY-BASED MODELS OF SEMICONDUCTOR PROCESSES AND DEVICES 9

1.2.1 The Physically Based Model of Impurity Diffusion in Silicon - Fick's Second Law Equation 11
1.2.2 The Physically Based Model of Semiconductor Device Behavior 12
1.2.3 The Poisson Equation 15

1.3 CLOSED-FORM ANALYTICAL SOLUTIONS VERSUS NUMERICAL SOLUTIONS OF THE PDES THAT ARE THE PHYSICALLY BASED MODELS OF MICROELECTRONICS 16

1.3.1 Closed-Form Analytical Solutions - Mathematical Aspects 16
1.3.2 Numerical Solutions - Mathematical Aspects 16
1.3.3 Comparing the Analytical and Numerical Solutions in Semiconductor Simulation Applications 17
1.3.4 Role of Analytical Solutions in Submicron Semiconductor Technology 17
1.3.5 Analytic Solutions of Fick's 2nd Law Equation 19

1.3.6 Analytical Solutions of the One-Dimensional Form of the Poisson Equation in the Depletion Approximation 23

1.4 CIRCUMSTANCES UNDER WHICH THE ANALYTIC SOLUTIONS OF THESE PDES BECOME INADEQUATE FOR SIMULATION PURPOSES 24

1.4.1 The Inadequacy of the Analytic Solutions of Fick's 2nd Law Equation for Simulating Diffusion in Submicron Device Structures 24
1.4.2 The Inadequacy of Using a One-Dimensional Form of the Poisson Equation and the Depletion Approximation to Model Submicron MOSFET Behavior 27

REFERENCES 28

Chap. 2 - NUMERICAL METHODS FOR SOLVING THE PARTIAL DIFFERENTIAL EQUATIONS WHICH MODEL THE PHYSICS OF SUBMICRON DEVICES AND PROCESSES 29

2.1 NUMERICAL SOLUTIONS OF DIFFERENTIAL EQUATIONS 31

2.2 NUMERICAL METHODS FOR SOLVING NONLINEAR PARTIAL DIFFERENTIAL EQUATIONS 32

2.3 GRID GENERATION AND REFINEMENT PROCEDURES 34

2.4 DISCRETIZATION OF DIFFERENTIAL EQUATIONS 36

2.4.1 Space Discretization 39
2.4.2. The Finite-Difference Method 39
2.4.2.1 The Difference Operator Method
2.4.2.2 Taylor Series.
2.4.2.3 One-Dimensional Finite Difference Approximations of Derivatives.
2.4.2.4 Two-Dimensional Space Discretization.
2.4.3 The Finite Element Method 45
2.4.4 Time Discretization 47
2.4.4.1 Explicit Time Discretization.
2.4.4.2 Implicit Time Discretization.

2.5 ORDERING THE SET OF DISCRETIZED PDEs INTO APPROPRIATE MATRIX FORM 50

2.6 LINEARIZING NONLINEAR ALGEBRAIC EQUATIONS 52

2.7 MATRIX METHODS FOR SOLVING SIMULTANEOUS LINEAR ALGEBRAIC EQUATIONS 55

2.8 VISUALIZATION OF THE RESULTS 57

2.9 DATA BASE OF THE SIMULATION PACKAGES 58

2.10 EXAMPLES OF THE PROCEDURES USED TO NUMERICALLY SOLVE THE MODEL PARTIAL-DIFFERENTIAL EQUATIONS OF MICROELECTRONICS 59

2.10.1 Outline of a Numerical Analysis Procedure for Solving the Two-Dimensional Poisson Equation to Simulate the Electrostatic Potential in a MOSFET 59

2.10.1.1 Example of How Device Simulation Programs are Used to Obtain the Two-Dimensional Solution of the Poisson Equation in a MOSFET Structure Operating in Subthreshold.

2.10.2 Example of a Numerical Procedure for Solving the One-Dimensional Form of Fick's 2nd Law Equation as a Model for Impurity Diffusion in Silicon 70

2.11 SYSTEMS OF COUPLED EQUATIONS 74

2.11.1 Gummel's Iterative Method 75
2.11.2 Newton's Direct Method 76

2.12 EVOLUTION OF MOSFET DEVICE SIMULATORS 77

2.12.1 CADDET 78
2.12.2 MINIMOS 78
2.12.3 GEMINI 78
2.12.4 PISCES-II 80
2.12.5 Three-Dimensional Device Simulators 80

REFERENCES 81

Chap. 3 - MOS TRANSISTOR DEVICE PHYSICS: PART 1 - BASICS MOS PHYSICS and THE MOS CAPACITOR **83**

3.1 BASIC MOS THEORY and PHYSICS of IDEAL MOS CAPACITORS (MOS-C) 84

3.1.1 The Structure of the MOS-C and the MOSFET 85

3.1.1.1 Comparison of MOS-Cs and Conventional Capacitor Structures.

3.1.1.2 Why is the MOS-C Studied First?

3.2 QUALITATIVE DESCRIPTION OF MOS-C DEVICE BEHAVIOR 89

3.2.1 Energy-Band Diagram Representation of Semiconductor Devices 90
3.2.2 Energy Band Diagram of the Ideal MOS Capacitor 91
3.2.2.1 Accumulation.
3.2.2.2 Depletion.
3.2.2.3 Inversion.
3.2.2.4 Strong Inversion.
3.2.2.5 Microscopic Perspective of the Mobile Charge Behavior in a MOS-C.

3.3 QUANTITATIVE DESCRIPTION OF MOS-C BEHAVIOR 96

3.3.1 The Electrostatic Potential in MOS Structures 98
3.3.2 General Approach to Deriving MOS-C Characteristics 100
3.3.3 Finding the Quantitative Characteristics of a MOS-C when the Depletion Approximation is Invoked 103
3.3.4 More Accurate and General Model for MOS-C Behavior 106
3.3.5 Charge-Sheet Model for Determining the Characteristics of the Ideal MOS-C 112
3.3.5.1 Approximation of Q_I in Weak Inversion.
3.3.5.2 Approximation of Q_I in Strong Inversion.

3.4 THRESHOLD VOLTAGE OF IDEAL MOS CAPACITORS 116

3.5 PHYSICS OF NON-IDEAL MOS CAPACITORS (n^+PolySi-SiO_2-Si MOS Capacitors) 117

3.5.1 Workfunction Difference Between Gate and Semiconductor 117
3.5.2 Oxide Charges and Traps 119
3.5.2.1 Mobile Ionic Charges.
3.5.2.2 Fixed Oxide Charge.
3.5.2.3 Interface Trapped Charge.
3.5.2.4 Oxide Trapped Charge.
3.5.2.5 Combined Effect of Oxide Charges on V_T of Submicron MOSFETs.

3.6 THE THRESHOLD VOLTAGE EQUATION OF Si-SiO_2 MOS CAPACITORS 127

3.7 VISUALIZATION OF MOS-C AND MOSFET PHENOMENA USING POTENTIAL VERSUS POSITION DIAGRAMS 129

REFERENCES 130

Chap. 4 - MOS TRANSISTOR DEVICE PHYSICS: PART 2 - LONG-CHANNEL MOSFETS 134

4.1 TERMINOLOGY OF MOSFET DEVICE STRUCTURES 135

4.1.1 Basics of MOSFET Operation 136

4.2 BIASING THE INVERSION LAYER (BODY EFFECT) 138

4.2.1 Qualitative Discussion of the Effect of Applying a Bias to the Inversion Layer of a MOS Structure 138

4.2.1.1 Case When Both V_{GB} and $V_{CB} = 0V$.

4.2.1.2. Case When $V_{GB} \geq V_T$ and $V_{CB} = 0V$.

4.2.1.3. Case When $V_{GB} = 0V$, But $V_{CB} > 0V$.

4.2.1.4. Case When $V_{GB} \geq V_T$ and $V_{CB} > 0V$.

4.2.1.5 Summary of the Behavior of the Three-Terminal MOS Structure and the Significance of Such Behavior on the Operation of the MOSFET.

4.2.1.6 Body Effect.

4.2.2 Quantitative Description of Effects of Biasing the Inversion Layer of MOS Structures 144

4.2.2.1 General Case.

4.2.2.2 Strong Inversion Case.

4.3 DEFINITION OF THE TERM: LONG-CHANNEL MOSFET 147

4.3.1 The Gradual Channel Approximation (GCA) 149

4.3.1.1 Verifying the Validity of the GCA as a Function of the Geometrical Factors of the MOSFET Device Structure and Bias Conditions.

4.3.2 Circuit Characteristics of Long-Channel MOSFETs 152

4.4 QUALITATIVE DESCRIPTION OF MOSFET OPERATION 153

4.4.1 MOSFET Behavior in Response to Gate Bias Alone 153

4.4.2 Qualitative MOSFET Behavior when both Gate and Drain Bias are Applied 158

4.5 QUANTITATIVE BEHAVIOR OF MOSFETS: DC CIRCUIT MODELS OF LONG-CHANNEL MOSFETS 162

4.5.1 General Approach to Deriving the DC Drain-Current Relationships in Long-Channel MOSFETs 163

4.5.2 Pao-Sah Model 166

4.5.3 The Charge Sheet Model 170

4.5.4 Variable Depletion-Charge (or Bulk-Charge) Model 173

4.5.5 Square-Law Model 175
4.5.6 Comparison of the Four MOSFET Circuit Models 177
4.5.6.1 MOSFET Models of the SPICE Circuit Simulator.

4.6 THRESHOLD-VOLTAGE CONTROL IN MOSFETS 182

4.6.1 Measuring Threshold Voltage 182
4.6.2 Adjusting Threshold Voltage 182
4.6.3 Ion Implantation for Adjusting Threshold Voltage 183
4.6.3.1 Impact of the V_T-adjust Implant on the Body Effect.
4.6.4 Impact of V_T-Adjust Implant on Other MOSFET Characteristics 188

4.7 SUBTHRESHOLD CURRENTS IN LONG-CHANNEL MOSFETs (I_{Dst} when $V_{GS} < |V_T|$) 191

4.7.1 Subthreshold Swing S_t 196
4.7.2 Gate-Induced Drain Leakage (GIDL) 198

4.8 SUMMARY OF LONG-CHANNEL MOSFET BEHAVIOR 200

REFERENCES 201

Chap. 5 - MOS TRANSISTOR DEVICE PHYSICS: PART 3 - THE SUBMICRON MOSFET 205

5.1 DEVICE CHARACTERISTICS OF LONG-CHANNEL VERSUS SHORT-CHANNEL MOSFETS 207

5.2 EFFECT OF GATE DIMENSIONS ON THRESHOLD VOLTAGE 208

5.2.1 Short-Channel Threshold-Voltage Shift 208
5.2.1.1 Numerical Solution of the 2-D Poisson Equation for Calculating ΔV_{TSC}.
5.2.1.2 Charge-Sharing Models.
5.2.1.3 Barrier-Lowering Models and DIBL.
5.2.1.4 Quasi-Two-Dimensional Analytical Solution for ΔV_{TSC}.
5.2.1.5 Measuring Surface-DIBL Effects.
5.2.1.6 Dependence of V_T on V_{BS} in Short-Channel MOSFETs.
5.2.2 Narrow Gate-Width Effects on Threshold Voltage 222
5.2.2.1 Devices Exhibiting Both Short- and Narrow-Width Effects.

5.2.3 Reverse Short-Channel Effects on Threshold Voltage 226

5.2.3.1 Model which Proposes RSCE is Due to Lateral Dopant Nonuniformity at the Channel Si-SiO_2 Interface (Arising from Enhanced Diffusion of Channel Dopants Caused by Interstitial Injection During Poly Reox, Salicide Formation, or Implant Damage).

5.2.3.2 Model Which Proposes RSCE Arises from Boron Segregation to Implant Damaged Regions at the Edge of the Channel.

5.2.3.3 Model Which Attributes RSCE to Transient Enhanced Diffusion of Channel Dopants to the Silicon Surface Arising from Implant Damage.

5.2.3.4 Model which Attributes RSCE to Silicon Interstitial Capture in the Gate Oxide.

5.3 SHORT-CHANNEL EFFECTS ON SUBTHRESHOLD SWING, S_t (Subsurface Punchthrough) 232

5.3.1 Experimentally Characterizing Punchthrough 240

5.3.2 Modeling Punchthrough 242

5.3.3 Modeling the Simultaneous Occurrence of Surface-DIBL and Subsurface-DIBL 245

5.4 SHORT-CHANNEL EFFECTS ON THE I-V CHARACTERISTICS OF MOSFETS OPERATED IN THE STRONG THRESHOLD REGIME 247

5.4.1 Impact of the MOSFET Electric Fields on Mobility 248

5.4.1.1 Models of Mobility Degradation Due to Gate-Induced Electric Fields.

5.4.1.2 Models of the Carrier-Velocity Dependence on Drain Electric Field (Velocity Saturation).

5.4.2 A Simple, Analytic DC Circuit Model of the Short-Channel MOSFET in Strong Inversion 254

5.4.2.1 DC Circuit Model of the Short-Channel MOSFET in the Linear Regime.

5.4.2.2 Effects of Scaling on g_{msat}.

5.4.2.3 Comparing the Values of I_D Calculated with the Long- and Short-Channel MOSFET Models as a Function of Channel Length.

5.4.3 Modeling the Saturation Region of Operation in Short-Channel MOSFETs 260

5.4.3.1 Modeling Short-Channel MOSFET Saturation Behavior Using the Abrupt Junction and Depletion Approximations.

5.4.3.2 Constant Field-Gradient Model.

5.4.3.3 Pseudo-Two-Dimensional Model.

5.5. SHORT-CHANNEL MOSFET MODELS USED IN THE *SPICE* CIRCUIT SIMULATOR 271

5.5.1 BSIM Model Equations 272
5.5.1.1 BSIM Threshold Voltage Equation.
5.5.1.2 BSIM Equations for I_D when the MOSFET is Operated in Strong Inversion.
5.5.1.3 BSIM Equations for I_D when the MOSFET is Operated in Weak Inversion (Subthreshold Regime).

5.6 MOSFET SCALING 274

5.6.1 Design of Submicron MOSFETs: Scaling Guidelines 274
5.6.1.1 Constant Electric-Field, Constant Voltage & Electrostatic Scaling.
5.6.1.2 Subthreshold Scaling.
5.6.1.3 Off-Current Scaling Approach.
5.6.2 Manufacturing Considerations in Submicron MOSFET Device Design 285

5.7 DESIGN OF SUBMICRON MOSFETS 286

5.7.1 Design Methodology Based on the Subthreshold Scaling Equation 286
5.7.2 Design Methodology Based on Required Off-Current Value 287

5.8 *p*-CHANNEL DEVICES IN CMOS 288

5.8.1 PMOS Devices with n^+-Polysilicon Gates 289
5.8.1.1 Punchthrough Susceptibility.
5.8.2 Why PMOS Devices with n^+-Polysilicon Gates and Implanted *p*-Surface Layers Exhibit Enhanced Susceptibility to Short-Channel Effects 291
5.8.2.1 Threshold-Voltage Roll-Off in Buried-Channel PMOS.
5.8.2.2 Subthreshold Swing (S_t) in Buried-Channel PMOS.
5.8.2.3 Punchthrough Susceptibility of Buried-Channel PMOS.
5.8.3 Techniques Used to Suppress Short-Channel Effects in Buried-Channel PMOS 305
5.8.3.1 Techniques Used to Keep the Implanted p-Layer as Thin as Possible.
5.8.3.2 Shallow p^+ Source/Drain Junctions to Suppress Punchthrough in PMOSFETs.
5.8.3.3 LDD Structures in PMOS to Limit Punchthrough.
5.8.3.4 New PMOS Device Structures Designed to Suppress Punchthrough.
5.8.4 PMOS Devices with p^+--Polysilicon Gates 311

REFERENCES 313

Chap. 6 - ISOLATION STRUCTURES IN SUBMICRON CMOS 321

6.1 CHARACTERIZING MOS ISOLATION

6.1.1 Test Structures for NMOS Isolation Structures 322
6.1.2 Punchthrough Prevention between Adjacent Devices in MOS Circuits 325

6.2 THE BASIC LOCOS ISOLATION PROCESS 330

6.2.1 Details of the Basic Semi-recessed LOCOS Process 332
6.2.1.1 Pad-Oxide Layer.
6.2.1.2 CVD of Silicon Nitride Layer.
6.2.1.3 Mask and Etch Pad-Oxide/Nitride Layer to Define Active Regions.
6.2.1.4 Channel-Stop Implant.
6.2.1.5 Problems Arising from the Channel-Stop Implants.
6.2.1.6 Grow Field Oxide.
6.2.1.7. Factors Which Impact Bird's Beak Length and Shape.
6.2.1.8 Use of High-Energy Ion Implantation to Introduce the Channel-Stop Dopant After Field Oxide Growth.
6.2.1.9 Strip the Masking Nitride/Pad-Oxide Layer.
6.2.1.10 Regrow Sacrificial Pad Oxide and Strip (Kooi Effect).
6.2.2. Planarizing Semirecessed LOCOS Structures 348

6.3 MODELING THE TOPOGRAPHY OF LOCOS STRUCTURES 349

6.3.1 Empirical Models of 2-D Thermal Oxidation 350
6.3.2 Physically Based Models of 2-D Oxidation 351
6.3.3 Simulating Thermal Oxidation in 3-D Using Physical Models 355

6.4 POLY-BUFFERED LOCOS 357

6.4.1. Modified Poly-Buffered LOCOS Processes 361
6.4.1.1 Framed-Mask PBL (FMPBL)
6.4.1.2 Polysilicon-Encapsulated LOCOS (PELOX)
6.4.1.3 Nitride-Clad LOCOS (NCL)
6.4.1.4 Mitigating the Field-Thinning Problem of LOCOS

6.5 SHALLOW TRENCH AND REFILL ISOLATION 367

6.5.1 Early Shallow Trench and Refill Isolation 368
6.5.2 Improved Shallow Trench and Refill Isolation Processes 371

6.6 CMOS ISOLATION TECHNOLOGY 373

6.6.1 Qualitative Description of Isolation in CMOS 374
6.6.1.1 Leakage Currents in CMOS Gates Due to Active Device Phenomena and Intra-Well Isolation Loss
6.6.1.2 Leakage Currents in CMOS Gates Due to Loss of Isolation Among Opposite-Type FETs
6.6.1.3 Why n^+-to-p^+ Isolation Spacing is Greater than n^+-to-n^+ Spacing in NMOS (or p^+-to-p^+ Spacing in PMOS)
6.6.1.4 Effects on n^+-to-p^+ Spacing of Dopant Redistribution at the Well Borders
6.6.1.5 Threshold Voltage Rolloff, DIBL, and Lateral Punchthrough in Submicron CMOS Parasitic Field FETs
6.6.2 Quantitative Modeling of CMOS Isolation 384
6.6.3 Experimental Characterization of Isolation in CMOS 387
6.6.3.1 Experimental Characterization of Isolation in n^+-to-n-well Parasitic FETs.
6.6.3.2 Experimental Characterization of Isolation in p^+-to-p-substrate Parasitic FETs.
6.6.4 Establishing the Minimum n^+-to-p^+ Isolation Spacing in CMOS from Modeling and Experimental Data 398
6.6.4.1 Isolation Spacing Design Rules for 0.35 and 0.25 μm CMOS.
6.6.5 Vertical Isolation in CMOS 402
6.6.6 LOCOS-Based Isolation in CMOS 405
6.6.7 Trench Isolation for CMOS 407

REFERENCES 414

Chap. 7 - THIN GATE OXIDES: GROWTH AND RELIABILITY **421**

7.1 GATE OXIDE CHARACTERISTICS NEEDED FOR SUBMICRON MOSFETS 422

7.2 PHYSICAL PICTURE OF SiO_2 AND THE Si/SiO_2 INTERFACE 423

7.2.1 Physical and Chemical Properties of SiO_2 423
7.2.2 The Si/SiO_2 Interface 425

7.3 CARRIER INJECTION IN THE Si/SiO_2 SYSTEM 429

7.3.1 Channel Hot Electrons (CHE) 430
7.3.2 Injection of Drain Avalanche Hot Carriers (DAHC) 434
7.3.3 Fowler-Nordheim (FN) Tunneling 435

7.3.4 Direct Tunneling 437

7.4 TESTS FOR GATE OXIDE RELIABILITY 438

7.4.1 Measuring Dielectric Strength 438
7.4.2 Measuring Time-Dependent Breakdown Behavior 440
7.4.2.1 Time-to-Breakdown under Constant-Voltage Stressing.
7.4.2.2 Time-to-Breakdown under Constant-Current Stressing.
7.4.2.3 Charge to Breakdown (Q_{BD}).
7.4.3 Mathematics of Reliability Characteristics 441
7.4.3.1 The Failure Unit (FIT).
7.4.3.2. The Cumulative Distribution Function.
7.4.3.3 Probability Density Function.
7.4.3.4 The Instantaneous Failure Rate.
7.4.3.5 The Mean Time to Failure.
7.4.3.6 The Weibull Cumulative Distribution Function.
7.4.3.7 The Log-Normal Cumulative Distribution Function.
7.4.3.8 Characterizing the Reliability of Oxides in Terms of the Mathematical Reliability Functions.

7.5 THE PHENOMENON OF OXIDE BREAKDOWN 448

7.5.1 Qualitative Physical Models of Intrinsic Oxide Breakdown 451
7.5.1.1 Hole Generation and Trapping Model.
7.5.1.2 Wolters' Electron Lattice-Damage Model.
7.5.2 Qualitative Models for the Physical Origin of B-Mode Failures in Oxides 458
7.5.2.1 Test Structure for Intrinsic Oxide Breakdown Measurements.
7.5.2.2 Damage at the Gate Periphery Resulting from the Polysilicon Etch Process.
7.5.3 Empirical Quantitative Model of Intrinsic Oxide Breakdown 462
7.5.4 Quantitative Models for B-Mode Failures 467
7.5.4.1 Reduced Barrier Height Model.
7.5.4.2 Oxide Thinning Model.
7.5.4.3. The Cumulative Defect Density Function $D(\Delta t_{ox})$.
7.5.4.4. Using $D(\Delta t_{ox})$ to Predict Oxide Breakdown Reliability.
7.5.4.5. Ramp-Voltage Breakdown Test for Obtaining $D(\Delta t_{ox})$ Data.
7.5.4.6. Correlating Ramp-Voltage Test Data and Constant-Voltage TDDB Data.
7.5.4.7 Interface Trap Generation as a Measure of Oxide Damage Due to Charge Injected into the Oxide.
7.5.5 Applications of the Oxide Breakdown Reliability Model 477
7.5.5.1 Projecting the Reliability of Gate Oxides with Various Thicknesses.

7.5.5.2 If a MOS IC is to have a Specific Lifetime (e.g., 10-Years), What is the Minimum Oxide Thickness that can be Used? (Assume the Maximum Cumulative Failure Percentage, Operating Temperature, and Oxide Area are Specified.)

7.6 SCREENING TESTS FOR OXIDE BREAKDOWN (BURN-IN) 482

7.6.1 Modeling the Effect of the Burn-In Stress on the Post-Burn-In Reliability of the Oxides 485
7.6.2 Procedure for Selecting an Optimal Set of Burn-In Conditions for Oxide Breakdown Screening 486
7.6.3 Example of Applying the Procedure for Determining an Optimal Set of Burn-In Conditions 489

7.7 OXIDE BREAKDOWN MODULE IN BERKELEY RELIABILITY TOOL (BERT) 492

7.8 MODELS OF THIN OXIDE GROWTH 492

7.9 TECHNOLOGY OF THIN OXIDE GROWTH 495

7.9.1 Oxidation Furnaces 495
7.9.2 Control of Gate Oxide Growth Rates 495
7.9.3 Factors Impacting Gate-Oxide Thickness Uniformity 496
7.9.4 Gate Oxide Growth Processes for 10-20 nm thick Gate Oxides on 150-200 mm Wafers 499
7.9.5 Alternative Gate Oxide Growth Processes 502
7.9.5.1 Stacked Gate Oxide Films (Thermally Grown Layer and CVD Layer).

7.10 GATE OXIDE DAMAGE FROM PLASMA PROCESSING 504

7.10.1 How Isolated Conductor Regions on a Wafer Surface Can be Charged Up by an RF Plasma 504
7.10.2 Causes of Non-Uniformities in Plasmas 506
7.10.3 Test Structures used to Characterize Plasma-Induced Oxide Damage 507
7.10.4 Electrical Tests for Measuring Oxide Damage 509
7.10.4.1 Wafer Mapping to Determine Positional Dependence of Damage on the Wafer Surface.
7.10.5 Oxide Damage Models in Various Plasma Etching Processes 511
7.10.5.1 Oxide Damage During Polysilicon or Metal Etch Processes.
7.10.5.2 Oxide Damage During Resist Stripping.
7.10.5.3 Radiation Induced Oxide Damage.
7.10.6 Procedures for Reducing Plasma-Induced Oxide Damage 514

REFERENCES 515

Chap. 8 - WELL FORMATION IN CMOS 523

8.1 *p*-WELL, *n*-WELL, and TWIN-WELL TECHNOLOGIES 524

8.1.1 *p*-Well CMOS 525
8.1.2 *n*-Well CMOS 527
8.1.3 CMOS on Epitaxial Substrates 528
8.1.4 Twin-Well CMOS 530
8.1.5 Summary of CMOS Well-Technology Issues 532

8.2 CONVENTIONAL FORMATION OF WELLS AND CHANNEL STOPS 532

8.2.1 Aligning Subsequent Patterns to the Wells 532
8.2.2 Twin-Well and Channel-Stop Formation 535
8.2.3 Split Well-Drive-In Process 538

8.3 RETROGRADE-WELL CMOS 539

8.3.1 Retrograde Well Technologies to Improve Isolation 546

8.4 FUTURE TRENDS IN WELL FORMATION 552

8.5 AFTER-GATE IMPLANTATION (AGI) 554

REFERENCES 555

Chap. 9 - HOT-CARRIER RESISTANT PROCESSING AND DEVICE STRUCTURES 559

9.1 MODEL FOR THE MAXIMUM LATERAL ELECTRIC FIELD $\mathcal{E}_{ymax}$ IN NON-GRADED-DRAIN MOSFETS 561

9.2 SUBSTRATE CURRENTS DUE TO HOT CARRIERS 563

9.3 MODELS OF HOT-CARRIER DEGRADATION PHENOMENA 566

9.3.1. Si-H Bond Breaking Model 569
9.3.2 Hot-Hole and Hot-Electron Trapping Model 569
9.3.3 Location of the Hot-Carrier Damaged Region 571

9.4 CHARACTERIZING MOSFET DEGRADATION DUE TO HOT-CARRIER EFFECTS 572

9.4.1 Device Parameters that are Impacted by Hot-Carrier Damage 573

9.4.2 How to Determine if a MOSFET Can be Operated with out Failing due to Hot-Carrier Effects Under Normal Operating Conditions? 574
9.4.2.1 Selecting a Criterion that Identifies Device Failure Due to Hot-Carrier Damage.
9.4.3 Accelerated Testing of Hot-Carrier Device Lifetime 576
9.4.4 Hot-Carrier Lifetime Under AC Stressing 580
9.4.5 Effect of Device Degradation on Circuit Performance 581
9.4.6 Hot-Carrier Reliability Simulation 582
9.4.6.1 SCALE.
9.4.6.2 Simulators that Relate the Device-Level Degradation Parameters to Circuit Performance (CAS and HOTRON).

9.5 DECREASING HOT CARRIER DEGRADATION BY MODIFYING THE MOSFET STRUCTURE 586

9.5.1 Phosphorus-Drain Structure 588
9.5.2 Double Diffused Drain (DDD) 588

9.6 LIGHTLY DOPED DRAIN (LDD) STRUCTURES 591

9.6.1 Modeling ε_{ymax} in LDD MOSFETs 592
9.6.1.1 Simple Model for Estimating ε_{ymax} in LDD MOSFETs
9.6.1.2 Analytical Model for Calculating ε_{ymax} in LDD MOSFETs Using the Pseudo-2D Approach to Solving the Poisson Equation.
9.6.1.3 Model for the Series Resistance Introduced by the n^- Region of the LDD.
9.6.2 Overview of the Process Sequence to Form LDDs 595

9.7 PERFORMANCE AND HOT-CARRIER LIMITATIONS OF THE CONVENTIONAL LDD 598

9.7.1 Performance Degradation in LDD MOSFETs 599
9.7.1.1 I_D Decrease Due to Added Resistance of n^- Regions.
9.7.1.2 Doping Compensation Effects that can Increase R_{n^-} in Deep Submicron LDD MOSFETs.
9.7.1.3 Increase of R_{sh} by Weak Overlap of the Gate and Drain.
9.7.1.4 The Impact of Poly Re-Ox on Weak Overlap in MOSFETs.
9.7.1.5 Off-Axis Implant Effects in MOSFETs, Leading to Weak Overlap Problems and Asymmetrical MOSFET Device Characteristics.
9.7.1.6 Implanter Issues that Impact the LDD Process Sequence.
9.7.1.7 Process integration issues which arise when an LDD is integrated into a CMOS Technology.
9.7.2 Hot Carrier Reliability Issues in Early LDDs 606

9.8 DISPOSABLE SPACERS 608

9.9 SUMMARY OF THE LIMITATIONS OF EARLY LDD DEVICES 611

9.10 SUMMARY OF THE METHODS TO BE USED IN DESIGNING EFFECTIVE LDD STRUCTURES 612

9.11 IMPROVED CONVENTIONAL (OXIDE-SPACER-BASED) LDD STRUCTURES 613

9.11.1 Increasing the Doping in the n^- Region of the LDD 614
9.11.1.1 Moderate Lightly Doped Drain (M-LDD).
9.11.1.2 Metal-coated LDD (MLD).
9.11.2 Tailoring the Vertical Doping Profile of the n^- Implant to Improve the Conventional LDD 616
9.11.2.1 Buried-LDD (B-LDD).
9.11.2.2 Profiled-LDD (P-LDD).
9.11.3 Improving the Short-Channel Performance of the Conventional NMOS LDD 621

9.12 FULLY OVERLAPPED LDD STRUCTURES 623

9.12.1 Inverse-T Gate LDD (ITLDD) and Gate-Drain Overlapped Device (GOLD) Structures 623
9.12.1.1 Inverse-T Gate LDD Structure.
9.12.1.2 Gate-Drain Overlapped LDD (GOLD).
9.12.1.3 Modified IT-LDD Structures.
9.12.2 Large-Angle-Tilt Implanted Drain (LATID) 630
9.12.3 High Dielectric Spacers 634
9.12.4 Asymmetrical MOSFET Structure: Halo-Source-GOLD (HS-GOLD) 636

9.13 HOT-CARRIER EFFECTS IN PMOS TRANSISTORS 637

9.13.1 Model for Hot-Carrier Damage in PMOSFETs 638
9.13.2 Hot-Electron Induced Punchthrough (HEIP) 638
9.13.3 Hot-Electron Induced Degradation in p^+-Poly PMOSFETs 641

9.14 STRENGTHENING OXIDES AGAINST HOT-CARRIER INJECTION 641

9.14.1 Minimizing Hydrogen and H_2O at the Si-SiO_2 Interface 643
9.14.1.1. The role of hydrogen in hot-carrier degradation.

9.14.1.2 Hydrogen introduced at high temperatures.
9.14.1.3 Moisture in Gate Oxides as a Contributor to Hot-Carrier Degradation.
9.14.2 Gate-Oxide Damage by Metal Etching 647
9.14.3 Nitrided Oxides 648
9.14.3.1 Thermal nitridation of SiO_2.
9.14.3.2 NH_3 Oxides as MOSFET Gate Dielectrics.
9.14.3.3 Low-Pressure and RTP NH_3 Oxides (Lightly Nitrided Reoxidized NH_3-Nitrided Oxides).
9.14.3.4 Oxynitridation of Silicon in N_2O Ambients.
9.14.3.5. Oxynitridation of Oxides in N_2O.
9.14.3.6 Fluorinated Oxides.

REFERENCES 661

***APPENDIX A* - MATRIX METHODS** 675

A.1 INTRODUCTION TO MATRICES 675
A.2 MATRIX MULTIPLICATION 677
A.3 MATRIX INVERSION 678
A.4 MATRIX METHODS FOR SOLVING SIMULTANEOUS LINEAR ALGEBRAIC EQUATIONS 680
A.5 NON-LINEAR ALGEBRAIC EQUATIONS 691

***APPENDIX B* - MAXIMIZING THE CIRCUIT PERFORMANCE OF LONG-CHANNEL MOSFETS THROUGH DEVICE DESIGN AND PROCESSING TECHNOLOGY** 695

B.1 CHARACTERISTICS OF DIGITAL IC MOSFETS 695
B.2 DESIRED CIRCUIT CHARACTERISTICS OF IC INTERCONNECTS 697

***APPENDIX C* - DERIVATION OF THE BSIM I_D MODEL** 705

***APPENDIX D* - DERIVATION OF THE DISCRETIZED 2-D POISSON EQUATIONS USING THE BOX-INTEGRATION METHOD** 708

INDEX 709

PREFACE

SILICON PROCESSING FOR THE VLSI ERA now represents a series of texts designed to provide a comprehensive and up-to-date treatment of this important and rapidly changing field. The volume in hand is the third of this series. Volumes 1 (*Process Technology*) and 2 (*Process Integration*) were published in 1986 and 1990, respectively. Volume 1 deals with the individual processes employed in the fabrication of silicon VLSI circuits, such as epitaxial growth, thermal oxidation of silicon, CVD and PVD of amorphous and polycrystalline films, diffusion and ion implantation of dopants in silicon, microlithography, and patterning technology. Volume 2 describes how the individual processes of Volume 1 are combined in various ways to produce MOS and bipolar VLSI and ULSI circuits. The task of integrating these various fabrication processes together is referred to as *process integration.*

Here, we treat the topics of submicron MOSFET device physics and the relationship between such device physics and submicron MOSFET fabrication. That is, an understanding of device physics has become even more important now that MOSFETs have crossed the long-channel frontier into the submicron realm. Device aspects that could once be ignored (because they cause only second-order effects in large MOS devices) are now significant. Many device effects observed in submicron MOSFETs are impacted by the process technology used to fabricate them. Some of such effects include: short-channel effect on V_T; RSCE (reverse-short-channel effect on V_T); DIBL (drain-induced barrier lowering); narrow-width effect; reverse-narrow-width effect; subsurface punchthrough in NMOSFETs; punchthrough in PMOSFETs; impact of the V_T-adjust implant on subthreshold swing, S_t; boron penetration of gate oxides in p^+-poly-gate PMOSFETs - and the polysilicon depletion effect that this causes; GIDL (gate-induced drain leakage); impact of process technology on the reliability and wearout of thin gate oxides, including damage from plasma processing; hot-carrier degradation and drain-engineered MOSFET structures developed to combat the problem (including LDDs, LATID, "halo" implants, and asymmetrical MOSFETs); well-engineering by high-energy implants (including retrograde-well CMOS); ROXNOX (re-oxidized-nitrided oxides); and very-lightly-nitrided gate oxides. These and many other phenomena associated with submicron MOS fabrication can only be understood in the context of short-channel MOSFET physics. Only with such an understanding can the relationship between circuit behavior, device design, and process technology of submicron MOSFETs be grasped. Consequently, by gaining such an understanding process engineers (and other microelectronic professionals) will be better able to contribute to the task of successfully designing and manufacturing submicron ICs. One of our main purposes is thus to provide a text that treats both the topics of submicron MOSFET device physics and the phenomena associated with fabricating such devices.

Chapter 1 introduces the process and device models employed in the early days of the semiconductor industry. Such models were based on empirical data and/or simplified physical equations (e.g., differential equations in their one-dimensional form). In the

submicron era, however, the more general forms of these differential equations must be used to obtain accurate predictive capability of both the fabrication processes and the device physics. Since this generally means that partial differential equations in two- or three-dimensional form must be solved, this can only be carried out by numerical analysis, performed with the aid of high-speed digital computers. Thus, chapter 2 outlines the methodology of this approach.

To facilitate the discussion of submicron MOSFET device physics, we divide the topic up into three chapters: The first of these, chapter 3, deals with basic MOS theory and the MOS capacitor. The next, chapter 4, covers long-channel MOSFETs and the circuit models developed to predict the drain current characteristics of such devices. Finally, chapter 5 describes the characteristics of the short-channel (i.e., submicron) MOSFET. This is the longest chapter of the book, and it includes the following topics: DIBL, subsurface punchthrough, the drain current in saturation, simplified short-channel MOSFET circuit models, MOSFET scaling, andsubmicron PMOSFETs.

Chapter 6 is concerned with isolation issues in integrated circuits with an emphasis on CMOS technology. Chapter 7 covers thin gate oxides, primarily with respect to their reliability and growth. Chapter 8 deals with well formation in CMOS. Chapter 9 focuses on hot-carrier effects in MOS ICs, as well as on device structures and processing techniques that mitigate the detrimental aspect of such phenomena on MOSFETs. Problems are included at the end of each chapter to assist readers in gauging how well they have assimilated the material in the text.

A book of this length and diversity would not have been possible without the indirect and direct assistance of many other workers. From an indirect perspective, virtually all the information presented here is based on the research efforts of countless numbers of scientists and engineers. Their contributions are recognized to a small degree by citing some of their articles in the references given at the end of each chapter. Likewise, the direct help came in a variety of forms, and was generously provided by many people. The text is a much better work as a result of this aid, and the author expresses heartfelt thanks to those who gave of their time, energy, and intellect.

Roy Montibon of Visionary Art Resources, Inc., Los Angeles, CA designed this cover, as well as those of Volumes 1 and 2.

Stanley Wolf

P.S. Additional copies of the books can be obtained from:

Lattice Press
Post Office Box 340-V
Sunset Beach, CA 90742

For your convenience an order form is provided on the final leaf of the book.

CHAPTER 1

THE ROLE OF PROCESS AND DEVICE MODELS IN MICROELECTRONICS TECHNOLOGY

Experimentation through trial-and-error (*empiricism*) is one way of advancing technology. Progress in such traditional materials-based industries as glass and steel relied on this approach until the early 20th century. However, as evidenced by the centuries-long evolution of these technologies, empirically driven progress is very slow. One reason is that new discoveries are codified merely as 'recipes' (i.e., as a set of explicit instructions for achieving a specific outcome). Unfortunately, recipes give no indication of the outcome if the variables of the procedure are altered. Thus, with empiricism alone, neither empirical results nor the 'recipes' point the way toward further advances.

The rate of technological progress can be radically increased by combining empirical investigations with scientific modeling. In this approach, theoretical behavior is inferred from experimental data, and models of the phenomena or systems being investigated are developed from these theories. Such models not only provide an understanding of the observed phenomena, but also offer a means for predicting behavior under conditions not yet experimentally attempted. Since scientific modeling has been one of the foundations of the rapid progress in microelectronics, we need to expand upon the role of such models, and how they are derived.

1.1 MODELS IN MICROELECTRONIC TECHNOLOGY

A *model*, in the technological sense, is *a schematic description or analogy of a phenomenon or system, that accounts for its known or inferred properties, and may be used for further study of its properties.* Whenever possible, models used in microelectronics are *physically based* (i.e., the phenomenon being modeled is represented by some well-understood physical effect). However, if the physics of the phenomenon being studied is unknown it may be necessary to resort to an *empirically based model.** In such cases, the relationship between the variables of the phenomena are experiment-

* The term *model* for an empirically derived description is probably not really appropriate. Maybe the term *empirical functional relationship* would be more apt. However, since the term *empirical model* is used frequently in the literature, we will use it as well.

ally determined. The quantitative empirical model in this case is often a closed-form mathematical expression fitted to the experimental data.

Models can contribute to technological progress in two ways:

1. Physically based models can provide a basic explanation or visualization of the phenomenon, device, or system under study, even if the effect being modeled it is not directly observable.

2. Both physically or empirically based models can serve as vehicles for *simulating* process or device behavior. That is, the behavior of a real process or device can be examined by studying the characteristics or operation of the analogous system (model). However, the simulation task is carried out by conducting 'paper' (or more recently 'computer') experiments. This can be comparatively less expensive and faster than performing actual experiments. Although, while simulation obviously does not replace fabrication, it does reduce the trial and error needed to converge to a stable process. Moreover, it allows exploration of the process window to help find an optimum device structure or set of process conditions. The major challenge is to develop models that can quickly, inexpensively, and accurately simulate actual semiconductor process and device phenomena.

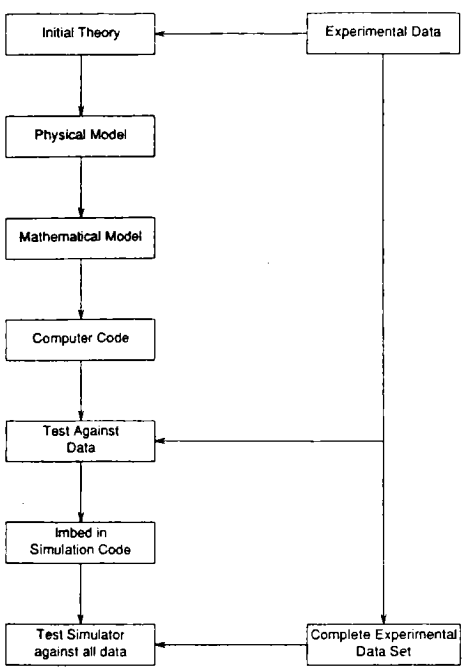


Fig. 1-1 Algorithm for creating new process models and embedding them in the process simulation code.[1] This paper was originally presented at the Spring 1989 Meeting of the Electrochemical Society, Inc. held in Los Angeles.

1.1.1 Development of Physically Based Models

Generally, physically based models are developed in three stages (Fig. 1-1). First, a concept of the mechanisms and relations which capture the essence of the observed phenomena must be deduced (this results in a *qualitative* model). In this form, the model can provide insight or even a visualization of the phenomenon or device being studied. This is especially important in microelectronics, where the phenomena being investigated are usually not directly observable. Two examples of such *qualitative models* are: (1) the depiction of the Bohr model of the atom as a nucleus surrounded by electrons occupying well-defined orbits (Fig. 1-2a); and (2) the description of drift current in a semiconductor bar as the motion of discrete particles of electric charge, e.g., electrons and holes, moving in response to an applied electric field (Fig. 1-2b).

In the second stage, the qualitative model has to be translated into a set of equations or computational operations (resulting in a *quantitative* model). To be useful for simulation applications, such quantitative forms of the model must be available. Examples of mathematical expressions that quantitatively depict some aspects of the two models above are: (1) the equation expressing the energy of an electron when it occupies one of

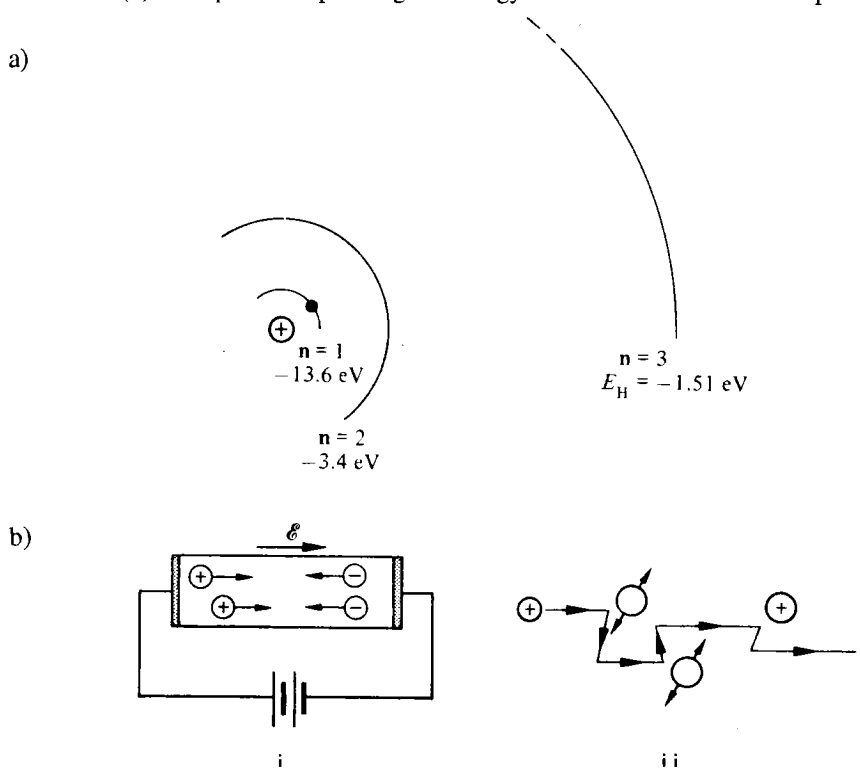


Fig. 1-2 (a) Depiction of the Bohr model of the atom as a nucleus surrounded by electrons occupying well-defined orbits; (b) Visualization of electric current (drift): (*i*) motion of carriers in a biased semiconductor bar; (*ii*) drifting hole on a microscopic scale.

the orbits in the model of the Bohr atom; and (2) the relationship between the electron drift current density, $J_{n\text{drift}}$ and the electric field $\mathcal{E}$, electron density n, electric charge per electron q, and electron mobility in a solid μ_n, or $J_{n\text{drift}} = qn\mu_n\mathcal{E}$.

Finally, the solution of the equations constituting the *quantitative* form of the model must be found and the results compared to data from actual experiments (to ensure that the simulation correctly emulates the observed phenomena).

In microelectronics, quantitative physical models frequently take the form of differential equations. Differential equations are classified according to the number of variables involved. If only one independent variable exists, it is known as an 'ordinary' differential equation. If the differential equation is expressed in terms of two or more independent variables, it is called a 'partial' differential equation (PDE). Many of the phenomena encountered in semiconductor device behavior or process technology are modeled by such *partial differential equations.*

It should also be mentioned that the task of simulating submicron device and process phenomena involves more than merely evaluating the model equations or carrying out computational operations. That is, simulating such behavior has become a complex discipline involving computer science, numerical analysis, and computer graphics (as described in chap. 2).

In addition, we should also note that simulation can play a role in aiding the manufacture of devices in several other ways. First, it can improve the probability of a first pass success for new products and processes. Second, the manufacturability of existing processes can be improved through *process centering* (i.e., finding the mix of process conditions that yields not only the correct result, but also the smallest variation from that result in the face of changes in the process conditions). In this same vein, by simulating the variation of device behavior versus changes in manufacturing conditions, the allowable limits of process variation at that point can also be identified.

1.1.2 Development of Empirically Based Models

1.1.2.1 Empirically Based *versus* Physically Based Models.

We noted that *physically based* models are developed by linking the behavior under study to an analogous physical phenomenon for which a quantitative relationship is known. For example, the motion of an electron in a vacuum subjected to an electric field can be modeled as the acceleration of a particle of mass m_e which is subjected to a force $q\mathcal{E}$ due to the electric field. Such models, when based on sound physical principles, are preferred over *empirically based* models (which provide quantitative predictive capability based only on empirically derived data). That is, physically based models are more likely to produce useful simulations even when the experimental conditions being simulated are outside of the range of conditions explored in actual experiments. In addition, when there are no generally available techniques for accurately measuring the data being simulated (the condition that exists, for example, in the case of two-dimensional doping profiles in submicron devices), obtaining simulation results from a model based on a physical foundation is the only avenue outside of empiricism for helping to design a new process or device structure.

Generally speaking, semiconductor *device* behavior can be more accurately simulated than the phenomena of *IC fabrication processes*. The physics of semiconductor devices are relatively well understood, allowing robust physically based models to be correctly implemented.* On the other hand, the physics of many IC processes is still not well understood, and physically based models of such processes may not be available. Hence, process simulation is more likely to be forced to rely upon empirical models.

Another point should be mentioned. The physically based models often rely on experimentally determined values of some parameter in the model equation. For example, the mobilities of the mobile charge carriers in the device equations (μ_n and μ_p), and the intrinsic diffusion coefficient D (in the Fick's 2nd Law equation) are experimentally determined. Although theoretical work is continuing to calculate these parameters from first principles, the simulation programs in current use depend on experimentally determined values of these parameters.

1.1.2.2 Deriving Empirically Based Models. As noted earlier, quantitative *empirically based models* are merely representations of experimental data, and have little, or no physical foundations. That is, to establish such models, data from previously conducted experiments is used to create such models in one of the following two forms:

1. The empirical results are assembled and stored in a *data base* in a computer. Sometimes the data is organized in tabular form, such as a *look-up table*. Tabular quantities however, are the least useful form for using data for simulation. For example, in such form no information is provided about how to carry out accurate interpolation should it be necessary to approximate an unknown value between two data points.

2. A closed-form equation relating the set of experimental parameters can be developed from the experimental data by curve fitting a mathematical function to a set of experimental results. That is, experimental results are plotted using an appropriate graphical format, and a closed-form mathematical expression is chosen to be a function that represents this plotted data. Curve fitting techniques are employed to find the parameter values which make the selected function best fit the experimental data points.

We should remark that there are two basic approaches for fitting data with an approximating function. The first involves passing an assumed polynomial through every data point. This approach is not used because an nth-order polynomial for $n + 1$ data points is too complex for large n values. In addition, experimental data are subject to errors, and thus passing a polynomial though every point is not appropriate.

* It should be noted, however, that obtaining a solution of the set of device equations is generally a complex procedure. Thus, while the path toward accurate device simulations is will marked, it may still be difficult to attain the goal in specific cases.

The second method is the standard approach of curve fitting data. The given data set is plotted and a mathematical function is selected to represent the plotted points.* The chosen function is fitted to the data set by having it pass as close as possible (but not necessarily through), every data point. The most popular method for finding the parameters of the assumed function that gives this best fit is known as the *least-squares method* (for it involves approximating a function such that the sum of the squares of the differences between the curve and the data points given is a minimum). There are other methods of curve fitting, all of which are denoted by the general term *regression analysis*. Figure 1-3 depicts a graphical interpretation of the least squares method.

If an attempt is made to fit a straight line of the form $y = a + bx$ to a set of data, the procedure for finding the most probable values for the coefficients a and b (i.e., the *regression coefficients*) is known as *linear regression*. If the assumed functional approximation is nonlinear, a *nonlinear regression* procedure must be used to find the regression coefficients.

1.1.2.3 Role of Empirical Models in Semiconductor Simulation. Despite the fact that they are not physically based, empirical models can still be very useful. First, there may be no other option available for simulating phenomena for which the physical basis is not yet understood. Second, when embedded as part of a process or device simulation program, empirical models can serve as convenient storehouses for experimental data. Third, obtaining simulation results from models of this type is usually quick and straightforward. Fourth, empirical models may be able to provide an exact simulation for some particular experimental conditions. Finally, even if the conditions being simulated fall between the actual experimental data points, interpolation is still likely to give reasonably accurate results.

However, since physical principles do not underlie the quantitative expressions which constitute these models, their main limitation is that the results they provide cannot be extrapolated to conditions outside of the ranges of previously conducted experiments. Hence, these models cannot be reliably used to characterize new processes beyond these ranges of process conditions.

It should also be emphasized that many, if not most, models used in semiconductor process and device simulation are, in fact, *semi-empirical* models. That is, where possible, phenomena are modeled with physically based equations and parameters relating to these phenomena. However, some effects are not well understood, or their parameter values cannot yet be derived from physical principles. Hence, empirical parameters are used in these instances. As an example, the MOSFET dc circuit models used in SPICE (described in chaps. 4 and 5) contain empirically determined parameters in the expressions they use to model the effects of carrier mobility degradation, carrier velocity saturation, and drain-induced barrier lowering on the drain current.

* For example, the plotted data is examined for obvious trends such as linear, quadratic, or higher order behavior. If logarithmic or exponential behavior is suspected, the data may be plotted on a semilog and/or log-log scale. Based on the trends observed, one of these functional forms is chosen as the function to be fitted to the data.

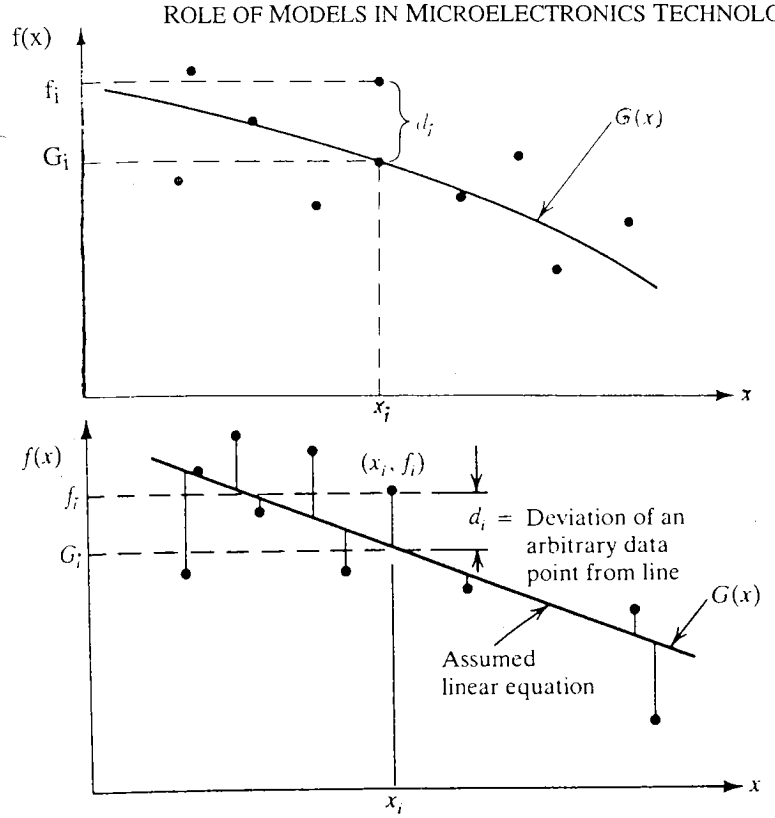


Fig. 1-3 (a) Graphical interpretation of the least-squares method. (b) Schematic representation of linear regression.

1.1.2.4 Examples of Quantitative Empirically Based Models. Some examples of empirically based semiconductor process models are the following:

1. When silicon is oxidized in dry O_2 to grow SiO_2 at thicknesses less than 350Å, the growth is 'anomalous' (i.e., it does not obey the conventional Deal-Grove model). By analyzing a vast amount of data, Reisman-Nicollian were able to derive an empirical model for this thickness regime.[2] Their thin-oxide-growth power-law model predicts that the oxide thickness t_{ox} is given by a general power law of the form, $t_{ox} = a\,(t_g + \tau)^b$, where a and b are constants, and t_g is the time for growth measured in a given experiment (Fig. 1-4a, see chap. 7). This result is contrasted with the Deal and Grove model, where oxide thickness is the solution of a quadratic equation.

2. When boron is ion implanted in single crystal silicon, channeling effects are observed (i.e., the penetration of some boron ions is deeper than in non-crystalline silicon). In an attempt to model these effects, an exponential tail of fixed decay length (independent of dose energy or crystal orientation) can be added

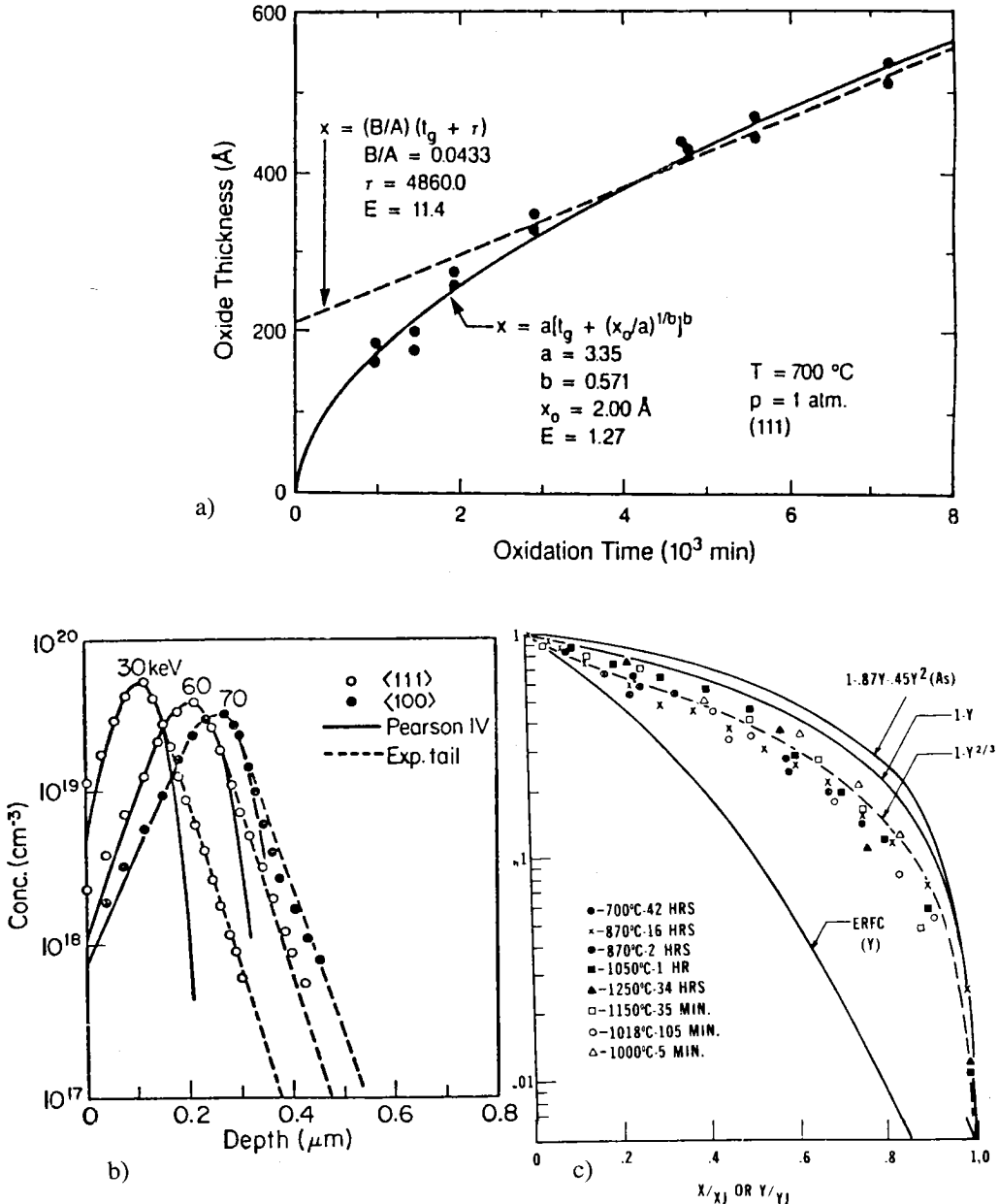


Fig. 1-4 (a) Fits to oxide thickness versus oxidation time using the Deal-Grove model and the power-law model expression.[2] Reprinted with permission of the Journal of Electronic Materials. (b) Boron as-implanted profiles in <111> and <100> silicon in a random direction. Pearson IV and modified Pearson IV distributions for representation of these profiles.[3] (© 1983 IEEE). (c) Boron diffusion profiles normalized to C_S and x_j.[4] (© North-Holland).

to the Pearson-IV distribution function curves used to simulate boron implantation processes into non-crystalline silicon (Fig. 1-4b). The decay length of this exponential tail is empirically determined to be 450Å. This empirical model is an option available in the SUPREM III program.[3]

3. The extent of boron or arsenic diffusion in silicon under high boron concentration (*extrinsic*) conditions can be simulated with a physically based model. However, this requires obtaining the solution of Fick's 2nd Law equation in nonlinear form by numerical analysis (see chap. 2). In some cases, one may wish to estimate extrinsic diffusion processes without having to resort to numerical analysis. A closed-form analytical solution that still yields a reasonably accurate result would be useful for such circumstances. Closed-form solutions to the one-dimensional diffusion equation, expressed as a polynomial have been empirically derived by fitting a polynomial expression to experimental data for extrinsic diffusion of B and As into Si (Fig. 1-4c).[4]

1.1.2.5 PREDICT-1: A Process Simulator Based on Empirical Models. Although virtually all process simulators (e.g., SUPREM-III, SUPREM-IV, FEDSS) use empirically based models to some degree, the process simulator PREDICT-1[5] has been organized to deliberately exploit the benefits of empirical models. First, the empirical models it employs are derived from a large data base of experimental results. Extensive analysis of sufficient data is conducted to ensure that valid model equations are extracted. Continued verification of the modeling software is performed with each addition to the data set. Second, since universal empirical models which can simulate all conditions usually do not exist, PREDICT-1 employs a decision tree architecture (Fig. 1-5). That is, when a particular process is to be simulated, the decision tree structure of the program selects the empirical model that has been shown to be valid over the range of processing parameters being simulated. Since the models can be separated as tree branches, it allows for their verification independently of other models.

1.2 MATHEMATICAL FORMULATION OF THE PHYSICALLY-BASED MODELS OF SEMICONDUCTOR PROCESSES AND DEVICES

Despite the fact that empirical models play a useful role in simulating semiconductor processes and devices, they are chiefly employed when physically based models are either too complex or not available. In fact, a major ongoing theme of semiconductor research is to develop new physical models to replace empirical models. We also noted that physically based models are used to provide insight about process physics and device behavior. Since the main purpose of our books is to discuss the physics of submicron devices and processing, the remainder of our discussion will be devoted to

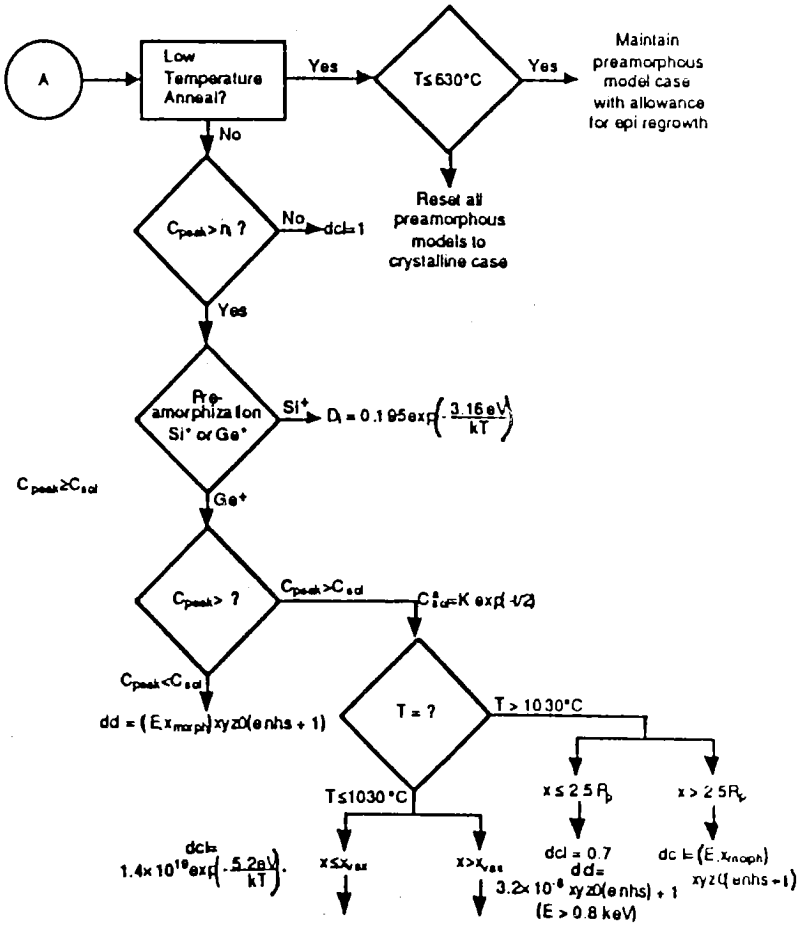


Fig. 1-5 Example of the "tree architecture" used in the PREDICT-1 simulator.[5] (© 1988 IEEE).

such physically based models. In this chapter, the emphasis will be on the mathematical forms such models assume, and the methods used to treat these mathematical relations.

As stated earlier, many of the fundamental equations utilized to provide quantitative relationships for the models in IC technology are *partial differential equations* (PDEs). Obviously, in order to use these equations for simulation purposes it must be possible to solve them. While finding the solution of such PDEs generally requires complex numerical procedures that must be carried out with the aid of a computer (as described in chap. 2), under some circumstances less elaborate methods can provide solutions. Specifically, by using suitable approximations, the form of the PDEs can be simplified to a degree that closed-form analytic solutions exist. Such analytic solutions are very attractive since they allow a particular regime of operation to be continuously explored (in a mathematical sense), and because they can provide solutions without costly and time-consuming computer calculations.

When such simplifying approximations are employed, however, the range of applicability of the solutions is reduced to a limited set of process conditions or device structures. However, it turns out that such simplified equation forms were adequate for semiconductor devices fabricated down to about the 2 μm device generations. On the other hand, to model the physics of smaller devices it became necessary to use the PDEs in their multidimensional forms. Coincidentally, at about the time the simplified forms of the model equations had to be abandoned, high-speed computing hardware became widely available. This was fortuitous, since such computing power was needed to implement the numerical procedures for solving the equations in their more complex form. This transition away from the simplified equations (and the resultant closed-form analytic solutions) and the implications this has on the approach that must then be used to simulate devices and processes should be described in more detail.

The remaining sections of this chapter will address this topic. We will proceed by first introducing the two PDEs in their general form, and then go on to answer the following questions:

1. How can the general forms of the model equations be simplified so that their closed-form solutions can be obtained?
2. Under what circumstances is it valid to invoke the approximations used to simplify the equations?
3. Why are the closed-form analytical solutions of the model equations so attractive, and what important roles do they still play in studying and simulating ICs?
4. Why do such simplified forms eventually have to be abandoned in order to retain the ability to accurately simulate submicron process and device behavior?

The answers to these questions will provide the background for the material of chap. 2. That is, the discussion of this chapter should make it clear why numerical procedures are necessary for simulating submicron process and device behavior.

1.2.1 The Physically Based Model of Impurity Diffusion in Silicon - Fick's Second Law Equation

Thermally driven *diffusion** of impurities in a semiconductor is one of the key steps in fabricating integrated circuits. A detailed description of the physics of dopant diffusion in silicon is presented in Vol. 1. Here, however, we merely introduce the mathematical formulation of the physical model of diffusion, namely, the PDE known as Fick's 2nd Law equation. It is presented at this point because we will use it as one of the two examples for illustrating the analytical and numerical methods needed to obtain solutions of such PDEs.

Assume that a wafer of silicon contains a known, nonuniform distribution of impurities (e.g., B, P, or As) at some initial time. It is observed that measurable

* *Diffusion* is defined as the net tendency of objects in random motion to move from regions of high concentration to low concentration.

redistribution of the impurities occurs as a result of heating the sample above about 600°C. How can this phenomenon be modeled?

The *qualitative physical model* of diffusion assumes that at high temperatures the impurities undergo random movement in the silicon lattice, and that such movement causes the impurities to redistribute themselves from regions of higher to lower concentration. In *quantitative* form, the model postulates that the extent of such redistribution can be found by solving Fick's 1st and 2nd Laws of Diffusion.

Specifically, the *quantitative model of diffusion* states that if the concentration of impurities (e.g., number/cm^{-3}) is known at all locations in a wafer at some initial time (the initial condition), and we wish to predict the concentration at a specific location at some later time t_1, this can be achieved by solving Fick's 2nd Law Equation for the concentration at that location and time. The Fick's 2nd Law equation, in its general three-dimensional form is

$$\partial C/\partial t = -\operatorname{div} F_D \qquad (1-1)$$

where the flux of impurities F_D according to Fick's first law is

$$F_D = -D \nabla C \qquad (1-2)$$

In these equations C is the impurity concentration in the silicon, and D is the concentration-dependent diffusion coefficient.

To be able to obtain specific solutions of the Fick's 2nd Law equation, in addition to the initial condition, we must also know: (1) the concentration values at the boundaries of the wafer during the time of the diffusion (the boundary conditions); and (2) the value of the diffusion coefficient D at each location and time in the wafer during the diffusion. Information about the value of D can be obtained from another model, which may be physically or empirically based.**

If the above information is known and a method is available for solving Eq. 1-1, then the simulated value of the impurity concentration can be obtained by applying the solution method to Eq. 1-1.

1.2.2 The Physically Based Model of Semiconductor Device Behavior

The goal of device simulation is to accurately predict the electrical behavior of semiconductor devices. For the purposes of circuit analysis and design, such behavior is generally expressed in terms of the variables of circuit theory (namely, the terminal currents and terminal voltages of the device). Examples of such circuit-related device behavior are: (1) the relationship between the current and voltage in a *pn* diode (one model of which is the *ideal diode equation*); and (2) the static output characteristics of bipolar transistor and MOSFET devices. In MOSFETs such output characteristics are

** In practice, as noted earlier, empirical values of D are used in such process simulators as SUPREM III. There is an attempt being made to develop physical based models of D.

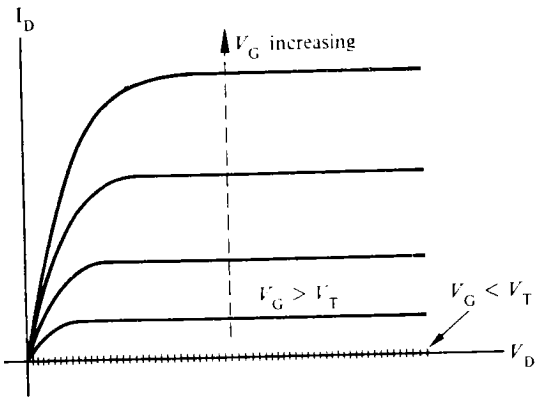


Fig. 1-6 Output I-V characteristics (I_D versus V_{DS}) of a MOSFET with V_{GS} as the parameter.

expressed as the dependence of the drain current I_D on such terminal voltages as V_{GS} and V_{DS}. Figure 1-6 is an example of a set of experimental I_D versus V_{DS} curves for a MOSFET (with V_{GS} as the parameter). An appropriate device model should be able to accurately simulate such a set of experimental I-V output characteristics.

The physically based model of semiconductor devices assumes that electrical device behavior, in general, can be simulated by solving a set of five fundamental, nonlinear, partial-differential equations (often referred to as the *classical semiconductor equations*). These constitute a complete set of equations which describe the carrier, current, and field distributions, and are written in general form as follows:

1. The Poisson equation: $$-\nabla^2\varphi = \nabla\mathcal{E} = q(n - p + N_D - N_A)/\kappa_{Si}\varepsilon_0 \qquad (1-3)$$

which describes the electrostatic potential (and the electric field) within the device. (Here φ is the electrostatic potential, $\mathcal{E}$ is the electric field, and N_D and N_A are the donor and acceptor concentrations.)

2. & 3. The continuity equations for electrons and holes;

$$\partial n/\partial t = (\nabla J_n/q) - R \qquad (1-4a)$$

$$\partial p/\partial t = (\nabla J_p/q) - R \qquad (1-4b)$$

where J_n and J_p are the electron and hole current densities, respectively, and R is the net recombination rate of the carrier being considered.

4. & 5. The equations for the electron and hole current densities;

$$J_n = qn\,\mu_n\,\mathcal{E} + q\,D_n\,\nabla n \qquad (1-5a)$$

$$J_p = q\, p\, \mu_p\, \mathcal{E} - q\, D_p\, \nabla p \qquad (1 - 5b)$$

In the first term on the right side of Eqs. 1-5a and b the current is proportional to the electric field and thus represents the drift of carriers with the electric field. In the second term the current is proportional to the gradient of the carrier concentration, e.g., ∇n. Hence, it represents the diffusion of carriers from regions of high to low concentration. Equations 1-5 are referred to as the *drift-diffusion current equations.*

This set of equations can be solved for dc or time-domain (transient) conditions. It is generally agreed that device modeling programs can do an excellent job of simulating device characteristics provided accurate information is available on the doping concentration and the device structure.

Some of the parameters of these five equations must be obtained by solving additional auxiliary equations, including the equations that provide the mobile charge concentrations n and p in the device and the equations that yield R. For example, the equations that give the concentration of mobile charges at *equilibrium*, n_o and p_o (assuming Boltzmann statistics) are most frequently expressed in terms of the intrinsic carrier concentration n_i, the *Fermi energy* (also called the *Fermi level*) E_F, and the *intrinsic Fermi level* E_i, as follows:

$$n_o = n_i \exp\,(E_F - E_i)/kT \quad \text{and} \quad p_o = n_i \exp\,(E_i - E_F)/kT \qquad (1 - 6)$$

However, since Eq. 1-3 is expressed in terms of the electrostatic potential φ instead of the energy E, it is useful to express Eqs. 1-6 in terms of φ and the Fermi potential φ_F instead of E_i and E_F. This can be done by noting that the electrostatic potential φ is related to the potential energy by $-q\varphi = E$ (see chap. 3 for more discussion on this). In addition, the electric field $\mathcal{E}$ is defined as the negative gradient of the electrostatic potential (written in one-dimension as $\mathcal{E} = -\,d\varphi/dx$), and $\mathcal{E}$ can also be expressed in terms of energy as represented using the energy-band representation, $\mathcal{E} = (1/q)(dE_i/dx)$. Thus, these two terms can be equated, as $-\,d\varphi/dx = (1/q)(dE_i/dx)$. This means we can define φ in terms of E_i, as $\varphi = -E_i/q$, and the *Fermi potential* φ_F in terms of E_F, as $\varphi_F = -E_F/q$. This allows us to express Eqs. 1-6 in another form, namely:

$$n_o = n_i \exp\,(q[\varphi - \varphi_F]/kT) \quad \text{and} \quad p_o = n_i \exp\,(q[\varphi_F - \varphi]/kT) \qquad (1 - 7)$$

However, under *nonequilibrium conditions* the Fermi level parameter is not meaningful. Thus, neither Eqs. 1-6 nor 1-7 are valid in such cases. Nevertheless, it is still possible to express n and p under conditions of nonequilibrium by defining two quantities (E_{Fn} and E_{Fp}) to replace the Fermi level in Eqs. 1-6, and another two quantities (φ_{Fn} and φ_{Fp}), to replace the Fermi potential in Eqs. 1-7. By doing this we can write:

$$n = n_i \exp\,[E_{Fn} - E_i]/kT = n_i \exp\,(q[\varphi - \varphi_{Fn}]/kT)$$

and $\qquad (1 - 8)$

$$p = n_i \exp\,[E_i - E_{Fp}]/kT = n_i \exp\,(q[\varphi_{Fp} - \varphi]/kT)$$

where E_{Fn} and E_{Fp} are called the *quasi-Fermi energies* for electrons and holes, respectively, and φ_{Fn} and φ_{Fp} are the corresponding *quasi-Fermi potentials*.

The term R in Eqs. 1-4a and 1-4b must be calculated by another auxiliary equation. In such indirect bandgap semiconductors as Si and Ge the recombination process is calculated using an equation derived by Shockley, Read, and Hall (and the process itself is frequently referred to as SRH recombination). This expression is

$$R = \frac{pn - n_i^2}{\tau_p (n + n_i) + \tau_n (p + n_i)} \qquad (1 - 9)$$

where τ_n and τ_p are the electron and hole lifetimes (which are independent of doping concentration at low doping levels, but decrease at high doping concentrations). Empirical expressions for τ_p and τ_n are used in process simulation programs for these parameters. At low doping, the values most commonly used for τ_n and τ_p in Si are 3.95×10^{-5} sec and 3.52×10^{-5} sec, respectively.

Note that for the case of low-level injection (i.e., when the excess minority carrier concentration, $\Delta n_p = n - n_o$ or $\Delta p_n = p - p_o$, is much less than the concentration of the dominant equilibrium species), then Eq. 1-9 is reduced for *p*-type material to

$$R \sim \Delta n_p / \tau_n \qquad (1 - 9a)$$

and for *n*-type material to

$$R \sim \Delta p_n / \tau_p \qquad (1 - 9b)$$

Equations 1-9a and 1-9b state that for low-level injection conditions the net recombination rate is proportional to the excess minority carrier concentration.*

1.2.3 The Poisson Equation

Of the five classical semiconductor device equations, we will focus on the Poisson equation (Eq. 1-3), where φ is the electrostatic potential, q the electronic charge, $\kappa_{Si}\varepsilon_o$ the permittivity of Si, N_D the donor doping concentration, N_A the acceptor doping concentration, n the electron concentration and p the hole concentration. We single out the Poisson equation for several reasons: (1) it is a fundamental physical equation, being essentially the third Maxwell's equation; (2) it relates the electrostatic potential and electric field to the charge density in a device; (3) as noted above, it is one of the 5 coupled equations that must be solved to simulate general semiconductor device operation; and (4) in it's two-dimensional form, it is a second-order, non-linear, time-

* The current density equations (Eqs. 1-5a & b) can be derived from the even more fundamental *Boltzmann transport equation (BTE)*. However, the classical hole-current-density and electron-current-density equations listed above have been adequate for most devices down to the mid-submicron range, and have been the work-horse model in nearly all of industry for device analysis of conventional FET and bipolar structures down to about 0.5 μm. Nevertheless, in order to simulate devices with minimum feature sizes smaller than 0.5 μm (and especially as device dimensions approach 0.25 μm and smaller),[6] it may be necessary to invoke the full BTE instead. This topic is beyond the scope of our text.

invariant, partial-differential equation. This makes it another useful example of a PDE that can be solved with numerical analysis techniques - in this case involving only spatial discretization.*

1.3 CLOSED-FORM ANALYTICAL SOLUTIONS *VERSUS* NUMERICAL SOLUTIONS OF THE PDEs THAT ARE THE PHYSICALLY BASED MODELS OF MICROELECTRONICS

Assume we have a differential equation that represents the quantitative physically based model of some phenomenon in a device or process. To obtain quantitative predictive results from the model it must be possible to solve the equation. Generally, two approaches are available for obtaining a solution: (1) using analytic techniques (i.e., techniques from calculus) to find a *closed-form analytical solution;* and (2) finding the solution using *numerical analysis.* Let us compare the solutions obtained using each of these approaches, first from a mathematical point of view, and then from a perspective of their usefulness in solving differential equations for simulation applications.

1.3.1 Closed-Form Analytical Solutions - Mathematical Aspects

A closed-form analytical solution of a differential equation is determined using *analytic* techniques, such as separation of variables followed by integration.** An analytic solution may consist of one or more functions from elementary calculus (including rational algebraic, trigonometric, exponential, or logarithmic functions). It may also take the form of an infinite series. In either case, when an analytic solution is substituted into the differential equation, an identity results. The analytic solution also makes it possible to find the *exact value* of the dependent variable for a given value of the independent variable. Finally, if the analytical solution consists of a continuous function over the range of interest, it can be used to provide a continuous set of dependent variable values.

1.3.2 Numerical Solutions - Mathematical Aspects

A numerical solution of a differential equation is obtained using numerical techniques instead of employing the techniques of calculus. For example, when an integration is called for in an analytic procedure used for obtaining the solution of a differential equation, a numerical approximation to the integration step may be used instead. Or, the differentials of the differential equation being solved may be replaced with finite difference expressions. Upon making such changes to the differential equation, the

* Note that Maxwell's third equation, $\nabla D = \rho$ becomes the Poisson equation if the permittivity is constant throughout the device region.

** *Analytic* is defined as: using, subjected to, or capable of being subjected to a methodology of calculus.

resulting equation can be solved using only algebraic techniques. Solutions obtained using such approaches give an *approximate* value of the dependent variable for a given value of the independent variable(s). In addition, values of the dependent variable are only provided for the discrete values of the independent variable chosen. However, if the error of the approximate value of the numerical solution is zero, the analytical and numerical solutions are identical. In chap. 2 we describe the numerical techniques used to model the PDEs of semiconductor devices and processes.

1.3.3 Comparing the Analytical and Numerical Solutions in Semiconductor Simulation Applications

Now that we have compared the two solution types from a mathematical perspective we can comment on their roles in simulation applications. First, based on the previous discussion, it is obvious that if an analytical solution of a PDE can be discovered, and this solution is found to give good agreement with experimental observations of the device or process phenomenon being simulated, the analytic solution of the *PDE in that form* will provide the most accurate simulation. However, if approximations had to be made to reduce the PDE to its simplified form so that an analytic solution could be found, and these approximations do not realistically reflect the actual conditions of the problem being considered, then the analytic solution may not provide the most accurate simulation of that problem. In that case, solutions obtained using numerical analysis offer the possibility of providing more accurate predictive capability.

That is, we stated that the numerical solution is an approximation of the analytical solution. But, if approximations with respect to actual device or process phenomena had to be made to simplify the PDE so that an analytic solution could be obtained, the analytic solution in that case *is also in fact approximate.* When the PDE is posed in a form which invokes fewer of these approximations, it more realistically mirrors the actual conditions in the device or process being simulated. In that case, it may no longer be possible to find an analytic solution of the PDE. To solve the equation in that form, numerical analysis is the only option. It is true that to obtain sufficient accuracy, extensive computation may be needed, but it will be possible for a solution to be obtained. For this reason, in simulation applications that require high accuracy in the submicron era, the closed-form analytic models are tending to be replaced by numerically obtained simulations - that is by models which can more accurately represent realistically complex problems: e.g., nonlinear, time-dependent conditions. In the last sections of this chapter we indicate under what circumstances the analytic solutions become inadequate for the most accurate device and process simulations.

1.3.4 Role of Analytical Solutions in Submicron Semiconductor Technology

Do not be misled from the above discussion that numerically obtained solutions will eliminate the need for analytic solutions in semiconductor technology. That is, analytic solutions still play many important roles in both visualization and simulation applications. First, analytic solutions can normally be obtained with much less

computation than can numerical solutions. (In the extreme solving three-dimensional forms of the semiconductor PDEs may even require the use of a supercomputer to obtain a numerical solution. Even the numerical solution of coupled two-dimensional PDEs can be a very computationally intensive task.) Thus, where it is possible to find analytic solutions that provide reasonably accurate predictive capability, they will be used. Several examples include:

1. Use of engineering (analytic based) models to predict I-V behavior of devices in circuit analysis simulators (SPICE). When simulating the transient response of digital circuits, the device currents and capacitances between various nodes for a given bias condition must be carried out thousands of times. Thus, a model that is computationally efficient, yet sufficiently accurate, must be available so that such a simulation can be quickly and inexpensively performed. Simple analytical models resulting from applying various approximations to the basic semiconductor device equations are widely used to calculate I_D to a first order. Higher order effects are then accounted for through the introduction of physical and empirical parameters into these models. These model parameters are optimally extracted from the data in a curve fitting fashion as described in section 1.1.1.2. Hence, such models are referred to as semiempirical analytical models. We will discuss the topic of the I_D models for MOSFETs in more detail in chaps. 3, 4, and 5.

2. Use of the analytical models to select device and process conditions in which it is known that such models provide good agreement with measured data. For example, such solutions may provide a reasonably accurate first order approximation of the doping profiles for some diffusion processes (e.g., a low-dose implant and subsequent drive-in). Such low-concentration profiles may be encountered in such applications as channel and field implants.

3. During the exploratory phase of new process or device development it is generally useful to examine the problem from the simplest and most intuitive point of view. In such cases, accurate simulation is not as important as short computation times, as the latter will permit a larger experimental space to be explored. If analytical models are available, they can be used for this purpose. Such models may yield both an initial estimate of the magnitude of the process parameters to be used and a simple picture of the physics behind the technique. Such insight may also help to design a numerical analysis approach that is computationally efficient. For example, initial simulations in one-dimension may be adequate to predict trends rather than exact values when various options are being examined. At that point the use of complex 2-D simulation programs could be a waste of time and computer resources. The use of more sophisticated tools might only be needed after the feasibility of an idea has been demonstrated.

4. Analytic solutions can also provide insight into the mechanisms being simulated that cannot be easily gained from examining numerical solutions. That is, by inspecting the mathematical function that represents the analytical model, trends about the phenomena being modeled can often easily be visualized. While

some of these insights may be quantitatively inaccurate, they are usually qualitatively correct. On the other hand, numerical analysis solves the differential equation that governs the phenomenon being examined one point at a time. Thus, it may not be easy to use such a solution to visualize trends for cases which have not been numerically treated.

In this sense analytic solutions can serve as a sort of bridge between the purely qualitative models and the more accurate simulation results obtained from numerical analysis. This is especially important when topics as complex as diffusion in semiconductors are first being introduced to students. For example, the Gaussian function as a representative of the diffusion profile after drive-in (and the concepts of $\sqrt{Dt}$ and the dependence of D on temperature), help develop a picture of the evolution of a doping profile with time and temperature T during a diffusion process. It also gives a rough visualization of the fact that the surface concentration decreases with time, and that the profile broadens.

It should also be noted that new analytic models can also be developed with the aid of numerical analysis techniques. Let us illustrate this procedure by considering the development of an analytical model of the threshold-voltage shift in a MOSFET as a function of decreasing channel length.

To begin, we recall that numerical analysis can be used to obtain solutions of the Poisson equation in two-dimensions. Such solutions may be employed to calculate the change in threshold voltage of short-channel devices. Comparisons of the numerically simulated results with experimentally measured shifts in V_T are then used to confirm the accuracy of the numerical simulation. This establishes the numerical simulation-based solutions as benchmarks for calculating V_T changes in short channel MOSFETs.

With this quantitative benchmark in hand, simplifications of the Poisson equation can be sought which will allow closed-form analytic solutions to be obtained. The analytic solutions of such simplified forms of the Poisson equation are then used to calculate the shift in threshold voltage as L is decreased. These results are compared to the results obtained from the numerical analysis to establish the range of parameters over which the analytic model is likely to yield accurate estimates of the V_T shift. If good agreement over some range of values is found, the analytic model developed becomes a computationally efficient vehicle for simulating the change V_T shift in short channel devices over that range of channel lengths and bias voltages.

The development of analytic solutions in this manner is described in more detail in chap. 5, where a variety of analytical solutions for such short-channel MOSFET effects as V_T roll-off (and roll-up), drain-induced barrier lowering (DIBL), gate-electric-field-induced carrier mobility degradation, and channel length modulation are presented.

1.3.5 Analytic Solutions of Fick's 2nd Law Equation

If the Fick's 2nd Law equation is simplified to a one-dimensional form, and the diffusion coefficient is constant everywhere throughout the device region during the time of the diffusion process, the more general form of the equation (Eq. 1-1) is reduced to the following form:

$$\partial C(x,t)/\partial t = D\,\partial^2 C(x,t)/\partial x^2 \qquad (1-10)$$

The equation in this form is known as the one-dimensional *simple* Fick's 2nd Law equation, and we observe this is a *linear* partial-differential equation. Analytical solutions are available for the simple Fick's 2nd Law equation under certain boundary and initial conditions. Two of these closely approximate the conditions that exist in two important types of diffusion processes that are performed during IC fabrication.

These two boundary conditions that permit a closed-form solution of the simple Fick's equation to be obtained are: (1) the impurity concentration at the surface of the Si (i.e., $x = 0$) during diffusion is constant at a value C_S, and the impurity concentration at large distances from the surface (i.e., $x = \infty$) is zero; and (2) the total concentration of impurities in the Si remains constant during the diffusion, and the impurity concentration at large distances from the surface (i.e., $x = \infty$) is again zero. The initial condition at $t = 0$ for both cases is the same; namely $C(x, 0) = 0$ (which states that the dopant concentration in the host semiconductor is initially zero). The solution for these two cases which satisfies both the initial and boundary conditions is:

$$C(x,t) = C_S \operatorname{erfc}[x/2\sqrt{Dt}] \qquad (1-11)$$

for the constant concentration C_S case, and

$$C(x,t) = C_S\,[\exp(-x^2/4Dt)] = (Q/\sqrt{\pi Dt})[\exp(-x^2/4Dt)] \qquad (1-12)$$

for the case of a constant dose Q per cm^2 of impurity in the silicon. The former case, in which Eq. 1-11 is called the *erfc solution* has been used to model the so-called *chemical predeposition* process, while the latter, in which Eq. 1-12 is called a *Gaussian solution*, the so-called *drive-in* step. These concentration distributions can be plotted on both linear and logarithmic scales, normalized to the surface concentration. As an example, this is done for three different diffusion times in Figs. 1-7a and 1-7b for Eq. 1-11, and in Figs. 1-8a and 1-8b for Eq. 1-12. Note that the penetration depth of the dopant increases only as the square root of the diffusion time (e.g., as shown in Fig. 1-8b, to double the depth of a given value of C(x), one must quadruple the diffusion time). It is also important to realize that in the case of the Gaussian distribution we have also introduced the time variable into our other normalizing parameter, the surface concentration C_S.

We should also note that both the erfc Eq. 1-11 and the Gaussian distribution Eq. 1-12 are functions of a normalized distance $x/2\sqrt{Dt}$. Thus, if we also normalize the dopant concentration with the surface concentration, each distribution can be represented with a single curve valid for all diffusion times. Such normalized impurity concentrations are shown in Fig. 1-9a on a linear axis, and in Fig. 1-9b on a log axis. The combination of parameters $2\sqrt{Dt}$ used to normalize the *x*-axis of Figs. 1-9a and 1-9b represents the characteristic diffusion length associated with a particular diffusion cycle and describes the depth of penetration of the dopant.

The closed-form solutions given by Eqs. 1-11 and 1-12 were widely used to model dopant diffusion in silicon devices during the 1960's. At that time bipolar device

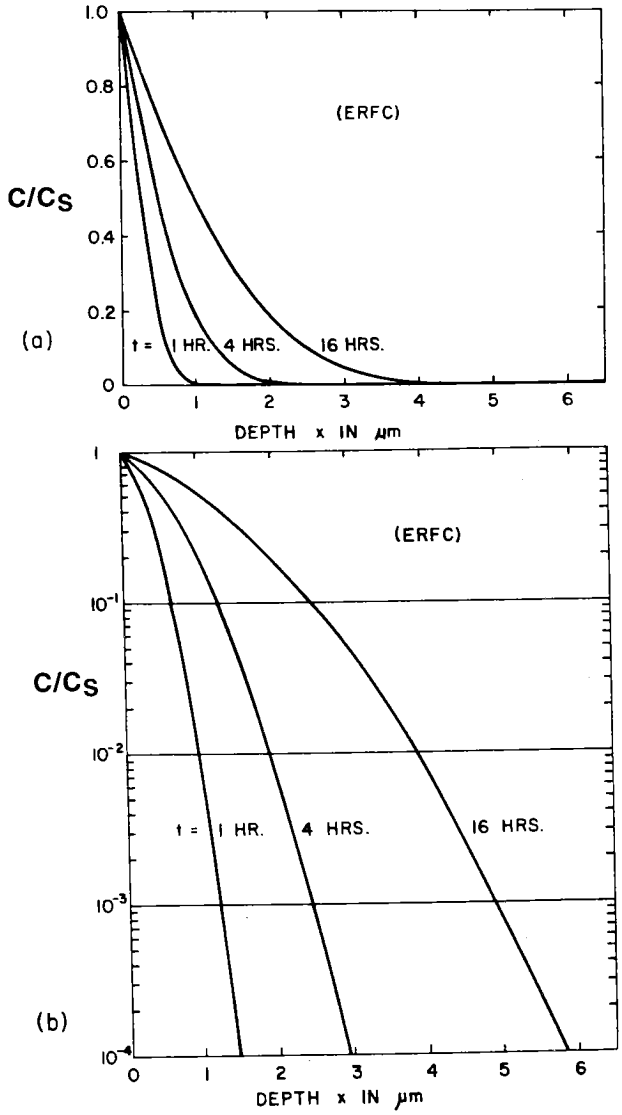


Fig. 1-7 Normalized impurity-concentration erfc profiles versus distance into the substrate for three different diffusion times (a) Linear presentation. (b) Log-linear presentation.

technology was still dominant and was being used to fabricate medium-scale-integration (MSI) ICs.* The bipolar devices were fabricated with a double-diffused process, and the analytical solutions of the 1-D simple Fick's 2nd Law equation given above were used to predict the process dependencies of such double-diffused profiles. For these early

* MOS technology had not yet overcome its isolation and threshold control issues, and for systems designed with MSI parts, the superior off-chip drive capabilities of the bipolar transistor provided critical system leverage.

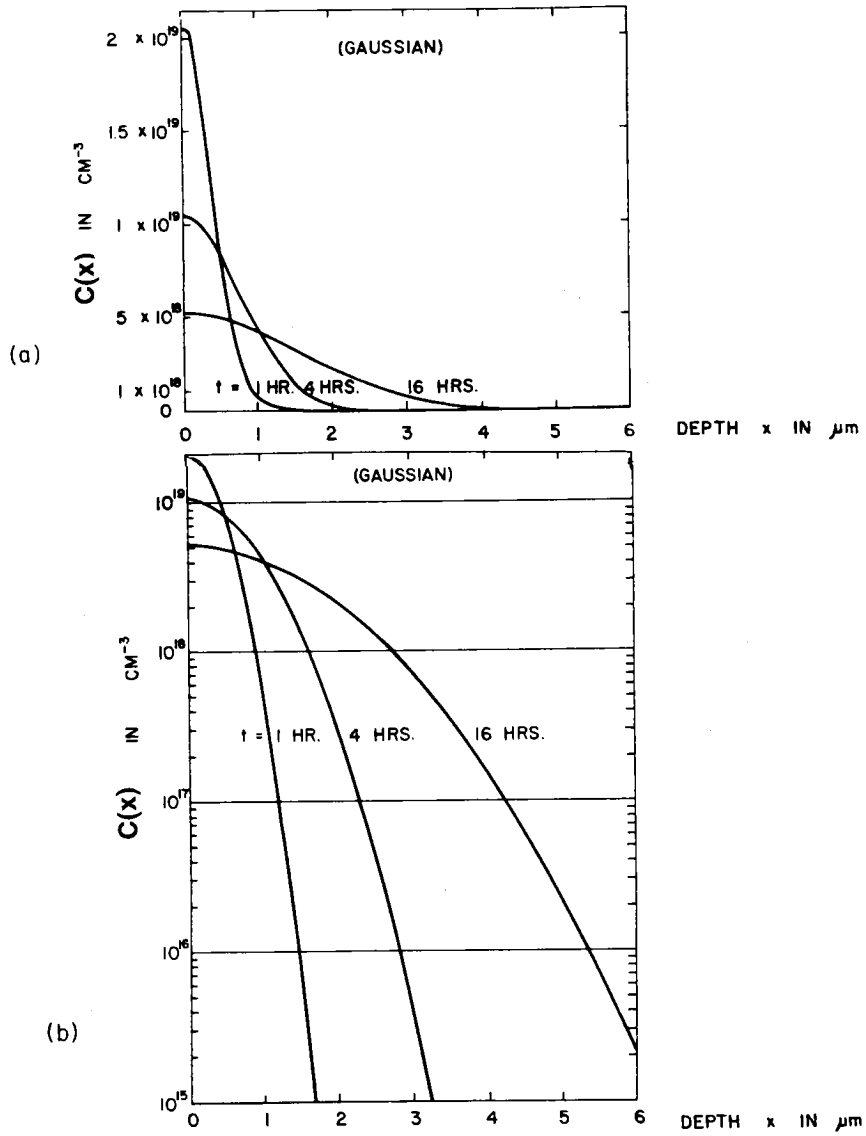


Fig. 1-8 Gaussian impurity-concentration erfc profiles versus distance into the substrate for three different diffusion times (a) Linear presentation. (b) Log-linear presentation.

bipolar device structures, the analytical solutions given by Eqs. 1-11 and 1-12 could predict the device impurity doping profiles (i.e., the one-dimensional variation of C versus depth at the conclusion of a particular diffusion process) with reasonable accuracy.**

** This was due primarily to the relatively gross feature sizes of those devices. This meant that: (1) The junction depths were sufficiently deep that the dopant surface concentrations during drive-in were reduced below the intrinsic carrier concentrations at the process

1.3.6 Analytical Solutions of the One-Dimensional Form of the Poisson Equation in the Depletion Approximation

In three- (or even two-) dimensional form, an exact analytical solution of the Poisson equation is not possible. Instead, the equation can only be approximately solved using numerical methods. Nevertheless, by making a number of assumptions (which happen to be appropriate for some important semiconductor device structures), the equation can be simplified to a degree where closed-form solutions can be obtained. This was the approach used for modeling large-dimension FET device structures in the early days of the semiconductor industry. It is, in fact, still the basis of the device equations used to

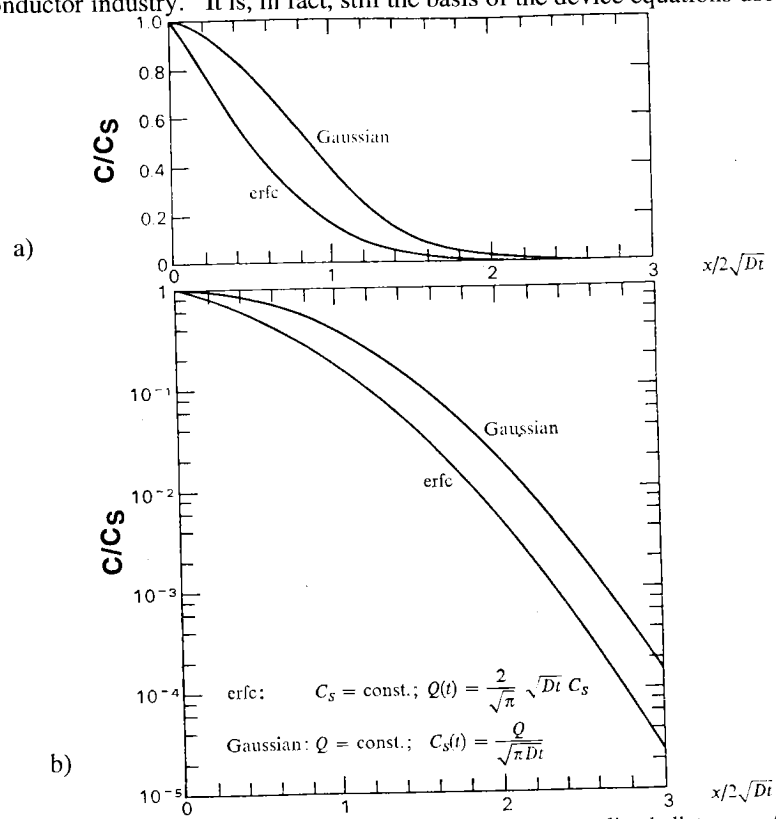


Fig. 1-9 The erfc and Gaussian impurity distributions versus normalized distance $x/2\sqrt{Dt}$. Note that by using this representation these distributions can be represented with a single curve valid for all diffusion times. C/C_S is displayed (a) on a linear axis, and (b) on a log axis.

* (cont.) temperatures. Thus, the approximations invoked to allow the Gaussian solution were valid during a large fraction of the diffusion time; (2) In planar transistor structures whose minimum feature sizes exceed 5 μm, two-dimensional diffusion effects occur over such a small fraction of the total device area that they have only a minor impact on measured device behavior. Hence, they can be ignored in the model.

model long-channel MOSFETs (as we shall describe in detail in chap. 4). Here let us briefly discuss the simplification procedure is it applies to case of MOS device structures.

In the ideal MOS *capacitor* structure described in chap. 3 (Fig. 3-1), the electric field is normal to the oxide surface, and hence only a one-dimensional variation in electrostatic potential φ exists (i.e., $\partial\varphi/dy$ and $\partial\varphi/\partial z = 0$). In addition, if we assume: (1) uniform doping exists in the semiconductor; (2) ($N_A >> N_D$); and (3) we invoke the depletion approximation ($n = p = 0$ in the depletion region), the Poisson equation in the semiconductor region of a MOS capacitor, under the oxide under depletion biasing becomes

$$d^2\varphi/dx^2 = q\, N_A/\kappa_{Si}\, \varepsilon_0 \qquad (1 - 13)$$

The Poisson equation in this form is an ordinary, linear differential equation that can easily be integrated twice to yield a closed-form solution of $\varphi(x)$ in the semiconductor. That is, assuming $\varphi = 0$ far below the wafer surface (e.g., at $x = \infty$),

$$\varphi\,(x) = \varphi_{surf}\,(1 - x/d)^2 \qquad (1 - 14)$$

where φ_{surf} is the value of $\varphi(x)$ at the surface ($x = 0$), and d is the depletion depth (i.e., the depth to which the mobile charge density is negligible, or where

$$d = [\kappa_{Si}\varepsilon_0(2\varphi_{surf})/kTN_A]^{1/2} \qquad (1 - 15)$$

Although Eq. 1-14 was derived here for the case of a MOS capacitor, we shall see in chap. 4 that the solution of as given by Eq. 1-10 is also used to calculate the electrostatic potential in long-channel MOSFETs, even when $V_{DS} \neq 0$. This is justified in such devices by invoking the *gradual-channel approximation* (see section 4.3.1, chap 4). When this approximation is valid the simplified MOSFET device models are able to simulate the experimentally observed drain current (I_D) values quite closely. However, as devices shrink to sizes in which they have submicron gate lengths, this approximation is no longer satisfied. In such cases if the simplified equation is used to simulate I_D, agreement with experimental results is poor, and the model must be modified to account for the discrepancy between predicted and measured behavior.

1.4 CIRCUMSTANCES UNDER WHICH THE ANALYTIC SOLUTIONS OF THESE PDES BECOME INADEQUATE FOR SIMULATION PURPOSES

1.4.1 The Inadequacy of the Analytic Solutions of Fick's 2nd Law Equation for Simulating Diffusion in Submicron Device Structures

If dopant diffusion exhibited only simple behavior, then erfc-type concentration profiles after constant-source depositions and approximately Gaussian-shaped profiles after ion implantation and diffusion would be found in silicon substrates. However, careful

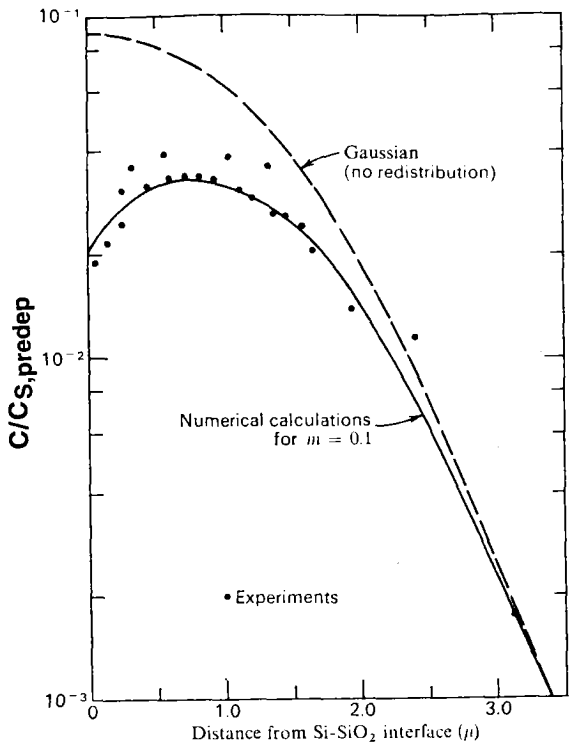


Fig. 1-10 Redistribution of diffused boron layer after thermal oxidation.[7]

comparison of experiments with the doping profiles predicted using the two classical solutions (Eqs. 1-11 and 1-12) revealed that several physical effects occurring during diffusion caused the profiles to be significantly different from those predicted by the closed-form functions. Furthermore, it became evident that such physical effects would play an increasingly greater role in the diffusion process as planar-process-fabricated device structures (both bipolar and MOS) were shrunk to the micron range and smaller. Some examples of the effects that cause the experimental doping profiles to deviate from those predicted by Eqs. 1-11 and 1-12, are:

1. If an oxide layer is grown during the drive-in step, the preference of the boron to be in SiO_2 rather than silicon gives rise to a doping profile in which the concentration of boron near the silicon surface is much lower than that predicted by the Gaussian solution (as shown in Fig. 1-10).

2. When the doping concentration is higher than the intrinsic carrier concentration at the diffusion temperature, the doping profile after predeposition does not match the erfc profile predicted by Eq.1-11, but instead exhibits a form as shown Fig. 1-10 for the case of boron diffusion. This apparently occurs because D becomes a function of the concentration ($D = f[C]$), and the simple

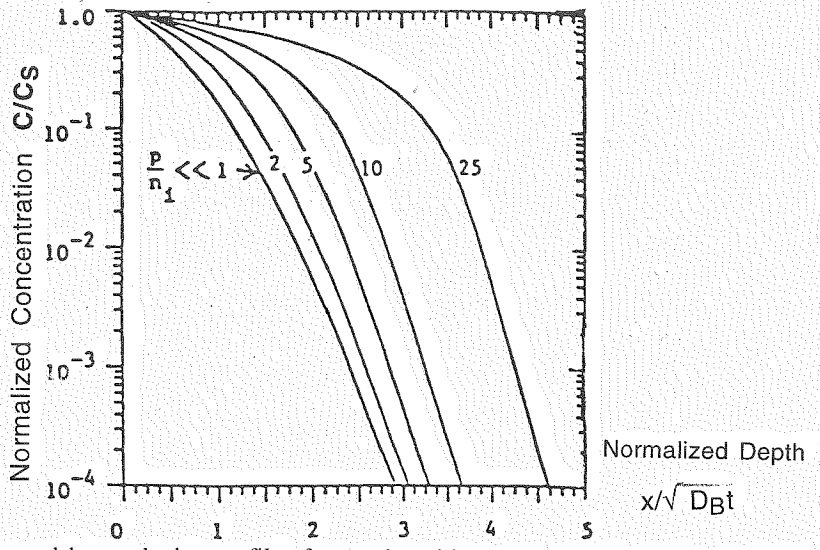


Fig. 1-11 Measured boron doping profile after predeposition step and the profile predicted using the erfc solution of the diffusion equation.[10] Reprinted with permission of the American Physical Society.

form of Fick's 2nd Law is no longer valid. Instead, the equation that must be solved is $\partial C/\partial t = \partial/\partial x\ [D(C)\ \partial C/\partial x]$, which is a non-linear PDE).

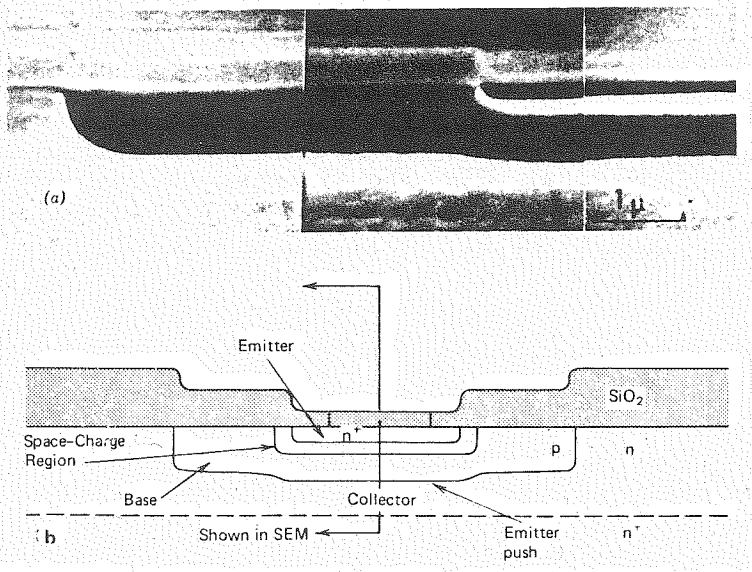


Fig. 1-12 (a) Scanning electron micrograph showing cross section through a bipolar transistor. (b) Sketch identifying the regions shown. The boron-doped base region has been pushed ahead (emitter push) by the concentration-dependent diffusion effects associated with heavy phosphorus doping in the emitter.[11] Reprinted with permission of John Wiley & Sons.

3. In the cross-section of a typical double-diffused bipolar device having a phosphorus-doped emitter (Fig. 1-12), another diffusion-related disparity is observed. The base-collector diffusion depth is deeper under the emitter than under the base regions. This indicates that the boron diffusion is affected by the high-concentration phosphorus emitter region (i.e., the boron diffusivity is now a function of the phosphorus concentration).

In device regions where these (and other) effects cause the diffusion coefficient to become position dependent, or where diffusion must be simulated as a two-dimensional effect, numerical methods with the aid of computers must be used to obtain solutions.

1.4.2 The Inadequacy of Using a One-Dimensional Form of the Poisson Equation and the Depletion Approximation to Model Submicron MOSFET Behavior.

In submicron MOSFETs the approximations used to reduce the Poisson equation to a one-dimensional form given as Eq. 1-13 are not valid, and thus Eq. 1-13 does not accurately model the electrostatic potential in submicron MOSFETs. First, in submicron MOSFETs the electric field has lateral components that cannot be ignored when a non-zero drain-source bias is applied. Second, the doping in the channel region is usually not uniform with depth below the oxide. Finally, the mobile charge concentrations in the depletion region are not zero (i.e., the depletion approximation does not hold, even when the device is biased in depletion).

Instead, the Poisson equation in at least two-dimensional form must be used to simulate φ and $\mathcal{E}$ in submicron MOSFETs. The 2-D form of the Poisson equation under the above conditions for the semiconductor region of the MOSFET is

$$\frac{\partial^2\varphi\,(x,y)}{\partial x^2} + \frac{\partial^2\varphi\,(x,y)}{\partial y^2} = \left(\frac{\rho}{\kappa_{Si}\varepsilon_o}\right) \qquad (1 - 16)$$

or

$$\frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2} = \left(\frac{-q}{\kappa_{Si}\varepsilon_o}\right)\left[-n_i e^{q(\varphi - \varphi_n)/kT} + n_i e^{q(\varphi_p - \varphi)/kT} - N_A + N_D\right] \qquad (1 - 17)$$

Note that the Poisson equation in this form relates the space-charge potential of mobile and fixed charges with mobile charge densities given by Boltzmann statistics. In this form it is a second-order, exponentially non-linear, partial-differential equation. The solution of Eq. 1-17 [i.e., φ (x,y) in the semiconductor region of the MOSFET] for an appropriate set of boundary conditions can only be found using numerical analysis. We will discuss such solution techniques of this equation in chap. 2.

REFERENCES

1. R.S. Fair, *Ext. Abs. Electrochem. Soc. Mtg,* Spring 1989, Abs. No. 141, May 1989, p. 197.
2. A. Reisman et al., *J. Electron. Mat.* **16**, p. 45, (1987)

3. C.P. Ho et al., "VLSI Process Modeling - SUPREM III," *IEEE Trans. Electron Dev.*, November 1983, p. 1438.
4. R.S. Fair, "Concentration Profiles of Diffused Dopants in Silicon,", in *Impurity Doping Processes in Silicon,* ed. F.F. Y. Wang, North-Holland, Amsterdam, 1981, chap. 7.
5. R.S. Fair, "Low-Thermal Budget Process Modeling with the PREDICT Computer Program," *IEEE Trans. Electron Dev.*, March 1988, p. 285.
6. D.C. Cole et al., "The Use of Simulation in Semiconductor Technology Development," *Solid-State Electronics,* **33**, No. 6, June 1990, p. 591.
7. T. Kato and Y. Nishi, *Jap. Jnl. Appl. Phys.*, **3**, 377 (1964).
8. R. Dutton, "Modeling Silicon Integrated Circuits," *IEEE Trans. Electron Dev.*, **ED-30**, September 1983, p. 975.
9. E.S. Meieran and T.I. Kamins, *Solid-State Electronics*, **16**, 545 (1973).
10. P.M. Fahey, P.B. Griffin, and J.D. Plummer, "Point Defects and Dopant Diffusion in Silicon," *Rev. Mod. Phys.*, April 1989.
11. R.S. Muller and T.I. Kamins, *Device Electronics for Integrated Circuits,* 2nd Ed., New York, John Wiley & Sons, 1986.

PROBLEMS

1. (a) Describe the difference between physically based and empirically based models. (b) Describe the difference between an analytical solution and a numerical solution of the partial differential equation which serves as the model of some process or device phenomenon. Elaborate on the perspective that many of the so-called physically based models of process and device phenomena should more aptly be termed hybrid (empirical-physical) models, i.e., insofar as they employ both physical models and empirically determined parameter values.

2. The intrinsic diffusion coefficient for boron in silicon is 2×10^{-14} cm^2/sec at 1000°C. Use this fact and the definition of D which states that $D = a(\Delta x)^2$ [where Δx is the lattice spacing in silicon, and a is the fraction of atoms in the volume of silicon (per unit area of lattice multiplied by a thickness Δx) which will move per unit time into an adjacent volume], to estimate the rate at which boron atoms move from lattice site to lattice site in Si at 1000°C.

3 The diffusion coefficient of boron in silicon can be written as $D_0\exp(-E_a/kT)$, where D_0 is 2 cm^2/sec, E_a is 3.5 eV, and $k = 8.62\times10^{-5}$ eV/K. Using this information, verify the value of D given in Problem 1.

4. Assume that inversion in a MOS-C occurs when the surface potential is twice the value of What is the maximum depletion width at room temperature where the *p*-type Si is doped at $N_A = 10^{16}$ cm^{-2}?

5. Calculate φ (x) versus x from x = 0 to x = d at 0.1 μm intervals for the structure given in Problem 4 when a gate bias is applied that just causes inversion. Plot these values, and compare with the plot shown in Fig. 3.9.

CHAPTER 2

NUMERICAL METHODS FOR SOLVING THE PARTIAL DIFFERENTIAL EQUATIONS WHICH MODEL THE PHYSICS OF SUBMICRON DEVICES AND PROCESSES

As described in chap. 1, the process and device models employed during the early days of the semiconductor industry were either: (1) *empirically based* or; (2) *based on simplified physical equations.* That is, one-dimensional forms of the fundamental differential equations (such as the Poisson equation or the diffusion equation) were used as the *quantitative physically based models* because exact closed-form solutions of such simplified equations exist, and these analytical functions can be evaluated without the aid of computers. However, it was also pointed out that such closed-form analytical solutions are limited in scope and applicability. Specifically, such elementary models become inadequate for predicting the behavior of submicron processes and devices. The simplifying approximations that permit the partial differential equations (PDEs) to be reduced to one-dimensional, linear, differential-equation forms (for which closed-form solutions exist) no longer match the actual conditions that exist in submicron devices.

Instead, to obtain accurate predictive capability in the submicron era it has become necessary to solve the PDEs in their nonlinear form in two or three dimensions. The only way to obtain the solutions of such PDEs is to use numerical analysis[1-4,20] - performed with the aid of high-speed digital computers, interactive computer graphics, and advanced discrete mathematical analysis methods. Unlike closed-form solutions, numerical solutions are not limited by the complexity of the differential equations or by the physical model they represent.

Some may still claim that numerical solutions of differential equations are inaccurate because they are merely an approximation to the exact solution. However - as was pointed out in chap. 1 - this view neglects to mention that the so-called exact solutions are generally constrained by various assumptions and simplifications. On the other hand, the accuracy obtained when using numerical methods is limited only by the power of the computer being used. While some numerical simulation applications require the use of supercomputers, simulation packages have been developed in recent years which

are capable of giving quite reasonable results on the more powerful desk-top computers and workstations.

The advantage of using numerical methods lies in their versatility and simplicity. Thus, it is now possible to tackle the most complex of engineering problems for which solutions were nowhere in sight even a few decades ago. It is likely that the semiconductor industry will rely ever more heavily on this type of simulation as the real cost of computing power continues to decrease.

Despite the fact that numerical analysis of process and device equations has become a basic methodology of research and development engineers, most textbooks on IC fabrication and device physics still focus their presentations on the closed-form models mentioned above. We have noted that this approach is no longer sufficient, and that results obtainable only from numerical solutions must now be used to explain process and device behavior. From this perspective it is important that the reader has some insight into the methods used in numerical analysis. We have therefore decided to include this chapter at the outset of the book as an introduction to the topic.

We also recognize that the users of the computer simulation programs are generally not specialists in most of the aspects involved with the procedure of numerical analysis. This is not surprising, considering the complexity of the assumptions, algorithms, and implementations of the programs they use. So, to impart some flavor of numerical analysis without overwhelming the reader with mathematical complexity, the discussion will be kept at a very basic level. However, a list of references is provided for those who wish to probe deeper.

The Synergy Between Submicron Devices & the Capability of Simulating their Behavior. The mathematical methods used to discretize the differential equations and to solve the large number of simultaneous discrete equations that result, require the manipulation of massive amounts of data. Hence, in practice, numerical analysis can only be carried out with the aid of high-speed computer hardware as well as software packages that have the ability to: (1) represent the structures of the devices in the computer; (2) discretize the appropriate differential equations that describe the process or device behavior; (3) solve the resulting set of discretized algebraic equations; and (4) display the solution results in a manner that allows the user to best interpret them (e.g., in tabular, graphical, contour plot, or 3-D-perspective form). Furthermore, such software packages can only be used effectively in device and process design by making the numerical computations transparent to the device or process engineer - while still allowing any interaction with the data. Currently available simulation programs leave the engineer free to focus on design issues instead of the numerical algorithms. Usually, the engineer need only understand of the physics of device or process to apply the simulation programs accurately. Nevertheless, by acquiring some knowledge of the numerical methods used in such simulation software packages, one becomes a much more astute user of such software.

If used properly, simulation can help in cutting development costs. One opinion is that as much as 40% cost saving is possible.[2] However, the cost savings depends considerably on the skill of the persons using the simulation programs. That is, to

obtain the maximum cost savings, experience and insight are still important. This is because the following tasks must be carried out by the user of the programs: (1) correctly specifying the relevant regions to be simulated in the complicated device structures; (2) selecting the appropriate material parameters and models; (3) varying appropriate process and device operating conditions to obtain optimum device performance; and (4) ensuring that accurate numerical answers are produced. If these tasks are not competently performed, the results of the simulation may either be less accurate, or the costs will be excessive. In addition, a basic understanding of the physics incorporated in the simulation program is helpful to the user in understanding the limitations of computer program. One of the goals of this text is provide such an understanding of the process and device physics.

The ability to accurately simulate submicron processes and behavior is based on a fascinating synergistic linkage between the advances in simulation of semiconductor devices and processes and improvements in computer power. That is, these two endeavors influence and complement each other. Simulating the fabrication and behavior of semiconductor devices aids in developing smaller and faster devices, which in turn, improves available computational resources. Greater computing power means more accurate simulations can be constructed, which are presently needed as devices shrink to submicron regimes, thereby requiring new physical effects to be taken into account.

2.1 NUMERICAL SOLUTIONS OF DIFFERENTIAL EQUATIONS

As we noted in chap. 1, the solution of some differential equations can be expressed in the form of analytical closed-form expressions. An example is Eq. 1-10, which is a solution of the one-dimensional Poisson equation of the form of Eq. 1-9. On the other hand, there are many differential equations for which no solution exists as a simple closed-form expression. A method for obtaining an approximate solution of such equations is through a numerical procedure. For example, for the following differential equation there is no simple formula (in terms of elementary functions)

$$\frac{dy}{dx} = (x^2 - y^2) \qquad (2 - 1)$$

However, one way to obtain a solution is to replace the differential equation by the *difference equation*

$$\Delta y = (x^2 - y^2)\,\Delta x \qquad (2 - 2)$$

For the solution starting at (0,2) (i.e., which represents a specific boundary condition for this equation) - that is, $y = 2$ for $x = 0$ - we select a Δx, say 0.1, and compute $\Delta y = (0 - 4)(0.1) = -0.4$, so that our next point is (0.1, 1.6). Here, with the same Δx, we find $\Delta y = (0.01 - 2.56)(0.1) = -0.255$, so that our next point is (0.2, 1.345). Thus, the value of y at each selected point x can be approximately computed using this method. The results of such a process are shown in Table 2-1 and Fig. 2-1.

SOLUTION OF $y' = x^2 - y^2$ WITH $\Delta x = 0.1$

x	y	x^2	y^2	$x^2 - y^2$	$\Delta y = (x^2 - y^2)\Delta x$
0	2	0	4	−4	−0.4
0.1	1.6	0.01	2.56	−2.55	−0.255
0.2	1.345	0.04	1.81	−1.77	−0.177
0.3	1.168	0.09	1.36	−1.27	−0.127
0.4	1.041	0.16	1.08	−0.92	−0.092
0.5	0.949	–	–	–	–

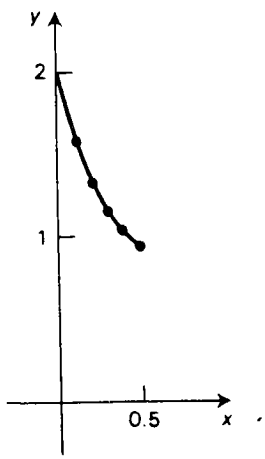


Fig. 2-1 Solutions of $dy/dx = x^2 - y^2$ found numerically, as in Table 2-1.

The simple numerical procedure illustrated above (i.e., finding the particular solution of a differential equation by replacing it with a difference equation, and then solving the difference equation numerically) is known as the *Euler method* of numerically solving differential equations. The Euler method, or some variant of it, is widely applied. To improve the accuracy of the approximate solutions obtained, one can decrease the size of Δx. The procedure can also be generalized to higher order differential equations. We will describe how similar numerical procedures can be used to solve the PDEs that are the physically based models of IC processes and devices.

2.2 NUMERICAL METHODS FOR SOLVING NONLINEAR PARTIAL DIFFERENTIAL EQUATIONS

Process and device simulators calculate values of solution variables at discrete locations within a domain. If these variables also change with time the solution at each location must also be determined at each discrete time interval. In process simulators the con-

centrations of dopants at each location and time step are usually the quantities being sought, whereas in device simulators it is the electrostatic potential φ and the mobile carrier concentrations n and p that are the quantities being determined. The values of φ, n, and p may then be used to calculate the currents flowing in the devices. The discrete points within the domain being simulated form the nodes of a mesh. Relationships between the values at discrete points are established by discretization procedures. The nonlinear coupling between equations is handled by an iteration strategy that involves solving sequences of linear subproblems. As a result, the numerical performance of a device simulator is determined by the combination of the characteristics of the mesh, the discretization procedure, the nonlinear iteration strategy, and the linear system solvers.

The overall approach to solving nonlinear PDEs using numerical analysis is as follows:

1. An appropriate grid for the device domain under consideration is specified, i.e., the domain is divided up into a sufficiently large number of discrete regions. If the PDE is also time dependent, the time interval of the process or device effect being analyzed is also broken up into a number of sufficiently small time increments.

2. The differential equation is transformed into an algebraic equation which is an approximation of the original differential equation. This allows a discrete algebraic equation to be written for each grid location or discrete region.

3. The resulting set of discrete equations is organized into a standard matrix format suitable for computer manipulation to solve the equations.

4. If these discrete equations are nonlinear, they must first be linearized.

5. The linearized set of equations are then solved, using linear matrix solution techniques.

6. If a time-dependent PDE is being solved, the time is advanced by an incremental step after the linearized equations have been solved for the previous time step. The PDE is then solved again at all the grid locations for the new time step, using the previous solution to find the solution during the next time interval. At the end of the last time increment, the variable values at each grid location represent the solution of the PDE at that time.

7. The final solutions of the PDE at each of the grid locations is finally processed into a form that is either: (a) suitable for visualization by the user; or (b) suitable as the input to the next process step or device event in the simulation, or as input into another simulator.

The balance of the chapter is devoted to describing each of these steps in more detail.

2.3 GRID GENERATION AND REFINEMENT PROCEDURES

The numerical solution of a PDE over some domain (i.e., the region chosen to represent the device geometry) requires that this domain be partitioned into a number of discrete subdomains. They should be small enough so that a sufficiently accurate approximation to the exact solution can be easily established over each subdomain. In the *finite difference* method, the simulation domain in two dimensions (typically a transistor cross section) is normally subdivided into regions defined by a *mesh* (or *grid*). In the simplest cases, this consists of a set of lines parallel to the coordinate axes. That is, in two-dimensions the simplest grid is a *rectangular grid.** The PDEs are then solved only at specific, discrete locations (e.g., in two-dimensions only at the points where the lines meet), and these locations are referred to as the *nodes* (or *mesh points*).

For example, if we put NX lines parallel to the y-axis and NY lines parallel to the x-axis we get NX*NY intersection points in the domain on which an approximate solution of the PDEs is sought (e.g., if NX = 20 and NY = 30, we seek solutions at 600 points). The first and last lines in both the horizontal and vertical directions coincide with the boundaries. Figure 2-2 shows a uniform finite-difference mesh superimposed over a rectangular simulation domain. As we shall see the grid spacing may also be non-uniform in the x- and y-directions.

In the *finite element* method the domain is also subdivided into smaller regions, called *elements.* The elements need not be rectangular, and in fact the most common shape is the triangle. The unknown parameter is approximated in each finite element with a smoothly varying function, usually an interpolation polynomial.

To obtain accurate approximations using numerical analysis, the grid spacing (or finite element area) must by sufficiently small. However, if too small a grid spacing is used, computation time is unnecessarily increased. Because of the simplicity of the algorithms, many of the early simulation programs used *uniform* meshes (in which the

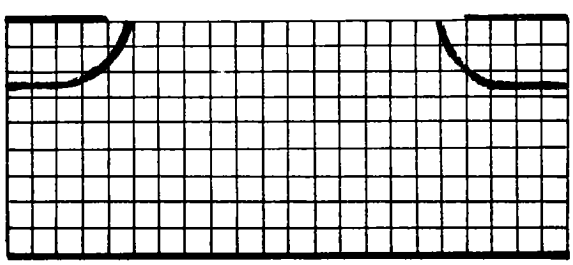


Fig. 2-2 (a) Uniform, rectangular, 2-D finite difference grid.

* The finite difference method is most easily applied to a rectangular grid, but can also be implemented on other (e.g., triangular) grids.

spacing is the same for all points). However, this benefit was gained at the expense of excessive computation time (because the smallest grid spacing needed in the entire domain had to be used everywhere). As a result, more sophisticated grid structures have been developed. The evolution of 2-D grid structures is depicted in Fig. 2-3.

The first improvement over the basic uniformly-spaced rectangular grid with a planar top surface (Fig. 2-3a) was the implementation of *variable* (or *nonuniform*) grid spacing. For example, when the potential in a MOSFET is found by solving the Poisson equation, it is observed that the variation of potential depends strongly on the local charge density. This implies that device regions where a large dopant-concentration gradient exists requires many grid points for accuracy. By using variable grid spacing many calculation points can be placed in regions of strongly varying doping concentrations (i.e., by employing a fine grid there). Conversely, fewer points can be used in less critical regions by specifying a coarse grid in such regions. A simple automatic grid generation scheme used in some second-generation MOSFET device simulators shows how non-uniform grid spacing can be exploited. That is, a fine grid is used at the silicon surface (with the minimum spacing being specified by the user) and the grid spacing is automatically increased geometrically toward the substrate. In the horizontal direction, fine grids are used in the channel and source/drain junction regions and a coarse grid in the middle of the channel (Fig. 2-3b).

The grid shown in Fig. 2-3b, however, does not permit simulation of the topology of the wafer surfaces found in VLSI devices because it assumes a *planar* surface. For example, LOCOS isolation structures and trenches cause the wafer surfaces in VLSI structures to be non-planar. To permit the effects of such non-planar structures to be simulated, a rectangular grid structure such as shown in Fig. 2-3c must be employed (together with numerical techniques that allow the PDEs to be solved on such a grid).

The incorporation of a non-planar surface does not yet produce a grid structure that is optimized for computation. That is, the disadvantage of the Cartesian grid is that when a finer mesh is needed in one region of the domain, this results in a finer mesh also being established in other parts where the extra refinement may not be necessary. On the other hand, a grid based on a triangular element shape (Fig. 2-3d) allows *local grid refinement.* That is, if a triangular grid is used, it is possible to achieve local refinement without adding unnecessary nodes in other parts of the device. Figure 2-4 illustrates how triangles which require refinement can be locally subdivided into four congruent subtriangles by joining the midpoints of the sides. These subtriangles may in turn be further refined. This permits the mesh to be refined adaptively as the computation progresses so as to limit the differences in the parameter being sought between any two mesh points to be less than a pre-specified limit. Adaptive meshes increase the complexity of the algorithm and may increase the solution time required, but provide a considerable improvement in accuracy and stability. Another advantage is that the triangular elements can work easily on arbitrary grid shapes (i.e., nonplanar device structures).

Three-dimensional device simulations require a 3-D grid. At present 3-D device simulators are actually based on 2-D device simulation programs which are applied to 3-D problems by replicating the 2-D model in the third dimension. Thus, the 3-D grid

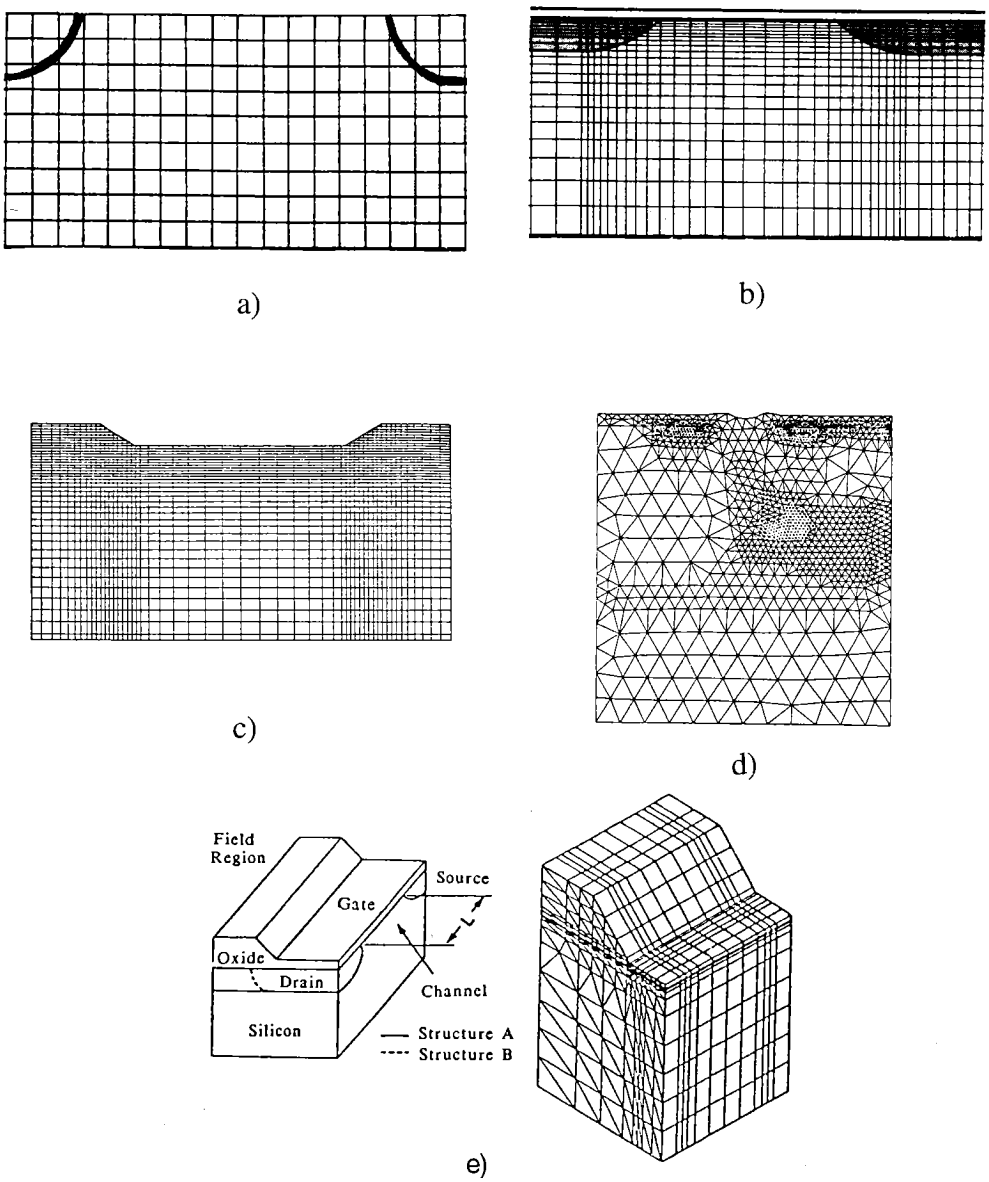


Fig. 2-3 Evolution of grids for numerical analysis of semiconductor PDEs. (a) Uniform rectangular (planar), 2-D grid. (b) Nonuniform rectangular (planar), 2-D grid. (c) Nonuniform rectangular non-planar 2-D grid. (d) Nonuniform triangular nonplanar 2-D grid. (e) Non-uniform triangular nonplanar 3-D grid.

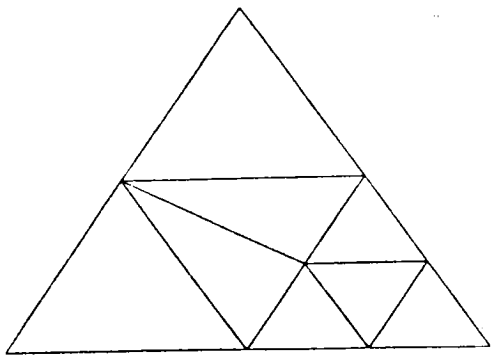


Fig. 2.4 How triangles which require refinement can be locally subdivided.

(Fig. 2-3e) is created by using a 2-D simulation program to create a 2-D triangular-based grid (as shown in Fig. 2-3d) in a single plane, and then this grid is replicated over many parallel planes to define the nodes in the third dimension. Several 3-D device simulators are commercially available, including DAVINCI™ (TMA Inc.), and HFIELDS-3D™ and THUNDER™ (Silvaco International).

The accuracy and efficiency of the discretization methods for solving PDEs can be significantly improved by using adaptive meshes. (An *adaptive mesh* is one that begins as a coarse mesh and is appropriately refined as the solution is evolved). An efficient way to implement adaptive meshes out is to provide a "self-generating" algorithm in the simulation software package. This algorithm automatically refines the grid on request by the user. The mesh is selected to minimize the discretization error and the number of mesh points required. Local refinement of the mesh may be based on the value of the truncation error or on changes in the doping concentration, potential, and carrier distributions (computed as the simulation progresses - that is, as the solution converges or as the distributions change with time). After the grid is adapted to the solution, the solution is computed again.

As an example, consider the adaptive triangular grid generator in PISCES II.[5] Starting with a list of boundary nodes specified by the user, a coarse grid is initially generated to cover the topology of the structure. This preliminary grid is then refined according to the doping profile before computing the equilibrium solution. The grid may be refined again at any time after this, using the variation of any physical variable to determine where the refinement is necessary. Note, however, that grid refinement is computationally expensive, and is thus usually updated only occasionally at the user's request. Otherwise it could completely dominate the solution time if it was invoked several times during each iteration

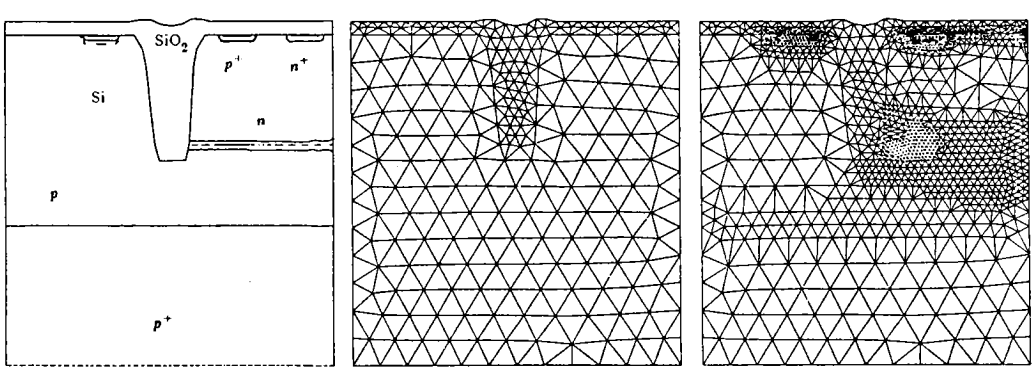
Figure 2-5a shows the cross-sectional geometry of an *n*-well CMOS device with trench isolation.[6] Figure 2-5b depicts the initial coarse triangular grid which was generated automatically by the device simulator Pisces-II. The formulation of the initial grid is based only on the device geometry. Using this initial grid, the triangles are

refined based on the input doping profile, or later, on preliminary or intermediate solutions for the potential and carrier concentrations. Fig. 2-5c corresponds to a refined grid during the solution process, obtained after several refinements using criteria on doping and potential.

2.4 DISCRETIZATION OF DIFFERENTIAL EQUATIONS

As just discussed, the numerical solution of semiconductor process and device equations requires that the simulation domain (chosen to represent the device geometry) be partitioned into a number of discrete regions which are small enough to allow the numerical method to achieve the required level of precision. The model PDEs are then approximated in each of the subdivided regions by a set of algebraic equations which involve only values of the continuous dependent variables at discrete points in the domain. This latter process is known as *space discretization of the PDEs* and results (in general), in a large set of non-linear algebraic equations with unknowns comprised of approximations of the continuous dependent variables at discrete points.

The model PDEs can also be either *time independent* (*time invariant*) or *time dependent*. In the first case, $\partial/\partial t = 0$, and hence only space derivatives appear in the PDE. In the second case, $\partial/\partial t \neq 0$, and hence the PDEs contain time derivatives as well as space derivatives. If the latter case, the time interval being simulated must be divided into an appropriate number of time steps, and the time derivatives must be discretized. That is, the relevant PDEs must be fully discretized in both time and space. Algebraic techniques are then employed to solve the set of spatially discretized equations at each time step. The time is advanced after the spatial solution is found for a given time step, and that solution is used to the find the spatial solution at the next time step. We discuss the details of space discretization in sections 2.4.1, 2.4.2 and 2.4.3, and the aspects of time discretization in section 2.4.4.



(a) (b) (c)

Fig. 2-5 (a) Cross-sectional geometry of an *n*-well CMOS device with trench isolation. (b) Initial coarse grid generated automatically by the device simulator PISCES. (c) Refined grid, based on either the input doping profile, or on preliminary solutions for the potential and carrier concentrations.[6] (© IEEE 1985).

2.4.1 Space Discretization

As noted earlier the continuous differential equations of IC process and device behavior must be transformed into discrete algebraic equations, so that they can be solved with numerical (i.e., algebraic) techniques. The two most important methods used for transforming these PDEs into discrete algebraic equations are the *finite difference method* and the *finite element method.* Either method can be used to discretize the *space derivatives* and they both have their advantages and drawbacks for this purpose. *Time derivatives*, however, are generally discretized using finite difference approximations.

Semiconductor simulation programs have been developed using both finite difference and finite element methods for space discretization. For example, in the process simulators SUPREM III, SUPREM IV and FINDPRO, finite differences are used to discretize the space derivatives of the diffusion equation. The device simulators GEMINI, MINIMOS, and PISCES-II also employ the finite difference method for space discretization. Programs using the finite element method for space discretization include FEDSS, FIELDAY, and some modules of SUPREM IV.

In our discussion we will provide more information on the finite difference method for space discretization. This is because the basis of finite difference schemes can be grasped in a rather straightforward manner, while the theory behind the finite element method is rather abstract (i.e., it is necessary to introduce concepts from functional analysis and variational methods in order to formulate the finite element algebraic equations). Since our intent is to introduce readers to the topics of numerical analysis (rather than to provide a rigorous treatment), we choose the simpler approaches to illustrate our examples whenever possible.

Nevertheless, it should be noted that the other numerical techniques being implemented in the latest simulation programs are also considerably more sophisticated than those used in the examples we present. Since the level of their mathematical complexity is far beyond the scope of our book, the ones we present allow the reader to gain a flavor of the techniques without being completely overwhelmed by the mathematics.

2.4.2. The Finite-Difference Method

There are also two ways in which the finite-difference technique is used to transform differential equations into their algebraic counterparts:

1. The *difference operator method*, in which the differential operators in the PDE are directly replaced by finite difference approximations.

2. The *box integration method*, in which the domain is divided into a set of subdomains, each one surrounding a node of the grid (see Appendix D). These subdomains partition the interior of the domain without overlap or exclusion. The finite difference approximation of the unknown (which in this case is the average value of the unknown over each subdomain) at each grid point (i.e., at each node) is obtained by integrating the PDE over the subdomain (i.e., the local area) surrounding the node. The integral for each node is then approximated to yield a discrete difference equation associated with each node.

We will provide more information on the *difference operator method* in this section because it is the simpler of the two. Nevertheless, finite-difference-based device simulators more often employ the *box integration method* (e.g., GEMINI and PISCES-II), primarily because it permits non-planar device structures to be handled more easily. In recognition of its importance, however, additional information on the box integration method is provided in Appendix D. This is intended as reference material for those who are interested the numerical procedures of box-integration-based device simulators.

In both methods, nevertheless, the resulting difference equations are applied at each node (except at the boundaries of the domain where boundary conditions specify the values of the unknowns at such boundary nodes), thereby relating the functional value at each node to those nearby. Consequently, a set of algebraic equations is developed which can be solved for the unknown functional values.

2.4.2.1 The Difference Operator Method. As noted above, the approach of the *difference operator method* to solving differential equations is to replace the continuous space derivatives of the PDEs by discretized finite-difference approximations (difference expressions). This transforms the continuous PDE into an algebraic difference equation for the function at a given node. By relating the functional value of the unknown at each node to those nearby, and using these values in the finite-difference approximation expressions, a difference equation is constructed at each node of the finite difference mesh of the domain. In this fashion, a set of algebraic equations is developed which relates the value of the unknown at each node to the values at neighboring nodes. Upon applying the proper boundary and initial values, this set can then be solved using algebraic techniques for the unknown functional values at each node. By using a rectangular grid, this finite-difference method becomes particularly well-suited for treating simple device geometries.

2.4.2.2. Taylor Series. The finite-difference expressions used to replace the differential operators in the course of discretizing a PDE can be derived from a truncated Taylor series. Since the Taylor series is also the basis of several other aspects of numerical analysis, we will briefly review the Taylor series and its properties here.

In general, the Taylor series is based on the assumption that any function can be expressed as a polynomial having an infinite number of terms. Consider the arbitrary function f(x) given in Fig. 2-6. Assume also that f(x) is a continuous, single-valued function of x with continuous derivatives, and has a known value $f(x_0)$ at $x = x_0$. If the functional value at a point $x = x_0 + \Delta x$ is derived, then we should be able to approximate $f(x_0 + \Delta x)$ by considering the slope of the function $df(x)/dx = f'(x)$ evaluated at $x = x_0$. Therefore the slope of the tangent line is easily evaluated in terms of the functional values $f(x_0)$ and $G(x_0 + \Delta x)$ to give

$$f'(x_0) = \frac{[G(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

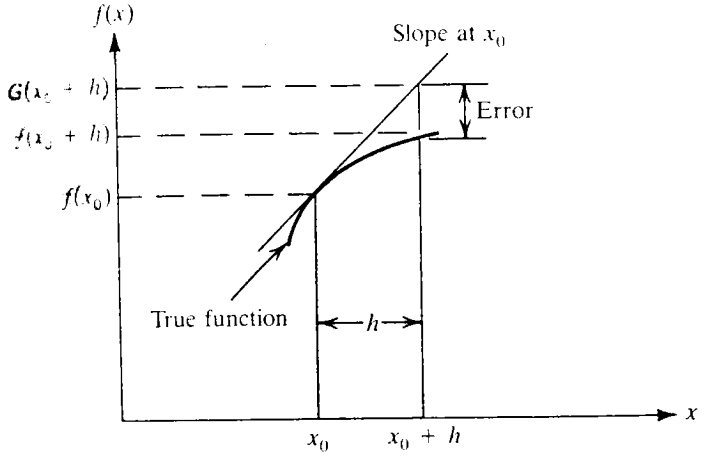


Fig. 2-6 Graphical illustration of functional approximation.

Solving for $G(x_0 + \Delta x)$ yields

$$G(x_0 + \Delta x) = \Delta x\, f'(x_0) + f(x_0). \tag{2 - 3}$$

Obviously $G(x_0 + \Delta x)$ is not exactly equal to $f(x_0 + \Delta x)$: however, it can be used to approximate its value. That is,

$$f'(x_0 + \Delta x) = G(x_0 + \Delta x) \pm \text{ERROR} \tag{2 - 4}$$

The reason that a ± sign is used is because if the true function is concave upward then the error is negative. Substituting Eq. 2-3 into Eq. 2-4 yields the following approximation:

$$f'(x_0 + \Delta x) = f(x_0) + \Delta x f'(x_0) \pm \text{ERROR} \tag{2 - 5}$$

Note that here $f'(x_0)$ is the first derivative of the function evaluated at $x = x_0$.

The Taylor series simply states that $f(x_0 + \Delta x)$ is an infinite series containing higher derivative terms. That is, the error term can be reduced to zero if the second, third, and infinite derivatives are known at x_0. Hence

$$f(x_0 + \Delta x) = f(x_0) + \Delta x f'(x_0) + \frac{\Delta x^2}{2} f''(x_0) + \ldots + \frac{\Delta x^n}{n!} f^n(x_0) + R_{n+1} \tag{2 - 6}$$

where $f'(x_0)$ is the first derivative of the function evaluated at $x = x_0$, $f''(x_0)$ is the second derivative evaluated at $x = x_0$, and so on, and

$$R_{n+1} = \frac{\Delta x^{n+1}}{(n+1)!} f^{n+1}(s), \quad x_0 \le s \le x_0 + \Delta x \tag{2 - 7}$$

Equation 2-6 is a Taylor series and Eq. 2-5 is then a truncated Taylor series. The term R_{n+1} is called the remainder term of the series.

2.4.2.3 One-Dimensional Finite Difference Approximations of Derivatives. Assuming we are still dealing with the function f(x) of the previous section, then $f(x_o + \Delta x)$ can again be expressed using 2-6. The first-order discrete approximations of the first derivative $f'(x_o)$ can be found using Eq. 2-6 by truncating the series after the second term. That is, if the Taylor series is truncated after the second term, and the remaining expression is solved, for $f'(x_o)$, we get:*

$$f'(x_o) = \frac{f(x_o + \Delta x) - f(x_o)}{\Delta x} \pm 0(\Delta x) \qquad (2 - 8)$$

where instead of using ERROR to denote the error term, we now use the symbol $0(\Delta x)$.**

If the Taylor series was expressed in terms of $f(x_o - \Delta x)$, it would be written as

$$f(x_o - \Delta x) = f(x_o) - \Delta x f'(x_o) + \frac{\Delta x^2}{2} f''(x_o) - \ldots - \frac{\Delta x^n}{n!} f^n(x_o) + R_{n+1} \qquad (2 - 9)$$

If $f'(x_o)$ was found in same manner by using Eq. 2-9 instead of Eq. 2-6, it would be expressed as

$$f'(x_o) = \frac{f(x_o - \Delta x) - f(x_o)}{\Delta x} \pm 0(\Delta x) \qquad (2 - 10)$$

The symbol $0(\Delta x)$ indicates that as Δx approaches zero, the error term becomes proportional to Δx. This further implies that more accurate solutions are obtained for smaller values of Δx.

* Note that here $f'(x_o)$ is the first derivative of the function evaluated at $x = x_o$.

** If the entire Taylor series - which contains an infinite number of terms - was used to approximate a function, the error between the approximation function and the true function would be zero. However, for practical reasons (including the cost of computation), only a few terms - generally two or three - are used in the development of the derivative approximations. When the discrete approximation of a derivative is obtained from such a truncated Taylor series, the difference between the value of the true function at x_o and that of the approximate function is known as the *error*. It is customary to express such errors in terms of the power of Δx. That is, in Eqs. 2-5 and 2-9 the error is assumed to be of the order of Δx, and is expressed as $0(\Delta x)$, whereas in Eqs. 2-11 and 2-12 the error is assumed to be of the order Δx^2 and is expressed as $0(\Delta x^2)$. This method of expressing the error allows the various forms of the difference formulas to be compared in terms of their order of error. That is, since it is almost always the case that $\Delta x \ll 1$, difference formulas having an error term of order $0(\Delta x^2)$ are more accurate than those having an error term of order $0(\Delta x)$.

The approximation to $f'(x_0)$ given by Eq. 2-8 is called a *forward-difference approximation* because the derivative is approximated by relating the base point x_0 to a point x *forward* of it by Δx ($x = x_0 + \Delta x$). On the other hand, the right-hand side of Eq.2-10 is called a *backward-difference approximation* because the derivative is approximated by relating the base point x_0 to a point x *backward* of it by Δx ($x = x_0 - \Delta x$). Note that Eqs. 3-8 and 3-10 are both first-order approximations. For the purposes of finite difference approximations we will assume that Δx is the incremental distance between mesh points. In a uniform mesh this distance is constant, but in a non-uniform mesh, Δx varies throughout the mesh.

Second-order approximations of $f'(x)$ and $f''(x)$ can also be obtained from the Taylor series. Such an expression for $f'(x)$ is found by truncating Eqs. 2-6 and 2-9 and after the third term. Then the truncated version of Eq. 2-9 is subtracted from that of Eq. 2-6, and we solve the result to yield:

$$f'(x_0) = \frac{f(x_0 + \Delta x) - f(x_0 - \Delta x)}{2\Delta x} \pm 0(\Delta x^2) \qquad (2 - 11)$$

Likewise, the second-order expression for $f''(x_0)$ is found by truncating Eqs. 2-6 and 2-9, then adding the two results, and finally solving the addend for $f''(x_0)$:

$$f''(x_0) = \frac{f(x_0 + \Delta x) - 2f(x_0) + f(x_0 - \Delta x)}{\Delta x^2} \pm 0(\Delta x^2) \qquad (2 - 12)$$

These last two expressions are also known as the *central difference* approximations of the derivatives. Note that the error term for the central difference approximations implies that the error is proportional to Δx^2, while that of the forward- and backward-difference approximations is proportional to Δx. For $\Delta x \ll 1$, this indicates that the error of the central difference approximations is smaller than that of either the forward or backward difference approximations, which in turn implies the central-difference derivative expressions are of order of Δx more accurate than either of the other two. Since they are more accurate, the central difference approximations are almost always used for discretizing the space derivatives of PDEs, with Eq. 2-11 used to replace first derivatives, and Eq. 2-12 to replace second derivatives. For PDEs containing time derivatives, however, the forward and backward difference approximations are frequently used to discretize the time derivatives, as we shall see in section 2.4.4.

2.4.2.4. Two-Dimensional Space Discretization. In a two-dimensional mesh, the distance between adjacent mesh points is often described using the following notation:

$$a_i = \Delta x = x_{i+1} - x_i \quad i = 0,1,2,3 \ldots n \qquad (2 - 13a)$$

$$b_j = \Delta y = y_{j+1} - y_j \quad j = 0,1,2,3 \ldots m \qquad (2 - 13b)$$

Using this notation for finite difference expressions in two dimensions, we get

$$f(x,y) = f_{i,j} \quad i = 0,1,2,3 \ldots n \quad j = 0,1,2,3 \ldots m \tag{2 - 14a}$$

$$f(x \pm \Delta x, y) = f_{i\pm 1}, \tag{2 - 14b}$$

$$f([x \pm \Delta x]/2, y) = f_{i\pm 1/2, j} \tag{2 - 14c}$$

$$f(x, [y \pm \Delta y]) = f_{i, j\pm 1} \tag{2 - 14d}$$

$$f(x, [y \pm \Delta y]/2) = f_{i, j\pm 1/2} \tag{2 - 14e}$$

where the relative positions of the points i,j, etc are expressed in Cartesian co-ordinates. This notation is depicted in Fig. 2-7 where the mesh points and 'half-points' are shown. The mesh points shown in Fig. 2-7 constitute the *five point difference approximation*, includes the four neighbor nodes which are horizontally and vertically adjacent to the center node.

It is often assumed that the derivatives of potential (in the Poisson equation) or impurity flux (in Fick's 2nd Law equation) are constant between mesh points. In these circumstances the central difference first derivatives at the half points (e.g., at i + 1/2,j and i,j + 1/2) are

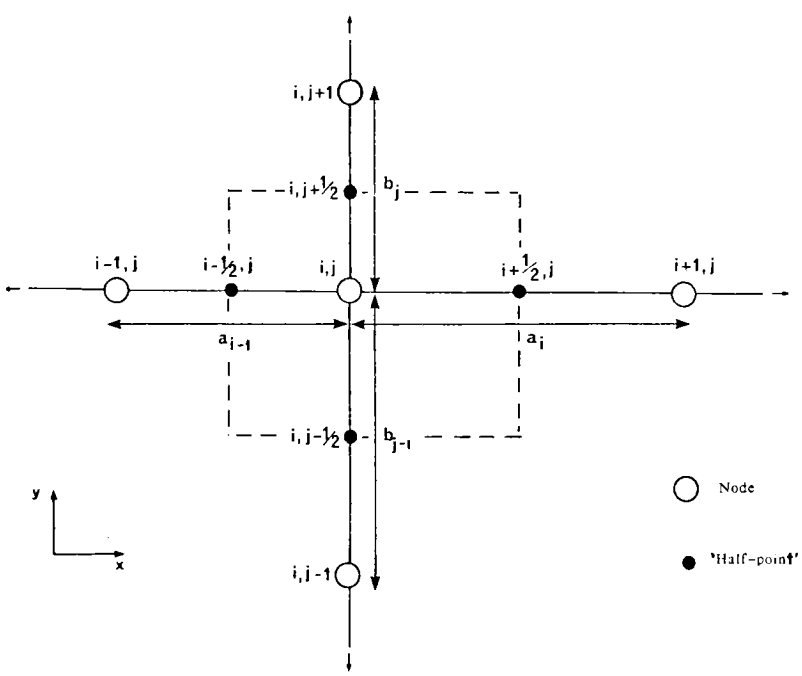


Fig. 2-7 Finite-difference mesh point notation for a rectangular grid (half-point).

$$\frac{\partial f}{\partial x}_{@\, i+\frac{1}{2}j} = \frac{f_{i+1,j} - f_{i,j}}{a_i} + 0(a_i^2) \qquad (2 - 15a)$$

$$\frac{\partial f}{\partial y}_{@\, i,j+\frac{1}{2}} = \frac{f_{i,j+1} - f_{i,j}}{b_j} + 0(b_j^2) \qquad (2 - 15b)$$

and the central difference second derivatives at these half points are

$$\frac{\partial^2 f}{\partial x^2}_{@\, i+\frac{1}{2}j} = \frac{f_{i+1,j} - 2f_{i+\frac{1}{2}j} + f_{i,j}}{a_i^2} + 0(a_i^2) \qquad (2 - 15c)$$

$$\frac{\partial^2 f}{\partial y^2}_{@\, i,j+\frac{1}{2}} = \frac{f_{i,j+1} - 2f_{i,j+\frac{1}{2}} + f_{i,j}}{b_j^2} + 0(b_j^2) \qquad (2 - 15d)$$

2.4.3 The Finite Element Method

The finite element method is the other most widely used spatial discretization procedure for numerically solving semiconductor differential equations. It is particularly useful for obtaining solutions for non-rectangular, irregularly shaped device geometries, where the inherent flexibility of the method can lead to a more efficient solution than can be obtained using finite difference methods. A major advantage of the finite element method is that there is no additional complexity introduced by using elements of different sizes. Another is that the domain can be subdivided in such a way as to increase the number of elements in regions where accurate results are desired and to reduce the overall number where the results are not so important. Hence, it is possible to reduce the total number of nodes needed for a given problem. The principal disadvantages of this method are: 1) initially more effort is required to implement the numerical algorithm; and 2) the stability and convergence criteria are not well understood at the present time.

A longer discussion of the finite element method applied to semiconductor simulation is given in Selberherr[2] and Snowden,[3] each of which devote a section 10-15 pages long to the subject. A introductory book on the application of the finite element method to electromagnetics is by Steele.[7] A more advanced and general treatment of the finite element method is given by Zienkiewicz.[8]

In the finite element method the spatial domain of interest is divided into small, basic elements. A physical object is thought of as composed of this finite number of pieces, each of which can be analyzed separately according to the laws of physics. Thus, each physical variable over the entire object can be expressed as a sum of finite elements, each element representing the contribution from one of the pieces. By combining the analyses of the separate pieces, a system of algebraic (usually linear) equations relating to the finite elements is deduced. Solution of this set of equations provides the desired approximate solution of the problem.

In IC simulation applications, the domain is the entire device geometry, and it is broken up into a mesh of fine pieces called *finite elements*, all having a standard shape (most often triangles). It is assumed that the exact solution of the differential equation in each region can be approximated by a simple, smoothly varying function over each element. Typically, interpolation polynomials (usually, first- to fourth-order polynomials) are selected to serve as the approximate solutions in each element. The approximate solutions in each element are also known as *partial solutions*. The total approximate solution for a parameter is given by the sum of the partial solutions over all the elements. For this to be possible, each approximation function is defined to be equal to zero for all finite elements except one. Therefore, when both sides of the differential equation are multiplied by the approximation functions, many matrix entries become equal to zero, allowing the resulting matrix to be solved efficiently. The accuracy of the method depends both on the complexity of the approximation functions used in each element and on the number of elements used to make the device geometry. A user may therefore increase the accuracy of the computation by using a greater number of finite elements or by using higher-order approximation functions.

The simplest, and most commonly used approximation function for a triangular element is a first-order polynomial, which, in Cartesian coordinates has the form

$$u_i^h(x,y) = Cx + By + A \qquad (2 - 16)$$

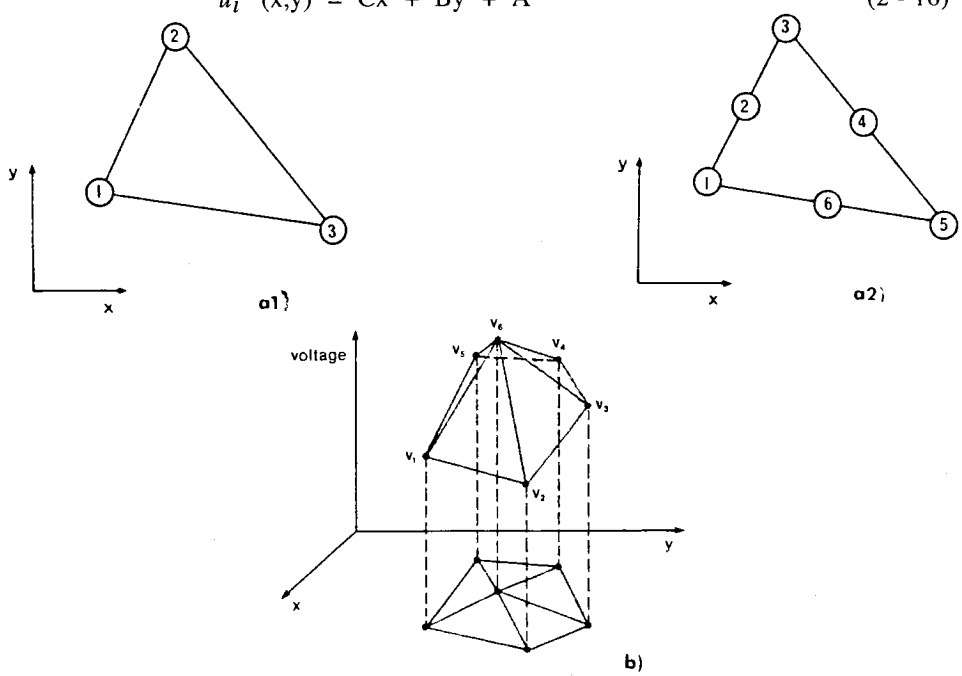


Fig. 2-8 Finite element discretization: (a) Triangular element of the finite-element grid [a1 - Linear approximation, a2 - Quadratic approximation]. (b) Solution is made continuous by setting the potential in common triangle vertices to be identical.

The coefficients of the polynomial are determined from three values of the approximate solution. In practice, these values are obtained at the nodes of the element for this linear approximation. For example, if the potential is defined over a first-order triangular finite element, the function has solution potential φ_1, φ_2, and φ_3 at the triangle vertices: the solution is interpolated linearly in between using the approximation polynomial (Fig. 2-8a). In other words, in this case the solution is approximated as a plane passing through the three nodal values of φ. As an example, in the solution surface over a mesh of five such first-order elements, the solution is made continuous by setting the potential in common triangle vertices to be identical (Fig. 2-8b).

A Galerkin matrix* is constructed for each finite element, and all these matrices are added together to create a single matrix equation for the entire device. Since the integral must be computed only over standard shapes, the computation of each separate Galerkin matrix becomes repetitive and is therefore greatly simplified. Similar types of matrix methods are used to solve the discretized equations generated using the finite element method as are used to solve the linearized equation sets resulting from the finite difference method (e.g., successive overrelaxation [SOR] methods).

2.4.4 Time Discretization

A PDE which has the time derivative on one and only one side of the equation is termed an *initial value problem*. The Fick's 2nd Law equation and the transient carrier-continuity equations (i.e., in which $\partial n/\partial t \neq 0$), are examples of such PDEs. To solve initial-value PDEs we begin by assuming that the solution is known at the starting time instant (i.e., the initial value is assumed known), and its evolution from that time on (i.e., the time derivative) can be obtained from the spatial variation.

To numerically solve such time-dependent PDEs, time derivatives must also be discretized. To do this, the total time interval of the process or device effect being simulated must first be divided up into an appropriate number of smaller time steps. The

* The *Galerkin method* is one of several methods for approximately solving differential equations based on the *calculus of variations* (others being: least squares, weighted residual, or Rayleigh-Ritz methods). When the Galerkin method is used to solve, for example the Poisson equation, the unknown potential distribution is approximated by writing it as a linear combination of known functions with unknown coefficients. This approximation is then substituted into the the original PDE, and both sides of the equation are multiplied by the full set of approximation functions. By integrating over the solution region a matrix equation is produced (a Galerkin matrix). Solving this matrix for the unknown coefficients provides the approximate potential. However, unless the solution domain is broken up into small finite elements, the Galerkin method as just described is not easy to solve with a computer solution. That is, the matrix entries are difficult to compute over regions of arbitrary shape and the matrix generated usually has no entries equal to zero. So, instead, the solution domain is subdivided into smaller regions of regular shape, and the Galerkin method is applied to each element. See Steele or Zienkiewicz or other books dealing with finite elements for more details on the Galerkin method.

solution of the variable is then found at each location for each time interval, with the solution of the PDE for the previous time step being used as the initial value for the next time step. After the unknowns are found for a given time step, time is advanced by an increment, and the solution is calculated again. This cycle is continued until the final specified time step is reached. Note that as these time-dependent equations are being solved it is assumed that the variables do not vary during the time interval over each individual subdomain; i.e., the concentrations of the various impurities present and the diffusivities in the Fick's 2nd Law equation, and φ, n, and p in the transient continuity equations keep their same values from the beginning to end of each time increment.)

Solution of such time-dependent PDEs can be very CPU intensive because of the number of times the PDE must be solved. It is common for transient device analysis to require 100 to 1000 times the CPU time needed for a single dc analysis of the same device. (Note, however, that if the static I-V curve - for example of a MOSFET - is needed, the value of I_D must be calculated for the desired number of points of the I-V curve, see for example, Fig. 4-14. Thus, for many steady-state cases the PDE may also have to solved a number of times.)

2.4.4.1 Explicit Time Discretization. It is possible to discretize time derivatives in various ways, depending on the choice of time points at which the spatial variation of the unknown variable is evaluated.[9] The most straightforward technique computes the unknown at the present time and thus yields equations which can be solved explicitly for the values at the new time step. The simplest *explicit* time discretization method is called the *forward-difference* (or *forward-Euler*) time-discretization method. The single unknown variable at the next time step is found explicitly from the value of the variable at the present time step. For example, in a one-dimensional diffusion equation (in which for simplicity we assume D is a constant) the concentration of impurities at position m at time step t^{k+1} (C_m^{k+1}) can be found from

$$\frac{[C_m^{k+1} - C_m^k]}{\Delta t} = D\frac{\partial^2 C_m^k}{\partial x^2} \qquad (2 - 17)$$

If the central difference second derivative is used to discretize the spatial derivatives (right hand side), and the grid spacing is uniform, we get

$$C_m^{k+1} = \left\{ [\frac{D\Delta t}{\Delta x^2}]C_{m-1}^k - (1 - [\frac{2D\Delta t}{\Delta x^2}])C_m^k + [\frac{2D\Delta t}{\Delta x^2}]C_{m+1}^k \right\} \qquad (2 - 18)$$

Now we can solve for C_m^{k+1} in terms of the known values of C_m^k from the previous (present) time step. In Fig. 2-9a, the reader can see that the forward Euler method connects mesh points C_{m-1}, C_m, and C_{m+1} at time step k to C_m at time step $k+1$.

The chief attraction of explicit time discretization is its simplicity. That is, at each time step neither a matrix solution nor a nonlinear iteration is needed to evaluate the unknown C_m^{k+1}. However, the approach is inherently unstable unless sufficiently small time steps are used. That is, the error caused by the discretization of the time

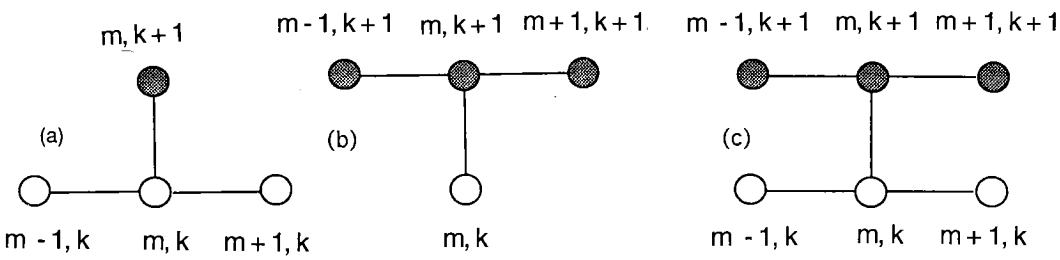


Fig. 2-9 Time discretization: (a) Forward-Euler (Explicit) Method; (b) Backward Euler (Implicit) Method; (c) Crank-Nicholson (Implicit) Method.

derivative can be cumulative in the forward Euler method, and is not bounded from one time step to the next. As a result, as time develops this accumulated error grows steadily and eventually yields a meaningless unphysical solution. In practice, to ensure sufficient accuracy while still guaranteeing stability with the forward-Euler method, unacceptably small time steps must be used to solve the diffusion equation and the current-continuity equations (e.g., of the order of 10^{-15} sec for doping concentrations of 10^{17} cm^{-3}). Use of such small time steps would entail the solution of the equations at an unnecessarily large number of time intervals to perform the simulation. As a result, the more complex (but also more stable) *implicit* time discretization methods are generally used instead.

2.4.4.2 Implicit Time Discretization. In *implicit time discretization* methods, the unknown parameters (e.g., the dopant concentrations at the next time step) are calculated using the values at the new time step as well as at previous time steps. In such methods, the discretization errors are partially canceled from one step to the next, and thus, accumulated errors can be limited to a certain, acceptably small value. Since the concentrations at the new time step are unknown, to find them requires the solution of a large system of equations at each time step, and may require large memory capacity. The long time needed to carry out the computations, however, is compensated by the greatly increased time step. Two of the most widely used implicit methods are the *backward Euler method* and the *Crank-Nicholson* method.

The first-order *backward-Euler method* is the most unconditionally stable of all the implicit methods. In the diffusion equation example above after time discretization using the backward-Euler method, the equation would be expressed as

$$C_m^{k+1} = C_m^k + \Delta t\{\frac{\partial}{\partial x} D [\frac{\partial C_m^{k+1}}{\partial x}]\} \qquad (2 - 19)$$

or

$$C_m^{k+1} = C_m^k + \left\{ \left(\frac{D\Delta t}{\Delta x^2}\right)C_{m-1}^{k+1} - \left(\frac{2D\Delta t}{\Delta x^2}\right)C_m^{k+1} + \left(\frac{2D\Delta t}{\Delta x^2}\right)C_{m+1}^{k+1} \right\} \qquad (2 - 19a)$$

In Fig. 2-9b, the reader can see that the backward-Euler method connects mesh point C_m at time step k to mesh points C_{m-1}, C_m, and C_{m+1} at time step $k+1$.

In the *Crank-Nicholson method* the time discretization is performed at the time step $k+1/2$ [i.e., the unknown variable is averaged between the value at previous time step (which is known) and the next time step (at which the value is unknown)]. The Crank-Nicholson method is attractive because it is second-order accurate in time, and also exhibits excellent stability. For example, if Eq. 2-17 was discretized using the Crank-Nicholson method, it would take the following form

$$C_m^{k+1} = C_m^k + \frac{1}{2}\left\{ \left(\frac{D\Delta t}{\Delta x^2}\right)C_{m-1}^k - \left(\frac{2D\Delta t}{\Delta x^2}\right)C_m^k + \left(\frac{2D\Delta t}{\Delta x^2}\right)C_{m+1}^k \right.$$

$$\left. + \left(\frac{D\Delta t}{\Delta x^2}\right)C_{m-1}^{k+1} - \left(\frac{2D\Delta t}{\Delta x^2}\right)C_m^{k+1} + \left(\frac{2D\Delta t}{\Delta x^2}\right)C_{m+1}^{k+1} \right\} \qquad (2 - 20)$$

In Fig. 2-9c, one can see that the Crank-Nicholson method connects mesh points C_{m-1}, C_m, and C_{m+1} at time step k to mesh points C_{m-1}, C_m, and C_{m+1} at time step $k+1$.

Another time discretization scheme that has been proposed is the *trapezoidal rule/backward difference-formula* (TR-BDF2). It is a composite method, i.e., it uses both implicit and explicit time discretization.[10]

2.5 ORDERING THE SET OF DISCRETIZED PDEs INTO APPROPRIATE MATRIX FORM

The set of discretized PDEs written for each node (using either the finite difference or finite element method) can be combined in matrix form so that a single matrix equation represents the entire system of equations. Matrix techniques can then be used to develop various solution techniques. For example, the discretized set of the Poisson equation may be more conveniently expressed in matrix form as

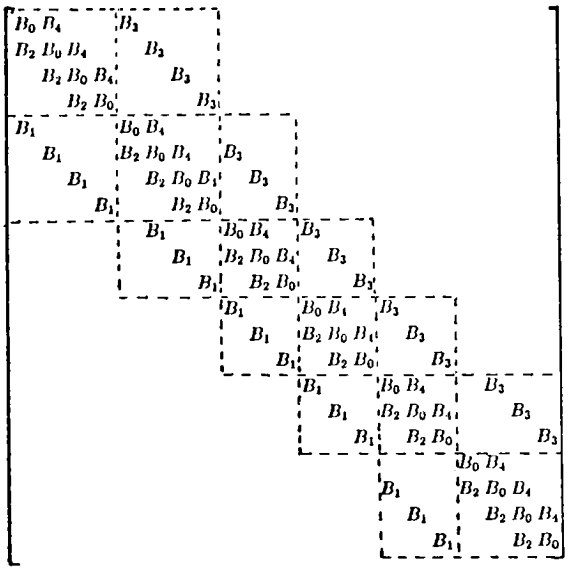
$$[\mathbf{B}]\{\varphi\} = \{\mathbf{Q}\} \qquad (2 - 21)$$

where $[\mathbf{B}]$ - in the finite difference approach - is the coefficient matrix for the finite difference equations, and $\{\varphi\}$ and $\{\mathbf{Q}\}$ are vectors consisting of the elements $\varphi_{i,j}$ and $Q_{i,j}$ for each node of the solution region where $\varphi_{i,j}$ is unknown. Equation 2-21 represents a system of m simultaneous equations, where m is the total number of elements in φ.

The matrix $[\mathbf{B}]$ has a maximum of five-non-zero elements on any row of the matrix, and is therefore termed *sparse*. Sparse matrix solution techniques have been developed for such linear systems of equations which significantly reduce the cost of computation by exploiting the sparsity of the given system. A variety of sparse matrix solution software

$$\begin{bmatrix} a_{11} & a_{12} & 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 & 0 \\ 0 & a_{32} & a_{33} & a_{34} & 0 \\ 0 & 0 & a_{43} & a_{44} & a_{45} \\ 0 & 0 & 0 & a_{54} & a_{55} \end{bmatrix}$$

a)



b)

Fig. 2-10 (a) Tri-Diagonal Matrix; (b) Penta-Diagonal Matrix.

packages have been developed to solve such sparse matrix systems. However, the form of the coefficient matrix must be such as to allow the sparse matrix techniques to be used to best advantage. That is, the coefficient matrix should ideally be organized so that a tridiagonal matrix is obtained for one-dimensional PDEs (such as the 1-D Fick's Second Law equation, Fig. 2-10a), and pentadiagonal matrices are formed for two-dimensional PDEs (such as the 2-D Poisson equation, Fig. 2-10b).*

The form of **[B]**, however, depends on the method of ordering the nodes within the solution region as well as on the equations associated with these nodes. The nodes can be sequenced in the order of i and $j - i$ being held constant at each of its values while j takes on all values in its range. This is illustrated in Fig. 2-11 for a simple node structure where the numbers define the node ordering. The elements $\varphi_{i,j}$ and $Q_{i,j}$ are placed in $\{\varphi\}$ and $\{\mathbf{Q}\}$ using the same order as the node sequence. This type of node ordering scheme is called a *natural by columns* ordering method.

More sophisticated ordering sequences can be produced using one of several algorithms developed by various workers (to take advantage of the sparse nature of the coefficient matrices). One that is well established for ordering the nodes of sparse matrices which are to be solved using direct methods so that a minimum number of non-zero entries occur after factorization (see section 2.7) is by Cuthill and McKee.[11]

* A set of tridiagonal linear algebraic equations are characterized by the unique form of having nonzero elements on the diagonal and a maximum of one element to the left and the right of the diagonal. The rest of the elements are zeros. The tridiagonal matrix of the coefficients of such a set of equations is shown in Fig. 2-10a.

1	6	11	16	21
2	7	12	17	22
3	8	13	18	23
4	9	14	19	24
5	10	15	20	25

Fig. 2-11 *Natural by columns* node ordering scheme.

2.6 LINEARIZING NONLINEAR ALGEBRAIC EQUATIONS

The discretized, ordered algebraic equations representing the quantitative model for the unknown parameter in the device domain are generally *nonlinear* algebraic equations. [Nonlinear algebraic equations are defined as those which contain powers and/or transcendental functions of the unknown variable(s).] Nonlinear algebraic equations cannot be solved by conventional linear algebraic techniques. Instead, a method such as the *Newton-Raphson (NR) iteration method* must be used to linearize such equations, after which linear algebraic techniques can be applied to solve them. (*Iteration* is defined as a computational procedure in which the desired result is approached through a repeated cycle of operations, each of which more closely approximates the desired result.) In such iterative linearizing methods, the nonlinear terms are first linearized, and an initial estimate of the unknown being sought is chosen. This initial estimate is substituted into the resulting equations. Solving the linearized equations with an estimate of the unknown value yields a new estimate of the unknown. This new estimate is then compared to the previous estimate. If the difference between the two is too large, the procedure is repeated, using the new estimate as the known value in the linearized equation in the next iteration. The iteration procedure is halted when the difference between the new and the previous estimate (hopefully) becomes acceptably small. If this condition is reached, it is said that the iteration procedure has converged to the solution. However, if after a number of iterations the new estimate still does not approach the value of the previous estimate to within the desired limit, the iteration method might not converge at all. The maximum number of iterations is usually set by the user and is more or less arbitrary. The convergence of the NR method is chiefly dependent on how close the initial estimate is to the final solution.

In the NR iteration method, linearization is achieved by replacing the nonlinear terms of the unknown variable of the equations with a first order Taylor series expansion about the most recent estimate of the unknown variable, or

$$f(x_0 + \Delta x) = f(x_0) + f'(x_0)\,\Delta x \qquad (2 - 22)$$

The iteration is started by choosing an initial estimate of the unknown. For each iteration, the equation is linearized by this method, and a linear solution technique is then used to solve the equation. The iteration is stopped when adequate convergence is achieved, or after the number of iterations specified by the user are performed and convergence has not yet been reached.

As an example of the NR method, consider a simple circuit containing a dc voltage source of value V_A (in volts), a resistor of value R_A (in Ω), and a *pn* junction diode as shown in Fig. 2-12. Assume that the polarity of the voltage source is such that the diode is biased well into forward-bias operation, and the steady-state I-V relation for the diode is then

$$I = I_0 \exp(qV/kT) \qquad (2 - 23a)$$

where I_0 is the diode saturation current, and V is the voltage across the diode. The Kirchhoffs-Current-Law equation for node 2 of this circuit is

$$(V_A - v_2)/R_A - I_0 \exp(qv_2/kT) = 0 \qquad (2 - 23b)$$

Since we have one nonlinear equation with one unknown (v_2) we can solve it using the NR iteration method. That is, we assume the initial value of v_2 is v_2^0 (e.g., $v_2^k @ k = 0 = v_2^0$ where here the superscript k denotes the number of the iteration). Then we apply a Taylor series expansion of the diode current I about this point, and retain only first-order terms. If in Eq. 2-22 I is $f(x_0 + \Delta x)$, and x_0 is now v_2^k and Δx is now $v_2^{k+1} - v_2^k$, $I = f(v_2^{k+1})$, and we get at the beginning of the first iteration,

$$I(v_2^1) = f(v_2^1) \sim f(v_2^0) - f'(v_2^0)(v_2^1 - v_2^0) \qquad (2 - 24a)$$

or

$$I \sim I_0 \exp(qv_2^0/kT) + (dI/dV @ v_2^k = v_2^0)(v_2^1 - v_2^0) \qquad (2 - 24b)$$

$$\sim I_0 \exp(qv_2^0/kT) + [q(v_2^1 - v_2^0)/kT]\, I_0 \exp(qv_2^0/kT) \qquad (2 - 24c)$$

With this linearization the Kirchhoffs Current Law equation at node 2 becomes for the kth iteration

$$(V_A - v_2^{k+1})/R_A - I_0 \exp(qv_2^k/kT) - [q(v_2^{k+1} - v_2^k)/kT]\, I_0 \exp(qv_2^k/kT) = 0 \qquad (2 - 25)$$

This equation is now linear in terms of the one unknown (v_2^{k+1}), and thus can be

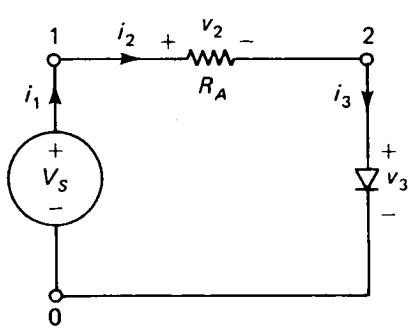


Fig. 2.12 Simple diode circuit to illustrate the method of nodal analysis when nonlinear elements are present.

solved to determine a new estimate for this voltage using linear algebraic techniques.

The NR iteration procedure can also be applied to a set of nonlinear algebraic equations expressed in matrix form. The first derivatives in the second term of the Taylor series are expressed in a matrix form known as the Jacobian matrix. That is, the *Jacobian* is a square matrix of partial derivatives, of the form

$$\begin{bmatrix} \frac{\partial Q_1}{\partial \varphi_1} & \frac{\partial Q_1}{\partial \varphi_2} & \cdots & \frac{\partial Q_1}{\partial \varphi_n} \\ \frac{\partial Q_2}{\partial \varphi_1} & \frac{\partial Q_2}{\partial \varphi_2} & \cdots & \frac{\partial Q_2}{\partial \varphi_n} \\ \frac{\partial Q_n}{\partial \varphi_1} & \frac{\partial Q_n}{\partial \varphi_2} & \cdots & \frac{\partial Q_n}{\partial \varphi_n} \end{bmatrix} \qquad (2 - 26)$$

Each of the terms of the Jacobian are evaluated at values of φ from the previous iteration.

If we use as an example the 2-D form of the Poisson equation, where φ is the unknown, the nonlinearized form of the matrix equation at the beginning of iteration k is

$$[\mathbf{B}]\,\{\varphi^{k+1}\} = [\mathbf{Q}^{k+1}] \qquad (2 - 27)$$

To linearize this equation using the NR iteration technique we replace $[\mathbf{Q}^{k+1}]$ (containing the nonlinear terms of φ) with

$$[\mathbf{Q}^{k+1}] = [\mathbf{Q}^{k}] + [\mathbf{J}^{k}]\,\{\varphi^{k+1} - \varphi^{k}\} \qquad (2 - 28)$$

so Eq. 2-27 becomes

$$[\mathbf{B}]\,\{\varphi^{k+1}\} = [\mathbf{Q}^{k}] + [\mathbf{J}^{k}]\{\varphi^{k+1} - \varphi^{k}\} \qquad (2 - 29)$$

where

$$\mathbf{J}^{k}(p) = \partial[\mathbf{Q}](p)/\partial\varphi\,(p)\ @\{\varphi^{k}\} \quad p = 1,2,3, \ldots N \qquad (2 - 30)$$

To express the matrix equation so that all the unknowns φ^{k+1} appear in the vector $\{\varphi^{k+1}\}$ we rearrange Eq. 2-29 to yield

$$[\mathbf{B} - \mathbf{J}^{k}]\,\{\varphi^{k+1}\} = [\mathbf{Q}^{k}] - [\mathbf{J}^{k}]\{\varphi^{k}\} \qquad (2 - 31)$$

Note that all of the terms in $\{\varphi^{k}\}$ are known, since they involve quantities calculated during the previous iteration. Thus, the set of nonlinear equations is linearized and a linear matrix solution technique can be used to find the values of $\{\varphi\}^{k+1}$. When this is found, the values of $\{\varphi\}^{k+1}$ are compared to those of the previous iteration according to

$$|\{\varphi\}^{k+1} - \{\varphi\}^{k}/\{\varphi\}^{k+1}| < \delta_1 \qquad (2 - 32)$$

where δ_1 is a specified error bound. If the condition is met this indicates that the iteration has converged. If not, the iteration procedure is repeated until it is met (or until the maximum number of allowed iterations have been performed without

convergence being reached). As noted earlier, the convergence of the NR method depends on how close the initial estimate is to the final solution. There are a number of ways for choosing a good initial guess to start the iteration, based either on physical reasoning (e.g., assuming that quasi-charge-neutrality exists at thermal equilibrium), or on some numerical techniques. Appendix A provides two worked out examples of the NR iteration method. An example of one of the equations of the matrix Eq. 2-31 is explicitly written out in Section 2.10.

2.7 MATRIX METHODS FOR SOLVING SIMULTANEOUS LINEAR ALGEBRAIC EQUATIONS

The solution of simultaneous linear algebraic equations is probably one of the most important procedures in modern engineering computations. For example, the finite element method and the finite difference method both involve the solution of systems of linear algebraic equations. Two classes of methods are available for solving a set of linear algebraic equations: *direct methods*, and *iterative methods*. Appendix A provides a basic introduction to both types of these methods, including simple worked examples for some of them.

The *direct methods* are based on the mathematical technique of matrix elimination. They include Cramer's Rule, Gauss elimination method, the Gauss-Jordan elimination method, and Cholesky's method (also called Crouts' method, or the *LU*-decomposition method). The most widely used direct method in process and device simulators is the latter, which is based on a symmetric variation of the Gaussian elimination technique. That is, in Cholesky's method *triangular factorization* is applied to the coefficient matrix e.g., [**A**] of the equation

$$[\mathbf{A}]\{\mathbf{x}\} = \{\mathbf{b}\} \tag{2 - 33}$$

which yields

$$[\mathbf{A}] = [\mathbf{L}]\,[\mathbf{L}^T] \tag{2 - 34}$$

where [**L**] is a *lower triangular matrix* and [$\mathbf{L}^T$] is the transpose of the [**L**] matrix (and [$\mathbf{L}^T$] is also an *upper triangular matrix* (see Appendix A for definitions of these terms). Using (2-34) in (2-33) we get

$$[\mathbf{L}]\,[\mathbf{L}^T]\,\{\mathbf{x}\} = \{\mathbf{b}\} \tag{2 - 35}$$

and by substituting $\{\mathbf{y}\} = [\mathbf{L}^T]\,\{\mathbf{x}\}$, we can find $\{\mathbf{x}\}$ by solving the triangular systems

$$[\mathbf{L}]\{\mathbf{y}\} = \{\mathbf{b}\} \tag{2 - 36}$$

and

$$[\mathbf{L}^T]\,\{\mathbf{x}\} = \{\mathbf{y}\} \tag{2 - 37}$$

However, upon factoring a sparse matrix using the Cholesky method, it suffers *fill-in* (i.e., [**L**] has nonzero entries in positions which are zero in the lower triangular part of [**A**]). By judicious reordering the equations of the set it is possible to end up with a triangular factor [**l**] which is just as sparse as the lower triangle of [**A**]. This will result

in significant savings in computer execution time and storage (assuming of course that sparseness is exploited). Algorithms have been developed to automatically perform this reordering, one of the best established being by Cuthill and McKee.

An example of a worked problem using Cholesky's method is given in Appendix A, section A-4.1.5. A text which describes the direct solution methods of sparse linear matrix equations (emphasizing the Cholesky method) has been written by George and Liu.[12] It also provides information on SPARSPAK, a software package of direct linear matrix equation solvers.

The *iterative methods* include Jacobi's method, the Gauss-Seidel method, the successive-overrelaxation (SOR) method, the alternating-direction implicit (ADI), and Stone's strongly implicit method. Varga's text is a good introduction to iterative methods for solving linear matrix equations.[13] A good (but compressed) survey on the subject of the various matrix methods for solving sets of sparse matrix linear algebraic equations is by Duff.[14]

If an iterative method is used to solve the linearized matrix equation, it can be written as

$$[\mathbf{A}]\{\mathbf{x}\} = \{\mathbf{b}\} \tag{2 - 38}$$

where the elements of Eqs. 2-31 and 2-33a are related by

$$[\mathbf{A}] = [\mathbf{B} - \mathbf{J}^k] \tag{2 - 39a}$$
$$\{\mathbf{x}\} = \{\varphi\}^{k+1} \tag{2 - 39b}$$
$$\{\mathbf{b}\} = \{\mathbf{Q}^k\} - \{\mathbf{J}^k\varphi^k\} \tag{2 - 39c}$$

The iterative techniques use the notation $\{\mathbf{x}\}^m$ to represent an approximation to $\{\mathbf{x}\}$, where the superscript m indicates that $\mathbf{x}^m$ is obtained as a result of the mth linear iteration used to solve Eq. 2-38. The linear iteration procedure is started with $\mathbf{x}^0 = \varphi^k$ as the initial selection of the approximation of the unknown vector, and is repeated until the approximation to $\mathbf{x}$ resulting from successive iterations satisfy the condition

$$|\{\mathbf{x}_i\}^{m+1} - \{\mathbf{x}_i\}^m / \{\mathbf{x}_i\}^{m+1}| < \delta_2 \tag{2 - 40}$$

The relative merits of direct and iterative methods often depend on the type of application. For systems of *moderate* size (e.g., the solution of the one-dimensional Fick's 2nd Law equation in SUPREM II), the use of direct methods is often advantageous. The large systems of simultaneous equations with sparse coefficient matrices obtained by finite-difference discretization are particularly well suited to solution by iterative methods. This is because iterative methods have smaller computer storage requirements, typically needing to store only $[\mathbf{A}]$ $\{\mathbf{b}\}$ and $\{\mathbf{x}\}^k$. On the other hand, when $[\mathbf{A}]$ is factored during a direct solution procedure, it typically suffers fill-in, so that the *filled matrix* must be stored. In addition, Gaussian elimination in its primary form requires more arithmetic operations than a single iteration cycle. Thus, if the matrix is large and the iteration converges to the desired accuracy in a small number of iterations, less storage is needed and considerable computing time can be saved by using an iterative method.

But, the newer direct methods require fewer operations since they take advantage of sparseness. Thus, a comparison of iterative versus direct methods of computation becomes quite complicated. In fact, unless the question of which class of method should be used is posed in a quite narrow and well-defined context, it is either very complicated or impossible to answer.[13]

2.8 VISUALIZATION OF THE RESULTS

To view 1-D or 2-D results from the data base, a postprocessor is needed to allow the calculated results to be presented in variety of formats at the user's option, including: linear, log-linear, contour, log-contour, or "bird's eye" (perspective). Here we will give some examples of the above formats. For a discussion of recent trends in scientific visualization for semiconductor technology simulation see ref. 15.

Figure 2-13a shows a entire MOSFET from a 3-D perspective. Next, in Fig. 2-13b we show the same device, also from 3-D perspective, but sliced down the middle of the

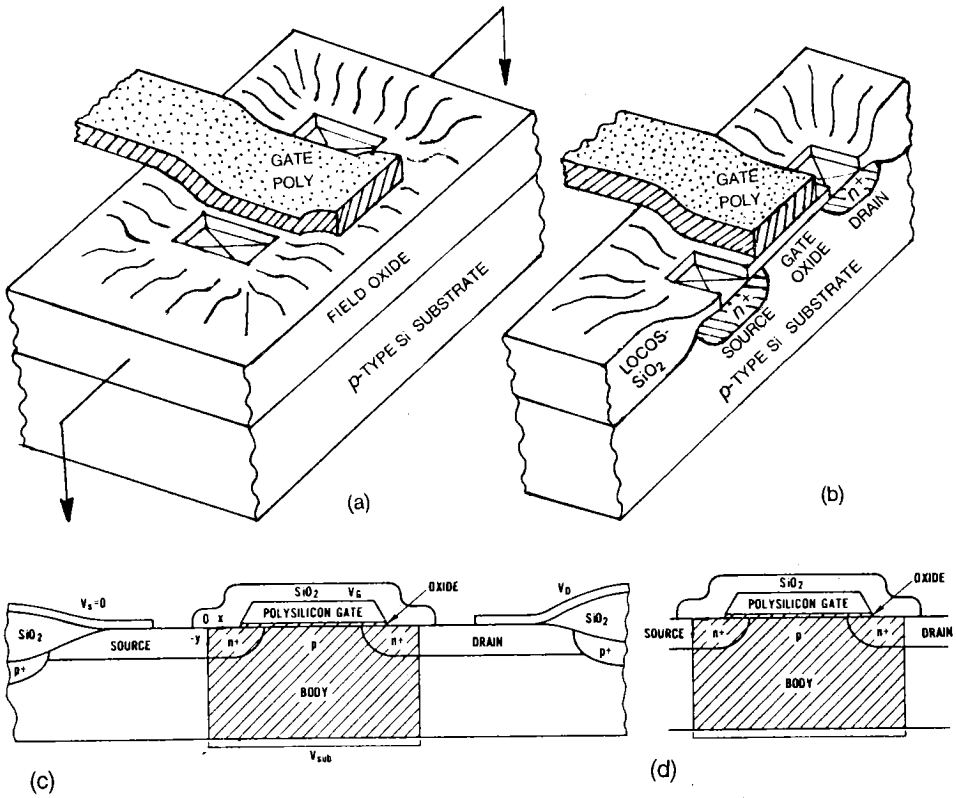


Fig. 2-13 (a) MOSFET structure from a 3-D perspective; (b) MOSFET from a 3-D perspective, but sliced down the middle of the channel so that a cross-section of the device is also visible; (c) 2-D cross-section of the MOSFET; (d) 2-D cross-section of the MOSFET showing only the active region.

channel so that a cross-section of the device along that direction is also visible. In Fig. 2-13c we show only this 2-D cross-section (including the surrounding field oxide region). Figure 2-13d shows the same cross-section, but excludes the field oxide region, so that only the active area of the device is depicted.

Next, we consider how the data obtained from a process or device simulation can be displayed. First we show a non-uniform mesh overlaying the device region (see Fig. 2-14a). We can display the simulated 2-D results of either doping concentration or the potential directly on the 2-D cross-section of the MOSFET active area using a contour plot format (Figs. 2-14b and 2-14c respectively). Note, however, that contour plots are somewhat cumbersome to interpret.

A graphical format that makes the 2-D variation easier to visualize is the 'bird's eye' (or *perspective*) surface plot. Note that this is still a plot of the 2-D dependence of the parameter of interest on position, but the third-dimension is used to allow the 2-D variation to be seen more vividly. We begin by taking the 2-D grid overlaying the active device region, and tilting it so it is lying flat (Fig. 2-14d). Next, we plot the value of the parameter being visualized as a point above the 2-D plane, with a height equal to the parameter value. When all these points are connected, a surface plot of the 2-D parametric variation is produced. We show such a surface plot of the doping concentration in the active region of a MOSFET in Fig. 2-14e1. (Here it is assumed that dopants have been placed under the oxide in the channel to adjust the threshold voltage, hence making the doping concentration in the channel near the surface higher than that deeper in the substrate.) In Fig. 2-14e2 we show a one-dimensional plot of the doping in the channel at the center of the device as a function of depth below the oxide. Again, higher doping near the surface is due to the V_T-adjust implant. In Fig. 2-14f1 we show the 2-D variation of the potential in this MOSFET (assuming a gate-source and drain-source voltage has been applied), again using such a 'bird's eye' perspective plot. In Fig. 2-14f2 the 1-D potential variation as a function of depth beneath the gate oxide at the center of the channel is known. Note that this plot is but one of many that can be extracted from the 2-D bird's eye plot.

If it is desired to show the doping variation in three-dimensions in a manner similar to the surface plot format, we could use color variation to represent change in doping concentration versus position, and this is one option being used in 3-D simulators.

2.9 DATA BASE OF THE SIMULATION PACKAGES

As we have noted, device and process simulation systems must operate on large volumes of data. This implies that large memory capacity must be available and that efficient data handling schemes must be established. Concerning the latter, a comprehensive, flexible data base is needed to keep from being overwhelmed by the mass of the simulation data. Included should be the capability for data storage and retrieval, and for pre- and post-processing of the data in an integrated, interactive environment. One manner of organizing such a data base is suggested by Cole et al.[15] In this suggested approach, the data is placed into five categories, or 'files':

1. A general *variable file* for parameters related to physical constants and coefficients needed by the models.

2. A *doping file* for impurity atom distribution (either a result of process simulation or created from a superposition of analytic functions).

3. A *geometry file* that contains the nodes & elements of the spatial discretization.

4. A *results file* that contains the discretized (nodal) values of such quantities as electronic potential, electron and hole concentrations, dopant atom concentrations, etc.

5. A *dictionary file* that provides a single indicator to any given simulation "run" and a cross-reference to data for that run contained in the other four data files mentioned above.

Other methods have also been proposed for handling the data. The most notable is that of Duvall, who proposed the Profile Interchange Format (PIF).[16] One of the objective of this data handling scheme is also to establish a standard protocol for exchanging data among simulation tools developed at different sites.

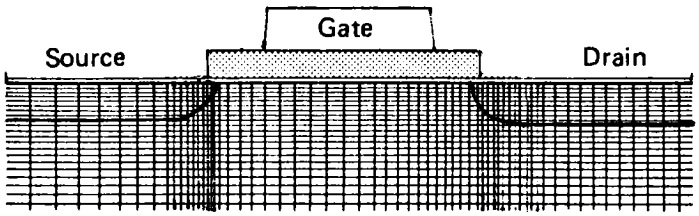
2.10 EXAMPLES OF THE PROCEDURES USED TO NUMERICALLY SOLVE THE MODEL PARTIAL-DIFFERENTIAL EQUATIONS OF MICROELECTRONICS

2.10.1 Outline of a Numerical Analysis Procedure for Solving the Two-Dimensional Poisson Equation to Simulate the Electrostatic Potential in a MOSFET

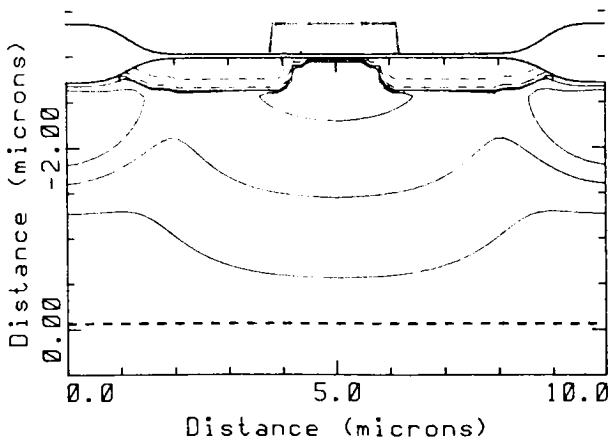
In this section we present an example of how the electrostatic potential as a function of position is simulated in an NMOSFET for a given set of bias conditions. This is done by solving the 2-D Poisson equation using numerical analysis techniques. We will assume that the geometry of the MOSFET in our example is as shown in Fig. 2-15. (Note that this device geometry is deliberately simplified to reduce the mathematical complexity of our example.) For the same reason, we also assume that $N_D = 0$. Finally, it is assumed the external dc biasing voltages (e.g., V_{GS} and V_{DS}) are restricted to values such that the MOSFET is operated only in the subthreshold region, and that these voltage values are given.

If the MOSFET is operated in subthreshold (see chap. 4), it can be assumed that the currents flowing in the device are small enough so that they do not change the electrostatic potential. Therefore, the solution of the Poisson equation alone yields an accurate simulation of the electrostatic potential in the silicon as a function of position $\varphi(x,y)$ in the device. To find $\varphi(x,y)$ we must set out to solve the Poisson equation in two dimensions.

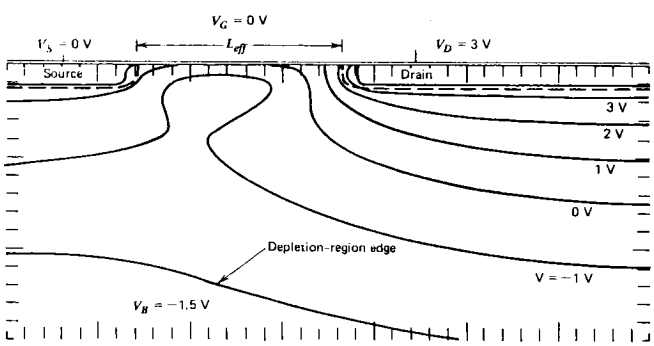
Since analytic closed-form solutions of the 2-D Poisson equation are not known, numerical techniques must be used to obtain the solution. Here we outline the sequence of mathematical procedures based on the *difference operator finite-difference method* for doing this. The following steps must be taken:



a)

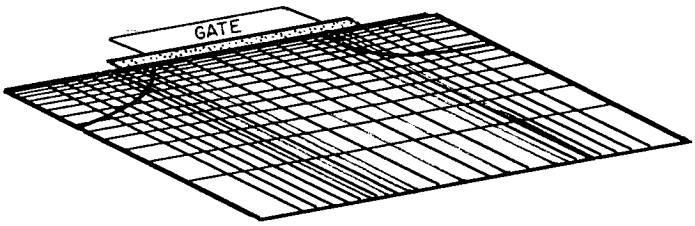


b)

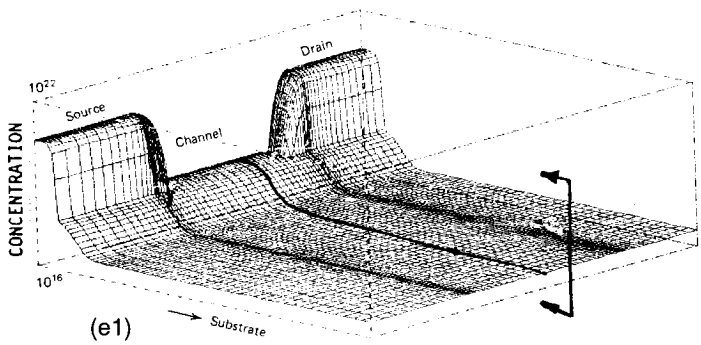


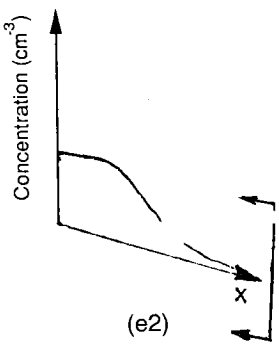
c)

Fig. 2-14 (a) Non-uniform mesh overlaying the device region of the cross-section of a MOSFET. (b and c) Contour plot format of the doping concentration (b), and the potential of the MOSFET of part (a).

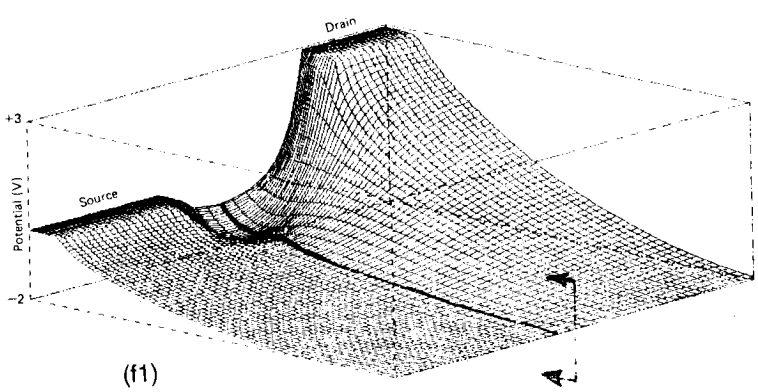


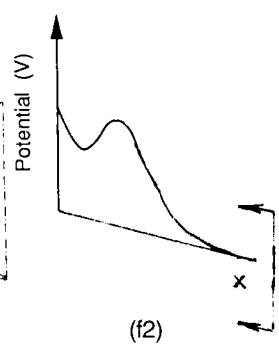
d)





(e)





(f)

Fig. 2-14 (cont.) (d) The mesh of part (a) lying on its back. (e and f) Perspective plot format used to depict the doping concentration (e1), and the potential (f1) of the same MOSFET of parts (b) and (c). Note That (e2) and (f2) show the 1-D variation of the doping concentration and potential versus depth (x) in the MOSFET at the center of the channel.

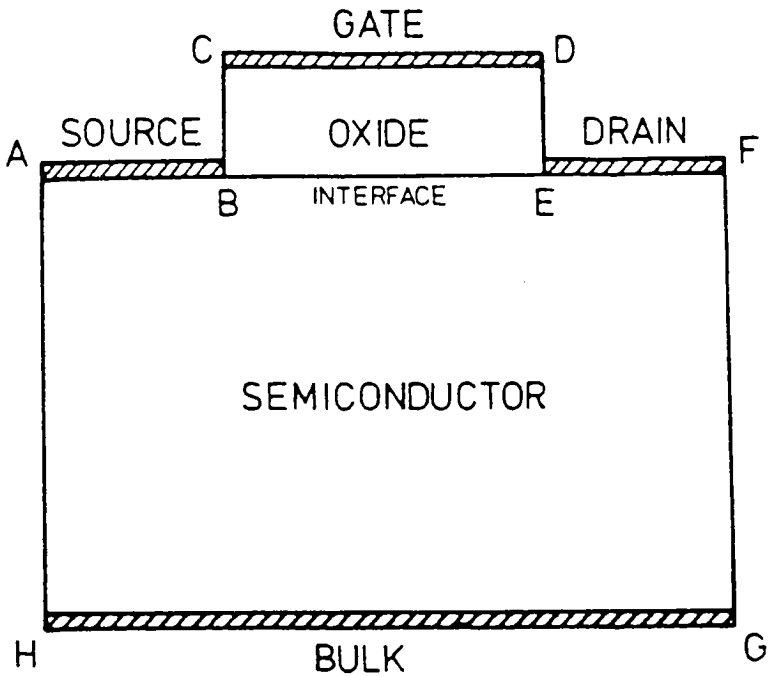


Fig. 2-15 The simulation geometry of the MOSFET for our example.

1. *State the Poisson equation in 2-D form for the Given Device Structure.* The Poisson equation in two dimensions for the silicon region of a MOSFET described above has the following form (i.e., assuming $N_D = 0$):

$$\frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2} = -\left(\frac{q}{\kappa_S\varepsilon_0}\right)\left[-n_i\, e^{q(\varphi - \varphi_n)/kT} + n_i\, e^{q(\varphi_p - \varphi)/kT} - N_A\right] \qquad (2-41)$$

where the right side expresses the charge in terms of the mobile charge concentrations (in terms of φ - assuming Boltzmann statistics), and the doping concentration N_A. The doping concentration as a function of position is assumed to be given. The Poisson equation is solved for the rectangular area A-F-G-H of the MOSFET in Fig. 2-15, which represents the silicon region. (In the area C-D-E-B, which represents the gate oxide, only the Laplace equation has to be solved - which is just the Poisson equation with the charge density equal to zero, $\partial^2\varphi/\partial x^2 = 0$ - because we assume that no charge centers exist in an ideal oxide, and because $\partial^2\varphi/\partial y^2 = 0$ [since there is no y- component of the oxide electric field in this structure].)

2. *Specify the Boundary Conditions.* In order to solve the 2-D PDE Poisson equation a set of spatial boundary conditions in φ is required. The boundary conditions for our example are assumed as follows: At the contacts (i.e., segment A-B: source contact, E-F: drain contact, and G-H substrate contact) a perfect ohmic contact is assumed to exist. The potential at each of these contact regions is thus kept fixed along the contact

length at a value equal to the applied bias plus the built-in potential between the source/drain region and the substrate (e.g., if the source and substrate are both connected to ground, the potential at A-B is set at 0V + V_{bi}). Boundary conditions where the value of the dependent variable itself is specified are referred to as Dirichlet boundary conditions.

At the vertical boundaries (A-H, F-G) the derivative of the potential perpendicular to the boundary (i.e., the lateral electric field) is specified (that is, so-called Neumann boundary conditions are specified here). These values are set at zero in our example, which assumes that zero current flows across these boundaries.

The gate potential boundary condition is again of the Dirichlet type, with the voltage being fixed along the gate at a value of applied bias minus the flat-band voltage (to account for the work-function difference and any oxide charge at the Si-SiO_2 interface). At the Si-SiO_2 interface the potential must obey Gauss's law (see Eq. 3-15), so the boundary condition at the Si-SiO_2 interface is (since $\mathcal{E} = -\partial\varphi/\partial x$):

$$\kappa_{ox}\,(\partial\varphi/\partial x)_{ox} = \kappa_{si}\,(\partial\varphi/\partial x)_{si} \qquad (2 - 42)$$

The solution of the Poisson equation must also satisfy continuity equations internally. The continuity of the potential φ must be maintained everywhere to achieve the physically realistic condition of finite stored electrostatic field energy.

3. *Specify an appropriate grid.* The potential which solves the Poisson equation as defined by Eq. 2-41 is a continuous function of the spatial coordinates. The numerical solution of the Poisson equation requires that the potential and charge concentrations be represented by discrete values at a network of nodes within the solution regions. Hence, a 2-D rectangular grid (which defines the domain of the device geometry) must next be selected (Fig. 2-14a). To improve the solution accuracy, nonuniform grid spacing in the vertical and horizontal directions is usually chosen for this problem - being coarse in regions away from the silicon surface and closely spaced where the dopant concentration varies strongly with position. Each intersection of grid lines represents one of the nodes at which the potential will be determined. The discrete approximation to the continuous Poisson equation relates the potential value at a node to the value at neighboring nodes. The resulting nodes are also sequentially numbered so that when the discretized algebraic equations at each node are established, they will be ordered in a standard format that facilitates the solution of the matrix equation. An example of the sequential node ordering for a simple node structure is shown in Fig. 2-11. Generally, more sophisticated node-ordering algorithms are used.

4. *Discretize the Poisson equation Using the Finite-Difference Method.* Since the electrostatic (non time-varying) potential is being sought, time discretization is not necessary. Hence, only spatial discretization of the Poisson equation needs to be carried out. In this example the '5-point' finite difference approximation of Fig. 2-7 is used. This is the simplest approach for two-dimensional problems since it includes only the four nearest-neighbor nodes (i.e., those horizontally and vertically adjacent to the center node). The resulting equation associated with the center node (at position [i,j]) depends on the potential ($\varphi_{i,j}$) and charge values ($n_{i,j}$, $N_{Ai,j}$) at this node and its four neighbors.

In its simplest form, the finite-difference equation for the Poisson equation for a two-dimensional uniform mesh is (using the central difference approximations for the second derivatives)*

$$\frac{\varphi_{i-1,j} - 2\varphi_{i,j} + \varphi_{i+1,j}}{\Delta x^2} + \frac{\varphi_{i,j-1} - 2\varphi_{i,j} + \varphi_{i,j+1}}{\Delta y^2} = \left(\frac{-q}{\kappa_{Si}\,\varepsilon_o}\right)\left[-n_i e^{q(\varphi_{i,j} - \varphi_n)/kT} + n_i e^{q(\varphi_p - \varphi_{i,j})/kT} - N_{Ai,j}\right] \qquad (2 - 43)$$

This is a nonlinear algebraic equation because φ appears on the right side in exponential form since $n_{i,j}$ and $p_{i,j}$ are expressed with Eqs. 1-6. The equation can nevertheless now be solved using algebraic techniques. Note that since the potential must be determined at each node, an equation at each node must be written (except for the nodes at which the potential is fixed by the externally applied voltage - i.e., the externally applied voltages becomes the boundary conditions at such nodes).

5. *Format the Set of Discretized Equations to Facilitate a Matrix Solution.* The result of the discretization procedure is a set of N simultaneous nonlinear algebraic equations, each having the form of Eq. 2-43. The solution of a large set of algebraic equations can be carried out using matrix methods. To implement this approach, the set of equations is represented as a single matrix equation, which can be written as

$$[\mathbf{B}]\,\{\varphi\} = \{\mathbf{Q}\} \qquad (2 - 44)$$

where $[\mathbf{B}]$ is the coefficient matrix of the difference expressions, and the vector $\{\varphi\}$ consist of the elements $\varphi_{i,j}$ for each node in the solution, and the vector $\{\mathbf{Q}\}$ contains the terms on the right hand side of the equation. (We noted earlier that the nodes of the grid can be ordered so that three of the five nonzero values in each row occur along the principal diagonal and its two adjacent diagonals (Fig. 2-10).* Thus, since most of the entries of $[\mathbf{B}]$ are zero, it is referred to as a *sparse matrix.* In addition, the elements $\varphi_{i,j}$ are placed in $\{\varphi\}$ using the same order as the node sequence, e.g. for the grid ordering scheme shown in Fig. 2-11, $\varphi_{3,4} = \varphi_{14}$ in $\{\varphi\}$.)

To put each equation into a format compatible with the matrix equation 2-44, we collect all the coefficients of the discrete dependent variables on the left-hand side of the

* In the case of a non-uniform mesh, the finite-difference representation of the 2-D Poisson equation is

$$\{[(\varphi_{i+1,j} - \varphi_{i,j})/a_i] - [(\varphi_{i,j} - \varphi_{i-1,j})/a_{i-1}]\}/([a_1 + a_{i-1}]/2)$$

$$+ \{[(\varphi_{i,j+1} - \varphi_{i,j})/b_j] - [(\varphi_{i,j+1} - \varphi_{i,j})]/b_{j-1}])/([b_j + b_{j-1}]/2)$$

$$= \left(\frac{-q}{\kappa_{Si}\,\varepsilon_o}\right)\left[-n_i e^{q(\varphi_{i,j} - \varphi_n)/kT} + n_i e^{q(\varphi_p - \varphi_{i,j})/kT} - N_{Ai,j}\right] \qquad (2 - 45)$$

equation. This puts the algebraic equation at each interior node (i,j) into the following form (for a uniform mesh)

$$\frac{\varphi_{i-1,j}}{\Delta x^2} + \frac{\varphi_{i,j-1}}{\Delta y^2} - \left(\frac{2}{\Delta x^2} + \frac{2}{\Delta y^2}\right)\varphi_{i,j} + \frac{\varphi_{i+1,j}}{\Delta x^2} + \frac{\varphi_{i,j+1}}{\Delta y^2}$$
$$= \left(\frac{-q}{\kappa_{Si}\varepsilon_o}\right)\left[-n_i e^{q(\varphi_{i,j} - \varphi_n)/kT} + n_i e^{q(\varphi_p - \varphi_{i,j})/kT} - N_{A_{i,j}}\right] \qquad (2 - 46a)$$

For example, the equation for node (3,4) in such a uniform mesh would be,

$$\frac{\varphi_{2,4}}{\Delta x^2} + \frac{\varphi_{3,3}}{\Delta y^2} - \left(\frac{2}{\Delta x^2} + \frac{2}{\Delta y^2}\right)\varphi_{3,4} + \frac{\varphi_{4,4}}{\Delta x^2} + \frac{\varphi_{3,5}}{\Delta y^2}$$
$$= \left(\frac{-q}{\kappa_{Si}\varepsilon_o}\right)\left[-n_i e^{q(\varphi_{i,j} - \varphi_n)/kT} + n_i e^{q(\varphi_p - \varphi_{i,j})/kT} - N_{A_{i,j}}\right] \qquad (2 - 46b)$$

and the coefficients of the terms on the left hand side would be placed in the coefficient matrix as shown below

$$\begin{array}{cc} \text{Column} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \text{Row} \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} & \begin{bmatrix} B_{11} & : & 0 & 0 & 0 \\ 0 & B_{22} & B_{32} & & 0 \\ 0 & B_{23} & 1/\Delta y^2 & & \\ 0 & 1/\Delta x^2 & -(2/\Delta x^2 + 2/\Delta y^2) & 1/\Delta x^2 & 0 \\ 0 & 0 & 1/\Delta y^2 & & \\ : & : & : & & \end{bmatrix} \end{array}$$

If we filled the rest of the coefficient matrix [**B**] with the other terms from all the other equations we would see that it is tridiagonal matrix (although when boundary conditions are included it contains at most 5 nonzero entries per row).

6. *Linearize the Set of Discrete Equations.* Each of the individual algebraic equations (such as 2-46) are still non-linear (i.e., the mobile charge densities in the right hand side n and p depend exponentially on φ). Thus, the set of equations must be linearized before linear matrix solution techniques can be applied. As described in section 2.6, the Newton iteration method can be used to linearize the nonlinear terms in the equation about the most recent approximation to φ. For each iteration Eq. 2-44 is linearized by such a technique and then a linear matrix solution technique can be used to solve the resulting *linear* matrix equation of the new approximation to φ. It is desirable that the initial values of $\varphi_{i,j}$ (e.g., $\varphi_{i,j}^o$) be selected so that the iteration is initially numerically stable. In one approach, the vector elements $\{\varphi^o\}$ are chosen such that each element of

* Sparse-matrix ordering techniques must be used to ensure the optimum ordering of the matrices before setting out to solve the matrix equation. There are a variety of ordering algorithms available to deal with sparse matrices, as mentioned earlier.

the vector $\{\mathbf{Q}\} = 0$. This yields a condition of charge neutrality at each node within the region, guaranteeing numerical stability of the initial iteration. The iteration procedure is halted when adequate convergence is achieved (i.e., when the change in one value of φ from one iteration ($\varphi_{i,j}{}^{k}$) to the next ($\varphi_{i,j}{}^{k+1}$) is smaller than some specified error bound, δ_1, as given by Eq. 2-32.

In our example, the nonlinearized form of the matrix equation at the beginning of iteration $k+1$ is

$$[\mathbf{B}]\,\{\varphi^{k+1}\} = \{\mathbf{Q}^{k+1}\} \tag{2 - 47}$$

To linearize this equation using the Newton iteration technique we replace $\{\mathbf{Q}^{k+1}\}$ (containing the nonlinear terms) with

$$\{\mathbf{Q}^{k+1}\} = \{\mathbf{Q}^{k}\} + [\mathbf{J}^{k}]\,\{\varphi^{k+1} - \varphi^{k}\} \tag{2 - 48}$$

so Eq. 2-47 becomes

$$[\mathbf{B}]\{\varphi^{k+1}\} = \{\mathbf{Q}^{k}\} + [\mathbf{J}^{k}]\{\varphi^{k+1} - \varphi^{k}\} \tag{2 - 49}$$

where the elements J^k of the Jacobian matrix $[\mathbf{J}^k]$ are

$$J^k(p,q) = \partial Q(p)/\partial\varphi(q)\,\big|_{\{\varphi k\}} \qquad p,q = 1,2,3, \ldots N \tag{2 - 50}$$

To express the matrix equation so that all the unknowns φ^{k+1} appear in the vector $\{\varphi^{k+1}\}$ we rearrange Eq. 2-49 to yield

$$[\mathbf{B} - \mathbf{J}^{k}]\,\{\varphi^{k+1}\} = \{\mathbf{Q}^{k}\} - [\mathbf{J}^{k}]\{\varphi^{k}\} \tag{2 - 51}$$

Note that all of the terms in $\{\varphi^k\}$ are known, since they involve quantities calculated during the previous iteration.

For the single equation at node (3,4), since $Q_{i,j} = -N_{Ai,j} - n_i \exp\,[q(\varphi_{i,j} - \varphi_n)/kT$ and

$$J_{3,4}{}^{k} = \partial Q_{i,j}/\partial\varphi_{i,j}\,\big|_{\varphi 3,4}{}^{k} = (q/kT)\, n_i \exp\,[q\,(\varphi_{3,4}{}^{k} - \varphi_n)/kT] \tag{2 - 52}$$

replacing $Q_{3,4}{}^{k+1}$ with $Q_{3,4}{}^{k} + [\partial Q_{i,j}/\partial\varphi_{i,j}\,|_{@\varphi 3,4}{}^{k}]\,[\varphi_{3,4}{}^{k+1} - \varphi_{3,4}{}^{k}]$, it becomes

$$\begin{aligned}
&\frac{\varphi_{2,4}^{k+1}}{\Delta x^2} + \frac{\varphi_{3,3}^{k+1}}{\Delta y^2} - \left(\frac{2}{\Delta x^2} + \frac{2}{\Delta x^2}\right)\varphi_{3,4}^{k+1} + \frac{\varphi_{4,4}^{k+1}}{\Delta x^2} + \frac{\varphi_{3,3}^{k+1}}{\Delta y^2} \\
&= \left(\frac{-q}{\kappa_{Si}\varepsilon_o}\right)\left[-n_i e^{q(\varphi_{3,4}^{k} - \varphi_n)/kT} + n_i e^{q(\varphi_p - \varphi_{3,4}^{k})/kT}\right] \\
&+ \left(\frac{-q}{\kappa_{Si}\varepsilon_o}\right)\left\{\left[\frac{q}{kT}\left(-n_i e^{q(\varphi_{3,4}^{k} - \varphi_n)/kT} + n_i e^{q(\varphi_p - \varphi_{3,4}^{k})/kT}\right)\right]\left[\varphi_{3,4}^{k+1} - \varphi_{3,4}^{k}\right] - N_{A3,4}\right\}
\end{aligned} \tag{2 - 53}$$

or

$$\frac{\varphi_{2,4}^{k+1}}{\Delta x^2} + \frac{\varphi_{3,3}^{k+1}}{\Delta y^2} - \left(\frac{2}{\Delta x^2} + \frac{2}{\Delta x^2} - \left[\frac{-q}{\kappa_{Si}\varepsilon_o}\right]\left[n_i e^{q(\varphi_{3,4}^k - \varphi_n)/kT} + n_i e^{q(\varphi_p - \varphi_{3,4}^k)/kT}\right]\right)\varphi_{3,4}^{k+1} + \frac{\varphi_{4,4}^{k+1}}{\Delta x^2} + \frac{\varphi_{3,3}^{k+1}}{\Delta y^2}$$

$$= \left(\frac{-q}{\kappa_{Si}\varepsilon_o}\right)\left[-n_i e^{q(\varphi_{3,4}^k - \varphi_n)/kT} + n_i e^{q(\varphi_p - \varphi_{3,4}^k)/kT}\right] \qquad (2 - 54)$$

$$+ \left(\frac{-q}{\kappa_{Si}\varepsilon_o}\right)\left\{ \left[\frac{q}{kT}\left(-n_i e^{q(\varphi_{3,4}^k - \varphi_n)/kT} + n_i e^{q(\varphi_p - \varphi_{3,4}^k)/kT}\right]\left[\varphi_{3,4}^k\right] - N_{A3,4}\right\}$$

Note that all of the terms on the right hand side of Eq. 2-54 involve φ at its previous iteration (φ^k) - which is known, while the terms on the left hand side involve φ^{k+1} - which is unknown.

Following this procedure at each node, the set of nonlinear equations for the unknown potential is thereby linearized. Linear matrix solution techniques can now be applied to the set to calculate φ^{k+1} at each node.

7. *Solve the Linear Matrix Equation.* As described in section 2.7 a number of matrix solution techniques have been developed to solve the resulting *linearized* (sparse) matrix equation. We also noted that such techniques can be grouped into two classes: direct methods and iterative methods. If a *direct* method is selected (e.g., Gaussian elimination, or some variation thereof), the solutions are checked against the previous nonlinear iteration estimate. If the solutions do not yet meet this convergence criterion, the nonlinear iteration procedure of step (5) is repeated until convergence is achieved. If an *iterative* technique is selected, the linear iteration is repeated until the iteration converges to a specified error bound δ_2, as given by Eq. 2-35. Then, the nonlinear convergence criterion is checked (Eq. 2-32), and if not met, another cycle through the linear iterative procedure is carried out. Thus, as shown in Fig. 2-18b, if an iterative technique is used to solve the linear matrix equation, this iteration is referred to as the *inner iteration* or *inner loop* of the algorithm, whereas the *non-linear iteration* is referred to as the *outer iteration* or *outer loop*.

8. *Display the Results in an Appropriate Graphical Format.* The values of φ computed for each of the grid points in the domain using this technique (the solution) can be represented in a number of ways. Since this example deals with a two-dimensional problem, the solution can be displayed in a number of ways, two of which are the *contour plot* (on which lines of equal potential are drawn through grid points of equal potential), and the *perspective graph* (in which the potential at each grid point is shown as height above that point on the grid, and the points are connected together to produce a surface in three dimensions, see Fig. 2-14).

2.10.1.1 Example of How Device Simulation Programs are Used to Obtain the Two-Dimensional Solution of the Poisson Equation in a MOSFET Structure Operating in Subthreshold. Procedures for obtaining the solution of the 2-D Poisson equation by numerical analysis akin to the one described in the previous discussion are typically carried out with the aid of high-speed computers. The software programs that enable the numerical analysis procedures to be performed are part of such device-simulation packages such as GEMINI, PISCES, MINIMOS, MEDICI, etc. Here we present a brief synopsis of how such a simulation package is used to obtain such a solution.

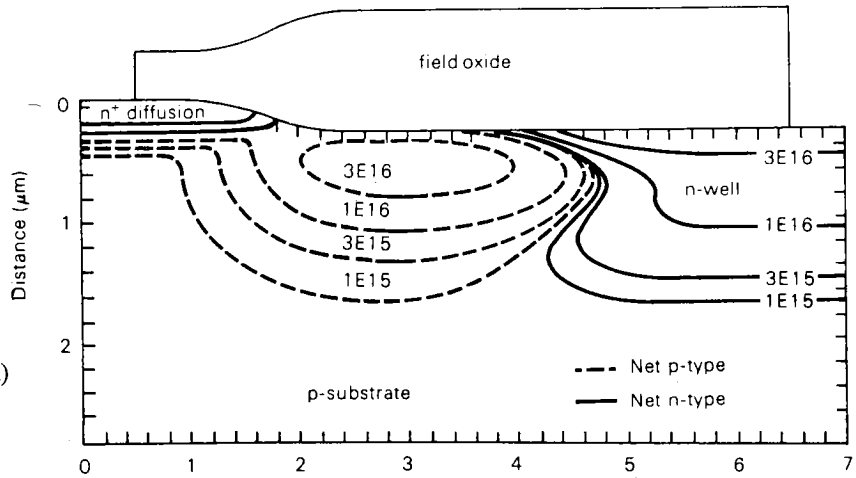
Before we begin this synopsis it should again be pointed out that 2-D finite-difference-based device simulators (e.g., GEMINI and PISCES) typically use the box-integration discretization method rather than the difference-operator method described in the previous section. This permits non-planar device structures to be handled more easily. The form of the algebraic equations that approximate the 2-D Poisson equation is nevertheless essentially the same as that obtained using the difference-operator method. Hence, the matrix techniques for solving them are the same as those described earlier in the chapter. The derivation of the finite difference equations using the box-integration method with a five-point difference approximation is given in Appendix D. There we also demonstrate how the box-integration approach is exploited to facilitate the simulation of non-planar device structures.

Various quantities must be specified as input to the simulation program so that the solution of the Poisson equation for a particular MOSFET structure can be carried out. These include: (1) the material properties of each of the device regions, namely the semiconductor (e.g., silicon), the gate dielectric (e.g., SiO_2), and the gate material (e.g., Al or n^+ poly); (2) the dimensions of the MOSFET (e.g., the gate length, and the lengths of the source and drain); (3) the gate oxide thickness t_{ox}; (4) the operating temperature T; (5) the doping concentrations at each point in the structure being analyzed (normally provided to the device simulator in the form of doping profile data obtained from a 2-D process simulator); and (6) the values of the potential at the boundaries of the solution domain (i.e., at the boundaries of the device structure, including at the source and drain regions, at the gate, at the backside of the device [substrate contact], and at the vertical sides of the device).

Figure 2-16a shows the simulated doping profiles (in contour form) in an example MOSFET structure obtained from a 2-D process simulator.[19] In this case a parasitic MOSFET is analyzed. Such parasitic devices occur in the field regions of *n*-well CMOS structures, with the n^+ region being the source, the *n*-well being the drain, the field oxide being the gate oxide, and an aluminum line running over the field oxide being the gate, see also Fig. 2-16c. The channel is the *p*-substrate region between the n^+ diffusion and the *n*-well, here with a nominal length of 2.4 μm. Notice that in this structure boron ions are implanted under the field oxide to form a *p*-type channel stop between active NMOSFETs. The values of the potential at the boundaries in this example are V_{SS} and V_{BS} = 0V, and $V_{GS} = V_{DS} = V_{DD}$ = 5V.

2-D device simulation programs (e.g., GEMINI, MINIMOS, or PISCES) are capable of using the doping profiles obtained from 2-D process simulators as their input.

a)



b)

```
COMMENT      N+ TO N-WELL MOSFET

STRUCTURE    TEMPERATURE=298 DATA.INP=IN.GEM

ELECTRODE    SOURCE ALUMINUM WIDTH=0.5
ELECTRODE    GATE ALUMINUM WIDTH=5 LEFTEDGE=1
ELECTRODE    DRAIN ALUMINUM WIDTH=0.5

END          STRUCTURE DEFINITION

SOLUTION     DATA.OUT=SAVEFILE.SAV

BIAS         SUBSTRATE POTENTIAL=0
BIAS         SOURCE POTENTIAL=5
BIAS         GATE POTENTIAL=5
BIAS         DRAIN POTENTIAL=5

END          SOLUTION DEF

COMMENTSTEP.VOL   MAXIMUM DELTA.VOL=2.0 STEPS=15 SUMM.OUT=STEP.STP
```

c)

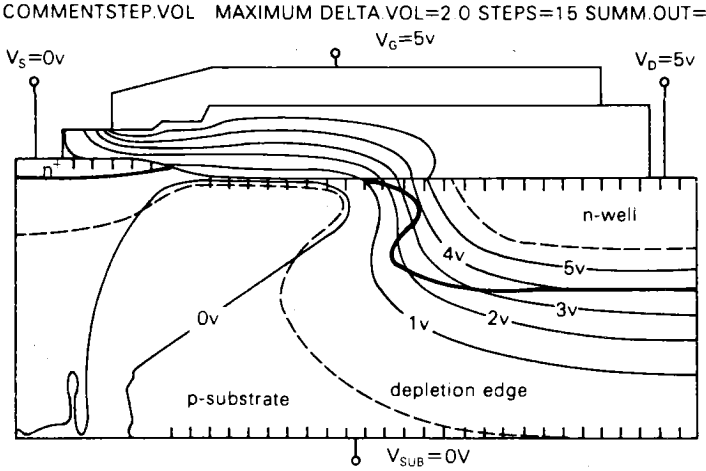


Fig. 2-16 Using GEMINI to analyze the potential in a parasitic MOSFET: (a) the simulated doping contours in an n^+ to n-well parasitic MOSFET; (b) the GEMINI input file; (c) the corresponding 2-D output potential plot as calculated with GEMINI.[19] (© 1986 IEEE).

Figure 2-16b lists the GEMINI input file for simulating the electrostatic potential as a function of distance (i.e., by solving the Poisson equation) for the device structure of Fig. 2-16a. Note that in the STRUCTURE statement of this file, the IN. GEM file used for DATA. INP in the GEMINI input file is from the file saved from running a 2-D process simulator. (That is, one of the last statements in the process simulator input file would be a SAVE statement instructing that the doping profile data is to be saved in the format of a file suitable as input to a particular device simulator - in this case to GEMINI). Also note that the dimensions and type of material of the electrodes at the source and drain contact and the gate are not specified by the process simulator input and hence must be defined. This is done using the three ELECTRODE statements in the example file given here.

Figure 2-16c depicts the output of the GEMINI program for the MOSFET structure of Fig. 2-16a. That is, here the form of the solution of the Poisson equation (the potential as a function of position in the MOSFET) is that of 2-D equipotential contours potential. The depletion region edge for the particular device structure and biasing is also shown (solid heavy black line).

2.10.2 Example of a Numerical Procedure for Solving the One-Dimensional Form of Fick's 2nd Law Equation as a Model for Impurity Diffusion in Silicon

In general, simulation of the redistribution of impurities in silicon under practical processing conditions involves the solution of the concentration-dependent nonlinear Fick's 2nd Law equation in a domain where one of the boundaries (the silicon-oxide interface) is continually and nonuniformly moving in space as a function of time. Analytical solutions of Fick's 2nd Law equation are only available for special, limited cases of process conditions (and usually only in one-dimension), and thus, this approach to obtain solutions of Fick's equation in the more general case equation is inadequate. For most practical process conditions, numerical analysis must be employed to obtain approximate solutions of the equation. Here we describe an example of a set of numerical procedures that can be used to solve the one-dimensional form of the equation.

We will follow the numerical analysis approach used in SUPREM II as described by Antoniadis.[17] In our example, however, the treatment is limited to the case of diffusion of boron or arsenic in silicon in an inert atmosphere. In this way we can maintain mathematical simplicity while still demonstrating the essential steps of the numerical analysis procedure.*

* If we treated the case of diffusion in an oxidizing ambient we would have had to include a boundary condition accounting for dopant segregation at the Si-SiO_2 interface, and would also have had to deal with the condition of a moving boundary (since Si is consumed during the diffusion process by the growth of an oxide on the Si surface). Even though we do not discuss these aspects of the simulation problem, they are, in fact, handled numerically in the SUPREM II and III simulators.

1. *State Fick's 2nd Law Equation in One-Dimensional Form.* The Fick's 2nd Law equation in its general three-dimensional form is

$$\partial C/\partial t = \text{div } F_D \qquad (2-55)$$

where the flux of impurities F_D according to Fick's first law is

$$F_D = -D \nabla C \qquad (2-56)$$

C is the concentration of dopants in the silicon, and D is the diffusion coefficient.

In one-dimension, F_D is written as:

$$F_D(x) = -D(x)\,[dC(x)/dx] \qquad (2-57)$$

and the Fick's Second Law equation has the following form

$$\partial C/\partial t = \partial/\partial x\,\{-D(x)\,[\partial C(x)/\partial x]\} \qquad (2-58a)$$

or

$$\partial C/\partial t = \partial/\partial x\,\{F_D(x)\} \qquad (2-58b)$$

The boundary and initial conditions depend on the particular diffusion process being simulated. For example, the Si surface may have a constant doping concentration during the time of diffusion being analyzed (predeposition step), or a drive-in step may be carried out, either in an inert environment (in which case dopant may be evaporating from the surface during the process), or an oxide may be being grown (in which case, dopant segregation effects at the surface will cause the Si surface concentration to change). Finally, the initial doping profile might have been introduced with a previous ion implantation step, which again would impact the boundary and initial conditions.

2. *Specify the Grid in One-Dimension.* The one-dimensional domain over which Eq. 2-58 is to be solved is partitioned into discrete cells as function of depth beneath the silicon surface (see Fig. 2-17). In the latest versions of SUPREM III, these cells can

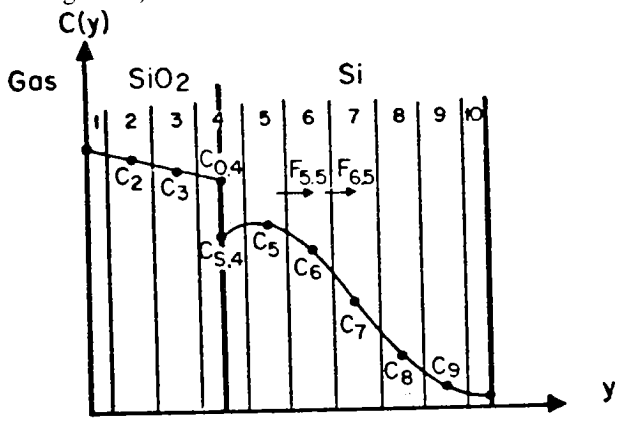


Fig. 2.17 Illustration of the partitioning of the simulation space into discrete cells with impurity fluxes across the cell boundaries.

have nonuniform sizes (specified by the user). The nodes at which the concentration C(x) is evaluated lie at the center of the cells, and these locations will be considered to be the points of the grid. The fluxes which give rise to the change in concentration in a cell, however, are evaluated across the cell boundaries (i.e., at the 'half-points' of the grid). The values of C and F_D are calculated using a discretized version of Eq. 2-58. For example, the concentration of impurities in some cell (say, cell 6, Fig. 2-17) is approximated by the concentration calculated at the grid point at x = 6. In the discretized equation, for purposes of calculation the concentration in each cell is also assumed to be constant throughout the cell. The change in number of impurities per unit area in a cell (ΔQ) from one time step to the next is found by computing the difference of the net fluxes of impurities crossing the cell boundaries (i.e., at x = 6.5 and x = 5.5) during that time interval (neglecting generation-loss terms). The change in the concentration is found by multiplying ΔQ by the width of the cell (e.g., $x_{6.5} - x_{5.5}$).

3. *Discretize the Fick's 2nd Law equation Using the Finite-Difference Method.* To approximate the change in concentration as a function of time we must discretize both the time and space derivatives in the Fick's 2nd Law equation. As discussed in section 2.4.4, time discretization can be done in a number of ways, using either *explicit* or *implicit* schemes. As was noted in section 2.4.4, however, the simple forward-Euler explicit time-discretization method is generally avoided because this approach can be unstable unless excessively small time steps are used.* Instead, the more complex, but also more stable and accurate, implicit time discretization methods are implemented in process (and device) simulators. In fact, the computation times for implicit methods are comparable to those when the explicit method is used, but accuracy in the former is better, and hence they are preferred.

In SUPREM II time-discretization is performed using the second-order implicit Crank-Nicholson technique. In this approach, the unknown parameter (e.g., C[x]) is averaged between its values at the previous step and the next time step. We will illustrate the Crank-Nicholson method for this equation after first describing the space discretization procedure.

To discretize the space derivatives of Eq. 2-58, we first deal with the outer partial derivative with respect to x in the right-hand side of Eq. 2-58b. That is, this operator is replaced with a first-order central difference approximation of the first derivative. Since we wish to evaluate the fluxes at the cell boundaries, we calculate their values at the half-points of the grid, e.g., a cell whose center is at x = i, has boundaries at i + 1/2 and i - 1/2. When this is done - for a cell whose midpoint is at x = i - the right-hand side of Eq. 2-58b becomes

$$\partial/\partial x \{F_D(x)\} \quad \text{---->} \quad -[F_D(x_{i+1/2}) - F_D(x_{i-1/2})]/(x_{i+1} - x_i) \qquad (2-59)$$

* Stability of the forward-Euler explicit method is guaranteed only if the dimensionless quantity $D\Delta t/(\Delta x)^2 = D\Delta t/[(x_{i+1} - x_i)(x_{i+1} - x_i)]$ in Eq. 2-59 does not exceed 0.5. This limit forces one to employ very small values of Δt, leading to very long computational times, since the equations must be solved at each time step.

For example, for cell 6, if the right-hand side of 2-59 is used in Eq. 2-58a, then 2-58b would be written as

$$dC(x = 6)/dt = -[F_D(x_{6.5}) - F_D(x_{5.5})]/(x_6 - x_5) \qquad (2 - 60)$$

where $F_D(x_{6.5})$ and $F_D(x_{5.5})$ are the impurity diffusive fluxes at the right and left-hand boundaries of cell 6, respectively. Next F_D is discretized and these discretized approximations of F_D are substituted into the right-hand side of 2-60. However, it should be noted that in SUPREM II a modified form of the Fick's First Law equation is actually discretized, and not Eq. 2-57. That is, Eq. 2-57 is modified before discretization by changing it to the following form:

$$F_D = -d/dx\,[D(x)\,C(x)] \qquad (2 - 61)$$

Note that this modification is valid for modeling boron and arsenic diffusion in silicon, and is used in SUPREM II for these dopants. For simulating phosphorus diffusion, however, this modification is not valid, and the correct (but more computationally lengthy) form (i.e., Eq. 2-57) should be used. This is done in SUPREM II when phosphorus diffusion is being simulated. For reasons of simplicity we nevertheless elect to use Eq. 2-61 in this example as the basis of the discretized forms of F_D.

According to the Crank-Nicholson method for time discretization, the value of F_D is calculated by taking the average of the first derivative approximations using C(x) at the previous time step ($C[x]^k$) and the next time step ($C[x]^{k+1}$). Hence, the appropriate discretized form of $F_D(x_{i+1/2})$ which is to be substituted in the right-hand side of 2-59 is

$$F_D\,(x_{i+\frac{1}{2}}) = -\frac{1}{2}\left[D(C^k_{i+\frac{1}{2}})\,\frac{C^k_{i+1} - C^k_i}{x_{i+1} - x_i} + D(C^{k+1}_{i+\frac{1}{2}})\,\frac{C^{k+1}_{i+1} - C^{k+1}_i}{x_{i+1} - x_i}\right] \qquad (2 - 62)$$

Then, the fully discretized form of Eq. 2-53a is

$$\frac{C^{k+1}_i - C^k_i}{\Delta t} = -\left(\frac{F_D\,(x_{i+\frac{1}{2}}) - F_D\,(x_{i-\frac{1}{2}})}{(x_{i+1} - x_i)}\right) \qquad (2 - 63)$$

or

$$C^{k+1}_i = C^k_i - \frac{\Delta t}{(x_{i+1} - x_i)}\left(F_D\,(x_{i+\frac{1}{2}}) - F_D\,(x_{i-\frac{1}{2}})\right)$$

Substituting for $F_D(x_i + 1/2)$ and $F_D(x_i - 1/2)$ we get

$$C^{k+1}_i = C^k_i - \frac{\Delta t}{(x_{i+1} - x_i)}\left(-\frac{1}{2}\left[D(C^k_{i+\frac{1}{2}})\,\frac{C^k_{i+1} - C^k_i}{x_{i+1} - x_i} + D(C^{k+1}_{i+\frac{1}{2}})\,\frac{C^{k+1}_{i+1} - C^{k+1}_i}{x_{i+1} - x_i}\right]\right.$$

$$\left. -\frac{1}{2}\left[D(C^k_{i-\frac{1}{2}})\,\frac{C^k_i - C^k_{i-1}}{x_i - x_{i-1}} + D(C^{k+1}_{i-\frac{1}{2}})\,\frac{C^{k+1}_i - C^{k+1}_{i-1}}{x_i - x_{i-1}}\right]\right) \qquad (2 - 64)$$

4. (a) *Format the Set of Discretized Equations to Facilitate a Matrix Solution.* (b) *Linearize the Matrix Equation.* (c) *Solve Using an Appropriate Linear Matrix Solution Technique.* An equation like Eq. 2-64 is formulated for each node of the grid. Thus, we have a set of algebraic equations for our solution space at each time step. Note that the unknown at the next time step (k+1) appears not only in the term describing the location of the cell of interest (e.g., for cell 6, i = 6 in the above equation), but also in the terms of the two neighboring cell concentrations (i - 1 and i + 1). Thus, when these unknowns are collected on the left-hand side of the equation, they will supply three coefficients to one of the rows of the coefficient matrix. The resulting equations are therefore mutually coupled and form a system of simultaneous algebraic equations. For each time step this set of equations must be solved. The result is a set of concentration values (one for each cell) for each time step. The final set of concentration values represents the approximate variation of doping concentration in one dimension after the completion of the diffusion process being simulated.

If the nodes are appropriately ordered, and the equation set is combined to form a matrix equation, the coefficient matrix [**A**] can be tridiagonal (sparse).

$$[\mathbf{A}]\,\{\mathbf{C}\}^{k+1} = \{\mathbf{M}\}^{k} \qquad (2\text{-}65)$$

where $\{\mathbf{M}\}^{k}$ is a vector containing concentration terms calculated from the previous time step, and hence are known. Now, matrix solution techniques, as described earlier, can be employed to efficiently solve Eq. 2-65. Note, however, that the diffusion coefficient D can also be a function of the impurity concentration; in that case, the system of equations becomes nonlinear. If the equation set is nonlinear, linearization techniques, such as the Newton iterative method described in section 2.6 must be used to linearize the set. Then, Newton iterations of the solution are performed until convergence of the results is achieved. In SUPREM II Antoniadis reports that a direct linear matrix solution technique (e.g., Gaussian elimination or variant thereof) is used to solve the linearized equation set.

5. *Display the results.* The results are a set of concentration values (one per cell) at each time step. Most commonly, these values are plotted (usually only the set of values at the last time step) versus distance in the silicon. An example of such a doping profile is shown in Fig. 2-18. Note that a 1-D doping profile like the one shown in Fig. 2-18 can also be extracted form a 2-D doping concentration surface plot (shown in Fig. 2-14e1).

2.11 SYSTEMS OF COUPLED EQUATIONS

We pointed out in chap. 1 that the quantitative model of semiconductor device operation is based on the set of coupled device PDEs (Eqs. 1-3 and 1-4). In addition, the modeling diffusion may also involve the solution of a set of coupled PDEs. Thus, we should briefly mention some of the issues involved in the solution of such coupled equations.

In general, two techniques were used in the 1970s to deal with a coupled set of PDEs. The algorithm flow diagram for these two techniques as applied to the coupled device

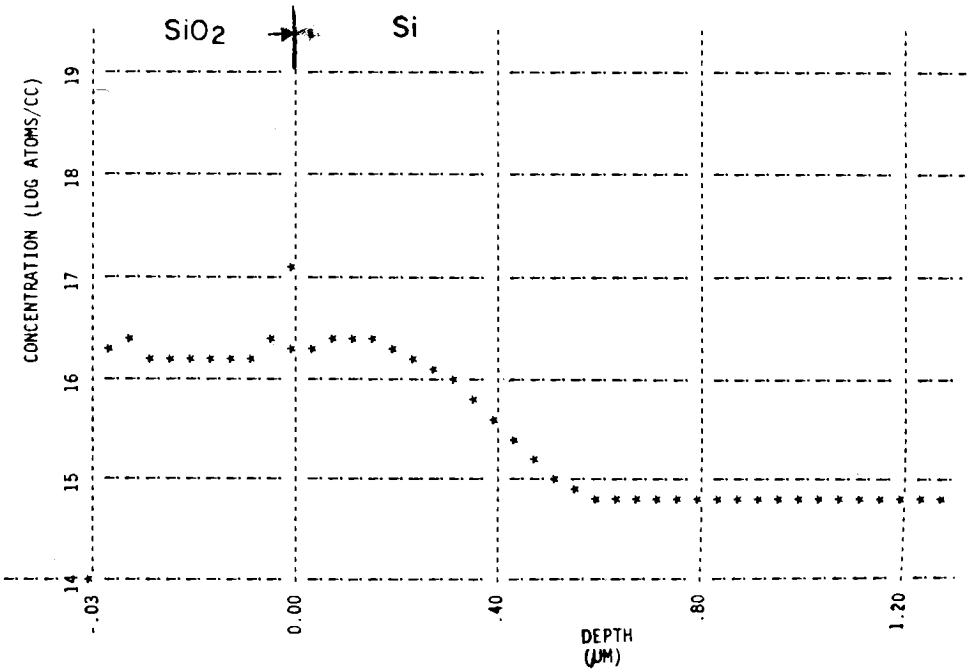


Fig. 2.18 SUPREM III output of boron distribution.

equations is shown in Fig. 2-19. These techniques are respectively: (1) *Gummel's iterative* (or *alternating*, or *decoupled*) method; and (2) *Newtons direct* method (also referred to as the *simultaneous*, or *coupled* method). The comparison between these two techniques is somewhat analogous to the comparison of direct versus iterative matrix solution methods for linear algebraic equations.

2.11.1 Gummel's Iterative Method

In *Gummel's method*, the equations are decoupled and solved sequentially and iteratively. As shown in Fig. 2-19b, Gummel's method begins with an initial guess of potentials and carrier concentrations. Then the Poisson equation is linearized and solved for the potential at each node of the mesh (e.g., using the Newton iteration method, or similar technique to linearize the discretized equations). The carrier concentrations are updated based on the new potentials and appropriate carrier statistics (Boltzmann or Fermi-Dirac), and the Poisson equation is solved again. This procedure is repeated until the potential is below some convergence criteria limit. This is the inner loop of the flow chart of Fig. 2-19b.

Next, the potentials obtained from the solution of the Poisson equation are applied to the current continuity equations. They are then solved for the carrier concentrations (keeping the potentials constant during this solution). The new carrier concentrations obtained require that the Poisson equation be solved again so that the algorithm loops back to the top of the flow chart (outer loop in Fig. 2-19b). Eventually, potentials and

carrier concentrations converge and are consistent with both the Poisson and continuity equations. A direct (e.g., Gaussian elimination) method is often used for each decoupled linearized matrix equation in Gummel's method.

This method is preferred for zero/reverse bias and low-current conditions, in which the equations are loosely coupled (such as the subthreshold region of operation in MOSFETs). Because less work is expended on each pass through the outer loop than in the Newtons direct method, much less memory capacity is required and superior execution speed is obtained. However, the convergence rate of Gummel's method can be very slow if the Poisson equation and the continuity equations are strongly coupled, making the solutions under these conditions very costly. In addition, convergence is not even always guaranteed. As a result, in such cases, the device simulation software packages switches over solving the equations with the Newton method.

2.11.2 Newton's Direct Method

Newtons direct method for solving the device equations involves appending the three matrix equations (the Poisson equation and the two current continuity equations) together to form a matrix that is nine times as large as in the Gummel's method (because the coupling variables, arising from the partial derivatives of all combinations of potential and carrier concentrations for adjacent nodes must be included). Because the matrix becomes so large for grids of any practical size, inverting it requires a large

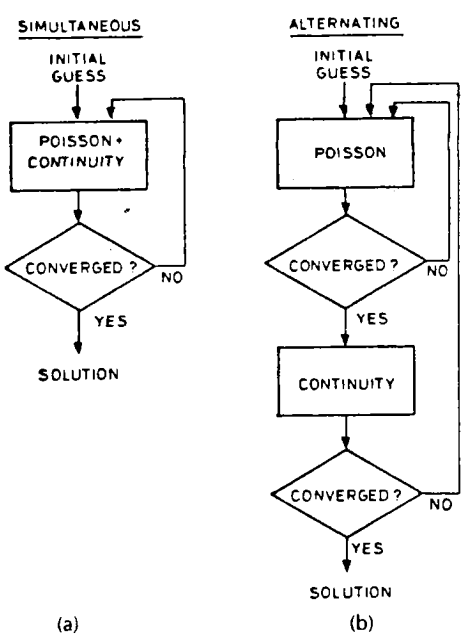


Fig. 2-19 Algorithm flow of simultaneous and alternating methods for solution of the coupled equations. (a) Simultaneous solutions. (b) Iterative solutions.[18] (© 1986 IEEE).

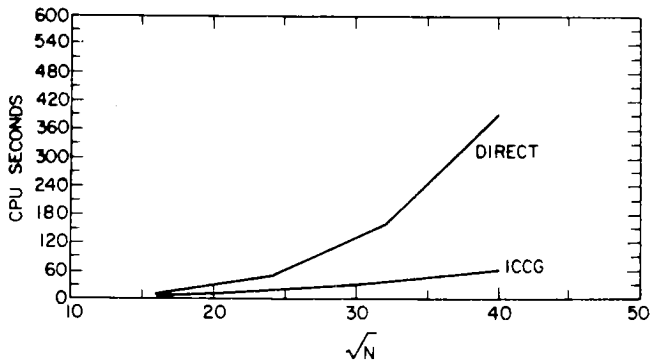


Fig. 2-20 Simulation speed comparison of direct and iterative solution versus square root of grid (VAX 11/780). (© 1986 IEEE).

memory capacity and much longer computing time than the Gummel's method. Hence, this method is results in more costly solutions at lower biases. However, it guarantees convergence at high bias levels. Consequently, it is invoked by the device simulation programs when device operating conditions are such that Gummel's method may not converge (i.e., when a MOSFET is in saturation), or when the equations are strongly coupled and convergence of Gummel's method is slow. The method is also needed for transient and ac analysis. Figure 2-20 illustrates the tradeoffs with number of grid points between direct and iterative solution methods (using an incomplete Cholesky conjugate gradient [ICCG] method as the iterative linear equation solution scheme).

Mixed mode methods for solving coupled systems of equations have also been investigated. Here, the tightly coupled equations are solved together and iterations are made between groups of loosely coupled equation solutions. Coupled PDEs also arise in simulating diffusion of impurities in silicon. Here, the diffusion equations for the impurities are coupled to those of the point defects. The same type of solution methods are applied to such coupled sets of PDEs as to the coupled device equations. A comparison of the estimated relative costs of simulating processing and device behavior using the numerical techniques described in this chapter are given in reference 18.

2.12 EVOLUTION OF DEVICE SIMULATORS

In Vol. 2, chap. 9 a synopsis of the evolution of process simulators is given. Here we trace the evolution of publicly (including commercially) available 2-D (and 3-D) MOSFET device simulators, as shown in Fig. 2-21. We will draw upon the material presented in the previous sections of this chapter as background material. Note that this discussion is not all inclusive. Specifically, a variety of other device simulators have been internally developed by large corporations and/or academic institutions, but since these are not available to the public either from academic or commercial vendors they are not described here.

2.12.1 CADDET. One of the first 2-D device simulators available to the public was CADDET. This simulator was developed by Hitachi and published in 1979.[21] It solves the Poisson equation and the continuity equation for electrons (single carrier analysis). Hence, it is suitable only for simulating FET devices. It uses a automatically generated, non-uniform rectangular grid, with a maximum of 2000 grid points. This early MOSFET simulator had many limitations, including the following: (1) the only doping profiles in the source/drain regions it could handle as input were uniform or Gaussian doping profiles (i.e., doping profiles generated by process simulators could not be used as input to CADDET); (2) only MOSFET structures with planar surface regions could be treated; (3) only single carrier analysis could be performed (i.e., could not simulate many bipolar device phenomena such as impact ionization or latch up, nor CMOS structures), and (4) only static (steady-state) device phenomena could be modeled.

2.12.2 MINIMOS. An improved 2-D FET simulator was introduced next called MINIMOS. It was published by Selberherr et al., at the Technical University of Vienna, in 1980.[22] MINIMOS solves the Poisson equation and both carrier continuity equations using a finite difference method. The program is easy to use and gained wide popularity. It can accept analytical doping profiles or doping profiles from SUPREM-III as input. It converts 1-D doping profiles to 2-D form suitable as input to the 2-D device simulation procedure. While it still finds much use and can be acquired from commercial vendors in its most recent version (MINIMOS4), it is slowly being replaced by more powerful device simulators such as PISCES-II because the following limitations make it less suitable for device structures of the 1990s: (1) MINIMOS can only handle MOSFET structures with planar surfaces; (2) it is a dc (steady-state) simulator, which means it cannot simulate transient MOSFET phenomena; (3) it uses a rectangular grid, which is not as computationally efficient as an irregular (e.g., triangular grid), as described in section 2.3; (4) because it is only a FET simulator, it cannot be used to simulate BiCMOS structures.

2.12.3 GEMINI. A 2-D Poisson solver, GEMINI, originally developed by Greenfield and Dutton at Stanford, was published in 1980.[23] As described earlier, GEMINI was the first 2-D device simulator developed to handle 2-D MOSFET structures with non-planar surfaces. That is, it solves the Poisson equation by using a box-discretization finite-difference method which permits non-planar surfaces to be treated. However, its simulations are valid only under a limited set of device operating conditions. That is, since it does not solve any of the carrier continuity equations, it must rely on analytical methods for calculating the drain current in MOSFETs from the solution that it generates of the Poisson equation. These analytical solution methods are only valid under operating conditions in low levels of current flow in the MOSFET.

The chief advantage that GEMINI offered for performing such device simulation was faster computation speeds than were possible from simulators that also had to solve the carrier continuity equations, and the ability to handle non-planar structures. Yet, advances in computing power has made the former advantage less important, and more powerful simulators such as PISCES-II also offered the ability to handle non-planar

Evolution of 2-D MOSFET Device Simulators

CADDET (1979)

* Solves Poisson Equation & One Carrier Continuity Equation
These equations are solved Iteratively (Gummel iteration)
* Can Only Handle JFET and MOSFET Structures
* Non-Uniform, Planar Rectangular Grid
* Source/Drain Impurity Concentration Profiles must be Gaussian Doping Profiles
* Uses Finite Difference Discretization
* Primitive User Interface

MINIMOS (1980)

* Solves Poisson Equation and 2 Carrier Continuity Equations
* Non-uniform Rectangular Planar Grids (But Offers Adaptive Gridding)
* Can Accept Arbitrary Impurity Concentration Profiles from Process Simulators (SUPREM III) or Doping Profiles Described with Analytical Functions as Input
* Uses Finite-Difference Box-Integration Discretization
Calculates Static-state solutions of I-V characteristics only
* User Friendly

GEMINI (1980)

* Solves Poisson Equation Only
(i.e., Solves for Electrostatic potential $\varphi[x,y]$, then calculates $n[x,y]$ and $p[x,y]$ from $\varphi[x,y]$)
* Non-uniform, non-planar rectangular grid
(Non-planar MOSFET structures can be handled)
* Uses Finite-Difference Box-Integration Discretization
Calculates drain current from analytical expressions (dc values)
* Solutions it provides are only valid for low V_{DS} values in inversion or in subthreshold.

PISCES (1985)

* Solves Poisson Equation and 2 Carrier Continuity Equations
* Non-uniform Triangular Non-Planar Grids
(Also permits Local Adaptive Refinement of Gridding)
* Can handle non-planar structures
* Can handle MOSFET and bipolar structures
* Can perform dc (static) and transient analysis
* Uses finite difference box-integration discretization scheme
* Commercial versions: MEDICI and S-PISCES-2B
* Can simulate:
- Latchup in CMOS
- Hot-electron simulations
- Substrate current in MOSFETs
- Impact ionization
- Ionizing radiation effects

Other Device Simulators: FIELDAY (2-D Finite Element Device Analyzer)
DAVINCI (3-D Device Simulator - Commercial)
THUNDER (3-D Device Simulator - Commercial)

Fig. 2-21 Evolution of Device Simulators.

devices. Thus, GEMINI, like MINIMOS has largely been supplanted by such more powerful simulators. The limitations of GEMINI were the following: (1) it does not provide accurate simulation of MOSFETs when large currents flow in the MOSFET (e.g., when the MOSFET is in inversion and a large V_{DS} is applied). This means that the important saturation region of operation cannot be simulated by GEMINI; (2) GEMINI uses a rectangular grid, which is not as computationally efficient as a triangular grid, and; (3) it does not handle transient cases.

2.12.4. PISCES. A 2-D device simulator, named PISCES-II (**P**oisson and **C**ontinuity **E**quation **S**olver) that is able to overcome the limitations of those just previously discussed, was published in 1984.[5,6] It was developed at Stanford by Pinto, Rafferty, and Dutton. PISCES-II solves both carrier continuity equations and the Poisson equation, and is capable of solving transient as well as steady-state problems. It accepts general doping profiles obtained from SUPREM-III, SUPREM-IV, or from analytic functions. It uses a box-discretization finite-difference, but with a non-uniform triangular grid. This permits non-planar device structures to be handled, and local adaptive gridding to be implemented. Unlike GEMINI, it can successfully simulate on-state MOSFETs in saturation. It also is capable of simulating bipolar transistor devices, allowing BiCMOS device structures to be simulated. Additionally, Fermi-Dirac statistics, incomplete impurity ionization, bandgap narrowing and Schottky barrier lowering are included. The recombination rate is modeled by both Shockley-Read-Hall and Auger band-to-band recombination. Mobility is modeled by Eq. 5-27 in chap. 5. Since it can handle transient problems, it can simulate such events as gate switching behavior, CMOS latchup, and single-event upset. (Note that comparably powerful 2-D device simulators based on finite-element discretization have been developed internally at IBM [FIELDAY],[24] the National Bureau of Standards [B2DE],[25] and other places.) An enhanced version of PISCES-II (dubbed MEDICI™) is now also commercially available from TMA Inc. PISCES-II is a more complex program than GEMINI or MINIMOS, but as computing power has continued to increase, results can now be obtained at reasonable cost with PISCES-II. Since PISCES is capable of performing the tasks of the earlier simulators and more, and the price and performance differential between them continue to decrease, it is increasingly replacing them.

2.12.5. Three-Dimensional Device Simulators. Several 3-D device simulators have been commercially released, including DAVINCI™ (TMA, Inc.), and THUNDER,™ and H-FIELDS-3D™ (Silvaco International). Triangular grids are created on a 2-D plane, and these grids are replicated many times in the third dimension to produce a 3-D grid structure. The program solves the semiconductor equations on each of the grids and assembles the solutions in a format that can be visualized in three dimensions.

REFERENCES

1. A.W. Al-Khafaji and J.R. Tooley, *Numerical Methods in Engineering Practice,* Holt, Rinehart and Winston, New York, 1986.
2. S. Selberherr, *Analysis and Simulation of Semiconductor Devices,* Springer-Verlag, Wien, New York, 1984.
3. C.M. Snowden, *Semiconductor Device Modelling,* Peter Peregrinus, London 1988.
4. P. Antognetti et al., Eds. *Process and Device Simulation for MOS-VLSI Circuits,* Martinus Nijhoff, Boston 1983.
5. M.R. Pinto, C.S. Rafferty, and R.W. Dutton, "PICSES II: Poisson and Continuity Equation Solver," Stanford University, Stanford CA 94305, September 1984.
6. C.S. Rafferty et al., "Iterative Methods in Semiconductor Device Simulation," *IEEE Trans. Electron Dev.* **ED-32**, October 1985, p. 2120.
7. C.W. Steele, *Numerical Computation of Electric and Magnetic Fields* (Van Nostrand Reinhold, New York, 1987.
8. O.C. Zienkiewicz, *The Finite Element Method,* McGraw-Hill, London 1977.
9. L. Lapidus and G.F. Pinder, *Numerical Solution of Partial Differential Equations in Science and Engineering,* John Wiley & Sons, New York, 1982.
10. R.E. Bank et al., "Transient Simulation of Silicon Devices and Circuits," *IEEE Trans. Electron Dev.* **ED-32**, October 1985, p. 1992.
11. E.H. Cuthill and J. McKee, "Reducing the bandwidth of sparse symmetric matrices," Proc. ACM Conference, 1969, p. 157.
12. A. George and J.W.H. Liu, *Computer Solution of Large Sparse Positive Definite Systems,* Prentice-Hall, Inc. Englewood Cliffs, N.J. (1981).
13. R.S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1962.
14. I.A. Duff, "A Survey of Sparse Matrix Research," *Proc. IEEE* Vol. 65, p. 500 1977.
15. D.C. Cole et al., "The Use of Simulation in Semiconductor Technology Development," *Solid-State Electronics,* **33** No. 6, June 1990, p. 591.
16. S.G. Duvall, *IEEE Trans. on Computer-Aided Design,* vol. **CAD-7**, no. 7, July 1988, p. 741.
17. D.A. Antoniadis, "One Dimensional Simulation of IC Fabrication Processes," in *Process and Device Simulation for MOS-VLSI Circuits*, P. Antognetti et al., Eds., Martinus Nijhoff, the Hague, 1983, p. 226.
18. R. Dutton and M. R. Pinto, "The Use of Computer Aids in IC Technology Evolution," *IEEE Proceedings,* Vo. 74, No. 12, Dec. 1986, p. 1730.
19. A.G. Lewis, J.Y. Chen, R.A. Martin and T.Y. Huang, "Device Isolation in High-Density LOCOS-isolated CMOS," *IEEE Trans. Electron Dev.*, **ED-34**, p. 1337, 1987.
20. R.W. Dutton and A. Yu, *Technology CAD: Computer Simulation of IC Processes and Devices.,* Klewer Academic Publishers, Boston 1993.
21. T. Toyabe and S. Asai, "Analytical Models of Threshold Voltage and Breakdown Voltage of Short-Channel MOSFETs Derived from Two-Dimensional Analysis," *IEEE Trans. Electron Dev.,* **ED-26**, p. 453, April 1979.
22. S. Selberherr, A. Schutz, and H.W. Potzl, "MINIMOS - A Two-Dimensional MOS Transistor Analyser," *IEEE Trans. Electron Dev.,* **ED-27**, p. 1540, Aug. 1980.

23. J.A. Greenfield and R.W. Dutton, "Nonplanar VLSI Device Analysis Using the Solution of Poisson's Equation," *IEEE Trans. Electron Dev.*, **ED-27**, p. 1520, Aug. 1980.
24. E.M. Butturla et al., "Finite-Element Analysis of Semiconductor Devices: The FIELDAY Program," *IBM J. Res. Develop.*, vol. **25**, p. 218, July 1981.
25. J.L. Blue and C.L. Wilson, "Two-Dimensional Analysis of Semiconductor Devices Using General-Purpose Interactive PDE Software," *IEEE Trans. Electron Dev.*, **ED-30**, p. 1056, Sept. 1983.

PROBLEMS

1. Determine a root of the following equation using the Newton-Raphson iteration technique:

$$f(x) = x^3 - 3.5x^2 + 2x - 10$$

Note that since this method does not always converge to a root if the initial guess is not sufficiently close to the actual root, it may be necessary to use a graphical technique to get an estimate of the root that will be sufficiently close to the actual root that the numerical procedure will converge to the root.

2. Calculate the boron doping profile and junction depth resulting from a predeposition process performed at 950°C for 30 minutes in a nonoxidizing ambient using the analytical solution of Fick's 2nd Law. Assume the substrate is *n*-type with a background doping level of 1.5×10^{16} atoms/cm^3 and the boron surface concentration reaches solid solubility (C_S = 1.8×10^{20} atoms/cm^3.) Repeat the problem using the SUPREM III process simulator. Compare the results and comment on any differences between the two doping profiles.

3. Calculate the diffusion profile and junction depth after the predeposition of Problem 1 is subjected to a non-oxidizing ambient drive-in at 1050°C for 60 minutes, using the analytical solution of Fick's 2nd Law for this purpose. Repeat this procedure using the SUPREM III process simulator and compare the results with those obtained using the analytical solution. Comment on why the analytical and numerical results of Problem 3 are in better agreement than the results obtained in Problem 2.

4. Consider an *n*-channel MOSFET made with an n^+ polysilicon gate, a gate oxide thickness of 40 nm, and a substrate doping of $N_A = 9 \times 10^{14}$ cm^{-3}. Assume a threshold-adjust implant of 3.2×10^{12} cm^{-2} at 130 keV is also used. (a) Use SUPREM III to simulate the channel doping profile and save an "export" file for further use. (b) Use the "threshold" statement in SUPREM to extract and plot V_T versus V_{BS} for increments of V_{BS} of 0.25V between 0 and 5V. (c) Using the "export" file from SUPREM III, load the results into MINIMOS or PISCES and compute the surface-channel charge over the range of ±0.2V within the value of V_T computed in part (b), by sweeping V_{GS} in 0.02V increments over this range (and assume V_{BS} = 0V).

5. Starting with the contour plot of potential shown in Fig. 2-14c, create for yourself a perspective plot of the potential using the same positional grid.

CHAPTER 3

MOS TRANSISTOR DEVICE PHYSICS: PART 1
BASICS MOS PHYSICS
and
THE MOS CAPACITOR

A basic goal of IC manufacture is to produce circuits having the highest performance at lowest cost. For digital CMOS ICs (the largest category of silicon ICs in the mid-1990s), the main performance criteria are speed, power dissipation, and device packing density. Reduction in device size has been the chief vehicle for achieving these goals. Smaller devices also produce the twin benefits of decreased die size (more chips per wafer and higher fractional yield) which exert downward pressure on die cost. However, the mere reduction of all device dimensions by a constant factor, for example, without paying attention to other processing parameters can adversely impact IC performance. That is, power dissipation may worsen because of increased device leakage currents, or circuit speed may be degraded. Reliability problems which afflict submicron MOSFETs might also be exacerbated, including hot-carrier degradation, gate-oxide wearout, and electromigration. As a result, it is necessary to employ more sophisticated scaling guidelines for device design to ensure that the shrunken devices exhibit electrical behavior comparable to that of previous generations of larger devices. In the case of MOSFETs, the aim of such scaling guidelines is to provide a design path to *short-channel devices* which continue to exhibit the ideal characteristics of *long-channel devices*. (These terms are defined later in the chapter.)

Our purpose here is to present the issues of submicron MOS device design for semiconductor process engineers. An understanding of device physics has become even more important now that MOSFETs have crossed the long-channel frontier into the short-channel (submicron) realm. Device aspects which could once be ignored (because they cause only second-order effects in large MOS devices) are now significant - such as source/drain doping profiles, channel doping profiles, and the device geometry around the gate periphery. Furthermore, if the degree of process control is not increased, variations in these parameters will become larger (on a percentage basis) as the devices become even smaller. Since the uniformity of device characteristics will be adversely impacted, process control must be improved. In addition, better processes will need to

be developed and devices will have to be designed with less process sensitivity. All of these issues can only be discussed in the context of short-channel MOSFET physics. By understanding the relationship between circuit behavior, device design, and process technology of submicron MOSFETs process engineers will be better able to contribute to the task of successfully designing and manufacturing submicron ICs.

The discussion of MOS device physics also serves to set the stage for some later topics of this book (and others of this series), namely the process technology developed to fabricate silicon ICs. To facilitate the understanding of the submicron MOSFET, we divide the topic of *MOS device physics* into three chapters. The first (chap. 3) deals with basic MOS theory and MOS capacitors The second (chap. 4) covers long-channel MOSFETs and the circuit models developed to predict the drain current characteristics of such devices. The last (chap. 5) describes short-channel (i.e., submicron) MOSFETs and issues of MOSFET device scaling.

In each of these chapters the following general approach is adopted. As each new topic is introduced it is first considered from a *qualitative* perspective, using visualization analogies whenever possible. This is intended to provide an intuitive, rudimentary grasp of the device phenomenon being presented. Then the same device effects are analyzed from a *quantitative* perspective, and mathematical relationships that describe the physical effects are developed. Where applicable, as a final task, quantitative models of the device behavior in terms of the terminal voltages and currents are derived. Where possible complicated mathematics and derivations are minimized to prevent them from standing in the way of conceptual explanations. Where such complications cannot be completely circumvented many of the intermediate steps are given. Hopefully this will make the material somewhat more digestible.

3.1 BASIC MOS THEORY and PHYSICS of IDEAL MOS CAPACITORS (MOS-C)

As noted in the introduction, the ultimate purpose of this and the next two chapters is to provide an understanding of short-channel MOSFET behavior (extending to devices with submicron channel lengths). Since the behavior of submicron MOSFETs is complex, the subject is best approached in a step-by-step fashion. That is, the device behavior of the simplest MOS structure (i.e., the two-terminal MOS capacitor, to be referred henceforth as the MOS-C) is considered first. The MOS-C exhibits behavior which is the foundation of both long- and short-channel MOSFET operation. Such behavior is easier to describe in the MOS-C since some of the device phenomena exhibited in more complex MOS structures (i.e., multi-dimensional electric fields and drain currents) are absent in the MOS-C. Once the reader acquires a good understanding of MOS-C behavior, these effects can be added, allowing a relatively straightforward explanation of long-channel MOSFET operation. Thereafter, submicron MOSFET behavior can be explored by showing how the long-channel MOSFET behavior is altered as the channel length is reduced. With this rationale in mind, we begin by describing the behavior of the MOS capacitor in this chapter - initially dealing with the *ideal* MOS capacitor, but later incorporating *nonideal* effects. We should also point out

that the MOS-C itself is quite a complex and subtle structure. In fact, a lengthy and rigorous text that deals comprehensively with these subtleties has been written by Nicollian and Brews.[1]

3.1.1 The Structure of the MOS-C and the MOSFET

Although we begin by studying the MOS-C, it is useful to compare its structure with that of the MOSFET. To do this, an even-more-fundamental configuration, the generic MOS structure is first defined. The *MOS* or *metal-oxide-semiconductor* structure is a three-layered object. It consists of a conducting layer (referred to as the *gate*), atop an insulator layer (usually silicon dioxide), and a semiconductor substrate beneath the insulator (termed either the *substrate* or *body*). The substrate can be doped either *n*-type or *p*-type. Note that the acronym MOS stems from the early days of MOS fabrication when all gates were made of metal (aluminum). While heavily doped polysilicon has replaced aluminum as the standard gate material, the name MOS continues to be used.

The structure of the *MOS capacitor* (or *MOS-C*) is shown in Fig. 3-1a. It consists of the generic MOS structure with terminals G and B connected to the gate and substrate, respectively, making it a *two-terminal device*. An external bias can be applied across these terminals, called the *gate-substrate (or gate-body) voltage* V_{GB}. In Fig. 3-1a1 we show a perspective view of a MOS-C having a *p*-type Si substrate. Then in Fig. 3-1a2 we show the cross-section of this device cut along the middle of the gate. This cross-section is shown lying flat Fig. 3-1a3, and in Fig. 3-1a4 the same structure with only the Si beneath the oxide is shown.

The structure of the *MOS field-effect transistor* (or *MOSFET*) is given in Fig. 3-1b. It has two regions of doping opposite that of the substrate, one at each edge of the MOS structure. These regions are called the *source* and *drain*, and a *pn* junction exists between them and the substrate. When terminals are connected to all the various regions of the MOSFET a *four-terminal device* results, with the terminals designated as G (gate), S (source), D (drain), and B (substrate).

We next observe that since one of these terminals can be designated as the common (or reference) terminal, three independent terminal voltages can be applied to the MOSFET (e.g., if B is chosen as common we will have V_{GB}, V_{SB}, and V_{DB}). However, only one significant current exists in an ideal MOSFET. That is, we assume that the gate current is zero ($I_G = 0$), and that the source and drain junctions are always kept under reverse bias during normal MOSFET operation. Since reverse bias current in a *pn* junction can be considered negligible (and $I_G = 0$), substrate current will also be inconsequential ($I_{sub} \approx 0$).* Thus, only the drain current I_D which flows between the source and drain in the MOSFET needs to be considered. In summary, three independent terminal voltages and one current (i.e., I_D) are generally associated with the operation of the ideal MOSFET. In Fig. 3-1b1 we show the MOSFET in perspective, then in cross-section (Fig. 3-1b2), and finally (in Fig. 3-1b3), with the cross-section lying flat.

Figure 3-1 also defines the axes to be used in our discussion of both the MOS-C and the MOSFET structures. That is, the *x*-direction is perpendicular to the Si-SiO_2

* We'll remove these restrictions in chaps. 7 and 9.

interface (*vertical* direction), with $x = 0$ at the Si surface. The y-direction is parallel to the Si surface in the direction from source to drain (*lateral* or *longitudinal* direction). Since the source and drain are separated by a distance L (termed the *channel length*), we usually choose $y = 0$ at the source end of the channel, and $y = L$ at the drain end. The z-direction is the other direction parallel to the Si surface (perpendicular to the y-direction), and defines the channel width.

3.1.1.1. Comparison of MOS-Cs and Conventional Capacitor Structures. It should be noted at the outset that the structure and behavior of the MOS-C differs from that of the more conventional metal-dielectric-metal capacitor structure. That is, in the latter structure both terminals are highly conductive, but in the MOS-C the semiconductor terminal is not of high conductivity. Consequently, if a potential difference is applied across the terminals of the conventional capacitor, an electric field arises within the dielectric, but is zero outside the dielectric. This implies that the "metal-dielectric-metal" capacitor stores all of the charge at the interfaces of the respective metal electrodes and the dielectric.

The situation is different in the MOS-C where one of the terminals of the capacitor is a semiconductor, not a metal. If we consider the case of a p-type semiconductor, then the application of a *negative* bias will induce a negative charge on the gate and a positive charge on the substrate. The high conductivity of the gate requires that the electric field within the dielectric again terminates at the gate-dielectric interface. Here, the charge on which the field lines terminate is the negative charge stored on the gate. At the substrate-dielectric interface, however, the semiconductor must supply positive charge to balance the gate's negative charge. This can readily occur because the majority mobile charge carriers in p-type Si are holes, implying that (1) they exist in great abundance and (2) that they are capable of moving to the interface in response to the negative bias on the gate. In this case, the p-type silicon MOS-C and the conventional capacitor structure behave identically. Furthermore, as we will describe in more detail in the next section, this bias condition is known as "accumulation" because it causes majority carriers to accumulate at the semiconductor surface.

On the other hand, if a *positive* gate bias is applied to this MOS-C, a somewhat different set of phenomena occur. That is, a positive gate potential produces a positive charge on the gate, and a negative charge must arise in the substrate to balance the gate charge. In this case, there are very few mobile negative charge carriers (electrons) in the p-substrate. Thus, their scarcity prevents them from supplying enough of the required negative charge, especially at small positive gate-bias values. Instead, the repulsion of holes (the majority carriers) by the positive bias drives the holes away from the substrate-dielectric interface. This forms a region depleted of mobile carriers (i.e., a *depletion region*) that penetrates from the surface into the substrate. The departure of holes from this surface region leaves it negatively charged because the immobile, ionized acceptor impurities remain within the region. This negative "background" charge thus becomes the negative charge induced by the positive gate potential. As a result, the electric field lines terminate on these negative charges. Unlike in the case of accumulation, however, the terminal charge of the MOS-C in the semiconductor is no

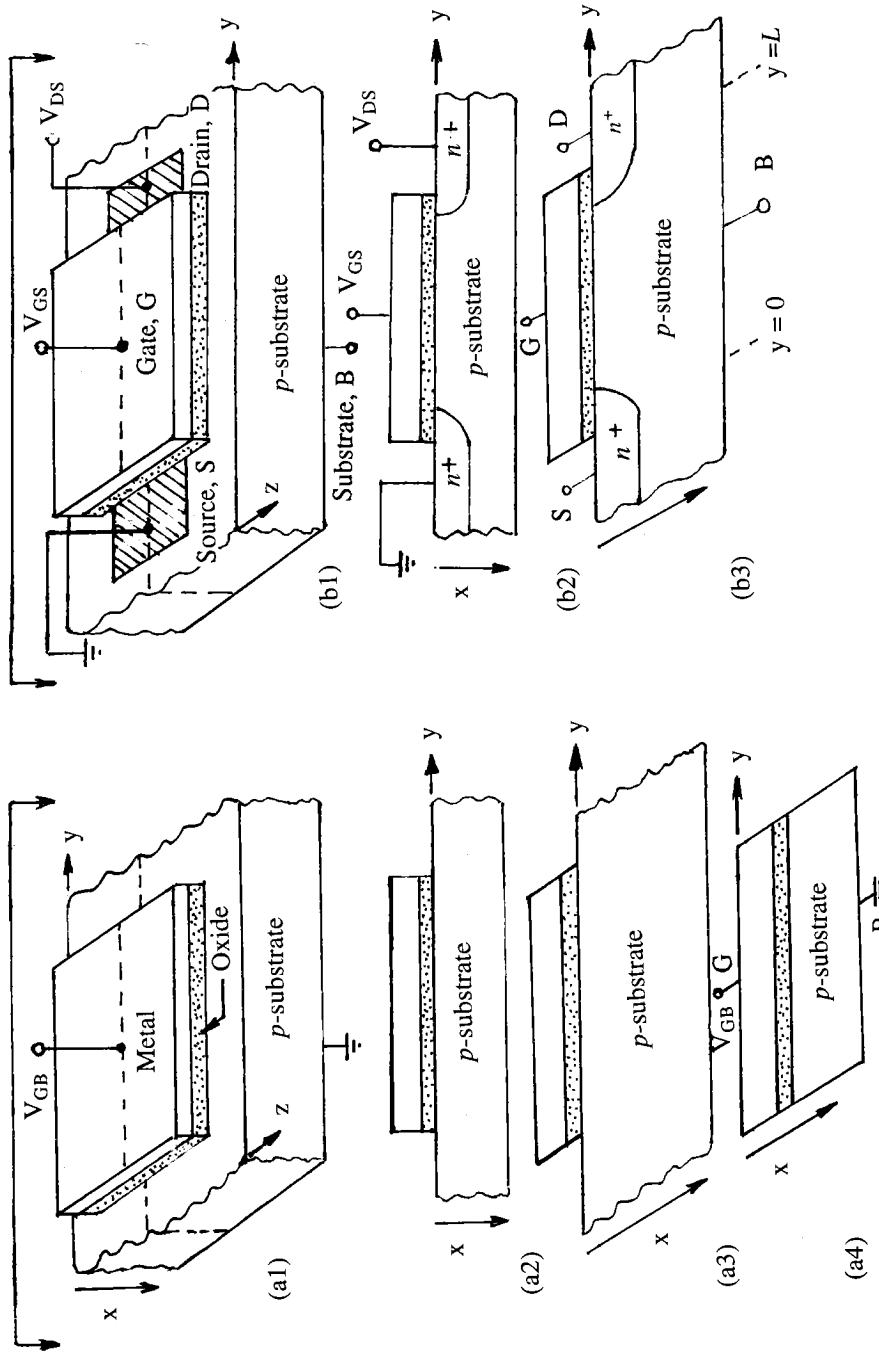


Fig. 3-1 (a1) Perspective view of the MOS-C, (a2) Cross-section of the MOS-C cut along the middle of the structure, (a3) Cross-section shown lying flat, (a4) Cross-section with only the Si beneath the oxide shown. (b1) Perspective view of the MOSFET structure, (b2) Cross-section of the MOSFET cut down the middle of the channel, (b3) Cross-section of the MOSFET lying flat.

longer confined to the substrate-dielectric interface (i.e., to the Si surface). Instead, the charge is spread over some distance extending to the depth of the depletion region. This added distance increases the "effective" dielectric thickness and thus decreases the capacitance of the structure. Thus, unlike a conventional capacitor, whose capacitance is constant regardless of the applied bias, the capacitance of the MOS-C varies with the applied bias under some bias conditions.

Note, however, that if the value of the positive gate bias becomes sufficiently large, a large number of minority carrier electrons will be attracted to the surface. They will collect there and come to dominate the negative charge in the substrate. This induced charge due to the electrons at the surface is called an *inversion layer.* Note also that these phenomena will be described in more detail in the upcoming section, but the purpose of briefly introducing them here is to point out the fact that MOS-Cs indeed behave in a somewhat different fashion than conventional capacitor structures.

3.1.1.2 Why is the MOS-C Studied First? As illustrated above, the MOS-C forms the heart of the MOSFET. Yet it is a fundamental structure in its own right, just as the *pn* junction is a fundamental structure apart from the bipolar transistor. Yet, much in the same way that understanding *pn* junctions is central to understanding the operation of the bipolar transistor, an understanding of the MOS-C is central to understanding of MOSFETs. Hence, we must study behavior of MOS-Cs first.

We can see this by noting that the basis of MOSFET behavior essentially involves two phenomena: (1) the action of mobile carriers in the semiconductor in response to the application of a gate bias; and (2) the transport of these mobile carriers between the source and drain (in the form of a drain current, I_D) in response to the application of a drain bias. Therefore, to quantitatively describe MOSFET operation, initially we need to establish a quantitative model which allows the first of these two phenomena to be treated (i.e., we need to be able to determine the mobile charge distribution in a MOS structure for a given gate bias). Once this charge distribution is known, it can be used to develop a model for the relationship between I_D and the terminal voltages (i.e., the gate and drain biases). Such a model for mobile-charge versus gate-bias is most easily developed in MOS-C structures for the following reasons:

1. We noted that the MOS capacitor is a two-terminal device. If we assume the oxide is a perfect insulator, then the current through the oxide I_G is zero for dc biases between the gate and substrate. Since this is the only possible current in the MOS-C - and its value in ideal structures is zero - the redistribution of charges in the semiconductor with respect to changes in V_{GB} can be analyzed without the complication of current flow. Furthermore, since no current flows under dc bias, the behavior of MOS-Cs in such cases represents an electrostatic problem. Thus, only one of the five classical device equations introduced in chap. 1 needs to be solved (i.e., the Poisson equation) to completely characterize the potential and mobile charge distribution in the ideal MOS-C (i.e., even if $V_{GB} \neq 0$).

2. The electric field in the MOS-C is perpendicular to the silicon surface, making it essentially uni-directional (vertical). The electrostatic potential in the device can

thus be found by solving the Poisson equation in its least complex (e.g., one-dimensional) form. On the other hand, if a bias is also applied between the source and drain of a MOSFET, the direction of the electric field in the semiconductor region of the MOSFET is no longer perpendicular to the silicon surface. That is, it generally contains components parallel as well as vertical to the silicon surface. Hence, to calculate the electrostatic potential in a MOSFET, it is in general necessary to solve the Poisson equation in its two- or three-dimensional form.

3. The absence of current flow in a MOS-C implies that equilibrium is maintained in the semiconductor region even if a dc V_{GB} is applied.* Thus, the equations used to calculate carrier concentrations in terms of equilibrium concepts (e.g., the Fermi level) can be used. This makes the solution of the Poisson equation simpler as well.

4. The MOS-C is widely used as the test structure for characterizing the reliability of the gate oxide film with respect to oxide breakdown and hot-electron degradation (as described in chaps. 7 and 9, respectively). Thus, the study of the MOS-C behavior is intrinsically useful for interpreting data obtained from such structures.

5. The quantitative relationships we develop for the mobile charge concentration in the channel region of the MOS-C under inversion are used in the I_D models of the long-channel MOSFET by invoking the *gradual-channel approximation.* This approximation will be examined in more detail in chap. 4.

3.2 QUALITATIVE DESCRIPTION OF MOS-C DEVICE BEHAVIOR

The behavior of the MOS-C from a qualitative perspective is considered first, namely a verbal description is given of how the mobile charge distribution in the semiconductor depends on the applied dc gate bias. To help visualize these phenomena we make use of energy-band and potential representations, as well as electric-field and charge distribution diagrams.

* The term *equilibrium* should be considered here to mean *thermal equilibrium,* as viewed from the microscopic perspective. That is, the net energy in the semiconductor is stored in the crystal-lattice vibrations (phonons) as well as by free carrier motion. Although external sources of energy (such as incident photoelectric radiation) may also excite a crystal, many situations exist where no such external forces act. In such cases, the total energy is a function of only the crystal temperature, and the semiconductor will spontaneously reach a state referred to as thermal equilibrium. In thermal equilibrium every process is exactly balanced by its inverse process. One such set of competing processes in a semiconductor involves carrier generation and recombination. At thermal equilibrium the rates of these two processes are equal (and are also assumed to be driven only by thermal processes). This makes it valid to use the equilibrium carrier concentration expressions (i.e., Eqs. 3-1a and 3-1b) to calculate the carrier concentrations in the MOS-C even when $V_{GB} \neq 0V$.

3.2.1 Energy-Band Diagram Representation of Semiconductor Devices

One of the most helpful tools for visualizing semiconductor phenomena is the *energy-band diagram.* In such diagrams, the vertical axis represents the energy possessed by electrons when they occupy energy levels at a given location in the device. (In one-dimensional representations, a location in the device is generally specified by a point x, which is designated as the distance along the horizontal axis of the diagram). Such one-dimensional representations are adequate for treating the MOS-C (with the *x*-direction representing the direction perpendicular to the silicon surface).

In the energy-band representation of a semiconductor (Fig. 3-2), E_C and E_V are the conduction and valence band edges, respectively. The intrinsic Fermi level E_i is at the middle of the band gap (i.e., $E_i \cong [E_C+E_V]/2$), corresponding to the Fermi-level of an intrinsic (undoped) semiconductor (Fig. 3-2a). In *n*-type doped silicon the Fermi level E_F is closer to E_C than to E_V (Fig. 3-2b), but in *p*-type it is closer to E_V (Fig. 3-2c) If an electron occupies a state at the bottom edge of the conduction band (E_C), its kinetic energy is defined to be zero (point A in Fig. 3-2d). As the kinetic energy of such an electron increases, it occupies energy levels higher in the conduction band (e.g., point B in Fig. 3-2d). If a semiconductor is in equilibrium, the Fermi level is invariant with position in the material, and this is depicted in a one-dimensional energy-band-versus-distance diagram as E_F being a straight, horizontal line (as in all of Fig. 3-2).

Assuming Maxwell-Boltzmann statistics, the electron and hole concentrations in a semiconductor at equilibrium (n_O and p_O, respectively) can be expressed in terms of quantities found in the energy-band diagram (E_F, E_i), as well as T, and n_i (the intrinsic carrier concentration at T) according to the following expressions:

$$n_O = n_i \exp (E_F - E_i)/kT \qquad (3 - 1a)$$

$$p_O = n_i \exp (E_i - E_F)/kT \qquad (3 - 1b)$$

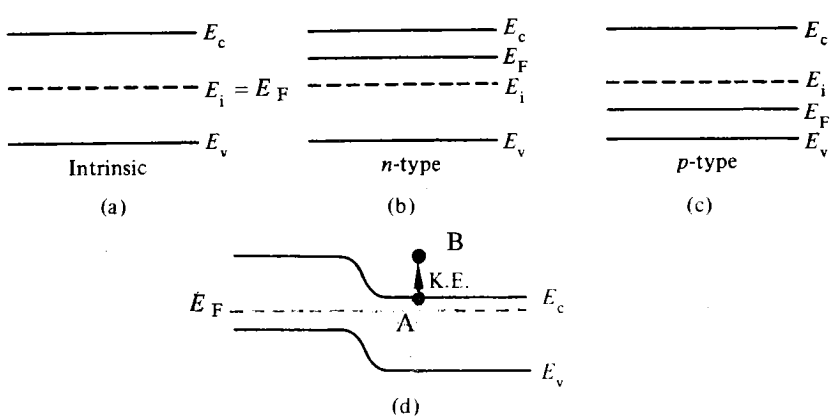


Fig. 3-2 Energy band representation of a semiconductor: (a) intrinsic; (b) *n*-type; (c) *p*-type; (d) Depicting the kinetic energy of an electron in the energy band diagram.

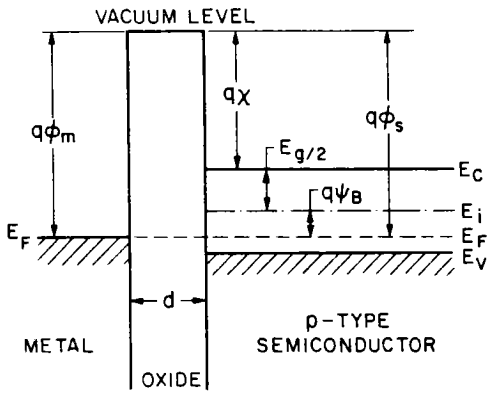


Fig. 3-3 Energy band diagram of an ideal MOS-C when V_{GB} = 0V (*flat-band condition*).

Note also that at equilibrium

$$n_o p_o = n_i^2 \qquad (3 - 1c)$$

3.2.2 Energy Band Diagram of the Ideal MOS Capacitor

Using the energy-band diagram representation we now consider the qualitative behavior of an *ideal* MOS-C subjected to dc bias. The following assumptions are assumed to be valid in an ideal MOS-C:

1. The energy difference between the work function of the gate material (metal), $q\varphi_m$ and that of the semiconductor, $q\varphi_s$ is zero (i.e., $q\varphi_{ms} = 0$);
2. No charges exist in the MOS-C under any biasing conditions, other than those existing in the semiconductor and those with equal but opposite sign on the gate, i.e., within the oxide there are no mobile charge carriers nor any charge centers;
3. No current can flow through the oxide under dc-biasing conditions ($I_G = 0$);
4. The semiconductor region is uniformly doped;
5. The gate layer is considered to be an equipotential region.

For our example we will consider an ideal MOS-C with a *p*-type semiconductor region.

First consider the case when no bias is applied (V_{GB} = 0V). The energy-band diagram for this case is shown in Fig. 3-3. We note that since equilibrium exists in the semiconductor and $q\varphi_{ms} = 0$ (Assumption #1), the Fermi level throughout the semiconductor must be invariant with position, and it must also line up with the Fermi level in the gate. This also implies that E_c, E_v, and E_i in the semiconductor are horizontal and straight (i.e., flat). An energy-band diagram in which E_c, E_v, and E_i are flat is said to be in a *flat-band condition. Thus, when V_{GB} = 0 in an ideal MOS capacitor, it's energy-band diagram is in a flat-band condition.* Note also that in *p*-type Si, E_F is below E_i, by an amount expressed as $|E_F - E_i|$.

Next let a dc bias be applied between the gate and substrate. Since $I_G = 0$ (Assumption #3), the silicon region remains in equilibrium, implying that E_F remains flat. This aspect of MOS-C behavior is unique, since the application of a dc bias across the electrodes of other semiconductor devices will cause at least some current to flow. Hence, MOS-C analysis remains simple under dc bias because the Fermi-level concept is still valid. Although the semiconductor region remains in equilibrium under dc bias (and E_F remains flat), the other bands are altered from that of the $V_{GB} = 0V$ case. This is seen in Fig. 3-4 where four different gate-bias conditions are depicted. Here in Fig. 3-4a we depict the MOS-C in cross-section, as was also shown in Fig. 3-1b2. This structure is then rotated as shown in Fig. 3-4b, and the energy-band and charge distribution versus x diagrams we show next refer to the MOS-C cross-section oriented in the manner of Fig. 3-4b.

3.2.2.1 Accumulation. In the first biasing case a *negative* bias is applied ($V_{GB}<0V$, Fig. 3-4c). The negative bias causes holes (the majority carriers in *p*-type Si) to be attracted to the silicon surface adjacent to the oxide. More holes accumulate near the Si surface than if $V_{GB} = 0V$. The effect of attracting additional majority carriers to the semiconductor surface is called *accumulation*. The conductivity of the region near the Si surface is increased by the presence of the extra holes.

Note that in the energy-band representation, extra holes at one location of a semiconductor (e.g., in this case, at the Si surface) implies that E_F must be closer to E_V than in a region containing fewer holes (e.g., the region deeper below the surface [bulk region]). However, since the *Fermi level* in the semiconductor must remain flat, the energy bands in the semiconductor (E_C, E_V, and E_i) near the surface must bend upward to reflect the presence of the extra holes at that location compared to the *bulk* (Fig. 3-4c1)

The applied gate bias also causes the Fermi levels in the gate and semiconductor regions to be separated by an amount equal to qV_{GB}, or

$$E_F(\text{gate}) - E_F(\text{semiconductor}) = -qV_{GB} \qquad (3 - 2)$$

Although the capacitance of the MOS-C varies with applied dc voltage under some bias conditions, in the accumulation mode it does not. That is, in this mode the capacitance per unit area remains constant as V_{GB} is changed, and is given by $C_{OX} = \kappa_{OX}\varepsilon_O/t_{OX}$, where κ_{OX} and t_{OX} are the dielectric constant and thickness of the oxide, and ε_O is the permittivity of free space.

We now introduce the *charge-distribution versus distance* diagram as an additional visualization tool. This allows us to depict the approximate charge distribution inside the MOS structure. For the case of accumulation, applying a negative bias causes more electrons to move onto the gate. Holes are attracted toward the silicon surface in response. Thus, the charge inside the device as a function of position can be approximated as shown in Fig. 3-4c2. Note that a squared-off (or block) representation of the charge is used in this simple plot. Such block approximations are only intended to be a qualitative representation of the actual charge distribution, which, in general, is not distributed so simply. When interpreting the magnitude and spatial extent of the charge using such simplified diagrams this fact should be kept in mind.

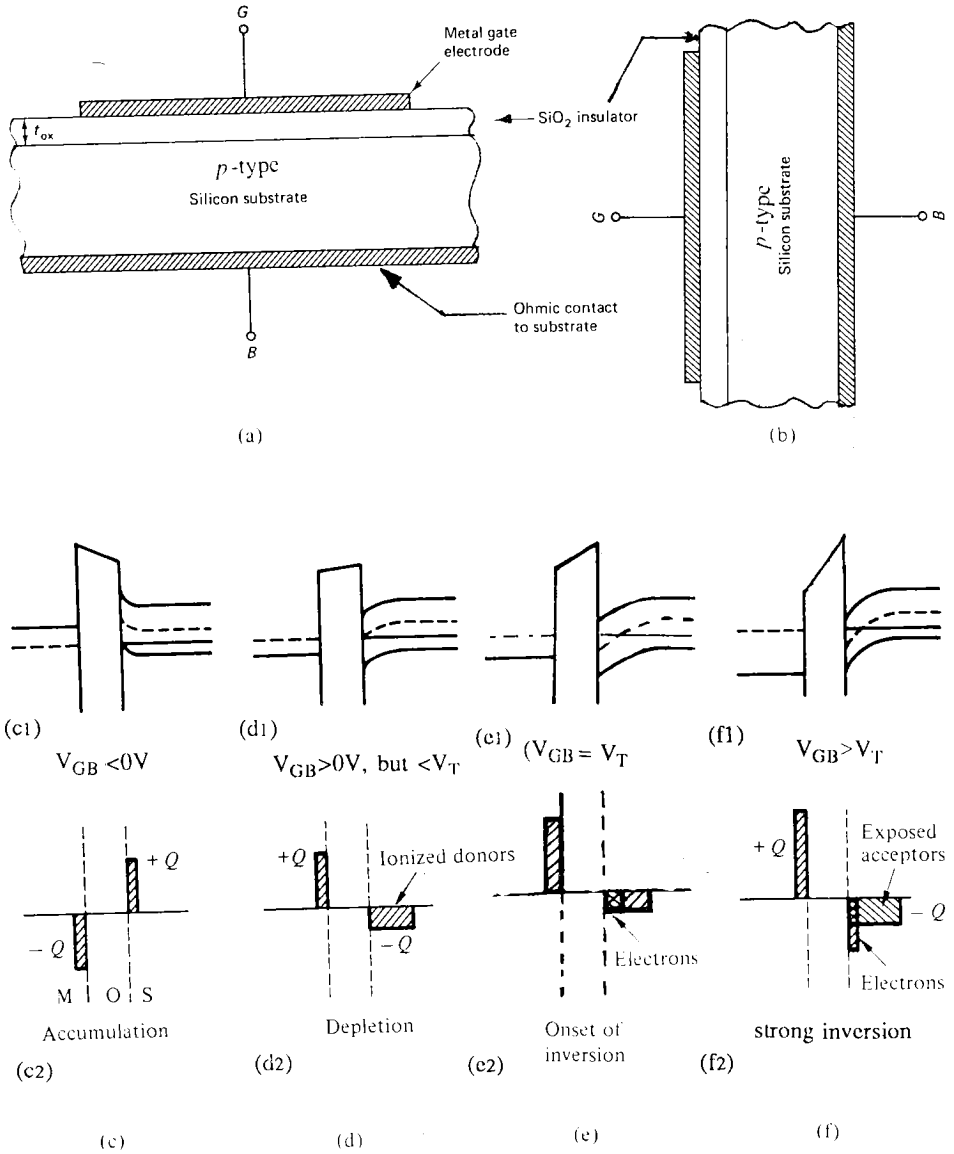


Fig. 3-4 (a) Cross-section of a MOS-C with p-substrate; (b) Cross-section of the same MOS-C, but rotated 90°; (c1) and (c2) Energy band and block charge diagrams when the MOS-C is biased to accumulation ($V_{GB}<0V$); (d1) and (d2) Same diagrams for depletion biasing ($V_{GB}>0V$, but $<V_T$); (e1) and (e2) Same diagrams at onset of inversion ($V_{GB} = V_T$); (f1) and (f2) Same diagrams at strong inversion ($V_{GB}>V_T$).

3.2.2.2 Depletion. In the second bias condition (Fig. 3-4d), a *small positive* voltage is applied to the gate of the *p*-type MOS capacitor. This causes holes in the semiconductor to be pushed away from the surface, creating a depletion region of width d, consisting primarily of negatively charged acceptor ions. In the energy-band representation of this phenomenon the bands bend downward near the surface (Fig. 3-4d1). In this mode of operation, known as *depletion,* the MOS capacitor exhibits two capacitances in series - the oxide capacitance per unit area C_{ox} and the depletion capacitance per unit area $C_d = \kappa_{si}\varepsilon_o/d$. While C_{ox} is constant, C_d is not. That is, C_d decreases as V_{GB} is made larger (more positive). The approximate charge distribution for this bias condition is shown in Fig. 3-4d2.

3.2.2.3 Inversion. If the positive voltage applied to the gate is increased, the bands are bent further downward, eventually causing E_F to cross over E_i and to be closer to E_c than to E_v at the surface. At this point the semiconductor surface is *inverted* (i.e., it has been changed from *p*-type to *n*-type) because there are now more electrons than holes at the surface.*

Another way to view this situation is to remember that the kinetic energy of an electron is zero when it occupies an energy state at the conduction band edge (corresponding to an energy E_c). Thus, under depletion or inversion biasing electrons near the Si surface having zero kinetic energy occupy energy states of lower *total* energy than electrons deeper in the Si having zero kinetic energy. Hence, when a positive bias is applied to the gate, electrons are more likely to fill such states near the surface. Consequently, electrons will be found at higher concentrations closer to the surface when $V_{GB}>0V$.

3.2.2.4 Strong Inversion. If V_{GB} is made even more positive, the bands will continue to bend further downward, until E_i at the surface is eventually below E_F by an energy equal to $|E_F - E_i|$. At this point the electron density per unit volume at the surface n_{surf} (located in a thin layer known as the *inversion layer*) is equal to the hole density in the bulk. This is known as the point of *onset of strong inversion*, and in this condition

$$E_i(\text{bulk}) - E_i(\text{surface}) = 2[E_i(\text{bulk}) - E_F] \qquad (3 - 3)$$

Solving this equation for E_i(surface), and using the result in Eq. 3-1a, the concentration of electrons at the Si surface at the onset of strong inversion, n_{surf}(inv) is

$$n_{surf}(\text{inv}) = n_i \exp[E_F - E_i(\text{surface})]/kT = n_i \exp[E_i(\text{bulk}) - E_F]/kT = p_{obulk} \approx N_A \quad (3 - 4)$$

Figure 3-4e1 is the energy-band diagram depicting the condition of onset of strong

* The surface is actually inverted as soon as E_F crosses over E_i. However, the density of electrons at the surface remains small until E_F is considerably greater than E_i. The mode of operation when the electron density is smaller than the hole density in the substrate bulk is referred to as *weak inversion.* Weak inversion is important in submicron MOSFET behavior and will be discussed in detail in section 4.7.

inversion. In this diagram the width of the depletion region is denoted by d_{max}. The block charge diagram at the onset of strong inversion is shown in Fig. 3-4e2. Note that the voltage applied between the gate and substrate at the onset of strong inversion is known as the threshold voltage, V_T.

For further increases in positive bias, the electron concentration in the inversion layer at the surface exceeds the bulk equilibrium hole concentration, $p_o = N_A$. This is referred to as *strong inversion.* The energy-band and block-charge diagrams depicting the strong inversion condition are shown in Figs. 3-4f1 and 3-4f2, respectively.

In general, the charges on the gate Q_G of a MOS-C must be balanced by the charges in the semiconductor Q_C

$$Q_G\ (C/cm^2) = -Q_C\ (C/cm^2) \qquad (3 - 5)$$

In strong inversion, the charge per unit area Q_C in a *p*-type semiconductor is the sum of the electrons in the inversion layer (Q_I) and the negative charges in the depletion region (Q_{Bmax}). Thus, in strong inversion

$$Q_G = -Q_C = -(Q_I + Q_{Bmax}) \qquad (3 - 6)$$

where $Q_{Bmax} = -qN_A d_{max}$. (Note that Q's in this chapter are expressed in units of charge/unit area - specifically, C/cm^2.) The total capacitance per unit area at strong inversion is the oxide capacitance C_{ox} in series with the depletion capacitance at the maximum depletion width in equilibrium $C_{dmax} = \kappa_{si}\varepsilon_o/d_{max}$. While the above discussion applies to a MOS-C with a *p*-type semiconductor, similar results can be obtained for an *n*-type semiconductor, with the polarity of the voltages and charges being reversed. A typical value of Q_I in a MOS-C or MOSFET in strong inversion is on the order of 10^{12} cm^{-2}.

3.2.2.5 Microscopic Perspective of the Mobile Charge Behavior in a MOS-C. It is also useful to consider from a (qualitative) microscopic perspective the phenomena involving charge carriers near the silicon surface of the MOS-C under conditions of depletion and inversion biasing. That is, when the MOS-C is thus biased, a potential difference between the surface and the bulk exists, giving rise to a potential energy barrier (PEB) with respect to the majority carriers. The majority carriers (holes in our example) are thereby repelled from the surface, and minority carriers (i.e., electrons in this example) are attracted toward it. In either inversion or depletion a region largely devoid of carriers is established (depletion region), and in this region thermal generation of carriers continuously produces holes and electrons in an attempt to re-establish the mobile charge concentration values present in the charge-neutral regions at equilibrium. The holes so produced, however, are continuously repelled from the surface, while the generated electrons are likewise drifted toward the surface.

In a state of thermal equilibrium, some holes in the bulk nevertheless still possess sufficient energy to overcome the PEB at the surface, and they are able to penetrate the depletion region so that recombination with the electrons attracted to the surface can occur. At thermal equilibrium, the rate of such recombination at the surface exactly equals the rate of electron generation in the depletion region, and the electron and hole

concentrations are thus maintained at a fixed, specific value, dependent on the temperature, doping concentration, and gate bias.

If the gate bias is instantaneously changed from one fixed value to another, the carrier concentrations must also change to reflect this new bias value. However, since an increase in the electron concentration relies essentially on their generation in the depletion region by thermal processes, some time must elapse before the new equilibrium electron concentration can be established. In fact, the response of a silicon MOS-C due to instantaneous changes in V_{GB} is relatively slow because thermal generation and recombination processes in the silicon substrate occur at low rates at room temperature. The situation that exists in a MOSFET, as we will describe in chap. 4, is quite different. That is, the presence of diffused regions of doping type opposite that of the substrate (i.e., the source/drain regions) permit electrons (under favorable bias conditions) to be injected into (or removed from) this surface region much more rapidly than if only thermal processes are responsible for contributing carriers (see section 4.2).

3.3 QUANTITATIVE DESCRIPTION OF MOS-C BEHAVIOR

We now consider the behavior of the ideal MOS-C from a quantitative perspective. Specifically, we seek to obtain mathematical relationships that permit us to calculate the following parameters in terms of the applied dc gate bias V_{GB}, the device structure parameters (i.e., t_{ox} and the doping concentration in the silicon), and the material parameters (i.e., ε_{si} and ε_{ox}):

1. The electrostatic potential as a function of position $\varphi(x)$;
2. The surface potential, φ_{surf}, which the potential at $x = 0$, or $\varphi(0)$
3. The electric field as a function of position $\mathcal{E}(x)$;
4. The mobile charge concentrations $n(x)$ and $p(x)$ in the silicon;
5. The bulk charge per unit area Q_B in the silicon;
6. The inversion layer charge per unit area Q_I

Later, we will treat the nonideal MOS-C in which we will also include the effects due to the properties of the gate material and the charges in the oxide and at the Si-SiO_2 interface. We will also describe how to calculate the threshold voltage of MOS structures.

Three quantitative models of MOS-C behavior will be developed, differing in their complexity, accuracy, and generality. The simplest MOS-C model, based on the *depletion approximation* will be derived first. It is also the least general and least accurate of the three models, and provides accurate values of the mobile charge concentrations in the silicon only beyond strong inversion. Nevertheless, this model offers easy-to-interpret analytical expressions of MOS-C behavior, and is the model developed in most basic device physics texts. It is also the basis of the simpler long-channel MOSFET models described in chap. 4.

The second MOS-C model we derive is the most general and accurate model (in that it is valid for the accumulation, depletion, and inversion regimes of MOS-C operation), but it provides these benefits only at the expense of much greater mathematical complexity. That is, the second model does not provide analytical expressions for the quantities listed above, but requires numerical analysis for specific values to be obtained. Despite its mathematical complexity, this model is presented for several important reasons:

1. The MOS-C parameters listed above when calculated using this model are so accurate compared to experimental measurements that the model represents a benchmark to which the results of the other models can be compared to establish their accuracy.

2. It is the basis for the most general long-channel MOSFET model we present in chap. 4 (i.e., the Pao-Sah model).

3. It illustrates the principle that numerical analysis must be relied upon to obtain the most accurate quantitative representations even for the cases of the MOS-C and the long-channel MOSFET.

4. It implies that numerical analysis must certainly be employed to obtain accurate predictions of submicron MOSFET behavior.

However, the complexity of this model is such that it is not feasible to use it for constructing models of MOSFET circuit behavior, nor does it provide analytical expressions for the MOS-C quantities listed above. Thus, quantitative insights about MOS-C behavior are difficult to extract from this model, making it a poor tool for learning about MOS-C behavior. As a result, the previous simple model and a third model (which we describe last), find much wider use than this one.

The third MOS-C model is intermediate in accuracy, generality, and complexity between the other two models. It is referred to as the *charge sheet model,* and it calculates the characteristics of the MOS-C quite accurately over the entire range of inversion (i.e., from weak to strong inversion). It requires more computational effort than the simple model, but far less than the general model. Unlike the general model it provides analytical expressions for the mobile charge density which also extend beyond the operating regimes treated by the simple model. These expressions provide quantitative insight about the behavior of device operating regions that become more important in short-channel MOSFETs.

Note that we will designate the externally applied (i.e., terminal) voltages with the symbol V (e.g., V_{GB}), while the voltage drops internal to the MOS structure will be expressed with the symbol φ (e.g., the voltage drop across the oxide φ_{ox} and the voltage drop across the semiconductor from the Si surface to the bulk, φ_{surf}). Note, however, that both the external and internal voltage drops have units of volts.

3.3.1 The Electrostatic Potential in MOS Structures

Before setting out to derive the expressions listed above, let us make some introductory remarks about the *electrostatic potential* φ. First, we note that since the bias voltages applied to MOS devices are expressed in units of volts, it would be preferable if the quantitative MOS-C relationships we seek were also expressed in the same units (rather than in terms of units of energy, which would result if we used the energy-band representation of MOS behavior). In addition, one of the five classical semiconductor device equations is the Poisson equation, and its solution yields the electrostatic potential in the device. Hence, two of the relationships we obtain when solving the 1-D Poisson equation in the semiconductor region of a MOS-C involve the electrostatic potential, i.e.,

a. The relationship between V_{GB} and the *surface electrostatic potential* φ_{surf} (henceforth referred to as the *surface potential*), and;

b. An expression for the electrostatic potential versus position $\varphi(x)$ [which uses φ_{surf} obtained from **a** as one of the boundary values for calculating this function].

It turns out that the electrostatic potential at any location in a MOS device is related to the potential energy an electron possesses at that location. That is, the electrostatic potential-difference $\Delta\varphi$ observed in moving from one position to another in the semiconductor (assuming the kinetic energy of the electron is the same at both positions) can be computed (in terms of an electron) from

$$\Delta\varphi = \Delta E/(-q) \qquad (3 - 7)$$

where q is the magnitude of the electron's charge. It is evident from Eq. 3-7 that if an electron undergoes a *potential energy* decrease, it means that a *potential* increase exists between two locations, because of the algebraic sign of the electron charge. Thus, for the case of an electron, an energy-band diagram can be converted to a potential diagram by merely inverting one of the bands E_c, E_i, or E_v (as shown in Fig. 3-5, which depicts both representations).* Note that *downward* band bending while going, for example, from the bulk to the surface in a MOS structure, implies an *increase* in the electrostatic potential when moving between those same locations.

Since the potential is a relative quantity, the zero value of φ can be arbitrarily chosen. In our discussion we select the zero value as the electrostatic potential in the bulk (i.e., at $x = \infty$). We also choose the energy level of the intrinsic Fermi level E_i in the bulk [E_i(bulk)] to designate this level of zero on an energy-band diagram. Thus, elsewhere, the value of $\varphi(x)$ is measured with respect to E_i(bulk) according to

$$\varphi(x) = [E_i(\text{bulk}) - E_i(x)]/q \qquad (3 - 8)$$

* Note that since an electron at E_c will have the same kinetic energy (zero) at any location, the bending of the E_c band directly indicates the change in potential energy versus location.

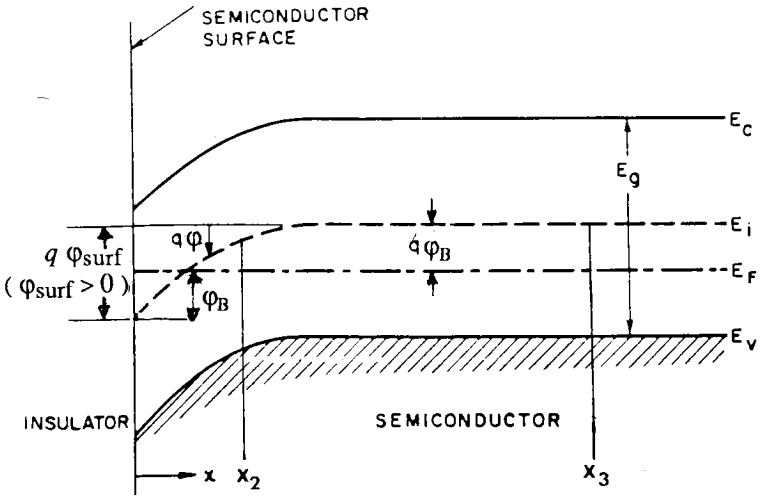


Fig. 3-5 The energy-band and potential diagram representations of the MOS-C (in this figure a MOS-C with a *p*-type substrate and biased to inversion is depicted).

as depicted in the figure. For example, in the bulk of the semiconductor (position x_3 in Fig. 3-5), $E_i(x_3) = E_i(\text{bulk})$. Hence, from Eq. 3-8 the potential at x_3 is 0. At x_2, however, $\varphi(x_2) \neq 0$, but instead is $\varphi(x_2) = [E_i(\text{bulk}) - E_i(x_2)]/q$.

Another consequence of choosing $\varphi\ (x = \infty) = 0$ is this: If we apply a bias to the gate of a MOS-C, part of the voltage will be dropped across the oxide φ_{ox} and part across the silicon φ_{si}, or,

$$V_{GB} = \varphi_{ox} + \varphi_{si} \tag{3-9}$$

If $\varphi(x = \infty) = 0$, then the potential at the surface of the silicon (φ_{surf}) represents the voltage drop from the bulk ($x = \infty$) to the silicon surface ($x = 0$) in a MOS structure. In that case we can write Eq. 3-9 as

$$V_{GB} = \varphi_{ox} + \varphi_{surf} \tag{3-10}$$

Now we see that if we can derive an expression for φ_{ox} in terms of φ_{surf}, an equation relating V_{GB} and φ_{surf} results. We show how this is done in the next section.

Let us also introduce another dimensionless parameter called the *doping parameter* φ_B which relates the potential of an electron at the Fermi level (in the bulk of the silicon in a MOS structure) to the doping concentration. The doping parameter is defined as

$$\varphi_B = [E_i(\text{bulk}) - E_F]/q \approx (kT/q)\ln(N_A/n_i)_{p\text{-type Si}} \approx (kT/q)\ln(n_i/N_D)_{n\text{-type Si}} \tag{3-11}$$

in the bulk region of the Si in the MOS capacitor, see Fig. 3-5). Note that $\varphi_B > 0$ for a *p*-type semiconductor and $\varphi_B < 0$ for an *n*-type semiconductor. For substrate doping concentrations usually encountered in MOSFETs (10^{14}-10^{17} cm^{-3}), the magnitude of φ_B is between $9kT/q$ and $16kT/q$ (where kT/q at 300K is ~0.025 V); that is, the magnitude of φ_B is between 225 and 400 mV at 300K for this range of doping concentrations.

Also, the value of the surface potential at the onset of weak and strong inversion can be expressed in terms of φ_B. That is, in section 3.2 we defined the onset of *weak inversion* as the condition when the electron concentration at the surface of *p*-type Si equals n_i. At that point, $\varphi_{surf} = \varphi_B$. Since the onset of *strong inversion* is the condition when the surface electron concentration equals N_A, at that point E_i (surf) - E_i (bulk) = 2 [E_F - E_i (bulk)]. By combining this fact with Eqs. 3-4 and 3-11, we see that the surface potential at the onset of strong inversion $\varphi_{surf}(inv) = 2\varphi_B$.

Note that it is also possible to express the equilibrium concentrations of electrons n_p (x) and p_p (x) in terms of $\varphi(x)$. That is, when the zero value of φ is defined as above

$$n_p(x) = n_{po} \exp[q\varphi(x)/kT] \quad (3 - 12a)$$

$$p_p(x) = p_{po} \exp[-q\varphi(x)/kT] \quad (3 - 12b)$$

where n_{po} and p_{po} are the equilibrium concentrations of electrons and holes in the bulk of the silicon (in this case in *p*-type Si). At the surface of the silicon, the concentrations n_{surf} and p_{surf} are given by

$$n_{surf} = n_{po} \exp(q\varphi_{surf}/kT) \quad (3 - 13a)$$

$$p_{surf} = p_{po} \exp(-q\varphi_{surf}/kT) \quad (3 - 13b)$$

Using Eq. 3-11 in Eq. 3-13a, we can express the electron concentration at the surface in terms of N_A and φ_B as:

$$n_{surf} = N_A \exp[q(\varphi_{surf} - 2\varphi_B)/kT] \quad (3 - 13c)$$

This expression is plotted versus φ_{surf} in Fig. 3-6. This figure reiterates several aspects of MOS behavior established with the energy-band representation of MOS characteristics. First, using Eq. 3-11 we see that when $\varphi_{surf} = \varphi_B$, n_{surf} becomes equal to the intrinsic concentration, n_i. Thus, $\varphi_{surf} = \varphi_B$ is defined as the limit point between the regions of depletion and weak inversion. Second, as φ_{surf} increases above φ_B, n_{surf} grows very rapidly. Eventually, when $\varphi_{surf} = 2\varphi_B$, $n_{surf} = N_A$. This is commonly defined as the demarcation point between weak and strong inversion. Finally, beyond this point (i.e., $\varphi_{surf}>2\varphi_B$), n_{surf} continues to increase rapidly, becoming much greater than N_A for small increases of φ_{surf} beyond $2\varphi_B$. This implies that in strong inversion the inversion layer charge density Q_I is generally much larger than the ionic charge density in the depletion region Q_B. An additional important point can also be inferred from this figure. That is, if the potential as a function of x is known we can calculate $n_p(x)$ from Eq. 3-12a.

3.3.2 General Approach to Deriving MOS-C Characteristics

Now we can set out to derive the expressions listed at the beginning of section 3.3. This can be done in four stages, assuming we are given V_{GB}, the oxide thickness t_{ox}, and the doping concentration N_A.

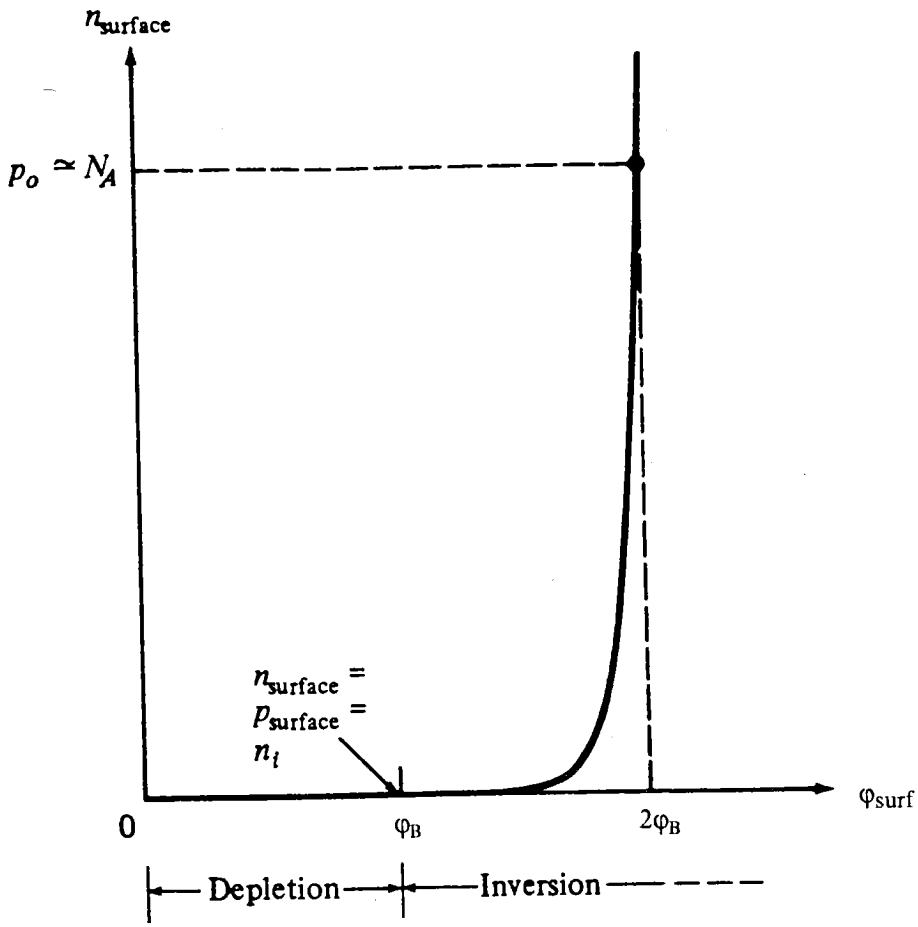


Fig. 3-6 Electron concentration at the surface of a MOS-C (with a *p*-type substrate) as a function of φ_{surf}.[6] Linear axes are used. Reprinted with permission of McGraw-Hill Book Co.

1. We first find a relationship between V_{GB} and φ_{surf} in the MOS structure. This is done by recalling that if we have an expression for φ_{ox} in terms of φ_{surf}, it can be substituted in Eq. 3-10 to yield an equation for φ_{surf} in terms of V_{GB}.

We begin the derivation of such an expression by noting that the oxide is considered to be an ideal insulator, with no mobile carriers or charge centers. Hence, the electric field in the oxide must be constant, or $\mathcal{E}_{ox} = - d\varphi_{ox}/dx =$ constant. Thus, the voltage across the oxide φ_{ox} is found from:

$$\varphi_{ox} = -\int_{-t_{ox}}^{0} \mathcal{E}_{ox}\, dx = t_{ox}\mathcal{E}_{ox} \qquad (3 - 14)$$

The next step is to relate $\mathcal{E}_{ox}$ to the electric field in the semiconductor. Since the charge per unit area located at the Si-SiO_2 interface of an ideal MOS-C is assumed to be zero, the electric displacement, $D = \varepsilon\mathcal{E}$ perpendicular to the interface, is continuous [i.e., $D_{ox} = D_{si}(x = 0)$]. The field in the oxide is therefore related to the field at the surface of the silicon $\mathcal{E}_{si}$ (x = 0) by the equation

$$\mathcal{E}_{ox} = (\kappa_{si}/\kappa_{ox})\, \mathcal{E}_{si}\ (x = 0) \qquad (3 - 15)$$

Now, using Eq. 3-15 in Eq. 3-14 we can write φ_{ox} as

$$\varphi_{ox} = t_{ox}\, (\kappa_{si}/\kappa_{ox})\, \mathcal{E}_{si}\ (x = 0) \qquad (3 - 16)$$

and this result can be substituted for φ_{ox} in Eq. 3-10. In order to complete our goal of finding an expression for φ_{surf} in terms of V_{GB} we see that an expression for $\mathcal{E}_{si}$ (x = 0) in terms of φ_{surf} is needed. Since $\mathcal{E}(x) = -\, d\varphi/dx$, such expression can be obtained by integrating the one-dimensional form of the Poisson equation once, using the appropriate total charge density $\rho(x)$, and the boundary conditions $\mathcal{E}\ (x = \infty) = 0$, and $\varphi\ (x = 0) = \varphi_{surf}$. We will carry out this integration for three cases in the upcoming sections.

2. We next solve the Poisson equation to find $\varphi(x)$. Since we have noted that the electric field is normal to the oxide surface in a MOS-C (neglecting fringing fields at the edges of the capacitor), we can consider the variation in electrostatic potential φ in a MOS-C to be one-dimensional (i.e., we assume $\partial\varphi/\partial y$ and $\partial\varphi/\partial z = 0$). Thus, it is only necessary to solve the Poisson equation in its one-dimensional form. We use the value of φ_{surf} from step **1** as one boundary condition [$\varphi\ (x = 0) = \varphi_{surf}$], and [$\varphi\ (x = \infty) = 0$] as the other, when we solve the Poisson equation to find $\varphi\ (x)$.

3. To calculate the mobile charge concentrations $n_p(x)$ and $p_p(x)$, we can use the values of $\varphi\ (x)$ found from step (2) in Eqs. 3-12.

4. To find expressions for Q_B and Q_I as a function of gate bias we proceed as follows:

a. First, we recall Gauss' law, which states that the charge contained in a volume equals the permittivity times the electric field emanating from the volume. Applying Gauss' law to a volume extending from the silicon surface to the field-free bulk region, which contains a total charge per unit area Q_C, we can write*

$$Q_C = -\, \kappa_{si}\varepsilon_o\, \mathcal{E}_{si}\ (x = 0) \qquad (3 - 17)$$

* Note that only the $\mathcal{E}$-field at the Si-SiO_2 interface (which represents the top surface of the volume in which the charge in the semiconductor exists) must be included in Eq. 3-17 because the $\mathcal{E}$-field does not emanate from any other surfaces of the volume. This is because: (1) the $\mathcal{E}$–field emanating from the bottom of the volume under consideration (i.e., at x = d) is assumed to be zero, and; (2) since the $\mathcal{E}$-field in the MOS-C is assumed to be unidirectional, $\mathcal{E}_y = \mathcal{E}_z = 0$. Thus, the $\mathcal{E}$-field emanating from the sides of the volume is also zero.

We also note from Eq. 3-16 and from the fact that $C_{ox} = \kappa_{ox}\varepsilon_o/t_{ox}$ that

$$\varphi_{ox} = t_{ox}\,(\kappa_{si}/\kappa_{ox})\,\mathcal{E}_{si}\,(x = 0) = \{\kappa_{si}\varepsilon_o\,\mathcal{E}_{si}\,(x = 0)\}/C_{ox} \quad (3 - 18)$$

Thus, we can also write

$$Q_C = -\varphi_{ox}\,C_{ox} \quad (3 - 19)$$

b. Next, we note that the charge per unit area in the semiconductor Q_C is the sum of the inversion layer charge per unit area Q_I and the ionic charge per unit area Q_B. So we can write

$$Q_C = Q_I + Q_B \quad (3 - 20)$$

In general, Q_I can be found by calculating the total mobile charge in the inversion layer (in our example, the electrons). In other words, to get Q_I, we could integrate Eq. 3-12a from x = 0 to the depth x_c at which the electron concentration becomes negligible, or

$$Q_I(x) = -q\int_0^{x_c} n(x)\,dx \quad (3 - 21)$$

Then, Q_B could be found by subtracting Q_I from Q_C.

However, this approach to calculating Q_I and Q_B is quite complex, in fact, it requires that $\varphi(x)$ be computed first, and then this result be used to integrate $n(\varphi)$. Both of these calculations, in general, require numerical integration. Although we shall outline this approach in more detail, we will also illustrate simpler methods for calculating Q_I and Q_B which still provide values accurate enough for many applications.

3.3.3 Finding the Quantitative Characteristics of a MOS-C when the Depletion Approximation is Invoked

Let us derive the quantitative characteristics of the ideal MOS-C under some assumptions that simplify the mathematical treatment outlined above, but still provide expressions that are accurate enough for many applications. The simplest of such quantitative analyses assumes that we deal with the ideal uniformly doped MOS-C, and this analysis is also the one carried out in most basic device textbooks. Here we invoke the following approximations: (1) the *depletion approximation* is assumed to be valid; (2) the *one-sided abrupt junction approximation* is assumed; and (3) other simplifying approximations are assumed which allow the depletion region width d, the surface potential φ_{surf}, and the inversion layer charge Q_I to be simply computed beyond the point of strong inversion.

The *depletion approximation* invoked in this case is based on the following simplifying assumptions:

1. The mobile charges in the depletion region are totally absent in the depletion region for values of V_{GB} smaller than those needed to cause inversion (i.e., $n = p = 0$ in the depletion region over this range of gate-bias values).

2. For values of gate-bias voltages beyond inversion, the depletion region is still totally devoid of mobile carriers, except in the inversion layer, which is assumed to be a very thin layer near the silicon surface.

The *one-sided abrupt-junction approximation* assumes that the depletion region extends to a depth d, beyond which the silicon returns abruptly to a condition in which the mobile carrier densities are again equal to their equilibrium values [i.e., at $x = d$, $p = N_A = p_{po}$, and $n = n_{po}$].

The simplifying assumptions dealing with the d (depletion region width), φ_{surf} (surface potential), and Q_I (mobile charges in the inversion layer) are the following:

1. Beyond strong inversion, additional positive bias on the gate simply draws more electrons into the inversion layer. This negative electron charge balances the additional positive gate charge, and the depletion-region width d no longer increases. That is, d remains fixed at the value attained at the onset of strong inversion, and this value is denoted as d_{max}.

2. Since the depletion region no longer increases once the inversion layer is formed, the bands of the energy band diagram are not bent any further with increasing positive bias. Hence, φ_{surf} also remains fixed (or is *clamped*) at the value it has at the onset of inversion, namely at φ_{surf} (inv) = $2\varphi_B$.

3. Beyond strong inversion, the ionic charge per unit area Q_B contained in the depletion region remains unchanged, since $d = d_{max}$. Hence, we can compute Q_B beyond inversion simply from a knowledge of N_A and d_{max}. Also if V_{GB} is given, and we know φ_{surf} beyond strong inversion, we can compute φ_{ox} from Eq. 3-10, and from this we can obtain Q_C. Then, Q_I can be found by subtracting Q_B from Q_C.

Using these assumptions, we set out to find the desired expressions. We initially solve the Poisson equation in the semiconductor region for $\varphi(x)$. That is, under depletion biasing conditions we have assumed $n = p = 0$, and so the charge density $\rho(x)$ is merely a constant term ($q\, N_A/\kappa_{si}\varepsilon_o$). In this case, the Poisson equation has the form:

$$\begin{aligned} d^2\varphi/dx^2 &= q\, N_A/\kappa_{si}\varepsilon_o \qquad & 0 < x < d \\ &= 0 & x > d \end{aligned} \qquad (3 - 22)$$

This is an ordinary, linear differential equation that can easily be integrated twice to yield a closed-form solution of $\varphi(x)$ in the semiconductor. That is, assuming the boundary conditions $\varphi(x = d) = 0$, and $\varphi(x = 0) = \varphi_{surf}$ we get,

$$\varphi(x) = \varphi_{surf}(1 - x/d)^2 \qquad (3 - 23)$$

where d is the depletion-region depth. The electric field $\mathcal{E}$ (x) is given by $-d\varphi/dx$ or

$$\mathcal{E}(x) = -d\varphi/dx = 2\,\varphi_{surf}(1 - x/d)(-1/d) \qquad (3 - 24)$$

and

$$d^2\varphi/dx^2 = 2\,\varphi_{surf}/d^2 \qquad \text{[from (3 - 23)]} \qquad (3 - 25)$$

$$= q\,N_A/\kappa_{si}\varepsilon_o \qquad \text{[from (3 - 22)]} \qquad (3 - 26)$$

Then from (3 - 25) and (3 - 26) we can find d in terms of φ_{surf} and N_A

$$d = [\kappa_{si}\varepsilon_o(2\varphi_{surf})/qN_A]^{1/2}. \qquad (3 - 27)$$

Beyond the point of strong inversion $\varphi_{surf}(inv) = 2\varphi_B$, and so d_{max} is then

$$d_{max} = (4\kappa_{si}\varepsilon_o\varphi_B/qN_A)^{1/2} \qquad (3 - 28)$$

Next, we solve Eq. 3-10 for φ_{surf}. In this case, let us initially assume that the MOS-C is biased into depletion, but below inversion. In that case, no inversion layer exists ($Q_I = 0$), thus, $Q_C = Q_B$ and Q_B is given by

$$Q_B = q\,N_A\,d \qquad (3 - 29)$$

Since $\varphi_{ox} = -Q_C/C_{ox} = -Q_B/C_{ox}$, by using Eq. 3-29 in this result, we get

$$\varphi_{ox} = -q\,N_A\,d/C_{ox} = \{[\kappa_{si}\varepsilon_o(2\varphi_{surf})qN_A]^{1/2}\}/C_{ox} \qquad (3 - 30)$$

Now if we let $\eta = \{[2\kappa_{si}\varepsilon_o qN_A]^{1/2}\}/C_{ox}$, we can write $\varphi_{ox} = \eta\sqrt{\varphi_{surf}}$. Substituting this into Eq. 3-10 we get

$$V_{GB} = \varphi_{surf} + \eta\sqrt{\varphi_{surf}} \qquad (3 - 31)$$

If we let $\alpha = \sqrt{\varphi_{surf}}$, we can write Eq. 3-31 as a quadratic equation

$$0 = \alpha^2 + \eta\alpha - V_{GB} \qquad (3 - 32)$$

We can now easily solve this for α, using the quadratic formula. Once we have α we can find φ_{surf}, since $\varphi_{surf} = \alpha^2$.

Now that φ_{surf} has been obtained, we can find Q_B (from Eq. 3-29) and $\varphi(x)$ (from Eq. 3-23). Below inversion we do not need to find n (x) and p (x), since for this case we have assumed them to be zero for x<d.

Beyond inversion we can calculate d_{max} from Eq. 3-28, and Q_{Bmax} from Eq. 3-29, (using $d = d_{max}$). Then, since $\varphi_{ox} = V_{GB} - 2\varphi_B$, we can find Q_C from

$$Q_C = -(V_{GB} - 2\varphi_B)\,C_{ox} \qquad (3 - 33)$$

Now, since $Q_I = Q_C - Q_{Bmax}$

$$Q_I = -[(V_{GB} - 2\varphi_B)\,C_{ox}] - q\,N_A\,d_{max} \qquad (3 - 34)$$

$$Q_I = -[(V_{GB} - 2\varphi_B)\, C_{ox}] + \sqrt{4\kappa_{si}\varepsilon_o\varphi_B q N_A} \qquad (3 - 35)$$

The results of this analysis are summarized in Fig. 3-7. Here the energy band diagrams of an ideal MOS capacitor (*p*-type Si substrate) are shown for the case of $V_{GB} = 0V$ (Fig. 3-7a). In Fig. 3-7b the case when sufficient gate voltage is applied to strongly invert the silicon surface is shown. Also depicted in these figures are the charge (Fig. 3-7c), $\mathcal{E}$-field (Fig. 3-7d) and electrostatic potential versus distance diagrams (Fig. 3-7e) for these two conditions. We note that since there are no charges in the oxide, the electric field is constant in the oxide, and that the slope of φ is thus constant in the oxide. In Fig. 3-7f we show the electrostatic potential versus distance diagram for an ideal MOS-C with an *n*-type substrate biased into strong inversion. It is useful to note that since a positive bias is needed to invert the *p*-type MOS-C, the φ versus x diagram curves upward, with the maximum value of φ at the surface. However, for the *n*-type MOS-C a negative gate bias is needed. Thus, in such cases the φ versus x diagram curves downward, with its minimum at the surface.

Another important point to which we will refer when discussing the so-called *gradual-channel approximation* (GCA) involves the gradient of the vertical electric field in the silicon, i.e., $\partial\mathcal{E}/\partial x$. In strong inversion $\partial\mathcal{E}/\partial x$ in the silicon is greatest in the inversion layer thickness, since $Q_I >> Q_B$ in strong inversion and the inversion layer is quite thin (~30-300Å thick). Hence, $\partial\mathcal{E}/\partial x$ can be estimated by approximating the vertical electric field value to be zero at the bottom edge of the inversion layer thickness t_{inv}. Then, $\partial\mathcal{E}/\partial x \sim [\mathcal{E}_{si}(x = 0) - 0]/t_{inv} = \mathcal{E}_{si}(x = 0)/t_{inv}$. We will make use of this result when examining the GCA in chap. 4.

3.3.4 More Accurate and General Model for MOS-C Behavior

The assumption that no mobile charges exist in the depletion region of the MOS-C (depletion approximation) does not accurately reflect the actual charge distribution condition in the silicon, especially below inversion. Thus, the expressions for $\varphi(x)$, φ_{surf}, $\mathcal{E}(x)$, and Q_I, obtained using this assumption are not accurate. In order to obtain a more accurate quantitative model of the MOS-C, we need to use the Poisson equation in which *p* and *n* have non-zero values. This will allow us to obtain more accurate expressions of the above quantities. As noted in the introduction of this section, this analysis is complicated. The complexity arises in part because it provides a model that is valid in all regimes (accumulation, depletion, and all ranges of inversion). Although the accumulation and depletion regions are not very important regimes of operation in most MOSFET applications, we nevertheless present this complex MOS-C model here for the several reasons listed in the introduction.

In this more accurate and general representation of the MOS-C, we treat the Poisson equation in the following form:

$$\frac{d^2\phi}{dx^2} = -\left(\frac{q}{\kappa_{Si}\varepsilon_o}\right)\left[-N_A + N_D + p - n\right] \qquad (3 - 36)$$

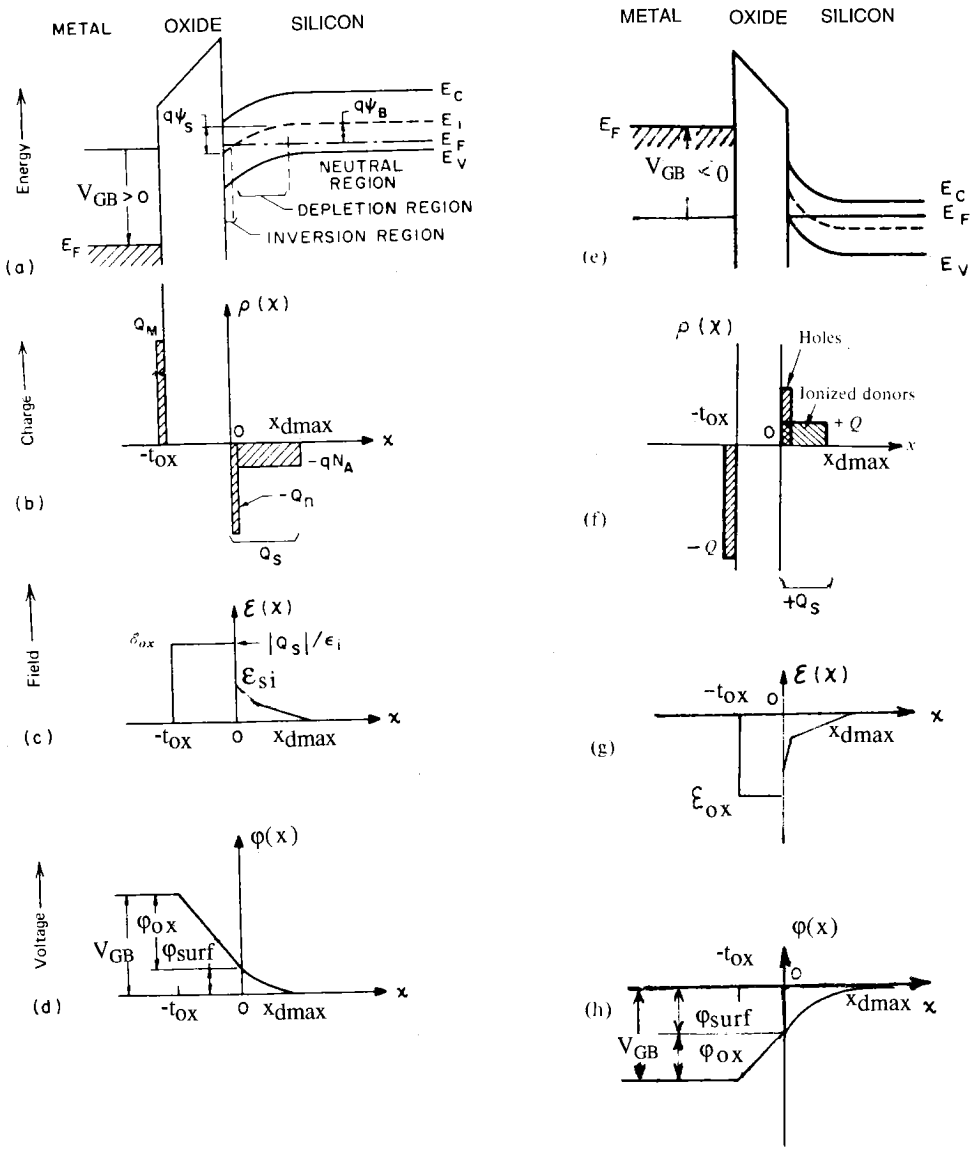


Fig. 3-7 Ideal *p*-type Si MOS-Cwhen biased into inversion [$V_{GB}>V_T$]: (a) Energy-band diagram; (b) Charge distribution; (c) Electric field; and (d) Potential distribution; (e) - (h) Comparable diagrams for an ideal *n*-type Si MOS-C.

Equations 3-12a and 3-12b for $n(\varphi)$ and $p(\varphi)$ are substituted for n and p to give

$$\frac{d^2\phi}{dx^2} = -\left(\frac{q}{\kappa_{Si}\varepsilon_o}\right)\left[-N_A + p_{po}\, e^{-q\varphi/kT} + N_D - n_{po}\, e^{q\varphi/kT}\right] \qquad (3 - 37)$$

In general, Eq. 3-37 can be integrated once to give an analytic expression for the electric field, $\mathcal{E} = -\, d\varphi/dx$. This is accomplished by first multiplying both sides of Eq. 3-37 by $2(d\varphi/dx)$. The left-hand side is then recognized as $(d/dx)(d\varphi/dx)^2$, and we can write the result as

$$\frac{1}{2}\frac{d}{dx}\left(\frac{d\phi}{dx}\right)^2 = -\left(\frac{q}{\kappa_{Si}\varepsilon_o}\right)\left[-N_A + p_{po}\, e^{-q\varphi/kT} + N_D - n_{po}\, e^{q\varphi/kT}\right]\left(\frac{d\varphi}{dx}\right) \qquad (3 - 38)$$

Integrating this expression from the bulk ($x = \infty$, where $\varphi = 0$) to an arbitrary point x, we get

$$\left(\frac{d\phi}{dx}\right)^2_{\infty} - \left(\frac{d\phi}{dx}\right)^2_{x} = -\left(\frac{2q}{\kappa_{Si}\varepsilon_o}\right)\int_0^{\varphi}\left[-N_A + p_{po}\, e^{-q\varphi/kT} + N_D - n_{po}\, e^{q\varphi/kT}\right]d\varphi \qquad (3 - 39)$$

If we note that $p_{po} \approx N_A$, then $n_{po} = n_i^2/N_A$. In addition, since $\varphi_t = q/kT$, and $\mathcal{E} = -\, d\varphi/dx$, and $\mathcal{E}_{si}(x = \infty) = 0$, the first term on the left hand side of Eq. 3-39 is zero. Thus, after integrating the right-hand side we get

$$-\frac{d\varphi}{dx} = \mathcal{E}_{si}(x) = \sqrt{\frac{2q\,N_A}{\kappa_{Si}\varepsilon_o}}\sqrt{\varphi_t e^{-\varphi/\varphi_t} + \varphi - \varphi_t + e^{-2\varphi_B/\varphi_t}\left(\varphi_t e^{-\varphi/\varphi_t} - \varphi - \varphi_t\right)} \qquad (3 - 40)$$

This result allows us to obtain the electric field at the surface ($x = 0$) if φ_{surf} is known

$$\mathcal{E}_{si}(x = 0) = \sqrt{\frac{2q\,N_A}{\kappa_{Si}\varepsilon_o}}\sqrt{\varphi_t e^{-\varphi_{surf}/\varphi_t} + \varphi_{surf} - \varphi_t + e^{-2\varphi_B/\varphi_t}\left(\varphi_t e^{-\varphi_{surf}/\varphi_t} - \varphi_{surf} - \varphi_t\right)} \qquad (3 - 41)$$

In addition, by applying Gauss' law we can use Eq. 3-41 to find Q_C - i.e., from Eq. 3-17 [$Q_C = -\,\kappa_{Si}\varepsilon_o\mathcal{E}_{si}(x = 0)$] we get:

$$Q_C = \sqrt{2q\,N_A\kappa_{Si}\varepsilon_o}\sqrt{\varphi_t e^{-\varphi_{surf}/\varphi_t} + \varphi_{surf} - \varphi_t + e^{-2\varphi_B/\varphi_t}\left(\varphi_t e^{-\varphi_{surf}/\varphi_t} - \varphi_{surf} - \varphi_t\right)} \qquad (3 - 42)$$

This equation is valid for all regimes (accumulation, depletion, and inversion). Figure 3-8 depicts the variation of total charge concentration Q_C with φ_{surf} as calculated using Eq. 3-42.[6] It provides a quantitative representation of the total charge dependence on φ_{surf} that was qualitatively described in section 3.2.* We see that for negative values of φ_{surf}, Q_C is positive, corresponding to the accumulation regime. For $\varphi_{surf} = 0$, we

* Note that in section 3.3.1 we also examined the variation of the mobile charge concentration at the silicon surface n_{surf} versus φ_{surf} for depletion and inversion ($\varphi_{surf} > 0V$).

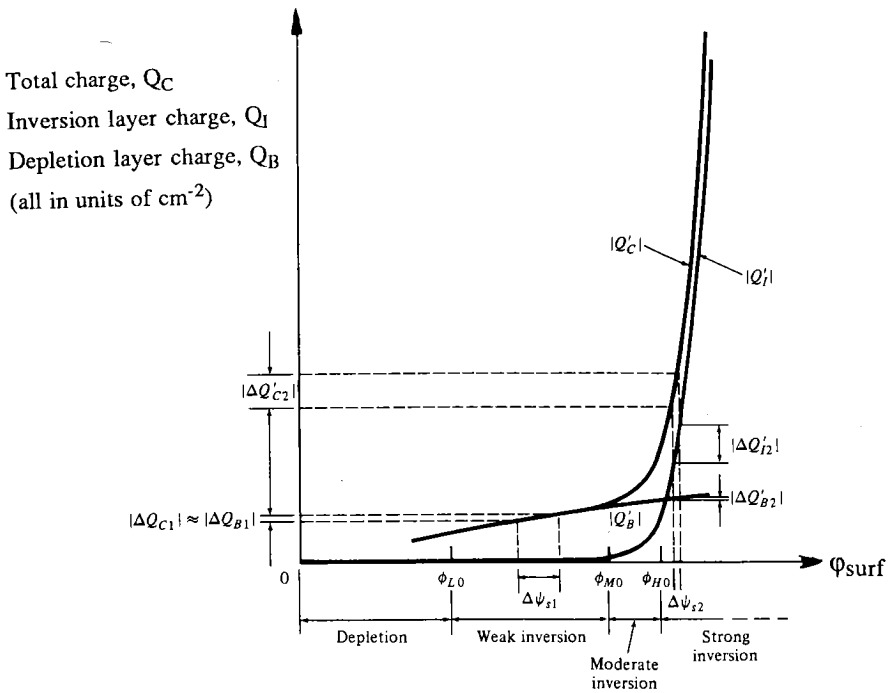


Fig. 3-8 Magnitude of inversion layer charge, depletion layer charge, and their sum (all per unit area) versus φ_{surf}. [6] Reprinted with permission of McGraw-Hill Book Company.

have the flat-band condition, in which $Q_C = 0$. When φ_{surf} is positive, and its value lies between 0 and $2\varphi_B$, Q_C is negative and is $\sim\sqrt{\varphi_{surf}}$. This means that in depletion and weak inversion the mobile charge density is much smaller than the ionic charge density. For $\varphi_{surf}>2\varphi_B$, strong inversion occurs, and $Q_C \cong -\exp(q\varphi_{surf}/2kT$, indicating that the mobile charge density is now much greater than the ionic charge density.

To calculate φ_{surf}, assuming we are given V_{GB}, C_{ox} and N_A, we use

$$V_{GB} = \varphi_{ox} + \varphi_{surf} \tag{3-43}$$

and

$$\varphi_{ox} = \{\kappa_{si}\varepsilon_o \mathcal{E}_{si}(x = 0)\}/C_{ox} \tag{3-44}$$

$$V_{GB} = \varphi_{surf} + \frac{\sqrt{2q N_A \kappa_{si}\varepsilon_o}}{C_{ox}}\sqrt{\varphi_t e^{-\varphi_{surf}/\varphi_t} + \varphi_{surf} - \varphi_t + e^{-2\varphi_B/\varphi_t}\left(\varphi_t e^{-\varphi_{surf}/\varphi_t} - \varphi_{surf} - \varphi_t\right)} \tag{3-45}$$

* (cont.) The conclusions we drew concerning the mobile charge concentration in the silicon in that case for these regions are reiterated here. However, in Fig. 3-8 the contribution of Q_B to the change in Q_C as φ_{surf} is varied is also included.

Note that to find φ_{surf} from this equation we must perform a numerical iteration.

To calculate Q_I accurately we must follow the procedure outlined in section 3.3.2. That is, we must use Eq. 3-21. Here we also change the variable of integration to φ

$$Q_I = -q\int_0^{x_c} n(x)\,dx = -q\int_{\varphi_{surf}}^{\varphi_c} \frac{n(\varphi)\,d\varphi}{d\varphi/dx} \qquad (3-46)$$

By using Eqs. 3-12a and 3-40 in Eq. 3-46 (since $\mathcal{E} = -d\varphi/dx$), we get

$$Q_I = q\int_{\varphi_{surf}}^{\varphi_c} \frac{n_{po}\, e^{\varphi/\varphi_t}}{\sqrt{\frac{2q\,N_A}{\kappa_{Si}\varepsilon_o}}\sqrt{\varphi_t e^{-\varphi/\varphi_t} + \varphi - \varphi_t + e^{-2\varphi_B/\varphi_t}\left(\varphi_t e^{-\varphi/\varphi_t} - \varphi - \varphi_t\right)}}\,d\varphi \qquad (3-47)$$

A similar procedure gives Q_B (consisting of ionized acceptor atoms and holes)

$$Q_B = q\,p_{po}\int_{\varphi_{surf}}^{\varphi_c} \frac{e^{-\varphi/\varphi_t} - 1}{\sqrt{\frac{2q\,N_A}{\kappa_{Si}\varepsilon_o}}\sqrt{\varphi_t e^{-\varphi/\varphi_t} + \varphi - \varphi_t + e^{-2\varphi_B/\varphi_t}\left(\varphi_t e^{-\varphi/\varphi_t} - \varphi - \varphi_t\right)}}\,d\varphi \qquad (3-48)$$

To get $\varphi(x)$ we must integrate the Poisson equation a second time. That is, to find φ at some point x we start with Eq. 3-40, separate variables and integrate from x to 0 on the right hand side, and from φ to φ_{surf} on the left:

$$-\int_{\varphi}^{\varphi_{surf}} \frac{d\varphi}{\sqrt{\frac{2q\,N_A}{\kappa_{Si}\varepsilon_o}}\sqrt{\varphi_t e^{-\varphi/\varphi_t} + \varphi - \varphi_t + e^{-2\varphi_B/\varphi_t}\left(\varphi_t e^{-\varphi/\varphi_t} - \varphi - \varphi_t\right)}} = x \qquad (3-49)$$

However, unlike the first integration which yielded the analytical expression for $\mathcal{E}(x)$ given by Eq. 3-40, this second integration cannot be carried out analytically. That is, we must use numerical procedures to calculate $\varphi(x)$ in this case. This was first done by Young in 1961.[2] Numerical integration must also be used to find the electron and hole densities using Eqs. 3-12 a and 3-12b. Sample plots of $[\rho/qN_A]$ versus x and $\varphi(x)$ versus x calculated using such numerical analysis techniques are shown for accumulation (Fig. 3-9a), depletion (Fig. 3-9b), onset of strong inversion, $\varphi_{surf} = 2\varphi_B$ (Fig. 3-9c), and strong inversion, $\varphi_{surf} = 2\varphi_B + 0.15V$ (Fig. 3-9d) for a MOS-C with $\varphi_B = 0.3V$.[3]

We observe from these curves that: (1) the mobile charge in the inversion layer is confined to a very thin layer near the surface; and (2) the depletion-region edge is fairly abrupt, with the charge concentration dropping to negligible values over a distance of ~0.5 μm. The first observation supports the use of the following two approximations

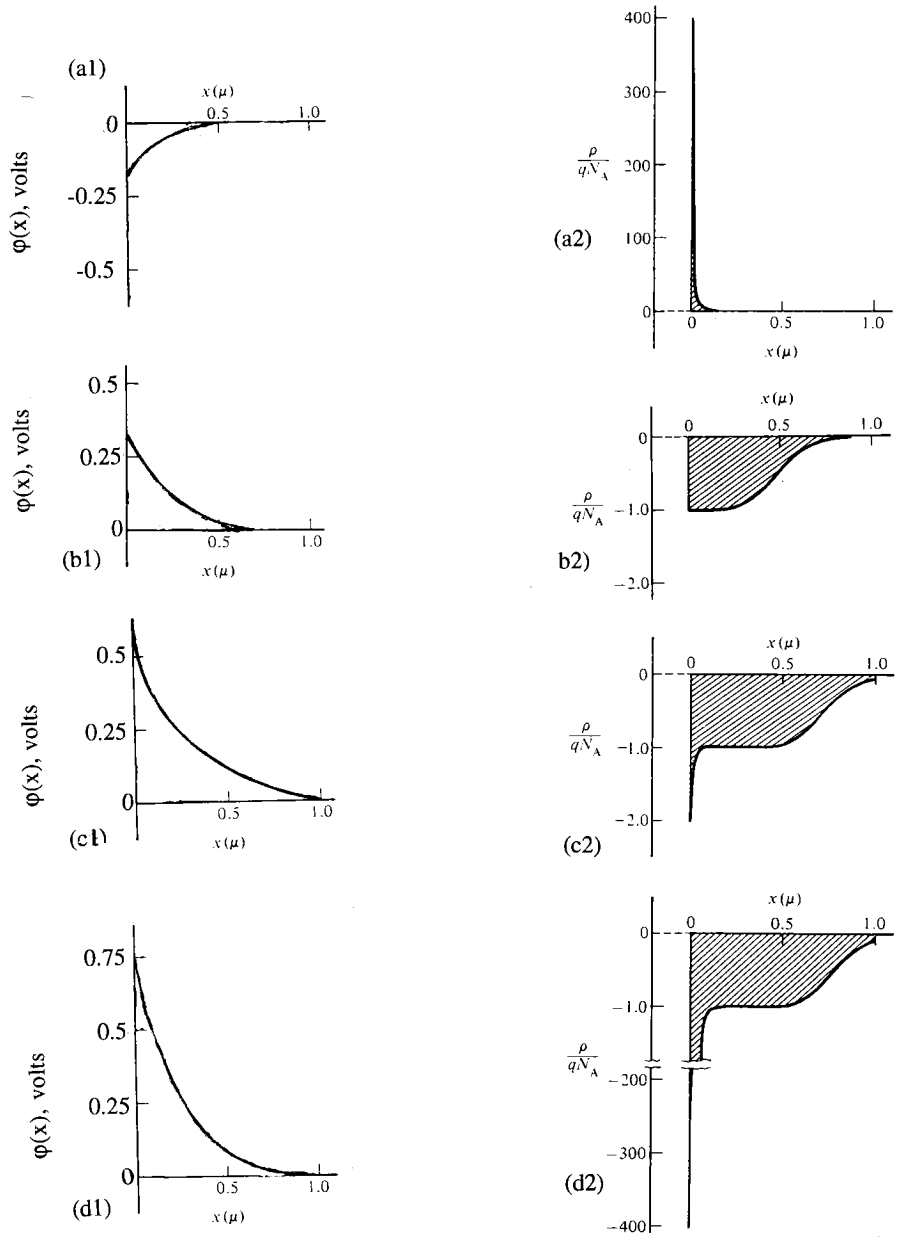


Fig. 3-9 Exact solution for the potential $\varphi(x)$ in volts and normalized charge density (ρ/qN_A), respectively, in the semiconductor region of a MOS-C with a *p*-type Si substrate, as a function of depth, x. Here we assume $\varphi_F = 0.3V$ and T = 300K: (a1 and a2) Accumulation ($\varphi_{surf} = -0.15V$); (b1 and b2) middle of depletion ($\varphi_{surf} = \varphi_F = 0.3V$); (c1 and c2) Onset of inversion ($\varphi_{surf} = 2\varphi_F = 0.6V$); and (d1 and d2) strong inversion ($\varphi_{surf} = 2\varphi_F + 0.15V = 0.75V$). The charge concentration ρ/N_A diagrams are drawn on a linear scale.[3] (Compare Figs. a2 - d2 to the simpler representation of charge in the silicon shown in Figs. 3-4c2 - f2.)

invoked in both the simplified and charge sheet models of the MOS-C: (a) the inversion layer can be treated as an infinitesimally-thin sheet of charge at the interface; and (b) the presence of mobile charge in the depletion region can be ignored (the depletion approximation). The second observation indicates it is reasonable to assume that transition from the bulk to the depletion region occurs abruptly.

In summary, the complicated procedure outlined above involves the following steps to get exact values of MOS-C parameters. For an ideal, uniformly doped *p*-type MOS-C, if we are given N_A, t_{ox}, and V_{GB}

1. Calculate φ_{surf} using Eq. 3-45.

2. With this value of φ_{surf} we can get Q_I from Eq. 3-47, $\mathcal{E}(x)$ from Eq. 3-40, and $\varphi(x)$ from Eq. 3-49.

3. Once $\varphi(x)$ is known we can calculate $n(x)$ and $p(x)$ using Eqs. 3-12a and b.

From the sample plots of the charge distribution we see that in accumulation and inversion the mobile charge resides in an extremely narrow portion of the semiconductor immediately adjacent to the oxide-semiconductor surface. This substantiates the validity of the assumption used in the depletion approximation analysis that the inversion layer charge is confined to a very thin layer near the silicon surface. Thus, a third model based on this approximation (that is intermediate in complexity between the two models just described) has been developed by Baccarani et al.,[4] and Brews.[5] It is called the *charge-sheet model* and will be presented next.

3.3.5 Charge-Sheet Model for Determining the Characteristics of the Ideal MOS-C

The third quantitative model of the MOS-C is the *charge-sheet model.* It is more accurate than the simple depletion approximation model, but does not require the numerical analysis of the general model to obtain Q_I. Since it includes mobile charge carriers in the Poisson equation it provides information about the charge distribution in the MOS-C below strong inversion (unlike the depletion approximation model, which assumes that no mobile charges are present below strong inversion). Since the charge-sheet model assumes that the mobile charge concentrations are zero in the depletion region outside of the thin layer at the surface (that would contain the inversion layer - see assumption 2 below), this model is not as accurate for characterizing the MOS-C in depletion or accumulation as the general model. However, since inversion is responsible for current conduction in MOSFETs, accurate quantitative characteristics of the depletion and accumulation regimes in MOS-Cs and MOSFETs are usually not of much concern. Hence, this does not represent a significant limitation for most applications.

To begin the derivation, we list the assumptions upon which the charge-sheet model is based:

1. Mobile charge can exist in the silicon beyond the onset of *weak* inversion i.e., when $\varphi_{surf}>\varphi_B$ (unlike the previous model that assumed that no mobile charge exists in the channel until strong inversion, i.e., until $\varphi_{surf}>2\varphi_B$);

2. All this mobile charge is confined to a negligibly thin inversion layer at the silicon surface, so that the remainder of the depletion region contains no mobile charges ($n = p = 0$, for $0 < x \leq d$);

3. The depletion region is defined by a sharp boundary at a depth d (i.e., an abrupt transition between the depletion region and the bulk exists), in the same manner as was invoked the depletion approximation model;

4. The voltage drop across the inversion layer in the x-direction is zero, hence V_{GB} (in an ideal MOS-C) is still the sum of φ_{surf} and φ_{ox};

5. The value of φ_{surf} beyond strong inversion is not clamped at exactly $2\varphi_B$. Instead its exact value can be calculated from the model equations;

6. The depletion width at strong inversion is calculated using Eq. 3-28, so that the ionic charge in the depletion region can be calculated using Eq. 3-29;

7. The value of Q_I can be found by subtracting Q_B from Q_C.

Since both holes and electrons are included in the charge density term of the Poisson equation, we start the development of the charge sheet model again with Eq. 3-37. The first integration of the Poisson equation is also carried out in the same way as in the general model analysis, yielding the identical value for $\mathcal{E}_{si}(x = 0)$. Q_C is also expressed by Eq. 3-42. However, since the goal of the charge sheet model is to provide accurate MOS-C expressions only for the case of inversion (including weak inversion), Q_C can be simplified for this operating range. That is, if $\varphi_{surf} \geq \varphi_B$ (where, as we noted in section 3.3.1, φ_B is $9\varphi_t$ to $16\varphi_t$ for the range of doping concentrations used in MOSFETs), several terms in Eq. 3-42 become insignificant and can be dropped, yielding

$$Q_C \approx \sqrt{2q\, N_A \kappa_{Si} \varepsilon_o} \sqrt{\varphi_{surf} + \varphi_t\, e^{(\varphi_{surf} - 2\varphi_B)/\varphi_t}} \qquad (3 - 50)$$

Now, since we assume that in the depletion region there are no mobile charges present (i.e., all the mobile charge is in the inversion layer, which is infinitesimally thin), the depth of the depletion region is again given by Eq. 3-27

$$d = [\kappa_{Si}\varepsilon_o(2\varphi_{surf})/qN_A]^{1/2} \qquad (3 - 51)$$

Then, Q_B will consist only of the uncovered acceptor atoms in the depletion region, and can again be expressed by Eq. 3-29, repeated here:

$$Q_B = q\, N_A\, d = (2\kappa_{Si}\varepsilon_o q N_A \varphi_{surf})^{1/2} \qquad (3 - 52)$$

To get Q_I we can now subtract Q_B from Q_C as given by Eq. 3-50 to get

$$Q_I = \sqrt{2q\, N_A \kappa_{Si} \varepsilon_o}\left(\sqrt{\varphi_{surf} + \varphi_t\, e^{(\varphi_{surf} - 2\varphi_B)/\varphi_t}} - \sqrt{\varphi_{surf}}\right) \qquad (3 - 53)$$

Also note that if we use Eq. 3-50 to approximate Q_C in Eq. 3-19 for φ_{ox}, and insert this into Eq. 3-10, the expression relating φ_{surf} and V_{GB} is now

$$V_{GB} = \varphi_{surf} - Q_C/C_{ox} \tag{3 - 54}$$

or

$$V_{GB} = \varphi_{surf} + \frac{\sqrt{2q\, N_A \kappa_S \varepsilon_o}}{C_{ox}} \sqrt{\varphi_{surf} + \varphi_t\, e^{(\varphi_{surf} - 2\varphi_B)/\varphi_t}} \tag{3 - 55}$$

We note, however, that this equation cannot be solved explicitly for φ_{surf}, but requires a numerical iteration approach to solve for φ_{surf}.

Nevertheless, the charge sheet model provides analytical expressions for Q_I and Q_B in a MOS-C that are quite accurate over the entire range of inversion (i.e., $\varphi_{surf} > \varphi_B$). This makes it more useful in many applications than the more general MOS-C model.

It should also be noted that when the equation for Q_I from the charge sheet model is plotted versus the surface potential (i.e., using a log-linear plot, as shown in Fig. 3-10),[6] two inversion regimes can be identified: (1) weak inversion; and (2) strong inversion.* In each of these regimes we can simplify the expression for Q_I given by Eq. 3-53.

3.3.5.1 Approximation of Q_I in Weak Inversion. First, we recall from section 3.3.4 that Q_I is much smaller than Q_B in weak inversion. Therefore, we can simplify Q_I by treating the term $\varphi_t \exp(\varphi_{surf} - 2\varphi_B)/\varphi_t$ in Eq. 3-53 as small compared to

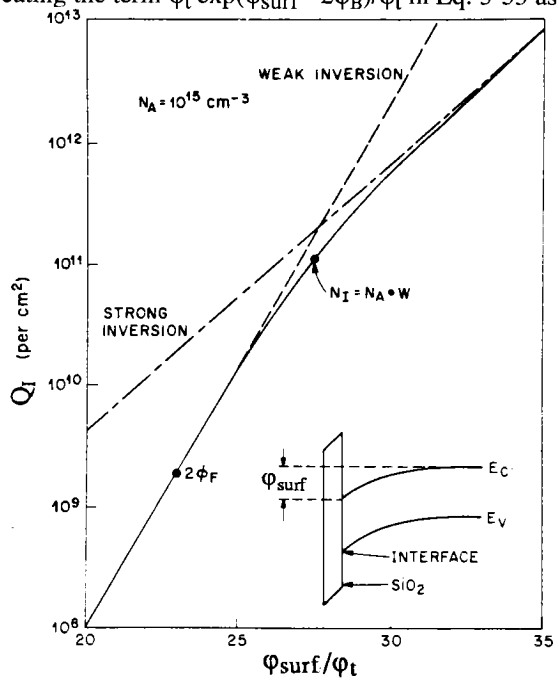


Fig. 3-10 Inversion layer carrier density per unit area versus φ_{surf}/φ_t as computed from Eq. 3- 53 and from approximations to this equation for the case of weak inversion (Eq. 3-58), and strong inversion (Eq. 3-62). [5] (© IEEE 1978).

the term φ_{surf}. Then, if we let

$$\zeta = \varphi_t \exp(\varphi_{surf} - 2\varphi_B)/\varphi_t \tag{3 - 56}$$

the first square root term in the parentheses in Eq. 3-53 becomes $\sqrt{\varphi_{surf} + \zeta}$. Since ζ is smaller than φ_{surf}, we can approximate the function $\sqrt{\varphi_{surf} + \zeta}$ by the first two terms of its Taylor expansion about $\zeta = 0$:

$$\sqrt{\varphi_{surf} + \xi} \approx \sqrt{\varphi_{surf}} + \frac{1}{2\sqrt{\varphi_{surf}}}\xi \tag{3 - 57}$$

Using Eq. 3-56 and Eq. 3-57 in Eq. 3-53 we get

$$Q_I \text{ (weak inv)} \approx \frac{\sqrt{2q N_A \kappa_S \varepsilon_o}}{2\sqrt{\varphi_{surf}}}\varphi_t e^{(\varphi_{surf} - 2\varphi_B)/\varphi_t} \tag{3 - 58}$$

This expression implies that in weak inversion the inversion layer charge is essentially an exponential function of φ_{surf}. Note from Fig. 3-10 that the weak inversion approximation of Q_I from Eq. 3-53 is quite accurate up to $2\varphi_B$, but becomes increasingly inaccurate beyond strong inversion.

We can easily solve for φ_{surf} in weak inversion as well. That is, since $Q_I << Q_B$, we can write Eq. 3-10 as

$$V_{GB} = \varphi_{surf} - Q_B/C_{ox} \tag{3 - 59}$$

or using Eq. 3-52

$$V_{GB} = \varphi_{surf} + [(\sqrt{2\kappa_{si}\varepsilon_o q N_A})/C_{ox}]\sqrt{\varphi_{surf}} \tag{3 - 60}$$

$$= \varphi_{surf} + \gamma\sqrt{\varphi_{surf}} \tag{3 - 60a}$$

where $\gamma = (\sqrt{2\kappa_{si}\varepsilon_o q N_A})/C_{ox}$. Equation 3-60a can be solved for φ_{surf} to yield

$$\varphi_{surf} = \left[-\frac{\gamma}{2} + \left(\frac{\gamma^2}{4} + V_{GB}\right)^{1/2}\right]^2 \tag{3 - 61}$$

To relate Q_I to V_{GB}, we can substitute the value of φ_{surf} from Eq. 3-61 into Eq. 3-58.

3.3.5.2 Approximation of Q_I in Strong Inversion. In strong inversion $Q_I >> Q_B$. As a result, the exponential term in Eq. 3-53 is the only significant term within the parentheses. Thus, Q_I in strong inversion can be well approximated as

$$Q_I \text{ (strong inv)} \approx (\sqrt{2\kappa_{si}\varepsilon_o q}\, n_i \exp(\varphi_{surf}/2\varphi_t) \tag{3 - 62}$$

The inversion-layer charge in strong inversion is thus seen to be an exponential function of the surface potential with a slope of $1/(2\varphi_t)$. Therefore, a small increase in the surface potential induces a large change in Q_I (strong inv). This implies that φ_{surf} will remain almost unchanged even in the face of a large increase in V_{GB} or Q_I beyond strong inversion. The approximation that φ_{surf} (inv) is clamped at $2\varphi_B$ made in the depletion approximation model is thus seen to be reasonably accurate for actual MOSFETs. If the gate voltage corresponding to value needed to produce φ_{surf}(inv) is

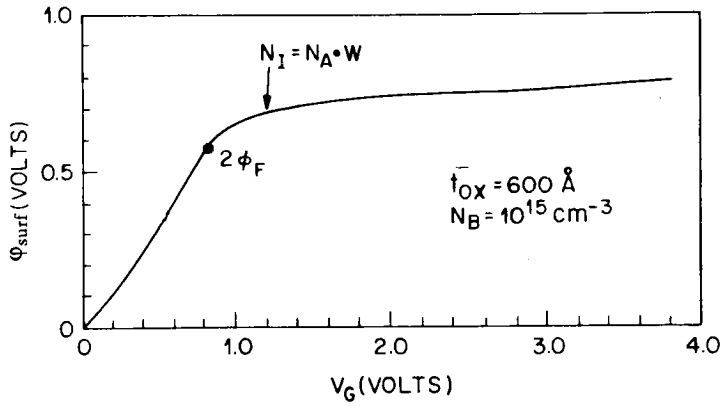


Fig. 3-11 φ_{surf} versus V_{GB} calculated from the charge sheet model for an oxide thickness of 60 nm, and a substrate doping level of $N_A = 10^{15}$ cm^{-3}. The dot indicates $\varphi_{surf} = 2\varphi_B$.[5] (© IEEE 1978).

termed the threshold voltage V_T (see next section for how to compute this value), then the Q_I (strong inv) dependence on V_{GB} can be approximated by

$$Q_I \approx - C_{ox} (V_{GB} - V_T) \qquad (3 - 63)$$

We should mention that φ_{surf}(inv) does change slightly as V_{GB} is increased beyond strong inversion, as shown in Fig. 3-11, where φ_{surf} from Eq. 3-55 is plotted versus V_{GB}. We see that φ_{surf} in strong inversion actually increases beyond $2\varphi_B$ by ~150 mV before leveling off. Thus, models of Q_I that use φ_{surf}(inv) = $2\varphi_B$ underestimate the actual value. This has an impact on the accuracy of predicted drain current values in MOSFET circuit models based on this approximation. We will discuss this further in chap. 4.

3.4 THRESHOLD VOLTAGE of IDEAL MOS CAPACITORS

Strong inversion in MOSFETs occurs in essentially the same manner as in MOS-Cs. In fact, one of the most important parameters of a MOSFET is the gate voltage at the onset of strong inversion. This parameter is known as the *threshold voltage* V_T. We can obtain an expression for V_T using an ideal MOS-C in which we assume the depletion approximation is valid. Then we follow the same analysis as in section 3.3.3.

* Note that a third inversion regime, intermediate to the two mentioned above, namely *moderate inversion*, has been defined by Tsividis (see Fig. 3-6).[6] This allows an even more careful representation of MOS behavior to be developed. Interested readers can consult ref. 6 for more details.

That is, we begin with Eq. 3-10:

$$V_{GB} = \varphi_{ox} + \varphi_{surf} \qquad (3 - 64)$$

At the *onset* of strong inversion (i.e., implying that the inversion layer is just at the verge of forming, so Q_I is still zero), the magnitude of the charge on the gate is balanced by just the charge in depletion region of the semiconductor Q_B due to the ionized dopants, or:

$$Q_G = - Q_B = q\, N_{SUB}\, d_{max} \qquad (3 - 65)$$

where N_{SUB} is the substrate doping (i.e., N_A in *p*-type and N_D in *n*-type substrates), respectively. Then, φ_{surf} at the onset of inversion = $2\varphi_B$. It can also be expressed in terms of d_{max} using Eq. 3-27. Setting these two equal, we have for a *p*-type substrate

$$\varphi_{surf}(inv) = 2\varphi_B = qN_A\,(d_{max})^2/2\kappa_{si}\varepsilon_o) \qquad (3 - 66)$$

Solving this for d_{max}, and substituting into Eq. 3-65 we get Q_B in terms of N_A and φ_B. Finally, using this expression for Q_B in Eqs. 3-19 and 3-20 ($\varphi_{ox} = -Q_B/C_{ox}$), and putting this into Eq. 3-64 we obtain the gate voltage value at the onset of strong inversion in an ideal MOS capacitor (V_T) expressed as a single equation in terms of t_{ox}, T, n_i, and N_A:

$$V_T \text{ (of an ideal MOS capacitor)} = 2\sqrt{\kappa_{si}\varepsilon_o q N_A \varphi_B}/C_{ox} + 2\varphi_B \qquad (3 - 67)$$

3.5 PHYSICS OF NON-IDEAL MOS CAPACITORS (n^+-PolySi-SiO_2-Si MOS Capacitors)

Although the ideal MOS capacitor is a useful structure for introducing MOS behavior, from the perspective of this text the most important MOS capacitor structure is the metal (Al, or heavily-doped polysilicon)-SiO_2-Si structure. In the metal-SiO_2-Si structure (and other "real" MOS capacitors), the work function difference φ_{ms} is virtually never zero, and various charges are also always present inside the oxide or at the Si-SiO_2 interface. These phenomena exhibited by "real" MOS capacitors significantly alter ideal MOS behavior. They are also inherent in "real" MOSFETs, and must therefore be included to accurately describe actual device behavior.

3.5.1 Workfunction Difference Between Gate and Semiconductor

We noted earlier that the energy-band diagram of an *ideal* MOS capacitor is in a flat band condition when $V_{GB} = 0$. In real MOS capacitors, however, a flat band condition with $V_{GB} = 0$ is essentially never observed. Instead, when $V_{GB} = 0$ the bands are almost always bent as a result of the work function difference between the gate and semiconductor regions, as well as the charges that exist in the MOS capacitor.

The *work function* of a material is defined as the minimum energy required to bring an electron from the Fermi level to the vacuum level. In MOS structures, however, it

is more useful to deal with the *modified* work function. This is defined as the minimum energy required to bring an electron from the Fermi level in the gate E_{Fm} or the semiconductor E_{FS} to the oxide conduction-band edge ($q\varphi_m'$ or $q\varphi_s'$, respectively), as shown in Fig. 3-12. Note also that $q\varphi_s' = q[\chi' + (E_c - E_F)_\infty] = q(\chi' + E_G/2 + \varphi_B)$ for a *p*-type semiconductor, where χ' is the *modified electron affinity* (defined as the energy difference between the conduction band edges of the semiconductor and the oxide). In the Si-SiO_2 case, $q\chi' = 3.23$ eV.

Figure 3-12a shows the energy-band diagram for an Al-SiO_2-*p*-type-Si structure before contact. For Al-SiO_2, $q\varphi_m' = 3.2$ eV, while for *p*-type Si-SiO_2 $q\varphi_s' = (3.23 + 0.55 + q\varphi_B) \sim 4.1$ eV. Obviously, $q\varphi_m'$ and $q\varphi_s'$ are not equal. Hence, if a MOS structure could be formed by bringing the 2 regions in contact, the effect of this difference on the energy-band diagram of the MOS capacitor would be as shown in Fig. 3-12b. That is, if such a MOS capacitor was unbiased ($V_{GB} = 0$), the Fermi levels would have to "line up" (be at the same energy level) everywhere in the capacitor, Fig.

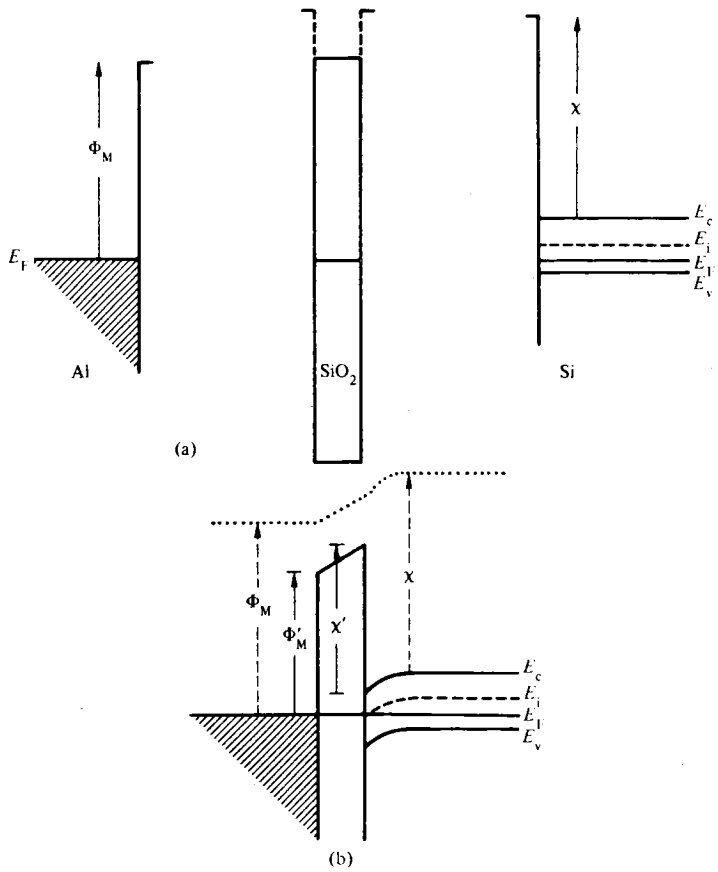


Fig. 3-12 The MOS capacitor structure and idealized energy-band diagram (a) before contact and (b) after contact.[12] Reprinted with permission of Addison-Wesley.

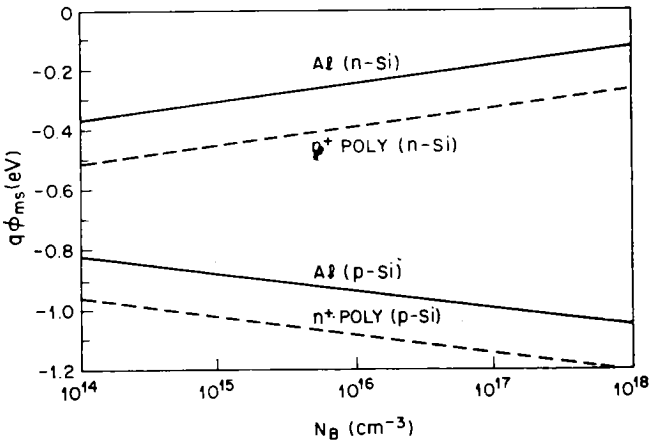


Fig. 3-13 Work-function difference ϕ_{ms} (V) versus doping for n^+ and p^+ polysilicon, and Al gate electrodes.[7] Reprinted with permission of Solid-State Electronics.

3-12). Since the modified work functions in this structure are different, the E_c, E_i, and E_v bands would be bent downward when $V_{GB} = 0$ to satisfy the requirement of a constant Fermi level. To achieve a flat-band condition in this structure, a negative voltage [equal to $(\varphi_m - \varphi_s)$], would have to be applied to the gate (in this example $\cong$ - 0.9 V).

In submicron ICs, however, the gate material is normally heavily-doped polysilicon. The Fermi level in heavily doped n-type polysilicon resides at the conduction band edge (and in heavily doped p-type poly, at the valence band edge). Hence, for an n-channel device with n^+ poly on a p-type substrate $\varphi_{ms}(n\text{-ch}) = (\varphi_m - \varphi_s) = (-E_G/2q) - \varphi_B$ (with its value between -0.9 to -0.95V for $10^{16}\,cm^{-3}<N_A<10^{17}\,cm^{-3}$ - the appropriate doping range for submicron technologies if $L>0.5\ \mu m$). Likewise, the p-channel φ_{ms} for an n^+ poly on an n-type substrate is $\varphi_{ms}(p\text{-ch}) = (\varphi_m - \varphi_s) = (-E_G/2q) + |\varphi_B|$, with its value between -0.15 to -0.2V for the same doping range. Finally, for a p-channel device having a p^+ poly on an n-substrate, $\varphi_{ms}(p\text{-ch}) = (\varphi_m - \varphi_s) = (E_G/2q) + |\varphi_B|$. Figure 3-13 shows the value of φ_{ms} for various substrate-doping values if an aluminum, n^+-poly, or p^+-poly gate electrode is used.

3.5.2 Oxide Charges and Traps

Four different types of charges are associated with the oxide and Si-SiO_2 interface as shown schematically in Fig. 3-14. The effective net (i.e., total) charge per unit area at the Si-SiO_2 interface (in C/cm^2) is denoted by the symbol Q_{tot}, whereas the net (total) number of charges per unit area (Q_{tot}/q) is given by N_{tot}. The standardized terminology and symbols used to differentiate the various individual charge types are as follows:

1. Mobile ionic charge per unit area Q_m (C/cm^2), and N_m (no. of mobile charges/cm^2);

2. Fixed oxide charge per unit area Q_f (C/cm^2), and N_f (no. of fixed oxide charges/cm^2);

3. Interface trap charge per unit area Q_{it} (C/cm^2), and N_{it} (no. of interface trapped charges/cm^2); D_{it} - number of interface trapped charges per unit area *and* energy (no. of interface trapped charges/cm^2-eV);

4. Oxide trapped charge Q_{ot} (C/cm^2), and N_{ot} (no. of oxide trapped charges/cm^2).

Since issues relating to the reliability of submicron MOSFETs also involve these charges, a brief review concerning the origin of the charges, their effect on device behavior, and techniques to minimize their presence is given here. A more detailed description of their characteristics can be found in Vol. 1 and refs. 1 and 7.

3.5.2.1 Mobile Ionic Charges. Mobile ionic charges Q_m in the oxide layer arise mostly from the presence of sodium or potassium ions, which have very high diffusivities in the oxide (even at temperatures of less than 200°C). Their presence leads to threshold instabilities and to a deterioration of the oxide reliability. After it was confirmed that sodium in even very small concentrations is ruinous to MOSFET operation, stringent procedures were developed to keep the level below $N_m = 10^9$ cm^{-2}. At such levels, the instabilities in V_T are negligibly small. Procedures to counteract Na contamination include: using clean processing techniques; employing gettering techniques; neutralization of the sodium in the oxide by growing the oxide in ambients containing chlorine compounds; and using SiN (or oxynitride) passivation films. (Such films are impervious to the transport of Na.) Tests for Na contamination are also routinely performed (using MOS-C monitor structures) to detect any accidental contamination. Steps can then be taken to find and eliminate the contamination source. More information on the details of these techniques can be found in Vol. 1, Chap. 7.

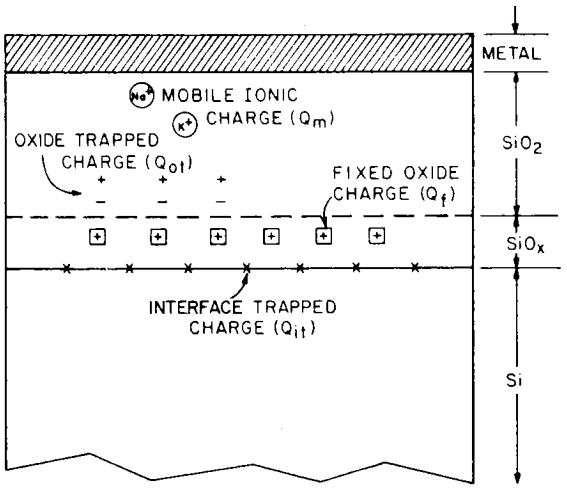


Fig. 3-14 Standardized terminology for oxide charges associated with thermally oxidized silicon.[13] (© IEEE 1980).

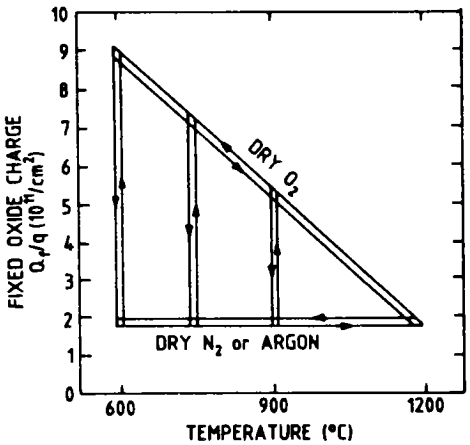


Fig. 3-15 The *Deal triangle,* which specifies the expected Q_f/q after dry O_2 oxidation and after inert ambient annealing.[9] Reprinted by permission of the publisher, The Electrochemical Society, Inc.

3.5.2.2 Fixed Oxide Charge. The fixed oxide charge Q_f (usually positive) is located in the oxide within approximately 30Å of the Si-SiO_2 interface. Since its density does not vary with surface potential (unlike interface trapped charge Q_{it}), it is termed *fixed* oxide charge. It is generally accepted that Q_f is associated with the structure of the interfacial transition region between the silicon and SiO_2. One hypothesis is that Q_f is due to *excess ionic silicon* which has broken away from the silicon proper and is waiting to react in the vicinity of the Si-SiO_2 interface when the oxidation process is abruptly terminated.

Regardless of the origin, it has been experimentally confirmed that the magnitude of Q_f is influenced by substrate orientation, oxidation temperature, and anneal conditions after oxide growth. These factors were carefully studied by Deal et al., and their results are summarized by the well-known "Deal-triangle" shown in Fig. 3-15. In this figure N_f (Q_f/q) is plotted for (111) Si against final oxidation or anneal temperature. It can be seen that the minimum N_f value for (111) material is ~$2x10^{11}$ cm^{-2}. The time required to reach a minimum value of N_f is less than 5 min at 1200°C and about an hour at 550°C. In addition, after an oxide growth step, wafers are annealed in an inert ambient (part of the annealing step occurring during the slow withdrawal from the furnace). It should be noted, however, that continued annealing in inert ambients (e.g., N_2) at high temperatures leads to a marked increase of N_f. Fig. 3-16 shows how N_f changes as a function of time during a nitrogen anneal at 950, 1100 or 1200°C.

The value of N_f also depends on the orientation of the Si surface, and is lowest on (100) orientations (in which the minimum N_f value is ~$1.1x10^{10}$ cm^{-2}). As the lowest values of N_f are desired, (100) material is used to build MOS ICs.

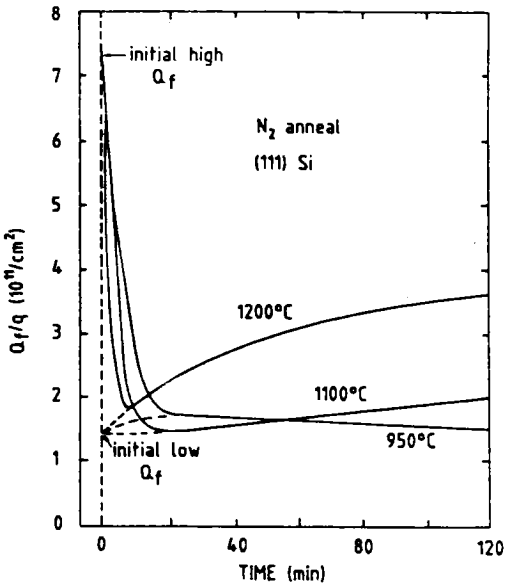


Fig. 3-16 Variation of fixed oxide charge as a function of time during nitrogen anneal at 950, 1100 or 1200°C.[10] Reprinted by permission of the publisher, The Electrochemical Society, Inc.

Since Q_f is located at the Si-SiO_2 interface, the change in V_T that it causes (ΔV_T) is simply calculated as the shift in V_T that would be produced by a sheet of positive charge placed at the interface, or

$$\Delta V_T = -Q_f/C_{ox} = q\, N_f/C_{ox} \qquad (3\text{-}68)$$

3.5.2.3 Interface Trapped Charge. Interface trapped charge Q_{it} arises from allowed energy states (also referred to as *surface states*, *interface states*, or *interface traps*) that somehow exist in the forbidden gap of the Si in the region very close to the Si-SiO_2 interface. Although models exist that detail the *electrical* behavior of interface traps, their *physical origin* has not been totally clarified. The weight of experimental evidence supports the view that the interface traps arise primarily from unsatisfied chemical bonds or so-called "dangling bonds" at the surface of the semiconductor. When a silicon lattice is abruptly terminated along a given plane to form a surface, one of the four surface-atom bonds will be left dangling as pictured in Fig. 3-17. It seems plausible that most, but maybe not all, of such Si-surface bonds would be tied up as a result of thermally growing an SiO_2 layer (Fig. 3-17b). The remaining dangling bonds might become interfacial traps. Other possible interfacial trap-formation effects include Si-Si bond stretching - due to the absence of an oxygen atom (Fig. 3-17c), Si-O bond stretching (Fig. 3-17d), and the presence of a metallic impurity at the Si surface (Fig. 3-17e).[11] These interface traps are also considered further in chaps. 7 and 8.

Based on the wide-ranging and degrading effects they cause in the operational behavior of MOSFETs (and the fact that Q_m and Q_f can generally be acceptably

minimized), interface trapped charges must be considered to be the most perfidious of the oxide charges. Because they are not usually caused by the presence of foreign impurities, but instead are due to a disruption of the periodicity of the crystal lattice, they are not associated with well-defined energy levels. Instead, interface trapped charges are distributed throughout the entire band gap at closely spaced levels. (Interface levels

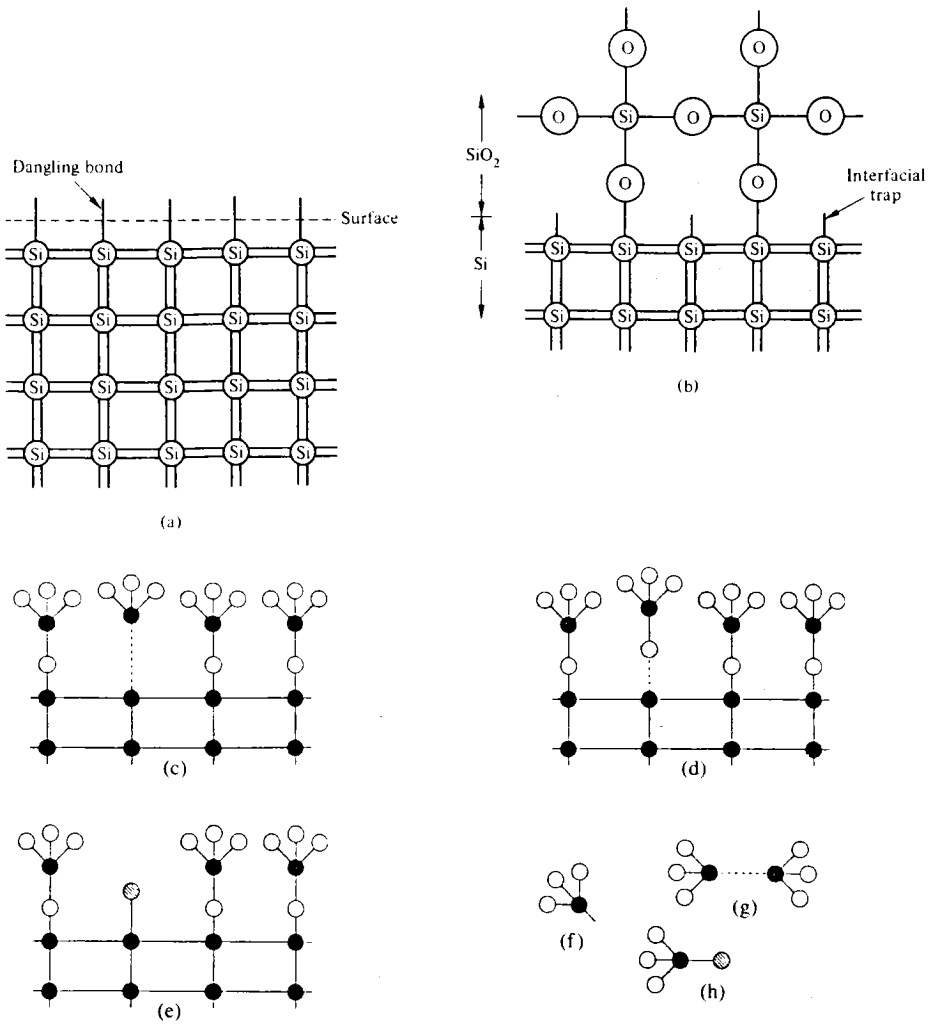


Fig. 3-17 Physical model for the interfacial traps. (a) "Dangling bonds" which occur when the Si lattice is abruptly terminated along a given plane to form a surface. (b) Post oxidation dangling bonds (relative number greatly exaggerated) that become the interfacial traps. (c) Si-Si bond stretching. (d) Si-O bond stretching. (e) presence of a metallic impurity at the Si surface. (f) - (h) possible defects in the SiO_2 bulk that may give rise to oxide trapped charge.

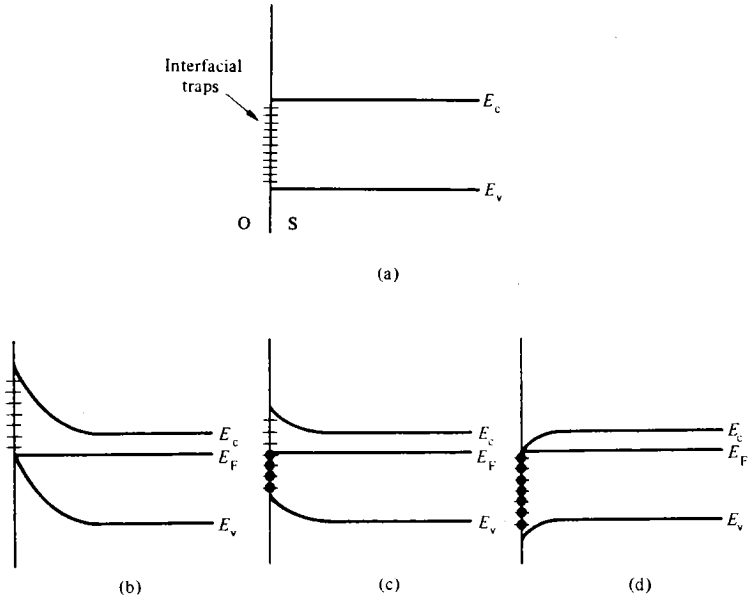


Fig. 3-18 (a) Electrical modeling of interfacial traps as allowed electronic levels localized in space at the oxide-semiconductor interface. (b) - (d) Filling of the interface levels under (b) inversion, (c) depletion, and (d) accumulation biasing in an *n*-type device. From R.F. Pierret, *Modular Series on Solid-State Devices*; *Vol. IV - Field Effect Devices.* Copyright, 1983, Addison-Wesley. Reprinted with permission.

can also occur at energies greater than E_C or less than E_V, but such levels are usually obscured by the much larger density of conduction- or valence-band states.) By using an energy-band diagram to show how these energy levels are distributed throughout the forbidden gap (Fig. 3-18a), we can more easily visualize their impact on a MOS-C.

That is, if an *n*-doped Si MOSFET (PMOS) is biased into inversion (Fig. 3-18b), the Fermi level at the surface will be close to E_V. In this case, the interfacial traps will mostly be empty, as they lie in energies above E_F. Furthermore, if these empty trap sites are assumed to be donorlike (that is, positively charged when empty and neutral when filled with an electron), the net charge per unit area associated with the interfacial traps (Q_{it}) will be positive. Changing the gate bias to achieve depletion conditions (Fig. 3-18c) positions E_F near the middle of the band gap at the surface. Since the interface levels always remain fixed in energy relative to E_C and E_V at the surface, depletion biasing obviously draws electrons into the lower interface state levels, and Q_{it} reflects the added negative charge. If the MOS capacitor is biased into the accumulation mode, most of the traps will be filled with electrons (Fig. 3-18d), and Q_{it} will approach its minimum value. The purpose of the above discussion is to show that the interface traps charge and discharge as a function of bias, thereby affecting the charge distribution inside the device, as well as the device characteristics in a complex but somewhat predictable manner.

Once again, however, since Q_{it} is located at the Si-SiO_2 interface, the change in V_T due to Q_{it} can be simply calculated from

$$\Delta V_T = -Q_{it}(\varphi_{surf}) / C_{ox} \tag{3-69}$$

Since Eq. 3-69 shows that Q_{it} depends upon the bias condition (and is therefore a function of φ_{surf}), its value will change as V_{GB} changes. Under conditions of inversion Q_{it} takes on its largest positive value, and decreases as the bias progresses from depletion to accumulation. The distribution of trap levels as a function of energy in the band gap has a U-shaped distribution (see Fig. 3-19). The minimum level of D_{it} at midgap is usually the density used to characterize their presence.

The overall value of Q_{it}, as well as the precise density of states as a function of energy, makes the value of D_{it} sensitive to a variety of process conditions used to fabricate the MOSFET. Nevertheless, certain definite trends concerning these values have been uncovered. First, Q_{it}, like Q_f, is greatest on (111) surfaces, and smallest on (100) surfaces, and the ratio of midgap-D_{it} values on the two surfaces is ~3:1. Second,

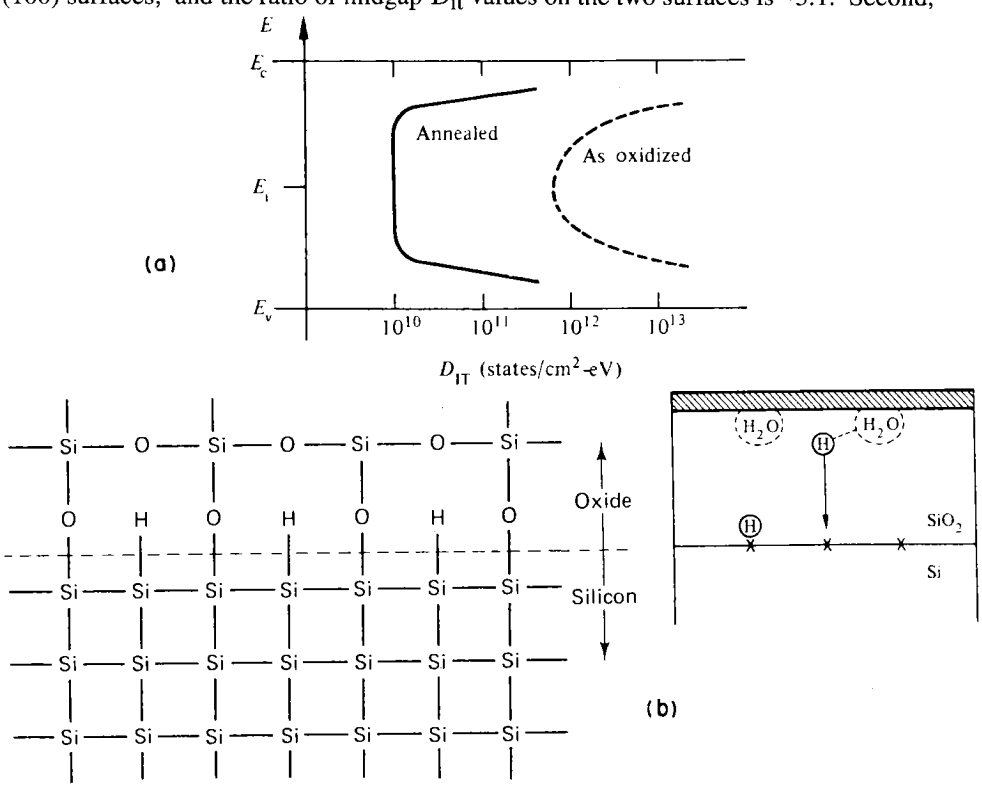


Fig. 3-19 (a) Energy distribution of interface states within the band gap. General form and magnitude of the interfacial trap density observed before and after interface state anneal. (b) Conceptual picture of how hydrogen is thought to passivate dangling bonds.

after oxidation in a dry O_2 ambient, the midgap value of D_{it} is relatively high ($\sim 10^{11}$-10^{12} states/cm^2-eV), with the values decreasing as the oxide temperature is increased (this is also the case with Q_f). However, annealing at lower temperatures in inert ambients *does not* reduce the value of D_{it}. Instead, annealing in the presence of hydrogen at relatively low temperatures (≤500°C) for 5-10 minutes minimizes D_{it}. D_{it} at midgap under an optimum set of processing conditions is $\sim 10^{10}$/cm^2-eV. See Fig. 3-19b for a conceptual picture of how hydrogen is thought to passivate the dangling bonds at the silicon surface and how this thus reduces the interface trap density.

3.5.2.4 Oxide Trapped Charge. The oxide trapped charge Q_{ot} is due to holes or electrons trapped in the *bulk* of the oxide, and hence can be positive or negative. Q_{ot} is associated with defects in the SiO_2, which may arise from ionizing radiation, hot carrier injection, or high currents through the oxide (e.g., due to Fowler-Nordheim tunneling - as described in chap. 7), or other reactions which either tend to break Si-O bonds in the oxide or otherwise lead to carrier trapping on sites or traps somehow already present in the oxide. Figure 3-17 f - h shows possible defects in the SiO_2 bulk that may give rise to such traps. For example, a trivalent Si atom in the SiO_2 will introduce a deep trap level in the SiO_2 bandgap ($\bullet Si{\equiv}O_3$), and stretched Si-O or Si-Si bonds (in the latter a missing bridging oxygen atom leads to a defect, ($O_3{\equiv}Si\bullet\bullet S{\equiv}O_3$) introduce a continuum of shallow or deep trap levels in the SiO_2 bandgap when located in the bulk of the SiO_2.

Q_{ot} can be annealed out by low-temperature treatments (450-550°C) although some neutral traps may remain. Therefore, oxide trapped charge has been considered relatively less important than the other oxide charges in just-completed MOSFETs. In early generations of MOSFET ICs, exposure of the devices to ionizing radiation encountered in space flights was the main concern involving Q_{ot}. However, more recently-introduced processing techniques such as ion-implantation, electron-beam evaporation of metal films, reactive-ion etching, and x-ray lithography are all potential sources for creating charged and neutral traps in the gate dielectric (all of which may not be completely annealed out at the end of the fabrication process). This may result in long-term reliability problems, especially if hot carriers are injected in these oxides and get trapped (see chap. 9). Another reliability concern involves gate oxide degradation due to high oxide electric fields (and subsequent Fowler-Nordheim tunneling currents in the oxide). This has become more of a concern as oxide thickness has been scaled down.

3.5.2.5 Combined Effect of Oxide Charges on V_T of Submicron MOSFETs. It is now possible to control the level of Q_m in modern IC fabrication processes so that it has an insignificant effect on V_T stability. Thus, Q_f and Q_{it} are the remaining important charges that must be controlled during fabrication. For this discussion, let the sum of Q_f and $Q_{it} = Q_{tot}$. Then, combining Eqs. 3-68 and 3-69,

$$\Delta V_T = - Q_{tot}/C_{ox} = - (Q_f + Q_{it})/C_{ox}. \qquad (3 - 70)$$

If $N_{tot} = Q_{tot}/q$ is maintained at $<5\times10^{10}$/cm^2, for oxides <200Å thick:

$$\Delta V_T = Q_{tot}/C_{ox} = qN_{tot}\, t_{ox}/3.9\, \varepsilon_o$$

$$= (1.6\text{x}10^{-19}\ \text{C})(5\text{x}10^{10}\ \text{cm}^{-2})(2\text{x}10^{-6}\ \text{cm})\ /(3.5\text{x}10^{-13}\ \text{F/cm})$$

$$= 0.05\ \text{V}$$

Thus, for routine submicron process technology, it may seem that these charges are no longer a major problem. However, this is not true as far as the problems of gate oxide reliability and MOSFET degradation due to hot-carrier injection effects are concerned. That is, due the high electric fields present in submicron MOSFETs, phenomena occur that give rise to the generation of additional Q_{ot} and Q_{it} during normal device operation. Thus, the value of these charges can rise during the lifetime of the device, and cause performance degradation or failure. This topic will be revisited in more detail in chaps. 7 and 9.

3.6 THE THRESHOLD VOLTAGE EQUATION OF Si-SiO$_2$ MOS CAPACITORS

To obtain a flat-band condition in a real (non-ideal) MOS capacitor, a voltage must be applied to the gate to overcome the effects of $\varphi_{ms} \neq 0$ and $Q_{tot} \neq 0$. The total voltage needed to offset these effects is referred to as the *flat-band voltage* V_{FB} and is given as

$$V_{FB} = \varphi_{ms} - Q_{tot}/C_{ox} \tag{3 - 71}$$

The gate voltage that must then be applied to a real MOS capacitor structure to achieve the onset of strong inversion (i.e., threshold voltage, V_T) must include the flat band voltage as well. Thus, the V_T equation for real MOS capacitors (and long-channel MOSFETs when the source-to-substrate voltage $V_{SB} = 0$V) is given as

$$V_{T\ (\text{uniform doping, long-channel})} = V_{T(ud,lc)} = V_{FB} + 2\sqrt{\kappa_{si}\varepsilon_o q N_{sub}\varphi_B}/C_{ox} + 2\varphi_B$$

$$= \varphi_{ms} - Q_{tot}/C_{ox} + 2\sqrt{\kappa_{si}\varepsilon_o q N_{sub}\varphi_B}/C_{ox} + 2\varphi_B \tag{3 - 72}$$

Figure 3-20 plots Eq. 3-72 for n^+ silicon-gate MOSFETs (both PMOS and NMOS) versus substrate doping concentration for several values of t_{ox} (i.e., 65, 25, and 15 nm thick) as the parameter, and assuming also that $Q_{tot} = 0$.

Note that Eq. 3-72 also assumes that the Si substrate is uniformly doped. In practical MOSFETs, however, V_T is normally adjusted by implanting a thin layer of impurity atoms (boron or arsenic) at the Si surface. Thus, the substrate becomes non-uniformly doped, and the V_T equation must be modified to take this into account. To a first order this is done by treating the implanted impurity atoms as if they are an additional layer of charge very close to the Si-SiO$_2$ interface. Their presence is accounted for by adding a term to the V_T equation similar to the term used to describe

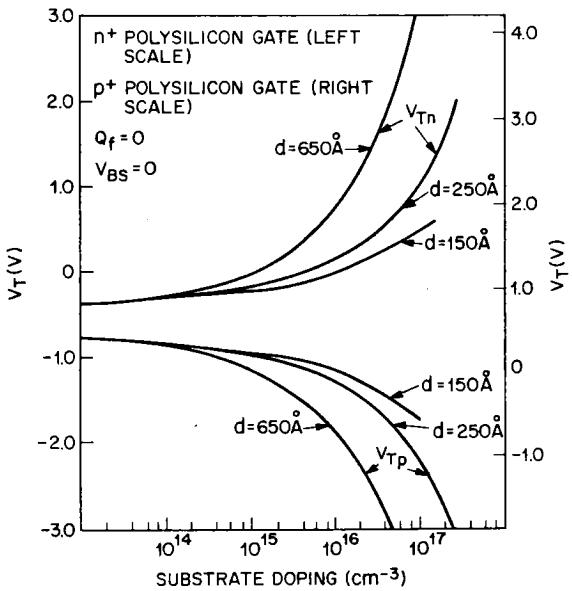


Fig. 3-20 Calculated threshold voltages of n-channel (V_{Tn}) and p-channel (V_{Tp}) transistors as a function of their substrate's doping, assuming n^+ polysilicon gate (left scale) and p^+ polysilicon gate (right scale). Curves for gate oxide thicknesses d of 150 Å, 250 Å, and 650Å are shown. From S.M. Sze Ed. *VLSI Technology,* 2nd Ed., Chap. 11, "VLSI Process Integration." Copyright, 1988 Bell Telephone Laboratories. Reprinted with permission.

the presence of oxide charge Q_{tot}. Details of V_T adjustment and the term used to modify Eq. 3-72 are discussed in chap. 4.

Example 3-1: Find $V_{T(ud,lc)}$ of an ideal MOS-C (φ_{ms} = 0V, Q_{tot} = 0) with a uniform substrate doping $N_A = 10^{15}/cm^2$. The gate-oxide thickness is t_{ox} = 15 nm.

Solution: $\varphi_B\,(sub) = (E_i - E_F)/q = 0.29$ V

$$\varepsilon_{ox} = 3.9\varepsilon_o = 3.5\text{x}10^{-13}\ \text{F/cm};\quad C_{ox} = \varepsilon_{ox}/1.5\text{x}10^{-6} = 2.3\text{x}10^{-7}\ \text{F/cm}^2$$

$$2\sqrt{\kappa_{si}\varepsilon_o q N_A \varphi_B} = 2\,[1.02\ \text{x}10^{-12}\ \text{x}\ 1.6\text{x}10^{-19}\ \text{x}\ 10^{15}\ \text{x}\ 0.29]^{1/2} = 1.4\text{x}10^{-8}\ \text{C/cm}^2$$

$$2\sqrt{\kappa_{si}\varepsilon_o q N_A \varphi_B}\,/C_{ox} = 1.4\ \text{x}\ 10^{-8}/2.3\ \text{x}\ 10^{-7} = 0.06\ \text{V}$$

$$V_{T(ud,lc,\ ideal)} = 0.58\text{V} + 0.06\text{V} = 0.64\ \text{V}$$

Example 3-2: Find $V_{T(ud,lc)}$ of a 'real' MOS-C just like the one of Example 3-1, except that it has a heavily-doped silicon gate (doped to $N_D = 10^{20}/cm^3$), and $Q_{tot} = q\,(5\ \text{x}\ 10^{10}/cm^2)$.

Solution: According to Eq. 3-72, the threshold voltage of the 'real' MOS-C is shifted from that of the 'ideal' MOS-C by the value of V_{FB}. This term is the sum of φ_{ms} + Q_{tot}/C_{ox}. Thus, since

$$\varphi_{ms} \text{ (from Fig. 3-13)} = -1.02 \text{ V}$$

and

$$Q_{tot}/C_{ox} = -1.6\text{x}10^{-19} \text{ x } 2\text{x}10^{10} / 2.3\text{x}10^{-7} = -0.01 \text{ V}$$

$$V_{T(ud,lc)} = V_{T(ud,lc,\ ideal)} + V_{FB}$$

$$= 0.64\text{V} + (-1.03\text{V}) = -0.39 \text{ V}$$

Note that a negative value of $V_{T(ud,lc)}$ is yielded, implying that this MOSFET would be ON at V_{GB} = 0V (and would therefore behave as a depletion-mode device). The parameters used in this calculation of $V_{T(ud,lc)}$ are typical of the NMOS device parameters encountered in the early days of MOS ICs. Hence, at that time it was not possible to easily and/or reliably fabricate enhancement-mode NMOS transistors. In chap. 4 we shall show how this problem was overcome by the development of ion implantation. With this process it became possible to introduce dopants into the channel region near the silicon surface, such that the threshold voltage is shifted from the value it has in the case of a uniformly doped substrate, to the desired value. This enabled enhancement-mode NMOSFETs to be manufactured with appropriate threshold voltage values.

The typical value of V_T in NMOSFETs is 0.7V. This means that V_T of the MOS-C structure of Example 3-2 must be shifted by 1.11V. Equation 4-73 indicates that a shallow implant with a dose of ~$1.6\text{x}10^{12}$ boron ions/cm^2 will produce such a V_T shift in the MOSFET of Example 3-2.

3.7 VISUALIZATION OF MOS-C AND MOSFET PHENOMENA USING POTENTIAL VERSUS POSITION DIAGRAMS

Many aspects of the behavior of the MOS-C and the MOSFET are described in terms of the variation of the electrostatic potential in the silicon as a function of position. By plotting such relationships in a graphical format, a visual model is obtained which helps to portray many of the phenomena associated with MOS-C and MOSFET behavior.

When describing the MOS-C most graphical representations are one-dimensional, reflecting the one-dimensional nature of the behavior of such structures [i.e., usually $\varphi(x)$ versus x plots are presented]. However, when the potential in the MOSFET is being depicted, a two-dimensional representation is needed to appropriately depict the MOSFET behavior. Thus, we have decided to illustrate the potential versus position in the MOS-C using both the 2-D plot (as is needed for the MOSFET), as well as the 1-D potential plots more usually encountered when MOS-C behavior is being studied. This is done in Fig. 3-21, in which both 1-D and 2-D potential plots for a MOS-C cross-section are depicted. We note that $\varphi_{surf} = \varphi\ (x = 0)$ is constant versus *y* in the MOS-C, and thus the 1-D representation of the potential in the MOS-C would be

adequate. However, when the 2-D potential versus position diagrams of the MOSFET are presented (i.e., in chaps. 4 and 5), it will be easier to interpret their details by comparing them to the simpler 2-D cases introduced here.

In Fig. 3-21a we show the 2-D cross-section of a *p*-substrate MOS-C, with a bias V_{GB} connected between the gate and body. In Fig. 3-21b we show the case of the *ideal* MOS-C described in Example 3-1. We observe that when $V_{GB} = 0$ (Fig. 3-21b1), $\varphi(x,y) = 0$ everywhere, corresponding to the flat-band condition in the energy-band representation. In Fig. 3-21b2 we portray the case of $V_{GB} = V_T = 0.64V$ for the MOS-C of Example 3-1. Here we note that $\varphi_{surf} = 2\varphi_B = 0.58V$, with the variation of $\varphi(x)$ expressed by Eq. 3-23.

In Fig. 3-21c we note that the *real* MOSFET of Example 3-2 is depicted for the bias condition of $V_{GB} = 0V$. Since we noted that $V_T = -0.39V$, implying that a negative voltage of -0.39V must be applied to cause a condition of $\varphi_{surf} = 2\varphi_B = 0.58V$, which corresponds to the onset of strong inversion. Thus, when $V_{GB} = 0V$, $\varphi_{surf} = 0.97V$, which is larger than $2\varphi_B$, as shown in Fig. 3-21c1.

In order for this MOS-C to exhibit a value of $V_T = 0.7V$, we indicated in the previous section that a dose of 1.6×10^{12} boron ions/cm^2 must be implanted into the channel region. If this is done, when $V_{GB} = 0V$, φ_{surf} will then equal -0.05V, and the potential versus position diagram will as shown in Fig. 3-21d1. Note that when $V_{GB} = 0.7V$, $\varphi_{surf} = 0.58V$, but the potential versus x variation is different from that of the uniformly doped case of Fig. 3-21b. The effect of the implant dose on the depletion region width of the channel and other MOSFET phenomena are discussed in chap. 4.

REFERENCES

1. E.H. Nicollian and J.R. Brews, *MOS Physics and Technology,* Wiley-Interscience, New York, 1982.
2. C.E. Young, "Extended Curves of the Space Charge, Electric Field, and Electrostatic Potential Inside a Semiconductor," *J. Appl. Phys.,* **32**, p. 329, (1961).
3. R.F. Pierret and J.A. Shields, "Simplified long-channel MOSFET theory," *Solid-State Electronics,* vol. 26, p. 143, 1983.
4. G. Baccarani et al., "Analytical i.g.f.e.t.model including drift and diffusion currents," *IEE Journal on Solid-State and Electron Devices,* vol. 2, p. 62, March 1978.
5. J.R. Brews, "A Charge-Sheet Model of the MOSFET," *Solid-State Electronics,* vol. 21, p. 345, 1978.
6. Y.P. Tsividis, *Operation and Modeling of the MOS Transistor,* McGraw-Hill, New York, 1987.
7. W.M. Werner, "The Work Function Difference of the MOS-System with Aluminum Field Plates and Polysilicon Field Plates," *Solid-State Electronics,* **17**, 769 (1974).
8. P. Balk Ed., *The Si-SiO$_2$ System,* Elsevier, New York, 1988.
9. B.E. Deal et al., "Characteristics of the Surface State Charge of Thermally Oxidized Silicon," *J. Electrochem. Soc.,* **114**, p. 266 (1967).
10. B.E. Deal et al., *J. Electrochem. Soc.,* **121**, p. 198C (1974).

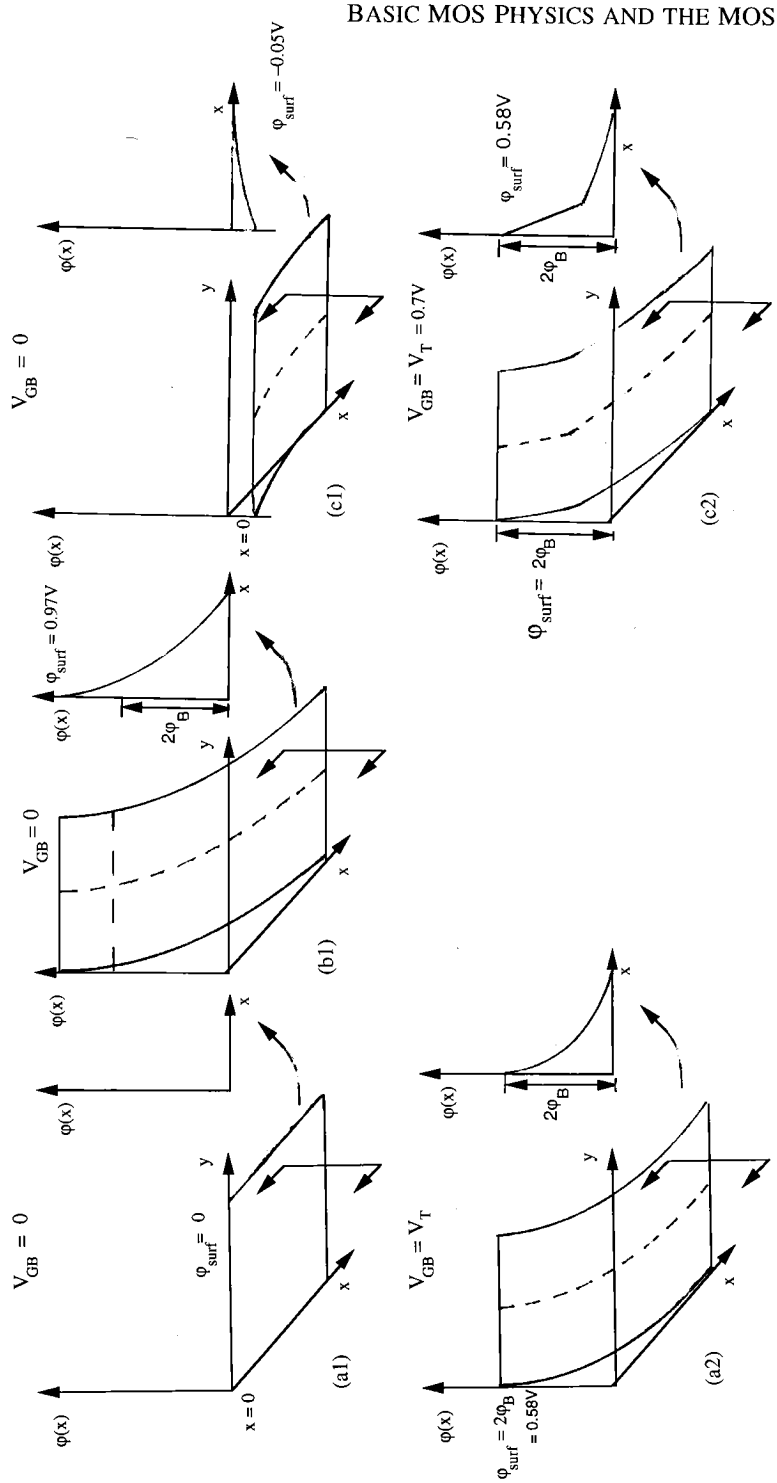


Fig. 3-21 Visualization of the electrostatic potential in *p*-type MOS-Cs with $N_A = 10^{15}$ cm^{-3}, and $t_{OX} = 150$Å. (a1) $\varphi(x,y)$ in an *ideal* MOS-C ($\varphi_{ms} = 0$ and $Q_{tot} = 0$) when $V_{GB} = 0$V (flatband condition); and (a2) when $V_{GB} = V_T$. (b1) $\varphi(x,y)$ in *real* MOSFET with uniformly doped substrate and n^+ polysilicon gate ($\varphi_{ms} = 1.02$V), and $Q_{tot} = 2\text{x}10^{10}$ cm^{-2}, when $V_{GB} = 0$V. (c1) $\varphi(x,y)$ in *real* MOSFET with n^+ polysilicon gate and V_T-adjust implant into substrate such that $V_T = 0.7$V, when $V_{GB} = 0$V. (c2) $\varphi(x,y)$ when $V_{GB} = V_T$.

11. P. Roblin, S. Samman, and S. Bibyk, "Simulation of Hot-Electron Trapping and Aging in nMOSFETs," *IEEE Trans. Electron Dev.*, **ED-35**, December 1988, p. 2229.

12. R.F. Pierret, *Modular Series on Solid-State Devices; Vol. IV - Field Effect Devices.*, R.F. Pierret and G.W. Neudeck, Eds., Addison-Wesley Publ. Co., Reading MA, 1983.

13. B.E. Deal, "Standardized Terminology for Oxide Charges Associated with Thermally Oxidized Silicon," *IEEE Trans. Electron Dev.*, **ED-27**, p. 606, (1980).

PROBLEMS

1. Explain why it is useful to study the behavior of the MOS-C prior to studying the behavior of the MOSFET.

2. Sketch the energy-band diagrams of a MOS-C with a heavily-doped *n*-polysilicon gate when (a) $V_{GB} = 0$, and (b) in the flat band condition when the Si substrate: (1) 1-Ω-cm *n*-type, and (2) 1-Ω-cm, *p*-type.

3. For an ideal MOS-C with $t_{OX} = 30$ nm and $N_A = 5\text{x}10^{15}$ cm^{-3}, find the applied voltage and electric field at the interface needed to: (a) make the silicon surface intrinsic; and (b) to bring about strong inversion.

4. Using Eq. prove that for $n_{surf} = 10N_A$, φ_{surf} is only 58 mV greater than $-\varphi_B$.

5. Calculate the flatband voltage and the threshold voltage (a) in 1 Ω-cm *p*-type silicon and (b) in 1 Ω-cm *p*-type silicon. The MOS-Cs for each case have : (i) aluminum gate for which $q\varphi_m = 4.3$ eV, (ii) $t_{OX} = 100$ nm; and (iii) the oxide is free of charge except for a surface density of $Q_f = 5\text{x}10^{10}$ cm^{-2}. The electron affinity of silicon is 4.05 eV.

6. Use the "threshold" statement in SUPREM III to extract and plot V_T vs. V_{SB} for increments of 0.25V between 0 and 5V for the device specified in Prob. 4. Compare the zero bias ($V_{BS} = 0$V) value with the two corresponding values of V_T found in Prob. 4.

7. Consider Poisson's equation in the substrate of a MOS-C. Using the condition that there is charge neutrality in the substrate, show that the impurity doping terms $N_D - N_A$ can be written as $n_i (\exp(u_B) - \exp(-u_B)] = 2 n_i \sinh u_B$, where u represents the potential normalized to (kT/q) and u_B is the substrate potential ($u_B = -q\varphi_B/kT$) in a *p*-type substrate.

8. Using the form derived in Problem 6, show that the Poisson equation can be written in terms of u as:

$$d^2u/dx^2 = (1/L_{Di})^2 [\sinh(u) - \sinh(u_B)]$$

where L_{Di} is the intrinsic Debye length

$$L_{Di} = (\varepsilon_{si}\, k\, T/2q^2 n_i)^{1/2}$$

9. Calculate the maximum depletion layer depth in a MOS-C with a *p*-type substrate doped uniformly at $2\text{x}10^{15}$ cm^{-3} if T = 77K and 300K.

10. Consider the establishment of a p-type inversion layer in a MOS-C. Show that for a nondegenerate n-type semiconductor, the surface charge density Q_C needed to make $p_{surf} = n_o$ (the equilibrium electron concentration in the bulk) is given by

$$Q_C = \sqrt{2\varepsilon_{si}\, q\, N_D\, (\varphi_B)}.$$

CHAPTER 4

MOS TRANSISTOR DEVICE PHYSICS: PART 2

LONG-CHANNEL MOSFETS

A description of the physics of the MOS capacitor (MOS-C, chap. 3) provides a solid foundation for analyzing the behavior of the MOSFET. Hence, in this chapter the device physics of long-channel MOSFETs is undertaken. In chap. 5 we examine the behavior of short-channel (i.e., submicron) MOSFETs.

Long-channel MOSFETs are discussed first for the following reasons:

1. The derivation of accurate long-channel-MOSFET drain-current models can be achieved by solving both the *Poisson equation* and the *drift-diffusion current density equations* in one-dimensional form, whereas accurate modeling of short-channel MOSFET behavior requires the solution of these equations in two (or even three) dimensions. Thus, it is logical to undertake the simpler analysis first.

2. The long-channel drain-current MOSFET models derived from the solution of the 1-D device equations can be expressed in analytical form. Hence, these models are widely used in circuit-design applications. In addition, quantitative insight about the device physics of MOSFET operation can also be obtained by examining these models. On the other hand, physically based analytical models of short-channel MOSFET behavior do not arise from the solution of the two-dimensional device equations, and thus it is harder to gain quantitative insight into MOSFET operation from them.

3. The analytical expressions for I_D from two of the four long-channel MOSFET models derived in this chapter are the I_D expressions used in SPICE for simulating MOSFET behavior in IC circuit design applications.

4. As MOSFETs are scaled down, they eventually become prone to 'short-channel' effects. In most cases, however, these effects are considered detrimental. That is, it is the task of the device designer to suppress them when scaling-down MOSFET structures. The goal is to produce 'short-channel' MOSFETs with 'long-channel' behavior. Thus, it is important to define long-channel behavior and to describe the physics that produce such characteristics in MOSFETs. This understanding provides insight into the measures that must be employed to suppress unwanted short-channel effects.

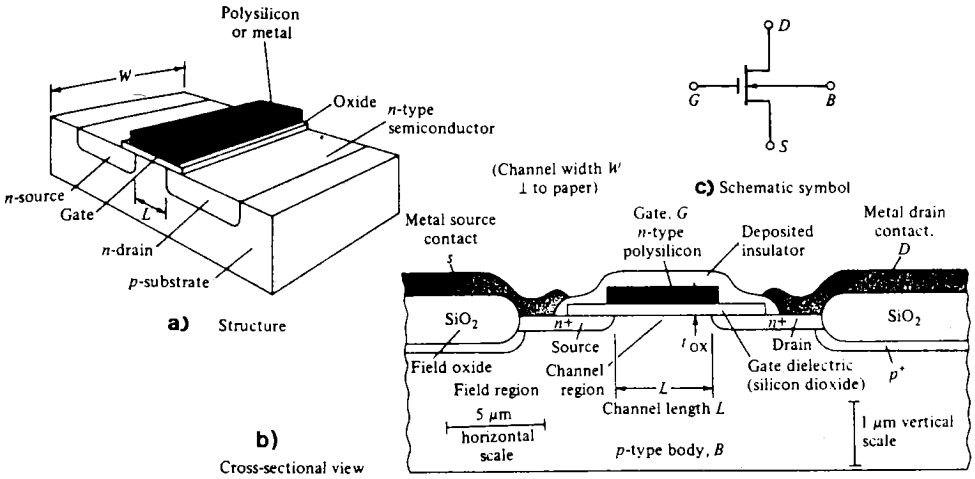


Fig. 4-1 (a) Structure of an MOS device. (b) Cross sectional view. (c) Schematic symbol.[1] From D. A. Hodges and H. G. Jackson, Analysis and Design of Digital Integrated Circuits, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission.

In this chapter the following topics dealing with MOSFETs are covered:

- **a.** Introduction to the terminology of MOSFETs and their behavior
- **b.** Effect of Applying a Bias to the Inversion Layer of the MOS-C
- **c.** Definition of the Long-Channel MOSFET and the Gradual-Channel Approximation
- **d.** Qualitative Description of Long-Channel MOSFETs
- **e.** Quantitative Models of dc Long-Channel MOSFET Behavior
  - **i.** Pao-Sah Model
  - **ii.** Charge-Sheet Model
  - **iii.** Variable-Depletion-Charge Model
  - **iv.** Square-Law Model
- **f.** Threshold-Voltage Control in Long-Channel MOSFETs
  - **i.** Body Effect
  - **ii.** Threshold-voltage adjustment by ion-implantation
- **g.** Subthreshold behavior of Long-Channel MOSFETs
  - **i.** Physics and modeling of subthreshold current flow
  - **ii.** Subthreshold swing
  - **iii.** Gate-induced leakage current (GIDL)

4.1 TERMINOLOGY OF MOSFET DEVICE STRUCTURES

Figure 4-1 shows the cross-section of the MOSFET, as introduced in chap. 3. The important difference between the MOS-C and the MOSFET is that in the latter a *pn*

junction exists at each side of the MOS-C channel region. Each of these junctions are also close enough to make contact with the inversion layer. By making external connections to each of these edge junctions, the MOSFET becomes a four-terminal device (with the terminals designated respectively as the *gate* G, *body* B, *source* S, and *drain* D). The MOS-C, on the other hand, remains a two terminal device (*gate* and *body*). The source and drain terminals of the MOSFET allow a bias V_{DS} to be applied across the ends of the channel region, giving rise to a *drain current* I_D.

As described in chap. 3, the inversion layer can be established by applying an appropriate bias to the gate. For the device shown in Fig. 4-1a, the induced charge in the channel inversion layer is *n*-type, and the device is known as an *n*-channel MOSFET (or simply an NMOS transistor). A *p*-channel MOSFET (or PMOS) is obtained by interchanging the *n* and *p* regions. Just as in the ideal MOS-C it is assumed that current in the gate oxide (*gate current,* I_G) is negligible. We also assume that the MOSFET is a symmetrical device, insofar as connecting the two junction terminals to a circuit (i.e., the drain current should remain the same if the two terminals are interchanged). Regardless of the way a MOSFET is connected, the terminal tied to the more positive voltage (in NMOS circuits) is defined to be the *drain*, and the other is the *source*.

4.1.1 Basics of MOSFET Operation

In simplest terms, the operation of a MOS transistor involves the application of an input voltage to the gate electrode. This establishes an electric field perpendicular (or transverse) to the Si-SiO_2 interface in the channel region of the device (Fig. 4-2a). The conductance of the channel region (in which the device output current I_D flows) can be modulated by varying this electric field. Since an electric field is responsible for controlling the output current flow, such devices are termed *field-effect transistors* (FETs).

If no gate bias is applied, however, the circuit path between source and drain consists of two back-to-back *pn* junctions in series. In this case, if V_{DS} is applied (such that the drain and source *pn* junctions are reverse biased), I_D will consist of only the reverse-bias diode leakage current, which is normally negligible.

When positive bias is applied to an NMOS transistor gate, however, electrons will be attracted to the channel region and holes will be repelled (as is the case in the MOS-C). Once the positive gate voltage becomes strong enough to form an inversion layer, an *n*-type channel is formed that connects the source and drain regions (Fig. 4-2a). A drain current I_D can then flow if a voltage V_{DS} is applied between the source and drain terminals. In the simplest analysis we assume that the voltage-induced *n*-type channel does not form unless the voltage applied to the gate exceeds the threshold voltage V_T.

MOS devices such as those just described (in which no conducting channel exists when $V_{GS} = 0$), are referred to as *enhancement-mode* (or *normally OFF*) transistors (see Fig. 4-2b). With NMOS enhancement-mode transistors, a positive gate voltage, V_{GS}, greater than V_T must be applied to create the channel (or to turn them ON), while to turn ON PMOS enhancement-mode devices, a negative gate voltage (whose magnitude is $>V_T$) must be applied. Note that in NMOS transistors a positive voltage must also

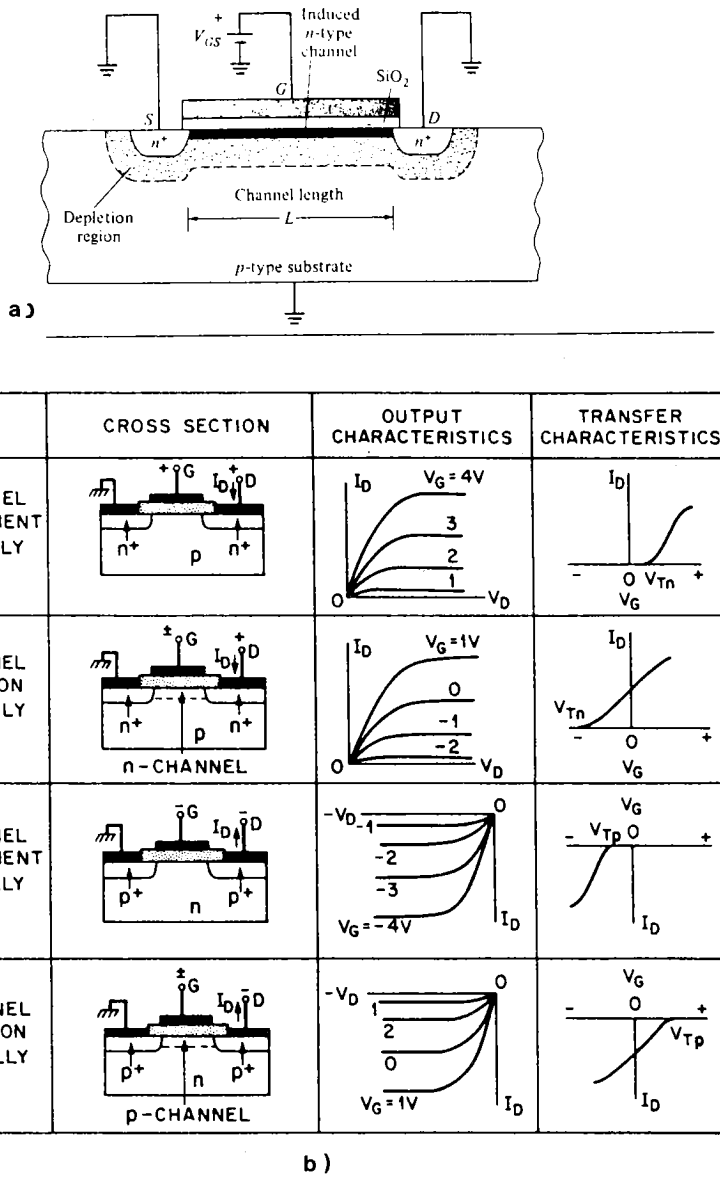


Fig. 4-2 (a) Idealized NMOS cross section with positive V_{GS} applied showing depletion regions and induced channel. From D. A. Hodges and H. G. Jackson, Analysis and Design of Digital Integrated Circuits, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission. (b) Cross sections and output and transfer characteristics of four types of MOSFETs. From S.M.Sze, Semiconductor Devices - Physics and Technology, 1985, John Wiley & Sons. Reprinted with permission.

be applied to the drain to keep the drain-substrate reverse-biased, while in PMOS devices this voltage must be negative.

On the other hand, it is also possible to build MOS devices in which a conducting channel region exists when $V_{GS} = 0V$ (see also Fig. 4-2b), and such MOS devices are described as being *normally ON*. Since a bias voltage to the gate electrode is needed to deplete the channel region of majority carriers, and thus turn them OFF, such devices are also commonly called *depletion-mode MOSFETs*. NMOS depletion-mode devices require a negative gate voltage to be turned OFF, while corresponding PMOS devices require positive gate voltages.

4.2 BIASING THE INVERSION LAYER (BODY EFFECT)

Before undertaking a detailed discussion of MOSFET behavior, we first study the consequences of *applying a dc bias to the inversion layer of a MOS-structure*. This topic is actually an extension of the subject of MOS-C behavior, but also serves as an important prelude to the discussion of MOSFET operation. Consequently, when we discuss the operation of the MOSFET we will refer back to the results developed in this section.

We examine the effect of biasing the inversion layer from both the qualitative and quantitative perspectives. In the latter we will show how the expressions for the mobile charge concentrations in the inversion layer of the MOS structure are altered when a such bias is applied, as well as how the application of a bias to the inversion layer modifies the threshold voltage equation derived for the MOS-C.

Biasing the inversion layer is accomplished by the application of a dc voltage between the substrate and a diffused region adjacent to, and in contact with, the inversion layer. This implies that a modified MOS-C structure must be used. Specifically, a diffused region must be added to the edge of the MOS-C channel region (see Fig. 4-3a), and a terminal must be connected to this region (designated by the letter C). The result, as shown, is the *three-terminal MOS device* depicted in Fig. 4-3a.

If we choose the substrate (body or B) terminal as the reference, the voltage applied between terminal C and the substrate (body) is denoted by V_{CB}, and the voltage between the gate (terminal G) and substrate by V_{GB} (see Fig. 4-3a).* For simplicity we also assume an ideal *p*-type Si-substrate MOS structure (i.e., $\varphi_{ms} = 0$ and $Q_{tot} = 0$).

4.2.1 Qualitative Discussion of the Effect of Applying a Bias to the Inversion Layer of a MOS Structure

If a MOS-C with a *p*-type substrate is biased into inversion by a gate voltage V_{GB}, a *pn* junction exists between the surface and the bulk of the silicon (as a result of the region near the silicon surface having been inverted to *n*-type). If there is an adjacent *n-doped* region in contact with the *induced* *n*-type inversion layer, it is possible to apply a bias to the induced junction by applying a bias V_{CB} between terminal C and B (the substrate)

* As elsewhere, we use the convention that the second letter of the voltage subscript designates the reference terminal.

as shown in Fig. 4-3a. (Note that here we have a MOS structure in which the induced junction is contacted only at one end of the channel.*) We also assume that the polarity of V_{CB} is such that the induced *pn* junction is either unbiased or under reverse bias (i.e., V_{CB} in Fig. 4-3a is $\geq$0V). We will refer to V_{CB} as the *channel voltage.*

Before presenting the equations relevant to this structure, let us first consider the qualitative effect of applying various V_{CB} bias values.

4.2.1.1 Case When Both V_{GB} and V_{CB} = 0V. In Fig. 4-3b we show the two-dimensional potential φ versus position diagram for the three-terminal MOS structure when both V_{GB} and V_{CB} = 0V. We note that the channel region is in a flat-band condition and that a built-in voltage φ_{bi} exists across the n^+p (doped) junction. We also recall that even if $V_{GB} \neq$ 0V, as long as V_{CB} remains at 0V, the structure is still in equilibrium, because I_G is assumed to be zero and there is no current in the n^+p junction if V_{CB} = 0V. Hence, this implies that when V_{CB} = 0V the channel region of this structure exhibits the same behavior as that of the MOS-C described in chap. 3.

4.2.1.2 Case When $V_{GB} \geq V_T$ and V_{CB} = 0V. If we now consider the case where V_{CB} = 0V, but $V_{GB} \geq V_T$, we encounter a situation as shown in Fig. 4-3c. That is, the channel region is now inverted, and a *field-induced junction* now exists between the *n*-type inversion layer and the underlying *p*-type substrate. Since a bias is not applied across the field-induced junction, it is in equilibrium and is again characterized by the same Fermi level as in the diffused junction. The surface depletion region in the field-induced junction under these conditions is also described by the equilibrium theory presented in chap. 3.

Note, however, that the potential difference between the inversion layer and the n^+ region is now much smaller than in the case when V_{GB} = 0V. Specifically, this potential difference is now ~(φ_{bi} - $2\varphi_B$). This implies that the inversion layer is in "communication" with the n^+ region adjacent to it. The term "communication" means that once the inversion layer is formed and a bias V_{CB} is applied between terminals C and B, this bias will appear across the entire length of the field-induced junction (i.e., no longer just across the doped n^+p junction at the edge of the channel). Thus, *if an inversion layer exists in the channel, then V_{CB} does have an influence on φ_{surf}.*

For example, if V_{CB} is suddenly allowed to increase from 0V to some non-zero value (while V_{GB} is held fixed), electrons will be observed to flow out of the inversion layer through the C terminal. Why this happens, and what the effects are on the behavior of the MOS structure are described next.

4.2.1.3 Case When V_{GB} = 0V, But V_{CB}>0V. If we next apply a non-zero V_{CB} bias while keeping V_{GB} = 0V, equilibrium will cease to exist in the three-terminal structure. This is because V_{CB} causes a small reverse-bias leakage current to flow in the n^+p junction. (Although this reverse-bias leakage current is extremely small, its

* However, in a MOSFET, contact to the inversion layer can be made at both ends of the induced junction since a diffused region exists at each end of the channel. In the MOSFET these two regions are called the source and drain, and the terminals are designated by the abbreviations S and D.

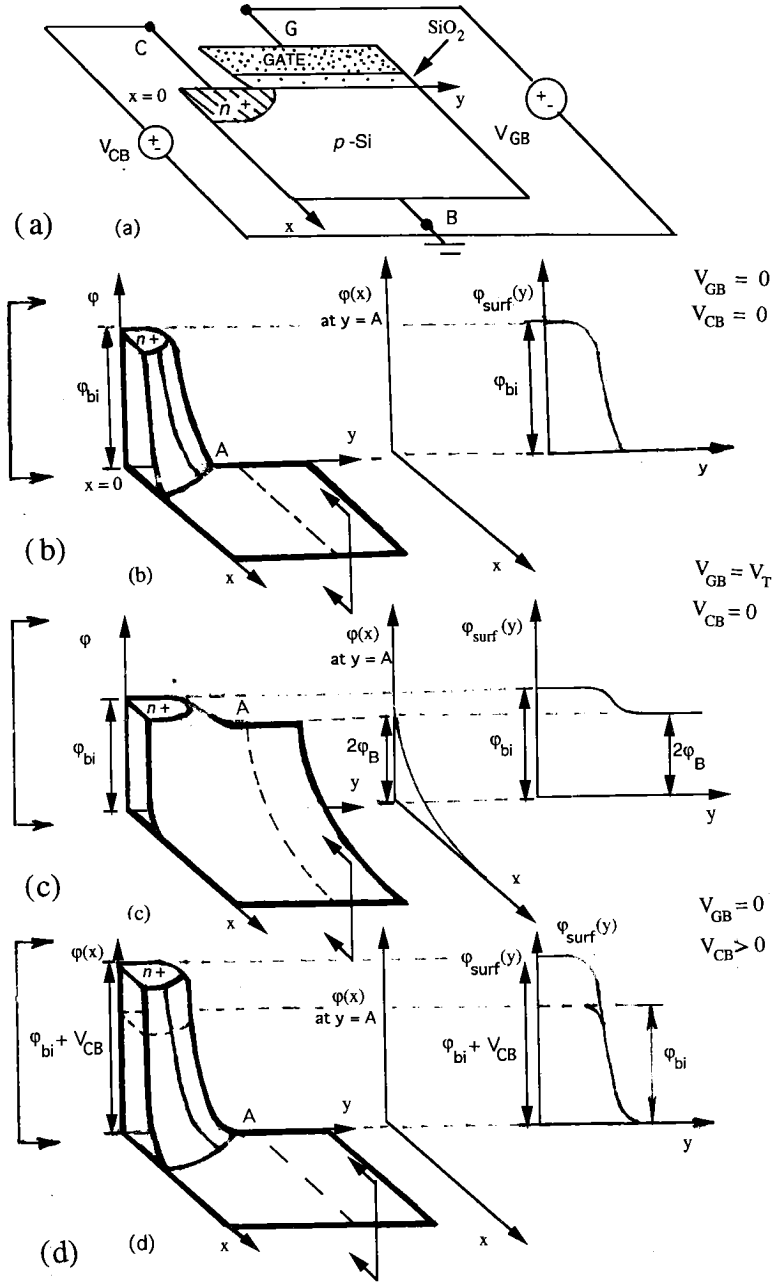


Fig. 4-3 (a) Applying a bias V_{CB} to an induced channel beneath the gate oxide by applying this bias V_{CB} between terminal C and the substrate of a three-terminal MOS structure. Two-dimensional potential φ versus position diagrams in the three terminal MOS structure when: (b) both V_{GB} and $V_{CB} = 0$; (c) $V_{GB}>V_T$ and $V_{CB} = 0V$; (d) $V_{GB} = 0V$ and $V_{CB}>0V$.

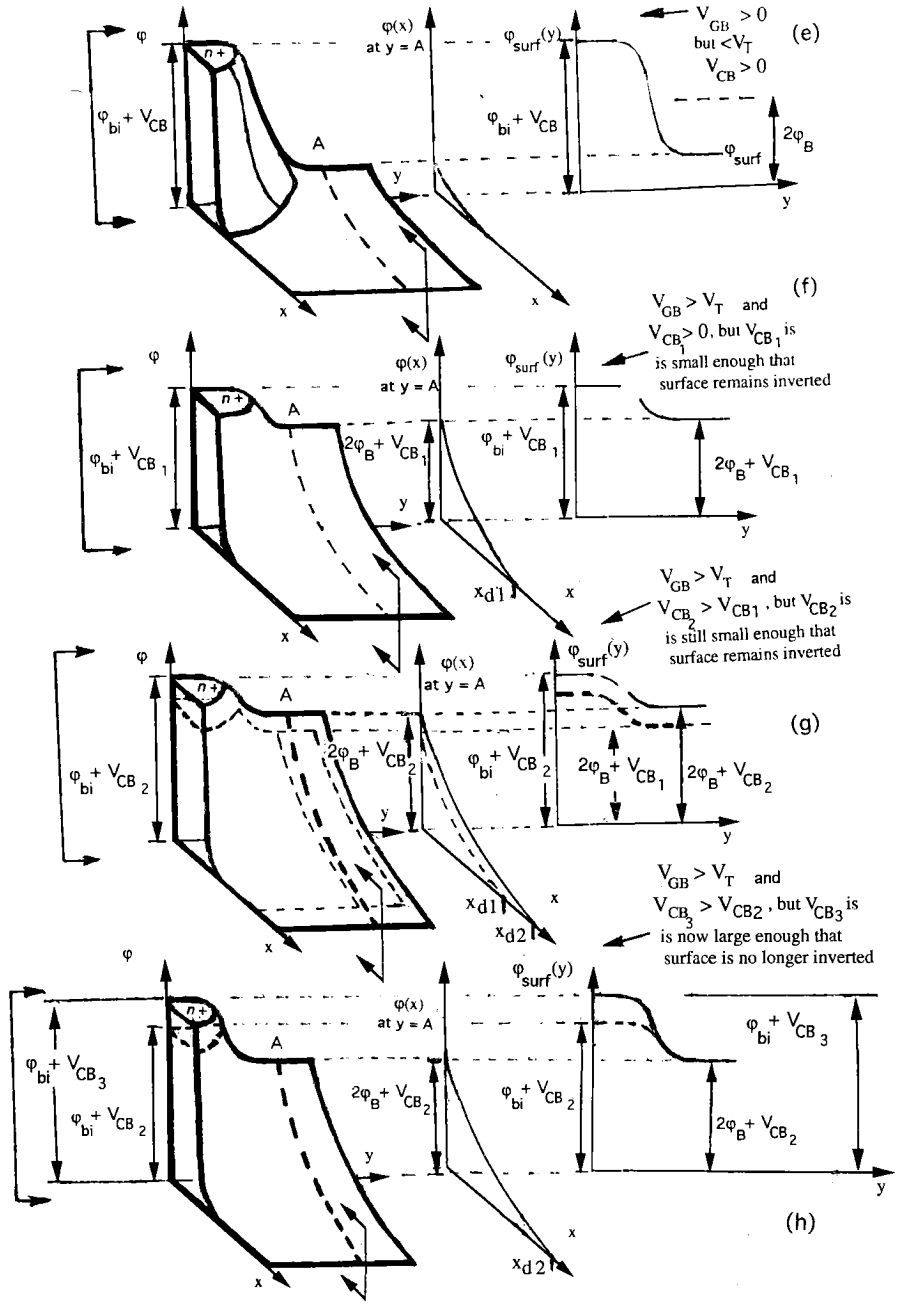


Fig. 4-3 (cont.) (e) $V_{GB}>0$, but is $<V_T$, and $V_{CB} >0$; (f) $V_{GB}>V_T$ and $V_{CB}>0$, but $V_{CB} = V_{CB1}$, which we assume is small enough so that inversion still exists. (g) V_{GB} is the same as in part (f), but $V_{CB} = V_{CB2}$, where $V_{CB2}>V_{CB1}$, but V_{CB2} is still small enough to keep surface inverted. (g) $V_{GB} = V_T$, $V_{CB} = V_{CB3}$ (large enough to suppress the channel inversion layer).

existence implies that the MOS structure is no longer in equilibrium because some of the carrier energy now depends on an externally applied force, namely the extra electric field created in the depletion region by V_{CB}.) However, if V_{GB} remains fixed at 0V, V_{CB} is dropped entirely across the n^+p junction. Hence, the surface potential along the channel is not altered from when V_{CB} = 0V. Instead, only the potential difference between the n^+ and p (substrate) regions is increased by V_{CB} (as shown in Fig. 4-3d), leaving the channel region to the right of the n^+p junction in a flat-band state. In addition, Fig. 4-3d indicates that *if no inversion layer is present, changing V_{CB} has negligible impact on the surface potential φ_{surf} along the channel.* This last statement also applies to the case illustrated in Fig. 4-3e (i.e., V_{GB}>0V, but is less than V_T, and V_{CB} is increased from 0V to some value greater than 0V). Here we observe that there is still no impact on φ_{surf} as long as no inversion layer is present in the channel.

4.2.1.4 Case When $V_{GB} \geq V_T$ and V_{CB}>0V. Let us now consider the case when $V_{GB}>V_T$ and a reverse-bias voltage V_{CB} is applied to the induced *pn* junction (such that the surface remains inverted even after V_{CB} is applied). Since the surface is inverted, the potential in the inversion region is very close to the potential value in the diffused region, as shown in Fig. 4-3f (and as exists in the case shown in Fig. 4-3c). That is, a point at the surface is at about the same potential (with respect to the substrate deep in the bulk) as a point in the n^+ region. Therefore, the surface is about as "attractive" for electrons as is the n^+ region. Consequently, electrons have no reason to prefer only the n^+ region as opposed to the "surface" at the right. As a result, the n^+ region can "communicate" with the inversion layer (implying that a bias applied to the n^+ region is also applied to the inversion layer). In addition, just as in the case of a MOS-C, the value of the surface potential φ_{surf} remains clamped at the value it has at the onset of strong inversion (i.e., for the case of Fig. 4-3f, at a value of $\varphi_{surf} \sim 2\varphi_B + V_{CB1}$). That is, as V_{GB} is increased beyond V_T, φ_{surf} remains unchanged, but the mobile charge concentration in the inversion layer continues to increase.

We also recall from *pn* junction theory that if a reverse-bias voltage is applied to a previously unbiased *pn* junction, the junction depletion-region width will increase. Thus, the depletion region on the *p*-side of the induced junction will also increase if a reverse-bias is applied between terminals C and B. However, if V_{GB} is kept fixed (such that $V_{GS} \geq V_T$) as the reverse-bias voltage V_{CB} is increased (see Fig. 4-3g), the charge on the gate terminal will remain the same as when V_{CB} = 0V. Therefore, the *total* charge in the semiconductor must also remain unchanged to maintain balance with the unchanged value of charge on the gate. However, since the application of V_{CB} widens the depletion region (thereby uncovering more charge in this region of the induced junction), fewer electrons are needed in the inversion layer to balance the unchanged gate charge. Consequently, electrons must exit the inversion layer, resulting in a lower level of inversion. (Looking at it in another way, the increased positive potential at terminal C will attract electrons from the inversion layer, which will start flowing toward the n^+ region, and from it into the top terminal of the voltage source.) However, if a sufficiently large value of V_{CB} is applied (while holding V_{GB} fixed), this effect can entirely extinguish the charge in the inversion layer. In other words, a gate voltage

capable of causing inversion in the three-terminal MOS structure when $V_{CB} = 0V$ may no longer be able to sustain inversion if a non-zero value of V_{CB} is applied.

As also shown in Figs. 4-3f and g, as long as the surface remains in inversion as V_{CB} is increased, the depletion region under the field induced junction increases (from x_{d1} in Fig. 4-3f to x_{d2} in Fig. 4-3g). This means that the bands are bent more, and thus φ_{surf} also increases. This supports the statement that as V_{CB} increases, the n^+ region communicates this change in V_{CB} to the field-induced junction. However, once a sufficiently large number of electrons depart from the inversion layer so that it essentially disappears, further increases in V_{CB} no longer influence the channel region, and the depletion region no longer widens (as shown in Fig. 4-3h), and remains fixed at a value dependent on V_{CB} (e.g., at x_{d2}, see Figs. 4-3g and h) as expressed by Eq. 4-8.

If a non-zero V_{CB} is applied and it is desired to re-establish inversion, φ_{surf} must be increased to a value larger than that needed when $V_{CB} = 0$. Specifically, at the onset of inversion, it is necessary to increase V_{GB} so that $q\varphi_{surf} \sim q(V_{CB} + 2\varphi_B)$. Note that the effect just described can also be exploited as a technique for raising the threshold voltage of a MOSFET, as will be discussed in section 4.6.2.

4.2.1.5 Summary of the Behavior of the Three-Terminal MOS Structure and the Significance of Such Behavior on the Operation of the MOSFET. In summary, it is possible to establish "communication" with the channel region (i.e., an external bias voltage can be applied to the channel region of a MOS structure) if an inversion layer has been induced there. On the other hand, if the channel region is not inverted, the application of a bias voltage to a diffused region adjacent to the channel region will have no effect on the surface potential in the channel. Such "communication" (or lack of) has the following impact on the behavior of a MOSFET:

1. If the gate voltage applied to a MOSFET is less than the threshold voltage, the device is said to be operating in the subthreshold regime. From our discussion, this means that if a bias is applied to the drain (V_{DS}) of a MOSFET being operated in subthreshold, this voltage will dropped entirely across the diffused n^+p junction (i.e., the drain-substrate junction), and it will have no impact on φ_{surf}. Hence, φ_{surf} will be determined by the gate potential, which is constant everywhere on the gate. Consequently, the potential drop along the y direction at the surface of the channel will be essentially zero. We will use this information when we discuss subthreshold current flow in MOSFETs in section 4.7.

2. If the channel is inverted, and V_{CB} is increased (and the channel remains in inversion for the range of applied V_{CB} values), the depletion region of the field-induced channel increases. If V_{GB} is held fixed, the number of electrons in the inversion layer will decrease as V_{CB} is increased. At some value of V_{CB} the inversion layer charge will entirely disappear. These effects occur when a drain bias V_{DS} is applied to a MOSFET. That is, for a fixed gate voltage (greater than the threshold voltage value), as V_{DS} is increased, the depletion region of

will correspondingly decrease, making it less inverted. Thus, in a MOSFET operating in inversion and with $V_{DS}>0$, the depletion region near the drain will be wider than at the source (i.e., assuming the source is the common terminal).

3. "Communication" with the channel also permits the diffused region to become an additional source (or sink) of mobile carriers to the channel (i.e., in addition to the carriers thermally generated in the depletion regions). As noted, two such diffused regions exist in a MOSFET, one which supplies mobile carriers (the source), and the other which removes them (the drain). Specifically, if a drain bias is applied to a MOSFET in inversion, the channel carriers will be attracted to and move into the drain. These are replenished by carriers entering the channel from the source. The resultant flow of charge in the channel constitutes a current in the drain terminal I_D. Since the channel is not a perfect conductor, this current also gives rise to a potential drop along the direction of the channel.

4.2.1.6 Body Effect. Until now we have chosen the substrate to be the reference terminal in the three-terminal MOS structure. However, when MOSFETs are used in circuits, the source terminal S (corresponding here to terminal C), is normally chosen to be the reference. Hence, *body effect* is the term used to describe the increase in the threshold voltage of a MOSFET (with reference to the source) upon the application of a reverse bias to the inversion layer. If, as shown in Fig. 4-4b, we choose terminal C to be the reference terminal (instead of the body terminal), the bias voltages are now applied between the gate and C (V_{GC}), and between the body and C (V_{BC}). In this case, for a given value of V_{BC}, a value of V_{GC} greater than that needed when $V_{BC} = 0V$ must be applied to the MOS structure to establish strong inversion. In section 4.2.2.2 we will derive an expression that determines the increase in V_T as a function of V_{BC}.

4.2.2 Quantitative Description of Effects of Biasing the Inversion Layer of MOS Structures

For the three-terminal MOS structure introduced above, it is possible to develop the quantitative relationships in a general manner for all cases of inversion. However, these can also be simplified somewhat for the more limited (but still most widely used form) of the strong inversion case. Note that in deriving these expressions we also include V_{FB}, which accounts for the work-function difference and oxide charge that can exist in 'real' MOS structures.

4.2.2.1 General Inversion Case. In the general inversion case (which covers weak as well as strong inversion), five equations characterize the phenomena in the channel region (i.e., the Si region in the vicinity of the Si-SiO_2 interface to the right of the n^+ junction in Fig. 4-3a) of three-terminal MOS structure (as they do for the MOS-C). If we invoke the charge-sheet and depletion approximations, four of these equations are identical to those for the expressions derived for the charge-sheet model of the MOS-C, and are repeated here for the readers convenience:

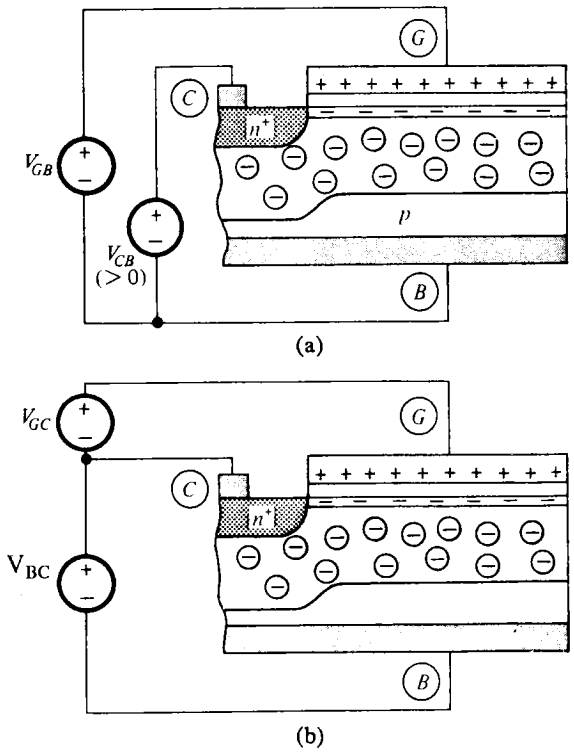


Fig. 4-4 Bias notation when: (a) terminal B is the reference terminal; (b) terminal C as the reference terminal instead of terminal B.[5] Reprinted with permission of McGraw-Hill Co.

$$V_{GB} = \varphi_{surf} + V_{FB} + \varphi_{ox} \qquad (3\text{-}10) \qquad (4\text{-}1a)$$

$$Q_C = Q_I + Q_B \qquad (3\text{-}20) \qquad (4\text{-}1b)$$

$$Q_C = -C_{ox}\varphi_{ox} \qquad (3\text{-}19) \qquad (4\text{-}1c)$$

$$Q_B = -\gamma C_{ox}\sqrt{\varphi_{surf}} \qquad (3\text{-}29) \qquad (4\text{-}1d1)$$

$$= [2qN_A\kappa_{si}\varepsilon_o(\varphi_{surf})]^{1/2} \qquad (4\text{-}1d2)$$

The fifth equation is a modified form of Eq. 3-53, in which we replace $2\varphi_B$ by $(2\varphi_B + V_{CB})$ to get

$$Q_I = \sqrt{2q\,N_A\kappa_{Si}\varepsilon_o}\left(\sqrt{\varphi_{surf} + \varphi_t e^{(\varphi_{surf} - 2\varphi_B - V_{CB})/\varphi_t}} - \sqrt{\varphi_{surf}}\right) \qquad (4\text{-}2)$$

Then, combining Eqs. 4-1a, 4-1c, and 4-1b we get

$$V_{GB} = \varphi_{surf} + V_{FB} - [Q_I(\varphi_{surf}) + Q_B(\varphi_{surf})]/C_{ox} \qquad (4\text{-}3)$$

Next, using Eqs. 4-1d1 and 4-2 in Eq. 4-3, V_{GB} can be written as

$$V_{GB} = \varphi_{surf} + V_{FB} + \gamma\sqrt{\varphi_{surf} + \varphi_t e^{(\varphi_{surf} - 2\varphi_B - V_{CB})/\varphi_t}} \qquad (4 - 4)$$

Now, if V_{GB} and V_{CB} are given, Eq. 4-4 can be solved iteratively for φ_{surf}. This value can then be inserted into Eq. 4-2 to calculate Q_I, or into Eq. 4-1d2 to calculate Q_B.

An expression for Q_I in terms of φ_{surf} and V_{GB} can also be derived from the above equations. That is, by inserting Eqs. 4-1c and 4-1d1 into Eq. 4-1b above, we get

$$- C_{ox}\varphi_{ox} = Q_I + \gamma C_{ox}\sqrt{\varphi_{surf}} \qquad (4 - 5)$$

Then, solving for φ_{ox}, and substituting the result into Eq. 4-1a, we obtain

$$Q_I = - C_{ox} (V_{GB} - V_{FB} - \varphi_{surf} - \gamma\sqrt{\varphi_{surf}}) \qquad (4 - 6)$$

4.2.2.2 Strong Inversion Case. For the three-terminal structure biased to strong inversion, the above general analysis can be simplified. That is, with $V_{CB} = 0V$, φ_{surf} of the three-terminal MOS structure at the onset of inversion is $2\varphi_B$. With a reverse bias applied to the induced junction, the surface potential at the onset of strong inversion becomes

$$\varphi_{surf}(inv) \approx 2\varphi_B + V_{CB} \qquad (4 - 7)$$

Since the applied V_{CB} extends the range of gate voltages under which the surface region remains in depletion, d_{max} is now larger. The expression for d_{max} for the case of an applied V_{CB} is (assuming the depletion approximation):

$$d_{max} = [2\kappa_{si}\varepsilon_o(2\varphi_B + V_{CB})/qN_A]^{1/2} \qquad (4 - 8)$$

and the depletion layer charge (per unit area), Q_{Bmax}, when $d = d_{max}$ is expressed as

$$Q_{Bmax} = [2q\kappa_{si}\varepsilon_o N_A (2\varphi_B + V_{CB})]^{1/2} \qquad (4 - 9)$$

The expression for the inversion layer charge per unit area Q_I in strong inversion in terms of V_{GB} and V_{CB} is

$$Q_I = - C_{ox} [V_{GB} - V_{FB} - 2\varphi_B - V_{CB} - \gamma (2\varphi_B + V_{CB})^{1/2}] \qquad (4 - 10)$$

Assuming again that Q_I is negligible at the onset of inversion, the total charge per unit area in the semiconductor is given by Q_B. Hence, the gate-substrate bias at the onset of strong inversion is

$$V_{GB} = V_{FB} + \varphi_{surf} + \varphi_{ox} = V_{FB} + (2\varphi_B + V_{CB}) + \gamma (2\varphi_B + V_{CB})^{1/2} \qquad (4 - 11)$$

Noting that $V_{GC} = V_{GB} - V_{CB}$, the threshold voltage, referenced to terminal C, is given by the expression

$$V_T \text{ (with ref. to term C)} = V_{T(ud,lc)} = V_{FB} + 2\varphi_B + \gamma(2\varphi_B + V_{CB})^{1/2} \quad (4 - 12)$$

Note that the term γ, which was defined in Eq. 3-60a as $[\sqrt{2qN_A\kappa_{si}\varepsilon_o}]/C_{ox} = [\sqrt{2qN_A\kappa_{si}\varepsilon_o}]t_{ox}/\varepsilon_{ox}$ is a factor which determines how much V_T will increase when V_{CB} is increased. Hence, γ has been given the name *body effect coefficient.* By inspecting the expression for γ we note that it predicts the body effect is stronger for heavier substrate dopings and/or thicker gate oxides. Conversely, the body effect can be made smaller by using lower doping concentrations and/or thinner gate oxides. Figure 4-5 shows the threshold voltage shift versus substrate bias for various γ values.[26]

4.3 DEFINITION OF TERM: LONG-CHANNEL MOSFET

It is appropriate at this point to provide a detailed definition of the term *long-channel MOSFET.* From a qualitative perspective, long-channel MOSFETs are assumed to have channel regions sufficiently long and wide so that "edge" effects along the four sides of the channel can be neglected. Insofar as the channel length is concerned, we assume the channel length L is much longer than the sum of the source and drain

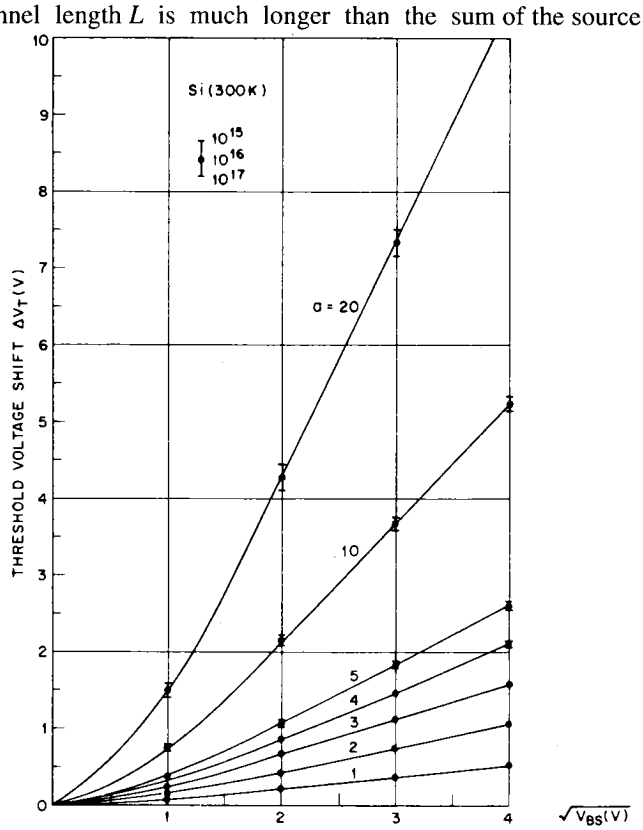


Fig. 4-5 Threshold voltage shift versus substrate reverse bias for various γ values.[2] Reprinted with permission of John Wiley & Sons.

depletion-region widths, or $L >> (d_{source} + d_{drain})$. The channel width, on the other hand is assumed to be much greater than the gate depletion-region depth, d (or Z>>d).

Even when the above criteria are well-satisfied, however, it is apparent that the electric fields and potential in the MOSFET still vary in at least two dimensions.* It seems that to accurately model device characteristics, the semiconductor device equations in their two-, or even three-dimensional forms must be solved. But in modeling long-channel MOSFET behavior this complexity can in fact be circumvented. That is, in such structures the solutions of the semiconductor device equations in their one-dimensional form yield drain-current device characteristics that are quite close to those obtained from measurements of actual long-channel MOSFETs.

Why is it possible to treat long-channel MOSFET behavior as a one-dimensional problem and still obtain relationships that give accurate estimates of I_D as a function of the applied the terminal voltages? First, consider the case of the MOSFET in which no drain bias is applied, and V_{GB} is the only external bias. We know that V_{GB} produces an electric field perpendicular to the device surface. Now, if $L >> (d_{source} + d_{drain})$ the junctions extend over such a small fraction of L that the $\mathcal{E}$-fields of these junctions have negligible impact on the $\mathcal{E}$-field due to the gate. Likewise, insofar as the channel width is concerned, if Z >>d the extension of the gate depletion region beyond the sides of the channel along which there is no source or drain hardly alters the total ionic charge uncovered by the gate voltage. Thus, the fringing field effect in the z direction can also be ignored. If both of these conditions are met, we can approximate the $\mathcal{E}$-field in a long-channel MOSFET when V_{DS} = 0V as being perpendicular to the surface everywhere along the channel. In this case, the one-dimensional analysis we applied to the MOS-C can also be used to determine Q_I, $\mathcal{E}$, φ and V_T in a long-channel MOSFET.

Next, if a drain bias V_{DS} is simultaneously applied, we assume that in long-channel MOSFETs the *gradual-channel approximation* is valid. This approximation assumes that the two aspects of drain current flow can be treated as strictly one-dimensional, independent problems. The two aspects are: (1) that the gate field acts approximately vertically to induce the channel (i.e., the mobile charge concentration that participates in I_D), and; (2) that the drain-to-source voltage acts approximately horizontally to move these mobile charges, and thus to drive the current I_D in the channel. The assumption that these two one-dimensional effects are independent of one another permits the results of the MOS-C analysis (shown to be valid in the case of a MOSFET with no drain bias), to also be applied when deriving the I_D-V_{DS} relationship of the MOSFET. Since the gradual-channel approximation is such a key simplification invoked when developing the models of long-channel MOSFET behavior, we need to examine it more rigorously.

* This is the case whether or not a drain current flows in the MOSFET. That is, if a gate bias is applied, but I_D = 0, there will still be electric fields arising from the source and drain junctions at the edges of the channel. If a drain current flows, it is obvious that there must be a component of the electric field along the y-direction. In fact, each of these cases should be examined separately when considering our definition in detail.

4.3.1 The Gradual Channel Approximation (GCA)

If a drain bias V_{DS} is applied to a MOSFET in inversion, drain current I_D flows. Owing to the channel's finite resistance and the passage of current, a potential variation occurs along the channel length (i.e., if the source is grounded, the channel potential near the source is close to ground, whereas near the drain it is close to V_{DS}). This causes the inversion-layer (mobile) charge density to decrease with position toward the drain because the gate is assumed to be an equipotential region (i.e., the voltage along the gate is invariant with position y along the channel), making the gate-to-channel voltage decrease with position toward the drain. Since I_D must remain constant at all points in the channel under dc biasing conditions, the electric field parallel to the flow of I_D ($\mathcal{E}_y$) must increase toward the drain to offset the decrease in mobile charge density. Thus, a gradient in the $\mathcal{E}_y$ value must exist in the channel (i.e., $\partial\mathcal{E}_y/\partial y \neq 0$). The *gradual-channel approximation* (GCA), however, assumes that this gradient $\partial\mathcal{E}_y/\partial y$ is much smaller than gradient of the $\mathcal{E}$-field in the x-direction, $\partial\mathcal{E}_x/\partial x$ (i.e., $\partial\mathcal{E}_x/\partial x >> \partial\mathcal{E}_y/\partial y$).

The two-dimensional form of the Poisson equation with $\partial^2\varphi/\partial x^2$ and $\partial^2\varphi/\partial y^2$ replaced respectively by $-\partial\mathcal{E}(x,y)/\partial x$ and $-\partial\mathcal{E}(x,y)/\partial y$) can be written as:

$$\frac{\partial\mathcal{E}(x,y)}{\partial x} + \frac{\partial\mathcal{E}(x,y)}{\partial y} = \frac{\rho(x,y)}{\kappa_{si}\,\varepsilon_o} \qquad (4 - 13)$$

If the GCA approximation is valid, the $\partial\mathcal{E}(x,y)/\partial y$ term in Eq. 4-13 can be neglected. Then, Eq. 4-13 can be approximated as a one dimensional differential equation with the form:

$$\frac{d\mathcal{E}_x(x)}{dx} \approx \frac{\rho(x,y)}{\kappa_{si}\,\varepsilon_o} \qquad (4 - 13a)$$

The GCA is generally valid when the channel of the MOSFET is long enough (e.g., L ~2 μm or longer), in which case the electric field strength along the y-direction $\mathcal{E}_y$ changes gradually from a small value near the source to values of the order of V_{DS}/L near the drain. Thus, neglecting the $\partial\mathcal{E}(x,y)/\partial y$ term in such devices is appropriate, and the one-dimensional form of the Poisson equation is sufficiently accurate for calculating Q_I used in the expressions for I_D. However, at large drain biases, the GCA always becomes invalid (even in long-channel MOSFETs), at least in a certain section of the channel near the drain where the gradient of the longitudinal field $\mathcal{E}_y$ becomes comparable to or larger than the gradient of the transverse field $\mathcal{E}_x$. In such cases, as will be described later, the drain current then becomes nearly independent of drain bias, and the MOSFET is said to be in *saturation*. This implies that under conditions of bias that cause saturation, the I_D behavior of long-channel MOSFETs can no longer be adequately modeled by invoking the GCA along the entire length of the channel. Instead, modifications to the long-channel models must be made (as described in chap. 5).

The importance of being able to invoke the gradual channel approximation should be emphasized, as this is the basis of the derivation of the long-channel MOSFET I_D-V_{DS} models. That is, by invoking the GCA, the problem of having to treat the semiconductor equations in their two-dimensional form to calculate I_D in a MOSFET is

avoided. Instead, the task is broken down into two one-dimensional analyses, each independent of the other. As a result, I_D is determined by the superposition of two one-dimensional solutions. Namely, by assuming that the *gradient* of the "horizontal" field in the channel is negligible, the one-dimensional *Poisson equation* can be solved to get the total charge in the silicon at each point y along the channel (i.e., the sum of the inversion layer charge Q_I [see Fig. 3-8] and the ionic charge Q_B in the depletion region). This means that the same formulas derived for the MOS-C are also applicable for calculating the local values of $Q_I(y)$ and $Q_B(y)$ in the long-channel MOSFET (provided, of course, that the effect of the channel voltage caused by the drain-source bias is also included in the MOS-C formulas). Then we note that within the conducting channel of a MOSFET the current flows almost exclusively in the y-direction. Thus, we can ignore the fact that there must be a slight x-component in the "horizontal" $\mathcal{E}$-field. This allows us to solve the one-dimensional *drift-diffusion current-density equation* in the y direction for I_D as a function of the terminal voltages [using the local $Q(y)$ values obtained from the solution of the Poisson equation in this current-density equation].

Note, however, that if the GCA is not valid, but it is still invoked in order to enable one-dimensional analysis to be used for calculating Q_I, it is unlikely that accurate values of I_D will be yielded This is the situation that arises in so-called short-channel devices (as described in chap. 5). In such cases, to obtain good agreement between calculated and measured values of I_D, the two-dimensional forms of the semiconductor equations must be solved (usually numerically).

4.3.1.1 Verifying the Validity of the GCA as a Function of the Geometrical Factors of the MOSFET Device Structure and Bias Conditions. The validity of the GCA for a particular MOSFET structure can be checked by making rough estimates of $\partial\mathcal{E}_x/\partial x$ and $\partial\mathcal{E}_y/\partial y$ values under given bias conditions. Since we are most interested in I_D when the MOSFET is in strong inversion, we will establish expressions that allow the GCA to be tested if the MOSFET is being operated in strong inversion.*

We begin this derivation by recalling that under strong inversion the inversion layer charge density Q_I is much larger than the ionic charge density Q_B in the depletion region (see section 3.3.3), and that the inversion layer is very thin (~30-300Å). Thus, in strong inversion $\partial\mathcal{E}_x/\partial x$ in the silicon is greatest in the inversion layer region (see Fig. 4-6). Now $\partial\mathcal{E}_x/\partial x$ can be estimated by approximating the vertical electric field value at the bottom edge of the inversion layer as zero. Using this assumption implies that $\partial\mathcal{E}_x/\partial x \sim [\mathcal{E}_{si}(x=0) - 0]/t_{inv} = [\mathcal{E}_{si}(x=0)]/t_{inv}$, where t_{inv} is the effective inversion layer

* Note that in weak in version the surface potential $\varphi_{surf}(y)$ along the channel in long-channel MOSFETs is almost constant (see section 4.7). Thus, $\partial\mathcal{E}_y/\partial y$ is very small, implying that $\partial\mathcal{E}_x/\partial x >> \partial\mathcal{E}_y/\partial y$, because $\mathcal{E}_x$ drops from its value at $x = 0$ to zero at the bottom edge of the channel depletion region, which is a relatively small distance. Thus, in long-channel MOSFETs the GCA is also valid in weak inversion. However, in short-channel MOSFETs, this will no longer necessarily be the case, as we shall see in chap. 5.

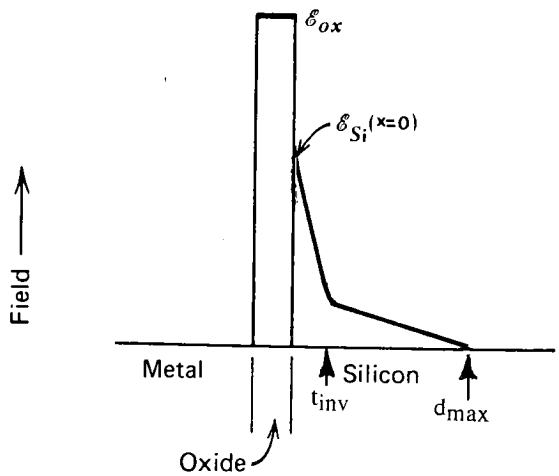


Fig. 4-6 The vertical electric field in a MOS-C (or MOSFET channel) when a large enough gate bias to cause inversion is applied. Note that since $Q_I > Q_B$ in inversion, that the gradient of the $\mathcal{E}$-field over the inversion layer thickness t_{inv} is larger than that over the depletion region width d_{max}, and we can therefore approximate the $\mathcal{E}$-field value at the bottom of the inversion layer thickness as being equal to zero for the purposes of estimating t_{inv}.

thickness. From Eq. 3-15, $\mathcal{E}_{si}(x=0) = \varepsilon_{ox}\mathcal{E}_{ox}/\kappa_{si}\varepsilon_o$, and from Eq. 3-14, $\mathcal{E}_{ox}$ is $\sim(V_{GS} - V_T)/t_{ox}$. Thus,

$$\frac{\partial \mathcal{E}_x}{\partial x} \approx \left(\frac{\varepsilon_{ox}}{\kappa_{si}\varepsilon_o t_{inv}}\right) \frac{(V_{GS} - V_T)}{t_{ox}} \qquad (4 - 14a)$$

Now, according to Eq. 3-12a, the concentration of electrons at some point in the semiconductor at equilibrium is exponentially dependent on the potential, with the exponential constant being $1/\varphi_t = kT/q$. This implies that the majority of the charges is contained within a distance from the surface over which φ drops by kT/q (i.e., since Eq. 3-12a predicts that if the electron concentration at the surface [x = 0] is n_{surf}, then at the point where the potential drops below φ_{surf} by an amount kT/q, n_{surf} will have dropped to a value 1/e of its value at x = 0, or to $0.37n_{surf}$). Thus, we assume that t_{inv} is roughly equal to that distance below x = 0. If we also assume that the vertical electric field $\mathcal{E}_x$ in the inversion layer is roughly the potential difference across the inversion layer $\Delta\varphi_{inv}$ divided by its thickness t_{inv} ($\mathcal{E}_x \sim \Delta\varphi_{inv}/t_{inv}$) we get $t_{inv} \sim (kT/q)/\mathcal{E}_{si}(x=0)$. Thus, substituting this expression for t_{inv} into Eq. 4-14a we finally obtain the following expression for roughly estimating $\partial\mathcal{E}_x/\partial x$:

$$\frac{\partial \mathcal{E}_x}{\partial x} \approx \left(\frac{\varepsilon_{ox}}{\kappa_{si}\varepsilon_o}\right)^2 \frac{q\,(V_{GS} - V_T)^2}{kT\, t_{ox}^2} \qquad (4 - 14b)$$

If we also assume that the value of $\mathcal{E}_y$ varies gradually from a small value near the source to values of the order of V_{DS}/L near the drain, $\partial\mathcal{E}_y/\partial y$ can be roughly estimated as $\partial\mathcal{E}_y/\partial y \sim V_{DS}/L^2$. Thus, if we assume that the GCA is valid when the ratio $(\partial\mathcal{E}_x/\partial x)/\partial\mathcal{E}_y/\partial y$ is large, then it is also valid if the following inequality is large

$$\frac{\partial\mathcal{E}_x/\partial x}{\partial\mathcal{E}_y/\partial y} = \left(\frac{\varepsilon_{ox}\,L}{\kappa_{si}\varepsilon_o}\right)^2 \frac{q\,(V_{GS} - V_T)^2}{kT\,t_{ox}^2\,V_{DS}} >> 1 \qquad (4 - 15)$$

For a MOSFET at 300K with $L = 1.0\ \mu m$, $t_{ox} = 300Å$, $V_{GS} - V_T = 0.5V$, and $V_{DS} = 0.5V$, the left-hand side of the inequality (4-15) is ~2300, indicating that the GCA is a very good approximation for such a MOSFET. This also implies that the GCA can be valid even in submicron MOSFETs, provided that $V_{GS} - V_T$ is not too small.

However, as we will see in chap. 5, $\mathcal{E}_y$ is quite non-uniform along the length of the channel at high values of V_{DS}, becoming much stronger near the drain. Furthermore, from Eq. 4-10 we note that the inversion layer charge density can be approximated as $Q_I \sim [V_{GS} - V_T - V(y)]$, if we replace V_{GB} with V_{GS}, and V_{CB} with $V(y)$ [where $V(y)$ is the channel voltage induced by V_{DS}]. However, this approximation is only valid when Q_I is positive, i.e., when $V_{GS} - V_T > V(y)$. Thus, if V_{DS} exceeds $V_{GS} - V_T$ (i.e., the condition which occurs in the region of the channel near the drain when the MOSFET is operated in saturation), the GCA becomes invalid in this region of the channel. [Because, when $V(y)$ is greater than $V_{GS} - V_T$, this implies that $\mathcal{E}_y > \mathcal{E}_x$, and that $\partial\mathcal{E}_y/\partial y$ is therefore at least comparable to, if not greater than, $\partial\mathcal{E}_x/\partial x$].

In summary, the GCA appears to be valid for MOSFETs operated in strong inversion even if the channel length is shorter than 1 μm - provided that V_{DS} is small enough that the MOSFET is not being operated in saturation (i.e., $V_{DS} \leq V_{DSsat}$, where V_{DSsat} is the voltage at the onset of saturation). On the other hand, if the MOSFET is operated in saturation, use of the GCA to calculate I_D will not yield accurate results, and alternative approximations (as described in chap. 5) must be employed to get good agreement with experimental I_D-V_{DS} characteristics.

4.3.2 Circuit Characteristics of Long-Channel MOSFETs

From the perspective of the MOSFET as a circuit component, long-channel behavior has been specified in a number of ways. A list of device parameters from a circuit perspective deemed to be characteristic of long-channel MOSFETs include the following:

1. The threshold voltage V_T is independent of channel length L.
2. The threshold voltage V_T is independent of drain bias voltage, V_{DS}.
3. The drain current in saturation I_{Dsat} is independent of V_{DS}.
4. The drain current has a linear dependence on $1/L$ (i.e., one quantitative criterion is that there should be no more than a 10% departure from linearity as the channel length is decreased).

5. The subthreshold current I_{Dst} is independent of drain bias.
6. The subthreshold swing S_t is independent of gate length.

As the behavior of the MOSFET is described later in this chapter, the terms associated with these device characteristics will be defined and discussed in more detail.

4.4 QUALITATIVE DESCRIPTION OF MOSFET OPERATION

The operation of the MOSFET is first described from a qualitative perspective. Specifically, our goal is to qualitatively explain the so-called *static characteristics* of the MOSFET; i.e., the drain current characteristics of the MOSFET as a function of the applied dc terminal voltages V_{GS} and V_{DS}. This behavior is normally depicted using an I_D-V_{DS} plot, with V_{GS} as the parameter.

For simplicity we consider an ideal (i.e., $V_{FB} = 0$) long-channel NMOSFET (*p*-type substrate) as our example device structure, with the source being chosen as the reference terminal (Fig. 4-4b). We also assume that the source-substrate and drain-substrate junctions are always kept under reverse bias and that the resulting reverse-bias leakage currents are negligible (this is valid as long as the bias voltage is kept smaller than the junction breakdown voltage), and that $V_{SB} = 0V$. We also use energy-band versus position and potential versus position diagrams to help depict the qualitative MOSFET behavior. Note that in the case of the MOS-C one-dimensional diagrams were adequate. Here, however, two-dimensional energy-band and potential diagrams must be used. Axes and terminal voltage designations are the same as in Fig. 4-2a.

4.4.1 MOSFET Behavior in Response to Gate Bias Alone

We begin by describing how a MOSFET behaves in response to a gate bias V_{GS} alone. In this case, the region under the oxide responds in the same way as does an ideal MOS-C. However, because the source and drain regions are present in a MOSFET the overall device behavior differs from that of a MOS-C with a V_{GB} applied.

To analyze the MOSFET behavior, let us assume that all three bias voltages are initially set to zero (i.e., $V_{GS} = V_{DS} = V_{BS} = 0V$). In this case a flat band condition in exists in the gated region of the MOSFET (as is also observed when $V_{GB} = 0V$ in an ideal MOS-C). The two-dimensional energy-band diagram corresponding to this condition is shown in Fig. 4-7b. We observe that E_c in the *x*-direction at the cut A-A is indeed flat. (The potential versus position diagram in the MOSFET for this bias condition is also shown in Fig. 4-8b). Assuming that $\varphi(x = \infty) = 0$, in this case $\varphi_{surf} = 0$ as well.)

At the surface ($x = 0$), however, a potential energy barrier of value qV_{bi} exists between the *p* region and each of the n^+ regions (where V_{bi} is the built-in potential of the *doped* pn^+ junction). This is shown in the E_c versus *y* direction plot for $x = 0$ in Fig. 4-7b. For this case (and for cases of when a reverse bias exists across the doped junctions) only the very small leakage current of a reverse-biased *pn* junction flows.

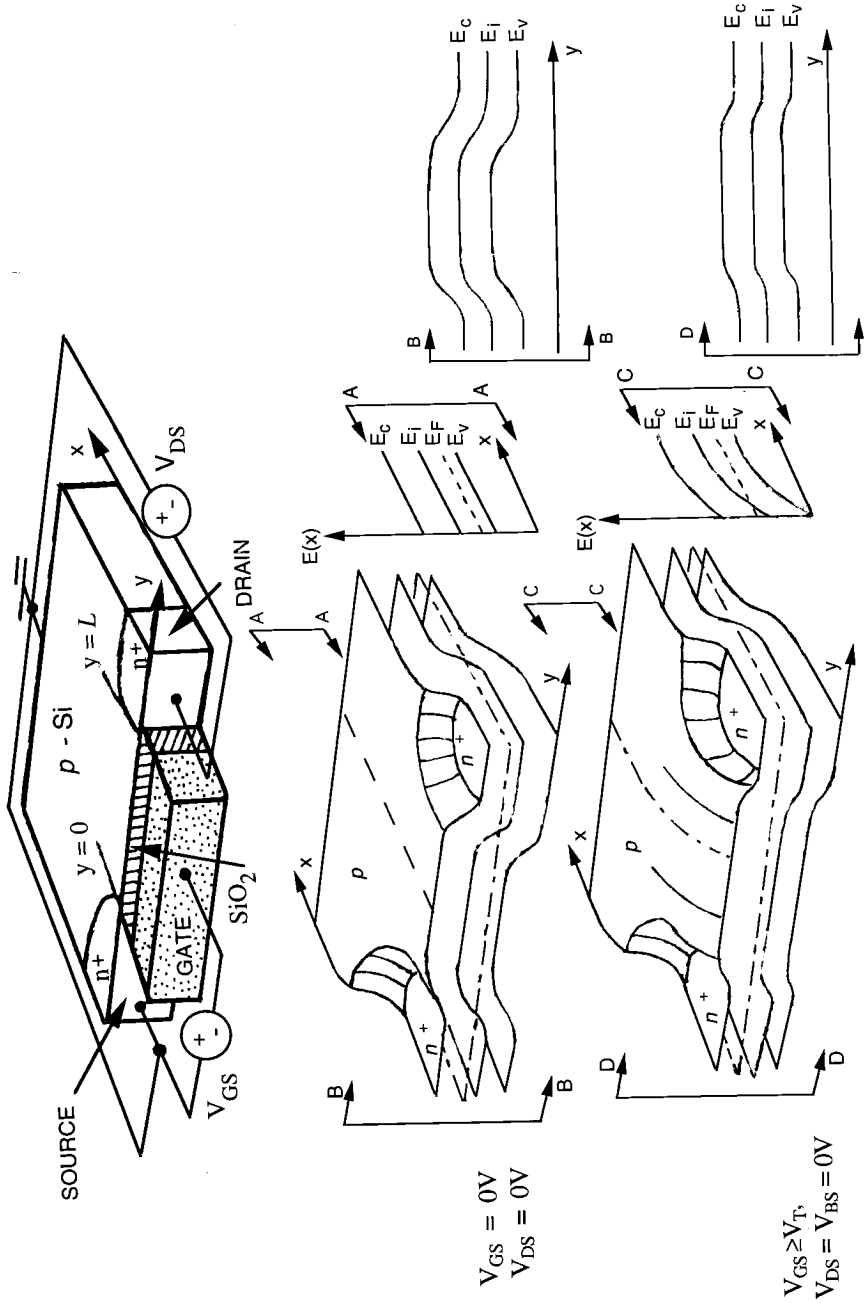


Fig. 4-7 (a) Cross section of an NMOS device. (b & c) 2-D perspective plots of the energy bands versus position of an ideal long-channel NMOS device when (b) V_{GS} = 0V (flat-band condition), and V_{DS} = 0V; and (c) $V_{GS} > V_T$, $V_{DS} = V_{BS}$ = 0V.

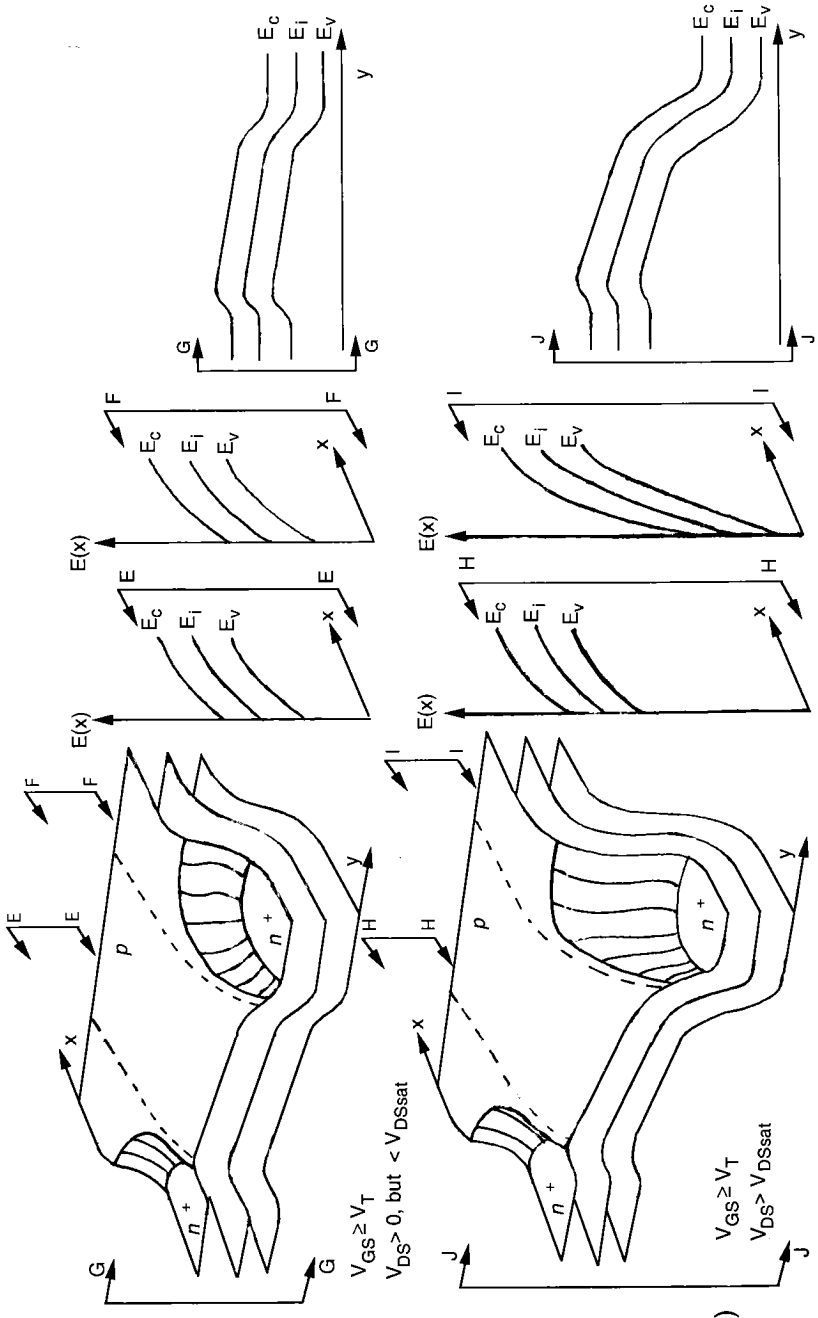


Fig. 4-8 (cont.) (d & e) 2-D perspective plots of the energy bands as a function of position an ideal long-channel NMOS device when: (d) $V_{GS} \geq V_T$ and V_{DS} is small (i.e., $< V_{DSsat}$); and (e) $V_{GS} \geq V_T$ and $V_{DS} > V_{DSsat}$. $V_{BS} = 0$ in both cases.

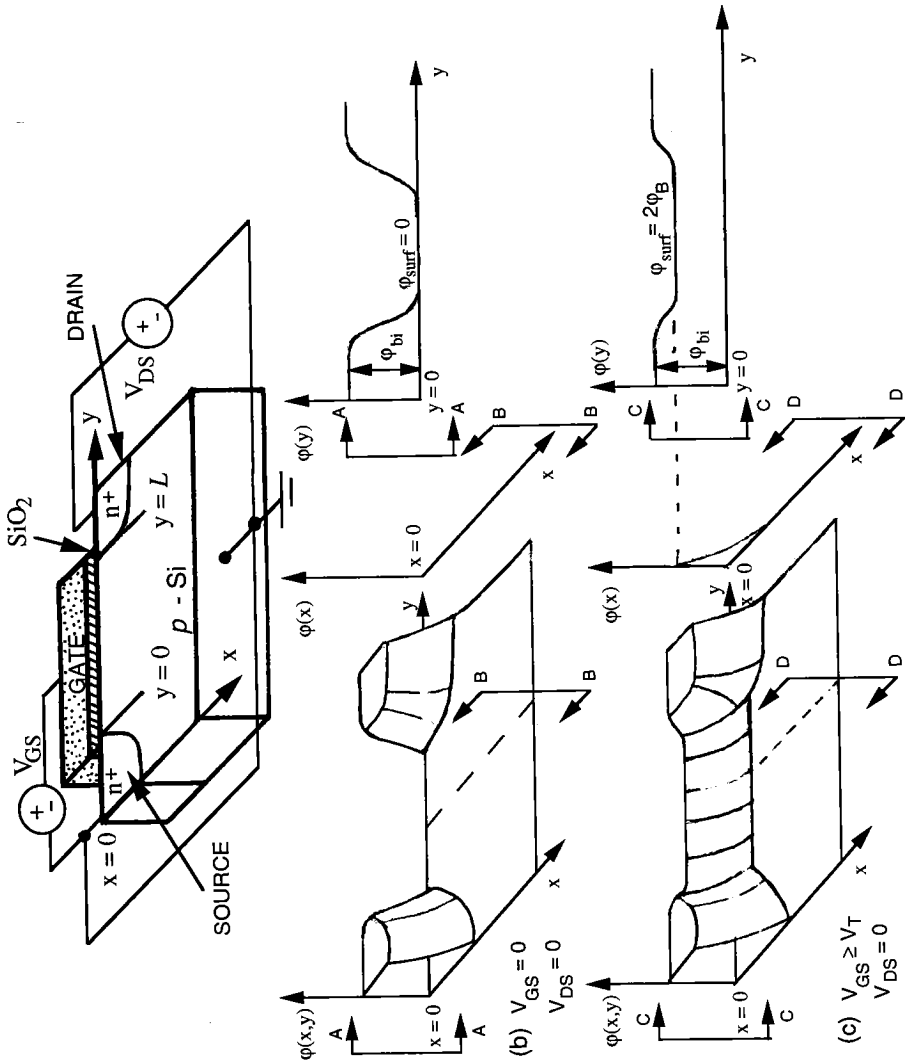


Fig. 4-8 (a) Cross section of an NMOS device. (b & c) 2-D perspective potential versus distance diagrams of an ideal NMOS device when: (b) $V_{GS} = 0V$ (flat-band condition), and $V_{DS} = 0V$; and (c) $V_{GS} \geq V_T$, $V_{DS} = V_{BS} = 0V$.

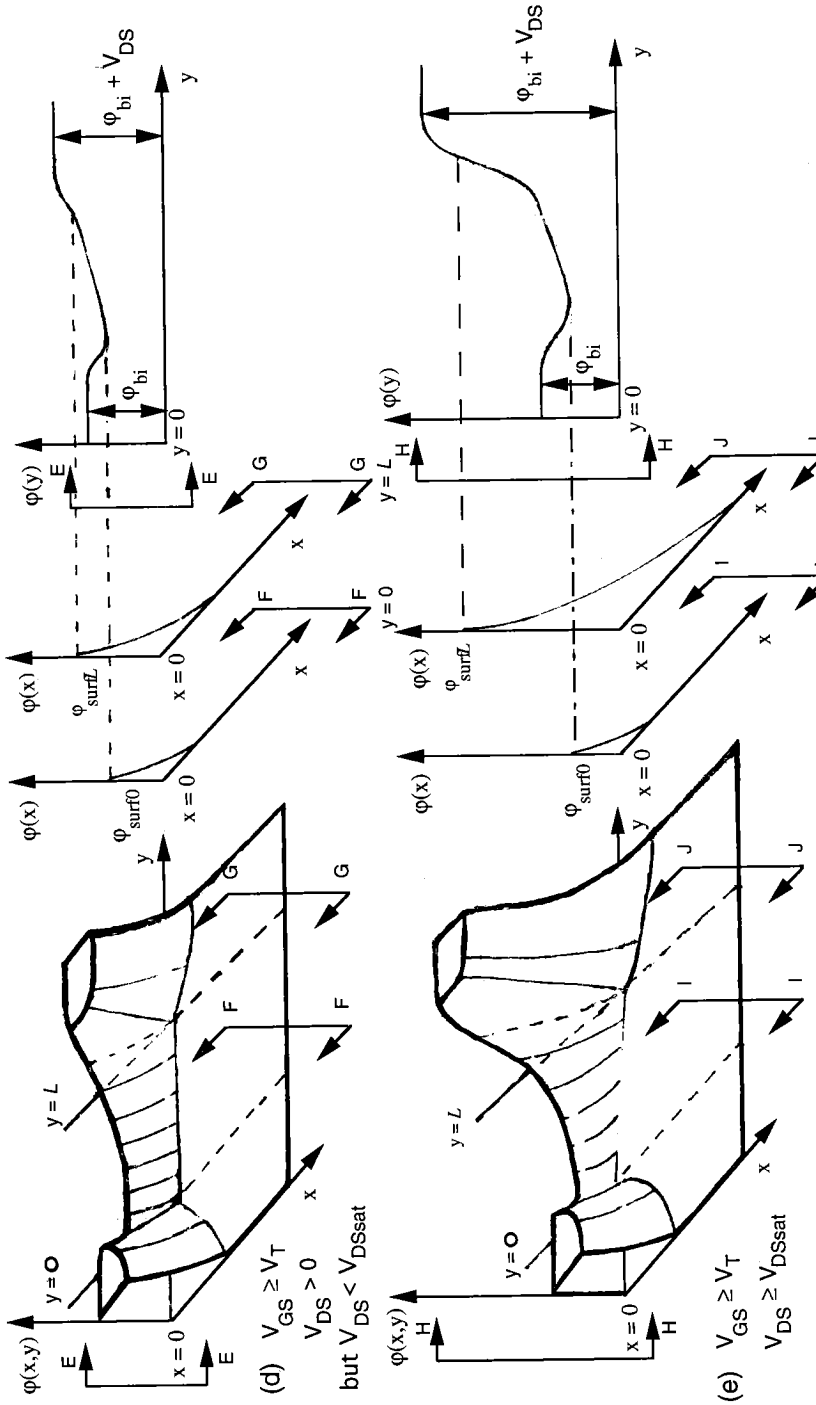


Fig. 4-8 (cont.) (d & e) 2-D perspective potential versus distance diagrams of an ideal long-channel NMOS device when (d) $V_{GS} \geq V_T$ and $V_{DS} > 0V$, but $< V_T$; and (e) $V_{GS} > V_T$, and $V_{DS} > V_{DSsat} \sim V_{GS} - V_T$.

Thus, the circuit path between the n^+ source and drain regions effectively appears as an open circuit. This means that even if a voltage is applied between the source and drain ($V_{DS} \neq 0V$), essentially no drain current will flow. Moreover, if we apply either a negative V_{GS} (or a positive V_{GS} such that $V_{GS} < V_T$,) the gated region contains either an excess of holes (*accumulation*, when $V_{GS} < 0V$) or a deficit of holes and very few electrons (*depletion*, when $V_{GS} < V_T$). In this case, the circuit path between source and drain still appears essentially as an open circuit. That is, no current flows in the drain terminal even if a drain voltage is applied. For all of these conditions, the MOSFET is said to be in an "OFF" state.

If a positive V_{GS} is applied such that $V_{GS} > V_T$, an inversion layer containing mobile electrons is formed adjacent to the Si surface throughout the entire gated region between the source and drain (here we still assume that $V_{DS} = V_{BS} = 0V$). Thus, the circuit path between source and drain now contains an induced "*n*-type" region (the inversion layer), as shown in Fig. 4-9. This means that a conducting path or *channel* exists between the source and drain. (However, if $V_{DS} = 0V$, no drain current will yet flow.) Now, if a drain voltage is applied (i.e., $V_{DS} > 0V$), drain current will flow. The device in this state is said to be "ON." Obviously, the larger the value of gate voltage (V_{GS}), the more electrons pile up at the Si surface, and the greater the conductance of this channel. This demonstrates that the input voltage to the MOSFET can induce a source-to-drain channel and will determine the maximum conductance of the channel.*

Figure 4-7c shows the energy band diagram of the MOSFET when $V_{GS} > V_T$ and $V_{DS} = V_{BS} = 0V$, and Fig 4-8c also shows the potential diagram for this case. Note that the bands along the *x*-direction in the gated region are no longer flat (see section B-B) and that $\varphi_{surf} > 2\varphi_B$ everywhere along the channel. The potential-energy barrier between the n^+ and *p* regions is also reduced (see the E_C versus *y* plot in Fig. 4-7c). Hence, only a very low barrier to current injection exists between the source and the channel. The barrier height is smallest at the surface with its value controlled by the bias on the gate. At strong inversion and beyond, the barrier height is assumed to be remain constant at $\sim q(V_{bi} - 2\varphi_B)$, because φ_{surf} is assumed to be clamped at $2\varphi_B$ in strong inversion. However, as V_{GS} is increased, the inversion charge density continues to increase, increasing the conductance of the channel

4.4.2 Qualitative MOSFET Behavior when both Gate and Drain Bias are Applied

We now consider the variation of I_D with V_{DS}. At first we assume that V_{GS} is kept fixed at a single value while V_{DS} is varied, and $V_{BS} = 0V$. Later we show how varying the value of V_{GS} impacts the I_D-V_{DS} characteristic. In this section we also assume that

* Note that when V_{GS} is positive but smaller than V_T, the gated region can be in state of weak inversion. In weak inversion only small, so-called *subthreshold currents* are able to flow between the drain and source. These subthreshold currents are not large enough to be useful as driving currents for most applications, but they are sufficiently large to cause troublesome leakage - which can complicate the design of some circuits. We will describe subthreshold currents in more detail in section 4.7 and in chap. 5, as part of discussion on the quantitative behavior of the long-channel MOSFET and the submicron MOSFET.

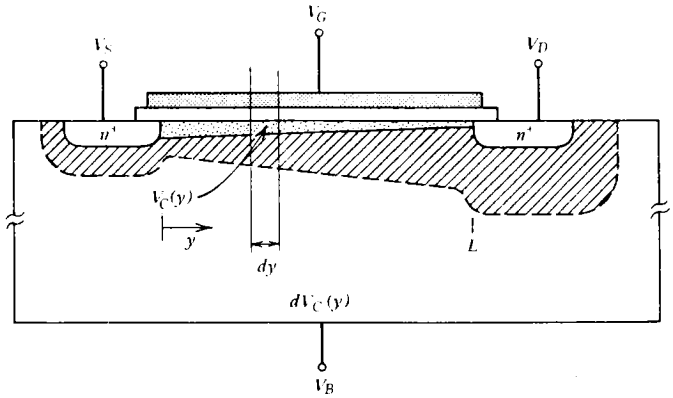


Fig. 4-9 Depletion region width variation from source to drain end of a MOSFET when V_{DS}>0V.

sufficient gate bias is applied to ensure that the MOSFET is being operated in inversion (i.e., $V_{GS} > V_T$). We then describe what happens as a dc V_{DS} is applied, with its value being increased in small steps, starting from V_{DS} = 0V.

As noted earlier, when V_{DS} = 0V, I_D = 0, even if $V_{GS} > V_T$ (see Fig. 4-10b). However, when V_{DS} is increased slightly, I_D starts to flow. When V_{DS} is small (i.e., a few tenths of a volt or less), the surface channel behaves like a simple resistor (Fig. 4-10c). That is, an I_D proportional to V_{DS} flows into the drain. If a plot of I_D versus V_{DS} is used to demonstrate the I_D-V_{DS} behavior, the response to small values of V_{DS} is shown as the line from the point 0 to the point A in Fig. 4-10a.

When V_{DS} is increased beyond a few tenths of a volt, the MOSFET exhibits a new phase of behavior. That is, current flowing in the channel gives rise to a significant voltage drop in the channel (unlike the case in the three-terminal structure introduced in section 4.2.1, where the channel voltage is constant along the entire channel length because essentially no current flows along it). Specifically, the *channel voltage* (introduced in section 4.2.1) is V_{DB}, at the drain end and V_{SB} at the source end. If V_{SB} = 0V, the channel voltage at the source end will equal zero. But even with V_{SB} = 0V, the channel voltage at the drain end will not be zero, since we assume V_{DS} >0V. As shown in Fig. 4-9 such a non-zero V_{DB} causes the depletion region under the gate to widen toward the drain end of the channel. As described in section 4.2.1, this implies that the mobile charge concentration in the inversion layer simultaneously decreases. The smaller number of carriers causes the channel conductance to decrease, which in turn is manifested as a smaller slope in the I_D-V_{DS} characteristic as V_{DS} is increased. As the drain voltage is further increased, the depletion region continues to widen and the slope of the I-V characteristic continues to decrease (see Figs. 4-10a and 4-10c).

The energy-band and potential diagrams of the MOSFET when both a small V_{DS} and a $V_{GS} > V_T$ are applied are shown in Figs. 4-7d and 4-8d, respectively. Even if the drain voltage is small enough so that the entire surface region remains in inversion, we see that near the drain the band bending is larger than at the source (see the E_C vs. y diagram in Fig. 4-7d). This implies that the depletion region under the gate is wider there. Since the gate-to-body voltage is constant across the entire gated region, less voltage is

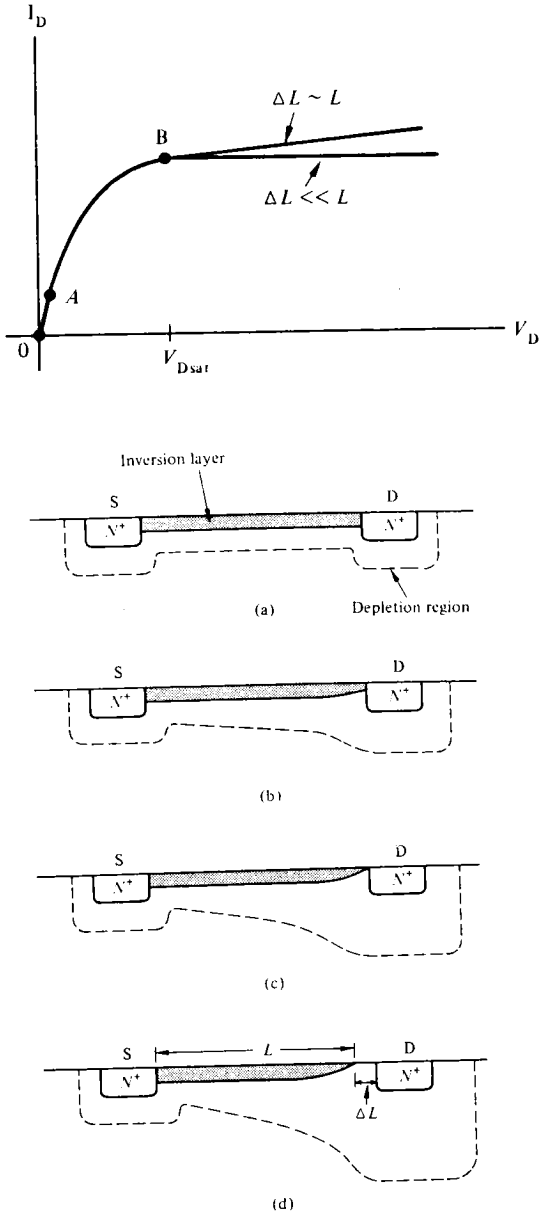


Fig. 4-10 (a) I_D-V_{DS} characteristic of a MOSFET being operated at a fixed V_{GS}, with $V_{GS}>V_T$. (b -e) Visualization of the various phases of $V_{GS}>V_T$ MOSFET operation: (b) V_{DS} = 0V; (c) channel (inversion layer) narrowing under moderate V_{DS} biasing; (d) pinchoff; and (e) post pinchoff ($V_{DS}>V_{DSsat}$).[30] Reprinted with permission of Addison-Wesley Pub. Co.

available near the drain to establish the inversion layer charge. Hence, inversion is less strong in the channel near the drain.

Eventually, if V_{DS} is sufficiently increased such that $V_{DS}>(V_{GS} - V_T)$, the inversion layer ceases to exist near the vicinity of the drain. The disappearance of the inversion layer due to an increase in V_{DS} is termed *pinch-off* (see Fig. 4-10d). The pinch-off point on the I_D-V_{DS} characteristic is shown as point B in Fig. 4-10a. We see that the slope of the I_D curve here has become approximately zero. Note, however, that pinch-off does *not* mean that I_D vanishes as V_{DS} exceeds the pinch-off voltage. In fact, it is the *gradual channel approximation* that is no longer valid in the region of the device in which pinch-off exists. That is, mobile charges still move through the depleted region near the drain, and drain current continues to flow. However, the equations for the charge distribution derived from the one-dimensional solution of the Poisson equation (which are valid for the case of the MOS-C) predict that no mobile charge exists in the pinch-off region. But in the depleted pinched-off region the one-dimensional Poisson equation is no longer valid. Instead, a two-dimensional numerical analysis (or an analytical expression based on a modified 1-D form of the Poisson equation - see section 5.4.3.3) must be used to calculate the potential and carrier concentrations in the pinched-off device region. A detailed numerical analysis of the saturation range has been carried out using the two-dimensional device simulator MINIMOS.[3] This work showed that for voltages greater than V_{DSsat} the current stream of the carriers moving into the pinched off region is no longer confined to the inversion layer region near the surface, but begins to move away from the surface and into the bulk. Note that this physical picture is plausible because once V_{DS} becomes larger than $(V_{GS} - V_T)$ the polarity of the vertical electric field in the pinched-off region is reversed. This tends to push any mobile carriers flowing in the channel *away* from the surface. Thus, in the region of the channel where $V_{DS}>(V_{GS} - V_T)$ the mobile carriers flow toward the drain *beneath* the surface.

Nevertheless, as the drain voltage increases beyond the pinch-off voltage V_{DSsat} the pinched-off section of the channel widens from just a point into a segment ΔL in length (see Fig. 4-10e). Most of the voltage drop in excess of V_{DSsat} is dropped across this pinched-off section ΔL. I_D remains approximately constant (*saturated*) for drain voltages greater then V_{DSsat} so long as $\Delta L<<L$. If ΔL becomes comparable to L (which is a condition that may occur in short-channel MOSFETs), the same voltage drop appears across a shorter channel $(L - \Delta L)$, and the post-pinch-off current I_{Dsat} in such devices will increase somewhat as V_{DS} is increased further beyond V_{DSsat}. Figures 4-7e and 4-8e show the energy-band diagrams and potential diagrams, respectively, for a long-channel MOSFET biased into pinch-off.

The response of the long-channel MOSFET to separate manipulation of gate and drain dc biases has been qualitatively described up to this point. This has allowed us to visualize the general variation of I_D with V_{DS} for a fixed value of $V_{GS}>V_T$. To allow the complete set of I_D versus V_{DS} characteristics to be obtained, we need to combine the results from the individual cases. These are as follows:

For $V_{GS}<V_T$, no channel exists and $I_D \approx 0$ for all drain biases below junction breakdown. For each value of $V_{GS}>V_T$, a characteristic of the form shown in

Fig. 4-10a will be observed. Since the conductance of the channel will increase as V_{GS} is made larger, the initial slope of I_D will also become steeper with increasing V_{GS}. Furthermore, since the inversion layer contains more electrons, a larger value of V_{DS} is needed to produce pinch-off (V_{DSsat} increases for increasing V_{GS}.). Based on these phenomena, the static I_D-V_{DS} characteristics assume the form shown in Fig. 4-11.

Let us next consider quantitative models which allow these static (i.e., dc) I_D-V_{DS} characteristics to be accurately simulated.

4.5 QUANTITATIVE BEHAVIOR OF MOSFETS: DC CIRCUIT MODELS OF LONG-CHANNEL MOSFETS

Although the discussion in the previous section provides a qualitative picture of how a MOSFET operates, we need to express this behavior in quantitative form. Specifically, to design circuits with MOSFETs we must be able to calculate how much current is carried in the "ON" state of a MOSFET and how much leakage current flows in the "OFF" state. In short, we need the current-voltage characteristics of the MOSFET, specifically the quantitative relationship between the drain current and terminal voltages.

In this section our goal is to derive relationships that allow the dc drain current I_D of long-channel MOSFETs to be quantitatively predicted. Such I_D-V_{DS} relationships are referred to as the *static (dc) circuit models* of the MOSFET. As is the case for the MOS-C, there exists a hierarchy of such long-channel MOSFET I_D-V_{DS} models in terms of their generality, accuracy, and complexity. Our discussion in this section will derive four of these models, using as our foundation two of the five classical semiconductor equations introduced in chap. 2: the one-dimensional Poisson equation, and the one-dimensional drift-diffusion current density equation (as well as the three models of the MOS-C derived in chap. 3).

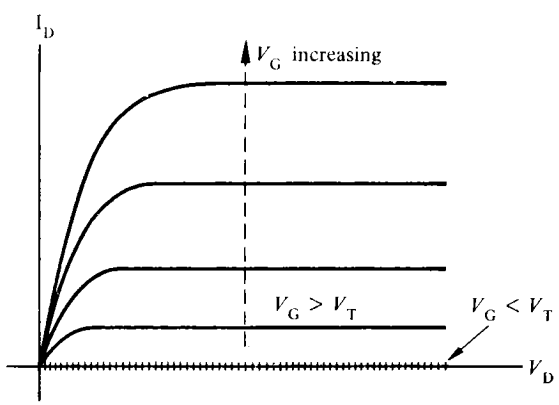


Fig. 4-11 General form of the I_D-V_{DS} characteristics expected from a long-channel MOSFET.

The first of these models is the *Pao-Sah* model, which is the most general and accurate of our four dc I_D long-channel MOSFET models. Like the general MOS-C model, it also the most computationally complex (in that it requires numerical analysis techniques to calculate I_D for a given set of terminal voltages). The three less-computationally complex long-channel models are derived from the Pao-Sah model by making suitable simplifying approximations. The first of these that we will study is the *charge-sheet MOSFET* model. It is based on the same assumptions used in the MOS-C charge-sheet model introduced in chap. 3. Unlike the Pao-Sah model, it provides an analytical expression for I_D in terms of V_{GB} and φ_{surf}, and thus involves much less computational effort. Nevertheless, it is only slightly less accurate than the Pao-Sah model provided the MOSFET is operated in inversion (i.e., this model is valid from weak through strong inversion). However, while this model provides an analytical expression for I_D, it still requires a numerical procedure to obtain the values of the surface potential needed to evaluate the I_D expression. That is, I_D is not explicitly stated in terms of the terminal voltages. (The terminal voltage values are used to first calculate φ_{surf}, which is then used to compute I_D.) Thus, the charge-sheet model is still too cumbersome to serve as a MOSFET model for most circuit-simulation applications.

By making suitable approximations that are valid over a more limited range of operation (namely, the strong inversion regime), two simpler long-channel MOSFET models have been derived that are extensively used in circuit analysis and design. They are the *variable depletion-charge* and the *square-law* models. These provide analytical expressions which give I_D explicitly in terms of the terminal voltages. We show how they can be derived from the charge-sheet model by making additional simplifying assumptions. At the end of the section we compare these four models in more detail.

4.5.1 General Approach to Deriving the DC Drain-Current Relationships in Long-Channel MOSFETs

The derivation of the dc drain current relationship recognizes that, in general, the current in the channel of a MOSFET can be caused by both drift and diffusion current. In an NMOSFET we can accurately assume that only electrons contribute to the drain current. Thus, we only need to use Eq. 1-5a to describe the drain current density J_D

$$J_D = J_n = q\ \mu_n\, n\, \mathcal{E} + q\, D_n\, \nabla n \qquad (4 - 16)$$

Within the conducting channel of a MOSFET the current flows almost exclusively in the y-direction, thus Eq. 4-14 can be simplified to the following one-dimensional form:

$$J_D \approx J_{ny} = q\, \mu_n\, n\, \mathcal{E}_y + q\, D_n\, dn/dy = -q\, \mu_n\, n\, d\varphi/dy + q\, D_n\, dn/dy \qquad (4 - 16a)$$

Since we also assume there are no current sources or sinks inside the device, the current flowing through any cross-sectional plane within the channel must be constant, having some value I_D. We also assume that the current density is constant in the z-direction, but not necessarily in the x-direction. Hence, if J_{ny} is integrated over Z we get

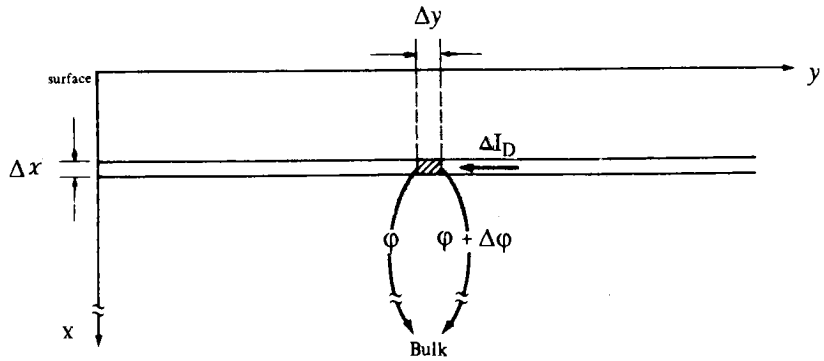


Fig. 4-12 The drain current I_D in a MOSFET can be modeled as the sum of currents in slices ΔI_D each having a thickness Δx and a width Z. These slices exist from the surface to the depth at which the electron concentration becomes negligible.[5] Reprinted with permission of McGraw-Hill Book Co.

$$I_D = -\int\int J_{ny}\ dx\ dz = -Z\int_0^{x_c(y)} J_{ny}\ dx \qquad (4 - 17)$$

If the flow of electrons can be considered to be laminar, we can assume that the total current I_D is the sum of elemental currents ΔI, each flowing horizontally in an inversion layer slice parallel to the surface, having width Z and depth Δx, as shown in Fig. 4-12. If we allow Δx to become a differential, the current in each slice dI_D is

$$dI_D = dI_{Ddrift}\,(x,y) + dI_{Ddiff}\,(x,y) \qquad (4 - 18a)$$

Since $\mathcal{E}(y) = -d\varphi/dy$, the drift component of Eq. 4-18a can be written as

$$dI_{Ddrift}\,(x,\ y) = q\ \mu_n\ Z\ n\ (x,\ y)\ \partial\varphi(x,\ y)/\partial y \qquad (4 - 18b)$$

The diffusion component of Eq. 4-18b can also be rewritten as,

$$dI_{Ddiff}\,(x,\ y) = -Z\,q\,D_n\ \partial n(x,\ y)/\partial y = -Z\,\mu_n\ kT\ \ \partial n(x,\ y)/\partial y \qquad (4 - 18c)$$

since $D_n/\mu_n = kT/q$.

An expression for the total drain current for each of these components can be obtained by integrating each expression over the depth from $x = 0$ to a point below the surface at which the electron concentration is negligible, x_c or

$$I_{Ddrift} = q\,\mu\,Z\frac{d\varphi}{dy}\int_0^{x_c(y)} n\,(x,y)\,dx \qquad (4 - 19)$$

$$I_{Ddiff} = -\mu Z kT \frac{d}{dy}\left[\int_0^{x_c(y)} n(x,y)\,dx\right] \qquad (4-20)$$

Since the charge per unit area due to carriers in the inversion layer, Q_I is given by

$$Q_I(y) = -q\int_0^{x_c(y)} n(x,y)\,dx \qquad (4-21)$$

Thus, Eqs. 4-19 and 4-20 can be written as

$$I_{Ddrift} = \mu Z(-Q_I)\frac{d\varphi}{dy} \qquad (4-22)$$

and

$$I_{Ddiff} = \mu Z\frac{kT}{q}\frac{dQ_I}{dy} \qquad (4-23)$$

Our goal is to derive a model in which I_D is expressed in terms of the terminal voltages V_{GS} and V_{DS}. To do this we must integrate the sum of Eqs. 4-22 and 4-23. The integration is carried out by changing the variables of integration by multiplying both sides of Eqs. 4-22 and 4-23 by (dy). Since we recognize that I_D is constant along the length of the channel, when we integrate the left hand side of the sum of Eqs. 4-22 and 4-23 (namely I_D[dy] from 0 to L) we get

$$I_D L = \int_0^L I_D\,dy \qquad (4-24)$$

Then we can write the total drain current I_D as

$$I_D = \frac{\mu Z}{L}\int_{\varphi_{surf0}}^{\varphi_{surfL}} (-Q_I)\,d\varphi + \frac{\mu ZkT}{Lq}\int_{Q_{I,source}}^{Q_{I,drain}} dQ_I \qquad (4-25)$$

We see that the first term on the right hand side of Eq. 4-25 is no longer integrated with respect to position, but with respect to voltage. (In the *variable-depletion charge model* [section 4.5.4, Eq. 4-58], and the *square-law model* [section 4.5.5, Eq. 4-62], we assume that only the drift current term is needed to describe I_D. Hence, only this integral needs to be carried out to calculate I_D. By changing the variable of integration from position to voltage this integration becomes easy to perform, yielding Eqs. 4-58 and 4-62.) The limits of integration of the first term of Eq. 4-25 are the value of the surface potential at the source end φ_{surf0} (which is usually set equal to zero), and the surface potential value at the drain end φ_{surfL} (which is usually set to V_{DS}). The limits of the integration

of the second term are the value of the mobile charge density at the source end of the channel $Q_{Isource}$, and the value at the drain end Q_{Idrain}.

4.5.2 Pao-Sah Model

The *Pao-Sah* model,[4] published in 1966, was the first advanced long-channel MOSFET model to be developed. While it retained the gradual-channel approximation, it did not invoke the depletion approximation and permitted carrier transport in the channel by both drift and diffusion current. The formulation of the I_D equation it develops is therefore quite general, but as a result requires numerical integrations in two dimensions. Such numerical complexity has limited the role of this model to theoretical analyses of the MOSFET. Other simpler models of the MOSFET have been developed for use in IC circuit design applications.

The Pao-Sah model follows the approach outlined in the previous section for deriving the I_D relationship from the semiconductor device equations. However, the derivation is usually carried out using a more compact formulation, in which the drift and diffusion current expression are combined into a single equation. When we derive the charge-sheet MOSFET model we will return to calculating both current components separately, as this results in a simpler derivation. Here, the opposite is true, and so we also elect the single equation approach to deriving the Pao-Sah model. Note that the presentation of the derivation in this section, and the next, closely mimics the method proposed by Tsividis.[5]

We begin by noting that the Pao-Sah approach invokes the gradual channel approximation (like all long-channel MOSFET models), and assumes that μ is constant along the channel. In our derivation we also assume that the substrate doping is uniform and that an n-channel MOSFET is being treated.

To simplify the expression for I_D so that it consists of one instead of two terms on the right hand side of Eq. 4-16, we start by noting that at the source end of the channel in the MOSFET (where $y = 0$), n can be expressed in terms of φ by

$$n\,(x, 0) = n_{po} \exp q[\varphi\,(x, 0) - V_{SB}]/kT \qquad (4 - 26)$$

and at the drain end by

$$n\,(x, L) = n_{po} \exp q[\varphi\,(x, L) - V_{DB}]/kT \qquad (4 - 27)$$

For an arbitrary position y along the channel, n is given by

$$n\,(x, y) = n_{po} \exp q[\varphi\,(x, y) - V_{ChB}(y)]/kT \qquad (4 - 28)$$

where $V_{ChB}\,(0) = V_{SB}$ and $V_{ChB}\,(L) = V_{DB}$. In Eq. 4-28, V_{ChB} represents the potential corresponding to the difference between the quasi-Fermi level of the electrons in the inversion layer and the quasi-Fermi level of holes in the bulk, as shown in Fig. 4-13. That is, because the source and drain will in general be biased with respect to the substrate, the channel is also at some potential $V_{ChB}\,(y)$ relative to the substrate. Clearly this voltage is a function of position y along the channel if $V_{DS} \neq 0$, because at the source end we have $V_{ChB}\,(0) = V_{SB}$ and at the drain end $V_{ChB}\,(L) = V_{DB}$.

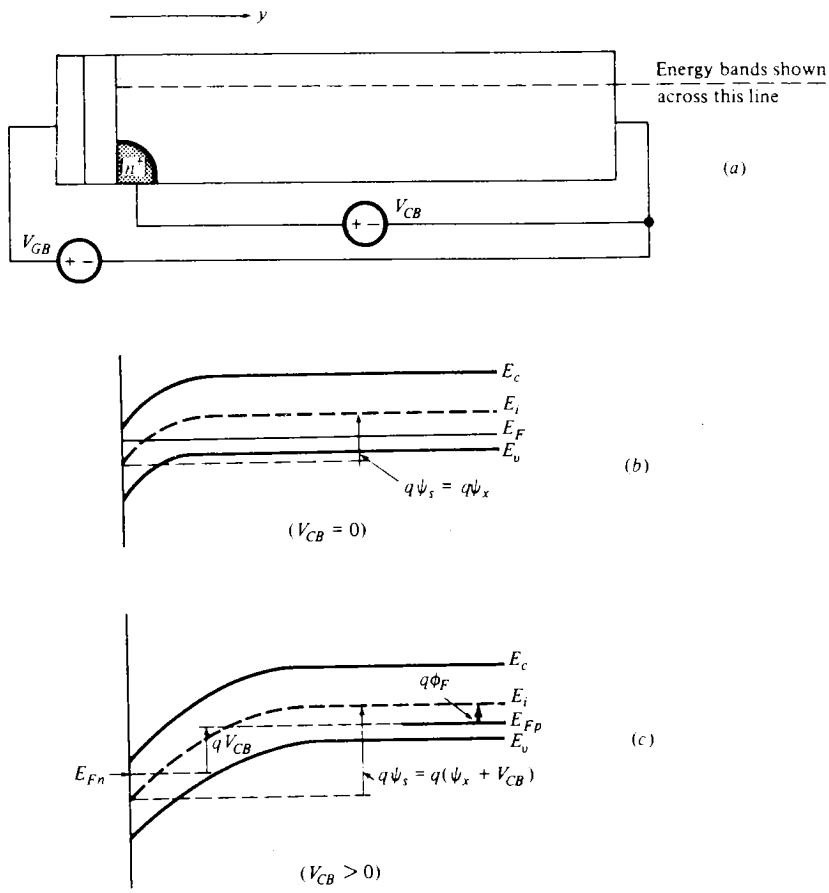


Fig. 4-13 Energy-band diagram of a three-terminal MOS structure with a bias V_{CB} (= V_{ChB} in a MOSFET) showing how the applied bias V_{CB} is related to the difference between the quasi-Fermi levels of the electrons in the inversion layer E_{Fn} and that of the holes E_{Fp}.[5] Reprinted with permission of McGraw-Hill Book Co.

If Eq. 4-28 is differentiated with respect to y we get

$$\frac{\partial n(x, y)}{\partial y} = \frac{q\, n(x, y)}{kT}\left[\frac{\partial \varphi(x, y)}{\partial y} - \frac{\partial V_{ChB}(y)}{\partial y}\right] \tag{4 - 29}$$

Using Eq. 4-29 in Eq. 4-18 we obtain

$$d\, I_{Ddiff}(x, y) = -(Z\, dx)\, q\, \mu\, n(x, y)\left[\frac{\partial \varphi(x, y)}{\partial y} - \frac{\partial V_{ChB}(y)}{\partial y}\right] \tag{4 - 30}$$

If Eq. 4-30 is inserted together with Eq. 4-17 into Eq. 4-16 it is seen that the $\partial\varphi/\partial x$ terms cancel out, leaving

$$d\,I_D\,(x, y) = (Z\,dx)\,q\;\mu\,n\,(x, y)\,\frac{dV_{ChB}(y)}{dy} \qquad (4 - 31)$$

As was done above for Eqs. 4-19 and 4-20, I_D from Eq. 4-31 can be obtained by integrating from the surface to a depth $x = x_c$ beneath which the electron concentration is negligibly small

$$I_D = \mu\,Z\,\frac{dV\,(y)}{dy}\,q\int_0^{x_c(y)} n\,(x, y)\,dx \qquad (4 - 32)$$

From Eq. 4-21 we can write Eq. 4-32 as

$$I_D = \mu\,Z\,(-Q_I)\,\frac{dV\,(y)}{dy} \qquad (4 - 33)$$

Now, integrating over the channel length and using Eq. 4-24, we get

$$I_D = \frac{\mu\,Z}{L}\int_{V_{SB}}^{V_{DB}} (-Q_I)\,dV \qquad (4 - 34)$$

We can now rewrite the expression for Q_I given by Eq. 4-21 as

$$Q_I = q\int_0^{x_c} n\,(x, y)\,dx = q\int_{\varphi_{surf}}^{\varphi_c} \frac{n\,(\varphi)\,d\varphi}{d\varphi/dx} \qquad (4 - 35)$$

If we define $\varphi_t = kT/q$ we can write $n\,(\varphi)$ in Eq. 4-35 as

$$n\,(\varphi) = N_A \exp\,[\varphi\,(x,y) - 2\varphi_B - V_{ChB}(y)]/\varphi_t \qquad (4 - 36)$$

because $p_{po} \approx N_A$, and $n_{po} = n_i^2/N_A$, and $\varphi_B = kT/q \ln\,(N_A/n_i)$. Thus,

$$(N_A/n_i)^2 = \exp\,(2\varphi_B)/\varphi_t \qquad (4 - 37)$$

Solving Eq. 4-37 for n_i^2 and substituting the result for n_i^2 into $n_{po} = n_i^2/N_A$, we get

$$n_{po} = N_A \exp\,(-2\varphi_B/\varphi_t). \qquad (4 - 38)$$

In addition, we recall that $\mathcal{E}(x) = -\,d\varphi/dx$. In chap. 3 we derived an expression for $\mathcal{E}(x)$ in a MOS-C in which both electrons and holes were allowed in the depletion region of the semiconductor. This expression, with φ being replaced by $\varphi \dashrightarrow \varphi + V_{ChB}$, is the expression we use for $d\varphi/dx$ in the denominator of Eq. 4-35. That is, $\mathcal{E}(x)$ is now

$$-\frac{d\varphi}{dx} = \mathcal{E}(\varphi) = \sqrt{\frac{2q\,N_A}{\kappa_{Si}\varepsilon_o}}\sqrt{\varphi_t e^{-(\varphi + V_{ChB})/\varphi_t} + \varphi - \varphi_t + e^{-2\varphi_B/\varphi_t}\left(\varphi_t e^{-(\varphi + V_{ChB})/\varphi_t} - \varphi - \varphi_t\right)} \tag{4 - 39}$$

Then substituting Eqs. 4-36 and 4-39 into Eq. 4-35 we obtain the *Pao-Sah I_D expression* *

$$I_D = \frac{q\,\mu\,Z\,N_A}{L}\int_{V_{SB}}^{V_{DB}}\int_{\varphi_{surf}}^{0}\left[\{e^{(\varphi - 2\varphi_B - V_{ChB})/\varphi_t}\}/\,\mathcal{E}(\varphi)\right]\,d\varphi\,dV \tag{4 - 40}$$

In this equation, we must first calculate the limit of integration φ_{surf} by using

$$V_{GB} = V_{FB} + \varphi_{surf} \tag{4 - 41}$$

$$+\ \gamma\sqrt{\varphi_t e^{-(\varphi_{surf} + V_{SB})/\varphi_t} + \varphi_{surf} - \varphi_t + e^{-2\varphi_B/\varphi_t}\left(\varphi_t e^{-(\varphi_{surf} + V_{SB})/\varphi_t} - \varphi_{surf} - \varphi_t\right)}$$

where $\gamma = \sqrt{2\kappa_{Si}\varepsilon_o N_A/C_{ox}}$. Note that this equation must be solved iteratively for φ_{surf}.

To solve for I_D once φ_{surf} is obtained, Eq. 4-40 must be solved numerically. This involves numerically evaluating the double integral in Eq. 4-40. This is a computationally lengthy procedure. Nevertheless, the Pao-Sah model has been found to accurately predict the I_D in long-channel MOSFETs.

Figure 4-14 is an example of the I_D-V_{DS} characteristics of a long-channel MOSFET

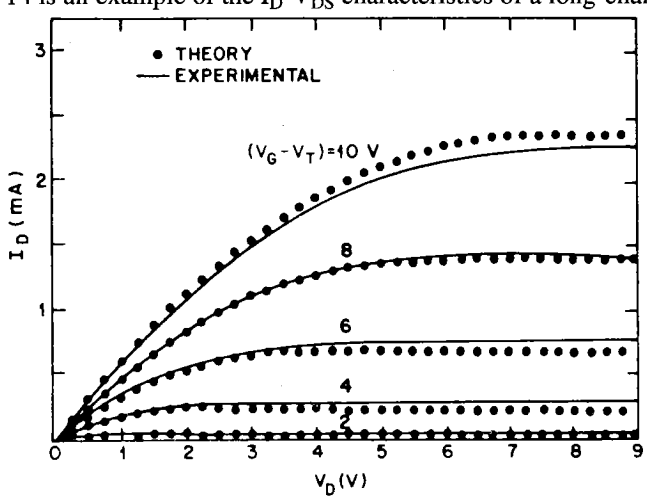


Fig. 4-14. Drain current in a long-channel *p*-channel MOSFET as calculated with the Pao-Sah model (dots) and as measured (solid lines).[4] Reprinted with permission of Solid-State Electronics.

* If $V_{BS} = 0V$, the limits of the voltage integral become 0 and V_{DS}, respectively.

calculated with the Pao-Sah model, together with experimentally measured results. We see that the agreement between measured and predicted I_D values is excellent.

The computational complexity of the Pao-Sah model, however, makes it impractical for use as a MOSFET model in circuit simulation applications. Thus, we show how simpler analytical models can be developed by limiting the validity of the model to reduced regimes of operation. We also mention that an approach has been proposed to reduce Eq. 4-40 to a single-integral formula (which, however, must still be evaluated numerically).[6] However, we do not discuss this approach here.

4.5.3 The Charge Sheet Model

The limited practical utility of the Pao-Sah model motivated a search for an approximate advanced analytical model that still was accurate over a wider range of operating conditions than the earlier *variable depletion-charge* (introduced in 1964) and *square law* models, which are only valid if the MOSFET is operated in strong inversion. The charge-sheet MOSFET, introduced separately by Bacarani[7] and Brews[8] in 1978, has become the most widely adopted long-channel MOSFET model that is accurate over the entire range of inversion.

The charge-sheet model of the long-channel MOSFET makes some simplifying assumptions so the numerical analysis needed in the Pao-Sah model to calculate I_D can be avoided. By doing so, it develops an analytical closed-form expression for I_D in terms of V_{GB} and the surface potential at the drain and source ends of the channel.* It is somewhat less general than the Pao-Sah model in that it is not valid if the MOSFET is operated in depletion or accumulation. However, it is only slightly less accurate than the Pao-Sah model if the MOSFET is being operated in inversion. Hence, by using this model it is possible to get quite accurate predictions of I_D in these regimes without having to resort to the complex numerical analysis required by the Pao-Sah model.

We introduced the charge-sheet approximation for the MOS-C in chap. 3. The MOSFET model is based on essentially the same approximations, with a few more added to account for the current flow behavior of the MOSFET. That is, the charge-sheet approximation basically assumes that the inversion layer (both in weak and strong inversion) can be treated as an infinitesimally thin sheet of charge at the Si surface, and that the depletion approximation is valid for the depletion region below the inversion layer. As such, the inversion-layer charge density Q_I can be found by subtracting the ionic charge in the depletion region Q_B from the total charge in the semiconductor Q_C.

The other assumptions of the charge-sheet model of the MOSFET are the following:

1. The gradual channel approximation applies to the MOSFETs being treated by this model.
2. The drain current consists of both drift and diffusion components.
3. The mobility μ in the channel is proportional to the electric field and is constant with position along the channel.

* Although, as we shall see, a numerical iteration procedure must be used to determine these surface potential values from the terminal voltages.

We begin the derivation of the charge-sheet I_D model by assuming that I_D consists of both drift and diffusion current components. Since we also assume that the mobile charge is located in a very thin region near the surface, all the current is located at one depth (i.e., there is no x-dependence of current). Hence, Eq. 4-16 can be expressed as

$$I_D(y) = I_{Ddrift}(y) + I_{Ddiff}(y) \tag{4 - 42}$$

Now, we consider each of these terms separately. To express I_{Ddrift}, we treat a small element of the inversion layer between y and Δy, as depicted in Fig. 4-15. The potential difference across this element is given by the difference of φ_{surf} at the points y and $y + \Delta y$, or $\varphi_{surf}(y + \Delta y) - \varphi_{surf}(y)$. Hence we can write an expression for the drift current using Eq. 4-22 with $\Delta\varphi(y)$ replaced by $\Delta\varphi_{surf}(y)$. In addition, we can express this equation in terms of the mobile charge density per unit area Q_I instead of in terms of n by multiplying the mobile charge density in the channel (qn) by the inversion layer thickness t_{inv}, or $Q_I = -q\, n\, t_{inv}$. Since $I_{Ddrift}(y) = A\, J_{drift} = Z\, t_{inv}\, J_{drift}$, $I_{Ddrift}(y)$ becomes

$$I_{Ddrift}(y) = \mu Z(-Q_I)\, \Delta\varphi_{surf}(y)/\Delta y \tag{4 - 43}$$

If we allow Δy to approach zero we get

$$I_{Ddrift}(y) = \mu Z(-Q_I)\, d\varphi_{surf}/dy \tag{4 - 44}$$

The diffusion component of Eq. 4-42 can likewise be rewritten as

$$I_{D\,diff}(y) = Z D\, dQ_I/dy = Z \mu (kT/q)\, dQ_I/dy \tag{4 - 45}$$

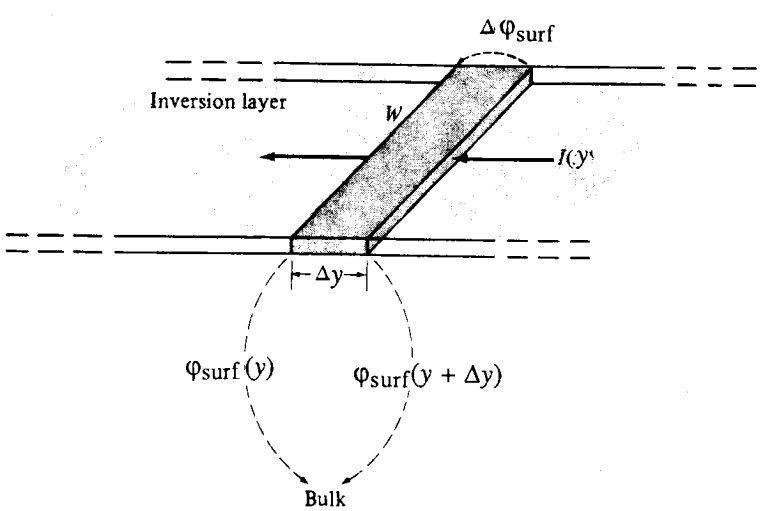


Fig. 4-15 Small element of the inversion layer in a long-channel MOSFET.[5] Reprinted with permission of McGraw-Hill Book Co.

In the case of dc biasing, the total current in the channel must be constant for all points in the y-direction. Thus, Eq. 4-42 can be written as

$$I_D\ (y) = \mu Z\,(-Q_I)\, d\varphi_{surf}/dy + Z\,\mu\,(kT/q)\, dQ_I/dy \qquad (4 - 46)$$

If we designate φ_{surf} at the source end of the channel ($y = 0$) by φ_{surf0}, and Q_I there by $Q_{Isource}$, and the corresponding quantities at the drain end of the channel ($y = L$) by φ_{surfL} and Q_{Idrain}, then by integrating from $y = 0$ to $y = L$ we get

$$\int_0^L I_D\, dy = \mu Z \int_{\varphi_{surf0}}^{\varphi_{surfL}} (-Q_I)\, d\varphi_{surf} + \frac{\mu Z kT}{q} \int_{Q_{I,source}}^{Q_{I,drain}} dQ_I \qquad (4 - 47)$$

or

$$I_D = \frac{\mu Z}{L} \int_{\varphi_{surf0}}^{\varphi_{surfL}} (-Q_I)\, d\varphi_{surf} + \frac{\mu Z kT}{L q} \int_{Q_{I,source}}^{Q_{I,drain}} dQ_I \qquad (4 - 48)$$

To perform the integrals in Eq. 4-47 and 4-48 we require an expression relating Q_I and φ_{surf}. Equation 4-6 derived in section 4.2.2.1 is the appropriate expression we seek, and is repeated here

$$Q_I = - C_{ox}(V_{GB} - V_{FB} - \varphi_{surf} - \gamma \sqrt{\varphi_{surf}}) \qquad (4 - 49)$$

Using this in Eq. 4-48 for the drift current component of the drain current, performing the integration from φ_{surf0} to φ_{surfL}, we get the following expression for I_{Ddrift}

$$I_{Ddrift} = (\mu C_{ox} Z/L)\, [(V_{GB} - V_{FB})(\varphi_{surfL} - \varphi_{surfO}) - (1/2)(\varphi_{surfL}^2 - \varphi_{surfO}^2)$$

$$- (2/3)\, \gamma\, (\varphi_{surfL}^{3/2} - \varphi_{surfO}^{3/2}) \qquad (4 - 50)$$

Using Eq. 4-49 for Q_I (in which we let $\varphi_{surf} = \varphi_{surfO}$ at the source and φ_{surfL} at the drain when we compute the values of Q_I at the limits of integration), in the term of Eq. 4-48 for I_{Ddiff}, we get

$$I_{Ddiff} = (\mu C_{ox} Z/L)(kT/q)(\varphi_{surfL} - \varphi_{surfO}) + (kT/q)\, \gamma\, (\varphi^{1/2}{}_{surfL} - \varphi^{1/2}{}_{surfO}) \qquad (4 - 51)$$

The last step needed when computing I_D with this model is to determine φ_{surfO} and φ_{surfL} from the terminal voltages V_{GB}, V_{SB}, and V_{DB}. It turns out that Eq. 4-4 is the appropriate expression for this task if φ_{surf} and V_{CB} are replaced with φ_{surfo} and V_{SB} at the source end of the channel, and with φ_{surfL} and V_{DB} at the drain end. Then we get

$$\varphi_{surfO} = V_{GB} - V_{FB} - \gamma\sqrt{\varphi_{surfO} + \varphi_t\, e^{(\varphi_{surf0} - 2\varphi_B - V_{SB})/\varphi_t}} \qquad (4 - 52a)$$

$$\varphi_{surfL} = V_{GB} - V_{FB} - \gamma\sqrt{\varphi_{surfL} + \varphi_t\, e^{(\varphi_{surfL} - 2\varphi_B - V_{DB})/\varphi_t}} \qquad (4 - 52b)$$

These equations for φ_{surfO} and φ_{surfL} however, are relatively complex transcendental equations which must be solved by iterative techniques to find φ_{surfO} and φ_{surfL}.

Now, I_D predicted by the charge-sheet model is a single expression given by the sum of Eqs. 4-50 and 4-51.

$$I_D\,(\text{charge sheet}) = I_{Ddrift}\,(\text{Eq. 4-50}) + I_{Ddiff}\,(\text{Eq. 4-51}) \qquad (4 - 53)$$

Experimental MOSFET characteristics are observed to agree very well with such predictions.

4.5.4 Variable Depletion-Charge (or Bulk-Charge) Model

As just described, the charge-sheet model for I_D in its full form includes drift and diffusion current components, making it accurate over the MOSFET operating regimes from weak to strong inversion. However, the fact that neither Eq. 4-50 nor 4-51 is explicitly expressed in terms of the terminal voltages makes the full charge-sheet formulation too cumbersome for use in circuit design applications.

Nevertheless, it is possible to simplify the charge-sheet model formulation to obtain analytical expressions that explicitly give I_D in terms of the terminal voltages. The price for such simplicity (while maintaining adequate accuracy) is that the resulting equations are only accurate over a limited operating range. Specifically, instead of being required to evaluate I_D from Eq. 4-53 (which is a single, relatively complex equation, valid from weak to strong inversion), we develop two less complex approximations to Eq. 4-53, one which is only valid over the weak inversion regime (subthreshold), and the other only over the strong inversion regime. In this section we derive the strong-inversion approximation to Eq. 4-53, which is termed the *variable-depletion-charge* or *bulk-charge* model, and was developed in 1964.[9] In section 4.7 we develop the simplified I_D equation valid for weak inversion. Note that in both of these models the charge-sheet and depletion approximations are still invoked.

We begin the derivation of the strong-inversion model by recalling that once a MOSFET is biased beyond the point of strong inversion, the diffusion current component of I_D is negligible compared to the drift current component; thus we can approximate I_D with only the drift current term of Eq. 4-53. In addition, we assume that beyond the onset of strong inversion φ_{surf} remains fixed. That is, the surface potential at the source and drain ends of the channel in strong inversion is assumed to be clamped at the value possessed at the onset of strong inversion. In that case, φ_{surf} can be approximated as

$$\varphi_{surfO} \approx 2\varphi_B + V_{SB} \qquad (4 - 54a)$$

and

$$\varphi_{surfL} \approx 2\varphi_B + V_{DB} \qquad (4 - 54b)$$

The final approximation we use to derive the variable-depletion-charge model is that if V_{GS} is biased below strong inversion (i.e., $V_{GS} < V_T$), I_D is assumed to be zero.

Invoking these approximations, I_D in a MOSFET can now be approximated by Eq. 4-50, with the values of φ_{surfO} and φ_{surfL} in this equation given by Eqs. 4-54a and 4-54b. Thus, by substituting these into Eq. 4-50 we get

$$I_D = \frac{\mu C_{ox} Z}{L}(V_{GB} - V_{FB})(V_{DB} - V_{SB}) - \frac{1}{2}[(V_{DB} + 2\varphi_B)^2 - (V_{SB} + 2\varphi_B)^2]$$

$$- \frac{2}{3}\gamma[(V_{DB} + 2\varphi_B)^{3/2} - (V_{SB} + 2\varphi_B)^{3/2}] \qquad (4 - 55)$$

By rearranging terms, this equation becomes

$$I_D = \frac{\mu C_{ox} Z}{L}(V_{GB} - V_{FB} - 2\varphi_B)(V_{DB} - V_{SB}) - \frac{1}{2}[(V_{DB}^2 - V_{SB}^2)$$

$$- \frac{2}{3}\gamma[(V_{DB} + 2\varphi_B)^{3/2} - (V_{SB} + 2\varphi_B)^{3/2}] \qquad (4 - 56)$$

We observe that I_D is now explicitly expressed in terms of the terminal voltages V_{GB}, V_{DB}, and V_{SB}. To put Eq. 4-56 into a form in terms of V_{GS}, V_{DS}, and V_{BS}, we recall that $V_{GB} = V_{GS} + V_{BS}$ and $V_{DB} = V_{DS} + V_{BS}$. This allows us to rewrite Eq. 4-56 as

$$I_D = \frac{\mu C_{ox} Z}{L}(V_{GS} - V_{FB} - 2\varphi_B)V_{DS} - \frac{1}{2}V_{DS}^2$$

$$- \frac{2}{3}\gamma\,[(V_{DS} + V_{BS} + 2\varphi_B)^{3/2} - (V_{BS} + 2\varphi_B)^{3/2}] \qquad (4 - 57)$$

If $V_{BS} = 0V$, Eq. 4-57 becomes

$$I_D = \frac{\mu C_{ox} Z}{L}(V_{GS} - V_{FB} - 2\varphi_B)V_{DS} - \frac{1}{2}V_{DS}^2 - \frac{2}{3}\gamma(2\varphi_B + V_{DS})^{3/2} - (2\varphi_B)^{3/2} \qquad (4 - 58)$$

Equation 4-57 (and 4-58) is sometimes referred to as the *Ihantola and Moll model*, since it was initially developed by them in 1964.[9] The name *variable-depletion charge model* signifies that the model computes Q_I as the difference between Q_C and Q_B (*bulk* charge), taking into account the variation of the depletion region width along the channel to compute Q_B. We will see that the square-law model assumes the depletion region width is constant from source to drain.

Equation 4-57 predicts that for a given V_{GS} the drain current first increases linearly with V_{DS} (the *linear* regime), then gradually levels off, approaching a saturated value (the *saturation* regime). Thus, it behaves quantitatively as depicted by the basic output characteristic of the idealized MOSFET of Fig. 4-11.

We noted in section 4.4.2 that as V_{DS} is increased, the output characteristic levels off. This leveling off occurs because as V_{DS} increases, the channel near the drain is driven toward pinch off. In terms of the gradual channel approximation, pinch off is the point at which Q_I drops to zero. (In fact, once Q_I gets very small, the gradual channel approximation breaks down and Eq. 4-57 no longer applies.) When the V_{DS} reaches the

value at which pinch off has been reached, the measured I_D value in long-channel MOSFETs *saturates*. That is, it becomes essentially constant as V_{DS} is increased further. The I_D value at that point is denoted as I_{Dsat}. Although Eq. 4-57 is no longer valid for this regime, we use the maximum value of I_D found using Eq. 4-57 as I_{Dsat}. We denote the V_{DS} value to just reach this condition as $V_{DS,sat}$.

Since $V_{DS,sat}$ is the V_{DS} value when I_D reaches its maximum, such a maximum can be found by setting $dI_D/dV_{DS} = 0$ from Eq. 4-57 and solving for $V_{DS} = V_{DS,sat}$

$$V_{DSsat} = V_{GS} - V_{FB} - 2\varphi_B + \frac{\gamma^2}{2} - \gamma\sqrt{V_{GS} - V_{FB} + V_{BS} + \frac{\gamma^2}{4}} \qquad (4 - 59)$$

The value of $V_{DS,sat}$ obtained from Eq. 4-59 is used in Eq. 4-57 to calculate I_{Dsat}.

Note that the linear and saturation regimes of MOSFET operation have been defined. Since we also assume in this model that $I_D = 0$ for $V_{GS}<V_T$, we define this bias range as the *cutoff* regime. In fact, a small but nonzero drain current exists even when $V_{GS} <V_T$, so it is more appropriate to refer to this regime as the *subthreshold* regime. Further details on this regime are given in section 4.7.

4.5.5 Square-Law Model

Equation 4-57 predicts I_D values in long-channel MOSFETs (operated in strong inversion) quite accurately. However, even this formulation is mathematically too complex for use in hand calculations. Furthermore, when it is necessary to simulate the circuit behavior of digital IC circuits containing a large number of MOSFETs, using Eq. 4-57 in SPICE significantly slows the simulation time. For such applications, a simplified version of Eq. 4-57 could be useful, even at the expense of somewhat less accuracy. This can be obtained if $2\varphi_B + V_{BS}>>V_{DS}$. In this case, we can use the binomial expansion to get

$$(2\varphi_B + V_{SB} + V_{DS})^{3/2} = (2\varphi_B + V_{SB})^{3/2}\left(1 + \frac{V_{DS}}{2\varphi_B + V_{SB}}\right)^{1/2} \qquad (4 - 60)$$

$$\approx (2\varphi_B + V_{SB})^{3/2} + \frac{3}{2}(2\varphi_B + V_{SB})^{1/2} V_{DS} \qquad (4 - 61)$$

Thus,

$$I_D = \frac{\mu C_{ox} Z}{L}\left(V_{GB} - V_{FB} - 2\varphi_B - [2\kappa_{si}\varepsilon_0 N_A (2\varphi_B + V_{BS})/C_{ox}]^{1/2} - V_{DS}/2\right)V_{DS} \qquad (4 - 62)$$

which can be written more succinctly as

$$I_D = [\mu C_{ox} Z/L]\,(V_{GB} - V_T - V_{DS}/2)V_{DS} \qquad V_{GS} \geq V_T \qquad (4 - 63a)$$

$$V_{DS} \leq (V_{GS} - V_T)$$

since $V_T = V_{FB} + 2\varphi_B + [2\kappa_{si}\varepsilon_0 N_A(2\varphi_B + V_{BS})/C_{ox}]^{1/2}$. Or, Eq. 4-63a can be written in an even more compact way (again assuming that $V_{BS} = 0$, in which case $V_{GB} = V_{GS}$),

as

$$I_D = k\,[(V_{GS} - V_T)V_{DS} - V_{DS}^2/2] \tag{4-63b}$$

where k (the so-called device transconductance parameter) is defined as

$$k = \mu\, C_{ox}\, Z/L \tag{4-64}$$

For very small V_{DS} values, Eq. 4-63 can be further approximated as

$$I_D = k\,(V_{GS} - V_T)\,V_{DS} \tag{4-65}$$

For $2\varphi_B + V_{SB} >> V_{DS}$, the condition for saturation becomes

$$-C_{ox}\,(V_{GS} - V_{FB} - 2\varphi_B - V_{DS,sat}) + (2\kappa_{si}\varepsilon_o q N_A[2\varphi_B + V_{BS}])^{1/2} = 0 \tag{4-66}$$

so

$$V_{DS,sat} \approx V_{GS} - V_T \tag{4-67}$$

In saturation, I_D in this model is given by Eq. 4-62 evaluated at $V_{DS} = V_{DS,sat}$; that is

$$I_{Dsat} = k/2\,(V_{GS} - V_T)^2 \qquad V_{GS} \geq V_T: \tag{4-68}$$

$$V_{DS} > (V_{GS} - V_T)$$

Since Eq. 4-68 predicts that I_{Dsat} will vary as the square of the gate voltage above threshold, Eq. 4-62 is termed the *square-law model.*

Equation 4-68 is often used as the basis for measuring V_T. In this technique, the gate and drain of the transistor are tied together as shown in Fig. 4-16, and then the drain current is measured as a function of applied drain voltage. Because V_{DS} is set equal to V_{GS} in this circuit, the transistor is in the saturated region of operation. Now a plot of

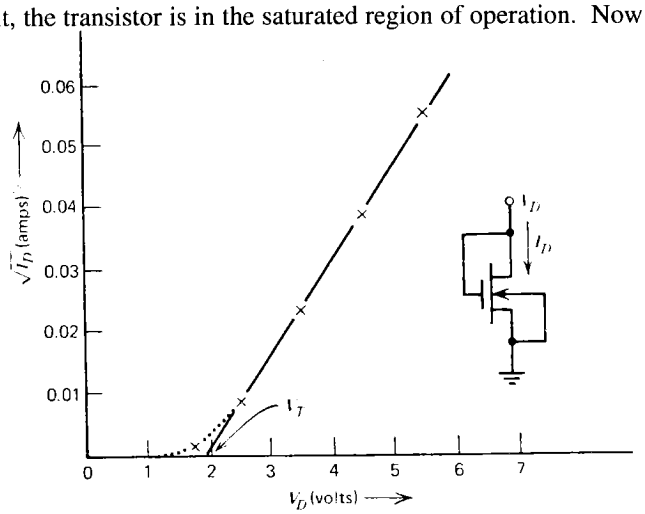


Fig. 4-16 Plot of $\sqrt{I_D}$ versus V_{DS} for n-channel MOSFET in saturation. Inset shows the circuit arrangement. The threshold voltage is indicated by the intercept of the straight line with the voltage axis.

$\sqrt{I_D}$ versus V_{GS} should be a straight line with a voltage axis intercept equal to V_T. In fact, plotting $\sqrt{I_D}$ versus V_{GS} is the most widely used method of determining V_T, yielding as will be discussed in section 4.6, the so-called *extrapolated threshold voltage* value.

From a physical perspective the square-law model can be derived by approximating that the channel depletion region has a constant width along the entire length of the channel (with a value equal to that near the source), even under $V_{DS} \neq 0$ biasing. This approximation is somewhat reasonable in strong inversion since the last term in the strong-inversion form of Q_I (Eq. 4-10), namely $[\sqrt{2\kappa_{si}\varepsilon_o q N_A (2\varphi_B + V_{ChB})}]$, is small compared to the other terms. Since it is the only term that depends on the depletion-region charge, and this charge is much smaller than the mobile charge concentration under strong inversion, the variation of this term will not change the value of Q_I much. That is, as long as the channel region remains inverted, Q_I does not vary much from the source to the drain ends of the channel. Hence, approximating this term by a constant value (namely with the charge that would exist if the depletion region had a constant width, equal to the value it at has at the source end) does not change the value of Q_I significantly. Yet it results in a much simpler form of the I_D model.

At the source the depletion region width is d_{max}(source), which can be calculated from Eq. 4-8

$$d_{max}\,(\text{source}) = [2\kappa_{si}\varepsilon_o(2\varphi_B + V_{BS})/qN_A]^{1/2} \tag{4 - 69}$$

Then

$$Q_B = [2\kappa_{si}\varepsilon_o N_A(2\varphi_B + V_{SB}]^{1/2} \tag{4 - 70}$$

and using Eq. 4-10

$$Q_I = -C_{ox}[V_{GB} - V_{FB} - 2\varphi_B - V_{ChB} - \gamma(2\varphi_B + V_{BS})^{1/2}]$$

$$= -C_{ox}[V_{GB} - V_T - V_{ChB}) \tag{4 - 71}$$

Thus,

$$I_D = \frac{\mu\, C_{ox}\, Z}{L}\int_{V_{SB}}^{V_{SB}+V_{DS}} (V_{GB} - V_T - V_{ChB})\, dV_{ChB} \tag{4 - 72}$$

Integrating 4-72 from source to drain gives 4-62.

4.5.6 Comparison of the Four MOSFET Circuit Models

Let us summarize the characteristics of the four long-channel MOSFET dc circuit models just described, and make some further comparisons about their relative accuracies and utility. Recall first off that all are based on the long-channel (GCA) approximation.

The most general and accurate of the models is the *Pao-Sah* model. While it retains the gradual-channel approximation, it does not invoke the depletion approximation and permits carrier transport in the channel by both drift and diffusion current. The formulation of the I_D equation it develops is therefore quite general, but also requires numerical integrations in two dimensions. Such numerical complexity has limited the role of this model role to theoretical analyses of the MOSFET. As a result the other three MOSFET models are more frequently used to calculate the I_D-V_{DS} characteristics.

The *charge sheet* model allows I_D to be calculated with much less computational effort than that of the Pao-Sah model and yet provides I_D values very close to those predicted by the more complex model over the regime of inversion (both weak and strong). The accuracy of the charge-sheet model when compared to measured dc I_D-V_{DS} characteristics of a MOSFET in inversion is about 1% or better. This model calculates the surface potential by solving the Poisson equation assuming the depletion approximation and the charge-sheet approximation. It also assumes that both drift and diffusion current can contribute to I_D. While not applicable to the depletion and accumulation regimes of operation, it is valid over all regimes of inversion. Although it accurately predicts I_D over these regions of operation, it requires values of the surface potential to be calculated at the source and drain ends of the channel. These equations for φ_{surf} are complex transcendental equations that require iterative techniques to solve them. As a result, the computational effort for determining I_D is still too great for most circuit simulation applications, and thus the charge sheet model has not been incorporated into such popular circuit simulation programs as SPICE.*

Since the charge-sheet model is too complex to be used in most circuit simulation applications, the simpler bulk-charge and square law models are used instead. They both assume that the MOSFET is biased to strong inversion, and that only drift current contributes to I_D. As long as the MOSFET is biased into strong inversion (V_H in Fig. 4-17) the results of the bulk-charge I_D equation agree very well with the charge-sheet model results. That is, to allow the simpler analytical expressions of the bulk-charge (Eqs. 4-57 and 4-59) and square-law models (Eqs. 4-63 and 4-68) to be used to accurately calculate I_D, the allowable regime of operation must be reduced to the strong-inversion regime. If the MOSFET is operated in weak inversion, an alternate analytical expression (as derived in section 4.7) must be used to get an accurate value of I_D in long-channel MOSFETs.

When comparing the utility of the *bulk-charge* and *square-law* models, the primary asset of the square-law model is its simplicity. It provides good insight into the device operation and is useful for obtaining first-order design data without an excessive amount of mathematical entanglement, particularly for the case of digital circuits. Most hand calculations for circuit design with MOSFETs therefore make use of the square-law model. It is also the model used for the simplest (Level-1) analyses of circuits containing MOSFETs in the SPICE program. The apparently arbitrary adjustment of k

* Note that the bulk-charge and square-law MOSFET models assume that the surface potential at inversion is fixed at $2\varphi_B + V_{CB}$. In reality this is the value of φ_{surf} at which the surface just enters strong inversion. If the electron charge in the channel is substantial, the surface potential must be greater than this value (but not much more than a few kT above $2\varphi_B + V_{CB}$). Nevertheless, both these models must therefore consistently underestimate the φ_{surf}.

Now, if φ_{surf} is underestimated, so is the magnitude of the depletion region charge Q_B. If Q_B is underestimated, then from Eq. 4-1b Q_I must be overestimated. The bulk charge model therefore consistently overestimates the amount of mobile electron charge in the channel, and thus overestimates the drain current - fortunately by less than 10 percent in most cases.

The charge sheet model, on the other hand does not make the above assumption, but instead calculates φ_{surf}. Hence, this error is avoided in the charge-sheet model approach.

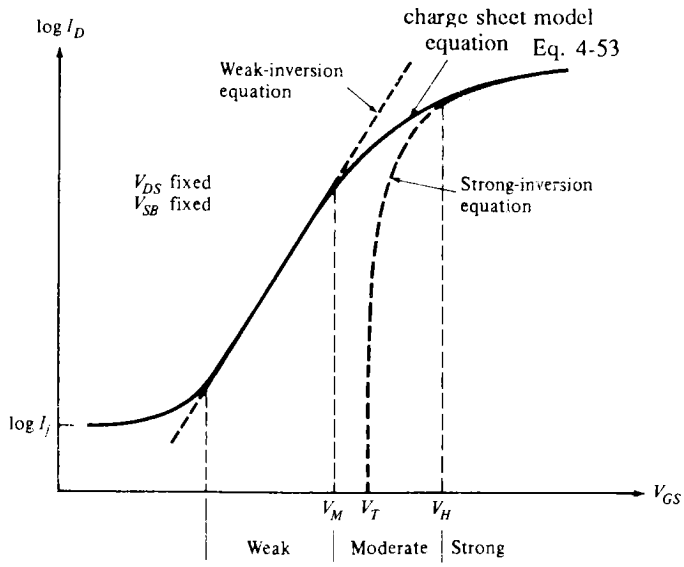


Fig. 4-17 Log I_D versus V_{GS} for fixed V_{DS} and V_{BS}, and comparison to weak- and strong-inversion equations.[5] Reprinted with permission of McGraw-Hill Book Co.

mentioned above is also not a practical problem since MOS processing procedures do not allow for the precise control of μ_n, and the parameter must be extracted from measurements.

There are many cases of circuit design for which the increased accuracy of the bulk-charge model is not necessarily an advantage when computational complexity is considered. Often, for example, the designer is interested only in MOSFET operation in the saturated region (i.e., only the value of I_{Dsat} is needed). To obtain I_{Dsat} from the bulk-charge mode, it is first necessary to calculate V_{DSsat} from Eq. 4-59 and then use this voltage value in Eq. 4-58 to calculate I_{Dsat}. This lengthy process is greatly simplified in the square-law model in which I_{Dsat} is calculated directly from Eq. 4-68.

The bulk-charge model is used when more accurate modeling of circuit behavior is needed. Measurements of I_D versus V_{DS} characteristics of long-channel MOSFETs are in reasonable agreement with the predictions of this model (i.e., usually within 10% of the measured data). However, the computational complexity is considerably increased, and hence it is typically implemented only in computer analyses. This model is available as the Level-2 model of the MOSFET in SPICE. Note that when comparing Eqs. 4-57 and 4-59 with 4-63 and 4-68, the added terms in the bulk-charge model equations are always negative and act primarily to reduce I_D and V_{DSsat} for a given set of operational conditions.

A plot of I_D versus V_{DS} (with V_{GS} as a parameter) for a long-channel NMOS transistor calculated using both models is shown in Fig. 4-18. As can be seen the square-law model predicts larger values of I_{Dsat} and V_{DSsat} for each value of gate

voltage. The general behavior of both sets of curves is, however, similar. Note especially that in the linear region of operation both models give the same results, but that in saturation - for the same value of k in both models - the bulk-charge model predicts a smaller value of I_{Dsat}.

4.5.6.1 MOSFET Models of the SPICE Circuit Simulator. The SPICE circuit simulator contains four MOSFET models which differ in the way they formulate the I-V characteristic. The simplest is the Level-1 model which is based on the rudimentary square-law model (Equations 4-63 and 4-68), with empirical adjustment. The most important adjustment it employs is the use of a coefficient in these equations to bring I_{Dsat} into congruence with experiment and the more accurate models. That is, to obtain satisfactory results in circuit analysis and design using this model, measured data on samples of MOSFETs must be obtained. Measured data is fitted to the threshold voltage equation and Eqs. 4-63 and 4-65. Specifically V_T and k are determined from data obtained as shown in Fig. 4-16. Note that the extrapolated threshold voltage value V_{T0} is used in SPICE because it is obtained by extrapolating the experimental curve of $\sqrt{I_D}$ versus V_{DS} (with $V_{DS} = V_{GS}$) to zero drain current. The subscript "0" in the V_{T0} symbol signifies that this is the extrapolated threshold voltage when $V_{BS} = 0V$.

The more-accurate Level-2 SPICE MOSFET model is based on the bulk-charge model equations (Eqs. 4-57 and 4-59). The Level-2 model also incorporates an approximate

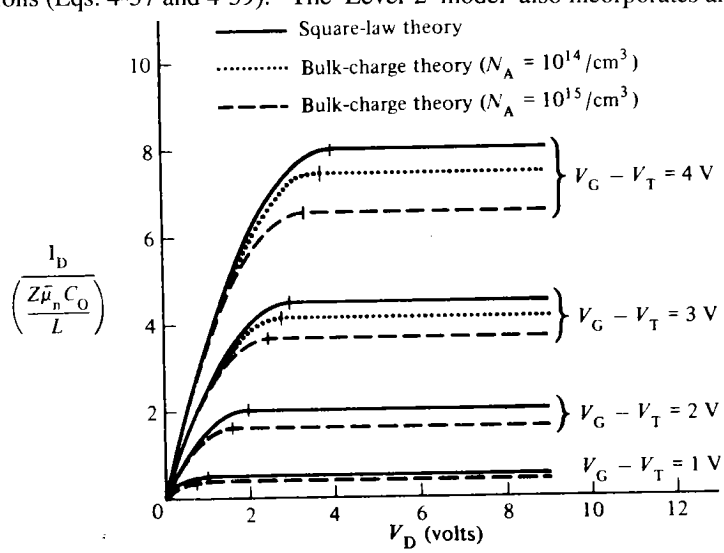


Fig. 4-18 Theoretical predictions of I_D versus V_{DS} using the bulk-charge model equations (solid lines) and the square-law model equations (dashed lines). I_D is taken to be equal to I_{Dsat} for $V_{DS}>V_{DSsat}$.[30] Reprinted with permission of Addison-Wesley.

* For example, in the Level-2 MOSFET model, the carrier mobility is not treated as a constant. Instead, the mobility is modeled as being impacted by the gate-induced electric field (see chap. 5). The value used in the Level-2 model μ_{eff} is found using the expression Eq. 3-26.

treatment of some of the short-channel effects discussed in chap. 5* and is capable of computing the subthreshold current (in the manner as described in section 4.7).

The basic device structure parameters required to describe a MOSFET when using either the Level-1 or Level-2 model are the substrate doping, the gate oxide thickness, and the channel length and width. SPICE computes the threshold voltage from N_A (NSUB in SPICE) and t_{ox} using Eq. 4-12 (assuming $V_{FB} = 0$). Alternatively, the threshold voltage (for the case of $V_{SB} = 0$) can be input directly by supplying a value to the model parameter V_{T0}, in which case the internally computed threshold voltage is overridden.

The Level-2 SPICE MOSFET model, being based on the bulk-charge MOSFET model, has a three-halves-power dependence of I_D, rather than the square-law dependence of the Level-1 SPICE MOSFET model. Because the numerical handling of a fractional power requires significantly more computation time than if a power in integer form must be computed, circuit designers sought a model that was more accurate than the Level-1 model, but was faster to execute than the Level-2 model. Hence, the Level-3 model was developed and incorporated into SPICE.*

The Level-3 model not only removed the computational shortcoming of the Level-2 model, but was formulated to deal empirically with many other effects as well, including short-channel and narrow-channel effects. In fact, the Level-3 model represents an attempt to pursue a semi-empirical modeling approach, i.e., which only approximates device physics and relies on the proper choice of the empirical parameters to accurately reproduce device characteristics. As a result, a more physically based short-channel MOSFET model has also been developed for SPICE, known as the Berkeley Short-channel IGFET Model (BSIM). It has been implemented into the SPICE program as the Level-4 MOSFET model. We will describe BSIM in more detail in chap. 5.

4.5.6.2 SPICE Models of the MOSFET Operated in Saturation. As we noted in section 4.4.2, the value of I_D may increase as V_{DS} exceeds the pinchoff voltage value V_{DSsat}. This was modeled by assuming that the pinched-off region with a length ΔL makes the channel length appear shorter (i.e., $L - \Delta L$). To predict I_D as V_{DS} exceeds V_{DSsat} the value of ΔL is needed. In the Level-1 and Level-2 MOSFET models

* The I_D equation that constitutes the dc long-channel Level-3 model of the MOSFET[10] in the SPICE program is derived from Eq. 4-57 by employing a three-term binomial-series expansion for the second term in Eq. 4-57. This yields the Level-3 MOSFET equation for I_D in the form:

$$I_D = \frac{\mu C_{ox} Z}{L}\left[(V_{GS} - V_T)V_{DS} - \left(\frac{1}{2} + \frac{\sqrt{\varepsilon_o q N_A/(-\varphi_B)}}{4 C_{ox}}\right)V_{DS}^2\right]$$

Notice by comparing it to the Level-1 expression, Eq. 4-63, that it differs only in the last term, a change that accounts for the ionic-charge variations that were ignored in the Level-1 model. Here, however, the square root term is only dependent on device structure parameters and not on V_{DS}, and hence this square-root term only needs to be evaluated once when calculating the I_D-V_{DS} characteristics.

SPICE offers two possible approaches to calculating ΔL. The first allows the *channel-length modulation parameter*, λ (LAMBDA) to be specified, in which case ΔL is obtained from the simple empirical expression $\Delta L = \text{LAMBDA} \cdot V_{DS}$. LAMBDA can be used in either the Level-1 or Level-2 model. The value of LAMBDA is typically in the range of 0.1 to 0.01 V^{-1}. Empirical values of $1/\lambda$ can be obtained by finding an approximate intercept with the V_{DS} axis of tangents to the I_D curves beyond V_{DSsat} in measured I_D-V_{DS} curves.

If the value of LAMBDA is not specified in the Level-2 model, however, then ΔL is computed by using Eq. 5-47, which is based on the 1-D Reddi-Sah model of the MOSFET operated in saturation (see chap. 5, section 5.4.3.1).

4.6 THRESHOLD-VOLTAGE CONTROL IN MOSFETS

4.6.1 Measuring Threshold Voltage

Equation 4-1 calculates V_T of real MOSFETs based on various device parameters. Yet, when devices are fabricated, they must be tested to ensure that the desired characteristics have been produced. Two approaches are used to obtain the V_T values of MOSFETs from measurements. The first yields an *extrapolated-V_T* value, and the second, a *constant-current-V_T* value.

As described in section 4.5.5, the extrapolated V_T is determined by connecting the drain and gate of a MOSFET together and measuring the drain current as a function of gate voltage (Fig. 4-16). Equation 4-68 predicts that a plot of $\sqrt{I_D}$ (with a linear I_D axis) versus V_{GS} should be linear. If $\sqrt{I_D}$ is plotted vs. V_{GS}, and the linear portion of the curve is extrapolated to intercept the voltage axis (corresponding to the condition that $V_{GS} = V_T$ in Eq. 4-68) the extrapolated-V_T value is yielded. (Note that the dotted section of the curve in Fig. 4-16 results from currents that flow below the threshold voltage. Such *subthreshold currents* are discussed in section 4.7.)

The constant-current-V_T is obtained by measuring the gate voltage at which a specific small drain current flows and this gate voltage is defined as V_T. For active MOSFETs, the gate voltage at which 1 μA of drain current flows per μm of drain width for V_{DS} = 0.5V is the most common criterion applied when specifying constant-current-V_T. The constant-current-V_T is generally measured in the IC production environment, while the extrapolated V_T is preferred when the threshold voltage parameter is needed for circuit simulation (e.g., in a SPICE analysis). Note also that the extrapolated V_T is typically 100 mV larger than the 1 μA/μm constant-current V_T.

4.6.2 Adjusting Threshold Voltage

As pointed out earlier, in many MOS IC applications it is critical to be able to establish and maintain a uniform and stable V_T. The factors that impact V_T are given in Eq. 3-72. An examination of each term will reveal the device parameters that can be adjusted to provide practical control of V_T.

The φ_{ms} term depends on the work-function difference between the gate, $q\varphi_m$(gate), and the semiconductor, $q\varphi_B$ (sub). While $q\varphi_m$(gate) for metal and heavily doped silicon

gates is constant, the parameter φ_B (sub) depends on the substrate doping - but only in a logarithmic manner. Hence, each factor-of-10 increase in substrate doping will change the φ_{ms} term by only 2.3 kT/q, or ~0.06 V (kT/q = 0.026 V at 300K). Thus, changes in the substrate-doping concentration produce changes in V_T through the φ_{ms} term which are too small to provide the required degree of threshold-voltage control. The next term ($2\varphi_B$) also changes only slightly as a result of changes in the substrate doping concentration (for the same reason given for the φ_{ms} term); thus, the $2\varphi_B$ term is also ineffective for controlling V_T.

Since every attempt is made to keep Q_{tot} as low as possible through various processing procedures, and C_{ox} is relatively large (since t_{ox} is very thin in submicron MOSFETs), the Q_{tot}/C_{ox} term is also very small in modern MOSFETs. Hence, this term must also be ruled out as a candidate for controlling V_T.

While is true that C_{ox} could be varied (primarily by changing t_{ox}), this is not a practical approach to controlling V_T in active devices, since t_{ox} is normally made as thin as possible to maximize I_D. (In the field regions, however, large V_{TF} values are needed to prevent inversion under the field oxide. A thick field oxide makes C_{ox} small, allowing V_{TF} to be increased Thus, C_{ox} is one of the parameters normally used to control V_{TF} in the field regions of the circuits.)

This leaves the $2\sqrt{\kappa_{Si}\varepsilon_o q N_{sub}\varphi_B}/C_{ox}$ term as the remaining candidate for controlling V_T in active devices. Since the change of V_T in this term depends on $\sqrt{N_{sub}}$, it indicates that V_T control is possible by exploiting this effect. In fact, Eq. 3-72 indicates that increasing the substrate doping (i.e., N_A in *n*-channel devices, and N_D in *p*-channel devices) will increase V_T. Figure 4-19 shows how V_T varies in MOSFETs (both PMOS and NMOS) with substrate doping concentration (assuming the substrate is uniformly doped - i.e., no threshold adjust implant is used - and that $Q_{tot} = 0$), for various gate-oxide thicknesses.

Merely increasing the substrate doping, however, is not desirable since it will adversely impact other MOSFET characteristics, such as lower junction-breakdown voltages, larger junction capacitances, and lower carrier mobilities. Yet, prior to the development of ion implantation in the early 1970s, adjustment of substrate doping was the only practical *processing* approach for significantly controlling V_T in active devices.

We showed earlier that a *circuit* approach based on applying a bias between the source and body (and hence known as body biasing) can also control V_T (section 4.2.2.2, Eq. 4-12). Generally, an increase in V_{BS} will raise V_T. Figure 4-20 gives an example of how the application of V_{BS} can change the V_T value in an NMOS device. The intercept of the channel conductance and the V_{GS} axis when $V_{BS} = 0$ in this example is +0.99V. As V_{BS} is increased from 0V to -5V, V_T increases up to +2.3V. However, body biasing is an added complication, and is now avoided whenever possible as a technique for adjusting V_T, in favor of the ion implantation V_T-adjust method.

4.6.3 Ion Implantation for Adjusting Threshold Voltage

The development of ion implantation for V_T adjustment removed the last obstacle to reliable production of *n*-channel devices for MOS ICs, as this procedure made it possible to select the substrate-doping value without having to consider its impact on V_T. That

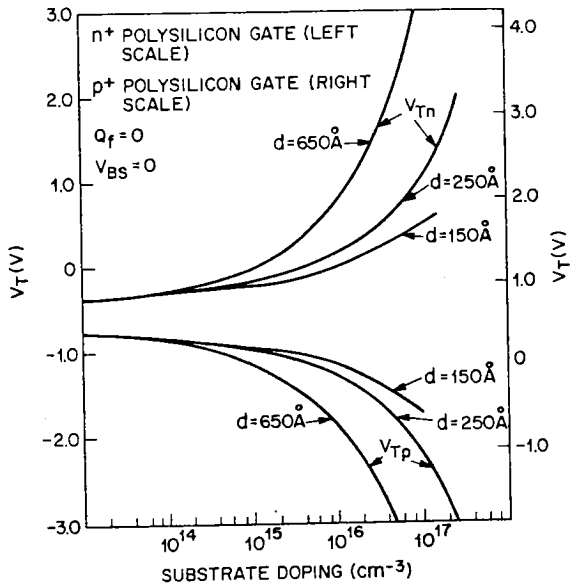


Fig. 4-19 Calculated threshold voltages of n-channel (V_{Tn}) and p-channel (V_{Tp}) transistors as a function of their substrate's doping, assuming n^+ polysilicon gate (left scale) and p^+ polysilicon (right scale). Curves for gate oxide thicknesses t_{ox} of 150 Å, 250 Å, and 650 Å are shown. From S. M. Sze, Ed., VLSI Technology, 2nd. Ed., Chap. 11, "VLSI Process Integration." Copyright, 1988 Bell Telephone Laboratories. Reprinted with permission of McGraw-Hill.

is, implantation can be used either to increase or to decrease (by compensation) the net dopant concentration at the silicon surface. As a result, substrate doping can be selected strictly on the basis of optimum device performance since V_T can now be set by the V_T-

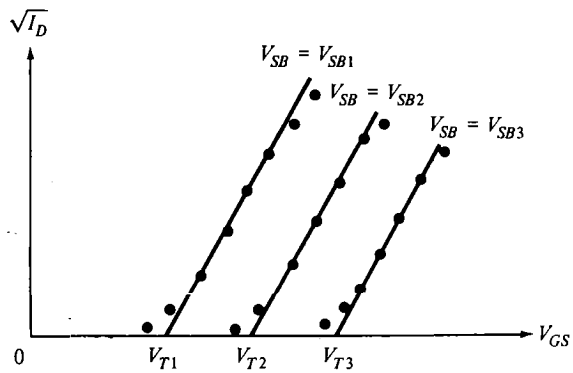


Fig. 4-20 Plot of $\sqrt{I_D}$ vs. V_{DS} for an NMOSFET showing the effects of a bias V_{BS} between the source and the substrate. The observed shift in the threshold is predicted by Eq. 3-71.[16] Reprinted with permission of John Wiley & Sons.

adjust implant process. In addition, since dopants can be selectively implanted into the *field regions*, high-performance NMOS circuits can be fabricated on lightly doped substrates, without the possibility of inadvertent inversion of the surrounding field regions.

As mentioned earlier, the V_T-adjust-implant technique involves implantation of boron, phosphorus, or arsenic ions into the regions under the gate oxide of a MOSFET. Boron implantation produces a positive shift in V_T, while phosphorus or arsenic implantation causes a negative shift. For shallow implants, the procedure has essentially the same effect as placing an additional sheet of "fixed" charge at the SiO_2-Si interface.

To first order, if the implant and substrate are of the same type (e.g., boron into a *p*-substrate), and the effective depth of the implant x_i (see Fig. 4-21) is less than d_{max}, the threshold-voltage change (ΔV_T) can be reasonably well estimated from[11]

$$\Delta V_T\,(II) = q\,D_I\,/C_{ox} \qquad (4 - 73)$$

Example 4-1: If the V_T of a uniformly doped NMOS transistor is -0.41V and it is to be used in an application that requires a V_T = 1.0V, calculate the B dose (to first order) needed to adjust $V_{T(nud,lc)}$ to this value.

Solution: The $V_{T(ud,lc)}$ of the device is -0.41 V. We wish to have a V_T of 1.0 V. Thus, the threshold voltage must be shifted (ΔV_T) by +1.41 V. Using Eq. 4-73, we see that the boron dose needed to cause this ΔV_T is

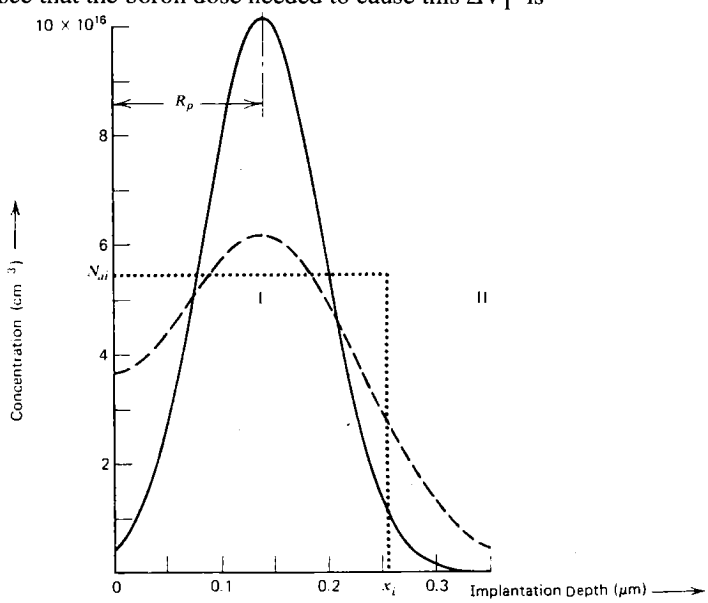


Fig. 4-21 Distribution of the implanted dopant (solid curve) as-implanted, and after activating anneal and diffusions (dashed curve). The dotted "box" curve shows the distribution used to calculate the threshold voltage. (After Rideout, Gaensllen & LeBlanc in ref. 27.)

$$D_I = \Delta V_T(II)\, C_{ox}/q = [1.41V \times 2.3\times10^{-7}\ F/cm^2]/1.6\times10^{-19}\ C$$

$$= 2.1\times10^{12}\ \text{boron atoms}/cm^2.$$

where D_I is the dose of the implanted ions (atoms/cm^2) which penetrate into the silicon. For example, Eq. 4-73 predicts that when $D_I = 5\times10^{11}$ ions/cm^2 and $t_{ox} = 25$ nm, a shift in V_T of 0.58 V will be produced. Thus, Eq. 3-72 must be modified to account for threshold-adjust implants, as (where *nud* stands for non-uniformly doped)

$$V_{T\,(nud,lc)} = V_{T\,(ud,lc)} + \Delta V_T(II) \qquad (4-74)$$

A more exact formula for calculating for V_T in this case (assuming $V_{BS} = 0$) is[12,13,14]

$$V_{T\,(nud,lc)} = V_{FB} + 2\varphi_B + qD_I/C_{ox} + \gamma_2\,(\sqrt{2\varphi_B + qD_I x_i/2\kappa_{si}\varepsilon_o}) \qquad (4-75)$$

where

$$\gamma_2 = (\sqrt{2q\kappa_{si}\varepsilon_o N_A})/C_{ox} \qquad (4-76)$$

The V_T-adjust implant is often done through an oxide layer.* The implant energy is selected to place the peak of the implant slightly below the oxide-silicon interface (Fig. 4-21). After the implant-activating anneal, the implanted distribution is broader than the as-implanted profile.

Calculating the effect of the implant on V_T is greatly simplified by approximating the actual distribution via a "box" distribution, as is also shown in Fig. 4-21 (in which the implanted dopant is assumed to have a constant density N_A' from the surface to the depth x_i.) Figure 4-22 plots V_T vs. effective implant depth x_i, and we can see that V_T does not change much as x_i is varied. Since the second term under the square root in Eq. 4-75 is the only one that depends on x_i, if we ignore it when calculating V_T the results are still adequate for many applications. Thus, the first-order approximation of ΔV_T given by Eq. 4-73 is useful for hand calculations of $V_{T(nud,lc)}$.

Another slightly more complicated approach to calculating V_T for this case is to solve the Poisson equation using a high-low step doping profile, which results in an effective concentration N_{eff} given by[28]

$$N_{eff} = N_A' \frac{\varphi_s}{\varphi_i}\left[1 - \frac{N_A}{N_A'} + \frac{N_A'}{N_A}\sqrt{1 + \frac{N_A}{N_A'}\left(\frac{\varphi_s}{\varphi_i} - 1\right)}\right]^2 \qquad (4-77)$$

where $\varphi_s = kT/q\ [\ln (N_A'N_A/n_i^2)]$ and $\varphi_i = qN_A'x_i^2/2\kappa_{si}\varepsilon_o$. Here φ_i is the potential corresponding to x_i of the step profile. N_{eff} is then substituted for N_A in the body effect coefficient term γ of Eq. 4-12, and then Eq. 4-12 with this value of N_{eff} is used to calculate the value of V_T in MOSFETs with implanted channels. This method for calculating V_T in MOSFETs with non-uniformly doped channel regions is often called the *doping transformation procedure,* and reportedly gives excellent results.

Ion implantation can also be used to fabricate depletion-mode MOSFETs. Depletion-mode NMOS devices (i.e., in which $V_T<0V$) are commonly used in NMOS logic circuits. In order for the required negative threshold voltage for a depletion-mode NMOS device to be produced, *n*-type impurities are implanted into the *p*-type substrate to form a built-in channel between the source and drain. The dose required to shift the threshold voltage by the desired value may also be estimated using Eqs. 3-72 and 4-73.*

4.6.3.1 Impact of the V_T-adjust Implant on the Body Effect. The implanted layer also impacts the variation of V_T with V_{SB}. This is shown in Fig. 4-23. We observe that for uniformly doped substrates [curves (a) and (b)] the variation of V_T with V_{SB} follows the data shown in Fig. 4-20. For shallow implants [curve (c)], the V_T vs. V_{SB} behavior approaches that for a delta function dose at the surface. That is, V_T is observed to be merely shifted by qD_I/C_{OX} from the V_T values given by curve (a), and thus curve (c) is parallel with curve (a). However, for a deeper implant [curve (d)], the initial variation in V_T with V_{SB} is comparable to that for a uniformly doped region of concentration N_A' [i.e., curve (b)]. However, as V_{SB} is increased, the channel depletion region width d approaches and soon exceeds x_i, and the variation of V_T with V_{SB} thereafter behaves as if the substrate is doped with the bulk concentration N_A [like curve (c)]. Since it is generally desirable to have V_T exhibit a low substrate bias sensitivity

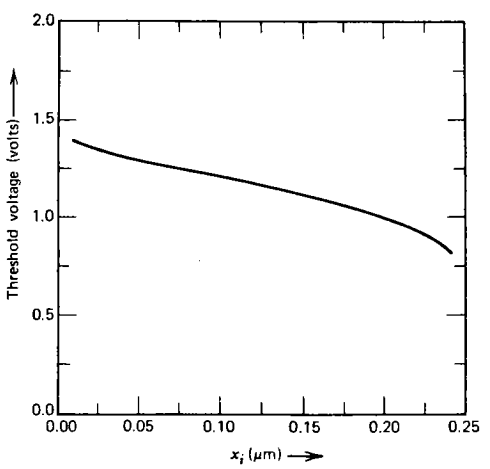


Fig. 4-22 Dependence of the threshold voltage on the effective implantation depth x_i as predicted by Eq. 4-75 for a MOSFET with an 87 nm-thick gate oxide, $Q_f/q = 10^{11}$ cm^{-2}, $D_I = 3.5\times10^{11}$ cm^{-2}, and $N_A = 2\times10^{15}$ cm^{-3}. Both the substrate and source terminal are assumed be grounded.[11] Reprinted with permission of John Wiley & Sons.

* Note that if x_i is greater than d_{max}, $V_{T(nud,lc)}$ can be calculated using Eq. 3-72, with N_A replaced by $N_A' = N_A + D_I/x_i$. However, this condition is generally avoided when choosing the V_T-adjust implant parameters, since the increased doping near the substrate edge of the source and drain junctions leads to higher junction capacitance, lower breakdown voltage, and a larger body effect.

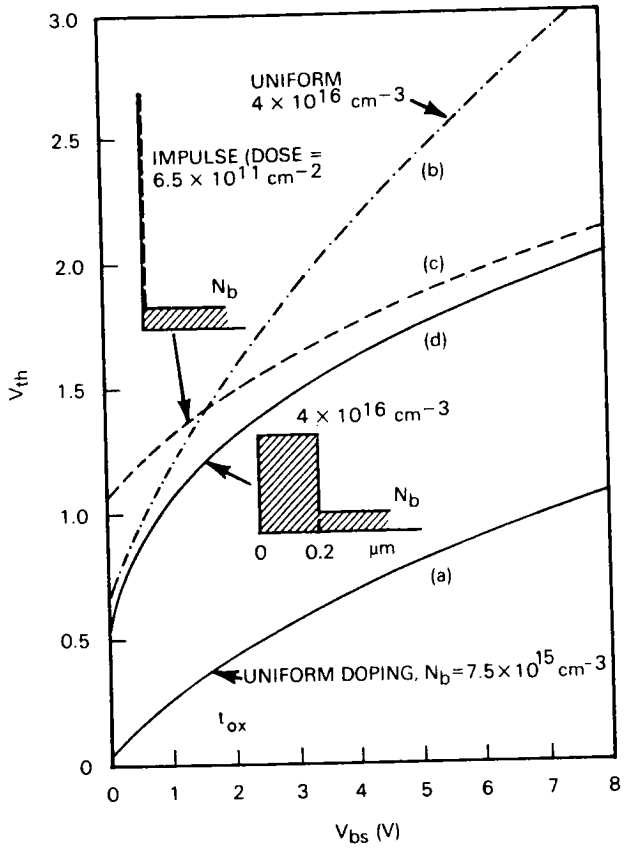


Fig. 4-23 Calculated substrate sensitivity for various doping profiles. (After Rideout, Gaensllen and LeBlanc in ref. 27.)

(i.e., small body effect), Fig. 4-23 also demonstrates that a shallow implant with an appropriate dose can raise V_T while maintaining a small body effect.

4.6.4 Impact of V_T-Adjust Implant on Other MOSFET Characteristics

By implanting ions near the Si surface, a nonuniformly doped channel region is produced. Besides shifting V_T, this alters the MOSFET behavior in several ways compared to devices having uniformly doped channels. First, the V_T-adjust implant introduces extra ions into the channel depletion region, causing its width to be modified. This effect is discussed here. In addition, the non-uniform doping profile changes the long-channel subthreshold characteristics (i.e., the slope of the log I_D vs. V_{GS} curves) as well as the punchthrough behavior of short-channel devices. These two latter effects will be discussed later, in chap. 5.

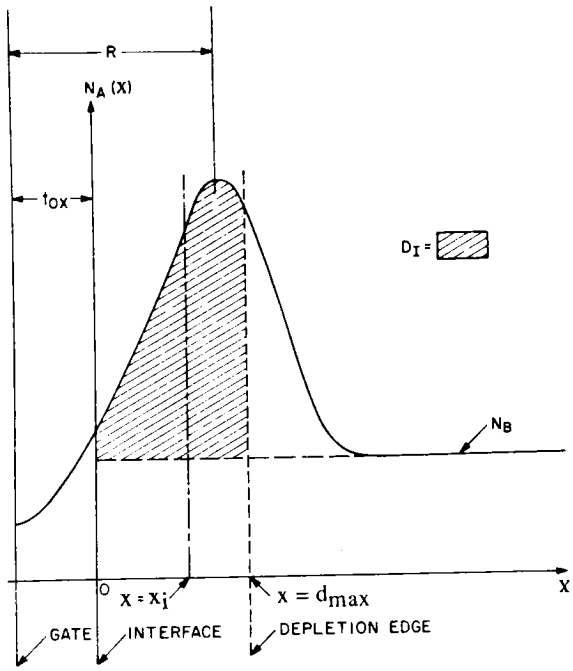


Fig. 4-24 Schematic of the threshold adjust doping profile in an implanted MOSFET. The shaded portion of the profile is the portion of the implanted dose in the depletion region, whose depth extends to d_{max}. The effective depth of the implant (or centroid) is labeled x_i.[8,15] Reprinted with permission of Solid-State Electronics.

However, all of the above effects depend only upon the so-called *exposed dose* D_I (i.e., the portion of the implant dose that ends up in the depletion region), and upon the *centroid* x_d of the exposed dose. That is, those implanted ions which do not penetrate the screening oxide or which penetrate beyond the edge of the channel depletion region have no impact. Thus, these terms must be defined. D_I (the shaded portion in Fig. 4-24), and x_d are given in terms $N_{sub'}(x)$ and N_{sub} by[15]

$$D_I\,(\text{cm}^{-2}) = \int_0^d [N_{sub'}(x) - N_{sub}]\,dx \qquad (4-78)$$

and

$$x_d = \frac{1}{D_I}\int_0^d [N_{sub'}(x) - N_{sub}]\,x\,dx \qquad (4-79)$$

where $N_{sub'}(x)$ is the true channel profile, including the implanted ions, and N_{sub} is the uniform background doping of the substrate, as depicted in Fig. 4-24.

The change in width of the channel depletion region is most easily illustrated by replacing the actual implant profile shown in Fig. 4-24 with a simpler delta-function equivalent implant. That is, the equivalent doping profile $N_{sub'}(x)$ is given by

$$N_{sub'}(x) = N_{sub} + D_I\,\delta(x - x_d) \qquad (4 - 80)$$

Using such a delta-function replacement for the actual profile, the effects of the V_T-implant are illustrated in Fig. 4-25. Figure 4-25a shows a potential versus distance diagram [$\varphi(x)$ versus x] of a long-channel MOSFET with a non-implanted (i.e., uniformly doped) channel region. The channel depletion region width is d and the band bending is φ_{surf}. In Fig. 4-25b, the effect of varying the *dose* D_I on the same depletion region is shown. The presence of D_I at x_d causes a larger voltage drop to occur between x_d and the surface, leaving a smaller portion of φ_{surf} to deplete the region behind the implant. Hence, the total width is reduced to $d_I = d_{DI}$. Doubling D_I further reduces the width (in Fig. 4-25b to d_{2DI}). Eventually, for a value of dose termed the *effective* dose, D_{eff}, d_I is reduced to a point $d_I = x_d$. Any increase of dose beyond D_{eff} simply is neutralized in the bulk, and the condition of $d_I = x_d$ is maintained. In other words, D_{eff} is the minimum dose needed to bring the depletion edge to that depth, making $d_I = x_d$.

Fig. 4-25c depicts the result of increasing implant *depth.* In this case, increasing x_d also increases the voltage drop between x_d and the surface, again reducing the width to a

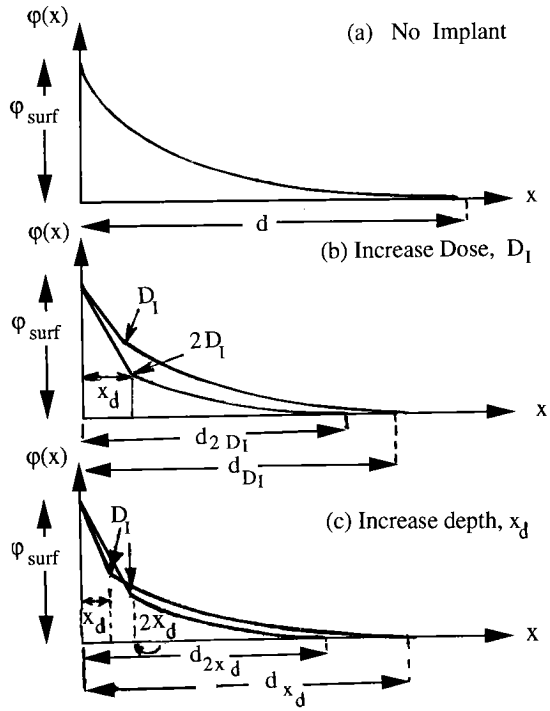


Fig. 4-25 The effects of a delta function implant upon the depletion width. (a) Unimplanted, (b) Effects of changing dose, and (c) Effect of changing depth.[13] (© IEEE 1978)

smaller d_I. As a deeper x_d is used, d_I moves toward the implant. For example, in Fig. 4-25c, we see that for an implant with depth x_d, $d_I = d_{xd}$. This is deeper than d_I due to a an implant with a depth $2x_d$ (here $d_I = d_{2xd}$). Ultimately, $d_I = x_d$, and $D_I = D_{eff}$, corresponding to this depth. Any further increase in x_d carries the depletion region edge with the implant. The depletion region edge becomes *clamped* to the implant.

The quantitative variation of d_I can also be simply expressed by again assuming a delta-function implant. That is, the depletion-region width in the presence of a V_T-adjust implant d_I is calculated from

$$d_I = \sqrt{2}\, L_D\, [\varphi_{surf}/\varphi_t - D_I x_d/(N_{sub} L_D^2) - 1]^{1/2} \qquad (4 - 81)$$

where L_D is the Debye length $\{[\kappa_{si}\varepsilon_o\varphi_t/qN_{sub}]^{1/2}\}$. Eq. 4-81 shows that the implant reduces the width, since in a uniformly doped channel the width d is given by

$$d = \sqrt{2} L_D\, [\varphi_{surf}/\varphi_t - 1]^{1/2} \qquad (4 - 82)$$

Likewise, D_{eff} is given by

$$D_{eff} = N_{sub}\, (d^2 - x_d^2)/2x_d. \qquad (4 - 83)$$

4.7 SUBTHRESHOLD CURRENTS IN LONG-CHANNEL MOSFETs (I_{Dst} when $V_{GS} < |V_T|$)

The small drain current which flows in the MOSFET channel below threshold (i.e., in weak inversion) is called *subthreshold current*, I_{Dst}. In most applications I_{Dst} is far too small to be useful as a drive current. However, it can represent an unwanted leakage current, especially in ICs designed for low-power applications. It can also give rise to premature discharging of memory-cell capacitors or dynamic-logic circuit nodes, preventing these circuits from operating properly. The issues involving leakage currents in MOSFETs will become even more important with the proliferation of battery-powered (i.e., portable) electronic systems. The common range of V_T in submicron digital CMOS ICs is 0.6-0.8V. Thus, when V_{GS} = 0V the MOSFETs in these circuits may be close to weak inversion. Consequently, for many applications (especially those requiring low standby-power dissipation), subthreshold leakage must be well characterized so that the total IC leakage current of the chip can be predicted during the design phase of the product.

In actual MOSFETs, when $V_{GS}<V_T$ the minority carrier concentration in the channel (e.g., n_{ps}) is small (but not zero). That is, as V_{GS} is increased from 0V and approaches V_T, the mobile charge concentration due to n_{ps} remains smaller than the ionic charge of the bulk doping concentration, but continues to increase as V_{GS} rises. In section 3.3.5 we defined *weak inversion* in terms of φ_{surf}, namely as $\varphi_B<\varphi_{surf}<2\varphi_B$.

The bulk-charge and square-law models of I_D are not suitable for treating subthreshold behavior because they were developed to describe MOSFET behavior only in strong inversion. That is, both models assume that $I_D = 0$ if $V_{GS}< |V_T|$. While the Pao-Sah model can provide accurate predictions of I_{Dst}, it requires too much computation to be practical for circuit simulation applications. Thus, the charge-sheet model becomes the candidate for characterizing I_{Dst}. We noted in section 4.5.3,

however, that the I_D expression of the charge-sheet model in its full form is also still too cumbersome for most circuit-design simulation tasks. Fortunately, by making some appropriate simplifying approximations, the general form of the charge-sheet model I_D equation can be reduced to a less complex formulation that still provides accurate predictions of I_{Dst} in long-channel MOSFETs.

Consider the case of a long-channel NMOSFET in which: (1) the source is grounded; (2) V_{GS} approaches (but remains less than) V_T, and; (3) $|V_{DS}| \geq 0.1V$. Now recall that in inversion (i.e., when $V_{GS} > |V_T|$), the free-carrier concentration n_{ps} in the channel is large (e.g., $n_{ps} \geq N_A$), and I_D in this case is dominated by drift current. However, when $V_{GS} < |V_T|$, n_{ps} is much smaller than N_A, but is not zero In addition, in weak inversion the drain voltage V_{DS} drops almost entirely across the reverse-biased drain-substrate depletion region.

Figure 4-26a illustrates the variation of inversion-layer charge concentration versus channel position in a long-channel NMOSFET in weak inversion. Figures 4-26b and c depict $\varphi_{surf}(y)$ and a 2-D perspective plot of the potential $\varphi(x,y)$ diagrams in weak inversion, respectively. We see that the band bending in the *x*-direction is too small to establish a strongly inverted channel [i.e., $\varphi_B < \varphi_{surf} < 2\varphi_B$], thus the channel mobile charge density Q_I is very small. In the*y*-direction most of the band bending occurs at the drain junction. Since the value of φ_{surf} along the *y*-direction remains essentially constant over most of the length of the channel, the $\mathcal{E}$-field along the *y*-direction is also very small [i.e., $\mathcal{E}(y) = -\partial\varphi_{surf}/\partial y$ is very small]. The combination of a small lateral $\mathcal{E}$-field acting on a relatively small number of free carriers implies that the drift component of I_{Dst} is negligibly small in weak inversion. In addition, in long-channel MOSFETs, the *gradient* of the $\mathcal{E}$-field in the *y*-direction along most of the channel length must also be very small. This means it is valid to apply the GCA when calculating I_{Dst} in such devices. In short-channel MOSFETs, however, the gradient of $\mathcal{E}(y)$ can no longer be neglected, even in weak inversion, and thus the long-channel formulation of I_{Dst} given here may not yield accurate estimates of I_{Dst} in short-channel MOSFETs.

On the other hand, we shall show that the *gradient of the free carriers* along the *y*-direction in the channel in weak inversion can be relatively large. Since diffusion current is proportional to the carrier-density gradient, in such cases carrier diffusion can produce a significant I_{Dst}. In fact, diffusion is the dominant mechanism that establishes subthreshold current flow in long-channel devices. The expression for the diffusion current component is given by:

$$I_{diff} = A\, J_{diff} = A\, [q\, D_n\, dn\,(y)/dy] = Z\, t_{inv}\, [\, q\, D_n\, dn\,(y)/dy\,] \qquad (4 - 84)$$

where D_n is the electron diffusion constant, A is the cross-sectional area of the electron flow, and t_{inv} is the inversion layer thickness (or $A = Z\, t_{inv}$). We can express this equation in terms of the mobile charge density per unit area Q_I instead of in terms of *n* by multiplying the mobile charge density in the channel (qn) by the inversion layer thickness t_{inv}. As shown in section 4.3.1.1, t_{inv} can be approximated as $(kT/q)\mathcal{E}_{si}$ (x = 0), where $\mathcal{E}_{si}$ (x = 0) is given by $\sqrt{2qN_A\varphi_{surf}/\kappa_{si}\varepsilon_o}$, so

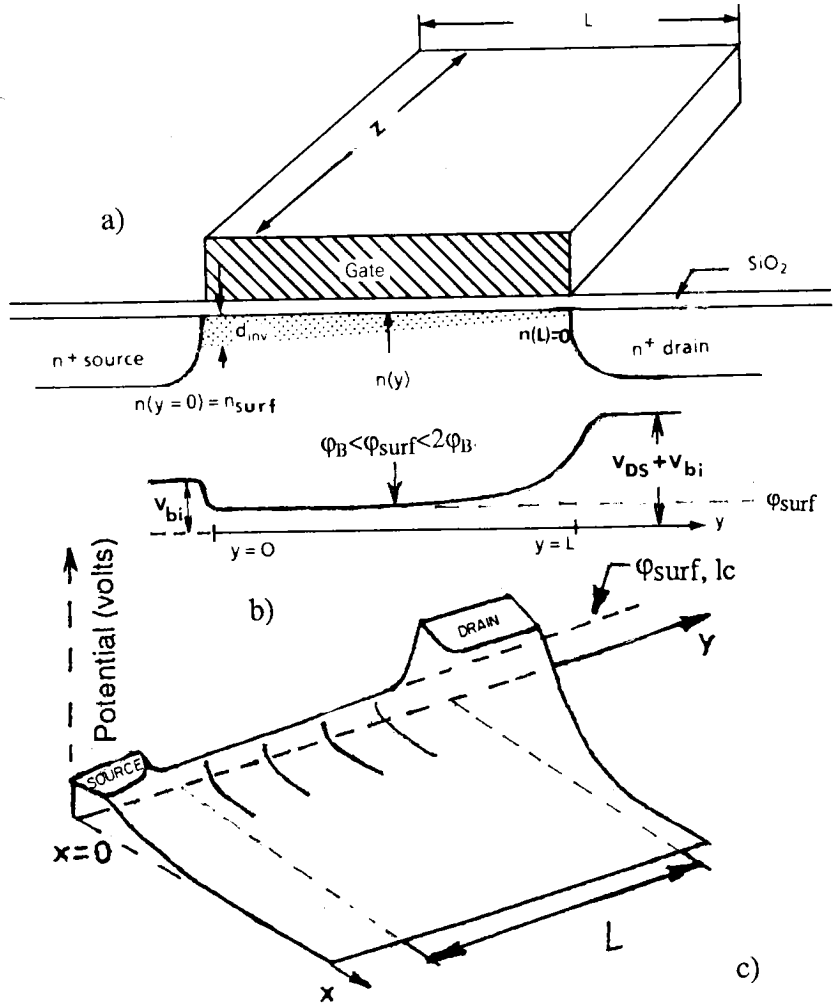


Fig. 4-26 (a) Cross-sectional area of an NMOSFET in weak inversion showing the variation of inversion layer charge concentration versus position along the channel; (b) Plot of $\varphi_{surf}(y)$ versus y for the device of part (a); (c) 2-D perspective plot of potential versus position in a long-channel NMOSFET in weak inversion.

$$t_{inv} \approx \frac{kT}{q} \frac{\sqrt{\kappa_{si}\varepsilon_o}}{\sqrt{2q\,N_A\,\varphi_{surf}}} \qquad (4 - 85)$$

In addition, the carrier concentration in the inversion layer can be found using Eq. 3-13 (with $n_{po} = n_i^2/N_A$), or $n = (n_i^2/N_A)\exp(q\varphi_{surf}/kT)$. Using these relationships, $Q_I(y)$ can be expressed as

By some algebraic manipulation this expression for Q_I can be shown to be identical to the equation for the mobile charge density per unit area Q_I in a MOS-C under weak inversion given in chap. 3 (i.e., Eq. 3-58, which we repeat here)

$$Q_I \text{ (weak inv)} \approx \frac{\sqrt{2q\, N_A \kappa_{Si} \varepsilon_o}}{2\sqrt{\varphi_{surf}}} \varphi_t\, e^{[\varphi_{surf}(GB) - 2\varphi_B]/\varphi_t} \qquad (4 - 86b)$$

where the surface potential in the MOS-C is established just by the gate-body bias V_{GB} (and $\varphi_t = kT/q$). We can include the effect of a bias between the channel and substrate in Eq. 4-85. That is, with the source grounded (i.e., $V_{BS} = 0V$, implying that $V_{DB} = V_{DS}$), Q_I at the source and drain ends of the channel in a MOSFET under weak inversion, can be written using Eq. 4-86a as follows:

$$Q_{I,source} \approx q \frac{kT}{q} \frac{\sqrt{\kappa_{Si} \varepsilon_o}}{\sqrt{2\, q\, N_A\, \varphi_{surf}(source)}} \frac{n_i^2}{N_A} \exp\left(\frac{q\, \varphi_{surf}(source) - 2\varphi_B}{kT}\right) \qquad (4 - 87a)$$

$$Q_{I,drain} \approx q \frac{kT}{q} \frac{\sqrt{\kappa_{Si} \varepsilon_o}}{\sqrt{2\, q\, N_A\, \varphi_{surf}(source)}} \frac{n_i^2}{N_A} \exp\left\{\frac{q\, [\varphi_{surf}(source) - 2\varphi_B - V_{DS}]}{kT}\right\} \qquad (4 - 87b)$$

where $\varphi_{surf}(GB) = \varphi_{surf}(source)$ in this case, as $V_{BS} = 0V$. Since the diffusion current is constant along the channel, the gradient of Q_I (i.e., dQ_I/dy) is also constant and can be approximated by the following expression, using Eqs. 4-87a and b:

$$\frac{dQ_I}{dy} = \frac{Q_{I,source} - Q_{I,drain}}{L} \approx \frac{Q_{I,source}}{L}\left[1 - \exp\left(\frac{-q\, V_{DS}}{kT}\right)\right] \qquad (4 - 88)$$

Then the expression for the diffusion component of the drain current in weak inversion I_{Dst} can be written as

$$I_{Dst} = D_n Z \frac{dQ_I}{dy} = \frac{D_n Z\, Q_{I,source}}{L}\left[1 - \exp\left(\frac{-q\, V_{DS}}{kT}\right)\right] \qquad (4 - 89)$$

Now, if $|V_{DS}| \geq 0.1V$, at room temperature the term $\exp(-qV_{DS}/kT)$ in Eq. 4-88 is smaller than e^{-4}. Hence, this term in Eq. 4-89 will be much smaller than 1. As a result, if V_{DS} exceeds a few φ_t (i.e., a few kT/q, or at 300K, $\geq 0.1V$), we can consider $Q_{I,drain}$ to be negligibly small compared to $Q_{L,source}$.

In such cases, the expression for I_{Dst} in long-channel MOSFETs can be approximated as $I_{Dst} \sim D_n Z Q_{Isource}/L$, or, in expanded form:

$$I_{Dst} \approx q\, D_n \frac{Z}{L} \frac{kT}{q} \frac{\sqrt{\kappa_{Si} \varepsilon_o}}{\sqrt{2\, q\, N_A\, \varphi_{surf}(source)}} \frac{n_i^2}{N_A} \exp\left(\frac{q\, \varphi_{surf}(source)}{kT}\right) \qquad (4 - 90)$$

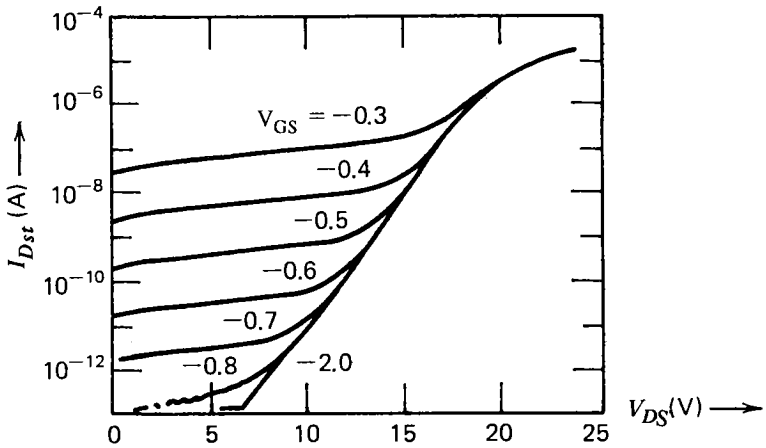


Fig. 4-27 Experimental low-current characteristics for a MOSFET with $L = 2.1$ μm, with the source and substrate both tied to ground. For lower values of V_{DS}, I_{Dst} is reduced by roughly one order of magnitude for each 0.1V reduction in V_{GS} below the threshold value. At higher values of V_{DS}, a subsurface current flows that is independent of V_{GS}.[24] (© IEEE 1974).

Since φ_{surf}(source) is roughly proportional to V_{GS} in the subthreshold region,[16] Eqs. 4-87 and 4-90 imply that $Q_{I,source}$ and I_{Dst}, respectively, are both exponentially dependent on V_{GS}.

In summary, Eq. 4-90 predicts that in long-channel MOSFETs I_{Dst} will exponentially increase with V_{GS} but will remain substantially independent of V_{DS}. Figure 4-27 illustrates this subthreshold behavior in a MOSFET with $L = 2.1$ μm. For values of V_{DS} smaller than the punchthrough voltage (i.e., roughly the V_{DS} value where the I_D curves intercept the more steeply rising portion of the curve in Fig. 4-27), I_{Dst} is seen to be essentially independent of V_{DS} and increases exponentially with V_{GS}.

It turns out, however, that devices with gate lengths of ~2 μm are at the transition between long-channel and short-channel device behavior. That is, when V_{DS} is sufficiently increased in MOSFETs with $L \leq 2$ μm, the source and drain depletion regions merge, and at that point I_{Dst} starts to become dependent on V_{DS} (as is also illustrated in Fig. 4-27 in the right side of the characteristics, where I_{Dst} no longer depends on V_{GS}). This short-channel effect, known as *punchthrough*, is discussed in chap. 5. In general, however, punchthrough must be prevented from occurring in normal device operation, even in short-channel devices. That is, at the maximum operating V_{DS} value the punchthrough current should be kept smaller than the long-channel I_{Dst} value. One simple rule of thumb to accomplish this in a MOSFET that employs a V_T-adjust implant is to set the substrate doping concentration N_{sub} at $>0.1N_I$, where $N_I = D_I/x_d$, and D_I and x_d are the threshold-adjust implant dose and the effective implant depth (or centroid), respectively.

It should also be noted that the V_T-adjust implant generally increases the subthreshold swing S_t, and this is one reason that S_t in real devices is larger (~100 mV/dec) than the theoretically predicted value of ~60 mV/dec at 300K. A more complete discussion of S_t is given in the upcoming section.

4.7.1 Subthreshold Swing S_t

Another important device characteristic in the subthreshold regime is the slope of log I_D versus V_{GS}. This characteristic indicates how effectively a MOSFET can be turned OFF as V_{GS} is decreased below V_T. The inverse of the slope is defined as the subthreshold swing S_t.[17] As voltage levels are reduced in an effort to save power and to use smaller devices, this characteristic becomes a limitation on how small a power-supply voltage can be used. In MOSFETs with uniformly doped substrates, S_t is calculated from

$$S_t = \ln 10(d\ln I_D/dV_{GS})^{-1} = 2.3\ (kT/q)\ (1 + C_d/C_{ox}) \qquad (4\text{ - }91a)$$

$$= 2.3\ (kT/q)\ [1 + \kappa_{si}t_{ox}/\kappa_{ox}d)] \qquad (4\text{ - }91b)$$

S_t is expressed in units of mV per decade. (The factor 2.3 comes from the conversion of ln(x) to 2.3 $\log_{10}[x]$). Let us consider the S_t parameter in more detail.

Ideally, an abrupt change in I_D should occur as V_{GS} passes through V_T. If a MOSFET exhibits a steep decline in I_D as V_{GS} is decreased below V_T, its S_t value will be small. Conversely, if a MOSFET structure is known to possess a small S_t, the implication is that only a small reduction of V_{GS} below V_T will effectively turn off the device; whereas if the device exhibits a large S_t value (meaning a gradual slope in log I_D vs V_{GS}) a significantly large I_{Dst} may still flow in the OFF state (e.g., when $V_{GS} = 0$).

According to Eq. 4-91b, at 300K in the ideal limit of t_{ox} --> 0, $S_t = (2.3kT/q)$ ~60 mV/dec. However, any nonidealities in an actual device (e.g., the presence of interface states, or a nonzero value of t_{ox}) cause S_t to become larger than the 60 mV/dec value (also, note that if T increases, so does the value of S_t). The S_t value of typical submicron NMOSFETs operating at 300K is ~100 mV/decade, and in such devices I_D will drop from 1 μA to 1 pA for a 0.6V decrease in V_{GS}. Thus, if a MOSFET exhibits a *constant-current-V_T* of 0.6V (i.e., $I_D = 1\ \mu A/\mu m$ when $V_{GS} = V_T = 0.6V$), when $V_{GS} =$ 0V, I_D will be reduced to 1 pA/μm. Thus, if S_t is known, Eq. 4-91a can be used to estimate I_{Dst} in long-channel MOSFETs in subthreshold operation (Fig. 4-28).[18] That is, circuit designers can readily calculate the gate bias required to keep the subthreshold leakage current below its specified maximum. Typically, to ensure that I_{Dst} will be negligibly small, the maximum bias applied to the gate when the device is in the OFF state should be kept at least 0.5 V below V_T.

Equation 4-91 also indicates that S_t can be made smaller by using a thinner t_{ox} or a lower substrate doping concentration (which will make the channel depletion-region width d larger). In addition, Eq. 4-91 implies that S_t increases with temperature (for example, at 100°C, the minimum theoretical value of S_t is increased to ~100 mV/dec).*

* Note that in a non-uniformly doped channel (e.g., one which uses a V_T-adjust implant), the expression for S_t is modified to

$$S_t = (kT/q) \ln 10\ [\ 1 + C_d/C_{ox}]/\{1 - [(\kappa_{si}/\kappa_{ox})(t_{ox}/L_D)]^2[C_d/C_{ox}]^2\} \qquad (4\text{ - }92)$$

Basically, this equation indicates that the V_T-adjust implant causes S_t to be somewhat increased (since the additional dopant in the channel reduces the depletion-region width, thus making C_d in Eq. 4-91 larger). However, if the implant is deep enough, the depletion region

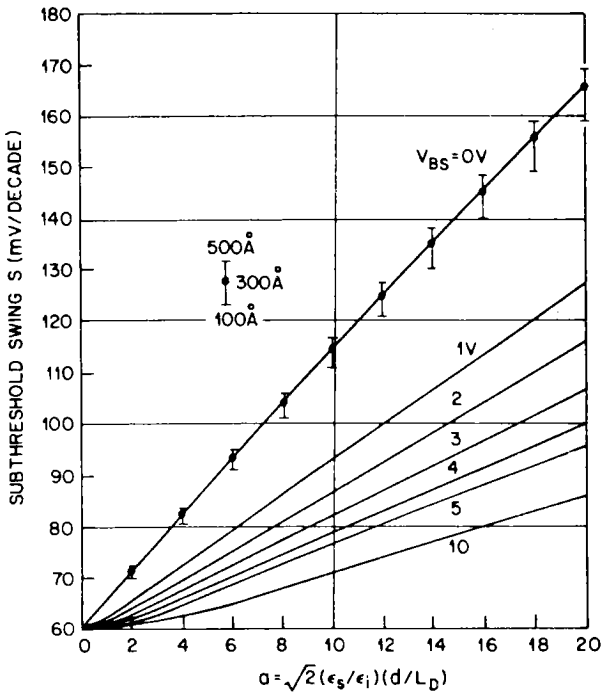


Fig. 4-28 Subthreshold gate swing S_t required to reduce the subthreshold current one decade versus the parameter $a = [\sqrt{2}\, t_{ox}/\kappa_{ox}q)][\kappa_{si}kT/\varepsilon_0 N_A)^{1/2}$. Vertical bars indicate the variation in S_t at constant a when the oxide thickness is varied from 500 to 100 Å, and doping level and mobility adjusted accordingly. $V_{DS} >> kT/q$. [13] (© IEEE 1979).

Finally, it should be noted that because the depletion width increases when a substrate bias is applied, the subthreshold swing decreases according to Eq. 4-91b. This effect is shown in Fig. 4-29.[17,24]

Note that when the channel length gets small, the values of I_{Dst} are larger than those predicted by the above model. This is due to so-called short-channel effects (discussed in chap. 5). As a result, measurements of S_t versus channel length are used to detect the onset of these short-channel effects (punchthrough and surface drain-induced barrier lowering). Since the measurable I_{Dst} is the sum of both the normal subthreshold and short-channel-induced components, an increase in the value of S_t will signal the onset of the latter.

* (cont.) width will eventually be equal to that of the unimplanted device, and S_t returns to the value exhibited by a device with no V_T-adjust implant. Nevertheless, the general dependences of S_t on t_{ox}, T, and channel substrate doping exhibited in devices with uniformly doped channels still apply.

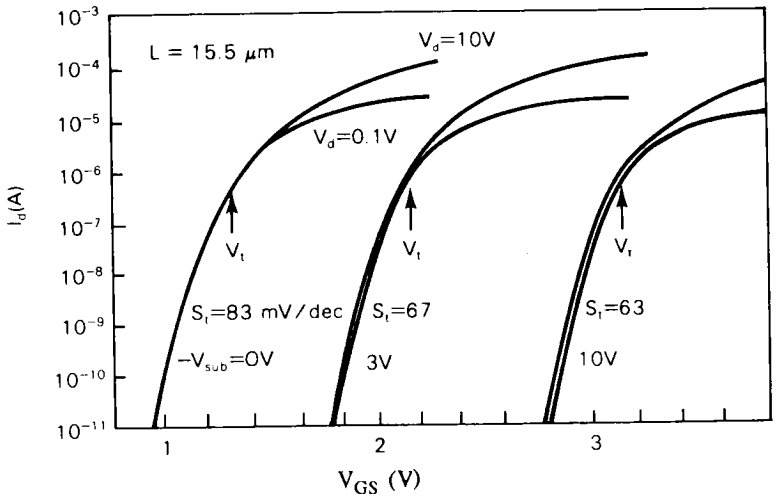


Fig. 4-29 Experimental subthreshold characteristics for a long-channel device (L = 15.5 μm).[24] (© IEEE 1974).

4.7.2 Gate-Induced Drain Leakage (GIDL)

Another form of leakage current observed in OFF-state MOSFETs is *gate-induced drain leakage (GIDL).*[19] The carriers responsible for GIDL originate in the region of the drain that is overlapped by the gate, and GIDL is occurs when the gate is grounded and the drain is at V_{DD}. A large electric field then exists across the oxide ($\mathcal{E}_{ox}$), which must be supported by charge in the drain region. This charge is provided by the formation of a depletion region in the drain as shown in Fig. 4-30a. If $\mathcal{E}_{ox}$ becomes sufficiently large, in addition to the depletion region, an inversion layer will attempt to form at the silicon surface of the drain. However, as the minority carriers arrive at the surface to form the inversion layer, they are immediately swept laterally to the substrate (which is a region of lower potential for the minority carriers). Hence, in this case, the depletion region under the oxide in the drain instead becomes a *deep depletion layer.** The zone near the surface where an inversion layer should be formed is referred to as an "incipient inversion layer." The current that flows as a result of the carriers being swept from this

* *Deep depletion* is a non-equilibrium condition where there are fewer minority carriers in a depletion region than would exist in equilibrium. Consider the case of an n^+MOS structure (i.e., the condition that exists in an NMOS drain region), that is biased so that an inversion layer of holes should exist at the surface (and a depletion layer exists below it). Under equilibrium conditions, this condition can persist indefinitely under appropriate dc bias. In the structure described above, a dc bias is applied that attempts to establish such an inversion layer of holes at the surface of the n^+drain. However, the potential of the substrate is lower for holes than the surface potential of the drain, and thus any holes arriving at the drain surface move laterally to the substrate region. Thus, the depletion region under the overlap area becomes a zone of deep depletion.

incipient inversion layer constitute the GIDL current.

The magnitude of GIDL current is independent of temperature, and hence it is believed that thermal generation of carriers in the incipient inversion region plays only a minor role in GIDL. Instead, it is proposed that when $\mathcal{E}_{ox}$ becomes large enough, the voltage drop in the deep-depletion layer becomes so large that various trap-assisted carrier-generation events become possible (Fig. 4-30b).[20] These include thermal emission of an electron from the valence band to a trap site (expressed as a vertical transition in the diagram) and then tunneling to the conduction band (horizontal transition), and trap-assisted tunneling from valence band to conduction band. For even larger fields (a few MV/cm), trap-free band-to-band tunneling may also occur (see also Fig. 4-30b).

Like other forms of leakage current, GIDL will contribute to standby power, giving rise to the previously mentioned problems of excessive heat dissipation at large device counts and heavy current drain in portable systems. Hence, GIDL must be controlled so that it does not exceed some specified maximum value (typically, ≤10 pA/μm).

In order to control GIDL, the most important mechanisms that may cause it need to be identified. This allows processes involving these mechanisms to be avoided or modified to minimize GIDL. For example, it has been observed that shallow junction

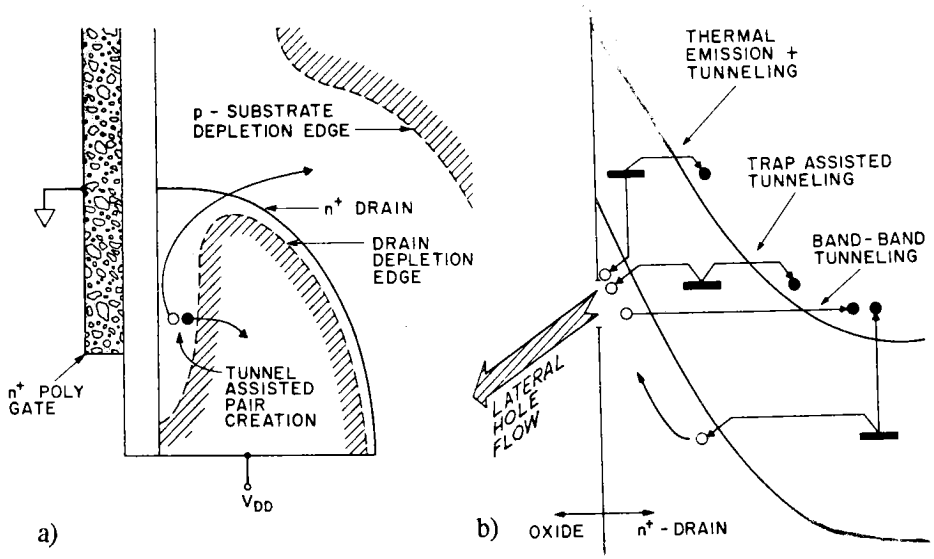


Fig. 4-30 (a) Gate induced drain leakage in the MOSFET. A schematic view of the gate-drain overlap region for a grounded gate and the drain biased at V_{DD}. Tunneling created pairs lead to a lateral hole flow in the n^+ drain. This flow prevents formation of an inversion layer inside the drain. (b) Various mechanisms that can contribute to gate-induced drain leakage. Vertical transitions are thermal-emission processes. Horizontal transitions are tunneling processes. Holes are prevented from forming an inversion layer by lateral transport to the substrate. This flow is normal to the plane of the diagram, as indicated by the slanted arrow, lateral hole flow.[20] From J.R. Brews, "The Submicron MOSFET," Chap. 3, in *High-Speed Semiconductor Devices*, Ed. Sze. Copyright 1990, John Wiley & Sons. Reprinted with permission.

fabrication using Ge preamorphization introduces bulk midgap traps that increase GIDL.[21] Other reports indicate that interface traps caused by hot carrier injection,[22] or Fowler-Nordheim tunneling[23] increase GIDL by stimulating additional tunneling.

There are several general procedures that appear to be effective in reducing GIDL. First, the oxide thickness can be increased to reduce $\mathcal{E}_{ox}$ for a given voltage. But this is usually not implemented, since a thicker gate oxide will adversely impact other device characteristics. Second, the trap density in the near-surface area can be reduced, but such trap elimination generally requires a carefully controlled fabrication sequence. Third, the doping in the drain can be increased, as this will decrease the depletion region width and the tunneling volume. Unfortunately, the latter approach favors abrupt junctions, because graded or lightly-doped-drain (LDD) structures will lead to a significant lateral extension of the drain within which the doping is lighter, inviting greater GIDL. Tradeoffs of field reduction against GIDL may need to be considered. On the other hand, it has also been reported that LDD devices can suppress GIDL by suppressing the lateral field.

Nevertheless, it has been reported that GIDL sets a limit on the power supply voltage V_{DD} of an LDD MOSFET of[25,29]

$$V_{DD} = 4 \times 10^6 \text{ V/cm} \times t_{ox} + 1.2\text{V} - V_{FB} \qquad (4 - 93)$$

A non-LDD MOSFET would have a much higher GIDL current. This implies that suppression of GIDL may be a strong reason to retain the LDD structure in MOSFETs even after V_{DD} is reduced to suppress hot-carrier effects. (Note that in Eq. 4-93 V_{FB} is ~0V for an n^+ polysilicon gate on an n^+ drain and about 1.1V for an n^+ polysilicon gate over a p^+ drain.)

4.8 SUMMARY OF LONG-CHANNEL MOSFET BEHAVIOR

The behavior of the long-channel as detailed in this chapter can be summarized by examining the I-V characteristics of the long-channel MOSFET in strong inversion (Fig. 4-31a), and in the subthreshold region of operation (weak inversion, see Fig. 4-31b).

In strong inversion we plot the output characteristics of the MOSFET (I_D versus V_{DS} with V_{GS} as the parameter) on an axis with linear I_D and V_{DS} scales. We see that for a fixed value of $V_{GS}>V_T$, the drain current first increases linearly with V_{DS}. Eventually, as shown in Fig. 4-31a, the drain current saturates. Beyond the value of saturation voltage (i.e., when $V_{DS}>V_{DSsat}$), I_D remains constant with increasing V_{DS}. The value of I_{Dsat} is also proportional to $(V_{GS} - V_T)^2$ - as predicted by charge-control model. The values of I_D versus V_{DS} and V_{GS} in the linear region, as well as the values of I_{Dsat} and V_{DSsat} can be calculated using any one of the four long-channel MOSFET models given in section 4.5, depending on the accuracy desired.

In the subthreshold region of operation, the behavior is described by plotting I_D (with a logarithmic scale) versus V_{GS} (with a linear scale). As shown in Fig. 4-31b, in long-channel MOSFETs the drain current I_D drops exponentially once V_{GS} becomes ~0.2V below V_T. Furthermore, in long-channel MOSFETs the slope of this log I_D-V_{GS}

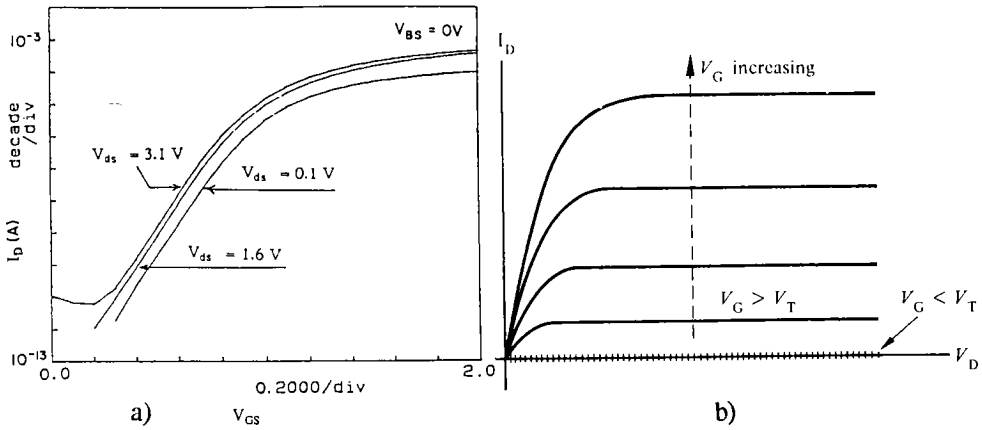


Fig. 4-31 (a) I_D versus V_{DS} characteristics of long-channel MOSFETs beyond strong inversion. (b) I_D versus V_{GS} characteristics of a long-channel MOSFET below threshold (weak inversion).

characteristic does not change with increasing V_{DS}, nor does the value of I_D increase at a fixed value of V_{GS} as V_{DS} is increased.

Finally, the threshold voltage of the long-channel MOSFET does not vary as L is decreased, nor does it change with increasing V_{DS}, although this V_T behavior is not indicated by either Fig. 4-31a or Fig. 4-31b.

REFERENCES

1. D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits,* McGraw-Hill Book Co., New York, 1983.
2. S. M. Sze, *Semiconductor Devices - Physics and Technology*, John Wiley & Sons, New York, 1985, p. 444.
3. S. Selberherr, A. Schutz, and H.W. Potzl, "MINIMOS-A two dimensional MOS analyzer," *IEEE Trans. Electron Dev.,* **ED-27**, p. 1540, 1980.
4. H.C. Pao and C.T. Sah, "Effects of Diffusion Current on Characteristics of Metal-Oxide (Insulator)-Semiconductor Transistors," *Solid-State Electron.* **9**, 927, (1966).
5. Y.P. Tsividis, *Operation and Modeling of the MOS Transistor,"* McGraw-Hill, New York, 1987.
6. R.F. Pierret and J.A. Shields, "Simplified long-channel MOSFET Theory," *Solid-State Electronics,* vol. 26, p. 143, 1983.
7. G. Baccarani et al., "Analytical i.g.f.e.t.model including drift and diffusion currents," *IEE Journal on Solid-State and Electron Devices,* vol. 2, p. 62, March 1978.
8. J.R. Brews, "A Charge-Sheet Model of the MOSFET," *Solid-State Electronics,* vol. 21, p. 345, 1978.
9. H.K. Ihantola and J.L Moll, "Design Theory of a Surface Field-Effect Transistor," *Solid-State Electronics* **7**, 423 (1964).

10. A. Vladimirescu, A.R. Newton, and D.O. Pederson, *SPICE Version 2G.2 Users Guide,* Electronics Research Laboratory, University of California, Berkeley.

11. R.S. Muller and T.I. Kamins, *Device Electronics for Integrated Circuits,* 2nd. Ed., New York, John Wiley & Sons, 1986.

12. V.L. Rideout, F.H. Gaensslen, and A. LeBlanc, "Device Design Consideration for Ion-implanted *n*-Channel MOSFETs," *IBM J. Res. Dev.,* p. 50, (1975).

13. J.R. Brews *IEEE Trans. Electron Dev.*, **ED-26(11)**, 1696 (1979),

14. D. Antoniadis "Calculation of threshold voltage in nonuniformly doped MOSFETs," *IEEE Trans. Electron Dev.*, **ED-31,** p. 303, March 1984.

15. J.R. Brews, "The Physics of the MOS Transistor," in *Silicon Integrated Circuits*, Part A, Ed. D. Kahng, New York, Academic Press, 1981.

16. S. M. Sze, *Semiconductor Devices: Physics and Technology,*, John Wiley & Sons, New York, p. 213, 1985.

17. J.R. Brews, J. "Subthreshold Behavior of Uniformly and Nonuniformly Doped Long-Channel MOSFET," *IEEE Trans. Electron Dev.,* **ED-26** (9) p. 1282, Sept. 1979.

18. *ibid.* , (**16**), p. 470.

19. T. Y. Chan et al., 1987 IEDM Tech. Digest, p. 718.

20. J.R. Brews, "The Submicron MOSFET," chap. 3 in *High-Speed Semiconductor Devices,* S.M. Sze Ed., Wiley- Interscience, New York, 1990.

21. D.-S. Wen et al., "Tunneling Leakage in Ge-PreAmorphized Shallow Junctions," *IEEE Trans. Electron Dev.,* **ED-35** (7), 1107 (1988).

22. C. Duvvury et al., "Leakage Current Degradation in *n*-MOSFETs Due to Hot-Electron Stress," *IEEE Electron Dev. Lett.* **EDL-9** (11), 579 (1988).

23. I. C. Chen et al., "Interface-Trap Enhanced Gate-Induced Leakage Current in MOSFET," *IEEE* Electron Dev. Lett., **EDL-10**, (5), 216 (1989).

24. R.R. Troutman, "Subthreshold Design Considerations for IGFETs," *IEEE J. Solid-State Circuits,* **SC-9**, p. 55, (1974).

25. S. Parke et al., "Design for suppression of gate-induced leakage in LDD MOSFETs."

26. *ibid.,* reference **16** p. 444.

27. V.L. Rideout, F.H. Gaensslen, and A. LeBlanc, *IBM J. Res. Dev.,* **19**, p. 50, (1975).

28. N.D. Arora, *Solid-State Electronics*, **30**, p. 559, (1987).

29. H.-J Wann, P.K. Ko, and C. Hu, "Gate-Induced Band-to-Band Tunneling Leakage Current in LDD MOSFETs," *Tech. Dig. IEDM,* 1992, p. 147.

30. R.F. Pierret, *Field Effect Devices*, Vol. IV Modular Series on Solid-State Devices, Addison-Wesley Publishing Co. Reading, MA., 1983.

PROBLEMS

Fig. 4-15 Small element of the inversion layer in a long-channel MOSFET. p. 108 Tsiv -change W to Z, φ_{surf} (x) to φ_{surf} (x + Δx), and $\Delta\varphi_{surf}$.5 Reprinted with permission of McGraw-Hill Book Co.

PROBLEMS

1. Another way to model the drain current in a MOSFET is to note that it can be represented as mobile charge induced in the channel divided by the *transit time* (or average time required for an electron to move from the source to drain). For an *n*-channel MOSFET with a 250Å oxide, 1 μm channel length, 10 μm channel width, and $V_T = 1V$, operated at $V_{GS} = 5V$ and $V_{DS} = 0.1V$, calculate the total electron charges in the channel and the electron transit time. Then, show that this charge divided by the transit time is equal to the current calculated at the linear region as given by Eq. 4-65.

2. Evaluate the coefficient *k* of Eq. 4-64 for a MOSFET in which $\mu_n = 580$ cm^2/V sec, Z/L = 1 and (a) $t_{ox} = 50$ nm, and (b) $t_{ox} = 10$ nm.

3. Account qualitatively for the differing drain current I_D predictions of the bulk-charge and square-law models of the long-channel MOSFET.

4. Describe how the saturation voltage V_{DSsat} is defined and derive $V_{DSsat} \sim V_{GS} - V_T$ as a first order expression.

5. The sheet resistances of the source (R_S) and the drain (R_D) are both parasitic resistances (see Appendix B). Discuss why R_S decreases I_D more than does R_D in the saturation region of operation. Also, derive an analytic expression of g_m in the linear region as a function of R_S and R_D.

6. If a circuit designer wants to ensure that a MOSFET exhibits no more than no more than 1 pA of leakage current in the off state, and also wants $V_T = 0.9V$ (where V_T is defined as the gate voltage that causes 1 μA current), what maximum S_t value must the device exhibit.

7. Verify that the expression for the capacitance of the surface space-charge region Cd in Eq. 4-91 is correct (i.e., that $C_d = \kappa_{si}\varepsilon_0/d$).

8. The I_D-V_{GS} characteristics for a long *n*-channel MOSFET are shown in Fig. 4P-1. Using the square-law model and information available from the graph, compute the drain current I_D when $V_{GS} = 3V$ and $V_{DS} = 5V$.

9. A MOSFET has been fabricated on a *p*-type substrate with doping concentration $N_A = 4\times10^{15}$ cm^{-3}. The gate oxide thickness is 50 nm, the gate is n^+ polysilicon, the oxide charge is $Q_{tot} = 2\times10^{10}$ cm^{-2}, and V_T-adjust implant of 2×10^{12} boron atoms/cm^2 is performed through the gate oxide. Assume that Z/L = 1 and $\mu_n = 800$ cm^2/V sec, and $V_{BS} = 0V$. (a) Calculate the threshold voltage of the device; and (b) Calculate V_{DSsat} and I_{Dsat} for the MOSFET using: (i) the bulk-charge model and; (ii) the square-law model. Take $V_{GS} = 5V$. Comment on the accuracy of the square-law model.

10. Explain why we can approximate the thickness of the inversion layer as the depth beneath x = 0 at which $\varphi(x)$ has dropped to a value of $0.38\varphi_{surf}$.

11. Explain why I_D does not become zero if $V_{GS}>V_T$ and $V_{DS}>V_{DSsat}$, as might be the conclusion based on the long-channel approximation used to calculated I_D.

12. Derive the I-V characteristics of a MOSFET with the drain and gate connected together and the source and substrate grounded. Can one obtain the threshold voltage from these characteristics?

13. An NMOSFET with $Z/L = 15$, $t_{ox} = 80$ nm and $\mu_n = 600$ cm^2/V sec is to be used as a controlled resistor: (a) What charge density is required for the device to exhibit a dc resistance of 3.5 kΩ between the source and drain in the saturation region? (b) What value of $(V_{GS} - V_T)$ is needed to obtain the desired resistance?

14. An Al gate enhancement-type NMOSFET has a substrate doping of $N_A = 6\text{x}10^{15}$ cm^{-3} and a gate oxide thickness of 100 nm, and $Q_{tot} = 5\text{x}10^{10}$ cm^{-2}. Determine V_T of this device. When the device is operated in the linear region with $V_{DS} = 0.5$V, calculate the gate voltage needed to obtain a drain current of 2 mA. Assume $L = 10$ μm, $Z = 100L$, and $\mu_n = 500$ cm^2/V sec.

CHAPTER 5

MOS TRANSISTOR DEVICE PHYSICS: PART 3

THE SUBMICRON MOSFET

The two major goals of MOSFET scaling are to increase the *density* and *speed* of the digital ICs in which such scaled-down devices are used. Increasing *density* obviously means using smaller channel lengths and widths. To increase the *speed* of digital ICs, the MOSFET saturation drain current I_{Dsat} must be increased (i.e., to allow faster charging and discharging of parasitic capacitances). From the long-channel MOSFET models given in chap. 4, a decrease in either the channel length L or gate oxide thickness t_{ox} will lead to an increase in I_{Dsat}. In fact, the long-channel MOSFET models predict that I_{Dsat} should continue to increase indefinitely as L and t_{ox} are decreased, seeming to imply that only the limitations of process technology (and not device effects) prevent the manufacture of even smaller, higher-performing MOSFETs. In addition, the MOSFET models of chap. 4 also accurately describe many other device characteristics of long-channel MOSFETs, including threshold voltage, subthreshold current, and the I_D-V_{DS} characteristics in strong inversion.

However, as process technology improved to the point where devices could be fabricated with gate lengths smaller than ~2 μm, it turned out that MOSFETs began to exhibit phenomena not predicted by the long-channel MOSFET models. Such phenomena were thus termed *short-channel effects.* One of the more surprising of such short-channel effects (which becomes especially pronounced as L decreases below 0.5 μm) is that the drain current in saturation I_{Dsat} shows far less increase as L is decreased than is predicted by the long-channel models (see Fig. 5-1).[1,47] In fact, in the simple model to be presented in section 5.4, I_{Dsat} is predicted to become independent of L in extremely small MOSFETs, approaching instead, a constant value given by $Q_I v_{sat} Z$ (where v_{sat} is the saturation velocity of the carriers, as described in section 5.4.1.2). In contrast, reducing t_{ox} (which has the effect of increasing the oxide field, and in turn causes Q_I to increase), yields a considerably greater increase in I_{Dsat}. Furthermore, as shown in Fig. 5-1 I_{Dsat} will also increase more rapidly as L is decreased if the MOSFET has a thinner gate oxide. This benefit provides even a greater impetus for making t_{ox} as thin as possible as gate lengths are decreased further in the drive for higher density.

Since the quest for higher density still requires L (and Z) to be further reduced, it will nevertheless be necessary to confront the other short-channel effects (besides the

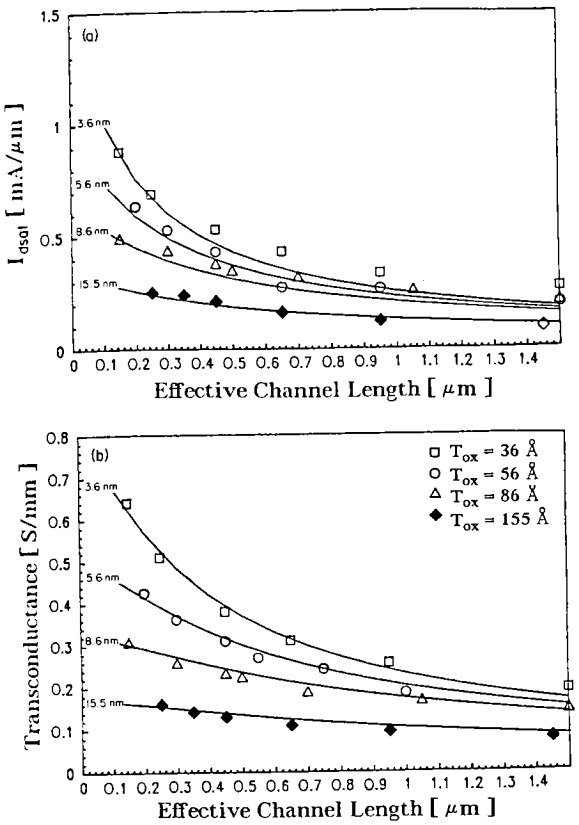


Fig. 5-1 Measured and calculated I_{Dsat} for an array of deep submicron NMOS transistors with four different gate-oxide thicknesses.[47] (© 1987 IEEE).

one of the I_{Dsat} dependence on L mentioned above). These other short-channel effects can be classified into two groups: (1) increased leakage current when the MOSFET is "off," and; (2) reliability problems associated with short-channel device structures.

Increased "off-state" leakage in short-channel MOSFETs is due to several phenomena, including: (a) lowering of the threshold voltage V_T as L is decreased and/or V_{DS} is increased; (b) the onset of punchthrough at smaller drain biases as L is decreased, and; (c) an increase in isolation leakage current as the isolation spacing is decreased. We will discuss V_T lowering and punchthrough in this chapter, and isolation leakage phenomena in chap. 6.

The reliability problems that arise in short-channel MOSFETs include: (a) thin gate-oxide breakdown (to be discussed in chap. 7); (b) device degradation due to hot-carrier effects (to be discussed in chap. 9); and (c) reliability problems associated with the interconnects between MOSFETs (e.g., electromigration failures in the metal lines). The latter problem is treated in Vol. 2, chaps. 3 and 4.

Table 5-1. Comparison of Long-Channel and Short-Channel MOSFET Characteristics

Long-Channel MOSFET Behavior	Short-Channel MOSFET Behavior
1. The threshold voltage V_T is independent of channel length L and width Z.	1. V_T decreases as L is decreased. V_T may also be impacted by changes in Z.
2. V_T is independent of drain bias voltage.	2. V_T decreases with increasing V_{DS}.
3. V_T depends on V_{BS} according to Eq. 4-12 (with V_{BC} in Eq. 4-12 replaced by V_{BS}).	3. V_T increases less rapidly with V_{BS} than predicted by Eq. 4-12.
4. The subthreshold current I_{Dst} increases linearly as L decreases.	4. I_{Dst} increases more rapidly than linearly as L decreases.
5. I_{Dst} is independent of drain bias.	5. I_{Dst} increases with increasing V_{DS}.
6. The subthreshold swing S_t is independent of gate length.	6. S_t increases with decreasing L.
7. The drain current in saturation I_{Dsat} is independent of V_{DS}.	7. I_{Dsat} increases as V_{DS} increases.
8. I_{Dsat} is proportional to $(V_{GS} - V_T)^2$	8. I_{Dsat} is proportional to $(V_{GS} - V_T)$.
9. I_{Dsat} is proportional to $1/L$.	9. As $L \rightarrow 0$, I_{Dsat} becomes independent of L.

In this chapter we will also cover several device design scenarios developed to keep leakage currents sufficiently small as the MOSFET is scaled down. In addition, we describe the issues related to submicron PMOSFET design.

5.1 DEVICE CHARACTERISTICS OF LONG-CHANNEL VERSUS SHORT-CHANNEL MOSFETS

From the perspective of MOSFETs as circuit components, long-channel behavior has been specified in a number of ways. As a result, short-channel device effects can be so designated when a deviation from long-channel MOSFET behavior can be correlated with a reduction in the gate length (and/or gate width) dimension. Table 5-1 gives a list of long-channel device characteristics that undergo variation as the gate dimensions are decreased. As the behavior of the short-channel MOSFET is described later in this chapter, the terms associated with these characteristics will be defined and discussed in more detail.

From Table 5-1 we can see that the short-channel effects can be divided into the following categories: (a) those that impact V_T (Items 1-3): (b) those that impact subthreshold currents (Items 4-6); and (c) those that impact I_D when the MOSFET is operated in saturation [i.e., $V_{DS}>(V_{GS} - V_T)$], namely Items 7-9. In this chapter we

discuss these effects in detail, as well as measures that can be adopted to minimize their impact on MOSFET behavior.

5.2 EFFECT OF GATE DIMENSIONS ON THRESHOLD VOLTAGE

In this section we will describe three short-channel effects on the threshold voltage of MOSFETs, namely: (1) the short-channel threshold-voltage shift; (2) Narrow gate width effects on threshold voltage; and (3) the reverse short-channel threshold voltage shift.

5.2.1 Short-Channel Threshold-Voltage Shift

Experimentally, it is observed that as the dimensions of the gate are reduced, the threshold voltage of MOSFETs become less well predicted by the long-channel V_T equation (Eq. 4-74). The error becomes significant as the *channel length dimension L* is reduced to less than 2 μm. Usually, V_T decreases as L is reduced.* In addition, in short-channel MOSFETs V_T also decreases as V_{DS} is increased. To get good agreement with measured data, a term ΔV_{Tsc} must be subtracted from the value of V_T obtained from Eq. 4-74. (In *n*-channel enhancement-mode MOSFETs the measured value of V_T becomes less positive than that predicted by Eq. 4-74, while in *p*-channel enhancement-mode MOSFETs, V_T becomes less negative.)

The decrease of V_T with L and V_{DS} in short-channel devices is crucial because enhancement-mode FETs in CMOS are generally designed to operate with V_T magnitudes of 0.6-0.8V. If the magnitude of V_T drops even slightly below its designed value, the device may exhibit excessive drain leakage current when V_{GS} = 0V. As discussed earlier, V_T values of 0.6-0.8V in long-channel NMOS devices with lightly doped substrates can only achieved by increasing the doping concentration at the surface of the channel (i.e., with a V_T-adjust implant). If short-channel effects reduce V_T below the long-channel values, the channel doping concentration must be increased even more to re-establish the desired V_T. Higher channel doping, however, generally degrades other device characteristics, such as carrier mobility and off-current. Thus, it is important to be able to model ΔV_{Tsc} to facilitate reliable tradeoffs between process and electrical parameters and leakage current levels.

The failure of Eq. 4-74 to give good agreement with measured V_T values in short-channel MOSFETs arises because this equation was derived assuming one-dimensional theory. That is, Eq. 4-74 assumes the space charge under the gate is controlled only by the vertical $\mathcal{E}$-field, $\mathcal{E}_x$. In that case, the channel depletion region charge will be influenced only by the charge on the gate. When the channel of a MOSFET is long, this is a reasonable assumption, as the influence of the drain and source junctions (and the influence of V_{DS}) on the quantity of channel charge can be neglected (Figs. 5-2a and b).

* In section 5.2.3, however, we will see that V_T may initially increase as L shrinks, before ultimately decreasing as described here. This unexpected effect is referred to as a *reverse short-channel effect* (RSCE), and will be described in section 5.2.3.

However, as L approaches the dimensions of the source/drain junction depletion-region widths, a greater part of the channel-depletion region begins to consist of the space charge in the junction depletion regions (Fig. 5-2c and d). Hence, less gate charge (and a smaller vertical $\mathcal{E}$-field) is needed to cause inversion in short-channel MOSFETs than in long-channel MOSFETs (assuming each has the same substrate doping). This makes it appear that a smaller V_T is needed to turn on a short-channel device. Furthermore, as the reverse bias applied to the drain (V_{DS}) is increased, the drain depletion region grows wider, leaving even less charge under the control of the gate.

There are four approaches that have been taken to establish quantitative values of ΔV_{TSC}. They are the following:

1. *Numerical analysis* is used to find solutions of the two-dimensional Poisson equation for calculating the electric field and potential in the channel region.
2. Analytical expressions for ΔV_{TSC} are derived using the *charge sharing model.*
3. Analytical expressions for ΔV_{TSC} are derived using the drain-induced barrier lowering model, in which the 2-D Poisson equation is simplified to a one-dimensional form, making it solvable with analytical functions.
4. Analytical expressions for ΔV_{TSC} are derived using the drain-induced barrier lowering model in which a *quasi 2-D* approach is employed to analytically solve the Poisson equation.

While the numerical solution approach is the most accurate, this technique is not feasible for circuit simulation and statistical modeling. As a result, substantial effort has been expended to develop analytical models which provide rapid, inexpensive, and sufficiently accurate simulation of ΔV_{TSC}. Hence, we devote most of our discussion to the latter three approaches.

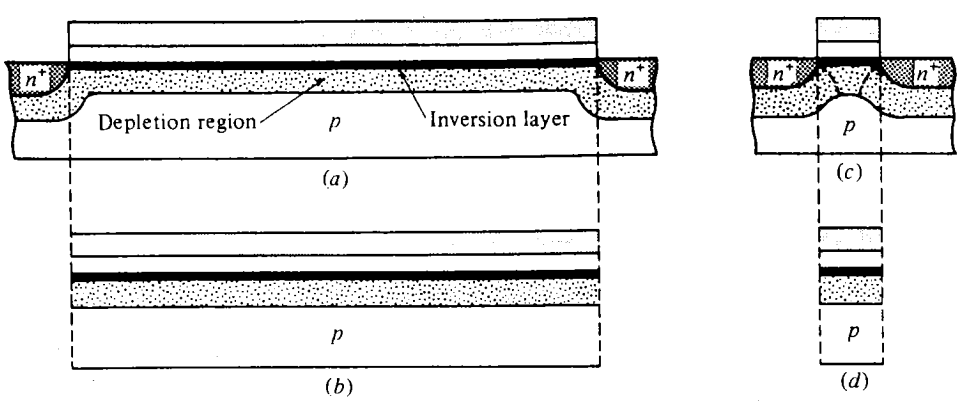


Fig. 5-2 (a) Long-channel transistor; (b) the channel of (a) with edge effects neglected; (c) a short-channel transistor; (d) the channel of (c) with edge effects neglected. From Y.P Tsividis, *Operation and Modeling of the MOS Transistor*, McGraw-Hill, New York, 1987. Reprinted with permission.

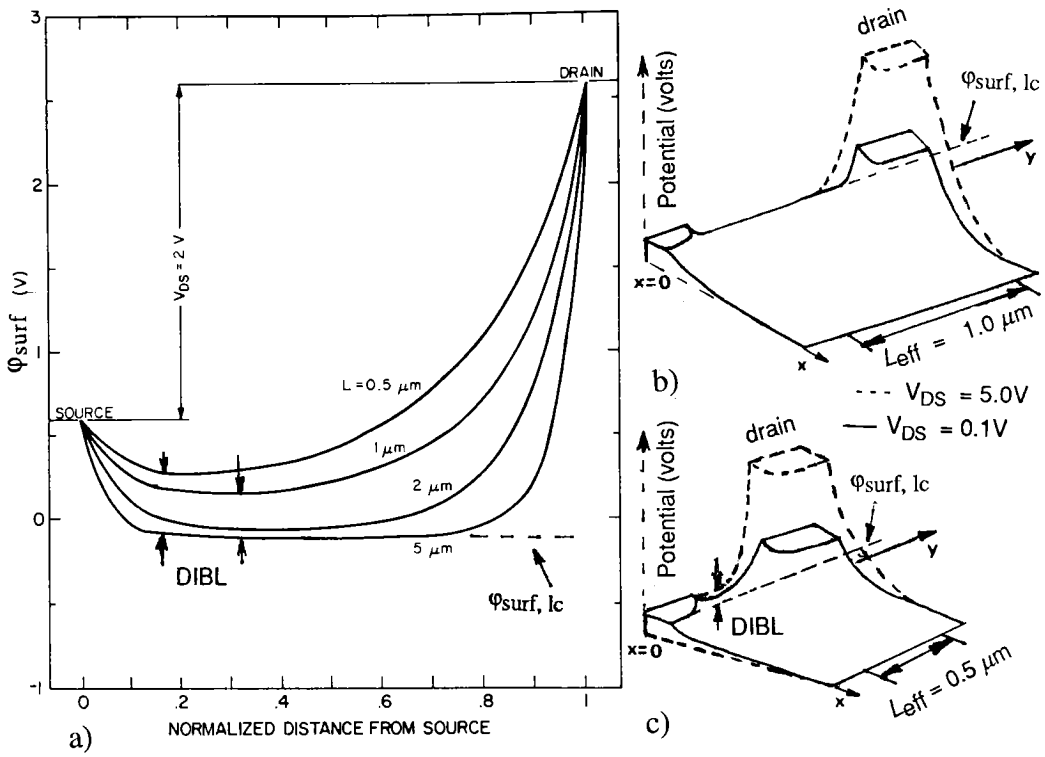


Fig. 5-3 Simulated values of the surface potential φ_{surf} along the channel for NMOSFETs with different channel lengths (note channel length is normalized). $N_A = 10^{15}$ cm^{-3}, t_{ox} = 500Å, $r_j = 0.33\mu$m, V_{GS} = 0V, V_{DS} = 2.0V, V_{BS} = 0V.[109] (© 1979 IEEE). (b & c) 2-D potential variation for NMOSFETs with (b) $L_{eff} = 1.0\ \mu$m, and (c) $L_{eff} = 0.5\ \mu$m (see also Fig. 5-5). Solid line, V_{DS} = 0.1V and dashed line, V_{DS} = 5.0V.

5.2.1.1 Numerical Solution of the 2-D Poisson Equation for Calculating ΔV_{TSC}. The most accurate method for simulating this effect is to solve the 2-D Poisson equation (i.e., as given here)

$$\frac{\partial^2\varphi(x, y)}{\partial y^2} + \frac{\partial^2\varphi(x, y)}{\partial x^2} = -\frac{\rho(x, y)}{\kappa_{si}\,\varepsilon_o} \qquad (5 - 1)$$

for the electrostatic-potential φ and the electric-field $\mathcal{E}$ in the channel region. While exact analytical solutions of the 2-D Poisson equation are not possible, numerical methods can be used to obtain solutions of Eq. 5-1. As noted in chap. 2, several software programs have been developed which allow these methods to be implemented.[2,21] Specifically, such numerical Poisson solvers are available in device simulators (e.g., GEMINI, MINIMOS, and PISCES-II). However, numerical solutions of 2-D partial differential equations are computationally intensive (and hence time-consuming and expensive), and therefore as noted, are not suitable for use in circuit simulation applications in which the electrical behavior of large circuits must be simulated. Nevertheless, they serve as accurate benchmarks of ΔV_{TSC} against which the

values derived from the analytical models can be compared. Thus, they are a valuable tool for helping to develop more accurate analytical ΔV_{TSC} models. Figure 5-3 shows such numerically calculated values of φ_{surf} versus y in NMOSFETs operated in subthreshold as L is decreased from 5.0 μm to 0.5 μm.

5.2.1.2 Charge-Sharing Models. The *charge-sharing* perspective assumes that not all of the electric-field lines emanating from the charge under the gate terminate on the gate charge. Instead, some terminate on space charge in the source and drain depletion regions. The depletion charge under the gate is thus fractionally established by the source and drain (and in this sense, the channel charge is considered to be "shared" among the gate, source, and drain charge, see Fig. 5-4). Hence, less gate charge is required to cause inversion, making V_T appear smaller. The fraction of the charge induced by the source and drain becomes significant as the channel length is of the order of the junction depletion region widths. Charge-sharing models seek to quantitatively determine the fractional charge induced by the gate.

In 1974, Yau proposed a simple, widely used analytical model based on the *charge-sharing* perspective.[3] This model assumes that the charge controlled by the gate is

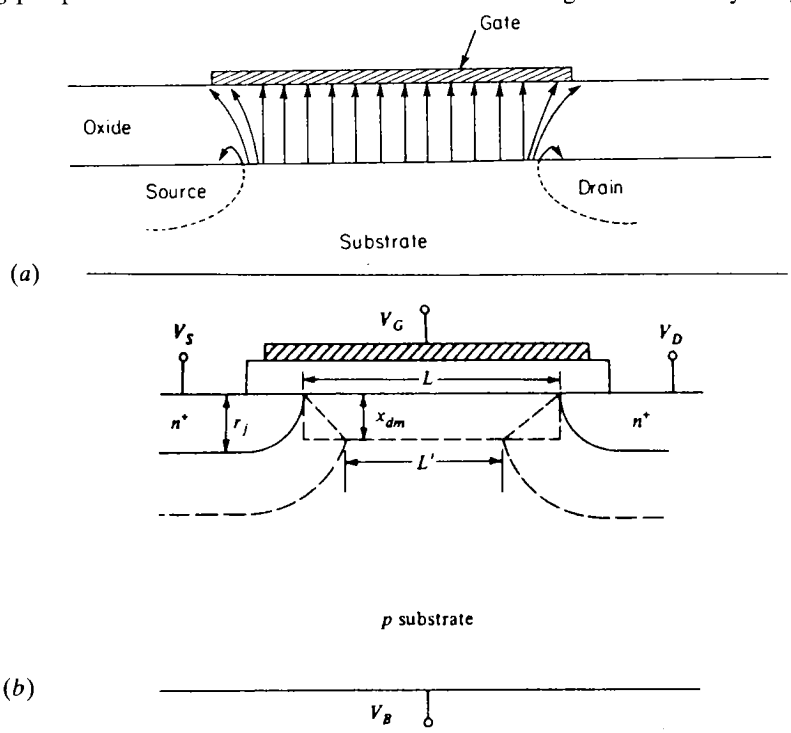


Fig. 5-4 (a) Schematic showing perturbation of the electric field lines in the oxide near the source and the drain.[5] (© 1985 IEEE). (b) Yau's model of charge sharing.[3] Reprinted with permission of Solid-State Electronics.

reduced to the charge in the trapezoidal region of the channel shown in Fig. 5-4b. The resulting analytical expression based on this assumption allows the value of ΔV_{Tsc} to be calculated in devices with uniformly doped channels:

$$\Delta V_{Tsc} = \frac{q\, N_{SUB}\, d_{max}\, r_j \left(\sqrt{\left\{ 1 + \frac{2\, d_{max}}{r_j} \right\}} - 1 \right)}{C_{ox}\, L} \qquad (5 - 2)$$

where d_{max} is the maximum width of the depletion region in the channel, and r_j is the junction depth of the source/drain regions. According to Eq. 5-2, as L decreases, ΔV_{Tsc} increases. Eq. 5-2 also implies that to keep ΔV_{Tsc} as small as possible, shallow source/drain junctions and thin gate oxides should be used ($C_{ox} = \varepsilon_{ox}/t_{ox}$). In addition, the channel depletion-region width should also be as small as possible. Although it will be seen that some of the assumptions used in Yau's model are generally invalid, the qualitative guidelines it yields for minimizing ΔV_{Tsc} appear to be correct as channel lengths are shrunk, regardless of the model employed for predicting ΔV_{Tsc}. Yau's model also permits qualitative insight of V_T lowering with decreasing L to be gained, even if the quantitative ΔV_{Tsc} values do not accurately reflect experimental observations.

By modifying Eq. 5-2, Yau's model can also be used to estimate the effect of drain voltage on ΔV_{Tsc}. That is, in Eq. 5-2 it is assumed that the width of the source, drain, and channel depletion regions are equal. However, under drain bias the drain depletion region grows wider. Thus these three depletion regions no longer have equal widths. To reflect the effect of drain-width variation on ΔV_{Tsc}, Eq. 5-2 is reformulated as follows:[4]

$$\Delta V_{Tsc} = \frac{q\, N_{sub}\, d_{max}\, r_j \left(\left[\sqrt{\left\{ 1 + \frac{2\, W_S}{r_j} \right\}} - 1 \right] + \left[\sqrt{\left\{ 1 + \frac{2\, W_D}{r_j} \right\}} - 1 \right] \right)}{2\, C_{ox}\, L} \qquad (5 - 3)$$

Despite the appealing simplicity of the charge-sharing models, they fail to give good quantitative agreement with measured V_T values in devices with very short channels (e.g., < 1 μm) or under large drain voltages. This apparently occurs because several of the assumptions used to derive the model are not valid, especially under short-channel and large V_{DS} conditions. These include the following:

1. In Yau's model the induced charge is arbitrarily apportioned among the gate, source, and drain, whereas an accurate apportionment can only be accomplished by numerical integration of Poisson's equation and graphical analysis.

2. In the long-channel MOSFET model we assume that in subthreshold operation the entire V_{DS} is dropped across the drain depletion region. Thus, the channel potential in the rest of the channel region is essentially independent of V_{DS} and depends only on the gate bias. This also implies φ_{surf} in this region is independent of position (i.e., its value is constant with position along the channel). However, as we shall see this assumption breaks down in the case of short-channel MOSFETs (and as is evident from the data shown in Fig. 5-3). Yau's model,

however, continues to assume that the potential in the channel region is constant with position. The portion of the charge that is shared is assumed to be distributed evenly under the gate to uniformly reduce the depletion charge density in the gate area. Obviously, this is not what actually happens physically.

These and other questionable assumptions of Yau's model are examined in ref. 5.

Nevertheless, despite the fact that charge-sharing models generally do not provide accurate quantitative results, they are still useful for providing first-order estimates of ΔV_{TSC}, as well as for helping to intuitively visualize the physical basis of changes in V_T as L is decreased.

5.2.1.3 Barrier-Lowering Models and DIBL. The *barrier-lowering* perspective introduced by Troutman in 1979, provides a more rigorous foundation for deriving analytical models of ΔV_{TSC}.[6] It seeks to determine the change in the potential energy barrier between the source and channel at the Si interface caused by the junction depletion region charges (which in turn arise from both the built-in *pn* junction $\mathcal{E}$-field and the reverse-bias drain voltage) as well as the gate charge.

Troutman made measurements of DIBL,[6] and he and others[1,2,6] performed numerical analysis of the Poisson equation (Eq. 5-1) to match experimental and theoretical results. As noted earlier, numerical analysis is necessary to obtain more accurate results. However, because of the extensive computational procedures required, this solution approach is not appropriate for applications involving computer-aided circuit design. As a result, other workers sought to develop approximate analytical solutions of Eq. 5-1 which would be suitable for such applications.

This is done by finding analytical models of the surface potential φ_{surf} which relate it to the MOSFET structural parameters (e.g., t_{ox}, r_j, L), substrate doping levels, and applied biases. The value of φ_{surf} from these models is used in the threshold voltage equation to calculate the modified V_T value. The mathematical approach is to try and discover approximate analytical solutions to the 2-D Poisson equation, generally by converting it to a form that contains a differential term in only one dimension. This approach may be fruitful in this case, since we are only interested in finding the minimum value of φ at the surface (i.e., of φ_{surf}) along the y-direction from source to drain, i.e., we want to determine $[\varphi_{surf}(y)]_{min}$. In most cases analytical solutions with approximate boundary conditions are proposed as the model. The accuracy depends on the degree of approximation in the boundary conditions.

Many DIBL-based analytical models have been developed. In general, they are more complex than the charge-sharing models. One of the simplest and most widely adopted is that of Hsu et al.[7] In this model, the 2-D Poisson equation is reduced to a 1-D equation by essentially approximating the $(\partial^2\varphi/\partial x^2)$ term of the 2-D equation as a constant. With the Poisson equation thereby converted into a 1-D form, it becomes possible to find an analytical solution, and Hsu's solution becomes the model for calculating ΔV_{TSC}. Jain and Balk extend this model to include the effect of non-uniform channel width.[8] They also point out that both the Hsu model and their extension are expected to be accurate only for small values of DIBL (i.e., for devices having channels that are not too short and in which V_{DS} is not too large). In addition, these models were derived for MOSFET structures with uniformly doped channel regions.

More recently, Skotnicki et al., reported another approach which allows the 2-D Poisson equation to be reduced to a 1-D equation using a so-called *voltage-doping transformation.* Using this technique they derived an analytical barrier-lowering model that is applicable to MOSFETs with non-uniformly doped channels.[9] Good agreement with measured data is also reported for devices with L as small as 0.8 μm (and in which V_{DS} values are as large as 3V). A final approach, as described in the next section, also provides good agreement with measured data over the ranges of voltages studied.

Nevertheless, the accuracy of the barrier-lowering-based analytical models has not yet been confirmed for predicting the value of ΔV_{Tsc} in all device structures. As a result, it may still be prudent to use empirical measurements and/or numerical analysis to obtain accurate estimates of barrier lowering in specific device structures.

5.2.1.4 Quasi-Two-Dimensional Analytical Solution for ΔV_{Tsc}. By using a quasi-2-D approach to solve the 2-D Poisson equation in the channel of the MOSFET, an analytical model has been derived that reportedly predicts ΔV_{Tsc} accurately, even for devices with channel lengths below 0.5 μm.[22] It begins by approximating the 2-D Poisson equation in the channel with the expression:

$$\frac{\varepsilon_{si}\, d_{max}}{\eta} \frac{d\mathcal{E}_y(y)}{dy} + \varepsilon_{ox} \frac{[V_T - V_{FB} - \varphi_{surf}(y)]}{t_{ox}} = q\, N_{sub}\, d_{max} \qquad (5\text{-}4a)$$

or

$$\frac{\varepsilon_{si}\, d_{max}}{\eta} \frac{d^2\varphi_{surf}(y)}{dy^2} + \varepsilon_{ox} \frac{[V_T - V_{FB} - \varphi_{surf}(y)]}{t_{ox}} = q\, N_{sub}\, d_{max} \qquad (5\text{-}4b)$$

where the parameters of this equation have been defined previously (except for η, which is a fitting parameter). The second term on the left-hand side is an approximation of the term $\partial\mathcal{E}_x/\partial x$ in the 2-D form of the Poisson equation. This quasi-2-D approach is also used in deriving a model for I_{Dsat} and thus, the details involved in transforming Eq. 5-1 into a form akin to Eq. 5-4 are given in section 5.4.3.

The solution to Eq. 5-4, namely $\varphi_{surf}(y)$, under the boundary conditions $\varphi_{surf}(0) = V_{bi}$ and $\varphi_{surf}(L) = V_{DS} + V_{bi}$ is:

$$\varphi_{surf}(y) = V_{sL} + (V_{bi} + V_{DS} - V_{sL}) \frac{\sinh(y/l)}{\sinh(L/l)} + (V_{bi} - V_{sL}) \frac{\sinh(L - y)/l)}{\sinh(L/l)} \qquad (5\text{-}5)$$

In Eq. 5-5 $V_{sL} = V_{GS} - V_T$, V_{bi} is the built-in potential between the source-substrate and drain-substrate junctions, and l is the characteristic length defined as

$$l = \sqrt{\frac{\varepsilon_{si}\, t_{ox}\, d_{max}}{\varepsilon_{ox}\, \eta}} \qquad (5\text{-}6)$$

Equation 5-5 can now be plotted for a sufficient number of points between $y = 0$ and $y = L$ (e.g., 10 or more). From such plotted data, and a fitted curve, the minimum value of $\varphi_{surf}(y)$ can be found. By subtracting the long-channel value of φ_{surf} at V_T from $\varphi_{surf}(y)_{min}$, we can find ΔV_{Tsc}. Figure 5-5 plots Eq. 5-5 for φ_{surf} at a given V_{GS}, and for two values of V_{DS} [i.e., $V_{DS} = 0.05V$ (dashed lines) and $V_{DS} = 1.5V$ (solid lines)] for

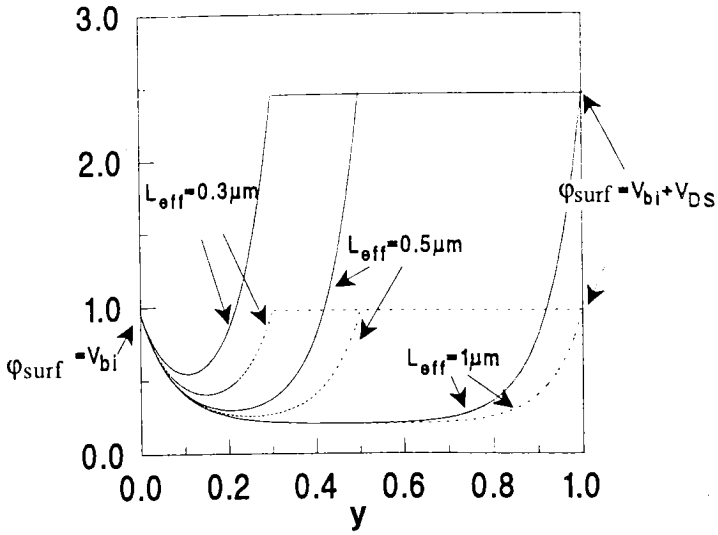


Fig. 5-5 (a) Calculated surface potential along channel for different channel lengths using Eq. 5-5. The device parameters are t_{ox} = 100Å, N_{sub} = N_A = 10^{16} cm^{-3}, n^+ = 10^{20} cm^{-3}, and η = 1, i.e., l = 0.1 μm.[22] (© IEEE 1993).

MOSFETs with different channel lengths. In this example, the device with L_{eff} = 1 μm has a region where φ_{surf} is constant as a function of position along a large fraction of the channel length (i.e., at a value of φ_{surf} ~0.2V in this figure), and this value is determined merely by the gate bias (and not V_{DS}). A constant value of φ_{surf} versus y is indicative of long-channel behavior and V_T in this device can thus be accurately found using Eq. 4-74.

In Fig. 5-5, devices with channel lengths 0.5 μm or smaller have a minimum value of φ_{surf} greater than the long-channel φ_{surf} value (i.e., $|2\varphi_B|$), and no region of constant φ_{surf} (y) is evident, even for small values of V_{DS}. (This is different than the charge sharing approach, in which a region of constant [but somewhat larger] φ_{surf} versus distance is assumed to exist even in very short length devices.) The results of Fig. 5-5 have been verified by 2D numerical simulation using such programs as MINIMOS and PISCES (e.g., as shown in plots of φ_{surf} versus y for various L values, see Fig. 5-3).

The minimum value of $\varphi_{surf}(y)$ is also seen in Fig. 5-5 to increase (i.e., the potential energy barrier for electron flow between source and drain will decrease) with decreasing channel length and increasing V_{DS}. If the minimum long-channel value of φ_{surf} is subtracted from the minimum value calculated from Eq. 5-5 we get an expression for ΔV_{Tsc}. When V_{DS} is small, this expression can be approximated as

$$\Delta V_{Tsc} = [2\,(V_{bi} - 2\varphi_B) + V_{DS}]\,(e^{-L/2l} + 2e^{-L/l}) \qquad (5-7)$$

Figure 5-6a shows the value of ΔV_{Tsc} versus L as calculated with Eq. 5-7 as well as with 2D numerical simulation. Agreement with the numerical simulation results is seen to be very good. It should also be noted that if $L > 5l$, then this equation reduces to a single exponential term, much like the one derived from the simpler barrier-lowering

models described in the previous section.* For most technologies l is about 0.1-0.15 μm, hence these simple models give accurate results for ΔV_{TSC} only if the channel length exceeds 0.5-0.8 μm.

When V_{DS} is not small, the expression for ΔV_{TSC} from this model is more complex, with its value when $L > 5l$ given approximately as

$$\Delta V_{TSC} = 3(V_{bi} - 2\varphi_B) + V_{DS})\, e^{-L/l} + 2\sqrt{(V_{bi} - 2\varphi_B)(V_{bi} - 2\varphi_B + V_{DS})}\, e^{-L/2l} \quad (5 - 8)$$

Eq. 5-8 reduces to Eq. 5-7 for large L/l and small V_{DS} as expected. Figure 5-6b shows the results of calculating V_T versus V_{DS} for $L_{eff} = 0.3$ μm with Eqs. 5-7 and 5-8 as well as with 2D numerical simulation. Eq. 5-8 is seen to compare well with the 2D numerical simulation results, but Eq. 5-7 overestimates ΔV_{TSC} at high V_{DS}. In Fig. 5-7 the value of V_T is calculated versus L using the charge sharing model, 2D numerical simulation, and Eq. 5-8. The results using Eq. 5-8 and numerical simulation are comparable. Finally, Fig. 5-8 compares the values of V_T versus L (with V_{BS} as the parameter) calculated with Eq. 5-8 and with measured V_T values in non-LDD

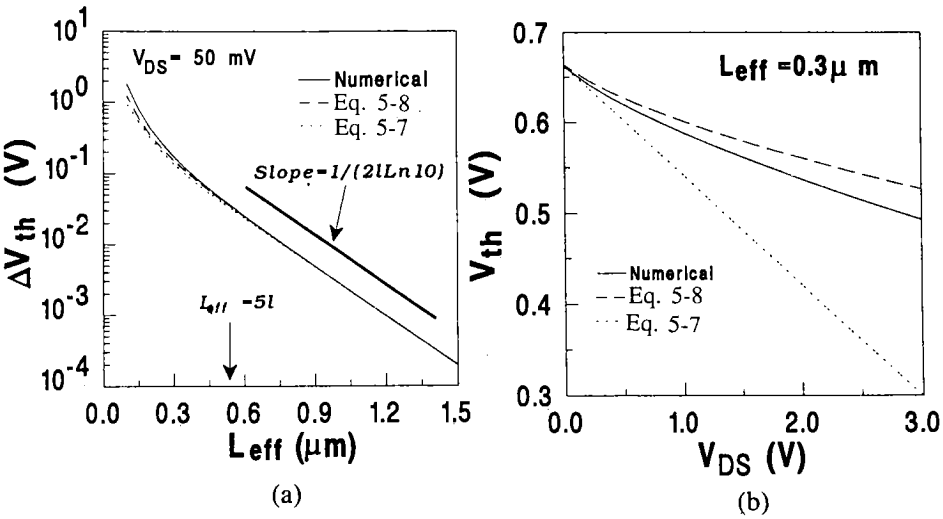


Fig. 5-6 (a) The calculated ΔV_T shifts versus channel length at V_{DS} = 0.005V. Note that the values obtained from the simple analytical models agree well with the numerical solution. The device parameters are the same as those in Fig. 5-5. Note that when $L > 5l$, all the curves have the same slope of l ($2l$ in Eq. 5-8). (b) Comparison between simple analytical solutions and numerical solution of threshold voltage versus drain voltage. The solution based on $y_O = L/2$ (Eq. 5-7) overestimates the variation in threshold voltage.[22] (© 1993 IEEE).

* If $L = 5l$ then the first exponential term in Eq. 5-7 is $e^{-2.5}$ while the second term is e^{-5}, which is much smaller than $e^{-2.5}$. Thus, for $L>5l$, the second exponential term in Eq. 5-7 can be neglected, leaving an expression with only one exponential term. For cases where L is smaller than ~$5l$, the second exponential term cannot be neglected, and the full Eq. 5-7 must be used to get accurate values of ΔV_{TSC} when V_{DS} is small.

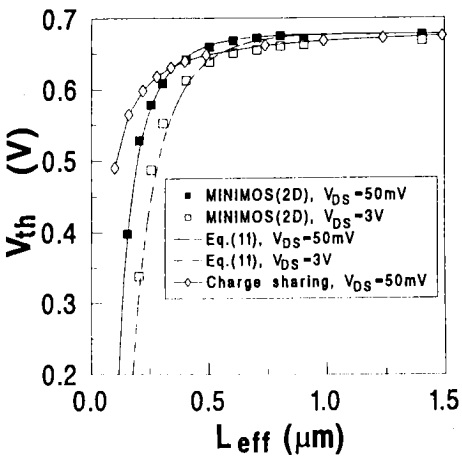


Fig. 5-7 Comparison of the V_T calculated using the charge sharing model, the two-dimensional numerical simulation (using MINIMOS), and the model of ref. 22. The device parameters are the same as those used in Fig. 5-5.[22] (© 1993 IEEE).

MOSFETs. Again, agreement is very good. The report describing this model also describes its application to LDD MOSFETs.

The characteristic length l defined in Eq. 5-6 needs to be more exactly determined in order for this model to give accurate values of ΔV_{TSC}. That is, due to the fitting parameter η in Eq. 5-6, the value of l from Eq. 5-6 only gives its correct order of magnitude. The exact value of l needs to be extracted from V_T data taken from actual MOSFETs.

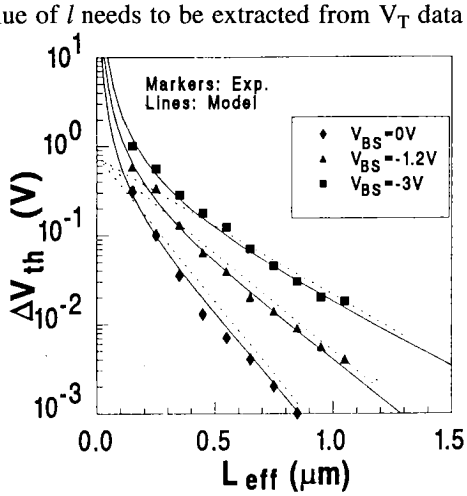


Fig. 5-8 ΔV_T versus L_{eff} at V_{DS} = 0.005V and different V_{BS} values for an non-LDD device. The solid lines are calculated results and the dashed lines are lines best fitting the experimental data of $L_{eff} >> 5l$. [22] (© 1993 IEEE).

An alternative form to Eq. 5-6 for expressing l which allows its exact value to be empirically extracted is based on the empirical expression of Brews et al[23] for the minimum channel length L_{min} a MOSFET must have in order to exhibit acceptable subthreshold behavior (see section 5.6.1.2 for more details on this expression):

$$L_{min} = 0.41\ (r_j\ t_{ox}\ d_{max}^2)^{1/3} \qquad (5 - 9)$$

If we assume L_{min} is about $4l$, using Eq. 5-9 yields an empirical expression for l as

$$l = 0.1\ (r_j\ t_{ox}\ d_{max}^2)^{1/3} \qquad (5 - 10)$$

According to this model, for an n-channel MOSFET with an n^+ poly gate, it can be shown that in order to maintain V_T at 0.7V, i.e.,

$$V_T = 0.7V = V_{FB} + 2\varphi_B + \frac{t_{ox}}{\varepsilon_{ox}} \frac{4\ \varepsilon_{si}\ \varphi_B}{d_{max}} \qquad (5 - 11)$$

it is necessary that

$$d_{max} = \frac{t_{ox}}{\varepsilon_{ox}} \frac{4\ \varepsilon_{si}\ \varphi_B}{(V_T - V_{FB} - 2\varphi_B)} \approx \frac{2\ \varepsilon_{si}}{\varepsilon_{ox}}\ t_{ox} \qquad (5 - 12)$$

By substituting Eq. 5-12 into Eq. 5-10, l can be rewritten as

$$l = 0.0007\ r_j^{1/3}\ t_{ox} \qquad (5 - 13)$$

where r_j and l are in μm and t_{ox} is in Å. On the other hand, for an n-channel MOSFET with a p^+ poly gate,

$$d_{max} = 4\ (\varepsilon_{si}/\varepsilon_{ox})\ t_{ox} \qquad (5 - 14)$$

(assuming $V_T = 1.2V$). Therefore one has $l \sim 0.0011\ r_j^{1/3} t_{ox}$. This helps explain why buried-channel devices, either NMOSFETs or PMOSFETs generally have worse short-channel effects than comparable surface-channel N- or PMOSFETs.

5.2.1.5 Measuring Surface-DIBL Effects. Surface-DIBL in MOSFET structures with a fixed L can be empirically characterized by measuring and plotting a set of log I_{Dst} versus V_{GS} curves with increasing values V_{DS} as the parameter (as shown in Fig. 5-9). When analyzing such a set of curves, it must first be ascertained that subsurface-punchthrough current is not flowing (see section 5.3 for a discussion on punchthrough). For example, in Fig. 5-9 there is no indication that such punchthrough current occurs over the applied V_{DS} range, since S_t remains constant as V_{DS} is increased.*

Once it is verified that punchthrough current is not significant, a quantitative estimate of surface-DIBL can be extracted from the shift in gate voltage ΔV_{GS} at a fixed drain current (typically 0.1 μA/μm) as the drain-source voltage is increased. That is, the

* If a device exhibits punchthrough current within the normal operating V_{DS} range, the structure must be redesigned to eliminate this behavior. Devices exhibiting punchthrough will exhibit excessive leakage current at $V_{GS} = 0V$ (i.e., when the device is nominally OFF).

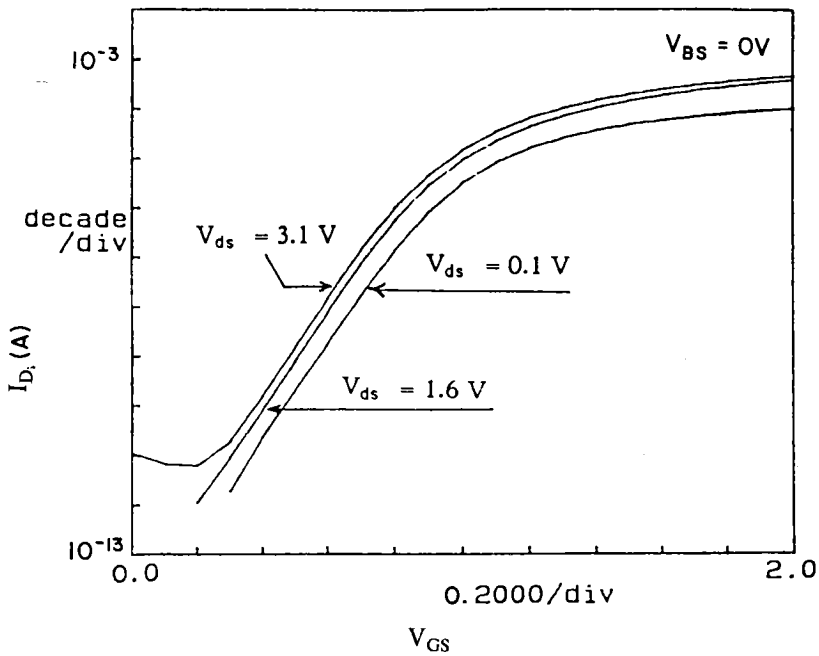


Fig. 5-9 Plot of log I_D versus V_{GS} for an NMOSFET operated below threshold, with V_{DS} as the parameter. The value of $\Delta V_{GS}/\Delta V_{DS}$ for a fixed value of I_D (e.g., 1×10^{-7} A/μm) is used to characterize DIBL.

ΔV_{GS} value represents a good first-order estimate of the expected shift in V_T as a function of drain bias. The shift in V_{GS} is normally expressed in normalized form as $\Delta V_{GS}/\Delta V_{DS}$. An acceptable value of $\Delta V_{GS}/\Delta V_{DS}$ is ~25 mV/V, depending on the circuit application and specific leakage requirements. $\Delta V_{GS}/\Delta V_{DS}$ is a strong function of channel length (Fig. 5-10) and must be guaranteed to at least 3-σ below the target channel length.

An alternative method of evaluating DIBL is to measure the change in drain current as V_{DS} is increased at some fixed gate voltage below V_T (e.g., the change ΔI_D in Fig. 5-10 when V_{DS} is increased from 0.1 [solid line in Fig. 5-10] to 5.0V [dashed line]). Because of its simplicity, this measurement is easily implemented on automated test equipment.

Another perspective on this last technique for measuring surface-DIBL is that such barrier lowering raises the electron density at the surface in subthreshold operation. That is, the electron density at the source end of the channel n_S(source) - where the surface potential is at a minimum - is increased due to surface-DIBL from its long-channel device value $n_S = (n_i^2/N_A)\exp(q\varphi_{surf}/kT)$ to

$$n_{SDIBL} = \left(\frac{n_i^2}{N_A}\right) e^{q\varphi_{surf}/kT}\, e^{q\Delta V_{TSC}/kT} \qquad (5\text{-}15)$$

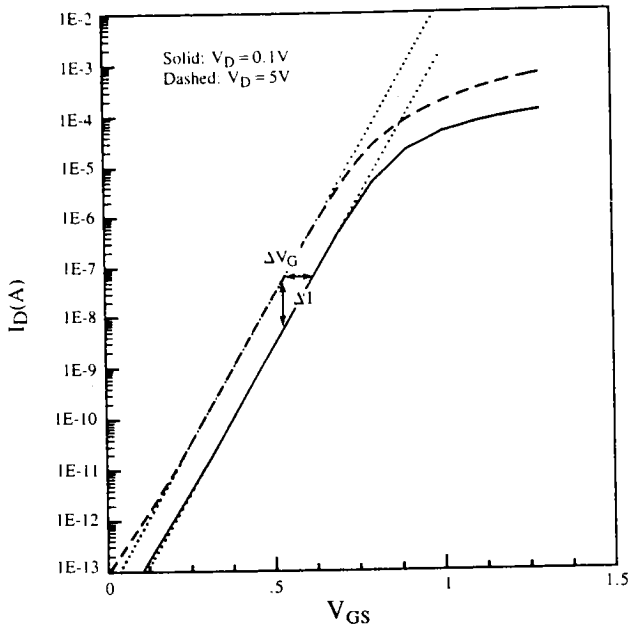


Fig. 5-10 Simulated values of $\Delta V_{GS}/\Delta V_{DS}$ (in mV/V) versus L_{eff}. From J. Teplik, Chap. 2, "Device Design," in *BiCMOS Technology and Applications*, A. Alvarez Ed., Klewer Academic Publishers, Boston MA, 1989. Reprinted with permission.

where ΔV_{TSC} will become larger as V_{DS} is increased. This also implies that surface-DIBL will cause I_{Dst} at a fixed value of V_{GS} and L to increase - by a factor of $\exp(q\Delta V_{TSC}/kT)$ - above that expected in a long-channel device.* Experimental data can be plotted to depict this effect in the following way: As shown in Fig. 5-11, log I_D is plotted versus V_{DS} with V_{GS} as the parameter. Figure 5-11a shows such characteristic curves for an NMOS device with $L = 2.1$ μm and Fig. 5-11b for a device with $L = 0.8$ μm. In both figures the curves exhibit two distinct regions. At low values of V_{DS}, I_D depends primarily on V_{GS} and increases only slightly with V_{DS} (*but it is this change in I_D that represents surface-DIBL*).** Eventually the I_D curves intercept a more steeply

* Note, however, that as V_{GS} is decreased, the value of n_s at the source will decrease as quickly with decreasing V_{GS} as in a long-channel device, because, the gate still has control of the charge giving rise to I_D. So, S_t will not change, even though V_T is smaller. But, when the device is in punchthrough, the gate no longer has the same control of charge flow as V_{GS} is reduced. Hence, drain current doesn't decrease as quickly with shrinking V_{GS}, or S_t gets larger in punchthrough.

** Note that according to the model of subthreshold behavior in long-channel MOSFETs at a fixed value of V_{GS}, I_{Dst} is independent of the drain voltage V_{DS} for V_{DS}>0.1V (see Eq. 4-90). But because of surface-DIBL, for a fixed value of L and V_{DS}, I_{Dst} increases as V_{DS} increases, as described in earlier sections.

rising curve, and for V_{DS} values beyond this point I_D is no longer influenced by V_{GS}, but instead rises quickly as V_{DS} is increased. The I_D behavior in this latter region is indicative that subsurface-DIBL (punchthrough) has set in, as discussed in section 5.3.

5.2.1.6 Dependence of V_T on V_{BS} in Short-Channel MOSFETs. In long-channel, uniformly doped MOSFETs the threshold voltage depends on the body-substrate voltage V_{BS} as given by Eq. 4-12 (repeated here for convenience)

$$V_T = V_{FB} + 2\varphi_B + \gamma(2\varphi_B + V_{BS})^{1/2} \qquad (5 - 16)$$

In short-channel MOSFETs, however, the effect on V_T when a substrate bias V_{BS} is applied is not well modeled by this expression. That is, as shown in Fig. 5-12, at shorter channels or higher drain biases, V_T is less sensitive to V_{BS} than predicted by Eq. 5-16. For example, in the case of $L = 1.5\ \mu m$ and $V_{DS} = 8V$ shown in Fig. 5-12, V_T becomes almost independent of V_{BS} at larger values of V_{BS}. In addition, for the case of $L = 0.7\ \mu m$, V_T becomes independent of V_{BS} for all values of V_{BS}.[10] This implies that the substrate tends to lose control of the channel (just as does the gate itself), at shorter channel lengths. The break in the slope of the $L = 1.5\ \mu m$ curves in Fig. 5-12 occurs with punchthrough of the depletion regions of the source and drain (see sec. 5.3 for details on punchthrough). This is because punchthrough decouples the edge of the depletion region from the gate, thereby decoupling the channel from the substrate.

The data shown in Fig. 5-12 also implies that this effect can be modeled numerically by using a smaller body effect coefficient γ (defined in Eq. 4-8a as $\gamma = [\kappa_{Si}\varepsilon_0/qN_A]^{1/2}$)

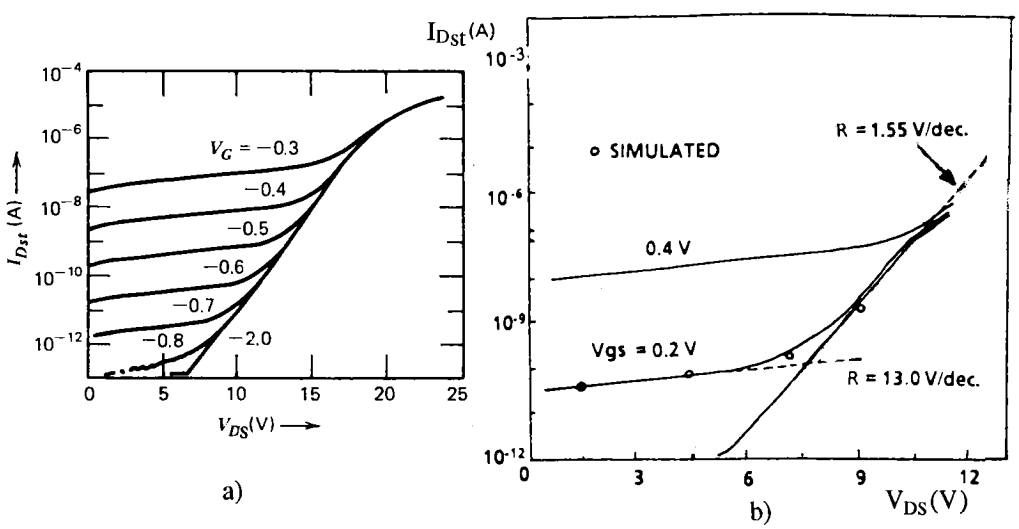


Fig. 5-11 (a) Experimental I_{Dst}-V_{DS} characteristics for a MOSFET with $L = 2.1\ \mu m$, $V_S = V_B = 0V$. For smaller values of V_{DS}, I_{Dst} is reduced by roughly one order of magnitude for each 0.1V decrease in V_{GS} below V_T. At higher values of V_{DS}, a subsurface punchthrough current flows that is independent of V_{GS}.[111] (© 1974 IEEE). (b) Experimental and simulated values of I_{Dst} versus V_{DS} for a 0.8 μm NMOSFET with Z = 20 μm.[74] (© 1988 IEEE).

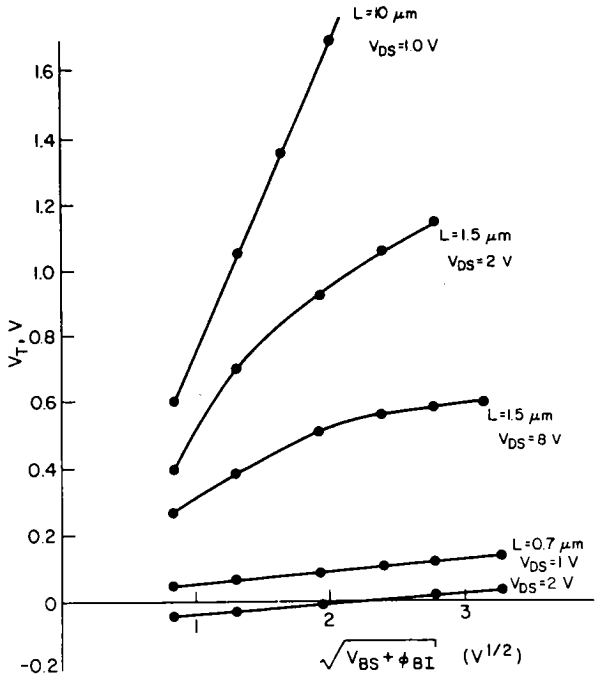


Fig. 5-12 Variation of V_T versus V_{BS} for three different channel lengths: $L = 10\ \mu m$ (long channel); $L = 1.5\ \mu m$ (short channel); $L = 0.7\ \mu m$ (punched through at $V_{BS} = 0V$). $N_A = 10^{15}$ cm^{-3}, $t_{ox} = 1000Å$, $r_j = 1\ \mu m$.[10] Reprinted with permission of *Solid-State Electronics*.

in the threshold voltage equation. Similar results have been obtained by Sheu et al.[11] They, however, elected to model this effect in the short-channel MOSFET SPICE model (BSIM, see section 5.5) by including another empirical term in the V_T equation (i.e., this new term is $P_2[2\varphi_B - V_{BS}]$, see Eq. 5-75). The coefficient P_2 of this term is empirically determined from measurements, and has a value of 0.8634 in BSIM.

5.2.2 Narrow Gate-Width Effects on Threshold Voltage

As device *widths* shrink, three additional effects begin to modify the expected V_T behavior of such *narrow-channel* MOSFETs. Although these effects are less critical than short-channel effects, they cannot be ignored. The first two cause measured values of V_T to *increase* above those calculated by Eq. 5-16 as the width is decreased (making the MOSFET appear narrower than its geometrical width).[12] The third causes V_T to *decrease*, in effect making the device appear wider. The first two effects are exhibited by MOSFETs fabricated with either raised field-oxide isolation structures (in which the field oxide is grown and then etched to expose the active device regions, Fig. 5-13a) or with semi-recessed LOCOS isolation structures (see chap. 6) as shown in Fig. 5-13b.

The third effect occurs in devices which use fully-recessed-LOCOS or trench-isolation structures (in which the trench is filled with an insulator).

The first effect is the opposite of the short-channel V_T shift effect. That is, at the source and drain ends of the channel (i.e., along the length dimension L), a depletion region exists at each *pn* junction (even at V_{FB}), requiring less voltage to be applied to invert the channel. As just mentioned, this leads to a *decrease* in V_T in *short-channel* MOSFETs. However, at the ends of the channel perpendicular to the drain current flow, no *pn* junction exists (i.e., along the edges of the device in the *width* dimension). Hence, a gate bias that repels majority carriers must not only deplete such carriers in the channel in the vertical direction, but also in the lateral direction beyond the thin gate oxide. Because the lateral depletion region width in this region extends beyond the gate, an effective bulk charge width larger than the actual width is apparently present. If the width Z is small, these side regions represent a large percentage of the depleted volume. Hence, a larger V_{GS} must be applied to invert the channel than if these edge regions did not have to be depleted. This leads to an *increase* in V_T.

The second effect arises from the encroachment of the channel-stop dopants under the edges of the sides of the gate. (In the case of NMOSFETs boron channel-stop dopants

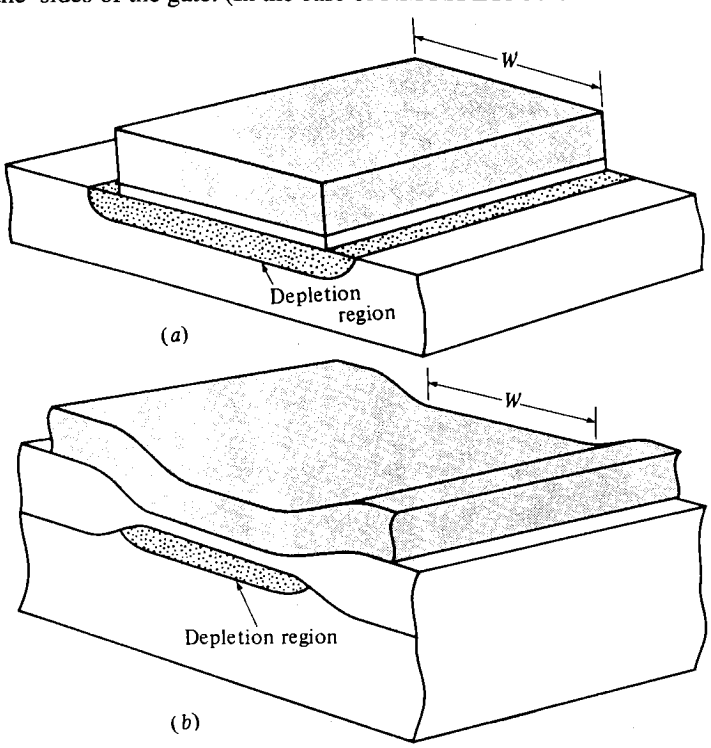
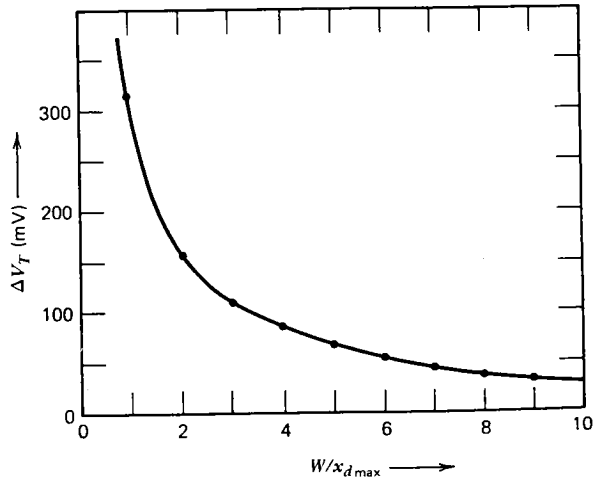


Fig. 5-13 Effect of fringing field on spreading the depletion region sideways. (a) In MOSFET with field oxide grown and then etched to expose the active region. (b) In MOSFET having a semirecessed LOCOS field oxide. From Y.P Tsividis, *Operation and Modeling of the MOS Transistor*, McGraw-Hill, New York, 1987, p. 192. Reprinted with permission.



5-14 Threshold voltage shift ΔV_T caused by this narrow-channel effect for a MOSFET with $N_A = 10^{15}$ cm^{-3}, and t_{OX} = 50 nm.[12] From G. Merkel, *Process and Device Modeling for IC Design,* (F. Van de Wield, Ed.), Noordhoff, Leyden, (1977). Reprinted with permission.

encroach under the gate edges. In the case of PMOSFETs, where a channel-stop implant may not be used, phosphorus pileup at the silicon surface during field oxidation produces *n*-type channel stop dopants at higher concentrations in the field regions. This also gives rise to subsequent encroachment into the active regions.) This has the effect of doping the channel at the edges along the width dimension more heavily than at the center (see Fig. 6-10). Again, a larger voltage must be applied to the gate to completely invert the channel than if the additional doping at the edges was not present (Fig. 5-14). From another perspective, the channel at the edges will be less turned on than the channel at the center. Thus, for the same (V_{GS} - V_T), significantly less current may be conducted in a narrow-channel MOSFET with channel-stop dopant encroachment than in one without such encroachment.

In practice, the latter (encroachment-related) effect is more severe than the edge-depletion-region effect, especially in devices with heavy channel-stop implants. An analytical expression that models both effects in semirecessed MOSFETs is given in ref. 13. For more precise calculations, however, numerical analyses are required.[14] Several methods are described in chap. 6 for reducing the channel-stop encroachment and thus reducing this narrow-width effect.

Because the third effect *decreases* V_T, is often called the *inverse narrow-width effect.* As pointed out earlier, the effect occurs in MOSFETs with trench or fully recessed LOCOS isolation structures, as shown in Fig. 5-15. (Note that the narrow width effects described earlier does not occur in trench isolation structures. Since there is no semiconductor region beyond the edges of the gate in the width dimension, there is no region that is subject to depletion.)

Figure 5-15a depicts the edge of the channel, the trench field-oxide, and the gate (which always overlaps the channel to some degree). When the gate is biased, the field lines from the overlapping region are focused by the edge geometry of the channel. Thus, at the edges of the channel, an inversion layer is formed at lower voltages than at the center. As a result, *less* bias overall must be applied to the gate to invert the channel across its full width. From another perspective, a *corner parasitic MOSFET* in parallel with the main device is established. The parasitic device turns on at voltages lower that the main channel, resulting in a "hump" in the drain current versus gate voltage curves (Fig. 5-15b).[15] Thus, the parasitic device increases the subthreshold leakage current of the active device.

An analytical expression that simulates the narrow-width effect in an ideal isolation trench structure is given in ref. 16. A more general analysis of the narrow width effect that covers both the first and third effects is carried out in ref. 17. An even more recent model of the inverse-narrow-width effect in trench isolated submicron MOSFETs, including the impact of non-uniformly doped channels was proposed by Chung and Li.[18]

However, the inverse-width effect is also sensitive to several other factors,[19] including:

1. The *doping concentration in the sidewalls of the silicon*, which can be impacted by channel-stop dopant redistribution, dopant segregation during field-oxide growth, or intentional doping of the sidewalls by such processes as large-tilt-angle implantation.

2. The trench isolation spacing.

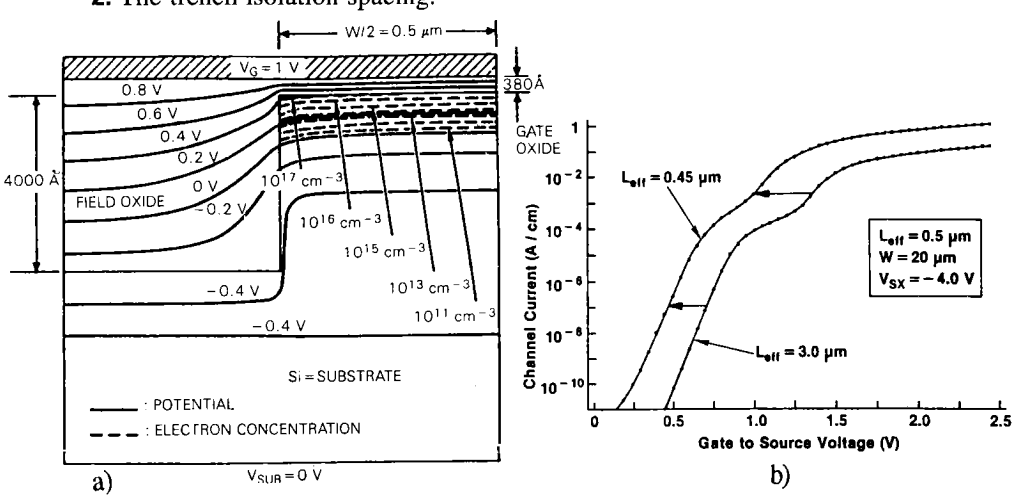


Fig. 5-15 (a) Structure and two dimensional contours of equipotentials and electron concentrations for a MOSFET employing a trench or fully-recessed oxide isolation structure.[112] (© 1979 IEEE). (b) "Hump" in subthreshold I_D-V_{GS} curve due to the inverse narrow width effect in trench isolated MOSFETs.[15] This paper was originally presented at the Fall 1988 Meeting of the Electrochemical Society held in Chicago IL.

3. The shape of the corner region at the edge of the gate. (If the top surface of the field oxide is depressed below the gate-oxide surface, the effect will be more severe than if the field oxide surface is raised above the gate oxide surface.)

Thus, if trench isolation is used, the planarization technology that determines the final wafer topology at the conclusion of the isolation structure fabrication will have to be tightly controlled to maintain precise control of V_T (see also chap. 6).

5.2.2.1 Devices Exhibiting Both Short- and Narrow-Width Effects. In devices in which both effects are capable of causing a shift in V_T, neither the short-channel nor the narrow-channel models alone suffice to accurately predict the resultant V_T. Some simplified analytical expressions employing a three-dimensional approximation of the depletion regions have been developed,[20] but for cases involving various device voltages and non-uniform channel doping concentrations, numerical analyses are necessary.

5.2.3 Reverse Short-Channel Effects on Threshold Voltage (RSCE)

In the previous sections it was implied that the V_T of a MOSFET decreases monotonically with decreasing channel length. This was explained by two-dimensional electrostatic field considerations, and various models for predicting such effects were described. However, in some situations it has been found that V_T initially *increases* with decreasing channel length (beginning when L ~2-3 μm), contrary to what is normally expected.[91,92] This phenomenon has been termed *reverse short channel effect (RSCE), V_T roll-up,* or *anomalous threshold behavior.* After V_T reaches a maximum value due to RSCE (at ~0.7 μm), it then declines as channel lengths are further decreased (*V_T roll off*). Apparently the two-dimensional effects that are responsible for V_T roll-off eventually compete with the effects causing V_T roll-up and the former come to dominate at the shortest channel lengths. (In fact, the ultimate rate of V_T roll-off with L_{eff} is reported to be much faster than can be explained by conventional models of laterally uniform channel doping). The combined RSCE and V_T roll-off effects result in a "hump" in the V_T versus L characteristics (Fig. 5-16). The initial rate of enhancement of V_T with decreasing L is also maximum at zero bias and decreases with increasing bias. As will be described in the next sections, different explanations have been given for the V_T roll-up effect. However, while several models which seem to provide good quantitative agreement with experimental observations of RSCE have been developed, a consensus has not yet been reached as to which physical mechanism(s) are responsible for the effect.

5.2.3.1 Model which Proposes RSCE is Due to Lateral Dopant Nonuniformity at the Channel Si-SiO_2 Interface (Arising from Enhanced Diffusion of Channel Dopants Caused by Interstitial Injection During Poly Reox, Salicide Formation, or Implant Damage). The first explanation of RSCE is based on the fact that this effect was initially observed in NMOSFETs fabricated with two boron channel implants (i.e., a

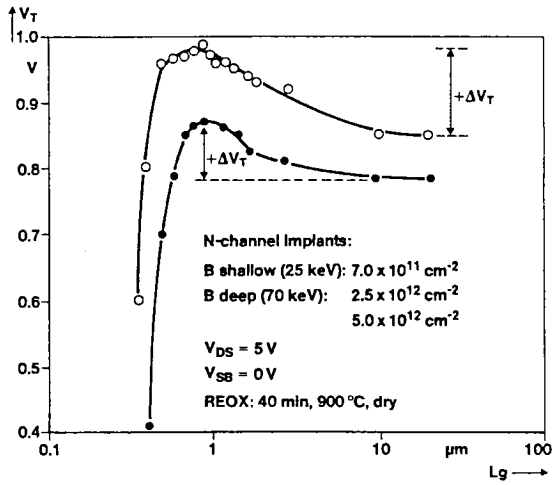


Fig. 5-16 Experimental *n*-channel threshold voltage versus channel length.[92] (© 1987 IEEE).

shallow boron implant for adjustment of the threshold voltage, as described in chap. 4, and a deeper boron implant for suppression of punchthrough, as described in section 5.3).[92] The doping profile as a function of x in the channel beneath the gate oxide resulting from these two implants is qualitatively depicted in Fig. 5-17a.

After the polysilicon gate is defined by etching and the source/drain regions are implanted, a thermal oxidation step (referred to as *reoxidation* or *re-ox*) is performed (see also chap. 9, sec. 9.7.1.4). This oxidation serves several purposes, including: (1) to re-oxidize the source/drain regions. (The source/drain implant is carried out with the gate oxide intact over these regions. The heavy implant is likely to damage and contaminate this oxide, and so it is stripped after the implant. The reoxidation step regrows a fresh oxide over the source/drain regions, as well as on and around the poly gate); (2) the thermal oxide on the poly serves as a diffusion mask which prevents boron dopants from the BPSG interlevel dielectric from diffusing into the n^+ poly during the BPSG flow and reflow processes; (3) to reduce the MOSFET gate-to-drain overlap capacitance and strengthen the gate oxide at the poly gate edge. The latter benefits are obtained because oxidation-induced encroachment gives rise to a so-called *gate bird's beak* - or *graded gate oxide* [GGO] - under the polysilicon edge, as shown in Fig. 5-17b. The thicker oxide due to GGO at the gate edge reduces the gate-to-drain overlap capacitance and relieves the electric-field intensity at the corner of the gate structure, thus enhancing the gate oxide integrity at its edge. However, the reoxidation step can also detrimentally impact the characteristics of MOSFETs which employ lightly-doped drains (LDD), and thus the re-ox step must be optimized for this type of device, as is described in more detail in chap. 9. Here only the proposed impact on V_T due to the re-ox step is considered.

As described in chap. 7. vol. 1, the injection of interstitials which accompanies thermal oxidation causes oxidation-enhanced-diffusion (OED) of dopants in the vicinity of the growing oxide (see Fig. 5-17b). In the re-ox step, the diffusion of dopants is enhanced not only under the oxidizing (i.e., source/drain) regions but also in the adjacent

channel region (due to lateral diffusion of the injected Si self-interstitials into the channel region). The OED causes boron to diffuse more rapidly. This brings more boron from the deep buried peak to the surface near the edges of the channel (because the injected interstitial concentration is highest near the channel edges). As a result, the boron surface concentration becomes larger at the channel edges than if diffusion of the implanted boron profile had not been enhanced (see Fig. 5-18). The increased concentration of the boron in these regions gives rise to an overall increase in V_T if L becomes small enough to make these regions a significant fraction of the channel length.

The V_T roll-up has also been observed in MOSFETs in which a salicide is fabricated[93] (i.e., a *salicide* is a self-aligned silicide formed on the source/drain regions to reduce the sheet resistance of the source and drain, see chap. 3, Vol. 2). Lu and Sung propose that during silicide formation vacancies are injected and these enhance the dopant diffusion of the boron in the dual-implant profile, again giving rise to a lateral nonuniform lateral distribution of boron along the channel surface. As in the OED model, this gives rise to an enhancement of V_T as the MOSFET channel length is decreased.

An analytical model of RSCE based on nonuniform lateral boron surface concentrations arising from either (or both) of the above effects has been developed by Arora and Sharma,[94] although they model the RSCE as being due to a nonuniform buildup of oxide charge along the Si-SiO_2 interface (see also section 5.2.3.4).

5.2.3.2 Model Which Proposes RSCE Arises from Boron Segregation to Implant Damaged Regions at the Edge of the Channel. Although the above models appear plausible, V_T roll-up has also been

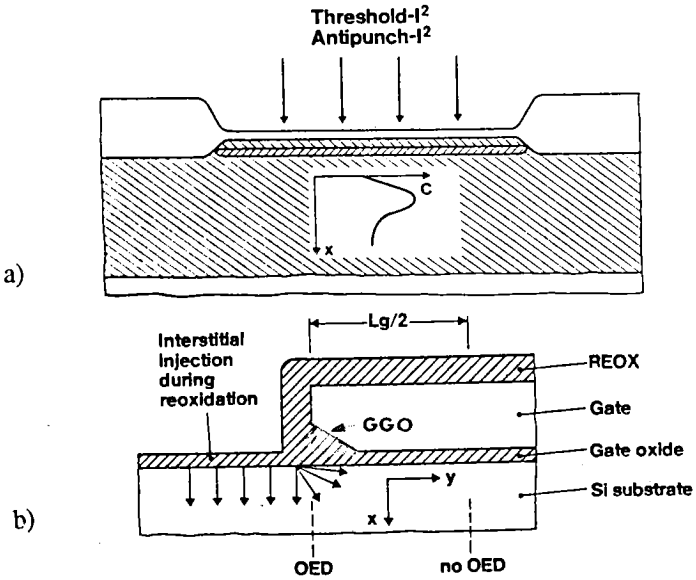


Fig. 5-17 (a) Vertical doping profile in an NMOSFET channel which uses a punchthrough implant. (b) Conceptual diagram of local interstitial injection during reox.[92] (© 1987 IEEE).

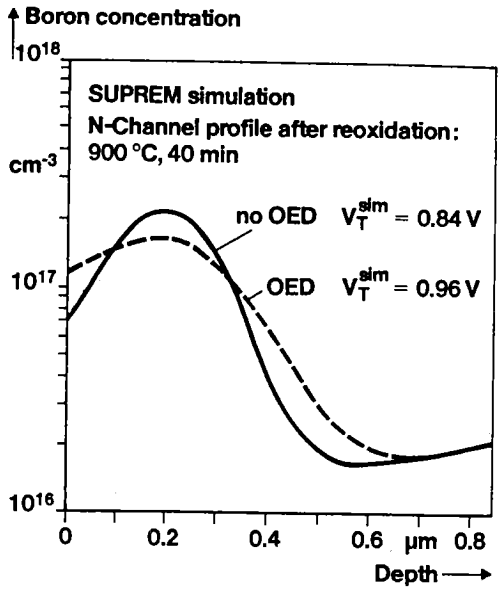


Fig. 5-18 Simulated vertical n-channel profiles at x = 0 and x = $L/2$ with and without oxidation-enhanced diffusion (OED).[92] (© 1987 IEEE).

observed in MOSFETs with uniformly doped channel regions (Fig. 5-19). In such MOSFETs the above models would predict that boron surface concentrations would *decrease* near the edges of the channels. (That is, boron would segregate into the growing oxide, and then boron from the center of the channel would diffuse into these depleted regions, resulting in a loss of boron from the channel region). Thus, the model of section 5.2.3.1 predicts that V_T would *decrease monotonically* with decreasing channel length in MOSFETs with nonimplanted (i.e., uniform) channel regions. Furthermore, when the channels get very short, (i.e., ≤0.3 μm) the observed V_T *roll-off* is more rapid than that predicted by the 2-D DIBL numerical analyses assuming a laterally uniform channel doping. As a result, another phenomenon has been suggested as being responsible for the RSCE (including both the initial V_T roll-up and the subsequent enhanced roll-off effects as the channel length is decreased), namely the depletion of boron from the channel region by strong segregation of the boron into the adjacent source and drain regions during the source/drain implant-activation anneal.[95, 96]

This segregation is thought to occur as a result of the presence of crystal defects caused by the source/drain implant step. (Such defects have, in fact, been detected using TEM micrographs of the source/drain regions after the source/drain implant step.) The boron is theorized to have a higher local solubility in the defective regions, and thus will diffuse to them during the anneal steps performed after the source/drain implant. Figure 5-19 shows the measured and calculated V_T versus L curves. The triangular data points in this figure are calculated using a 2-D device simulator assuming a uniform

lateral doping profile in the channel, and these show the conventional roll-off versus L as described in section 5.2.1. The data points indicated by the squares and filled circles show the V_T versus L values calculated using the same device simulator, but with nonuniform lateral boron channel doping (see Fig. 5-20). The open-circle data points are the experimental data.

The V_T roll-up is well predicted by this model, as is the enhanced V_T roll-off for channel lengths shorter than about 0.3 μm. (This model also gives good agreement with RSCE in PMOSFETs.) For MOSFETs in which $L \leq 0.3$ μm the center of the channel is so small that the channel region depleted of boron extends across the entire channel length (see, for example, case where $L = 0.2$ μm in Fig. 5-20), and there is not sufficient dopant present in this region to cause the peaks of the boron concentration at the channel edges to rise to a degree that would offset the enhanced V_T decrease due to the depletion effect.

The defects that give rise to the lateral dopant redistribution are observed to be eventually extinguished by a higher temperature anneal step (e.g., 900°C in N_2 for 5 min). Thus, it has been suggested that recrystallization of the defective regions using RTP could be carried out, such that the anneal would take place over a sufficiently short time as not to permit lateral dopant redistribution to occur.[95] If this is possible MOSFETs could be fabricated which would exhibit neither the V_T roll-up nor enhanced roll-off effects. Another suggestion to alleviate the enhanced roll-off effect is to use a halo implant to increase the boron dopant concentration near the source and drain such that the B depletion would be counteracted.[96]

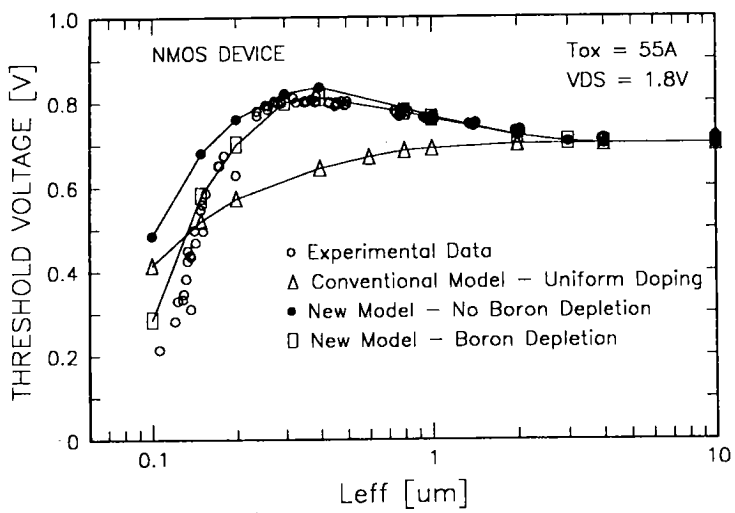


Fig. 5-19 Measured and modeled V_T versus L curves. The conventional model results assume laterally uniform channel doping. The new model results assume box approximations for the lateral channel doping profile shown in Fig. 5-20. The boron depletion case assumes that the bulk doping is depleted into the source and drain regions within 0.1 mm of their edges.[95] (© 1993 IEEE).

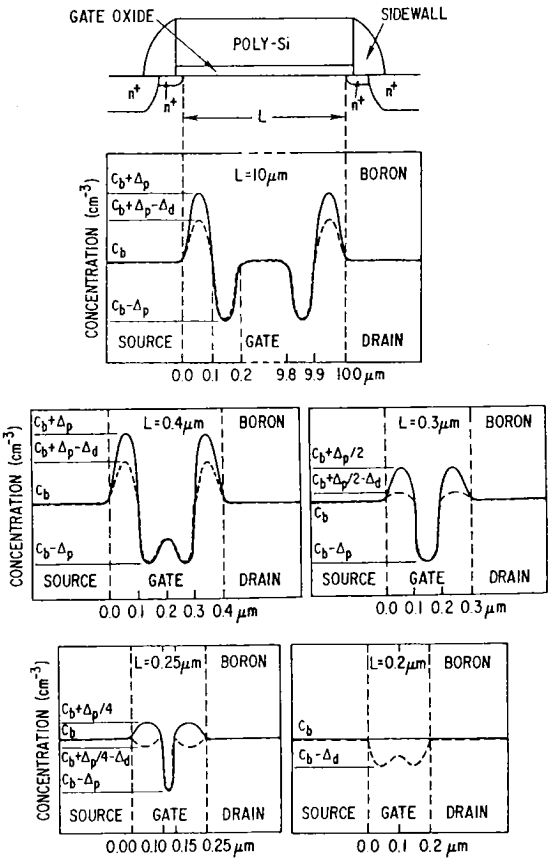


Fig. 5-20 Conceptual schematic of the lateral channel doping profile of the present model for different channel lengths. The solid lines represent the case of no bulk depletion while the dotted lines represent the case of bulk depletion.[95] (© 1993 IEEE).

5.2.3.3 Model Which Attributes RSCE to Transient Enhanced Diffusion of Channel Dopants to the Silicon Surface Arising from Implant Damage. Another model that explains the occurrence of RSCE in MOSFETs with uniformly doped channel regions was developed by Rafferty et al.[98] In this model, it is proposed that the effect is associated with damage from the LDD implant. That is, point defects (interstitials) are assumed to be created by the LDD implant. They diffuse laterally into the channel region during the relatively low-temperature step used to deposit the CVD interlevel dielectric layer (needed to create the gate-sidewall spacers). As these point defects are primarily created at a depth corresponding to the end-of-range damage, the interstitial concentration peak exists beneath the surface. Thus, a retrograde profile of interstitials in the channel will arise. Since the dopant diffusivity is proportional to the interstitial concentration, the

retrograde interstitial concentration profile will cause a net boron diffusion flux toward the surface (i.e., during the source/drain activation anneal) even if a flat boron dopant profile exists in the vertical direction. As a result, boron piles up at the surface, leading to an increase in V_T. Experimental confirmation of such an impurity pile-up is found in SIMS analysis of flat profiles subjected to silicon damage and subsequent anneal.[99] The diffusion of boron toward the channel surface from this effect is a transient phenomenon because the interstitials are eventually annihilated during the anneal. Good agreement between this model and experimental data is also reported if the transient diffusion is modeled using a coupled impurity/point defect simulator.

5.2.3.4 Model which Attributes RSCE to Silicon Interstitial Capture in the Gate Oxide. Jacobs et al. note that V_T can also be shifted by a change in fixed oxide charges Q_f at the Si-SiO_2 interface.[97] Thus, they propose that RSCE in the absence of a vertical gradient in the channel dopant profile can be explained by the trapping of injected silicon interstitials by the gate oxide. These captured interstitials introduce the negatively charged complex in the oxide required for the observed V_T roll-up). The oxide surface is known to be an effective sink for interstitials, and the process steps which supply the excess interstitials can be those mentioned above. The result is a position-dependent oxide charge distribution along the Si-SiO_2 interface, similar to the assumption of an exponentially decaying fixed charge variation made by Arora when deriving the analytical model of RSCE mentioned in section 5.3.2.1.[94]

5.3 PUNCHTHROUGH (Subsurface-DIBL) IN SHORT-CHANNEL MOSFETS

In chap. 4, section 4.7 we described the nature of subthreshold current in long-channel MOSFETs. We noted that in weak inversion I_{Dst} flows at the surface of the channel region, and its value is calculated from Eq. 4-89. We also saw that the subthreshold-swing parameter (S_t) is an important characteristic of subthreshold behavior, and in long-channel, uniformly-doped-substrate devices it can be modeled using Eq. 4-91.

In short-channel MOSFETs, however, if V_{GS} is fixed while V_{DS} is increased, larger values of I_{Dst} values than are predicted by the long-channel device equations are observed. This larger subthreshold current is due in part to increased I_{Dst} flowing at the surface as a result of surface-drain-induced barrier lowering (DIBL), as described in section 5.2). If only surface DIBL is occurring in a MOSFET, S_t values are observed to remain unchanged. However, another short-channel effect known as *subsurface punchthrough* can also give rise to an increase in I_{Dst}, and one manifestation of this effect is an increase in the measured value of S_t. That is, the gate has less control of the subsurface punchthrough current than it has over the surface subthreshold current. Thus, once a MOSFET enters subsurface punchthrough and such current starts to flow, the S_t value of the MOSFET becomes larger than if only normal subthreshold current is flowing. Since the measurable drain current is the sum of both current components,

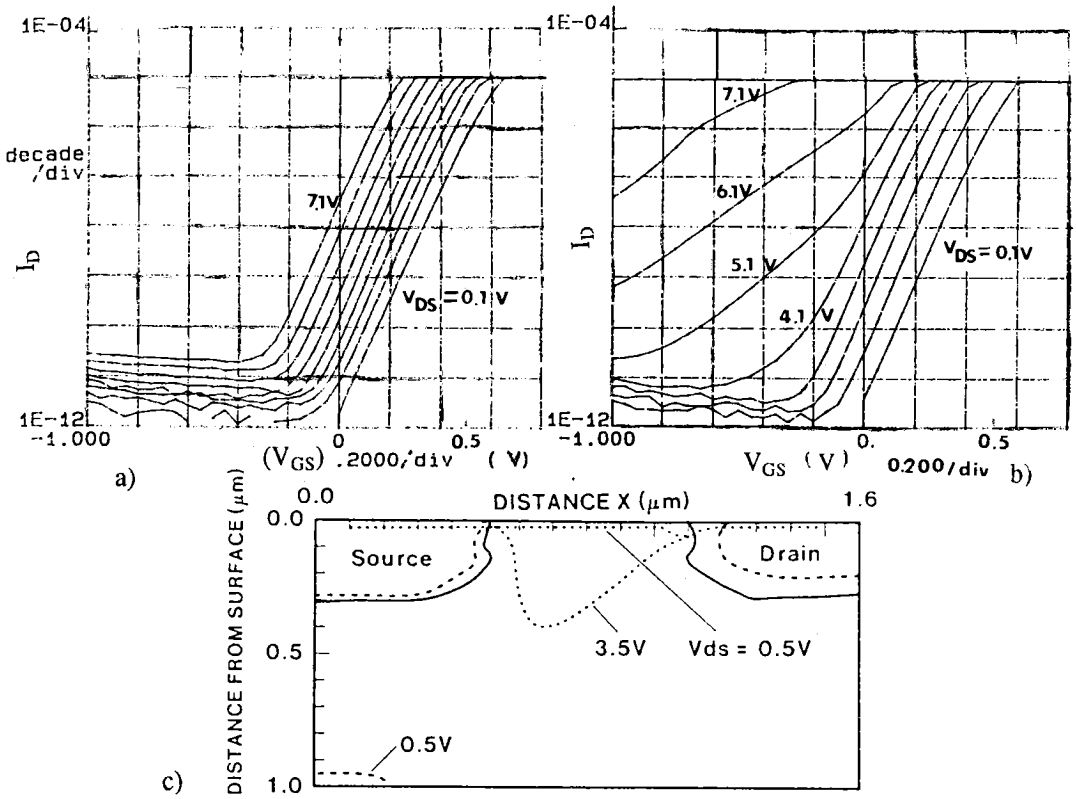


Fig. 5-21 Two sets of log I_D versus V_{GS} curves for MOSFETs, with V_{DS} varying in 1V increments. (a) No indication of punchthrough is observed, as S_t remains constant for all values of V_{DS}. (b) Onset of punchthrough occurs as V_{DS} reaches ~6V, as evidenced by an increase in S_t that point. (c) Punchthrough current paths (dotted) in a surface-channel NMOSFET with V_{DS} = 0.5 and 3.5V. Solid line shows the depletion region edge. The current path transfers from the surface to bulk as V_{DS} is increased.[74] (© 1988 IEEE).

one technique for detecting the onset of punchthrough in a MOSFET is to monitor S_t as V_{DS} is increased. .

For example, in Fig. 5-21 two sets of I_{Dst} versus V_{GS} curves with V_{DS} as the parameter are shown. In the MOSFET characterized by Fig. 5-21a S_t remains constant as V_{DS} is increased, indicating that no punchthrough is observed up to the maximum applied V_{DS}. In Fig. 5-21b S_t increases as V_{DS} is increased beyond ~4V, indicating that punchthrough sets in at about this value.

Punchthrough is a phenomenon associated with the merging of the source and drain depletion regions in the MOSFET. That is, as the channel gets shorter, these depletion region edges get closer (assuming that the channel-region doping is kept constant as L is decreased). When the channel length is decreased to roughly the sum of the two junction depletion widths, punchthrough is established. (Note that punchthrough may

even exist in a short-channel device at $V_{DS} = 0$, depending on L and the substrate doping.) Nevertheless, since the depletion regions of a *pn*-junction widen as reverse-bias is increased, all MOSFETs would eventually enter punchthrough if a high enough V_{DS} could be applied. However, in MOSFETs with $L>2.0$ μm, breakdown of the drain-substrate junction generally sets in before this punchthrough voltage is reached. As a result, in practice, punchthrough is not a limiting factor in long-channel digital MOSFET design. In shorter-channel devices (i.e., $L\leq 2.0$ μm), however, punchthrough does represent a serious limitation.

Another factor also impacts punchthrough behavior in *n*-channel MOSFETs. That is, in submicron NMOSFETs a V_T-adjust implant is used to raise the doping of the surface-channel region above the doping of the substrate. (This can also be viewed as an increase in the electrostatic potential at that point near the source.) Thus, the source/drain depletion-region widening is greater beneath the surface due to the lighter doping there. A punchthrough condition would therefore be first established below the surface in such devices, and hence this effect is referred to as *subsurface punchthrough* (see Figs. 5-22 and 5-23). Note that in PMOSFETs fabricated with n^+ polysilicon gates the punchthrough problem occurs in a somewhat different manner, which we will describe in a later section dealing with this issue (i.e., section 5.8.2.3).

Once subsurface punchthrough occurs, any increase in the applied reverse-bias drain voltage (beyond that required to establish punchthrough) will lower the potential energy barrier for majority carriers in the source.[6] With a diminished barrier height, a larger number of source-region majority carriers (e.g., electrons in an NMOSFET) have enough energy to be injected from source to substrate. Some of these injected electrons are collected by the drain, causing I_{Dst} to increase. The component of I_{Dst} which flows as a result of subsurface punchthrough is known as *punchthrough current.* When designing MOSFETs one criterion of a successful device design is that the punchthrough current should be insignificant under normal bias conditions. That is, if sufficient punchthrough current is observed, the device structure must be redesigned to effectively suppress such current.

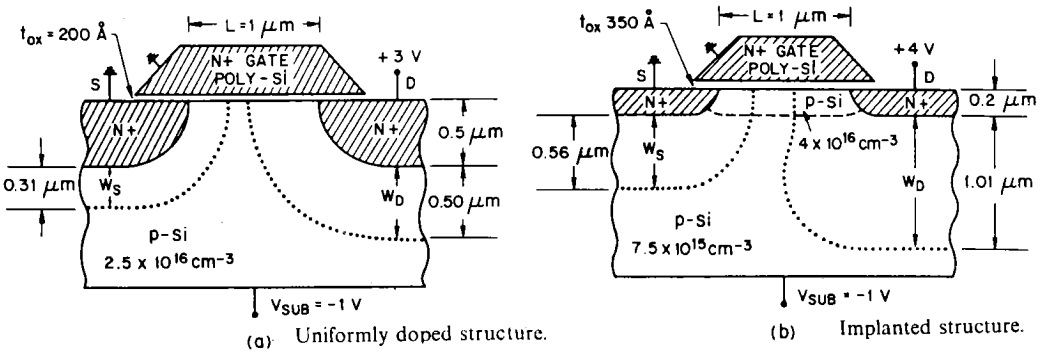


Fig. 5-22 (a) NMOSFET without V_T-adjust implant. (b) NMOSFET with V_T-adjust implant, showing the bulging of the drain depletion region beneath the surface as V_{DS} is increased, due to the lower doping concentration in the substrate below the depth of the V_T-adjust implant.

Two-dimensional simulations of the potential in the channel regions of short channel NMOSFETs with V_T-adjust implants support the view that depletion-region widening is greater beneath the surface. In addition, they indicate that the point of maximum source-substrate potential energy barrier lowering is also beneath the Si-SiO_2 interface (generally at approximately the same depth as the source/drain junction depths). Figures 5-23a, c, e and g show a "bird's eye" or perspective view of the electrostatic potential variation in a short-channel MOSFET computed using such 2-D numerical analyses, as V_{DS} is increased from V_{DS} = 3V to V_{DS} = 9V. Figures 5-23b, d, f and h show contour plots of these same results. We can see that when the device enters punch-through the location of the electrostatic potential saddle point is beneath the surface, and the value of the potential at this saddle point may be higher than the potential minimum at the surface. Since electrons follow the potential peak (see Figs. 5-23g and i), in such cases the component of current due to subsurface punchthrough will be larger than the channel-current component, and the subsurface current will flow along the path indicated by the subsurface potential ridge connecting the source and drain. Figure 5-21c shows a simulation of the subthreshold current paths in an NMOSFET with L = 0.6 μm for surface-DIBL (i.e., V_{DS} = 0.5V) and for subsurface-DIBL (punchthrough, V_{DS} = 3.5V). This shows that the subthreshold current path transfers from the surface to the bulk as V_{DS} increases, which supports the view that at low V_{DS} surface-DIBL is responsible for the increase in I_{Dst} in short-channel NMOSFETs, but that at high V_{DS} punchthrough becomes the primary cause of subthreshold leakage.

Subsurface punchthrough exerts the biggest impact on circuit operation when the MOSFET is OFF, since a tiny punchthrough current in this case can still represent intolerable leakage. However, the device parameter most commonly used to characterize punchthrough behavior is the *punchthrough voltage* V_{PT} defined as the value of V_{DS} at which I_{Dst} reaches some specific magnitude (e.g., 1 nA/μm) when V_{GS} = 0. The goal in designing a MOSFET is to have V_{PT} exceed any possible operating V_{DS}. Thus, V_{PT} is measured when the device is in subthreshold, as will be discussed in upcoming paragraphs. Note that V_{PT} can be roughly estimated by determining the value of V_{DS} at which the sum of the depletion widths W_S under the source and W_D under the drain (using the value of the doping concentration in the substrate as N_{SUB}) equals the channel length L, or from the expression[24]

$$V_{PT} \propto N_{SUB} (L - r_j)^3 \qquad (5 - 17)$$

Nevertheless, no accurate analytical model exists for estimating V_{PT} as a function of implant dose and depth. A full numerical solution of the 2-D Poisson equation must thus be obtained to predict the variation of V_{PT} for each device structure.

Several measures can be taken to prevent short-channel MOSFETs from entering subsurface punchthrough. First, the substrate doping can be increased to decrease the depletion-layer widths. A simple rule-of-thumb proposed by Klaasen[25] is to select a minimum substrate doping value according to $N_{SUB}>N_{ch}/10$, where N_{SUB} is the uniform bulk-substrate doping/cm^3 and N_{ch} is the average doping/cm^3 in the implanted region of the channel (e.g., D_I/x_C). This approach, together with a single V_T-adjust implant can produce appropriate threshold and punchthrough voltages for NMOSFETs

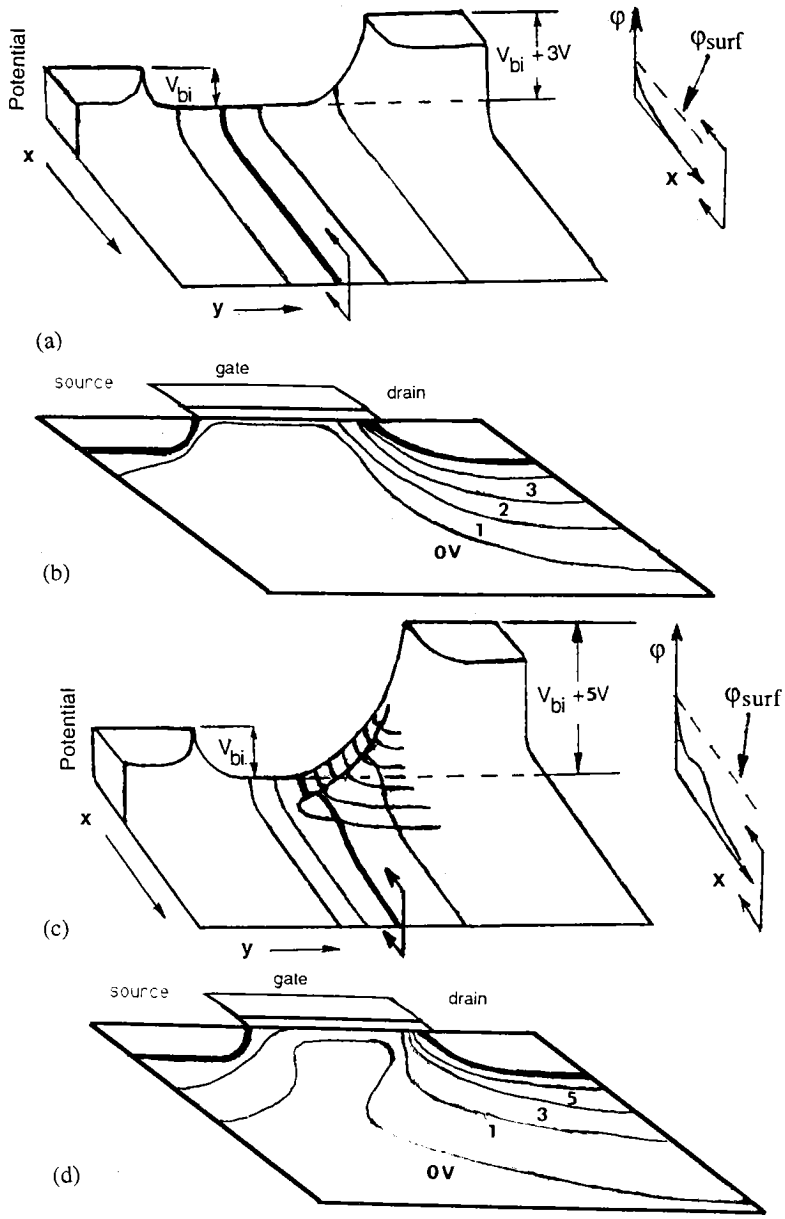


Fig. 5-23 (a and c) 2-D perspective plots of the potential in an *n*-channel MOSFET with a drain bias of (a) 3 V and (c) 5 V, respectively. Channel length = 1 μm. (b and d) 2-D contour plots of the same cases. From K. M. Cham, et al., *Computer Aided Design & VLSI Device Development*. Copyright 1986 Kluwer Academic Publishers. Reprinted with permission.

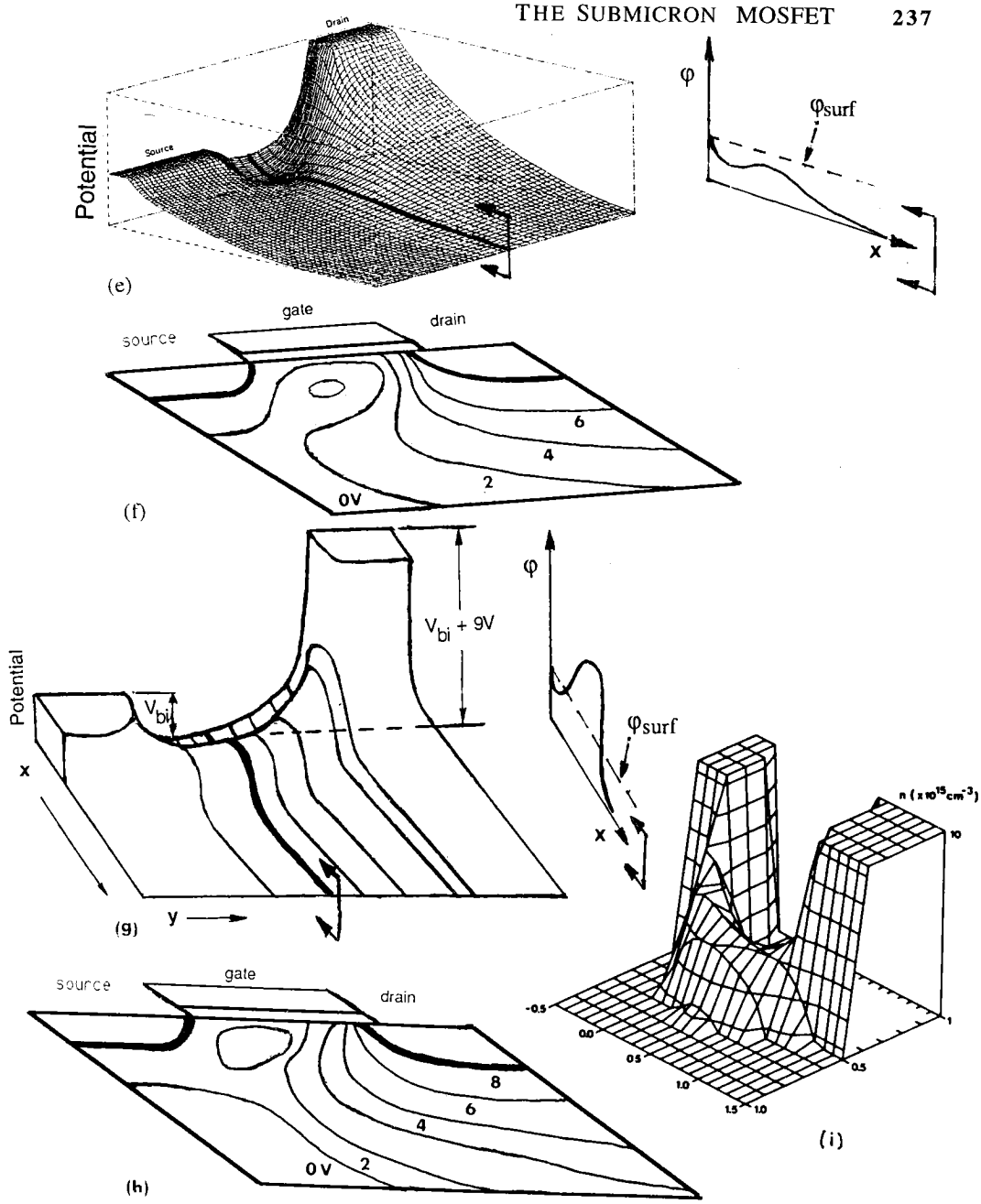


Fig. 5-23 cont. (e and g) 2-D perspective plots of the potential in an n-channel MOSFET with a drain bias of (e) 7 V and (g) 9 V, respectively. Channel length = 1 μm. (f and h) 2-D contour plots of the same cases. From K. M. Cham, et al., *Computer Aided Design & VLSI Device Development*. Copyright 1986 Kluwer Academic Publishers. Reprinted with permission. (i) Electron concentration of the MOSFET in punchthrough, i.e., case g & h.

with drawn gate lengths of 1.3 μm nominally operated at 2.5V power supply (e.g., a single 20 keV B^+ implant at ~$9x10^{11}$ cm^{-2}). A similar process has also been applied to 1.25 μm CMOS devices using a 5V supply. Although some short-channel effects are observed, V_T and V_{PT} voltages are reasonable. However, for devices with drawn gate lengths of ~1 μm (L_{eff} ~0.55 μm), a single V_T/punchthrough implant cannot adequately satisfy all requirements, especially if high resistivity (>10 Ω-cm) substrates are used.

As a result, another approach using an additional implant known as a *punchthrough stopper implant* (or *punchthrough implant,* PTI) must be employed. The second implant peak concentration is located at a depth near the bottom of the source-drain regions. Fig. 5-24a qualitatively depicts such a MOSFET structure. The additional doping introduced by such an implant (e.g., boron in NMOSFETs) reduces the lateral widening of the drain-depletion region below the surface without increasing the doping under the junction regions. With such implants, the component of the punchthrough current can be suppressed to well below the long-channel I_{Dst} value in MOSFETs down to 0.5 μm, while still maintaining a lower value of N_{SUB} (for reduced values of C_j). However, precise control of the placement and dose of this implant within the channel region is required to achieve the intended enhancements. For example, Fig. 5-24b shows the as-implanted doping profile of the channel with (dashed line) and without (solid line) a punchthrough implant (i.e., prior to annealing the implants). In addition, it is important to keep the punchthrough implant profiles from spreading out appreciably during annealing (see Fig. 5-24c), since the additional doping above the well's value adds to the source/drain-to-well capacitance and the transistor's threshold voltage sensitivity to back biasing of the*p*-well. Thus, the post channel implant thermal cycles should be kept to a minimum. An example of such a punchthrough implant for a 1 μm NMOSFETs is a 70 keV B^+ implant at $6x10^{11}$ cm^{-2}.

A third technique is to locally implant *p*-type dopants (in NMOSFETs) under the lightly doped tip region of the LDD as shown in Fig. 5-25a, and such implants have been termed a "halo" implants. (The use of halo implants as punchthrough stoppers in PMOSFETs is discussed in section 5.8.1.1.) Because the implanted dopant raises the doping concentration only on the inside walls of the source/drain junctions, the channel length can be decreased without needing to use a higher-doped substrate. That is, punchthrough does not set in until a shorter channel length because of the "halo." A penalty in increased sidewall junction capacitance is paid, but this cost is smaller than that incurred should the bottom area of the junctions lie above heavily doped substrate material. The maximum halo doping is determined by the onset of low-level avalanche breakdown of the pn^+ junction as is shown in Fig. 5-25b.

A halo NMOSFET with L = 0.35 μm, t_{ox} = 15 nm, and operated at 3V has been reported.[26] The junction depth r_j and *p*-halo peak doping were ~0.14 μm and ~$3x10^{17}$ cm^{-3}, respectively. A more recent report applied a large-angle tilt (LAT) implant of boron ions in NMOS and phosphorus in PMOS (Fig. 5-25c) to form a halo-like structure in each device (called a *self-aligned pocket implantation*). A tilt angle of 25-30° was used to implant the pocket dopants after gate formation, so that the gate structure acted as a self-aligned implant mask. Devices with channel lengths as small as 0.2 μm that still exhibited long-channel subthreshold V_T behavior were fabricated.[27]

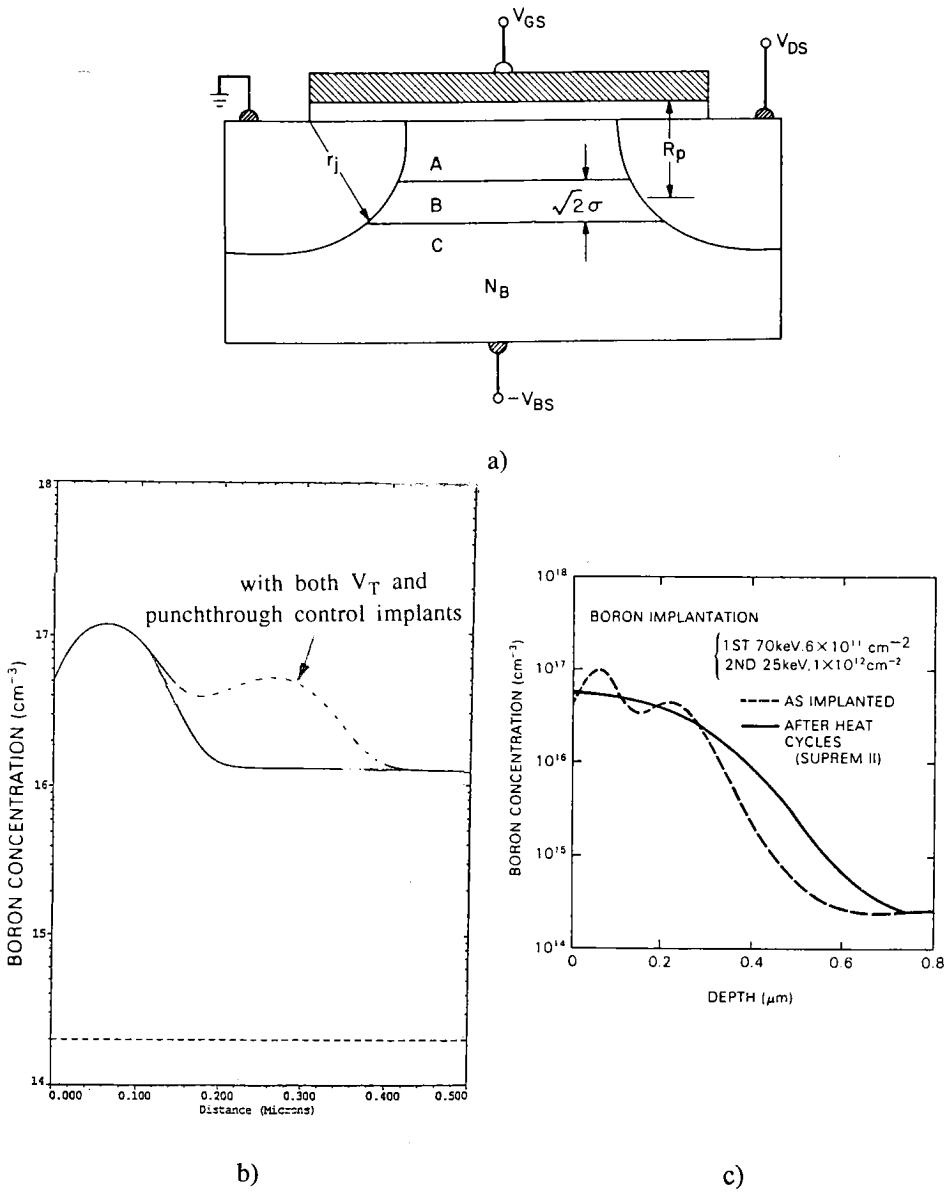


Fig. 5-24 (a) An implanted NMOS structure with three regions of different doping. Level A contains the dopant from the V_T implant. Level B contains the dopant from the punchthrough control implant, considered to be a Gaussian centered at a distance R_p from the gate, with a straggle σ. Level C is the original substrate. (b) As-implanted channel doping profiles with (dashed line) and without (solid line) punchthrough implant. (c) Example of simulated channel impurity profile of an NMOSFET with both V_T and punchthrough control implants - as-implanted (dashed lines), and after anneal (solid line).[113] (© 1982 IEEE).

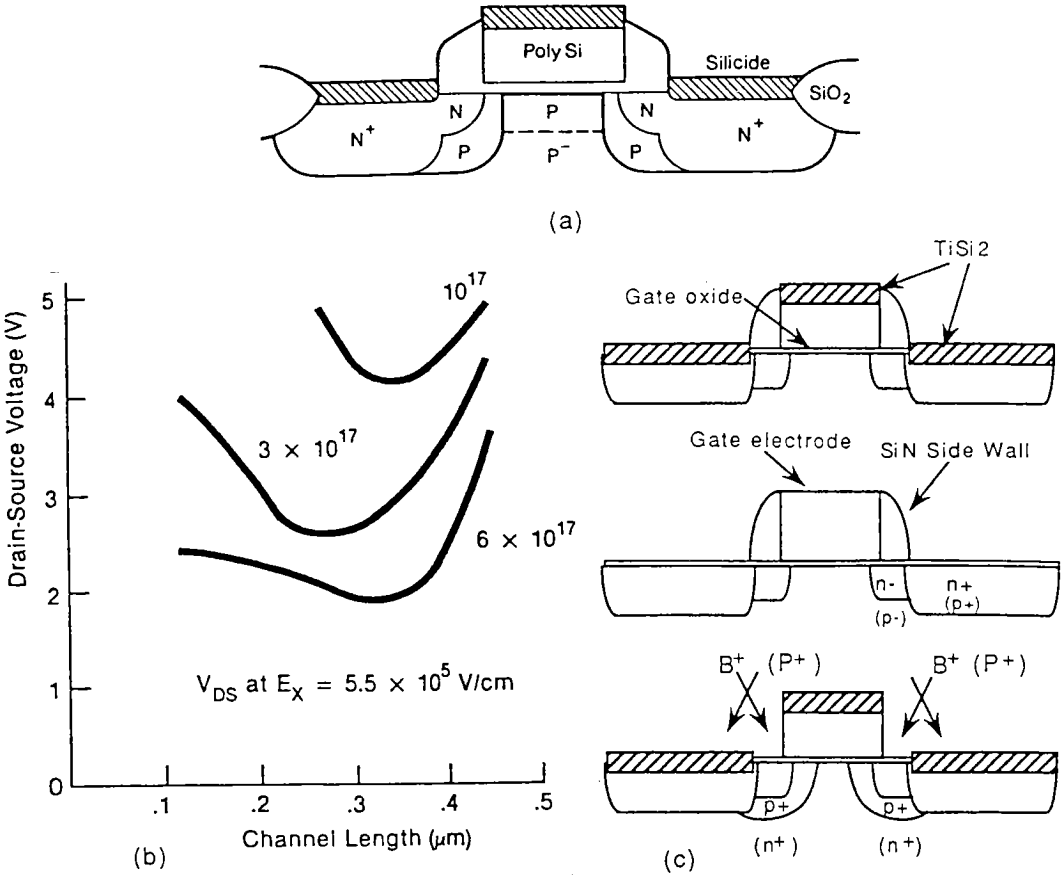


Fig. 5-25 (a) "Halo" NMOS structure.[26] (© 1985 IEEE). (b) V_{DSmax} limited by avalanche breakdown versus channel length for various peak halo doses.[26] (© 1985 IEEE). Process for producing "halo" NMOS structure using a Self-Aligned Pocket Implantation [SPI] technology is shown.[27] (© 1991 IEEE).

5.3.1 Experimentally Characterizing Punchthrough

The onset of punchthrough in a device can be measured in several ways. As mentioned at the outset of the discussion, for a specific device (i.e., one that has a single, specific channel length and channel doping profile), the onset of punchthrough over the operating range of V_{DS} can be monitored by measuring log I_{Dst} vs. V_{GS} for different V_{DS} values (Figs. 5-21a and b). If the curves look like those of Fig. 5-21a, the device does not exhibit subsurface punchthrough up to the maximum V_{DS} applied (since the slope of the log I_{Dst} curves remain constant as V_{DS} is increased). The curves of Fig. 5-21b however, indicate that punchthrough is setting in, with its onset marked by the V_{DS}-curve having a larger S_t value than that exhibited at small V_{DS}. (Note that since punchthrough current flows in parallel with ordinary I_{Dst}, S_t will noticeably increase

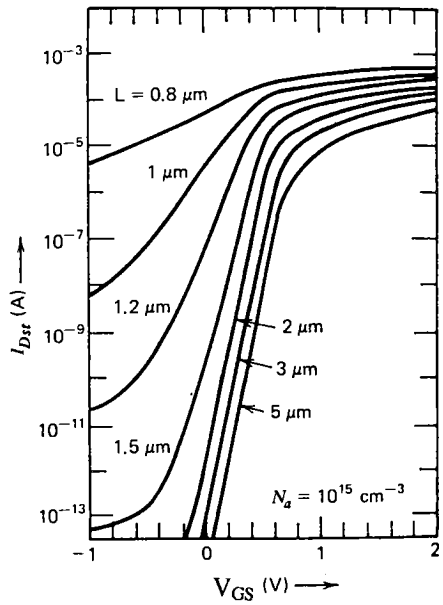


Fig. 5-26 Simulated behavior of I_{Dst} versus V_{GS} for various channel lengths, $r_j = 0.33$ μm, $t_{ox} = 500$Å, $V_{DS} = 2$V, and $V_{BS} = 0$V.[109] Reprinted with permission of *Solid State Electronics.*

once punchthrough current exceeds a small fraction of the long-channel I_{Dst} value.) Punchthrough voltage V_{PT} (as defined earlier) can also be extracted with this procedure by continuing to increase V_{DS} until the specified I_{Dst} is reached at $V_{GS} = 0$.

One way to detect the onset of punchthrough in a family of devices with varying channel lengths is to plot log I_{Dst} vs. V_{GS} for a fixed V_{DS} for the devices of different channel lengths on a single graph (Fig. 5-26). A second method is to measure S_t at a fixed V_{DS} for devices with different channel lengths, and then to plot these values as shown in Fig. 5-27. Here the impact of a punchthrough implant on S_t versus channel length is depicted. A third way is to measure I_D vs V_{DS} at $V_{GS} = 0$, and to pick off the V_{DS} value when I_{Dst} reaches the specified value (e.g., 0.1 μA/μm).

An example of the impact of a punchthrough implant (as well as the impact of depth and dose) is shown in Fig. 5-28. Here a 1.2-μm MOSFET with a substrate doping of 1.9×10^{15} cm^{-3} (but without a punchthrough-stopping implant) shows a very large S_t value (and hence a large I_{Dst} when $V_{GS} = 0$ V - curve A). The large I_{Dst} at $V_{GS} = 0$V indicates that the device is in punchthrough even when $V_{DS} = 0$V.[28] Implants of boron with a dose of 8×10^{11} atoms/cm^2 and different energies are then performed in an attempt to produce a V_T and an S_t equal to the values exhibited by a long-channel (7.8 μm) device (curve B). If the punchthrough implant is too shallow, it has the effect of increasing the V_T of the device to well beyond the desired value (curve C). When the energy is increased so that the implant is sufficiently deep, S_t drops to a value comparable to that exhibited by the long-channel device (curves D and E). At the same time, the surface concentration remains essentially unchanged, so that V_T is not

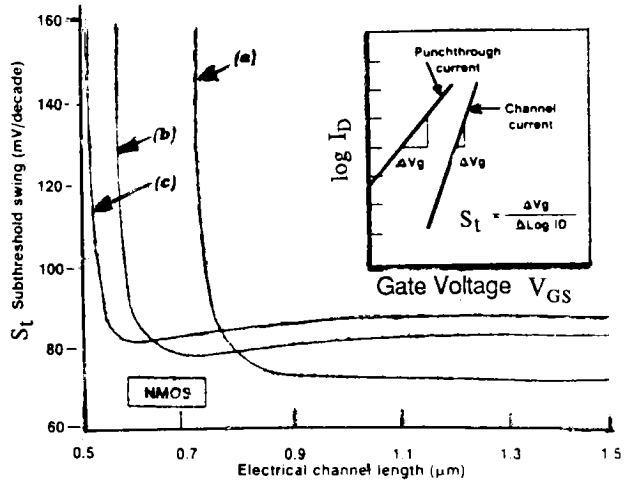


Fig. 5-27 Subthreshold slope versus electrical channel length for NMOS devices V_T = 0.7V, having a common threshold adjustment implant and punchthrough implant doses of: (a) zero; (b) $2x10^{11}$ cm^{-2}; (c) $3x10^{11}$ cm^{-2}.[29] Reprinted with permission of Semiconductor International.

appreciably shifted. Note, however, that S_t in the punchthrough implanted device is somewhat larger than that of the long-channel device.

In another example, the S_t of a device without a punchthrough-prevention implant is measured as its length is varied (Fig. 5-27).[29] At an L_{eff} of ~0.85 μm the S_t starts to increase, indicating that punchthrough current begins to dominate I_{Dst}. By adding an implant that places boron atoms in subsurface region below the V_T-adjust implant, the punchthrough component of I_{Dst} is suppressed so that it is not observed until L_{eff} becomes nearly as small as 0.5 μm. Again, the price of an increased punchthrough implant dose is an increase in S_t. This tradeoff is discussed in more detail in the section on device scaling based on I_{off} (that is, on I_{Dst} at V_{GS} = 0).

5.3.2 Modeling Punchthrough

As discussed in the previous sections, punchthrough behavior can be measured in a number of ways, including measurements which determine V_{PT}. It is also often convenient to use punchthrough voltage (V_{PT}) as a device parameter during the tasks of device scaling and device design. Unfortunately, the amount of drain current used to define V_{PT} is not universally standardized. Depending on the particular circuit application, this current level may vary from 1 pA for a DRAM to 1 nA for a static logic circuit. As a result, it would instead be useful to be able to predict the actual value of subthreshold surface-DIBL or punchthrough current in a MOSFET from the design parameters of the device prior to its fabrication. This, in fact, can be accomplished by using a 2-D device simulator which numerically solves both the Poisson equation and the carrier continuity equations to find φ, n, and p, and then uses these values to calculate the currents using the drift-diffusion current equations. Such an

entirely numerical approach, however, does not offer quantitative insight into the phenomenon of punchthrough, nor does it provide any information on trends in the device behavior as the terminal voltages are varied. However, a popular, hybrid approach that first employs a numerical method to calculate φ and then uses this solution in an analytical expression to calculate I_{Dst} (instead of first solving the carrier continuity equations) has been implemented. This technique has been widely adopted for analyzing subthreshold behavior in MOSFETs because it offers some insight into the device physics, gives accurate results (provided its use is restricted to low-current operating conditions), offers a very computationally efficient approach to device analysis, and can handle both surface-DIBL and subsurface-DIBL (punchthrough) effects in MOSFETs.

This method uses the numerical techniques described in chap. 2 (with the aid of a computer and a 2-D Poisson solver software program), to find the solution of the 2-D Poisson equation. The electrostatic potential found in this manner is used in an analytical expression to calculate the value of I_{Dst}. That is, the value of the electrostatic potential at its minimum point, together with its variation in the vicinity of this minimum point, is used in the calculation of I_{Dst}, as described below. This technique was implemented in the device simulation program called GEMINI. The program has been used to analyze many aspects of MOSFET IC behavior including many of the

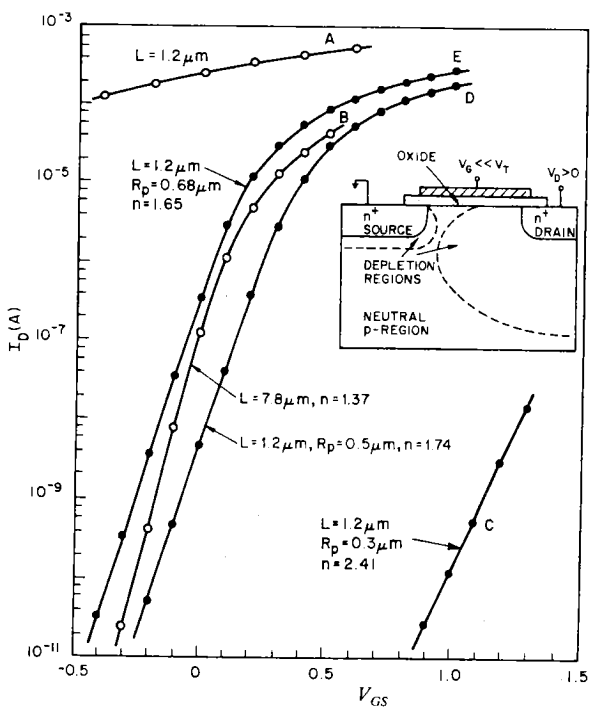


Fig. 5-28 Drain current versus gate voltage for *n*-channel devices with a substrate doping of 1.9×10^{15} atoms/cm^3, source/drain junctions are 0.47 μm deep, t_{ox} = 575Å, and V_{DS} = 5V, and V_{BS} = 0V. Devices A and B have no channel implant, and devices C and E have a boron channel implant of 8×10^{11} atoms/cm^2 at various energies.[28] (© 1978 IEEE).

topics discussed in this chapter and elsewhere in the book.[2,5,74,114,115,116]

The analytical expression for I_{Dst} used in this approach is

$$I_{Dst} = -q D_n Z \frac{t^*}{L^*} \frac{n_i^2}{N_A} \exp\left[\frac{q}{kT}(\varphi^* - V_S)\right] \left\{ 1 - \exp\left(\frac{-qV_{DS}}{kT}\right) \right\} \qquad (5 - 18a)$$

Note that this equation has the same basic form as the general expression derived for I_{Dst} in long-channel MOSFETs in chap. 4 (i.e. Eq. 4-89). When $V_{DS} \geq 0.1V$ (the only operating condition of practical interest), Eq. 5-18a reduces to the following expression, which now has a form similar to that of Eq. 4-90 (but with some important differences):

$$I_{Dst} = -q D_n Z \frac{t^*}{L^*} \frac{n_i^2}{N_A} \exp\left[\frac{q}{kT}(\varphi^* - V_S)\right] \qquad (5 - 18b)$$

That is, in Eq. 5-18, φ^* is the value of the potential at the point P between the source and drain where it has its minimum value (i.e., along the surface when surface-DIBL dominates I_{Dst}, or at the saddlepoint in the bulk region when subsurface-DIBL [punchthrough] dominates I_{Dst}), and V_S is the source potential. At P the lateral electric field (i.e., the direction of the field along the direction of current flow) becomes zero, and any current at that point must thus be due entirely to carrier diffusion. Point P is taken as the origin of a coordinate system (r,θ) with the r axis oriented along the current path and the θ axis oriented in the orthogonal direction at P (see Fig. 5-29). The

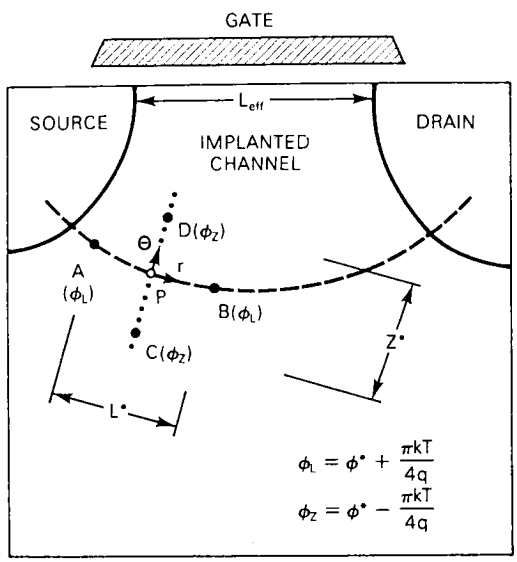


Fig. 5-29 Cross-section of a MOSFET with a bulk current path showing the two axes (r and θ) used to calculate the current. The points defining the approximate effective length L^* and effective thickness t^* are shown.[2] (© 1980 IEEE).

electrostatic potential φ^* at P corresponds to a potential minimum along the $\boldsymbol{r}$ axis and to a potential maximum along the θ axis. The value of φ^* used in Eq. 5-18b is determined by numerically solving the 2-D form of the Poisson equation in the device region. The functional variation of the electrostatic potential at some point $(\boldsymbol{r},\theta)$ in the vicinity of point P with respect to the value φ^* is $h\ (\boldsymbol{r},\theta)$. Hence, the electrostatic potential at point $(\boldsymbol{r},\theta)$ can be written as $\varphi(\boldsymbol{r},\theta) = \varphi^* + h\ (\boldsymbol{r},\theta)$.

In addition, the inversion layer thickness t_{inv} used in Eq. 4-90, is replaced in Eq. 5-18 by t^*, and the channel length L with L^*. These two terms represent, respectively, the *effective length* and *effective thickness* of the electron path near the point P, which controls the current transport. Note that when Eq. 5-18 is used to model surface-DIBL, the values of t^* and L^* are the same as in Eq. 4-90 (i.e., t^* and L^* in Eq. 5-18 in this case become t_{inv} and L). Hence, the only difference between 5-18 when it is modeling surface-DIBL-induced I_{Dst} and Eq. 4-90 is the use of φ^* instead of φ_{surf} (which was calculated using the analytical solution of the 1-D form of the Poisson equation).

On the other hand, when Eq, 5-18b is used to model punchthrough, t^* and L^* take on somewhat different meanings. Although the detailed derivations of their expressions is given in ref. 2, as shown in Figs. 5-29a and b, L^* is the distance along the $\mathbf{r}$ axis between points A and B where $\varphi = \varphi^* + \pi kT/4q$, and t^* is the distance along the θ axis between C and D where $\varphi = \varphi^* - \pi kT/4q$. The functional variation of φ about point P found from the 2-D solution of the Poisson equation is used calculate t^* and L^*.

Results obtained with this approach have been compared to solutions found from solving the carrier continuity equations as well as the Poisson equation, and with experimental data. Generally, good agreement has been found, if the operational limits of the approach are respected.[2,115]

5.3.3 Modeling the Simultaneous Occurrence of Surface-DIBL and Subsurface-DIBL

As a final note, we should recognize that up to now surface-DIBL and subsurface-punchthrough effects have been treated as separate, unconnected phenomena. However, when a device exhibits severe short-channel behavior, both phenomena are likely to be occurring simultaneously (see, for example, Figs. 5-23e & f), and the effects may be coupled. It may not even be possible to distinguish which one is dominant, either from a model or from measurements. In fact, the concepts of V_T and ΔV_{Tsc} break down in devices that exhibit severe DIBL (surface and/or subsurface), because the log I_D vs. V_{GS} plots depend sensitively on the value of V_{DS}. That, is if a V_T is defined as the V_{GS} at which a specific I_D flows, in severe short-channel-effect devices this will not be a unique value because the same specified current will be obtained at different gate voltages if V_{DS} is changed. It is for this reason that a major goal of submicron MOSFET design is to produce devices without severe subthreshold short-channel behavior, and instead exhibit subthreshold behavior comparable to that of long-channel MOSFETs.

Nevertheless, it would still be highly useful to predict the degree to which each of these effects is occurring in a MOSFET, because each surface and bulk mechanism needs a different strategy for its suppression. One paper by Skotnicki et al. offers an analytical model that allows both surface and bulk punchthrough currents to be calcu-

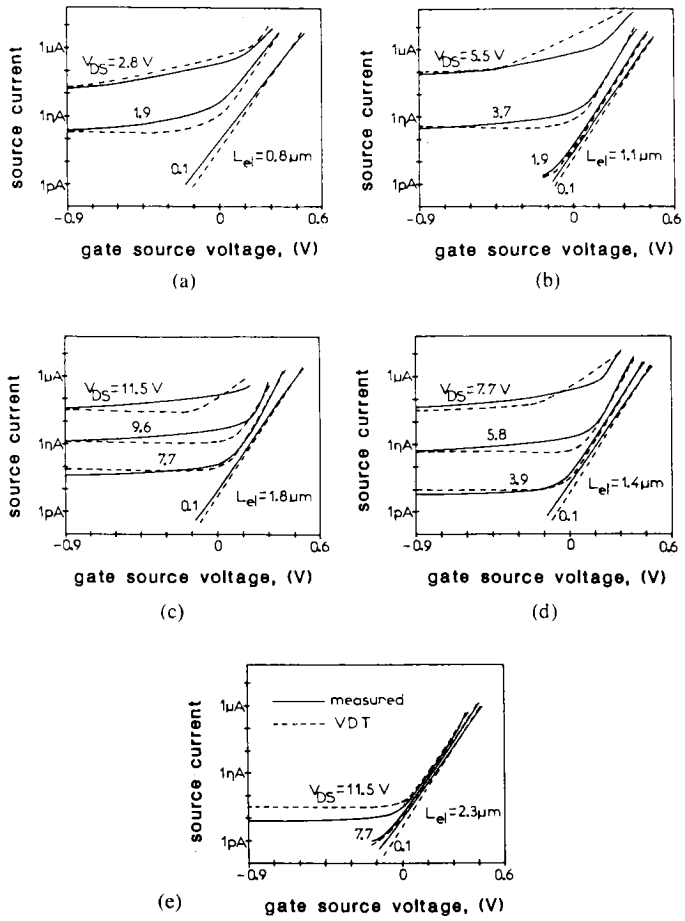


Fig. 5-30 Measured (solid line) and analytically calculated (dashed lines) I_{DSt} versus V_{GS} characteristics at V_{BS} = 0V and various V_{DS} values for a series of MOSFETs with electrical channel lengths from 0.8 to 2.3 μm. The parameters in the VDT calculations are r_j = 0.27 μm, t_{ox} = 280Å, and Z = 50 μm.[30] (© 1988 IEEE).

lated simultaneously.[30] The same voltage-doping transformation (VDT) method introduced in ref. 9 is used to reduce the 2-D Poisson equation to a 1-D form. Using the VDT-method, the actual barrier heights at the surface and in the bulk can be calculated as a function of the applied voltages and channel length using the long-channel expressions where the doping concentrations in the channel N_{ch} and N_{SUB} are replaced by reduced values N_{ch}^* and N_{SUB}^*. With the barrier heights known, the surface and bulk punchthrough currents can be determined. (This model is also valid in MOSFETs with non-uniformly doped channels.) Results reported show that good agreement with measured subthreshold currents is obtained for devices with gate lengths as small as 0. 8 μm and with V_{DS} as large as 5.5V (see Fig. 5-30).

5.4 SHORT-CHANNEL EFFECTS ON THE I-V CHARACTERISTICS OF MOSFETS OPERATED IN THE STRONG INVERSION REGIME

Short-channel effects significantly alter the dc I_D-V_{DS} characteristics of long-channel MOSFETs being operated in strong inversion in three ways. First, the combined effects of reduced gate length and gate width produce a change in V_T. Second, the mobility of the carriers in the channel is reduced by two effects, which in turn reduces I_D. These two effects are the *mobility-degradation factor* (due to the gate field), and the *velocity-saturation factor* (due to the lateral [channel] field). Third, the channel length is modulated by the drain voltage when the device is in saturation (i.e., when $V_{DS} > [V_{GS} - V_T]$), causing an increase in I_{Dsat} with increasing V_{DS} (*channel-modulation* effect). Figure 5-31 shows the I-V characteristics of an MOS device.[31] The curves in Fig. 5-31a are those of an ideal long-channel MOS device, while those in Fig. 5-31b show the effect of adding the channel modulation factor. Figure 5-31c shows the effect of adding the velocity saturation factor to the curves of Fig. 5-31b.

Here we describe some relatively simple models which allow an analytical dc model of the I_D-V_{DS} characteristics of short-channel MOSFETs to be obtained. These analytical models allow circuit designers to estimate such dc parameters as I_D and g_m for various bias conditions using only hand calculations. They also allow an intuitive understanding of the short-channel electrical behavior to be developed, making such models useful for

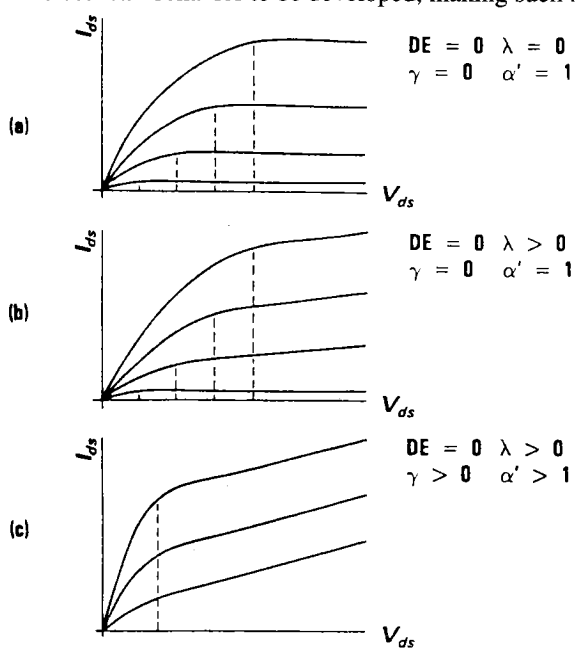


Fig. 5-31 The I-V curves of an MOS device showing the effects of progressively increasing short-channel behavior on the operation of the device in inversion. (a) Long-channel behavior; (b) with channel-length modulation; (c) addition of velocity saturation.[31] (© 1986 IEEE).

device designers and process engineers who need to relate device and process parameters to circuit parameters. Finally, they allow I_D to be calculated merely from a knowledge of device width Z, channel length L, the device threshold voltage V_T, the gate oxide thickness t_{ox}, the source/drain junction depth r_j, and with (V_{GS} - V_T) as the independent terminal voltage parameter. On the other hand, to use the elaborate models provided in circuit simulators (such as BSIM the short-channel MOSFET model in SPICE, see the next section), it is necessary to determine a large number of model parameters. These model parameters must be derived from device measurements, often with the help of automated parameter extraction tools.

5.4.1 Impact of the MOSFET Electric Fields on Mobility

In deriving the long-channel dc MOSFET models (chap. 4) we assumed that the channel mobility was constant as a function of position along the channel. We also ignored any effect on the mobility due to the transverse and longitudinal electric fields acting on the carriers. In fact, both the gate-induced electric field (i.e., also called the vertical, normal, or transverse electric field), and the drain-induced electric field (i.e., referred to also as either the lateral, horizontal, or longitudinal electric field) influence the velocity of the moving carriers in the inversion layer of the channel. Since each effect can be modeled separately, this is done first. Later we show how both effects can be incorporated into the MOSFET models, allowing their impact on the I_D-V_{DS} curves to be calculated. Before discussing how these electric fields alter the velocity of channel carriers, let us briefly review the concept of carrier mobility in a semiconductor.

The electron and hole mobilities (μ_n and μ_p, respectively), are a measure of the ease of carrier motion in a semiconductor crystal. In the semiconductor bulk (i.e., at a point far away from the semiconductor surface), the carrier mobilities are typically determined by the amount of lattice scattering and ionized impurity scattering taking place inside the material. For a given temperature and semiconductor doping, these bulk mobilities are well-defined and documented material constants. (In lightly doped silicon at 300K, μ_n and μ_p are ~1350 cm^2/V sec and ~500 cm^2/V sec, respectively.)

Carrier transport in the MOSFET, however, primarily occurs in the surface inversion layer. The lateral electric field due to V_{DS} acts to accelerate the carriers parallel to the Si-SiO_2 interface, whereupon they suffer scattering much as in the bulk. The transverse field, however, also causes the carriers to be accelerated toward the Si-SiO_2 surface. Thus, the carriers near the surface experience additional motion-impeding collisions with the silicon surface (see Fig. 5-32). As a result, the carrier mobilities at the surface are observed to be lower than in the bulk. (In the substrate doping range N_A, $N_D < 10^{17}$ cm^{-3}, the mobilities of electrons and holes in the inversion layer at $V_{GS} = V_T$ are μ_{no} ~670 cm^2/V sec and μ_{po} ~160 cm^2/V sec, respectively.) The surface-scattering mechanisms include phonons, coulombic scattering sites, and surface roughness (see chap. 7). For a good quality Si-SiO_2 interface in strong inversion, phonon scattering is generally accepted to be the dominant scattering mechanism.

At low electric fields, the drift velocity of carriers in the bulk of a semiconductor is proportional to the electric field. The mobility is the proportionality constant in the

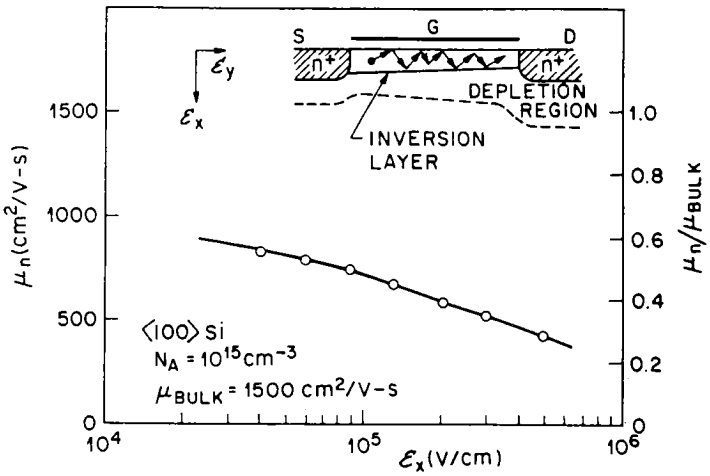


Fig. 5-32 Visualization of surface scattering at the Si-SiO_2 interface.[119] Reprinted with permission of the American Physical Society.

equation which relates the drift-velocity to the electric field (at low $\mathcal{E}$–field values). However, experimental data indicate that if the electric fields get too large (i.e., >~$3x10^3$ V/cm in the bulk), the drift velocity no longer increases linearly with increasing field. In fact, at sufficiently high fields, the drift velocities approach a saturation velocity (~10^7 cm/sec for electrons in bulk Si).

Electrons and holes drifting in the channel also reach a saturation velocity if the lateral electric fields are high enough. In fact, in MOSFETs operating in the saturation regime, carriers moving in the pinched off region of the channel travel at this saturation velocity. The surface saturation velocity of electrons is estimated to be between $6x10^6$ to $1x10^7$ cm/sec, while that of holes is between 4-to-$8x10^6$ cm/sec.[32] These saturation velocities also appear to be relatively independent of the gate (transverse) field. In the next sections we discuss the quantitative models of both mobility degradation due to gate-induced field and the velocity-lateral field relationship (velocity saturation).

5.4.1.1 Models of Mobility Degradation Due to Gate-Induced Electric Fields. As the gate bias is increased, more carriers are drawn closer to the oxide-semiconductor interface. In addition, the stronger transverse (vertical) electric field acts on these carriers. This combination of effects is thought to enhance the phonon scattering of the carriers near the surface, which thereby lowers the average carrier mobility. According to Eq. 4-65 when V_{DS} is small, i.e., $V_{DS}<<(V_{GS} - V_T)$, we can determine the mobility of the electrons in the MOSFET channel by measuring the slope of the I_D-V_{DS} characteristic at fixed V_{DS}.. When this is done, a reduction in the mobility of carriers near the surface is observed (compared to that in the bulk). Not only is the surface electron mobility under low lateral electric fields (μ_{no}) found to be only one-half to two-thirds that of the low-field bulk electron mobility, but μ_n is also observed to decrease as the transverse field is increased.

An early, widely used empirical model introduced by Crawford for calculating the degree of degradation due to the transverse field (i.e., yielding an effective mobility μ_{eff})

is[33]

$$\mu_{eff} = \mu_o / [1 + \theta \mathcal{E}_x] \quad (5 - 18)$$

$$= \mu_o / [1 + \theta \; C_{ox} \; (V_{GS} - V_T) / \kappa_{si}\varepsilon_o] \quad (5 - 19)$$

where μ_o is μ_{no} for electrons and μ_{po} for holes, $\mathcal{E}_x$ is the transverse field in the silicon at the Si-SiO_2 interface, and θ is an empirical parameter found to be technologically and substrate-bias dependent. While it is difficult to assign a coherent physical interpretation to θ, it's value has been empirically determined to be 10^{-7}-10^{-6} cm/V.

To improve the agreement between calculated and experimental values, an alternative empirical expression to that of Eq. 5-18 for the effective mobility was proposed, namely[34]

$$\mu_{eff} = \mu_o \, (\mathcal{E}_{crit}/\mathcal{E}_{eff})^{UEXP} \quad (5 - 20)$$

where $\mathcal{E}_{crit}$ and UEXP are empirically determined constants. This expression was used together with the results of Sabnis and Clemens,[35] who showed in 1978 that the channel carrier mobility on the (100) surface of silicon is a unique function of the *effective transverse field*, $\mathcal{E}_{eff}$ (which is defined as the field averaged over the electron distribution in the inversion layer). The expression of Sabnis and Clemens for $\mathcal{E}_{eff}$ is given by

$$\mathcal{E}_{eff} = [\, Q_B + (Q_I/2) \,] / \kappa_{si}\varepsilon_o \quad (5 - 21)$$

Equation 5-21 can be understood as follows: The electric field at the Si-SiO_2 interface $\mathcal{E}_{si}$ (x = 0) can be written in terms of the total charge per unit area in the Si

$$\mathcal{E}_{si} (x = 0) = Q_C / \kappa_{si}\varepsilon_o \qquad \text{(Eq. 3-17)} \quad (5 - 22)$$

If an inversion layer exists, $Q_C = Q_B + Q_I$. If no inversion layer exists, $Q_C = Q_B$. Thus, $\mathcal{E}_{si1}(x = 0) = (Q_B + Q_I)/\kappa_{si}\varepsilon_o$ and $\mathcal{E}_{si2}$ (x = 0) = $Q_B / \kappa_{si}\varepsilon_o$ when an inversion layer is present, or not, respectively. Also, if an inversion layer exists, the electric field at the boundary of the inversion layer x_i and the depletion region is the same as at x = 0 if no inversion layer exists, or $\mathcal{E}_{si}$ ($x = x_i$) = $Q_B / \kappa_{si}\varepsilon_o$. Thus, the effective transverse electric field within the channel can be approximated as the average of $\mathcal{E}_{si1}(x = 0)$ and $\mathcal{E}_{si}$ ($x = x_i$), or

$$\mathcal{E}_{eff} = [\mathcal{E}_{si1}(x = 0) + \mathcal{E}_{si} (x = x_i)] /2 \quad (5 - 23)$$

If we express Eq. 5-23 in terms of Q_I and Q_B we get

$$\mathcal{E}_{eff} = [Q_B + Q_I + Q_B] / 2\kappa_{si}\varepsilon_o = [\, Q_B + (Q_I/2) \,] / \kappa_{si}\varepsilon_o \quad (5 - 24)$$

Eq. 5-24 can be approximated as

$$\mathcal{E}_{eff} \cong C_{ox}\,[V_{GS} - V_T]/\,(2\kappa_{si}\varepsilon_o) \qquad (5 - 25)$$

Then, if Eq. 5-25 is substituted into Eq. 5-20 the following expression is obtained

$$\mu_{eff} = \mu_o\,\{\,\kappa_{si}\varepsilon_o U_{crit}\,/\,(C_{ox}\,(V_{GS} - V_T)\}^{UEXP} \qquad (5 - 26)$$

where U_{crit} is $2\mathcal{E}_{crit}$. Equation 5-26 is the model used to calculate μ_{eff} in the Level-2 MOSFET model in SPICE.*

In 1986 Liang et al[36] developed a modified form of the model of Eq. 5-26 for μ_{eff} also based on the concept of $\mathcal{E}_{eff}$. Their equation has the form:

$$\mu_{eff} = \mu_o\,/\{1 + (\mathcal{E}_{eff}/\mathcal{E}_{crit}]^{\nu}\,\} \qquad (5 - 27)$$

where $\mathcal{E}_{crit}$ and ν are empirically determined constants. Values for these parameters are shown in Table 5-2. Universal curves for μ_{eff} for electrons and holes at room temperature as a function of $\mathcal{E}_{eff}$ calculated using Eq. 5-27 are shown in Fig. 5-33.

Eq. 5-27 can also be expressed in a more useful form that explicitly relates $\mathcal{E}_{eff}$ to the device parameters and terminal voltages:

$$\mathcal{E}_{eff} \cong [(V_{GS} - V_T)/\,6t_{ox}] + [(V_T + V_a)/\,3t_{ox}] \qquad (5 - 28)$$

where it has been semi-empirically determined that $V_a \cong 0.5V$ in typical n^+ polysilicon gate MOSFETs (e.g., for NMOS and PMOS devices in an n^+ polysilicon gate CMOS technology with $V_{Tn} = 0.7V$ and $V_{Tp} = -0.7V$).[37] Note that in this equation it is assumed that $Q_I = C_{ox}\,(V_{GS} - V_T)$. Since this assumption overestimates Q_I at low $(V_{GS} - V_T)$,[38] the $\mathcal{E}_{eff}$ according to Eq. 5-28 would be higher than its actual value. But, the error is small for $(V_{GS} - V_T)>1V$.

Using Eq. 5-27 and the parameter values in Table 5-2, the carrier mobilities for NMOS and PMOS devices are plotted in Fig. 5-34 for several gate oxide thicknesses as a function of $(V_{GS} - V_T)$. These curves are useful for rough estimates of device performance or as first-order calibrators in process monitoring and control. A theoretical explanation of the electron and hole mobility dependencies on the transverse electric field has been developed by Lee et al.[39]

5.4.1.2 Models of the Carrier-Velocity Dependence on Drain Electric Field (Velocity Saturation). As we noted earlier, the carrier velocities (v) near the Si-SiO_2 interface tend to saturate at high lateral electric fields, in a manner qualitatively similar to velocity saturation in the bulk. There have been

* In the Level-1 MOSFET model, however, the effective electron mobility is not explicitly expressed in terms of the field, but is treated as a fixed value.

Table 5-2. Parameters for the Effective Mobility Models for Electrons and Holes[36]

	Electron (surface)	Hole (surface)	Hole (buried channel)
μ_0 (cm^2/ V sec)	670	160	290
$\mathcal{E}_0$ (MV/cm)	0.67	0.7	0.35
ν	1.6	1	1

several empirical models developed to describe the dependence of the carrier velocity on lateral field (ν-$\mathcal{E}_y$).[40,41,42] Two of the most widely used are

Model I $$\nu = \mu_{eff}\,\mathcal{E} \,/\, \{1 + (\mathcal{E}/\mathcal{E}_{sat})^{\alpha}\}^{1/\alpha} \qquad (5 - 29)$$

where $\mathcal{E}_{sat}$ and α are empirically determined parameters (i.e., $\alpha = 2$ for electrons and $\alpha<2$ for holes). In Model II, the α term in Model I is taken to be unity, allowing Eq. 5-29 to become

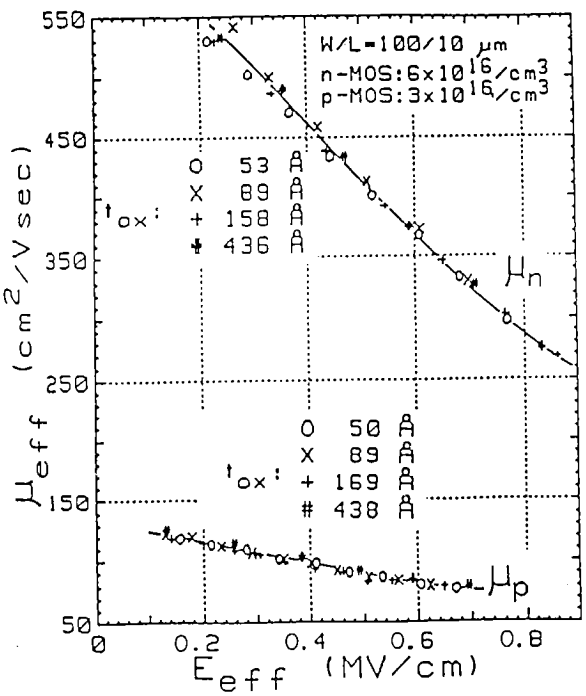


Fig. 5-33 Measured universal μ_{eff} versus $\mathcal{E}_{eff}$ curves for electrons and holes in the inversion layer.[36] (© 1986 IEEE).

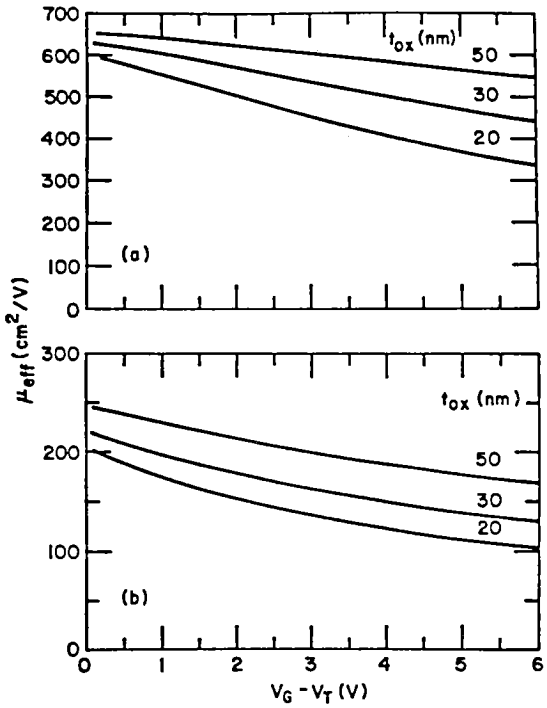


Fig. 5-34 Calculated μ_{eff} of current carrier versus (V_{GS} - V_T) for three different values of t_{OX}: (a) NMOS device and (b) PMOS devices. The "buried-channel" parameters in Table 5-2 are used for PMOS. (From [36] © 1986 IEEE) and[46] © 1988 IEEE).

Model II $$v = \mu_{eff}\,\mathcal{E} \,/\, \{1 + (\mathcal{E}/\mathcal{E}_{sat})\} \qquad (5 - 30)$$

A comparison of these two models with experimental data for electrons is shown in Fig. 5-35.[43] It is seen that Model I fits the measured data quite well, while Model II significantly underestimates the velocity at moderate fields (the same value of saturation velocity is used in both models plotted in Fig. 5-35). On this basis Model I would be preferred for estimating v. However, if Model I was inserted into an expression to calculate I_D, a solution could only be obtained by carrying out a prohibitively complex calculation.[44] Thus, to improve the accuracy of the carrier velocity model without incurring the complexity of Model I, a third, simpler, two-region piecewise model can be used. This model is defined by

Model III $$v = \mu_{eff}\,\mathcal{E} \,/\, \{1 + (\mathcal{E}/\mathcal{E}_{sat})\} \qquad \mathcal{E} < \mathcal{E}_{sat} \qquad (5 - 31a)$$

$$= v_{sat} \qquad \mathcal{E} > \mathcal{E}_{sat} \qquad (5 - 31b)$$

By using this model in the expression which calculates I_D a closed-form solution of the I_D-V_{DS} characteristics of the MOSFET is obtained. In addition, the carrier velocity saturates at $\mathcal{E}_{sat}$. $\mathcal{E}_{sat}$ can be expressed in terms of v_{sat} and μ_{eff} by matching the carrier velocities given by Eqs. 5-31a and 5-31b at $\mathcal{E} = \mathcal{E}_{sat}$. Therefore,

$$v_{sat} = \mu_{eff}\, \mathcal{E}_{sat} / [1 + (\mathcal{E}_{sat} / \mathcal{E}_{sat})] \qquad (5 - 32)$$

or

$$\mathcal{E}_{sat} = 2\, v_{sat} / \mu_{eff}. \qquad (5 - 33)$$

This model is also plotted in Fig. 5-35 over a wide range of electric fields. Note that it slightly overestimates the carrier velocity near $\mathcal{E}_{sat}$.

5.4.2 A Simple, Analytic DC Circuit Model of the Short-Channel MOSFET in Strong Inversion

Now that we have described the influence of the transverse and lateral fields on the velocity of the channel carriers, we can use the models of these effects to modify the dc circuit models of the MOSFET operating beyond threshold. In this way we develop a simple analytical dc circuit model that is valid for short-channel MOSFETs beyond threshold, and which can be used for first-order, hand calculations of I_D-V_{DS} and g_m for devices with channel lengths down to ~1 μm.[45,46]

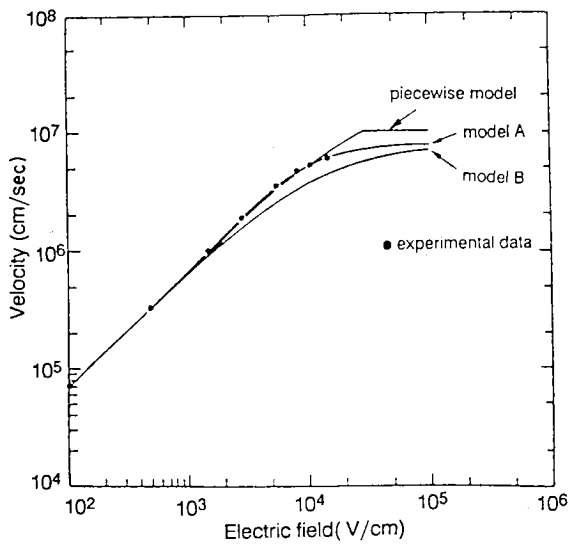


Fig. 5-35 Comparison of velocity-field models: models A and B as given in Eqs. 5-29 and 5-30, respectively. The piecewise model is given in Eq. 5-31, and μ_0 = 710 cm^2/V-sec, v_{sat} = 1x10^7 cm/sec.[38] Reprinted with permission of *Solid State Electronics.*

We first note that the factors resulting from both mobility degradation due to transverse fields and velocity saturation effects are outside of the integral in the equations used to calculate I_D (see Eqs. 4-16 and 4-19). Hence, they merely serve as multiplicative factors if incorporated into any of the long-channel dc models described in chap. 4. This means that any one of the long-channel models could be used to demonstrate the impact of these mobility effects the I_D-V_{DS} characteristics. Here, for mathematical simplicity, we elect to use the square-law model of section 4.5.5 for this purpose. To complete the dc short-channel MOSFET model, however, we will later still need to add some discussion of the short-channel effects that occur when the MOSFET is operated in saturation (i.e., the effects of channel-length modulation).

5.4.2.1 DC Circuit Model of the Short-Channel MOSFET in the Linear Regime. As noted above, all of the long-channel MOSFET dc model equations can be modified in the same manner to take into account both effects described above. For simplicity, the square-law long-channel model is chosen to be modified. In the linear-region, the modified (i.e., short-channel) version of dc I_D-V_{DS} equation of the square-law model (Eq. 4-68) is

$$I_D = \frac{\mu_{eff} C_{ox} \frac{Z}{L}}{(1 + \frac{V_{DS}}{\mathcal{E}_{sat} L})} \left[(V_{GS} - V_T) V_{DS} - \frac{V_{DS}^2}{2} \right] \qquad (5 - 34)$$

Taking the ratio of the two currents given by Eqs. 5-22 and 4-68, we get

$$I_D\,(5\text{-}34)/I_D\,(4\text{-}68) = (\mu_{eff} / \mu_n)\,[1/\{1 + (V_{DS}/\mathcal{E}_{sat}L)\}] \qquad (5 - 35)$$

Note that since $\mathcal{E}_{sat}$ and L appear as a product term in both Eqs. 5-22 and 5-23, a MOSFET is "long-channel-like" if either $\mathcal{E}_{sat}$ or L is large. Therefore, a MOSFET with a thinner gate oxide is more "long-channel-like" because the thinner oxide will produce a larger transverse field, making the carrier mobility lower so that $\mathcal{E}_{sat}$ is high.

The saturation drain voltage V_{DSsat} can next be calculated by equating the current at the drain to the current given by Eq. 5-34 when $V_{DS} = V_{DSsat}$. The current at the drain is given by the product of the carrier density and the drift velocity. The charge density at the drain, assuming that the GCA is still valid, is $ZQ_I = C_{ox}(V_{GS} - V_T - V_{DSsat})$ and the drift velocity is v_{sat}. Therefore the current is

$$I_D = Z\, v_{sat}\, C_{ox}\, [V_{GS} - V_T - V_{DSsat}] \qquad (5 - 36a)$$

$$= \frac{Z}{L\, \mu_{eff}\, C_{ox}} [V_{GS} - V_T - (V_{DSsat}/2)]\, V_{DSsat} \left[\frac{1}{1 + (V_{DS}/\mathcal{E}_{sat}\, L)} \right] \qquad (5 - 36b)$$

After some algebra, an expression for V_{DSsat} is obtained

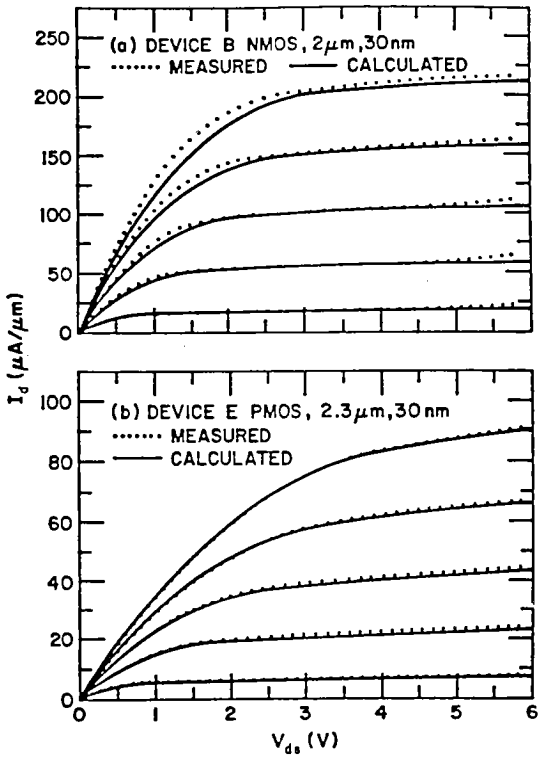


Fig. 5-36 Measured and calculated I-V characteristics for a (a) NMOS device and (b) a PMOS device.[46] (© 1988 IEEE).

$$V_{DSsat} = \frac{\mathcal{E}_{sat}L\,[V_{GS} - V_T]}{(V_{GS} - V_T) + \mathcal{E}_{sat}L} \qquad (5 - 37)$$

where $\mathcal{E}_{sat}$ is the critical field for velocity saturation (~4×10^4 V/cm).

The saturation drain current can be determined by substituting Eq. 5-37 into Eq. 5-36a to get

$$I_{Dsat} = Z\,v_{sat}\,C_{ox}\,\frac{[V_{GS} - V_T]^2}{(V_{GS} - V_T) + \mathcal{E}_{sat}L} \qquad (5 - 38)$$

The small-signal saturation transconductance g_{msat} can be found by differentiating Eq. 5-36a with respect to V_{GS}:

$$g_{msat} = Z\,v_{sat}\,C_{ox}\,[1 - (\partial V_{DSsat} / \partial V_{GS})] \qquad (5 - 39)$$

Equations 5-36, 5-38, and 5-39 form the core equations of this simple model for calculating the dc drain characteristics for short-channel MOSFETs. Some modeling results using these equations are shown in Figs. 5-36a and b.[46] (Note that the I_D values for $V_{DS} > V_{DSsat}$ in these curves are calculated using a saturation region model to be

described in the next section.) Good agreement with measured data is seen in these figures, and other work has shown good agreement for devices with channel lengths as small as 0.2 μm.[47] For example, agreement between calculated and measured values of g_m is better than ±6 percent over most of the gate voltage range for channel lengths from 3 μm down to 1 μm.

Equation 5-37 indicates that as the channel length is decreased, the channel field at the drain end will reach $\mathcal{E}_{sat}$ at a lower value of V_{DS}. Thus, V_{DSsat} is predicted to decrease with decreasing L as well. When the gate oxide thickness is made smaller, μ_{eff} decreases because $\mathcal{E}_{eff}$ is larger (see Eq. 5-27). Consequently, $\mathcal{E}_{sat}$ will increase, and so will V_{DSsat}. MOSFETs with thin gate oxides thus appear more "long-channel-like," as was discussed earlier.

The short-channel dc model of the MOSFET given by Eq. 5-36 also predicts four effects that impact the value of I_{Dsat} in short-channel devices not addressed by the long-channel dc models. First, mobility degradation reduces I_{Dsat} in short-channel devices by a factor (μ_{eff}/μ_0). Since the effective mobility decreases as V_{GS} increases, this indicates that I_{Dsat} can be significantly reduced for large gate voltages. Second, I_{Dsat} is also reduced by velocity saturation effects - to a value equivalent to the one predicted by the long-channel models (which assume a constant value of mobility, e.g., 4-68), except that the channel length appears longer than L by the factor $V_{DS}/\mathcal{E}_{sat}$.

Third, Eq. 5-36 indicates that if the product $\mathcal{E}_{sat}L$ is large compared to $(V_{GS} - V_T)$, then I_{Dsat} varies quadratically with $(V_{GS} - V_T)$. On the other hand, if $(V_{GS} - V_T) >> \mathcal{E}_{sat}L$ then I_{Dsat} becomes linearly proportional to $(V_{GS} - V_T)$. Since one of the criteria in Table 5-1 for long-channel behavior is that $I_{Dsat} \propto (V_{GS} - V_T)^2$, long-channel behavior will be exhibited by devices in which either L or $\mathcal{E}_{sat}$ is large. On the other hand, if L or $\mathcal{E}_{sat}$ are both small, such that $\mathcal{E}_{sat}L << (V_{GS} - V_T)$, the saturation regime behavior of the MOSFET will be "short-channel-like." Note that even a very short-channel device will exhibit some long-channel behavior if $(V_{GS} - V_T)$ is small enough. The drain current under this situation is given by

$$I_{Dsat} = [\mu_{eff} Z C_{ox} / 2L] (V_{GS} - V_T)^2 \quad \text{if } \mathcal{E}_{sat}L >> (V_{GS} - V_T) \qquad (5 - 40)$$

In shorter channel MOSFETs there is an observable trend toward a linear variation of I_{Dsat} with $(V_{GS} - V_T)$ instead of the quadratic dependence, as is predicted without velocity saturation. This behavior is illustrated in Fig. 5-37a.

The fourth effect predicted by Eq. 5-36 (as noted in the introduction of this chapter) is observed in the limit of very strong velocity saturation (i.e., when the channel length is very short, such that V_{DSsat} approaches zero). In such cases, Eq. 5-36 indicates that the drain current is independent of L, and I_{Dsat} is given by

$$I_{Dsat} = [v_{sat} Z C_{ox}] (V_{GS} - V_T) \qquad (5 - 41a)$$

Under this condition, even the carriers at the source [whose density is given by $C_{ox}(V_{GS} - V_T)$] are velocity saturated.

However, I_{Dsat} of a device with practical geometries is always lower than the maximum predicted by Eq. 5-41b. As a result, an ideality factor K has been defined as a useful figure of merit which indicates how close I_{Dsat} in a real device comes to that in an ideal device. This factor is inserted in Eq. 5-41b to yield

$$I_{Dsat} = K\,[v_{sat}\,Z\,C_{ox}]\,(V_{GS} - V_T) \quad \text{where } K < 1 \qquad (5\text{ - }41b)$$

Then if we compare Eqs. 5-36 and 5-41b, we obtain a definition of as

$$K = (V_{GS} - V_T - V_{DSsat})/(V_{GS} - V_T) \qquad (5\text{ - }41c)$$

This factor is a very handy parameter for device performance evaluation. That is, if an NMOS device with t_{ox} = 60 nm and L = 1 μm saturates at a V_{DS} of 1.3V when $(V_{GS} - V_T) = 4.3$, the value of K is (4.3 - 1.3)/4.3 = 0.72. This implies that shrinking the device would improve its driving current by no more than 30%. The K factor given by Eq. 5-41c is plotted against L for two different oxide thicknesses in Fig. 5-37c. The

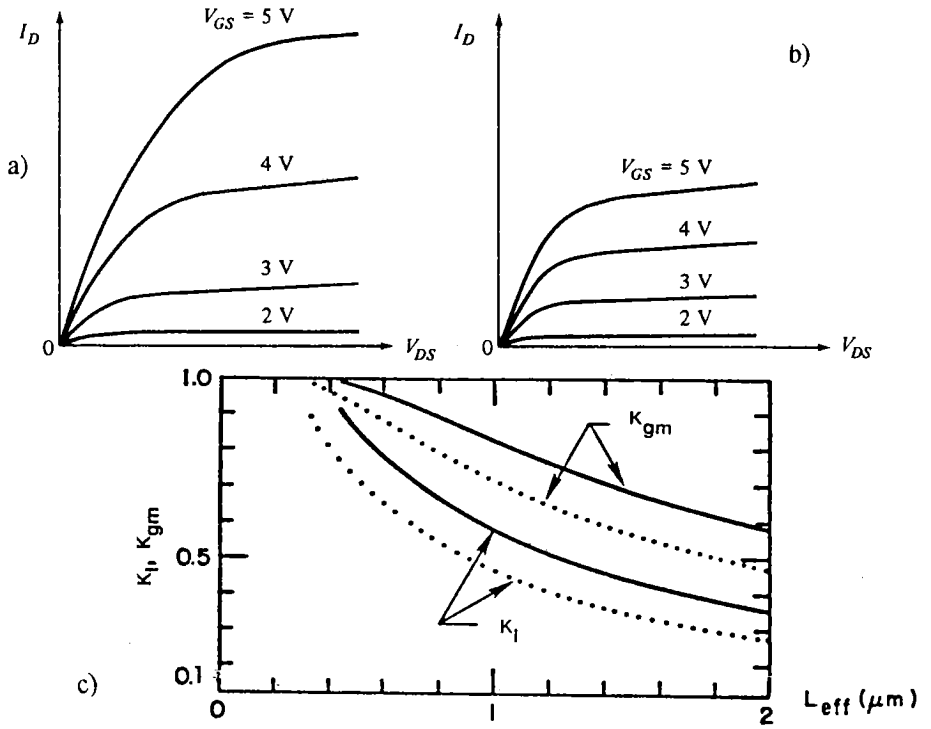


Fig. 5-37 Device characteristics (a) in the absence and (b) in the presence of velocity saturation effects. From Y.P Tsividis, *Operation and Modeling of the MOS Transistor*, McGraw-Hill, New York, 1987, p. 180. Reprinted with permission. (c) The calculated drain current ideality factor K_I versus L for different gate-oxide thicknesses. Solid lines: t_{ox} = 250Å; dotted lines t_{ox} = 100Å.[46] (© 1988 IEEE).

K of the thicker oxide device saturates at longer channel lengths. As a result, decreasing the channel length of a thick oxide MOSFET cannot significantly improve the current drive. On the other hand, as L becomes very short (e.g., of the order of 0.25 μm), K approaches unity independent of t_{ox}. Thus, I_{Dsat} becomes proportional to C_{ox}. Since, in MOSFETs with very short channels, I_{Dsat} becomes independent of channel length, the chief means of increasing I_{Dsat} will be to make t_{ox} as thin as possible (i.e., the device driving current will increase linearly with C_{ox}). Data and numerical simulation results support this argument.

As a last comment, note that because the value of v_{sat} for electrons and holes is comparable, n-channel and p-channel MOSFETs tend to perform similarly under velocity saturation, other things being equal. This is not the case under normal operation. That is, because I_D is proportional to μ_0 in normal operation, and in this case the value of μ_0 for holes in p-channel devices is one-half to one-fourth the corresponding value of μ_0 for n-channel devices.

5.4.2.2 Effects of Scaling on g_{msat}. From Eq. 5-39 the maximum available value of g_{msat} for a given t_{ox} is given by

$$g_{msat} = Z\, v_{sat}\, C_{ox} \qquad (5 - 42)$$

This is the ultimate relationship for very-short-channel MOSFETs. However, the g_{msat} of a device with practical geometries with finite channel length is always lower than this maximum. In general,

$$g_{msat} = K_{gm}\, Z\, v_{sat}\, C_{ox}, \qquad \text{where } K_{gm}<1 \qquad (5 - 43)$$

Compared to Eq. 5-39, K_{gm} can be written as

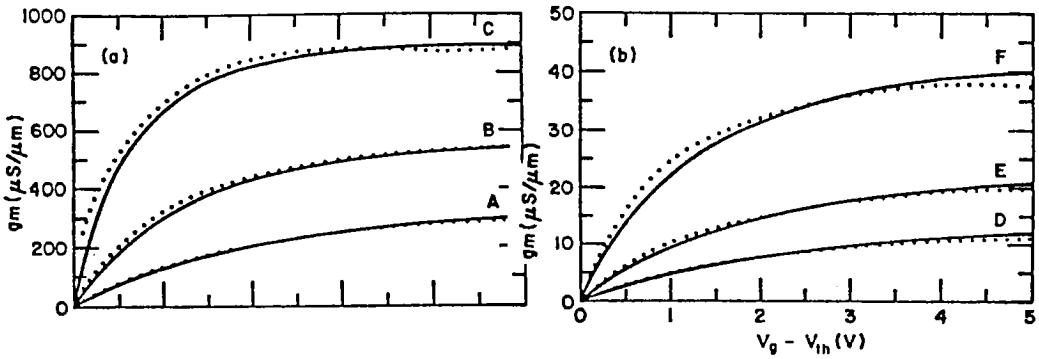
$$K_{gm} = [1 - (\partial V_{DSsat} / \partial V_{GS})] \qquad (5 - 44)$$

It is observed that g_{msat} of a short-channel device often peaks at some moderate gate bias, beyond which it appears to either remain flat or sometimes decrease slightly, as shown in Fig. 5-38. This is a consequence of the term $\partial V_{DSsat}/\partial V_{GS}$ in Eq. 5-44, which becomes almost constant beyond a gate bias roughly equal to $\mathcal{E}_{sat}L$.

5.4.2.3 Comparing the Values of I_D Calculated with the Long- and Short-Channel MOSFET Models as a Function of Channel Length. The values of I_{Dsat} as the channel length is reduced from 5 μm to 0.3 μm as calculated both by the charge-control dc circuit model of the long-channel MOSFET (Eq. 4-68), and the short channel dc circuit model (Eq. 5-38, whose basis is also the charge-control model) are given in Table 5-3. We observe that Eq. 4-68 predicts larger values of I_{Dsat} than those calculated with Eq. 5-38 (and these in turn are much closer to the measured values). This reiterates the importance of choosing the correct device model when performing technology scaling studies.

Table 5-3 Comparison of Drive Current Values versus Gate Length with the Long-Channel (Eq.4-68) and Short-Channel MOSFET Models.

Technology/Gate Length (μm)	5	3	2	1	0.5	0.3
Gate Oxide (nm)	100	60	40	20	12.5	7.5
Calaculated I_{Dsat} @ V_{GS} = 5V Long-channel Model (μA/μm)	1	2.61	6.25	20	99.3	206
Calaculated I_{Dsat} @ V_{GS} = 5V Short-Channel Model (μA/μm)	1	2.18	4.12	10.5	19.9	30



(a) (b)

Fig. 5-38 Measured and calculated g_m vs (V_{GS} - V_T) @ V_{DS} = 5V for (a) NMOS; & (b) PMOS devices with L = 3 μm (A & D); L = 2 μm (B & E); and L = 1 μm (D & F).[46] (© 1988 IEEE).

5.4.3 Modeling the Saturation Region of Operation in Short-Channel MOSFETs

For drain voltages in excess of the pinch-off voltage (V_{DSsat}) the gate-to-drain voltage drops below the voltage value needed to establish an inversion layer (i.e., it becomes smaller than the threshold voltage). Hence, long-channel MOSFET theory implies that the inversion layer no longer exists in this case, and that there are no longer any mobile carriers present in the channel at such locations. The portion of the channel extending back from the drain over which the inversion layer is extinguished is called the *pinched-off region.* The length of the pinched off region is ΔL, and its length increases as V_{DS} grows beyond V_{DSsat} (see Fig. 5-39). In addition, in saturation most of the drain voltage in excess of V_{DSsat} is dropped across this region. From a modeling perspective this makes it appear as if the normal channel length (in which the GCA is still applicable, and in which mobile charges still exist) is growing shorter. The apparent shrinkage in channel length from this effect is known as *channel-length modulation.*

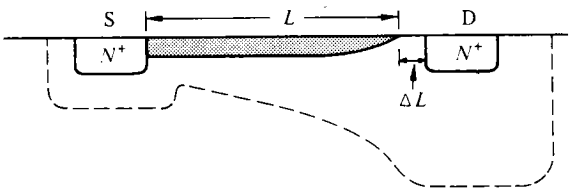


Fig. 5-39 Visualization of the velocity saturation region in a MOSFET, having a length ΔL.

In long-channel MOSFETs the length of the pinched-off region is much smaller than the original channel length (i.e. $\Delta L \ll L$), even at large values of V_{DS}. Thus, I_D remains approximately constant for drain voltages in excess of V_{DSsat}. In short-channel MOSFETs, however, ΔL can be comparable to L. Thus, the same voltage drop (V_{DSsat}) will appear across a substantially shorter channel ($L - \Delta L$). Hence, when $V_{DS} > V_{DSsat}$ in short-channel MOSFETs, I_D can increase significantly with increasing V_{DS} (as is seen in Fig. 5-31b). Since I_{Dsat} is proportional to $1/L$, the equation of drain current I_D in saturation has the general form of

$$I_D\,(V_D > V_{DSsat}) = I_{Dsat}\,[L/(L - \Delta L)] \qquad (5 - 45)$$

To be able to predict the increase in I_D in short-channel MOSFETs, the value of ΔL as a function of the terminal voltages is needed. The most general approach to calculate ΔL, however, is to simultaneously solve the current transport equation (Eq. 1-4), and the two-dimensional form of the Poisson equation (Eq. 1-3) in the pinch-off region. Unfortunately, this general approach requires numerical solution of these equations, using such 2-D device simulators as MINIMOS or PISCES (see chaps. 1 and 2). To obtain an analytical expression for ΔL some approximations must be made to reduce the 2-D Poisson equation to a 1-D form. Analytical solutions of the Poisson equation can then be found, and these solutions can be used to derive analytical expressions for ΔL. The analysis used to derive models for ΔL will also yield expressions that calculate the value of the channel electric field. These can be employed to find the magnitude of the maximum channel $\mathcal{E}$-field strength $\mathcal{E}_{ymax}$, a quantity that is needed when the subject of hot-carrier effects in MOSFETs is considered (see chap. 9).

For the purposes of modeling the MOSFET in saturation, the most common approach for reducing the 2-D Poisson equation to a 1-D form *assumes that* $\partial\mathcal{E}/\partial x$ (i.e., *the gradient of the electric field in the vertical (x) direction*) is much smaller than $\partial\mathcal{E}/\partial y$. In such cases, the $\partial\mathcal{E}/\partial x$ term in the 2-D Poisson equation can be dropped, turning a 2-D partial differential equation into the form of a one-dimensional (ordinary) differential equation. We note that this approach is opposite to the one taken in the GCA approximation. Namely, in the GCA approach the $\partial\mathcal{E}/\partial y$ term is considered to be much smaller than the $\partial\mathcal{E}/\partial x$ term, and thus it is the term dropped from the Poisson equation. Here we will derive three analytical models for ΔL. The simplification of dropping $\partial\mathcal{E}/\partial x$ entirely from the 2-D Poisson equation is used in the first two.

5.4.3.1 Modeling MOSFET Saturation Behavior Using the Abrupt Junction and Depletion Approximations. The first and simplest approach to modeling saturation behavior in MOSFETs divides the MOSFET channel into two regions: one in which the long-channel I_D models are still assumed to be valid, and the pinched-off region. This simple model also invokes three approximations in the pinched-off region, which are:

a. Only the drain field is assumed to be responsible for determining the current flow in the pinched-off region. (That is, the vertical field due to the gate is ignored.) In addition, the drain field is assumed to have only a horizontal component, $\mathcal{E}_y(y)$. Furthermore, its magnitude is assumed to be constant as a function of x. The map of the drain field in the pinched-off region is thus qualitatively depicted in Fig. 5-40. The strength of the field is indicated by the length of the arrows, and the field direction at each point by the direction of the arrow.

b. No mobile charge is present in the pinched-off region. That is, only the ionic charge [uniformly distributed with a concentration $\rho(x,y) = -qN_A$] is assumed to be present (i.e., the depletion approximation described in section 3.2.3 is once again invoked here).

c. The abrupt junction approximation is also invoked. This implies that $\mathcal{E}_y = 0$ at the leading edge of the pinched-off region $y = (L - \Delta L)_+$ (because the electric field at the edge of the depletion region in any abrupt junction is assumed to be zero).

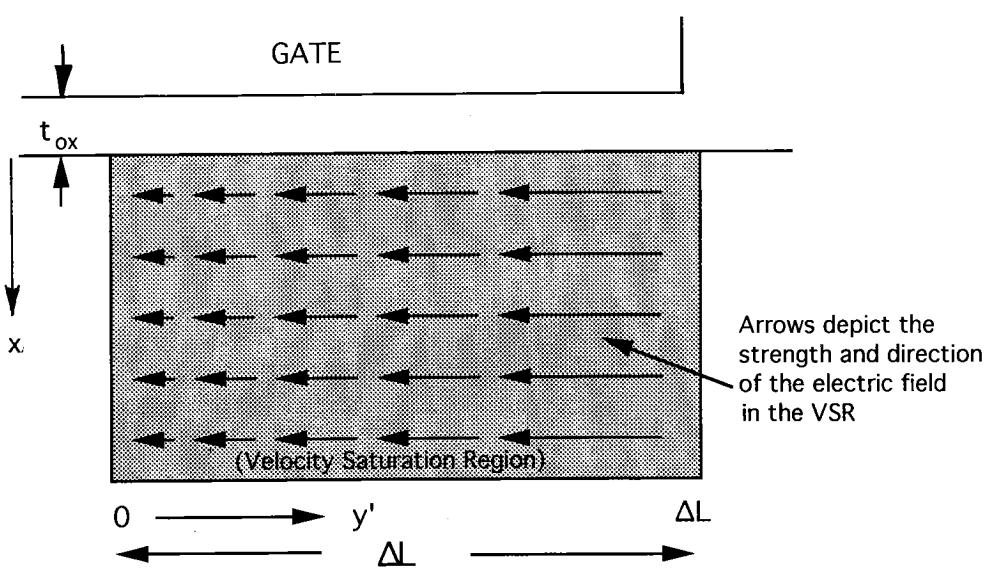


Fig. 5-40 Map of the drain field in the VSR region.

Using these approximations the Poisson equation is reduced to the form:

$$\frac{d\mathcal{E}_y(y)}{dy} = \frac{q N_A}{\kappa_{si}\,\varepsilon_o} \qquad (5-46)$$

The expression for calculating the channel length modulation is found from Eq. 5-46 to be

$$\Delta L = \sqrt{\frac{2\varepsilon_{si}}{q N_A}(V_{DS} - V_{DSsat})} \qquad (5-47)$$

This equation for calculating ΔL was introduced by Reddi and Sah in 1965.[48]

Using this method to estimate ΔL, however, leads to some physical inconsistencies in the value of $\mathcal{E}_y(y)$ as well as yielding values of ΔL that substantially overestimate its length (i.e., it predicts that ΔL can be larger than L, even in 2 μm MOSFETs under normal operating conditions). A more complex model must therefore be invoked to reconcile these apparently unphysical results, as well as to improve the predictive capability of the ΔL model. Before we describe such more-complex models let us first discuss the shortcomings of the Reddi-Sah ΔL model in more detail.

The variation of the channel electric field when the device is saturated provides some insight into the shortcoming of this ΔL model. We calculate the channel field using the approximations that led to the charge-control long-channel dc MOSFET model (adopted here again for the sake of mathematical simplicity). The channel field $\mathcal{E}_y(y)$ for $y < L - \Delta L$ can be found by substituting Eq. 4-63 into Eq. 4-22. Combining with the pinchoff conditions that led to Eq. 5-47, we get

$$\mathcal{E}_y(y) = \frac{V_{GS} - V_T}{2(L - \Delta L)} \frac{1}{1 - [y/(L - \Delta L)]} \qquad \text{for } y < L - \Delta L \qquad (5-48)$$

$$= \frac{q N_A}{\varepsilon_{si}} [y - (L - \Delta L)] \qquad \text{for } y > L - \Delta L \qquad (5-49)$$

These two equations for $\mathcal{E}_y$ are plotted as a function of y in Fig. 5-41. Note from Eq. 5-49 that $\mathcal{E}_y$ increases linearly with y in the pinched-off region. In addition, from Eq. 5-48 we see that the model predicts $\mathcal{E}_y$ goes to infinity at $y = L - \Delta L$. This occurs because at the point where the pinched-off begins (i.e., $y = L - \Delta L$), the inversion-layer charge concentration Q_I becomes zero, as required by the pinchoff (and depletion-approximation) condition. To maintain current continuity when the mobile-charge concentration approaches zero, the electric field must approach infinity. In addition, Fig. 5-41 shows that $\mathcal{E}_y$ is discontinuous at $y = L - \Delta L$ (i.e., $\mathcal{E}_y$ is infinite at $[L - \Delta L]_-$, and is zero at $[L - \Delta L]_+$). Both of these predictions are, of course, unphysical. The problem lies with the pinchoff condition. That is, the concentration of mobile charges in the pinched-off region never actually becomes zero because of carrier velocity saturation. As a result, to derive a more accurate and physically correct model of the

MOSFET in saturation, a somewhat more complex formulation of the Poisson equation than Eq. 5-46 must be employed.

It should be noted that a completely empirical approach based on the Reddi-Sah model of ΔL was developed by Frohman-Bentchkowsky and Grove.[49] In this approach, Eq. 5-47 is still used, but two terms - each containing an empirically determined constant - are added as corrections to this equation to improve the accuracy with which it predicts the value of ΔL. One of the correction terms is attributed to the fringing field between the gate and sidewall of the drain and the other to the fringing field between the gate and the end of the channel. These effects are not considered in the basic model represented by Eq. 5-47. Agreement with experiment is good for channel lengths from 2 to 4 μm and $N_{SUB} \sim 10^{15} - 10^{16}$ cm^{-3}.

5.4.3.2 Constant-Field-Gradient Model. The prediction of an infinite electric field strength at $y = L - \Delta L$ arises from the assumption that no mobile charge exists in the depletion region. If we remove this assumption (approximations **b** of the Reddi-Sah model), and also approximation **c** (which leads to the nonphysical discontinuity in the $\mathcal{E}$-field value at the leading edge of the pinched-off region), but retain approximation **a**, the formulation of the second of the three saturation models is obtained. That is, we now assume that a uniform mobile charge concentration (with a density N_m) exists to a depth r_j in the pinched-off region, and also that these carriers are velocity saturated. Hence, it is appropriate to rename the pinched-off region as the *velocity saturation region* (VSR). Now, the mobile charge density in the VSR can be expressed in the following form[50] *

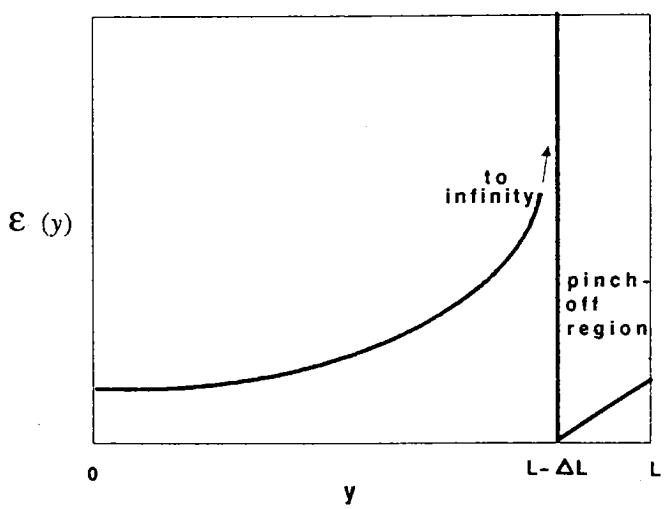


Fig. 5-41 General behavior of $\mathcal{E}(y)$, the channel field, for a device in saturation according to the long-channel model.[45] Reprinted with permission of Academic Press.

* Note that Eq. 5-50 is similar to the more familiar drift current expression $I_{drift} = J_{drift} A = q\, n \langle v \rangle A$, where N_m corresponds to n, v_{sat} to $\langle v \rangle$, and $r_j Z$ to A.

$$q\,N_m = (A_{ph}\,I_D) / v_{sat}\,r_j\,Z \qquad (5 - 50)$$

where I_D in this region is assumed to consist of velocity saturated carriers distributed uniformly into the bulk to a depth equal to the source/drain junction depths r_j, and A_{ph} is a fitting parameter. With the total charge in the VSR now expressed as the sum of the ionic and mobile charges, the one-dimensional form of the Poisson equation (Eq. 5-46) becomes

$$\frac{d\mathcal{E}_y(y)}{dy} = \frac{q\,N_A}{\kappa_{si}\,\varepsilon_o} + A_{ph}\frac{I_D}{\kappa_{si}\,\varepsilon_o\,r_j\,v_{sat}\,Z} \qquad (5 - 51)$$

Using this equation, we can solve for the value of $\mathcal{E}_y(y)$ in the VSR. When removing approximation **c** we instead apply the boundary condition that $\mathcal{E}_y(0) = \mathcal{E}_{sat}$ at the leading edge of the VSR (instead of the boundary condition that $\mathcal{E}_y = 0$ at $[L - \Delta L]_+$, as is done in the Reddi-Sah model).

Since the terms on the right hand side of Eq. 5-51 are constant as a function of y, this model indicates that the gradient of the channel electric field in the y-direction in the VSR, $d\mathcal{E}_y(y)/dy$, is also a constant (hence the name, the *constant-field-gradient model*). As a result, $\mathcal{E}_y(y)$ is predicted by this model to increase linearly with y in the VSR (as is also predicted by Reddi-Sah model). Since more charge is now assumed to exist in the VSR, however, the magnitude of the linear gradient in this model is larger. We plot this result over the channel position ΔL as the dotted line in Fig. 5-42.[54]

The channel field $\mathcal{E}_y(y)$ as calculated from a numerical solution of the two-dimensional form of the Poisson equation is also plotted in Fig. 5-42.[55] The numerical simulation predicts that $\mathcal{E}_y(y)$ exhibits a near-exponential-dependent rise towards the drain, implying that the constant-field-gradient model significantly underestimates the peak channel electric field. Because hot-carrier effects are extremely sensitive to the magnitude of the peak channel field, it has been concluded that this model is not sufficiently accurate for use in analyzing hot-carrier problems. Consequently, a more accurate analytical model of the saturation region is still needed.

5.4.3.3 Pseudo-Two-Dimensional Model of the MOSFET Operating in Saturation. The chief reason the constant-field-gradient model underestimates the maximum value of $\mathcal{E}_y$ in the VSR is that it ignores the effect of the gate field in the analysis. A pseudo-two-dimensional approach, which includes the effect of this field, has therefore been developed to obtain a more accurate analytical model of ΔL and $\mathcal{E}_y$ in the VSR. From these relationships a model for I_{Dsat} as a function of V_{DS} can also be obtained. The original pseudo-2-D model was introduced by Elmansy and Boothroyd in 1977.[51] Later, Ko et al. developed an enhanced version[52,53] that has also been applied to modeling hot-carrier effects and subthreshold MOSFET behavior (as outlined in section 5.2.1.4). Here we present a simplified version of this model, adapted from the description by Ko in references 52 and 53.

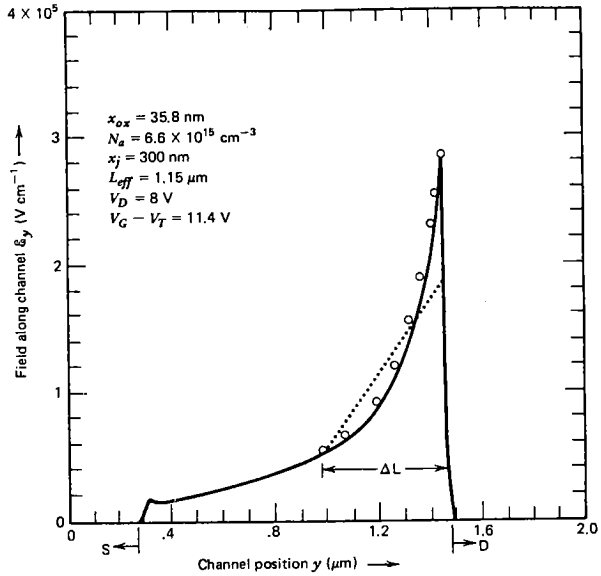


Fig. 5-42 Calculated channel field region versus position in the channel near the drain end for three models (t_{OX} =358Å, N_A = 6.6x10^{15} cm^{-3}, r_j = 0.3 μm, L_{eff} = 1.15 μm, V_{DS} = 8V, (V_{GS} - V_T) = 1.14V. Electrons in the channel have saturated velocities over the region ΔL which extends from the channel pinch-off point to the drain. From R.S. Muller and T.I. Kamins, *Device Electronics for Integrated Circuits*, 2nd Ed. Copyright John Wiley & Sons, New York, 1986. p. 491, Reprinted with permission.

In this approach we again remove approximations **b** and **c**, but also modify approximation **a** of the Reddi-Sah model. In Fig. 5-43a a schematic diagram of the VSR is shown. As in the previous approach the mobile carriers in this region are assumed to be velocity saturated and are uniformly distributed in the VSR to the depth of the source and drain junctions r_j (with a concentration N_m). The VSR is thus approximated as a box bounded on the sides by ABCD. This box encloses all of the VSR mobile charges and most of the ionic charges of interest. The electric field and voltage at the leading edge of the VSR (point A) are assumed to have a values $\mathcal{E}_{sat}$ and V_{DSsat}, respectively. Since in steady-state the total number of mobile charges in the VSR is constant with time the phenomenon being analyzed is essentially an electrostatic problem, solvable using the Poisson equation.

Approximation **a** is modified as follows: First, let the coordinate system be selected such that $y = 0$ at A and $y = \Delta L$ at D. Next, we see that the charge enclosed in the VSR is the sum of the mobile charge and ionic charge concentrations, qN_A and qN_m, respectively, or:

$$\rho\,(x,y) = qN_A + qN_m \qquad (5 - 53)$$

The drain $\mathcal{E}$-field is still assumed to have only a horizontal component $\mathcal{E}_y$ that varies with position y, but not with position x. The map of the drain field component in the

VSR bounded by ABCD is shown in Fig. 5-43b. The effect of the field due to the gate charge, however, is now also included in the Poisson equation. This field is assumed to have only a vertical component $\mathcal{E}_x$ which varies with position y (i.e., $\mathcal{E}_x$ gets weaker as the drain is approached). $\mathcal{E}_x$ has a value at the Si-SiO_2 interface given by $\mathcal{E}_x(0,y)$ and is assumed to be zero at the bottom edge of the depletion region [$\mathcal{E}_x(d_{max},y) = 0$]. The value of $\mathcal{E}_x(r_j,y)$ is also assumed to be close to zero, even though r_j is not as deep as d_{max}. (This latter assumption is plausible since the MOSFET is biased into strong inversion, making qN_m in the VSR larger than qN_A. Thus, most of the total charge in the VSR region of the channel [i.e., from $y = 0$ to $y = \Delta L$], is confined within the depth from $x = 0$ to $x = r_j$). *It is now assumed that the value of* $\partial\mathcal{E}_x/\partial x$ *at each point y can be replaced by the average value of* $\partial\mathcal{E}_x/\partial x$ *from* $x = 0$ *to* $x = r_j$. This average value is given by [$\mathcal{E}_x(0,y) - \mathcal{E}_x(r_j,y)/r_j$]. Since $\mathcal{E}_x(r_j,y)/r_j \sim 0$, the average value of $\partial\mathcal{E}/\partial x$ can be approximated as just $\mathcal{E}_x(0,y)/r_j$.

The above modifications to approximation **a** result in the following important conclusion: *The term* $\partial\mathcal{E}/\partial x$ *in the Poisson equation can be approximated as* $\mathcal{E}_x(0,y)/r_j$. With this approximation, the 2-D Poisson equation which governs the relation between the electric-field and the charge density in the VSR, can be expressed as follows (with $\mathcal{E}_x(0,y)$ being written as $\mathcal{E}_{si}(0,y)$):

$$\frac{d\mathcal{E}_y(y)}{dy} + \frac{\mathcal{E}_{si}(0,y)}{r_j} = \frac{q\,N_A}{\kappa_{si}\,\varepsilon_o} + \frac{q\,N_m}{\kappa_{si}\,\varepsilon_o} \qquad (5 - 54)$$

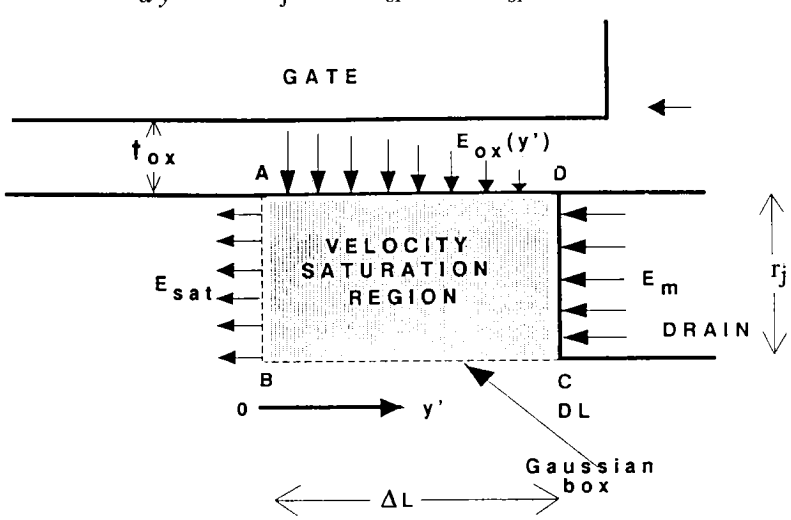


Fig. 5-43 Schematic diagram to illustrate the analysis of the velocity saturation region (VSR). From P.K. Ko, "Approaches to Scaling," Chap. 1, in *Advanced MOS Device Physics*, N.G. Einspruch and G. Gildenblatt, Eds., Vol. 18 VLSI Electronics. Academic Press 1989. Reprinted with permission.

Thus, by using this modified approximation, not only is the Poisson equation reduced to a 1-D form (i.e., dependent only on y), but in addition, the effects of the both the vertical (gate) and the horizontal (drain) fields in the VSR are included. Consequently, the name chosen for this model (i.e., the *pseudo-2-D model*) is appropriate.

Now, we can further simplify Eq. 5-54 by noting that $\mathcal{E}_{si}$ $(0,y)$ can be described in terms of $\mathcal{E}_{ox}$ (according to Eq. 3-15) as

$$\mathcal{E}_{si}(0,y) = (\varepsilon_{ox}/\kappa_{si}\varepsilon_o)\, \mathcal{E}_{ox}\,(0,y) \qquad (5 - 55)$$

Substituting the right hand side of Eq. 5-55 into Eq. 5-54 for $\mathcal{E}_{si}$, and then recalling that we since we can express $\mathcal{E}_{ox}$ $(0,y)$ as

$$\mathcal{E}_{ox}(0, y) = [V_{GS} - V_{FB} - 2\varphi_B - V(y)]/t_{ox} \qquad (5 - 56)$$

[see Eq. 3-14, with $\varphi_{ox} = V_{GS} - V_{FB} - 2\varphi_B - V(y)$], we can make these both of changes in Eq. 5-54 to get

$$\frac{d\mathcal{E}_y(y)}{dy} + \frac{\varepsilon_{ox}}{\kappa_{si}\,\varepsilon_o}\,\frac{1}{t_{ox}}\,\frac{[V_{GS} - V_{FB} - 2\varphi_B - V(y)]}{r_j} = \frac{q\,N_A}{\kappa_{si}\,\varepsilon_o} + \frac{q\,N_m}{\kappa_{si}\,\varepsilon_o} \qquad (5 - 57)$$

Since $V(y = 0) = V_{DSsat}$ we can use to Eq. 5-56 to write

$$\mathcal{E}_{ox}(y = 0) = \frac{1}{t_{ox}}[V_{GS} - V_{FB} - 2\varphi_B - V_{DSsat}] \qquad (5 - 58)$$

In addition, since the GCA approximation is still valid at the leading edge of the VSR ($y = 0$), the total charge in the silicon (Q_C) at that point can still be calculated by Eq. 4-1c ($Q_C = \varphi_{ox}C_{ox}$), where $C_{ox} = \varepsilon_{ox}/t_{ox}$ and $\mathcal{E}_{ox} = \varphi_{ox}/t_{ox}$. Thus, since Q_C (cm^{-2}) is ($qN_Ar_j + qN_mr_j$), and $Q_C = \mathcal{E}_{ox}\varepsilon_{ox}$, we can also write

$$\mathcal{E}_{ox}(y = 0) = (q\,N_A\,r_j + q\,N_m\,r_j)/\varepsilon_{ox} \qquad (5 - 59)$$

Hence,

$$\frac{\varepsilon_{ox}}{\kappa_{si}\,\varepsilon_o}\,\frac{1}{t_{ox}}[V_{GS} - V_{FB} - 2\varphi_B - V_{DSsat}] = \frac{q\,N_A}{\kappa_{si}\,\varepsilon_o}\,r_j + \frac{q\,N_m}{\kappa_{si}\,\varepsilon_o}\,r_j \qquad (5 - 60)$$

Now we can substitute $(\varepsilon_{ox}/\kappa_{si}\varepsilon_o t_{ox} r_j)[V_{GS} - V_{FB} - 2\varphi_B - V_{DSsat}]$ for the right hand side of Eq. 5-57, allowing Eq. 5-57 to be simplified to

$$\frac{d\mathcal{E}_y(y)}{dy} = \frac{\varepsilon_{ox}}{\kappa_{si}\,\varepsilon_o\,t_{ox}\,r_j}\,[V(y) - V_{DSsat}] \qquad (5 - 61)$$

or

$$\frac{d\mathcal{E}_y(y)}{dy} = \frac{[V(y) - V_{DSsat}]}{l^2} \qquad \text{where } l^2 = \frac{\kappa_{si}\varepsilon_o}{\varepsilon_{ox}}\,t_{ox}\,r_j \qquad (5 - 62)$$

or

$$\frac{d^2[V(y) - V_{DSsat}]}{dy^2} = \frac{[V(y) - V_{DSsat}]}{l^2} \tag{5 - 63}$$

where the parameter l is the effective length of the VSR.

A qualitative physical interpretation of equations 5-62 and 5-63 can be given as follows: At $y = 0$, where $V(0) = V_{DSsat}$, the charge in the VSR is still controlled entirely by the vertical field ($\mathcal{E}_x$). Since, for conditions of dc current flow the total charge density (i.e., the sum of the mobile and ionic charge densities) must remain constant along the VSR toward the drain, the total electric field at each point y must also remain constant. However, the voltage $V(y)$ increases toward the drain. Thus, from Eqs. 5-55 and 5-56, we see that $\mathcal{E}_x$ must decrease toward the drain, and so must the amount of charge in the VSR controlled by it. According to the Poisson equation, this charge must be taken up by the gradient of $\mathcal{E}_y$. Thus, the gradient of $\mathcal{E}_y$ must increase toward the drain. In Eq. 5-62, the RHS is the amount of charge released by $\mathcal{E}_x$ as a result of the rise in the channel voltage [equal to $V(y) - V_{DSsat}$], while the LHS is the corresponding increase of the gradient in $\mathcal{E}_y$ in order to support these charges.

The general solution of Eq. 5-63 is:

$$[V(y) - V_{DSsat}] = A\, e^{y/l} + B\, e^{-y/l} \tag{5 - 64}$$

By applying the boundary conditions: $V(0) = V_{DSsat}$, and $\mathcal{E}_y(0) = [\partial V(y)/\partial y]_{@y=0} = \mathcal{E}_{ysat}$, the coefficients in Eq. 5-64 are found to be $A = l\mathcal{E}_{sat}/2$ and $B = -\, l\mathcal{E}_{sat}/2$. Thus, we can write the expressions for $V(y)$ and $\mathcal{E}_y(y)$ in the VSR as

$$V(y) = V_{DSsat} + l\,\mathcal{E}_{sat} \sinh(y/l) \tag{5 - 65a}$$

and

$$\mathcal{E}_y(y) = dV/dy = \mathcal{E}_{sat} \cosh(y/l) \tag{5 - 66a}$$

Note that near the drain end of the VSR

$$V(y) \approx V_{DSsat} + l\,\mathcal{E}_{sat} \exp(y/l) \tag{5 - 65b}$$

and

$$\mathcal{E}_y(y) \approx [(l\mathcal{E}_{sat})/l] \exp(y/l) \tag{5 - 66b}$$

At the drain end of the channel $V(y)$ and $\mathcal{E}_y$ have their maximum values V_{DSmax} and $\mathcal{E}_{ymax}$, as given by

$$V_{DSmax} = V_{DSsat} + l\,\mathcal{E}_{sat} \sinh(\Delta L/l) \tag{5 - 67}$$

and

$$\mathcal{E}_{ymax} = \mathcal{E}(y = \Delta L) = \mathcal{E}_{sat} \cosh(\Delta L/l) \tag{5 - 68}$$

Equations 5-67 and 5-68 can be combined to yield an expression for ΔL

$$\Delta L = l \ln \frac{\{[V(y) - V_{DSsat}]/l\} + \mathcal{E}_{ymax}}{\mathcal{E}_{sat}} \tag{5 - 69}$$

and

$$\mathcal{E}_{ymax} = \left[\frac{[V_{DS} - V_{DSsat}]^2}{l^2} + \mathcal{E}^2_{sat}\right]^{1/2} \tag{5 - 70a}$$

Equations 5-66 to 5-70 describe important aspects of the behavior of the MOSFET in saturation. First, Eq. 5-66b indicates that $\mathcal{E}_y$ grows almost exponentially towards the drain, as was predicted by 2-D numerical simulation. In Fig. 5-42, Eq. 5-66a is plotted along with the results of the 2-D numerical simulation. Close agreement is observed. However, the parameter l must be adjusted to give approximately the same peak channel field. We indicate below how this has been done using both experimental data and numerical simulation.

From Eq. 5-70, $\mathcal{E}_{ymax}$ is seen to depend explicitly on $[V(y) - V_{DSsat}]$, but the channel length L only influences $\mathcal{E}_{ymax}$ indirectly. Furthermore, if V_{DS} is more than about 2V greater than V_{DSsat}, $(V_{DS} - V_{DSsat})/l >> \mathcal{E}_{sat}$. In this case Eq. 5-70a for $\mathcal{E}_{ymax}$ is simplified to

$$\mathcal{E}_{ymax} \approx [V_{DS} - V_{DSsat}]/l \tag{5 - 70b}$$

From experimental data and 2-D device simulations using MINIMOS the following expressions for l that appear valid over a wide range of t_{ox}, N_A, and r_j have been deduced:[1]

$$l \approx 0.22\, t_{ox}^{1/3}\, r_j^{1/3} \qquad \text{if } t_{ox} \geq 15 \text{ nm} \tag{5 - 70c1}$$

and

$$l \approx 1.7\text{x}10^{-2}\, t_{ox}^{1/8}\, r_j^{1/3} \qquad \text{if } t_{ox} < 15 \text{ nm} \tag{5 - 70c2}$$

where l, r_j and t_{ox} are in centimeters.

By substituting Eq. 5-70c1 into Eq. 5-70b a simple, explicit expression for the maximum channel electric field is obtained

$$\mathcal{E}_{ymax} \approx [V_{DS} - V_{DSsat}]/0.22\, r_j^{1/3}\, t_{ox}^{1/3} \tag{5 - 71}$$

where V_{DSsat} is found using Eq. 5-37 and $\mathcal{E}_{sat}$ in that equation is taken to be the critical field for velocity saturation (~$5\text{x}10^4$ V/cm). Equations. 5-71 and 5-37 can thus be used to estimate the peak channel electric field $\mathcal{E}_{ymax}$ as a function of device dimensions and bias voltages in conventional (i.e., non-graded drain) MOSFETs.

For example, by substituting some realistic values of t_{ox} and r_j into Eq. 5-71 (i.e., using values typical of 1-μm CMOS technology, namely, t_{ox} = 250Å and r_j = 0.2 μm), we get l = 135 nm, and $1/l = 7.5\text{x}10^4\ \text{cm}^{-1}$. At low $(V_{GS} - V_T)$ values, $V_{DSsat} \sim 0$. Using $\mathcal{E}_{sat} = 5\text{x}10^4$ V/cm and V_{DS} = 5V, from Eq. 5-71 we get $\mathcal{E}_{ymax} = 3.85\text{x}10^5$ V/cm. Since $\mathcal{E}_{ymax}$ is proportional to $(1/r_j^{1/3}\, t_{ox}^{1/3})$, it is apparent that device scaling will magnify the high-field problems unless the power-supply voltage is also correspondingly reduced, or the MOSFET device structure is modified to reduce $\mathcal{E}_{ymax}$.

With an expression for ΔL also in hand (Eq. 5-69), a relationship for the drain current in saturation based on the pseudo-2D saturation model can now be derived, namely for $V_{DS} > V_{DSsat}$:

$$I_D(\Delta L) = I_{Dsat}\left[\frac{L}{L - \Delta L}\right]\left[\frac{V_{DSsat} + \mathcal{E}_{sat} L}{V_{DSsat} + \mathcal{E}_{sat}(L - \Delta L)}\right] \tag{5 - 72}$$

5.5. SHORT-CHANNEL MOSFET MODELS USED IN THE *SPICE* CIRCUIT SIMULATOR

The simple, analytical short-channel MOSFET model described in section 5.4 is adequate for hand calculations. However, for computer-aided circuit analysis and design a more accurate model is needed, which also retains the benefit of simplicity to allow rapid circuit-simulation execution times. The Level-1 and Level-2 MOSFET models in SPICE are not suitable since they do not work well for small geometry transistors. Likewise, the Level-3 model relies too heavily on a semi-empirical approach. Thus, a simple and accurate, physically based, short-channel MOSFET model suitable for circuit simulation programs was needed.

Such a model has been developed and is incorporated in the circuit simulator SPICE. This model, named BSIM (Berkeley Short-channel IgFET Model), was introduced in 1987.[56] It is a modified version of the Bell Labs model CSIM with substantial enhancements (Compact Short-Channel IgFET Model).[57] BSIM is the Level-4 MOSFET model in SPICE, and BSIM2 (introduced in 1990) is the Level-5 MOSFET model.[58] By 1993, BSIM3 was developed and incorporated into SPICE.

BSIM is capable of simulating the drain current in short channel MOSFETs in both the strong-inversion and weak inversion regimes. It can model such short-channel effects as channel-length shortening, channel narrowing, mobility degradation, drain-induced barrier lowering, nonuniform channel doping for ion-implanted devices, subthreshold slope, and geometric dependencies. Terms with strong physical meaning are employed to model the fundamental physical effects while empirically derived parameters are judiciously used to embrace less-well-understood (and generally subtle) device characteristics. This approach is well-suited for circuit-analysis purposes, especially as two- and three dimensional small geometry effects become more important.

BSIM2 is an extension of the original BSIM model, but BSIM3 is essentially a new model. That is, while BSIM has about 60 parameters (17 for modeling the dc characteristics) and BSIM2 about 120, BSIM3 has 38 (the Level-1 and Level-2 SPICE models described in chap. 4 each have a total of about 15). It is generally conceded that while the Level-1 and Level-2 models had too few parameters, BSIM and BSIM2 had too many. BSIM3 is a compromise that seeks to use just the right number. In addition, many BSIM and BSIM2 parameters were put into the model strictly for curve fitting and did not have much physical basis. This makes it difficult to generate fast and slow models using BSIM or BSIM2. For the most part, the BSIM3 parameters are based on reasonable physical assumptions, and almost all have some connection with the real world, making the results of BSIM3 easier to interpret.

BSIM also includes a dedicated parameter extraction program, which together with the parameter-extraction system hardware (consisting of three major elements: a desktop computer, an H-P 4145 semiconductor parameter analyzer, and a probe station) allows the BSIM model parameters for a family of devices to be automatically extracted from measurements of actual test devices.

5.5.1 BSIM Model Equations

The formulation of BSIM can be broken down into the following aspects: (1) an equation for the threshold voltage V_T of the short-channel MOSFET; (2) equations for the operation of the short-channel MOSFET in strong inversion; and (3) equations for the operation in weak inversion (subthreshold regime).

5.5.1.1 BSIM Threshold Voltage Equation. The equation used to model V_T in long-channel MOSFETs (i.e., Eq. 4-12, with V_{BC} replaced with V_{BS}), is modified in BSIM to allow the threshold lowering effect observed in short-channel MOSFETs to be modeled. The BSIM V_T equation is

$$V_T = V_{FB} + 2\varphi_B + \gamma(2\varphi_B + V_{BS})^{1/2} - K_T(2\varphi_B - V_{BS}) - \eta V_{DS} \qquad (5 - 73)$$

We observe that in addition to the three parameters V_{FB}, φ_B, and γ, Eq. 5-73 contains two more for modeling V_T, namely K_T and η. K_T is the source and drain depletion charge sharing coefficient, and η is the drain-induced barrier lowering coefficient (which is roughly inversely proportional to the channel length L). Together, K_T and η model the threshold lowering effect.

5.5.1.2 BSIM Equations for I_D when the MOSFET is Operated in Strong Inversion. The BSIM I_D equations contain three additional parameters to allow the mobility degradation and velocity saturation effects to be modeled. These are the vertical field mobility degradation coefficient U_0, the velocity saturation coefficient U_1, and the carrier mobility μ_0. (Note that μ_0 in BSIM is a function of the substrate and drain biases.) In order to speed up the circuit simulation time, instead of using I_D in the form of the ionic charge model (i.e., of the form of Eq. 4-57 - which has a 3/2 power dependence of the drain current on the drain and substrate bias), a numerical approximation to the terms of Eq. 4-57 that contain V_{BS} is used. The resulting BSIM equations for I_D in the linear and saturation regions are:

Linear Region: [$V_{GS} > V_T$), and $0 < V_{DS} < V_{DSsat}$]

$$I_D = \frac{\mu_0}{[1 + U_0(V_{GS} - V_T)} \frac{C_{ox}\frac{Z}{L}}{(1 + \frac{U_1}{L}V_{DS})}\left((V_{GS} - V_T)V_{DS} - \frac{aV_{DS}^2}{2}\right) \qquad (5 - 74)$$

where

$$a = 1 + \frac{\gamma}{2\sqrt{2\varphi_B - V_{BS}}}\left[1 - \frac{1}{1.744 + 0.8364\,(2\varphi_B - V_{BS})}\right] \qquad (5 - 75)$$

If we compare Eq. 5-74 with the linear I_D equation of the square-law model (Eq. 4-63b), we see that Eq. 5-74 is similar but that it also has some extra terms. The first fractional term of Eq. 5-74 models the vertical field mobility degradation effect, and has the same form as Crawford's mobility expression (Eq. 5-18). The denominator of the second fractional term models the velocity saturation effects. The last term of Eq. 5-74 is like the last term of Eq. 4-63b except that it also contains the empirical parameter 'a.' This formulation allows the terms of Eq. 4-57 containing V_{BS} to be handled via a numerical approximation, rather than by having to compute terms to the 3/2 power (as would have to be done if Eq. 4-63 was used here instead to compute I_D). The values of I_D computed with this approximation are quite close to the values that would be computed if the Eq. 4-57 was used instead (over a reasonable range of V_{BS} and V_{DS}). See Appendix C to see how this approximation is derived, starting from Eq. 4-57.

Saturation Region: $[V_{GS} > V_T)$, and $V_{DS} \geq V_{DSsat}]$

$$I_D = \frac{\mu_o}{[\,1 + U_o(V_{GS} - V_T)} \; \frac{C_{ox}\frac{Z}{L}(V_{GS} - V_T)^2}{2aV} \qquad (5 - 76)$$

where

$$V = \frac{1 + v_c + \sqrt{1 + 2\,v_c}}{2} \quad \text{and} \quad v_c = U_1(V_{GS} - V_T)/a \qquad (5 - 77)$$

and

$$V_{DSsat} = (V_{GS} - V_T)/aV \qquad (5 - 78)$$

5.5.1.3 BSIM Equations for I_D when the MOSFET is Operated in Weak Inversion (Subthreshold Regime). In BSIM the total drain current is modeled as the linear sum of the strong and weak inversion drain current components ($I_{D,S}$ and $I_{D,W}$, respectively) to allow proper matching of these two components at a transition point close to the threshold voltage. The weak inversion component is expressed as

$$I_{D,W} = (I_{exp}\, I_{limit})/(I_{exp} + I_{limit}) \qquad (5 - 79)$$

where

$$I_{exp} = \frac{\mu_o\, C_{ox}\, Z}{L}\left(\frac{kT}{q}\right)^2 e^{1.8}\, e^{(q/kT)(V_{GS} - V_T)/\eta}\left[\,1 - e^{-V_{DS}(q/kT)}\right] \qquad (5 - 80)$$

and

$$I_{limit} = \frac{\mu_o\, C_{ox}\, Z}{L}\left(3\,\frac{kT}{q}\right)^2 \qquad (5 - 81)$$

The factor $e^{1.8}$ is empirically chosen to achieve best fits in the subthreshold characteristics with minimum effect on the strong-inversion characteristics. The

subthreshold parameters η_O, η_B, and η_D are used to model the subthreshold slope coefficient η, as:[59]

$$\eta = \eta_O + \eta_B V_{BS} + \eta_D V_{DS} \qquad (5 - 82)$$

The subthreshold swing S_t is expressed as (Eq. 4-91)

$$S_t = (2.3kT/q)\,\eta = (2.3kT/q)\,(1 + C_d/C_{ox}) \qquad (5 - 83)$$

where η is the subthreshold slope coefficient calculated from Eq. 3-82. As discussed in chap. 4 and earlier in this chapter, the subthreshold swing is a function of both substrate and drain biases, and this bias dependence is modeled in BSIM by the linear relation given by Eq. 5-82.

5.6 MOSFET SCALING

A roadmap of the scaling of MOSFET devices was proposed by Hu in 1993.[24] He assumed that a new generation of technology will continue to be developed every three years, with perhaps a slow-down to four years beyond the 0.35 μm generation. His proposal for MOSFET scaling is outlined as follows, with the result summarized in Table 5-4.

1. The IC industry/market will agree on the next power-supply voltage standard V_{CC} several years in advance of introducing a technology/product using that voltage.

2. For a given V_{CC}, the thinnest gate oxide (i.e., from manufacturability, reliability and GIDL considerations) will be used to get maximum I_D.

3. Junction depths will be scaled aggressively to keep the short-channel effects (i.e., subthreshold leakage current) within a desired limit.

4. V_T of general purpose technology will remain basically unchanged. A significant reduction in V_T is unacceptable for channel subthreshold leakage.

5. The well doping concentration or punchthrough implant dose will be increased, and the gate length may be chosen to be larger than the minimum feature size in order to achieve acceptable leakage and standby current.

6. Drain engineered structures will be used as necessary to meet the constraints of hot-carrier reliability, breakdown voltage, and GIDL.

5.6.1 Design of Submicron MOSFETs: Scaling Guidelines

The fundamental issue of downsizing MOSFETs is to preserve long-channel characteristics after miniaturization. Several approaches have been proposed as roadmaps for designing submicron MOSFETs (L<1 μm) so that they exhibit such behavior.

Table 5-4 MOSFET Technology Projection[24] (High-Speed Scenario)

Minimum Feature Size (μm)	0.5	0.35	0.25	0.18	0.13	0.09
Year of Introduction	1991	1994	1997	2001	2005	2009
SRAM Density	4M	16M	64M	256M	1G	4G
V_{CC} (V) (+10% for max V_{CC})	5	3.3	3.3	3.3	2.2	2.2
MOSFET Gate Oxide (nm)	13.5	9	8	7	4.5	4
Junction Depth (μm)	0.15	0.15	0.1	0.08	0.08	0.07
Effective Channel Length (μm)	0.40	0.30	0.25	0.2	0.15	0.13
Threshold Voltage (V)	0.7	0.7	0.7	0.7	0.7	0.7
NMOS I_{Dsat} @ $V_{GS} = V_{CC}$ (mA/μm)	0.64	0.48	0.55	0.65	0.51	0.57
PMOS I_{Dsat} @ $V_{GS} = V_{CC}$ (mA/μm)	0.31	0.22	0.26	0.32	0.24	0.28

Three of the most important scaling methodologies are the following: (1) constant electric-field scaling and its derivatives (constant voltage and constant electrostatic scaling); (2) scaling based on subthreshold behavior; (3) scaling to achieve a desired value of subthreshold current, I_{off}. Because subthreshold conduction is the dominant short-channel performance issue that limits MOSFET scaling,[60] the latter two have more applicability, but for completeness we also include a brief discussion of approach #1.

5.6.1.1 Constant Electric-Field, Constant Voltage & Electrostatic Scaling. The method of *constant electric-field scaling* for designing submicron MOSFETs was proposed by Dennard et al. in 1974.[61] In this approach, a successful larger device structure is selected (e.g., from the previous generation of ICs) and all its dimensions and voltages are reduced by a constant scaling factor λ (>1). The new device dimensions are then $L' = L/\lambda$, $t_{ox}' = t_{ox}/\lambda$, and $W' = W/\lambda$, while the operating voltages are $V' = V/\lambda$. With this approach, operating speed increases, component density increases, and power density remains constant. Unfortunately, current density also increases by the scaling factor (a condition that aggravates electromigration failure in metal interconnect lines). In addition, the threshold voltage V_T' is reduced to V_T/λ, which makes the subthreshold current larger at $V_{GS} = 0$. The worst shortcoming, however, is that the power-supply voltage must be reduced by the same factor as the channel length. This is not compatible with circuit requirements such as noise margin, current driving capability, and power-supply voltage compatibility from one generation to the next. Thus, this method has not been adopted in its original form.

Indeed, scaling of the power-supply voltage has not been implemented in digital ICs down to device dimensions of 0.8 μm. Instead, it has been maintained at 5.0 V, primarily to permit the next generation of semiconductor products to "plug in" to existing systems. Only for devices with smaller dimensions has a reduced power-supply voltage been implemented (3.3V). This has led to the consideration of *constant-voltage scaling* as a miniaturization alternative. In this approach, device dimensions are scaled by the same factor λ as in constant-$\mathcal{E}$-field scaling, but the applied voltages remain constant. While this solves the circuit incompatibility problems, it leads to higher electric fields inside the device. These high fields cause mobility degradation, hot-carrier effects, oxide tunneling, and other reliability problems.

A compromise approach named *constant-electrostatic scaling* (or *quasi-constant voltage scaling*) was thus proposed by Bacarani et al.[62] Dimensions are still scaled by λ but the potentials are allowed to be scaled by a separate factor κ (e.g., $\kappa = \sqrt{\lambda}$). The operating voltages can then correspondingly decrease to $V' = V/\kappa$. This has allowed suitable fabrication parameters for MOSFETs to be derived for devices down to channel lengths of 0.25 μm (some degree of power supply voltage reduction is used).

However, even with the latter method, the depletion regions of the source and drain junctions scale by a factor less than λ, and hence devices specified according to this approach are subject to such two-dimensional problems as punchthrough and DIBL. In addition, subthreshold turn-off does not scale because it is dominated by the exponential dependence on potential of the carrier density (and the subthreshold current is dominated by a diffusion rather than a drift component). Thus, the subthreshold current does not scale with field. Consequently, the scaled down device fails to turn off at $V_{GS} = 0$ (because the voltage swing needed to drive the device "off" - which is in the range of 0.5-0.8V - cannot be scaled because the subthreshold portion of this swing $\{\Delta I_{Dst}/\Delta V_{GS}\}$ is invariant). As a result, the target V_T value of a new, reduced-design-rule process will not decrease with the other physical parameters relative to an existing, larger-design-rule process. Hence, for ICs in which subthreshold leakage currents must be minimized, even this scaling method is not adequate.

5.6.1.2 Subthreshold Scaling. Another approach to scaling which overcomes the above difficulty was proposed by Brews et al.[63] They introduced an empirical formula which determines the minimum gate length a MOSFET can have so that its subthreshold behavior remains insensitive to drain bias. That is, by using this formula, a constraint upon each single combination of device parameters is identified, such that if this constraint is satisfied with this combination of parameter values, the resulting device exhibits long-channel subthreshold behavior.

We recall that when long-channel devices are operated in subthreshold, I_{Dst} is independent of drain-to-source voltage once V_{DS} exceeds a few kT/q (see Eqs. 4-87 and 4-89). Thus, the criterion selected to represent acceptable long-channel behavior was that no more than a 10% change in drain current could occur for a 0.5V change in V_{DS} (e.g., I_{Dst} should not vary more than 10% as V_{DS} is increased from 0.5 to 1.0V).

Brews et al. found empirically that such subthreshold long-channel behavior will be preserved provided the channel length L exceeds a minimum value L_{min}

$$L \geq L_{min} = A\,[r_j\, t_{ox}\, (W_S + W_D)^2\,]^{1/3} \qquad (5 - 84)$$

where, r_j is the junction depth, t_{ox} is the oxide thickness, W_S and W_D are the depletion widths of the source and drain (in a one-dimensional, abrupt junction for uniform doping). With r_j, W_S, and W_D in microns, and t_{ox} in angstroms, the fitting parameter is $A = 0.41$ Å$^{-1/3}$. The substrate doping dependence is contained in the depletion-width equations, inasmuch as W_S and W_D (e.g., in an NMOSFET) are given by

$$W_S = \sqrt{2\kappa_{si}\varepsilon_o V_{bi}/qN_A} \qquad (5 - 85)$$

$$W_D = \sqrt{2\kappa_{si}\varepsilon_o(V_{bi} + V_{DS})]/qN_A} \qquad (5 - 86)$$

According to Eq. 5-84, miniaturization requires shallower junctions (r_j), thinner oxides (t_{ox}), lower voltages, or heavier doping in the channel region (i.e, this will reduce $W_S + W_D$).

A set of experimental and computer-generated points were fitted to Eq. 5-84. The validity of the equation was established for gate lengths as small as 0.3 μm and for device parameters in the ranges of 100Å$\leq t_{ox} \leq$1000Å, 10^{14} cm$^{-3} \leq N_A \leq 10^{17}$ cm^{-3}, 0.18 μm$\leq r_j \leq$1.5 μm, and for drain voltages up to 5V. Figure 5-44 shows a plot of Eq. 5-84 for L_{min} (solid line) compared with experimental measurements and computer simulations. For the ranges of device parameters examined, Fig. 5-44 indicates that Eq. 5-84 provides a reliable guideline for scaling MOSFET gate lengths down to 0.3 μm.

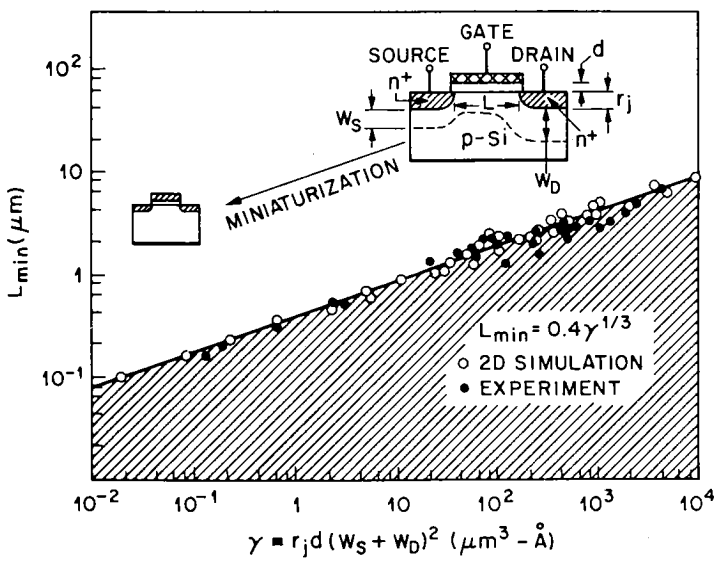


Fig. 5-44 Equation 5-84 for L_{min} compared with experimental measurements and computer simulations. The criterion for L_{min} is that for $L = L_{min}$ a 10% variation in drain current occurs as V_{DD} varies from 0.5 to 1.0V in the subthreshold regime of gate biases. The units of the abscissa reflect the choice of units: d in Å, the other variables in μm.[63] (© 1980 IEEE).

This approach is also more flexible than the earlier ones because not all device parameters need to be scaled by the same factor λ. Instead, various device parameters can be adjusted independently as long as the value of L_{min} remains the same. This is shown in Fig. 5-44, which also plots L_{min} versus γ [where $\gamma = r_j\, t_{ox}(W_S + W_D)^2$]. All devices with channel lengths lying above the line exhibit long-channel subthreshold behavior, while those below the line (crosshatched area) exhibit short-channel subthreshold behavior.

As an example, this figure shows that in order to design a MOSFET with L_{min} = 0.5 μm, γ should be 2. Once γ is determined, r_j, t_{ox}, W_S, and W_D can be chosen so that the resulting γ does not exceed 2. Unlike electrostatic scaling, it is not necessary to start with an optimal large device and then scale it down by a single constant factor. Rather, any device that fits the criterion will be suitable. In addition, it allows voltages to be kept fixed or only reduced modestly, provided the other factors are changed to compensate.

To some degree Eq. 5-84 appears plausible. For example, it seems reasonable that L_{min} should increase if the depletion widths increase, because such an increase enlarges the two-dimensional regions controlled by the source and drain at the expense of the region controlled by the gate. Similarly, increased junction depth increases the influence of source and drain regions, so we expect that shorter-channel devices will require smaller values of r_j (i.e., shallower junctions). Finally, as t_{ox} is increased, the screening effect of the gate is reduced, and the field lines emanating from the source and drain penetrate further into the channel region, and fewer of them terminate on the gate, causing L_{min} to increase. However, while these simple ideas appear to support the validity of Eq. 5-84, the authors of this approach in fact admit that they do not know why it works.[65] In addition, Eq. 5-84 is known to be valid only over the fitted range, and has not been tested below 0.3 μm.

Nevertheless, Eq. 5-84 is useful in keeping DIBL under control. On the other hand, this approach still does not incorporate all of the necessary tradeoffs to handle such problems as hot-carrier degradation, or circuit-related constraints such as noise margin or off-current requirements. Consequently, the reduced devices designed with this criterion still need adjustments, especially as L gets deeper below 1 μm.

Note as well, that in modern MOSFETs the substrate doping is not uniform. That is, ion implantation is used to increase the doping in the region between the source drain so that the doping there is heavier than in the substrate region beneath the source and drain. This allows a shorter source-to-drain spacing. With many MOSFET designs this surface doping is nearly uniform out to the junction depth. Thus, in Eq. 5-84 it is this doping concentration (i.e., N_A in Eqs. 5-85 and 86) that must be used to calculate W_D and W_S.

In addition, another aspect of modern NMOSFETs is the formation of shallow junction extensions by implantation using the gate as a mask (the LDD structure). It is the depth of this shallow junction extension that is used in Eq. 5-84. (The entire length of the junction is not made shallow, to avoid junction series resistance and junction leakage problems.)

Finally, Eq. 5-84 does not take into account the lateral ballooning of the drain depletion region in the lightly doped substrate beneath the source and drain regions. That is, this effect may allow bulk (subsurface) punchthrough to occur between drain and source. To prevent such subsurface punchthrough the substrate doping cannot be made as low as one would like in an attempt to reduce junction capacitance and threshold variation under source-to-substrate bias. A rough guideline for selecting the minimum value of substrate doping to avoid subsurface punchthrough has been suggested by Klaassen[64] to be

$$N_{SUB} \geq N_{ch}/10 \qquad (5 - 87)$$

where N_{SUB} is the substrate doping and N_{ch} is the doping in the region between the source and drain. Equation 5-87 sets a lower bound on the junction-to-substrate capacitance.

5.6.1.3 Off-Current Scaling Approach. Another scaling approach employing a more flexible criterion for acceptable "off"-state behavior (but which is also somewhat more complex than the previous subthreshold approach), has been recently proposed.[66] We will provide a brief outline of this method in this section. The crux of the procedure involves finding a channel doping profile which yields an acceptable combination of off-current I_{off} (I_D when $V_{GS} = 0V$) and V_T in a device with as small an L as possible. The following three parameters are varied to arrive at the appropriate channel doping profile: the exposed dose D_I and centroid x_c of the V_T implant, and the substrate doping N_{SUB}. The profile is initially determined assuming that the device exhibits long-channel behavior (i.e., no significant punchthrough should be observed in the subthreshold operating region). However, the acceptable value of "off"-current in the long-channel device ($I_{off,lc}$) is deliberately picked to be smaller than I_{off} in the short-channel device being designed - in anticipation of the increase in I_{off} due to surface-DIBL in the shrunken device. The estimate of the increase in $I_{off,lc}$ due to surface-DIBL depends on the DIBL model chosen by the designer. To ensure that the subthreshold punchthrough current remains insignificant at maximum V_{DS}, a minimum N_{SUB} is identified. Finally, in the process of selecting appropriate values of D_I and x_c, the criterion that the channel depletion-region width d should be a minimum is applied (to keep ΔV_{Tsc} as small as possible when L is shrunk, as suggested by either Eq. 5-2 or barrier-lowering analysis).

In the report that describes this scaling strategy, the example to which it is applied is a symmetrical CMOS process (i.e., in which NMOSFETs are built with n^+-poly gates and PMOSFETs with p^+ poly gates), a technological approach that is gaining more popularity as L is decreased to 0.5 μm or smaller. Let us now also consider this scaling strategy as applied to a symmetrical CMOS process in more detail.

First, the approach assumes that the following parameters are set by system and processing constraints: V_{DD}; the type of gate material (e.g., heavily doped poly); and minimum values of t_{ox} and r_j. The minimum manufacturable L is also limited by the lithographic process. However, I_{off}, V_T, and the actual value of L are left as parameters that can be traded off later, based on the specific circuit or system application

requirements. The tradeoffs between V_T and long-channel I_{off} ($I_{off,lc}$) values are carried out with the aid of families of curves that relate V_T and $I_{off,lc}$ for a fixed t_{ox} (with N_{SUB} as a parameter). Figure 5-45 is an example of one such family of curves, in which t_{ox} = 100Å. Notice that the x-axis is expressed in $N_{off,lc}/N_{mg}$ instead of $I_{off,lc}$. However, since $I_{off,lc}$ is directly proportional to $N_{off,lc}$ (the latter being defined as the free-carrier density in the channel per unit area - e.g., per cm^{-2} - when V_{GS} = 0V) according to

$$I_{off,lc} = \mu\,(Z/L)(kT/q)qN_{off,lc} \qquad (5 - 88)$$

$I_{off,lc}$ is implicitly expressed in these curves. In Eq. 5-88, Z and L are the device width and length, respectively. In addition, N_{mg} is the free-carrier concentration per unit area in the channel at midgap,* and this value can be found from

$$N_{mg}\ (cm^{-2}) = n_i\,(\kappa_{si}/\kappa_{ox})(2kT/E_g)\,t_{ox} \qquad (5 - 89)$$

N_{mg} is therefore the free-carrier density in the channel at the point when the surface is undergoing its field-induced reversal of polarity (e.g., from p-type to n-type). Thus, N_{mg} represents the smallest free-carrier concentration in the channel in weak inversion. As is also indicated by Eq. 5-89, N_{mg} is independent of N_{SUB}. In an Si-SiO_2 MOS device at 300K, Eq. 5-89 can be written to explicitly express the relation between N_{mg} and t_{ox} as

$$N_{mg}\ (@\ 300K,\ in\ cm^{-2}) = 10^{10}(cm^{-3})(11.8/3.9)\ 22.4\ t_{ox}(cm) = 1.33x10^9\ t_{ox}(cm).$$

Thus, for a MOSFET with t_{ox} = 15 nm = $1.5x10^{-6}$ cm

$$N_{mg} \sim 2x10^3\ cm^{-2}.$$

Thus, if $I_{off,lc}$ is given, and it is desired to relate this value to a point on a curve in which one axis is expressed in terms of $N_{off,lc}/N_{mg}$ (such as Fig. 5-45), it can be done as follows:

Example 5-1: If a maximum off-current $I_{off,lc}$ at 300K of 10 pA is required in an NMOSFET with L = 3 μm, Z = 6 μm, t_{ox} = 15 nm, and μ_n = 400 cm^2/Vsec, find the value of $N_{off,lc}/N_{mg}$ that corresponds to this $I_{off,lc}$ value.

Solution: Since $N_{off,lc} = I_{off,lc}/[\mu_n\,(Z/L)\,(kT/q)q]$

$$= 1x10^{-11}\ A\ /\ [400\ cm^2/Vsec\ (6/3)\ (0.025V)\ 1.6x10^{-19}C]$$

$$= 3.1x10^7\ carriers/cm^2$$

* *Midgap* is the state of a MOSFET in which band bending is causing E_F at the Si surface to co-incide with the energy level at the middle of the bandgap (which is also where E_i resides). Also at midgap, $\varphi_{surf}/\varphi_t = \ln(N_{SUB}/n_i)$ as shown in Fig. 5-46. In the special case of midgap occurring at V_{GS} =0V, N_{mg} is the free-carrier concentration (cm^{-2}) in the channel for that condition.

and since N_{mg} (300K, t_{ox} = 15 nm) = $2x10^3$ cm^{-2},

$$N_{off,lc}/N_{mg} = 1.5x10^4$$

To arrive at the V_T values exhibited in a family of curves such as those in Fig. 5-45, we follow the general procedure given in section 4.6.3. That is, the value of $V_{Tud,lc}$ (assuming a uniformly doped channel region with a doping of N_{SUB}) is calculated using Eq. 3-71, and this is added to $\Delta V_T(II)$ using Eq. 4-73, to get $V_{Tnud,lc}$ (Eq. 4-74 or Eq. 4-75) However, to accurately determine $\Delta V_T(II)$ the value of both D_I and x_c must be specified. In addition, the impact of N_{SUB} must included.

That is, in the simple long-channel analysis of a MOSFET in section it was assumed that $V_{Tud,lc}$ can be shifted by choosing D_I and x_c independent of N_{SUB}.

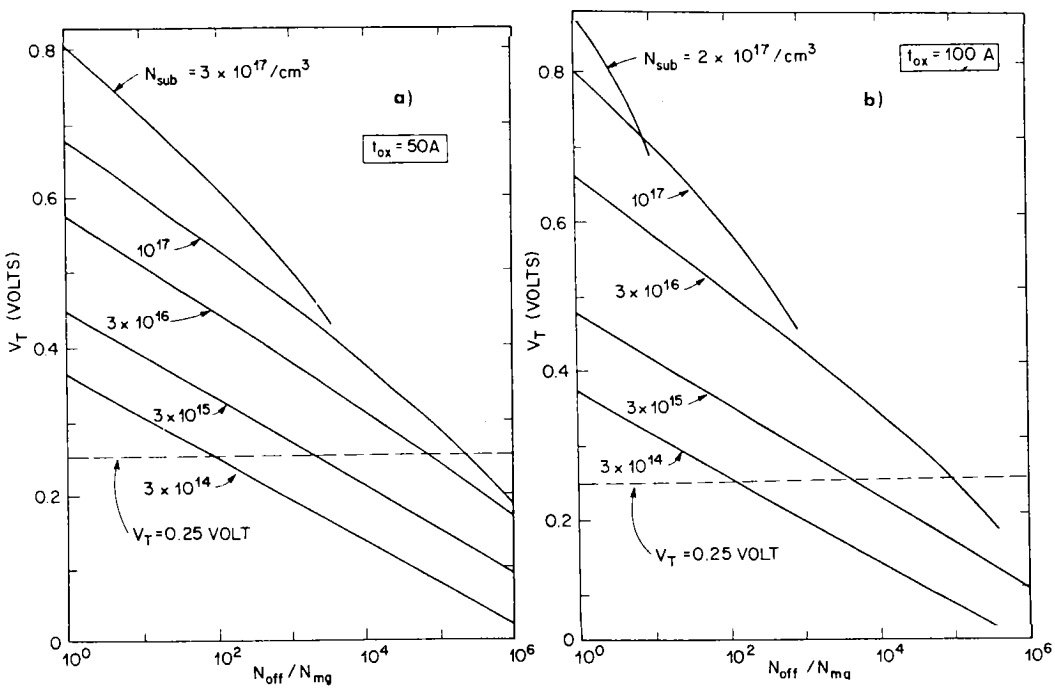


Fig. 5-45 Threshold voltage for implanted structures with the indicated substrate doping as a function of N_{off}/N_{mg}. The dose and centroid of the implants are chosen for minimum depletion width. (a) t_{ox} = 50Å; (b) t_{ox} = 100Å. From J.R. Brews, K.K. Ng, and R.K. Watts, Chap. 1, "The Submicrometer Silicon MOSFET," in *Submicron Integrated Circuits*, R.K. Watts, Ed., Wiley-Interscience, New York, Copyright 1989 p. 34. Reprinted with permission.

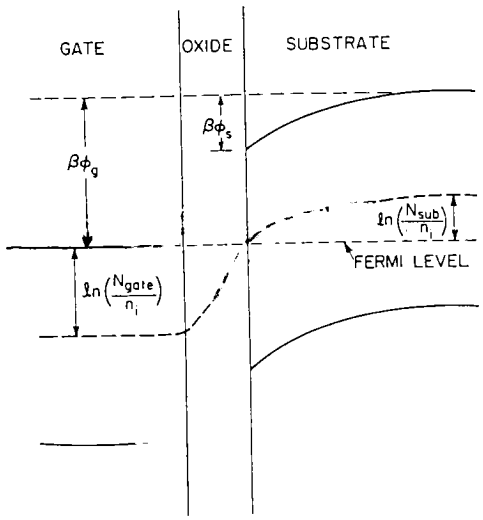


Fig. 5-46 An MOS structure at zero gate bias with midgap level and the Fermi level at the surface being equal at zero bias (*midgap* occurs).

However, for the following reasons N_{SUB} cannot be ignored, especially when designing short-channel devices: First, a minimum N_{SUB} must be identified to prevent subsurface punchthrough, e.g., $N_{SUB} \geq N_{ch}/10$. (Note that in this analysis, the effect of a punchthrough implant is not considered, and hence only the substrate doping plays a role in preventing subsurface punchthrough.) Second, the choice of N_{SUB} impacts d_I (the channel depletion region width in the presence of an V_T implant) in that larger values of N_{SUB} will cause d_I to decrease (see Eq. 4-80). To make d_I smaller (i.e., to keep ΔV_{TSC} as small as possible, as described earlier), larger N_{SUB} values are needed.* Third, N_{SUB} impacts $N_{off,lc}$, because S_t depends on d (see Eq. 4-91b). In general, increasing the value of N_{SUB} will cause the value of S_t to increase. If V_T remains the same, the consequence of an increase in S_t is a larger $N_{off,lc}$ (as is also evident in Fig. 5-45). Obviously, the first two issues call for large N_{SUB} values, while the third calls for smaller values. Hence, tradeoffs may be necessary.

Values for the exposed dose D_I and the centroid x_c that will produce a minimum value of d are plotted versus $N_{off,lc}/N_{mg}$ with N_{SUB} as a parameter, and t_{ox} = 50Å in Fig. 5-47. The value of D_I is obtained from the following equation:

* The equations used to generate the V_T versus $N_{off,lc}/N_{mg}$ curves are derived by invoking the constraint that the channel depletion width in weak inversion (d) must be a minimum. This constraint is derived from Eq. 5-2, and the fact that we want ΔV_{TSC} to be as small as possible. Since it is assumed that the device will employ the smallest t_{ox} and r_j values allowed by process technology, the last parameter in Eq. 5-2 that can reduce ΔV_{TSC} when the device is shrunk to its minimum *L*, is d_{max}.

$$qD_I = (Q/Q_{mg})\,(\kappa_{ox}\varepsilon_o/t_{ox})\,(kT/q)\,(\beta E_g/2q) \qquad (5\text{ - }90)$$

The value of Q/Q_{mg} is found in terms of $N_{off,lc}/N_{mg}$ with the aid of Fig. 5-48.* (It turns out that for most cases of interest the substrate doping independence of D_I assumed in section does in fact continue to hold to larger values of N_{SUB} when the devices have thin oxides. Thus, Eq. 5-90, which is independent of N_{SUB}, continues to be valid for such cases). Once D_I is known, its value is used to find x_c from

$$x_c/d = -D_I/(N_{SUB}d) + \{(D/N_{SUB}d)^2 + 1\}^{1/2} \qquad (5\text{ - }91)$$

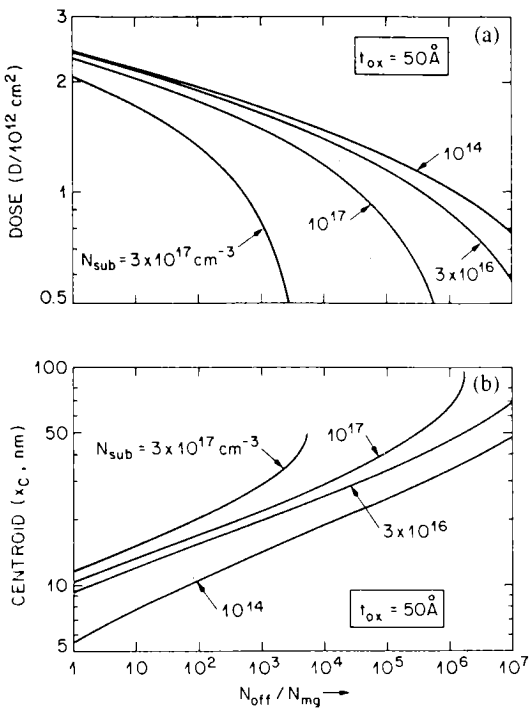


Fig. 5-47 (a) Dose in units of $10^{12}/cm^2$ as a function of N_{off}/N_{mg} with substrate doping as parameter. (b) Centroid or depletion layer width $x_c = w$ in nanometers versus N_{off}/N_{mg}. In both figures t_{ox} = 50Å. From J.R. Brews, K.K. Ng, and R.K. Watts, Chap. 1, "The Submicrometer Silicon MOSFET," in *Submicron Integrated Circuits*, R.K. Watts, Ed., Wiley-Interscience, New York. Copyright 1989 p. 28. Reprinted with permission.

* Note that Q here is a dimensionless surface charge defined as by

$$Q = (D_I/N_{SUB}L_D) + (N_{off,lc}/N_{sub}L_D) + (d/L_D) \qquad (5\text{ - }92)$$

where L_D is the Debye length, and $Q_{mg} = (\kappa_{ox}L_DE_g/2kT\kappa_s t_{ox})$. But for the purposes of our abbreviated description of this scaling strategy (described more completely in ref. 65), it suffices to note that Q/Q_{mg} can be determined in terms of $N_{off,lc}/N_{mg}$ by using Fig. 5-48.

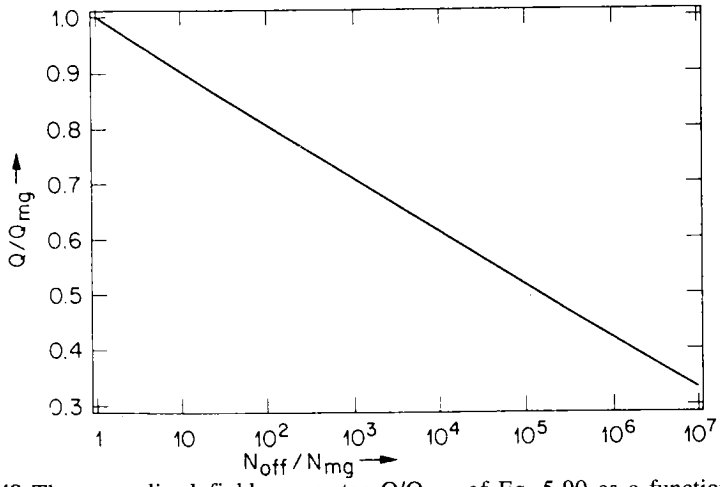


Fig. 5-48 The normalized field parameter Q/Q_{mg} of Eq. 5-90 as a function of zero-bias carrier density to midgap carrier density ratio N_{off}/N_{mg}. From J.R. Brews, K.K. Ng, and R.K. Watts, Chap. 1, "The Submicrometer Silicon MOSFET," in *Submicron Integrated Circuits*, R.K. Watts, Ed., Wiley-Interscience, New York, Copyright 1989 p. 28. Reprinted with permission.

Once V_T is determined for an acceptable value of $N_{off,lc}$, a check is made to ensure that the chosen N_{SUB} value exceeds the minimum substrate concentration identified as being able to provide adequate subsurface punchthrough protection. If N_{SUB} is found to be large enough to prevent punchthrough, the design process can proceed to the next step - barrier-lowering analysis. (However, if the chosen value of N_{SUB} cannot prevent punchthrough, a larger N_{SUB} value may need to be accepted, which in turn may possibly require an adjustment of the maximum acceptable I_{off}.)

The barrier-lowering analysis examines the effect of shrinking L to the minimum value that can be fabricated with the best lithographic process at hand. At that length the value of ΔV_{Tsc} is calculated, using the DIBL model of choice. The value of the long-channel off current $I_{off,lc}$ in the device designed using the strategy outlined above now becomes the basis for determining the short-channel device I_{off}, according to

$$I_{off} = I_{off,lc} \exp (q\Delta V_{Tsc}/kT) \qquad (5 - 93)$$

If I_{off} is less than or equal to the target value, the device with that value of L will succeed in meeting the circuit requirements and is deemed nominally manufacturable. If I_{off} is too high, either the channel length will need to be increased, or a larger value of I_{off} will need to be accommodated by changing the circuit design.

The term *nominally manufacturable* is used, because as the channel length is reduced, the sensitivity of V_T to channel length variations increases. As already noted, the impact of channel length variation is to require a more conservative setting of nominal-I_{off} so that devices with smaller than nominal V_T values will still have acceptable off-currents. As the channel length is reduced, at some point this lowered I_{off} estimate will encounter

the same punchthrough limit already discussed. If channel length is under very tight control, then V_T variations from length variation remains small, and only variation due to changes in V_{DS} will matter. However, it is more common that channel length variations will dominate the choice of L_{min} as illustrated in the next section.

5.6.2 Manufacturing Considerations in Submicron MOSFET Device Design

As noted above, the "off"-current value in submicron MOSFETs must be maintained over the range of manufacturing tolerances. For example, assume that a typical 0.8 μm NMOS transistor (i.e., the drawn channel length L is 0.8 μm) has an L_{eff} of 0.7 μm and a 3-sigma manufacturing tolerance on patterned gate length of ± 0.12 μm. Hence, the shortest L_{eff} in a device that still meets the manufacturing tolerance (i.e., patterned gate length is 0.68 μm) will be 0.58 μm. At this gate length, the CMOS devices will still have to possess sufficient V_T control, satisfactory hot-carrier reliability, and acceptable punchthrough characteristics. To enable a MOSFET to meet these requirements, N_{SUB} will need to be raised to a value higher than that needed for the nominal device length. As a result, the performance of the nominal gate length device ($L = 0.8$ μm) will be degraded through increased source/drain junction capacitance and subthreshold turnoff.

The importance of manufacturing control on patterned gate length is illustrated in Fig. 5-49.[67] This figure plots the NMOS and PMOS device leakage in the "off" state

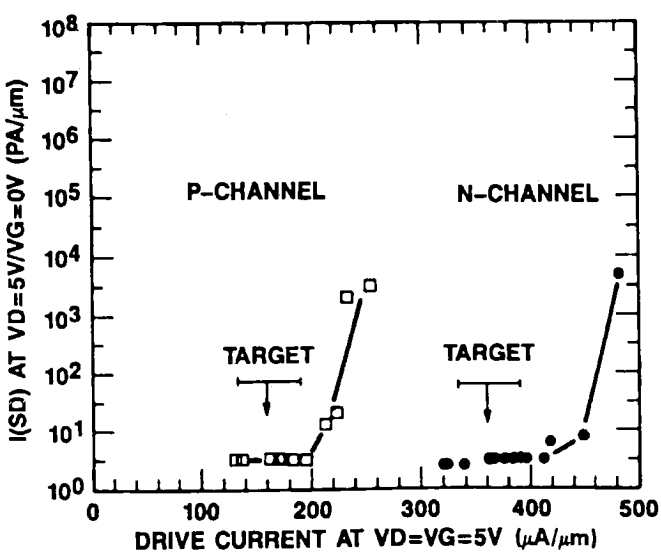


Fig. 5-49 NMOS and PMOS drive current, defined as the drain current for 5V applied to the gate and drain, versus leakage in the off state with 0V on the gate and 5V on the drain. The parameter is gate length, with a target value of 0.8 μm. From R.A. Haken et al., "BiCMOS Process Technology," Chap. 3, p. 85, in *BiCMOS Technology and Applications,* A. Alvarez Ed. Klewer, Norwell MA, Copyright 1989. Reprinted with permission.

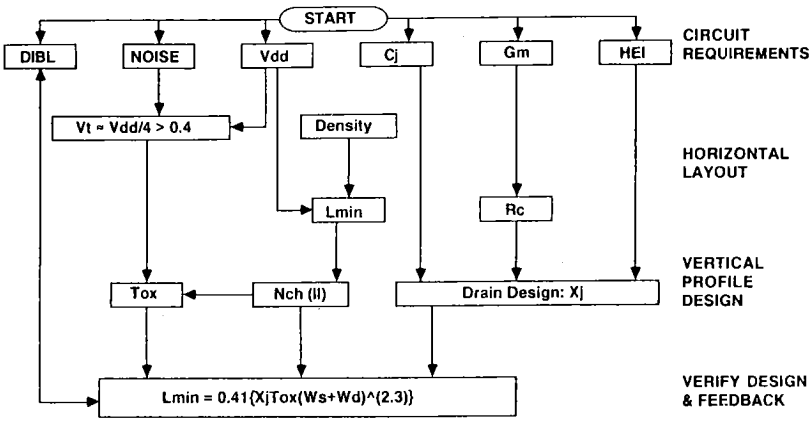


Fig. 5-50 First order MOS device design methodology. Horizontal design generally precedes the vertical profile design. The horizontal design is set to optimize intrinsic device delay (L_{eff}, R_{SD}). Vertical profile design is set to minimize short-channel effects.[68] (© 1988 IEEE).

(in this figure I_{SD} with 0V on the gate and 5V on the drain), versus drive current, defined as I_D when $V_{DS} = V_{GS} = 5V$. Although I_D is the x-axis variable, the actual parameter being varied is L, since I_D is inversely proportional to L. In this example, for the n-channel device, when $L = 0.8$ μm ($L_{eff} = 0.7$ μm), I_D is 400 μA/μm. The device design has been optimized for a patterned gate length of 0.8 μm with a manufacturing 3-σ control on L of ± 0.12 μm. According to Fig. 5-49, as the 3-σ manufacturing control limit is reached, the device leakage current increases to ~10^4 pA/μm (i.e., from Fig. 5-49 the leakage starts to rapidly rise when $I_D = 480$ μA/μm, which means it is occurring when $L = 0.68$ μm, or $L_{eff} = 0.58$ μm). Such high levels of device leakage might cause circuits to become nonfunctional, or could cause them to fail a particular circuit specification, such as maximum standby power.

5.7 DESIGN OF SUBMICRON NMOSFETS

We outline two approaches for the design of submicron NMOSFETs that have appeared in the literature. The first is fundamentally based on Eq. 5-87, while the second relies on the off-current scaling approach.

5.7.1 Design Methodology Based on the Subthreshold Scaling Equation

Figure 5-50 shows a flow-chart of the steps taken in the subthreshold-scaling approach to designing submicron MOSFETs in CMOS.[68] It assumes that the following circuit constraints on the device are defined (as dictated by the application being served): the power-supply voltage V_{DD}, the required drive current I_D in μA/μm (or transconductance,

g_m), the required noise margin, the minimum intrinsic gate delay, and the maximum allowed device degradation (long-term) due to hot-carrier effects (HEI in the diagram).

The given V_{DD} and noise-margin constraint determine the choice of V_T. V_{DD} and the lithographic-processing capability limit the minimum channel length L_{min}. By thinning t_{ox}, I_D can be increased, but minimum t_{ox} is set by V_{DD} and processing constraints.

The maximum allowed parasitic resistance R_S depends on g_m (which is really the inverse of R_{ch}). For example, R_S may need to be kept smaller than 10% of R_{ch}. This may then dictate the minimum horizontal dimensions of the drain and source regions, and also the depth r_j of the source and drain junctions. The required R_S value may also determine the doping profile of the tip region of the LDD structure and may dictate whether or not salicided source/drains are necessary. The minimum gate delay τ_d will be impacted by the intrinsic device characteristics (i.e., I_D through L_{min}, and C_L through C_{gate} and C_j). By minimizing gate area and source/drain junction areas, smaller C_{gate} and C_j can be realized. Lower substrate doping also decreases C_j.

The selection of the channel doping concentration (N_{ch}) will be impacted by the substrate concentration and the V_T-adjust implant. V_T and N_A will determine the V_T adjust implant dose. The drain design (e.g., an LDD) is impacted by the desire to keep hot-carrier degradation to an acceptable value (see chap. 9). The depletion region widths of the source and drain junctions - W_S and W_D - are determined by N_{ch} and the drain structure design.

Once the values of t_{ox}, r_j, and W_S and W_D are determined, they are inserted into Eq. 5-84. If the right hand side of the equation is a number smaller than L_{min} with this group of parameters, then according to past experience it is likely that a MOSFET with this L_{min} can be successfully fabricated (i.e., the device should satisfy the subthreshold leakage current criterion given in section 5.6.1.2). Hence, if some acceptable combination of manufacturable parameters meeting this criterion can be found, fabrication success is likely. If not, a larger value of L_{min} may have to be accepted.

5.7.2 Design Methodology Based on Required Off-Current Value

The flow chart of the I_{off} design strategy of submicron MOSFETs is shown in Fig. 5-51.[69] In summary, a MOSFET is designed by selecting what appears to be an appropriate set of V_T and I_{off} values, applying the principles of long-channel MOSFET physics to derive the channel doping profile. Then the device length is shortened and punchthrough analysis and barrier-lowering analysis is performed on the shortened device to ensure that the I_{off} circuit specification is still met. If not, the design procedure must be repeated, trading off V_T and I_{off} against channel length until the best set of parameters for the application at hand is achieved. Details of the various steps in the flow chart were described in section 5.7.1.

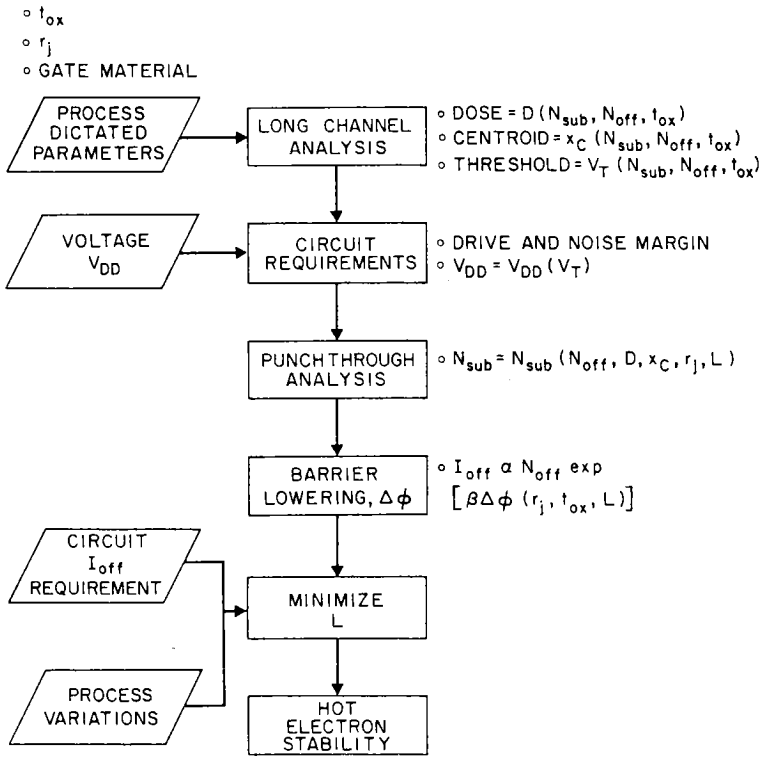


Fig. 5-51 A flow chart of the I_{OFF} strategy. From J.R. Brews, K.K. Ng, and R.K. Watts, Chap. 1, "The Submicrometer Silicon MOSFET," in *Submicron Integrated Circuits*, R.K. Watts, Ed.,Wiley-Interscience, New York, Copyright 1989, p. 24. Reprinted with permission.

5.8 SHORT-CHANNEL PMOS DEVICES IN CMOS

Early MOS ICs were built using PMOS technology, primarily because enhancement-mode MOSFETs with suitable V_T values could only be fabricated as *p*-channel devices using aluminum or n^+-doped polysilicon gates and uniform lightly doped *n*-substrates. That is, at that time it was not possible to produce enhancement-mode *n*-channel MOSFETs using n^+-poly or Al gates on a uniform, lightly doped *p*-substrate because such structures exhibit a negative threshold voltage (see Fig. 3-20). However, with the advent of ion implantation it became possible to build NMOSFETs with positive V_Ts by adjusting channel doping profiles. Thus, enhancement-mode and depletion-mode NMOSFETs could then be fabricated with little extra difficulty. Since *n*-channel transistors have greater drive current (and hence speed), NMOS naturally replaced PMOS as the dominant digital IC technology. During this era of NMOS dominance, interest in *p*-channel MOSFETs languished. However, when CMOS arose to become the main

p-channel MOSFETs languished. However, when CMOS arose to become the main technology for VLSI in the late 1980s, the need for enhancement-mode p-channel MOSFETs returned.

The fabrication of p-channel devices with short channels in CMOS, however, presents some unique problems which arise from having to build both NMOS and PMOS devices on the same chip. The problems revolve around the choice of a doping type for the polysilicon gate electrode and the impact that this choice has on the threshold voltage and other characteristics of PMOS devices.

5.8.1 PMOS Devices with n^+-Polysilicon Gates

To achieve high drive current it is necessary to make the threshold voltage of a MOSFET as close to 0V as possible (with the minimum value limited by noise margin and subthreshold-leakage considerations). When MOS IC technologies were initially being developed V_T values with a magnitude of ~1V were acceptable. In early PMOS (and early p-well CMOS) it was relatively straightforward to build p-channel devices with V_T = -1V since the doping in the n-substrates was in the range of $1\text{-}3\times10^{15}$ cm^{-3}. As seen in Fig. 3-20, V_{Tp} of PMOS devices made on such substrates is -0.8V to -1.0V, which is well within the acceptable V_T range.

However, in n-well CMOS (which was developed after p-well CMOS), as well as in twin-well CMOS, the situation changes. That is, in n-well CMOS the doping in the well is about 10x the doping in the substrate, making it at least 1×10^{16} cm^{-3}. Twin-well CMOS becomes the well-architecture of choice as the gate lengths decrease below about 1 μm. For 1 μm MOS devices (both p- and n-channel) the substrate doping density must be $2\text{-}3\times10^{16}$ cm^{-3}. Thus, as shown in Fig. 3-20, the V_T values of PMOS devices made with n^+-poly gates in either n-well or twin-well CMOS will be at least -1.5V, whose magnitude is too large. To make matters worse, in PMOSFETs with $L \leq 1$ μm, punchthrough effects are more severe than in comparably sized NMOSFETs (primarily due to the inability to make p^+ source/drain junctions as shallow as n^+ junctions, as will be described later). To suppress such punchthrough current it is necessary to increase the n-doping in the substrate (i.e., by an extra deep phosphorus implant into the channel of PMOS devices). The resultant increase in channel doping makes short-channel PMOS devices even more strongly enhancement mode (i.e., V_T is increased even more).

For optimal logic-gate performance the threshold voltages of the n- and p-channel devices in CMOS circuits should also have comparable magnitudes. In addition, to allow for maximum current-driving capability, they should also be as small as possible. Thus, for 5V CMOS technologies, desirable threshold voltages are 0.6 to 0.8V for V_{Tn} and -0.6 to -0.8V for V_{Tp}. Furthermore, as mentioned above, the most common choice for the gate material has historically been heavily doped n-type polysilicon. The work function of n^+ polysilicon is ideal for n-channel devices since these yield threshold voltages of less than 0.7 V for reasonable values of channel doping and oxide thicknesses.*

Figure 3-20 plots V_T for long-channel, uniformly doped MOSFETs with n^+ poly gates (left scale)[70] as a function of p-substrate doping (assuming the substrate is uniformly doped and that $Q_{tot} = 0$). The value of V_{Tn} is less than 1.0V for gate-oxide thicknesses of 25 nm or less and substrate dopings of less than 10^{17} cm^{-3}. Thus, V_{Tn} can easily be adjusted to the desired values by implanting boron into the channel.

However, when n^+ poly is the gate electrode of a PMOS device, adjustment of V_{Tp} to -0.7 V is not as simple, especially as the PMOS channel lengths shrink below about 2 μm. First, as we noted above, the magnitude of V_T of PMOS devices when the substrate doping exceeds $2x10^{16}$ cm^{-3} is significantly greater than -0.8V (e.g., -1.4V or larger). Even higher values of substrate doping will be needed as the channel lengths are decreased below 1 μm, implying that V_T would be even larger in submicron PMOS devices with uniformly doped substrates. In addition, such short-channel PMOSFETs are more prone to punchthrough than NMOSFETs of the same channel length. This means that the substrate doping concentration will need to be increased even further in submicron PMOS devices to suppress these effects (e.g., by performing an additional deep n-type [phosphorus] implant into the channel region).

The increased susceptibility to punchthrough of scaled-down PMOS devices with uniformly doped channels is caused in part by the dopant used to form the source and drain regions of the PMOS device, i.e., boron. Boron diffuses much more quickly than the dopant used to form the source and drain junctions in NMOSFETs (arsenic). As a result, the junction depths of the p^+ regions in 1 μm PMOS are ~0.35 μm, whereas the n^+ regions in 1 μm NMOS are shallower (typically only ~0.2 μm deep). The deeper junctions accentuate the punchthrough susceptibility. The extra n-implant used to suppress punchthrough mentioned above must also place the additional doping at a depth below the source and drain depths to be effective (e.g., with a peak at ~0.15 μm and increased doping above the background level to a depth of ~0.5 μm). This extra implant, however, not only increases the V_T value to about -2V, but also increases the source/drain-to-substrate capacitance.

In any case, it is obvious that the magnitude of V_T in short-channel PMOS devices of advanced CMOS technologies needs to be reduced. However, implanting the n-doped substrate with more n-type dopant would only *increase* the magnitude of V_{Tp}, rather than bringing its value closer to the desired -0.7V. To reduce the magnitude of V_{Tp} when an n^+ poly gate is used, a thin sheet of *negative* charges must therefore be placed at the Si-SiO_2 interface. This is accomplished by implanting a shallow layer of boron atoms, which as we know become negatively charged acceptor ions in the Si lattice. The boron dose must be heavy enough to overcompensate the n-surface so that a p-region (depleted of holes at V_{GS} = 0V) is formed. This shifts V_{Tp} toward more positive values. In practice, the boron dose depends on the substrate doping, as shown

* Note that the polysilicon layer may be combined with a layer of silicide for sheet resistance reduction; since the polysilicon remains as the underlayer of the polycide, the work function of the gate electrode will not be changed.

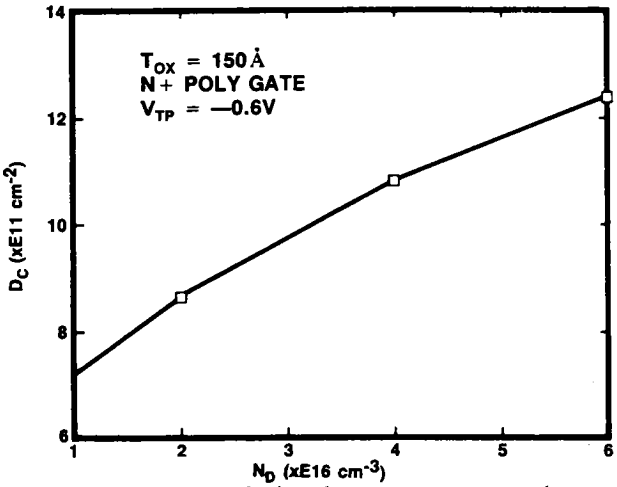


Fig. 5-52 Simulation of the counter-doping dose versus n-type substrate concentration for a fixed p-channel threshold voltage of -0.6V.[115] (© 1984 IEEE).

in Fig. 5-52 for devices with t_{ox} = 15 nm in which a V_{Tp} = -0.6V is desired. Typical doses of such implanted layers are of the order of 5×10^{11} cm^{-2} at 25 keV.

The fact that boron is implanted to adjust both V_{Tn} and V_{Tp} in CMOS circuits with n^+ poly gates suggests that a single implant could be used instead of two separate implants. Figure 5-53 shows that this can be accomplished if the appropriate background dopings are chosen for the substrate and the well.[71] On the other hand, it may be decided to use two separate implants in order to achieve better short-channel behavior through individual optimization of the n- and p-channel devices. In the latter case, if the boron dose needed to adjust V_{Tp} is smaller than the dose needed to adjust V_{Tn}, the former implant can be made into both the p- and n-channel regions without use of a mask. Then the second implant for adjusting V_{Tn} is done with a mask that protects the n-well region. (A third boron implant may also be performed at that time to suppress punchthrough in the n-channel devices.) In other processes, however, it may be necessary to use a separate mask for both the V_{Tp}- and V_{Tn}-adjust implants.

5.8.2 Why PMOS Devices with n^+ Poly Gates and Implanted p-Surface Layers Exhibit Enhanced Susceptibility to Short-Channel Effects

The result of implanting boron atoms to adjust V_T is that such PMOS devices exhibit worse short-channel effects than NMOSFETs with the same L. This susceptibility arises mainly because the implanted boron layer has a finite thickness (i.e., instead of being an infinitely-thin sheet of charge at the surface, as would exist in the ideal case). Figure 5-54 illustrates this fact, in which the channel doping profile of a PMOS device with such a boron implant is seen to extend well below the channel surface. The penetration of the p-layer into the n-substrate gives rise to two phenomena not observed

in surface-channel NMOSFETs: (1) a *pn* junction is formed beneath the gate oxide in the channel, extending to a depth Y_j (as can also be seen in Fig. 5-54); and (2) the potential minimum is moved away from the surface. The latter effect gives rise to important device-behavior effects that differ from those of surface-channel MOSFETs. We explain this behavior by first comparing the location of the potential minimum in such boron-implanted-layer PMOSFETs to its location in surface-channel MOSFETs.

Figure 5-55 shows the one-dimensional potential diagrams of two PMOS devices, one with a p^+-polysilicon gate and a uniformly doped *n*-substrate (3.4×10^{16} cm^{-3}), and the other with a n^+-polysilicon gate and a shallow boron V_T-adjust implant. In both devices $V_{GS} = 0V$, and $V_{DS} = -5V$. The potential profile is taken at the point along the channel at which the longitudinal field vanishes (to match the fact that in a MOS-C there is also no longitudinal electric field). In the p^+-poly gate device the potential minimum is located at the surface, implying that such a device will operate much like the surface-channel NMOSFETs described in chap. 4. In the PMOS structure with the n^+-poly gate and boron V_T-adjust implant, however, we see that the potential minimum is away from the Si surface. The mobile carriers (holes) will thus be found at highest density near the location of the potential minimum. This implies that, unlike in the

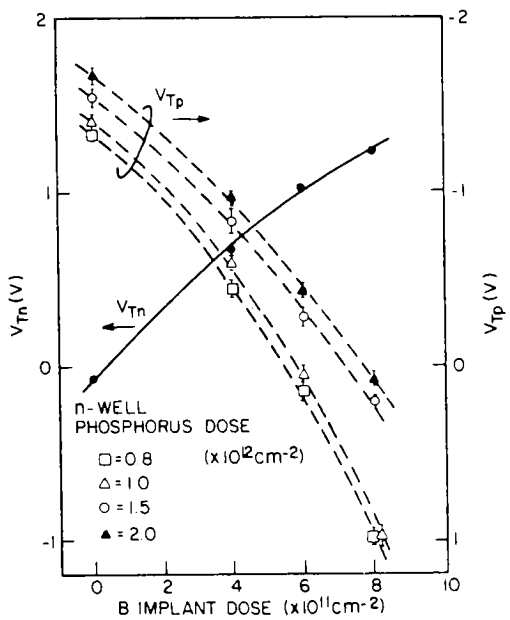


Fig. 5-53 Threshold voltages of *n*-channel (V_{Tn}) and *p*-channel (V_{Tp}) transistors as a function of boron threshold-adjustment dose. The CMOS structure uses an *n*-well implanted into a *p*-substrate (whose doping level is 6×10^{14} atoms/cm^3). V_{Tp} results are shown for various implant doses of the *n*-well.[71] (© 1980 IEEE).

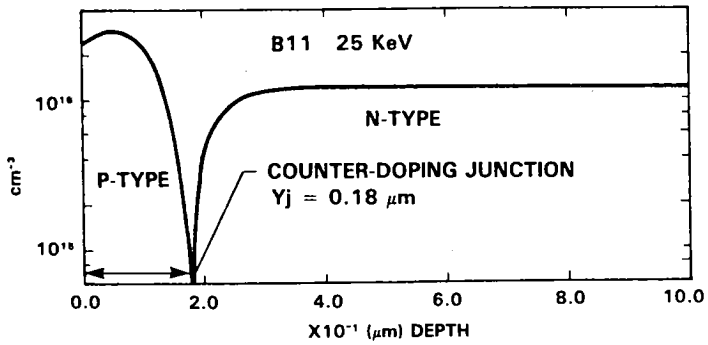


Fig. 5-54 Simulated channel profile for a p-channel MOSFET with an implanted B dose at the channel surface.[115] (© 1984 IEEE).

case of conventional surface-channel devices, more current will flow in a channel *beneath* the surface of the device than at the surface.[72] Such PMOS devices are therefore referred to as *buried-channel* (BC) transistors. As also seen in Fig. 5-56 (which gives the calculated variation of the potential as a function of distance below the surface as the channel *pn* junction depth is varied), the potential minimum moves further into the substrate as the thickness of the implanted p-layer is increased. The mobility of carriers in the buried channel is ~15% higher than carriers at the surface, and thus BC MOSFETs exhibit a somewhat larger drive current than comparable surface channel FETs. While this is an advantage, it is outweighed by several other disadvantages exhibited by BC FETs. Let us examine how the presence of the channel junction and

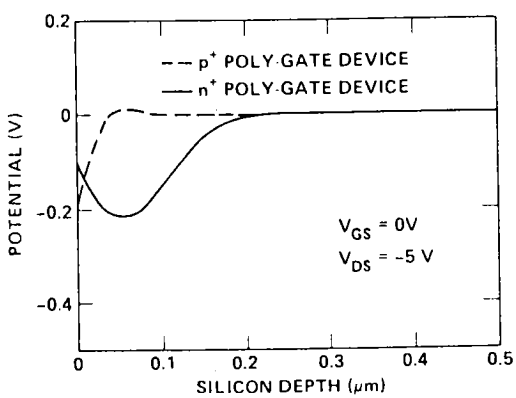


Fig. 5-55 Simulations of the electrical potential perpendicular to the surface in the channel region of two long-channel PMOS devices using GEMINI and SUPREM III simulations of devices with n^+ polysilicon and p^+ polysilicon gates with effective gate lengths of 20 μm.[114] (© 1985 IEEE).

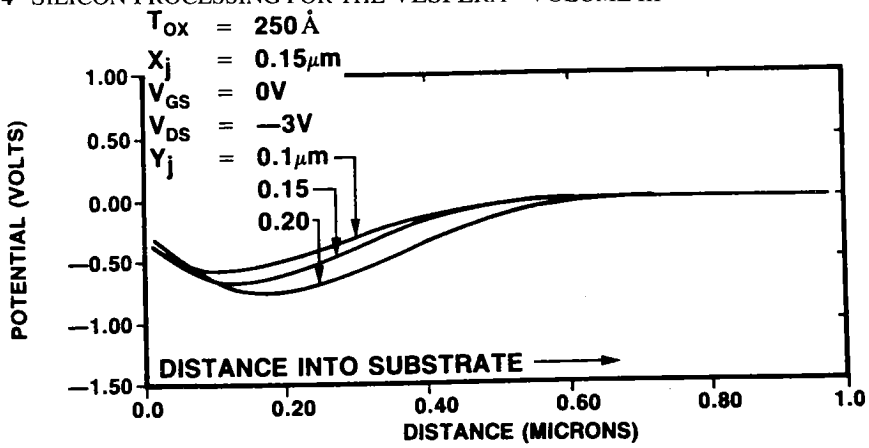


Fig. 5-56 Channel potential profile versus vertical distance for a p-channel MOSFET with Y_j = 0.1, 0.15, and 0.2 μm, and V_{DS} = - 3V and V_{GS} = 0V.[115] (© 1984 IEEE).

the shifting of the potential minimum beneath the Si surface worsen the susceptibility of the PMOS device to short-channel effects.

5.8.2.1 Threshold-Voltage Roll-Off in Buried-Channel PMOS.

As the channel length of MOSFETs is decreased, the threshold voltage decreases from the value it exhibits in long-channel devices. This effect, as described in section 5.2.2, is referred to as *V_T-roll-off.* The V_T-roll-off effect is more severe in BC PMOS devices than in surface channel FETs (Fig. 5-57a).[114] This is due to an effect in BC FETs that does not occur in surface-channel FETs. That is, the implanted p-layer is a lightly doped p-region that extends entirely between the p^+ source and drain regions of the device. Holes diffuse from the regions of high concentration into the lightly doped layer as depicted in Fig. 5-57b.[106] The diffusion length d_p is given by[107]

$$d_p = L_D \tan^{-1} [(N_p^+/N_p^-)^{1/2}] \qquad (5 - 94)$$

where

$$L_D = [(2 \kappa_{si}\varepsilon_0 kT)/(q^2 N_p^-)]^{1/2} \qquad (5 - 95)$$

and N_p^+ and N_p^- are the doping levels in the p^+-source/drain and p-layer regions, respectively. If the diffusion length is smaller than half the channel length, the net effect of this hole diffusion is to increase the effective positive charge in the channel, thus making its threshold voltage smaller in magnitude. As the channel length is decreased, the proportionate contribution of the diffused mobile charge to the total channel charge increases, and as a result, the threshold is shifted to smaller values. Eventually, at short enough channel lengths, the diffused hole distribution from the source meets that from the drain.[108] When that occurs, the potential barrier between the source and drain is reduced, giving rise to poor subthreshold characteristics.

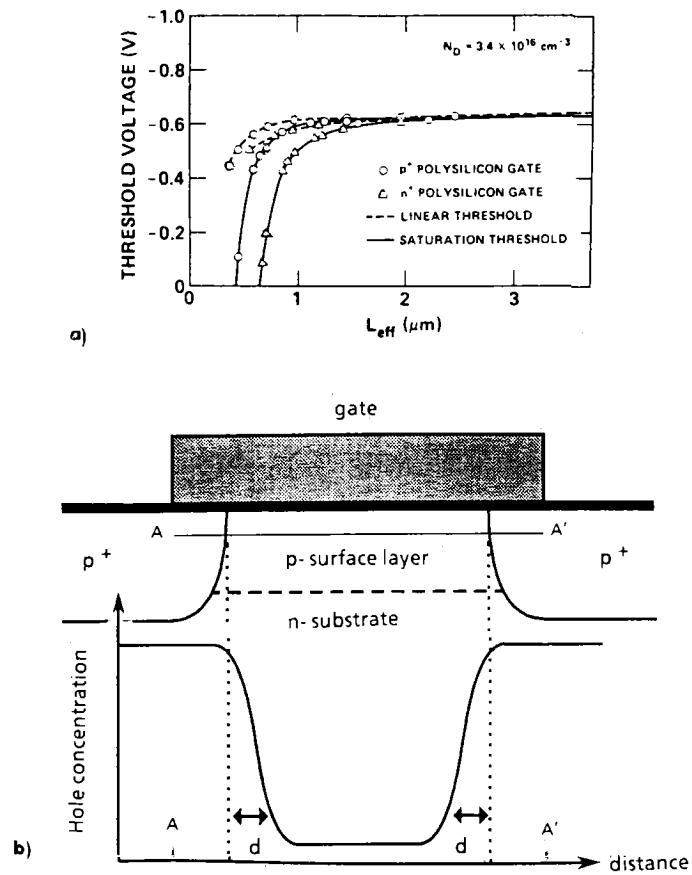


Fig. 5-57 (a) Threshold voltage variation as a function of channel length L_{eff} in surface-channel (p^+-poly gate) and buried-channel (n^+-poly gate) PMOSFETs.[114] (© 1985 IEEE). (b) The origin of short-channel threshold shifts in buried-channel PMOS transistors. From A.G. Lewis and J.Y. Chen, "Current Trends in MOS Process Integration," in N.G. Einspruch and G. Gildenblatt, Eds. *Advanced MOS Device Physics,* Academic Press, San Diego CA, 1989, p. 58. Reprinted with Permission.

Experimental values of V_T in BC PMOS devices with a gate oxide thickness of 15 nm versus channel length are shown in Fig. 5-58a. As in Fig. 5-57a we see that the magnitude of V_T decreases as L is reduced. Simulation studies indicate that the severity of the V_T roll-off is increased as Y_j is made deeper (see Fig. 5-58b). This is due to the fact that more holes are able to diffuse into the thicker p-layer of the channel region from the source/drain regions as Y_j is made deeper, accentuating the roll-off effect. As in surface channel devices, V_T roll-off is decreased as t_{ox} is made smaller (Fig. 5-58c). Thus, we conclude that smaller Y_j and t_{ox} values are needed to mitigate the V_T-roll-off effect in BC PMOS.

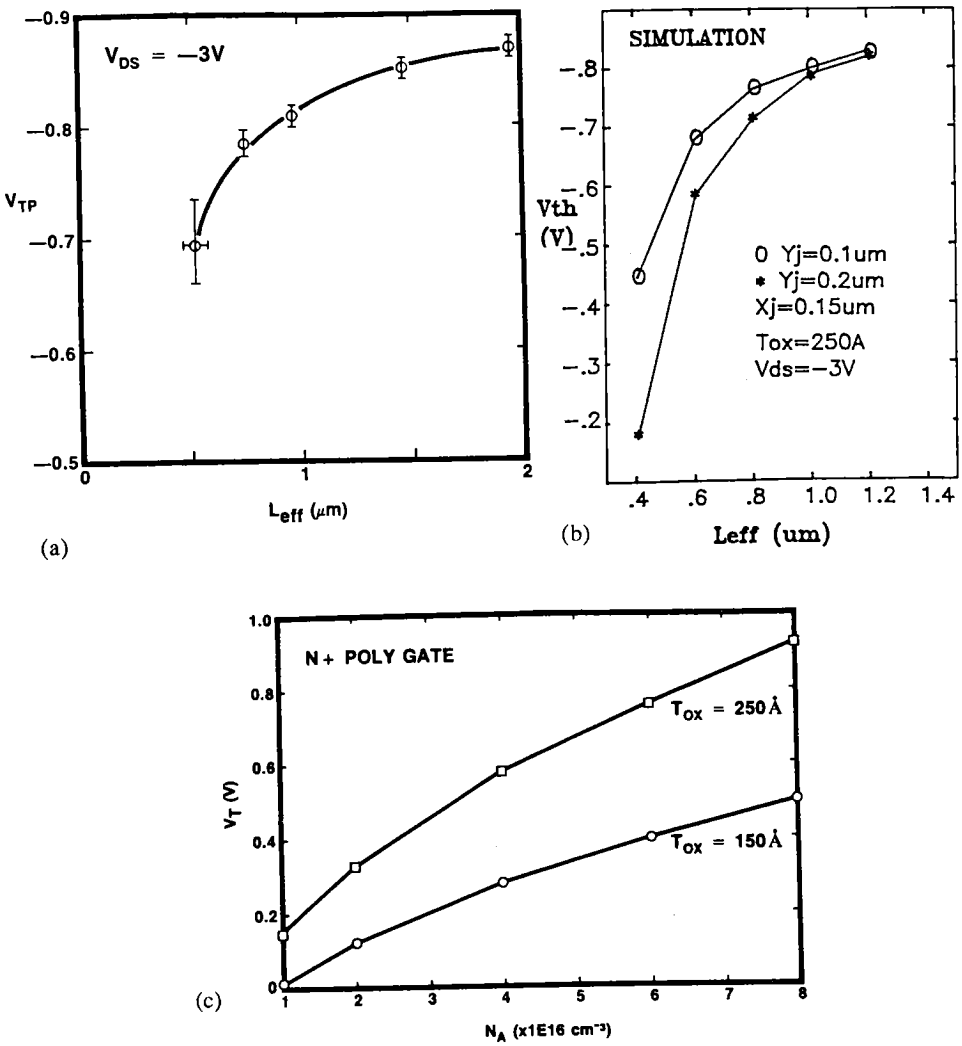


Fig. 5-58 (a) Experimental data of *p*-channel threshold voltage versus L_{eff} for gate oxide thickness of 15 nm. (b) Simulation of threshold voltage versus L_{eff} for *p*-channel MOSFETs with different Y_j. (c) Simulated *p*-channel threshold voltage versus L_{eff} for two different gate oxide thicknesses. From K. M. Cham, et al., *Computer Aided Design and VLSI Device Development*. p. 187-8, Copyright 1986 Kluwer Academic Publishers. Reprinted with permission.

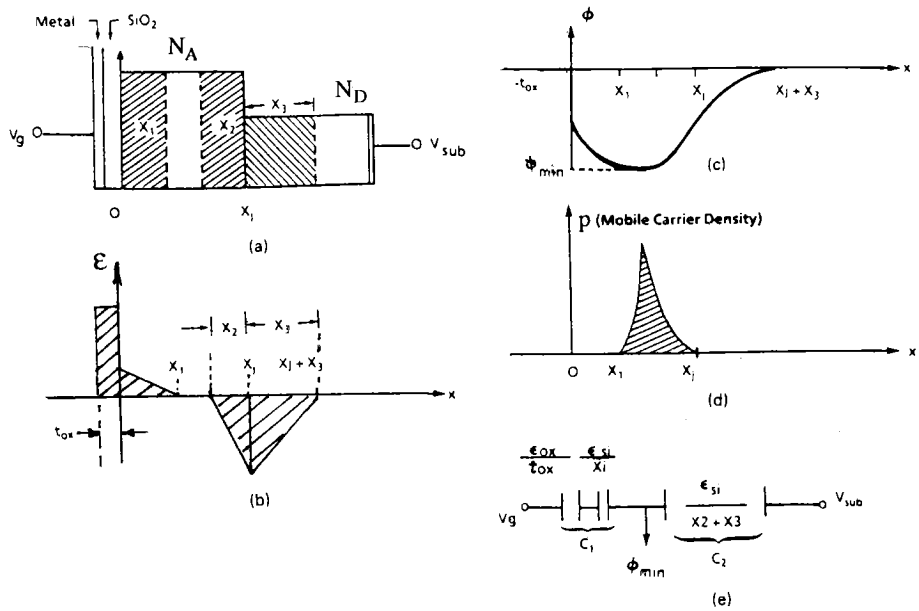


Fig. 5-59 (a) Doping profile and depletion regions; (b) electric field; (c) electron potential; (d) mobile carriers; and (e) equivalent capacitances in sub-pinchoff condition of a buried-channel PMOSFET.[120] (© IEEE 1978).

5.8.2.2 Subthreshold Swing (S_t) in Buried-Channel PMOS. The PMOSFETs used in CMOS must be enhancement-mode devices; i.e., they need to be OFF when $V_{GS} = 0V$. If such PMOSFETs are buried-channel devices, the entire implanted *p*-layer in the BC PMOS devices must be depleted of holes when $V_{GS} = 0V$. On the other hand, upon applying a gate voltage equal to the threshold voltage, the *p*-layer of such devices is no longer depleted. That is, holes will have returned to the layer, forming a *p*-type conducting channel between the source and drain. But this channel will exist beneath instead of at the surface. Thus, for gate voltages greater than V_T, two depletion regions will be present in this *p*-layer; one that is formed between the surface and the buried channel (to a depth x_1) and the other due to the channel *pn* junction, as is shown in Fig. 5-59a. The BC MOSFETs thus have two capacitances in addition to the oxide capacitance $\kappa_{ox}\varepsilon_o/t_{ox}$; namely the depletion capacitance from the surface ($\kappa_{si}\varepsilon_o/x_1$), and the depletion capacitance from the *pn* junction ($C_2 = \kappa_{si}\varepsilon_o/[x_2 + x_3] \sim \kappa_{si}\varepsilon_o/x_3$). The potential minimum is located in the position between these two capacitors, as shown in Fig. 5-59e. As also noted earlier, the location of the potential minimum causes most of the free carriers to appear in the bulk Si rather than at the Si-SiO_2 interface. It is expected that since the free carriers are now farther away from the surface they will be harder to modulate by the gate voltage. (In the same way that

carriers in a surface-channel MOSFET with a thicker gate oxide are harder to modulate - because of reduced gate-to-channel capacitive coupling - than in one with a thinner gate oxide.) Hence, to achieve the same degree of turn-off, the gate voltage will need to be reduced more than in the case of a surface-channel device. Thus, we expect that the turn-off characteristic will be less sharp, which means that S_t will be larger.

Quantitatively the buried channel turn-off characteristic can be modeled by an expression that gives the subthreshold swing S_t in a structure with two such depletion regions, i.e.,

$$S_t = 2.3\ (kT/q)\ (1 + C_2/C_1) \qquad (5 - 96)$$

$$= 2.3\ (kT/q)\ [1 + (\kappa_{ox}\varepsilon_o x_1 + \kappa_{si}\varepsilon_o t_{ox})/(\kappa_{ox}\varepsilon_o x_2 + \kappa_{ox}\varepsilon_o x_3)] \qquad (5 - 97)$$

where C_1 is the oxide capacitance (C_{ox}) in series with the depletion capacitance from the surface ($\kappa_{si}\varepsilon_o/x_1$), and C_2 is defined above. Let us compare Eq. 5-96 with Eq. 4-91 (which gives S_t for the case of surface-channel MOSFETs, and is repeated here for convenience).

$$S_t = \ln 10(d\ln I_D/dV_{GS})^{-1} = 2.3\ (kT/q)\ (1 + C_d/C_{ox}) \qquad (5 - 98)$$

By noting that C_1 is smaller than C_{ox}, and C_2 is typically comparable in value to C_d, we see that S_t is expected to be larger in BC devices. Experimental results confirm that S_t is worse (i.e., larger) in buried channel devices, especially in short-channel PMOSFETs. That is, S_t values in buried-channel PMOSFETs range from 0.5 to 1.5 V/decade, compared to the 70-120 mV/decade S_t values observed in active NMOSFET devices. Figure 5-60 compares the subthreshold behavior of the two 0.8 μm BC PMOS

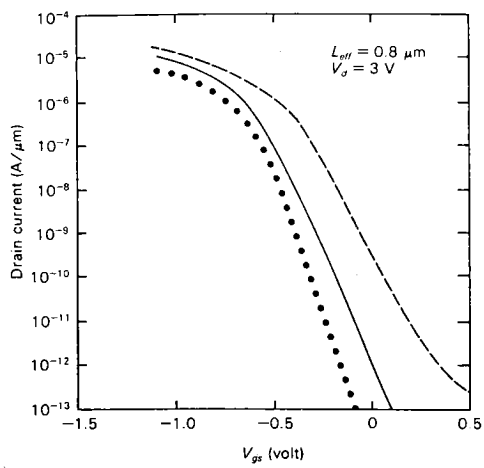


Fig. 5-60 Subthreshold I-V curves in PMOS device in CMOS. Dashed lines: buried channel PMOS in a lightly doped substrate. Solid lines: buried-channel PMOS in a retrograde *n*-well. Dotted lines: surface channel NMOS drawn for comparison.[72] (© IEEE 1982).

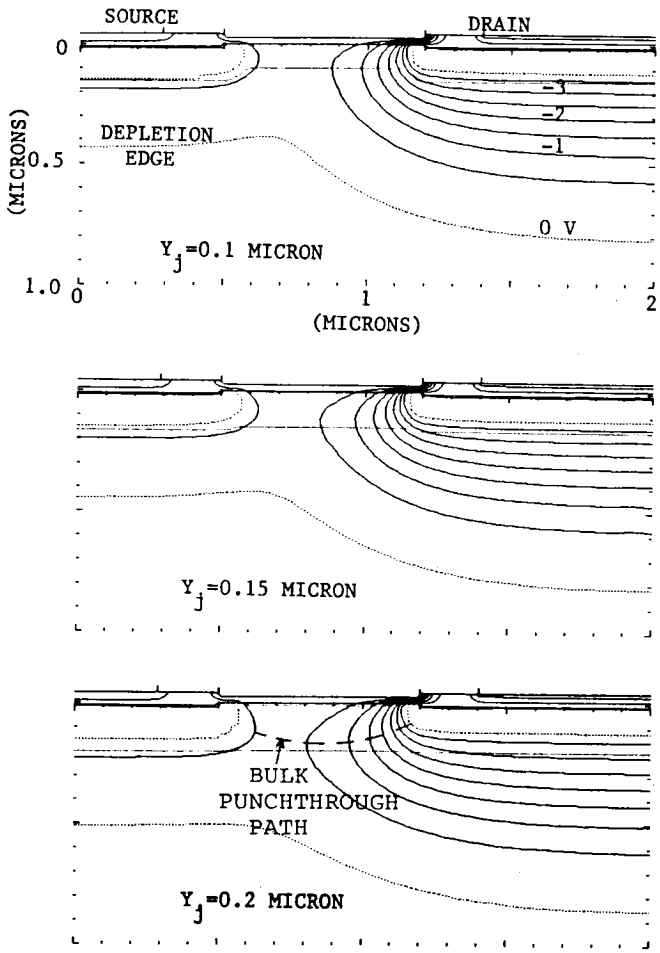


Fig. 5-61 Simulated 2-D potential profile for *p*-channel MOSFET with Y_j = 0.1, 0.15, and 0.2 μm. V_{GS} = 0V and V_{DS} = - 3V. [115] (© 1984 IEEE).

devices with that of a comparable surface-channel NMOSFET. (Note that the V_{GS} scale for the NMOSFET should be +1.5V to -1.5V from left to right.) It indicates that S_t is larger for the BC PMOS devices.

The effect on S_t of the depths of the channel and source/drain junctions Y_j and r_j, respectively, and of the gate-oxide thickness t_{ox} has also been investigated with simulation studies. Figure 5-61 illustrates the 2-D simulated potential profiles for a 0.5 μm *p*-channel MOSFET as Y_j is increased from 0.1 to 0.2 μm. It is observed that as Y_j is increased the potential contours spread more from drain to source, and the bulk punchthrough path passes through the depletion region near the *pn* junction. Figure 5-62 shows the calculated values of S_t for three Y_j values (i.e., 0.1, 0.15, and 0.2 μm), as r_j is varied in a BC PMOS device with an L_{eff} = 0.5 μm, a t_{ox} = 25 nm,

bulk punchthrough path passes through the depletion region near the *pn* junction. Figure 5-62 shows the calculated values of S_t for three Y_j values (i.e., 0.1, 0.15, and 0.2 μm), as r_j is varied in a BC PMOS device with an L_{eff} = 0.5 μm, a t_{ox} = 25 nm, and V_{DS} = -3V. Each curve corresponds to a fixed value of Y_j. The threshold voltage [defined as the gate voltage at which the drain current is 100 nA*(Z/L)] is specified at each point. The data shows that S_t is strongly dependent on Y_j and r_j. That is, as Y_j is increased for a fixed value of r_j, S_t increases. For a fixed value of Y_j, S_t also increases as r_j is increased. In Fig. 5-63 the simulated value of S_t for two different gate oxide thicknesses is shown as Y_j is varied. It is seen that S_t improves for thinner gate oxides, but the dependence of S_t on Y_j is only slightly reduced. The overall conclusion from these studies is that small values of Y_j, r_j, and t_{ox} are also needed to realize small values of S_t in submicron BC PMOS devices.

5.8.2.3 Punchthrough Susceptibility of Buried-Channel PMOS. The larger S_t values exhibited by BC PMOS devices implies that they are susceptible to larger leakage currents in the OFF state. In fact, such relatively large leakage currents are commonly observed in short PMOSFETs, especially at high drain bias. For example, it has been shown that in submicron PMOS transistors (L_{eff} = 0.8 μm), I_{Dst} will increase by two orders of magnitude if V_{DS} is increased from -1V to -10V.[74] In this case, if I_{Dst} was 1 nA at -1V, it would increase to 0.1 μA if V_{DS} was raised to -10V. It would be useful to be able to predict the value of I_{Dst} prior to fabricating the device.

As described in section 5.3.1 an increase in S_t has been correlated to the onset of subsurface punchthrough in surface-channel MOSFETs. The increase in S_t in BC PMOS devices as a result of various device parameters (i.e., L, Y_j, r_j) discussed in the previous section can also be considered from the perspective of the phenomenon of punchthrough (subsurface-DIBL), which was introduced in section 5.3. However, we will see that the subthreshold leakage behavior in BC PMOSFETs differs from that observed in surface-channel FETs. To help explain this difference, we will once again

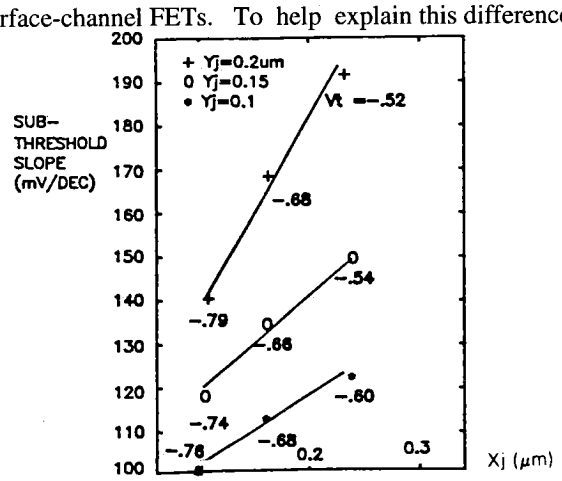


Fig. 5-62 Simulated subthreshold slope versus r_j and Y_j for *p*-channel MOSFETs in which L_{eff} = 0.5 μm, t_{ox} = 250Å, and V_{DS} = -3V. V_T is the threshold voltage.[115] (© 1984 IEEE).

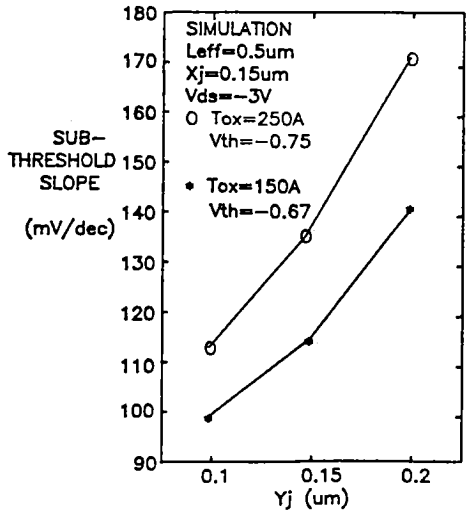


Fig. 5-63 Simulated subthreshold slope versus Y_j and t_{ox}.[115] (© 1984 IEEE).

employ the punchthrough modeling technique described in section 5.3.2, but here it will be applied to quantitatively model I_{Dst} in BC PMOSFETs. The results of this analysis will permit us to compare the subthreshold leakage behavior of surface-channel and buried channel MOSFETs.

We recall that punchthrough current is determined by carrier injection over the potential energy barrier at the saddle point of the electrostatic potential in the bulk (see Fig. 5-64).[117] Therefore, analysis of the channel potential, especially the change of the potential at the saddle point P of the BC PMOS device with V_{DS} is necessary. This analysis can be performed with the aid of 2-D device simulators,** and here we present the results of one report that performed such a study using the device simulator GEMINI.[74] The simulated potential along the direction from source to drain of a BC PMOSFET in an *n*-well with $L_{eff} = 0.55$ μm as V_{DS} is increased is shown in Fig. 5-65. It is observed that the magnitude of φ^* (the potential at the saddle point P), begins to drop as soon as V_{DS} is increased above 0V. This is due to the fact that there is no *pn* junction along the channel region in a *p*-channel BC MOSFET. (Conversely, in the case of short, surface-channel MOSFETs, at low drain-bias values the V_{DS} drop occurs almost entirely across the drain-channel junction, and the barrier at the source does not decrease as long as V_{DS} remains small. Only when V_{DS} becomes large enough to cause subsurface-DIBL does the barrier begin to drop as V_{DS} exceeds this value.) The decrease of the potential barrier height ($\varphi^* - V_S$) in a BC PMOSFET as V_{DS} is increased is plotted in Fig. 5-66. Here we see that ($\varphi^* - V_S$) is linearly proportional to the increase in drain voltage, even at low values of V_{DS}.

** Note that an entirely analytic model of punchthrough in PMOSFETs has also been published by Skotnicki et al.[117]

Once (φ^* - V_S) is found for a given value of V_{DS}, it can be used to calculate I_{Dst} with the aid of an analytic expression (Eq. 5-18b) as discussed in section 5.3. This expression for I_{Dst} is repeated here for the case of a PMOS device:

$$I_{Dst} = - q\, D_p\, Z \frac{t^*}{L^*} \frac{n_i^2}{N_D} \exp\left[\frac{q}{kT}(V_S - \varphi^*)\right] \qquad (5 - 99)$$

To exploit the results plotted in Fig. 5-66, we express Eq. 5-99 in logarithmic form as

$$\log (I_{Dst}) = \{\log [- q\, D_p\, Z\, (t^*/L^*)\, (n_i^2/N_D)]\} - q(\varphi^* - V_S)/k \qquad (5 - 100)$$

or

$$\log (I_{Dst}) = A - q\varphi_b/kT \qquad (5 - 101)$$

where $A = \log [- q\, D_p\, Z\, (t^*/L^*)\, (n_i^2/N_A)]$, and $\varphi_b = (\varphi^* - V_S)$. Now, Fig. 5-66 shows that the decrease in φ_b is linearly proportional to V_{DS}, i.e.,

$$\varphi_b = (\varphi_{bo} - b\, |V_{DS}|) \qquad (5 - 102)$$

where φ_{bo} is the barrier height at $V_{DS} = 0V$, and b represents the slope of the curve of Fig. 5-66 (i.e., $d\varphi_b/dV_{DS}$). We also note that the values of both φ_{bo} and b can be determined from Fig. 5-66. Then we can write Eq. 5-101 as

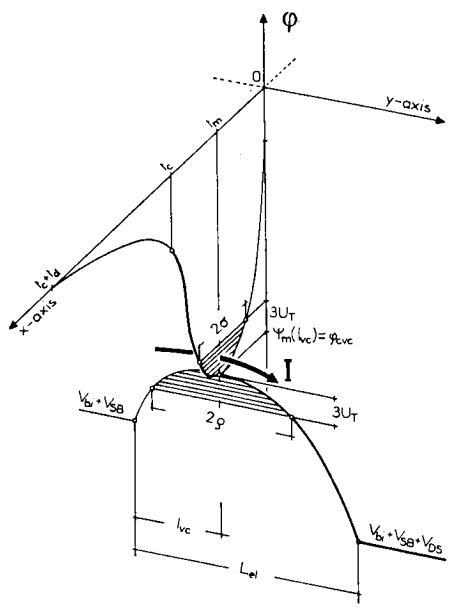


Fig. 5-64 Perspective view of the electrical potential distribution in two perpendicular cross sections of the channel of a buried-channel PMOSFET. Shadowed regions are those corresponding to t* and L^* in Eq. 5-99.[117] (© 1989 IEEE).

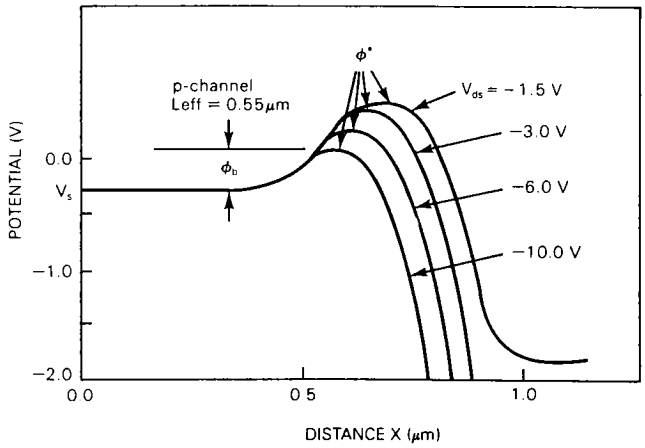


Fig. 5-65 Simulated potential distributions along the buried channel of a PMOSFET for various drain voltages.[74] (© 1988 IEEE).

$$\log (I_{Dst}) = A - q(\varphi_{bo} - b\,|V_{DS}|)/kT \tag{5 - 103}$$

or

$$= A' + V_{DS}/R \tag{5 - 104}$$

where $A' = A - q\varphi_{bo}/kT$ and $R = kT/bq$. The parameter R is the incremental voltage that the drain can sustain before I_{Dst} increases by another order of magnitude ($R = \partial V_{DS}/\partial \log I_{Dst}$). The larger the value of R, the lower the DIBL rate, hence the better the MOSFET is able to maintain low subthreshold leakage as the drain bias rises.

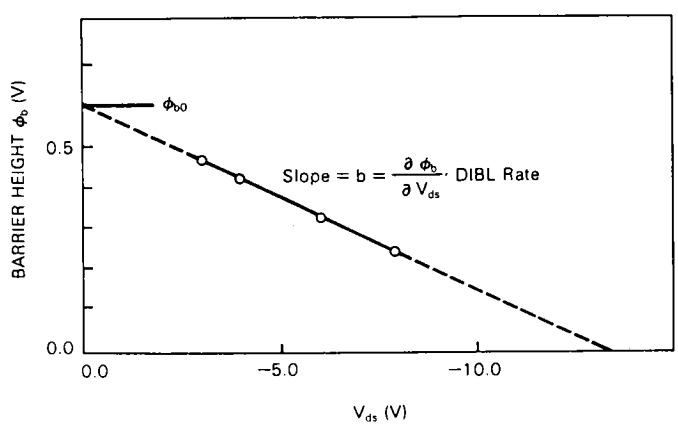


Fig. 5-66 Potential barrier height as a function of drain voltage.[74] (© 1988 IEEE).

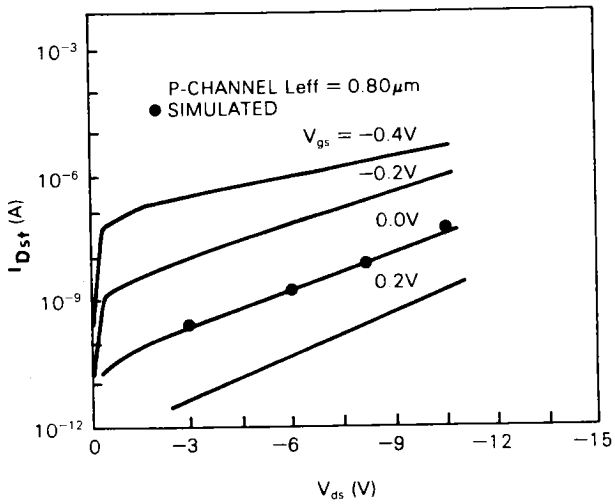


Fig. 5-67 Subthreshold I_{Dst}-V_{DS} characteristics for a submicron PMOSFET.[74] (© 1988 IEEE).

Equation 5-104 implies that the relationship between log I_{Dst} and V_{DS} should be linear in BC PMOSFETs operating in subthreshold. Figure 5-67 shows the measured log I_{Dst}-V_{DS} curves for a 0.8 μm BC PMOSFET, together with some simulated I_{Dst} values at V_{GS} = 0V. The agreement between the experimental and simulated data is good. We also see that R in this device is about -4.4V/decade for V_{GS} = 0V. This means that if I_{Dst} is 10 pA at V_{DS} = -1V, it will increase to 10 nA at V_{DS} = -10V.

Now we can compare the I_{Dst} behavior of SC and BC MOSFETs with respect externally applied biases. We note that in the case of surface channel MOSFETs two regions of different slope are observed in the log I_{Dst} versus V_{DS} curves (see Fig. 5-11). At low values of V_{DS} the value of R is large (e.g., in Fig. 5-11b R = 12.6V/decade when V_{DS} <6.0V), but when V_{DS} exceeds about 8.0V, the value of R drops to about 1.0V/decade. This change implies that in surface-channel MOSFETs there is a transfer from surface-DIBL to subsurface-DIBL (punchthrough) - e.g., in the device of Fig. 5-11b at about 6.0V. Also, up to some value of V_{DS} (e.g., V_{PT}), the drain bias has little effect on I_{Dst}. In BC PMOSFETs, however, the value of R is relatively constant (e.g., it has a magnitude of R = 4.4V/decade for all values of V_{DS} in the case of the device characterized in Fig. 5-67). This implies that the nature of subthreshold leakage current in BC PMOSFETs does *not* undergo a change from surface- to subsurface-DIBL, but instead the DIBL phenomenon in a BC device is a combination of both effects. Furthermore, the intermediate value of R also supports the view that *both* the gate and the drain voltages affect the potential of the p-type layer, and thus both can impact I_{Dst} at all values. Finally, this indicates that the measured slope of log I_{Dst} versus V_{DS} (defined as $1/R$), reflects the effect of V_{DS} on φ_b through DIBL, and permits a comparison of the subthreshold leakage behavior of various PMOS device structures.

5.8.3 Techniques Used to Suppress Short-Channel Effects in Buried-Channel PMOS

5.8.3.1 Techniques Used to Keep the Implanted *p*-Layer as Thin as Possible. The most obvious solution to suppressing short-channel effects in PMOS devices is to increase the PMOS device channel length. This is done in many CMOS technologies, and it is the reason that the minimum channel lengths of PMOS devices are often larger than those of the NMOS devices on the same chip. But we also noted that to achieve small values of S_t, less V_T-roll-off, and larger punchthrough voltages in BC PMOS devices, the *p*-buried layer depth Y_j must be as shallow as possible. That is, the thinner the *p*-layer, the more the device resembles a surface channel FET. Several methods have been reported to realize shallow *p*-layer implant profiles, including the following:

1. Using the lowest practical implant energy, which in ordinary medium-current implanters is typically 20-25 keV.

2. Implanting BF_2^+ (which produces shallower boron layers than implantation with elemental boron, see Vol. 1, chap. 9).

3. Using a high-energy *n*-implant (e.g., P^+ at 400 keV) to form a retrograde *n*-well in the regions where the PMOS devices are formed. This places more *n*-type dopant atoms beneath the channel *pn* junctions, and the more heavily doped regions absorb the drain voltage in a shorter distance, while simultaneously squeezing the channel *pn* junction toward the surface, as shown in Fig. 5-68.

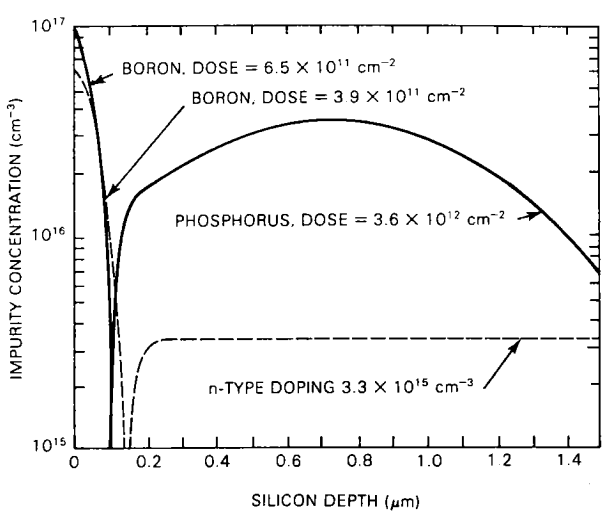


Fig. 5-68 Simulated doping profiles in PMOS channel regions: dashed lines lightly doped substrate; solid lines, retrograde *n*-well.[72] (© 1982 IEEE).

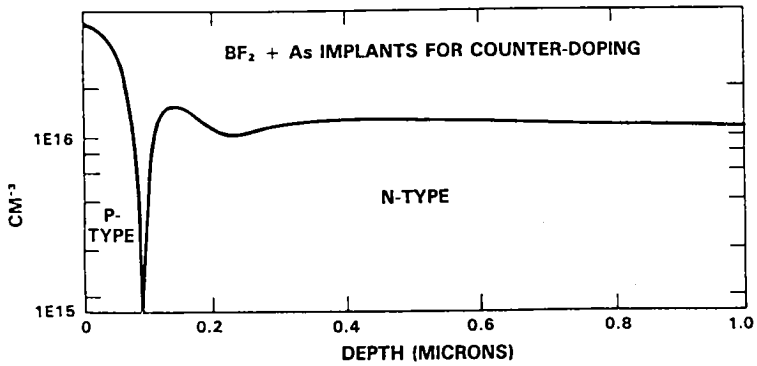


Fig. 5-69 (a) Simulated channel profile for p-channel with BF_2 and As counterdoping.[115] (© 1984 IEEE).

4. An extra n-type implant (typically arsenic) can be performed into a conventional n-well or n-substrate, with it's peak located below the p-layer immediately following the BF_2^+ implant (Fig. 5-69a). The n-doping under the p-layer is thus raised, making the junction depth lie further up the tail of the boron distribution (and thereby making it shallower). The arsenic diffuses relatively slowly, and thus will largely remain in the vicinity of where it was implanted. This forms a steep profile at the counter-doping junction, and maintains a shallow junction depth. The simulation shown in Fig. 5-69a indicates that Y_j with such an arsenic implant is ~0.09 μm. Such an extra implant can also be used with a retrograde n-well.

Two drawbacks of this technique are that since the boron channel implant is compensated by an arsenic implant very good process control must be maintained for it to be successful, and it also requires an extra masking step to prevent the arsenic implant from entering the n-channel transistor regions.

To prevent a shallow implanted-boron layer from growing thicker, it is necessary to use a reduced thermal budget in order to restrict boron diffusion following the implant. Specific steps for restricting boron redistribution include the following:[76]

1. Implanting the boron through the gate oxide. This avoids the oxidation-enhanced diffusion of boron that would occur during the growth of the gate oxide if the implant were performed first. In this case, it is necessary to prevent the gate oxide from becoming contaminated during implant, either by material sputtered by the beam line or by vaporized resist material used as a mask against the implant. To prevent such contamination, a thin layer of polysilicon may be

deposited on the gate oxide prior to the implant (in fact, immediately after the oxide is grown).[77]

2. After the implant has been performed, the remainder of the polysilicon film is deposited. In order to avoid the 900°C thermal cycle that would have to be used if the poly were doped following deposition, this polysilicon is doped *in situ* at ~600°C with phosphorus.

3. A BPSG glass layer is used as the dielectric between the gate and the first level of metal. A significantly lower temperature can be used to flow BPSG than PSG (e.g., 900°C versus 1000°C). A lower temperature cycle can thus be used to smooth the surface topography (flow step) and gently taper the contact holes after etch (reflow step).

5.8.3.2 Shallow p^+ Source/Drain Junctions to Suppress Punchthrough in PMOSFETs. As noted in our discussion of MOSFET scaling in section 5.6.1.2, reducing the source/drain junction depths also decreases the susceptibility to short-channel effects. By reducing the vertical depth of the source/drain junctions the lateral spread of the dopants in these regions beneath the gate also becomes smaller. Hence, a longer effective channel length is possible for a fixed drawn gate length. In addition, the lateral spread of source and drain depletion regions below the surface layer makes a MOSFET more prone to punchthrough. Consequently, shallow source and drains effectively suppress subsurface punchthrough paths.

In NMOSFETs, it is easier to fabricate shallow n^+ source/drain regions (i.e., ≤ 0.2 μm in depth) because dopants with heavy ionic masses (usually arsenic) are implanted to form these regions. They yield implant profiles with small projected ranges and little channeling. Furthermore arsenic has a low diffusivity, so that little redistribution occurs during postimplant annealing. In PMOSFETs, however, the only practical p-type dopant for forming the p^+ regions is boron, and junctions formed with boron are much deeper than those formed with arsenic for several reasons. First, the projected range of the low-mass boron is relatively larger, even at low implantation energies. Ion channeling also causes the implanted boron profile to have a long tail, even if the target wafers are tilted to minimize this effect. Thus, the minimum junction depth which results when boron is implanted into a lightly doped n-type crystalline silicon wafer is about 0.35 μm.

The high diffusivity of boron compounds the problem. That is, in CMOS processes both the n- and p-implants that form the NMOS and PMOS source and drain regions, respectively, must be annealed. The anneal used for the arsenic implant will cause significant (>0.1 μm) redistribution of the boron implant. In CMOS technologies with channel lengths of 2 μm or longer, this issue does not arise because the boron can be implanted after the arsenic implantation and anneal steps (i.e., allowing the boron to be annealed only for the minimum time and temperature needed for full activation). Thus, boron redistribution after implantation can be minimized. However, in submicron CMOS processes it is generally not possible to perform the boron implant after the n^+ arsenic implant. This is because submicron NMOS devices usually use LDD structures,

in which the arsenic implant is performed after spacers have been formed. The boron p^+ implant, however, must be carried out before the spacers are formed for two reasons. First, this ensures that the boron implant sufficiently counterdopes the n^- implant (used to form the LDD tip region) everywhere in the areas of the PMOS source and drain (i.e., the n^- implant is typically done unmasked to maintain process simplicity). Second, this also guarantees that the p^+ doping will extend under the gate. Thus, in CMOS processes that use LDDs in the NMOS devices the p^+ regions must experience the same anneal cycle as the n^+ regions. Alternative approaches have been proposed to avoid subjecting the implanted boron to the n^+ anneal, as will be described below.

Techniques to decrease the p^+ junction depths fall into two groups: (1) those which set out to reduce the junction depth of the implanted boron profile; and (2) those which attempt to reduce the diffusion of the boron during the post-implantation anneals.

The approaches mentioned earlier when discussing techniques to reduce the depth of the as-implanted boron region in the channel of PMOS devices also apply to reducing the source/drain junction depths. These include use of BF_2^+ instead of atomic boron as the implant species, and implanting through a screen oxide. Pre-amorphizing the wafer surface region into which the junction dopants will be implanted to a depth of several tenths of a micron (with an earlier implant of silicon or germanium) has also been investigated (see Vol. 2, chap. 3). The amorphous layer effectively prevents channeling, and subsequent annealing at a low temperature is sufficient to achieve recrystallization by solid-phase epitaxy. Use of very-low-energy ion implanters (<5 keV) for implanting the boron has also been studied, as have techniques for implanting the boron into silicide layers formed above the source/drain regions and then diffusing the boron into the substrate from these silicide layers.

Several techniques to limit the redistribution of boron after implant have also been suggested. The first involves use of rapid thermal processing instead of a conventional furnace treatment following the implant. Relatively high temperatures are employed to realize adequate dopant activation, while keeping the times short enough so that boron redistribution is minimized. The second makes use of lower temperature flow and reflow steps by employing BPSG as the interlevel dielectric material instead of PSG. The third modifies the process sequence so that the boron implant is done after the arsenic implant. If the spacers required for NMOS LDD formation are in place, the boron implant must be sufficiently redistributed to ensure that the p-regions extend under the p-channel gates.[105] On the other hand, some process sequences have been reported in which the spacers are removed prior to implanting the boron. Other materials than SiO_2 must be used for such removable spacers (e.g., silicon nitride,[102] TiN,[103] or polysilicon[104]) since removing SiO_2 spacers would significantly thin the field oxide. Concerns about damaging the gate oxide at the gate edge have been raised if dry etching is used for such spacer removal. In the reports describing silicon nitride and TiN removable spacers, wet etching is used to remove them.

5.8.3.3 LDD Structures in PMOS to Limit Punchthrough. Another way to achieve a shallow junction in a PMOS device is to use an LDD structure. The junction depth of the lightly-doped LDD tip region is shallower than the heavily doped

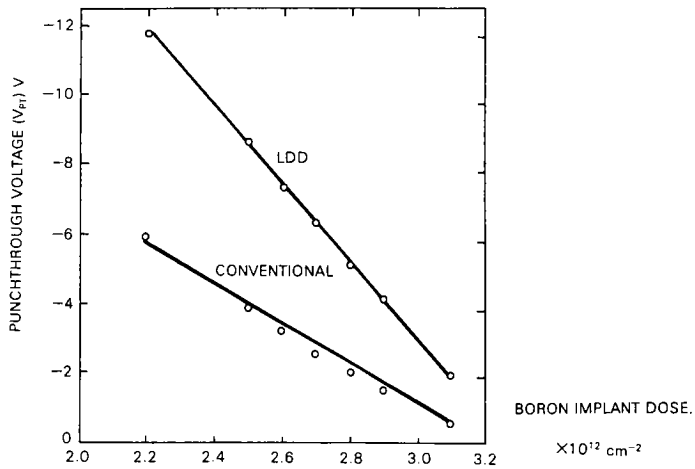


Fig. 5-70 Punchthrough voltage as a function of boron implant dose for a 0.4 μm PMOSFET with and without LDD structure.[75] (© 1986 IEEE).

p^+ source/drain junction depth implanted at the same energy. Figure 5-70 indicates that the punchthrough voltage V_{PT} is increased by 2-3V when an LDD is incorporated into a PMOS device. (The LDD structure, of course, may also be used to increase the hot-carrier resistance of the PMOS devices, especially as the channel length is decreased below about 0.8 μm.) The LDD, however, will generally reduce I_D due to a rise in the series resistance of the drain region. Thus, a tradeoff analysis between an increase in V_{PT} and a decrease in the drain current must be made.

Adding an LDD structure to a PMOS device in CMOS involves relatively straightforward changes to the process sequence. The gate sidewall spacers are already available from the creation of NMOS LDDs, but there is some increase in process complexity, (especially if the former process used a maskless n^- implant to form the tip regions of the NMOS LDD). That is, to form both PMOS and NMOS LDDs four masking steps are generally needed (n^-, p^-, n^+, p^+), rather than just two. However, a two-mask process, which employs removable polysilicon spacers, has also been reported.[104]

5.8.3.4 New PMOS Device Structures Designed to Suppress Punchthrough. New PMOS device structures have also been proposed to minimize punchthrough current. For example, a small segment of an n-type region can be created by ion implantation between the p-type S/D and the channel as a punchthrough stopper (refer to Fig. 5-71). This structure has been termed a double-implanted or *halo structure*. One report of such a halo device with $L = 0.5$ μm used a phosphorus implant dose of 1×10^{12} cm^{-2} at 130 keV.[78] A sidewall spacer was used to form the lateral offset of the n-type region in this device. However, since the p^+ source/drain is offset from the gate by the sidewall spacer, this method of forming the local n-regions degrades the drive current and impairs hot-electron immunity.[79] Furthermore, as this technique also places a highly doped n-region adjacent to a highly doped p-region, the maximum halo doping

is determined by the onset of low-level avalanche at the chosen supply voltage and channel length.

A large-tilt-angle implant can be used to fabricate halo-type structures without a sidewall spacer, thereby gaining the benefits while avoiding the problems of the structure of Fig. 5-71.[79] The dopant is placed precisely between the p^+ source and the buried p-channel junction. The lateral penetration and depth of the n-type region are controlled by the implant energy and tilt-angle. The peak concentration is set by the dose. A typical implant sequence to form such a *large-angle tilt implanted punchthrough stopper* (LATIPS) structure as shown in Fig. 5-72 is: (1) implant phosphorus at 90 keV with a dose of $2x10^{13}$ cm^{-2} at 25° tilt, in two segments, with the target rotated 180° between segments; then (2) implant BF_2^+ (source/drain) at 40 keV (no tilt) to a dose of $3x10^{15}$ cm^{-2}.[80]

Devices fabricated using the LATIPS process were compared to conventional PMOSFETs. LATIPS devices showed dramatically improved punchthrough characteristics for physical gate lengths as small as 0.5 microns. As an example, the OFF-state leakage current was eight orders of magnitude smaller (at -3V drain bias) than that of a conventional control device fabricated without the LATIPS structure. This dramatic improvement in punchthrough resistance was achieved without extreme scaling of the source/drain junction depth or the n-well background doping. That is, the p^+ source/drain junction depth was about 0.2 μm and the n-well concentration was $1x10^{16}$ cm^{-3}. Furthermore, the benefits were obtained without significant degradation of device performance. That is, no gate-oxide degradation was observed as a result of using BF_2^+ source/drain implants at 8 keV, even at doses of $5x10^{15}$ cm^{-2}. Better high-current drive capability than conventional devices was also exhibited, as well as a smaller body (back-gate bias) effect.

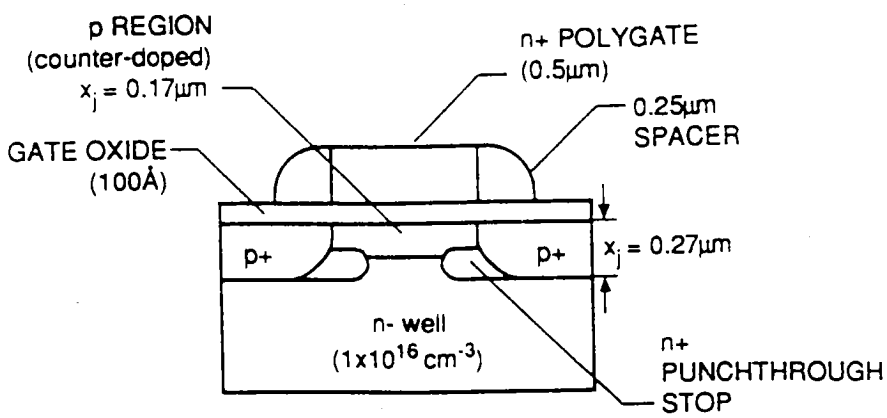


Fig. 5-71 PMOS Punchthrough-stop structure.[78] (© 1985 IEEE).

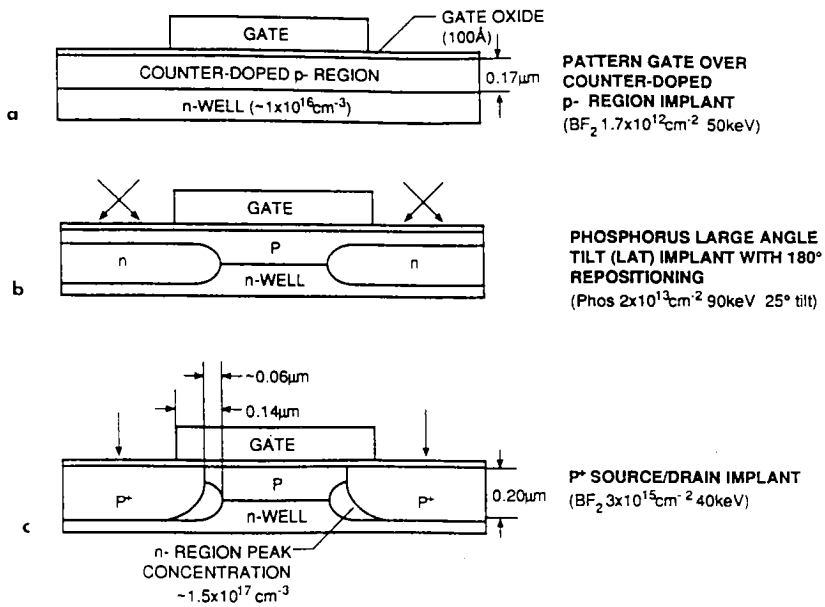


Fig. 5-72 LATIPS process sequence for PMOSFET: (a) Pattern gate over an n-well region that has been conterdoped with a p-type implant (BF_2); (b) Large angle tilt implant with phosphorus; (c) p^+ Source/Drain implant (BF_2).[79] (© 1988 IEEE).

5.8.4 PMOS Devices with p^+-Polysilicon Gates

The use of a boron implant to adjust V_{Tp} becomes less feasible as devices use even thinner gate oxides, since larger doses of boron are needed. Y_j thus becomes deeper, and the punchthrough problem worsens. In addition, buried channel devices are subject to more process variability stemming from the difficulty of controlling the compensating implant. Solutions involving the use of gate electrodes other than n^+ polysilicon must therefore be explored. One alternative is to use p^+ polysilicon for PMOS devices. This approach appears to become necessary for CMOS and BiCMOS once L becomes smaller than 0.5 μm.

When p^+ polysilicon is used for the gate material, V_{Tp} is shifted from the values that occur when n^+ polysilicon is used. Figure 3-20 also plots V_{Tp} as a function of substrate doping when p^+ polysilicon is employed (using the scale on the right side of the figure). V_{Tp} can be seen to be less negative than -0.7 V over the substrate doping range of 10^{15}-10^{17} cm^{-3}. Thus, V_{Tp} can easily be made more negative through the implantation of phosphorus or arsenic into the channel. If p^+ polysilicon is used throughout the circuit, however, the NMOS device then has to be overcompensated in order for V_{Tn} to be reduced to sufficiently small values. (Note that some reports have been published on the use of p^+ polysilicon alone.[81,82]) This implies it would be ideal

to use both n^+ and p^+ poly gates on the same chip (with n^+ poly for NMOS devices, and p^+ poly for PMOS devices).[83]

Such a dual-doped poly approach, however, adds process complexity, and also introduces other problems. One drawback of is that the sheet resistance of p^+ poly is ~3x higher than that of an n^+ poly film of the same thickness, because the mobility of holes is ~3x smaller than that of electrons in Si. A second problem arises from the need to connect the two types of poly (e.g., at the input of an inverter). Such problems occur when a silicide overlayer (or strap) is used to make the connection between n^+ and p^+ poly (as a method to avoid a separate, space-consuming metal contact). Because the silicide strap provides a very rapid diffusion path for boron and arsenic, one type of poly can be counterdoped by the other when the device is subjected to high-temperature excursions. This counterdoping can occur to the degree that a region of poly can change doping types (i.e., from n^+ to p^+). In such cases, the threshold voltage of devices with counterdoped poly will be shifted from their designed value.

If the processing temperature is limited to 800°C after the two types of poly have been connected by the silicide, such lateral diffusion does not produce significant shifts in V_T.[84] On the other hand, temperatures of 900°C will produce sufficient counter-doping to significantly shift V_T. Since one of the last high temperature steps in CMOS is activation of the source/drain implants, this would mean either using a lower-temperature activation step, or performing the step prior to formation of the silicide layer.

Another approach is to form a polysilicon electrode with an overlying conductor layer that suppresses such counterdoping. One such structure is a W-TiN-poly electrode structure.[85] The thin (30-nm) TiN film acts as a diffusion barrier to the dopants in the poly and also prevents reaction between the poly and the W. Very little of the dopant present in the poly is found to diffuse into the W, even following an anneal at 900°C for 1 hour.

Another problem encountered with p^+ poly gates when a thin gate oxide is used is poor V_T process control in the PMOS devices, due to penetration of the boron into the oxide (or further, into the silicon). It is reported that boron will penetrate gate oxides that are ≤12.5 nm thick during a 900°C 30-minute post-implant anneal in N_2.[86] This implies that if p^+ poly gates are used, a lower processing temperature may be needed. If too low a temperature is used, however, the boron implanted into the polysilicon will not be sufficiently redistributed. The polysilicon dopant concentration at the polysilicon-gate oxide interface could thus end up being less than the mid-$10^{19}/cm^3$. This would make the work function of the polysilicon different from the desired degenerately doped value, creating V_T control problems in the MOS devices.

It has also been found that the presence of fluorine in the gate oxide worsens the boron penetration problem (Fig. 5-73).[87,88] Such fluorine can be introduced into the gate oxide if the PMOS source drain regions are implanted using BF_2^+. Elemental boron is therefore considered inherently superior to BF_2^+ as the implant species for surface-channel PMOS devices in CMOS technologies that use p-doped polysilicon. A

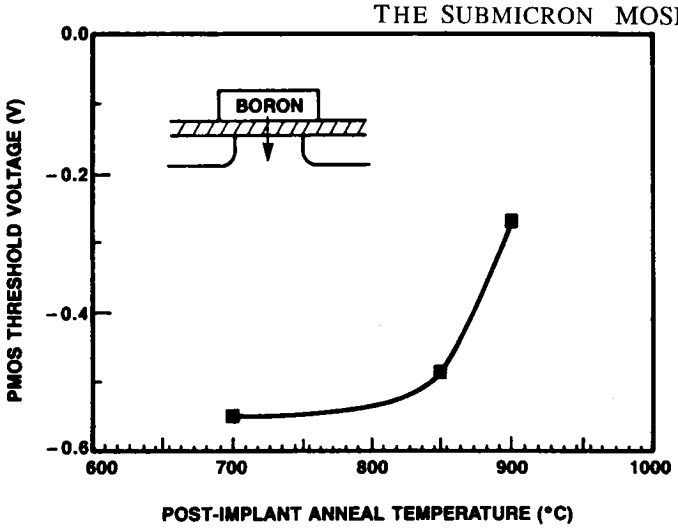


Fig. 5-73 Boron penetration through the gate oxide causes V_T in BF_2 implanted PMOSFETs to shift positive at anneal temperatures above ~800 °C.[87] (© IEEE 1989).

study of enhanced boron diffusion through thin SiO_2 layers in a wet oxygen atmosphere is also reported in reference 89.

It also appears that nitrided oxides and reoxidized nitrided oxides (see chap. 9) are effective in blocking boron diffusion.[90] For this, and other reasons, they are being extensively studied for use as candidates to replace SiO_2 as the gate insulators in dual-gate CMOS processes. An alternative approach to introducing nitrogen into the gate oxide to block boron penetration is through ion implantation. Recently, ion implantation of nitrogen into the gate *p*- and *n*-doped poly electrodes of a 0.25 μm CMOS process has been reported. The heavy nitrogen dose (1×10^{15} cm^{-2}) resulted in the gate oxide incorporating high concentrations of nitrogen. This had the effect of suppressing boron penetration of the gate oxide.[110]

Hot carrier effects in p^+-poly 0.5-μm PMOSFETs have also been examined. Stress experiments showed significantly more electron trapping in short stress times than is observed in NMOSFETs. The device degradation due such trapping can be severe enough that it may pose a problem in such devices.[118]

REFERENCES

1. J.E. Chung et al., "Performance and Reliability Design Issues of Deep Submicron MOSFETs," *IEEE Trans. Electron Devices* **ED-38**, March 1991, p. 545.
2. J. Greenfield and R. Dutton, "Nonplanar VLSI Device Analysis Using the Solution of Poisson's Equation," *IEEE Trans. Electron Dev.*, **ED-27**, p. 1520 (1980).
3. L..D. Yau, "A Simple Theory to Predict the Threshold Voltage in Short-Channel IGFETs," *Solid-State Electronics,* **17**, p. 1059, (1974).

4. W. Fichtner and H.W. Poetzl, "MOS Modelling by Analytical Approximations - Subthreshold Current and Subthreshold Voltage," *Int. J. Electronics*, **46**, p. 33, Jan. 1979.

5. C.R. Viswanathan, B.C. Burkery, G. Lubberts, and T.J. Treadwell, *IEEE Trans. Electron Dev.*, **ED-32**, p. 932, (1985).

6. R.R. Troutman, "VLSI Limitations from Drain-Induced Barrier Lowering," *IEEE Trans. Electron Dev.*, **ED-26**, p. 461, (1979).

7. F.-C. Hsu et al., *IEEE Trans Electron Dev.*, **ED-30**, p. 571 (1983).

8. S.C. Jain and P. Balk, "A Unified Analytical Model for DIBL ... in Short-Channel MOSFETs," *Solid-State Electronics*, **30** (5), p. 503, May 1987.

9. T. Skotnicki et al., *IEEE Electron Dev. Letts.*, **EDL-9**, p. 109, March 1988.

10. G.W. Taylor, *Solid-State Electronics*, **22**, p. 701, (1979).

11. B.J. Sheu, D.L. Scharfetter, P.K. Ko, and M.-C. Jeng, "BSIM: Berkeley Short-Channel IGFET Model for MOS Transistors," *IEEE J. of Solid State Circuits*, August 1987, p. 558.

12. G. Merkel, *Process and Device Modeling for IC Design*, (F. Van de Wield et al., Eds.), Noordhoff, Leyden, (1977), p. 705.

13. L. A. Akers et al., "A Model of a Narrow-Width MOSFET including Tapered Oxide and Doping Encroachment," *IEEE Trans. Electron Dev.* **ED-28**, p. 1490, 1981.

14. C.R. Ji and C.T. Sah, "Two-Dimensional Analysis of the Narrow-Gate Effect in MOSFET," *IEEE Trans. Electron Dev.*, **ED-30**, P. 635, June 1983.

15. T. Furukawa and J.A. Mandelman, "Process and Device Simulation of Trench Isolation Corner Parasitic Device," *Ext Abs. Mtg. Electrochem. Soc.*, Fall 1988, p. 329.

16. L. A. Akers, "The Inverse Narrow-Width Effect," *IEEE Electron Dev. Letts.*, **EDL-7** (7), p. 419, July 1986.

17. E.H. Li et al., "The Narrow-Channel Effect in MOSFETs with Semi-Recessed Oxide Structures," *IEEE Trans. Electron Dev.*, **ED-37** (3) p. 692, March 1990.

18. S. S. Chung and T.-C. Li, "An Analytical Threshold-Voltage Model of Trench-Isolated MOS Devices with Nonuniformly Doped Substrates," *IEEE Trans. on Electron Dev.*, March 1992, p. 614.

19. K. Ohe et al., "Narrow-Width Effects of Shallow Trench-Isolated CMOS with n^+-Polysilicon Gate," *IEEE Trans. Electron Dev.*, **ED-36** (6), p. 1110, June 1989.

20. G. Merkel, "A Simple Model of the Threshold Voltage of Short and Narrow Channel MOSFETs," *Solid-State Electronics.*, **23**, p. 1207, December 1983.

21. S. Selberherr et al., "MINIMOS - A Two-Dimensional MOS Transistor Analyzer," *IEEE J. Solid State Circuits*, **SC-15**, p. 605 (1980).

22. Z.-H. Liu et al., "Threshold Voltage Model for Deep-Submicrometer MOSFETs," *IEEE Trans. Electron Dev.*, **ED-40**, January 1993, p. 86.

23. J.R. Brews, W. Fichtner, E.H. Nicollian, and S.M. Sze, "Generalized Guide for MOSFET Miniaturization," *IEEE Electron Dev. Letts.*, **EDL-1**, p. 2, 1980.

24. C. Hu, "Future CMOS Scaling and Reliability," *Proceedings of the IEEE*, **81**, p.682, May 1993.

25. F.M. Klaasen, "Design and Performance of Micron-Sized Devices," *Solid-State Electron.*, **21**(3), p. 565 (1978).

26. C.F. Codella and S. Ogura, "Halo Doping Effects in Submicron DI-LDD Device Design," *Tech. Dig. IEDM*, p. 230, 1985.

27. A. Hori et al., "A Self-Aligned Pocket Implantation (SPI) Technology for 0.2 μm-Dual Gate CMOS," *Tech. Dig. IEDM,* p. 641, 1991.

28. H. Nihara et al., "Anomalous Drain Current in NMOS and its Suppression by Deep Ion Implantation," *Tech. Dig. IEDM,* p. 487 (1978).

29. L.C. Parrillo, "CMOS Active and Field Device Fabrication," *Semicond. International,* April 1988, p. 64.

30. T. Skotnicki, G. Merkel, and T. Pedron, "A New Punchthrough Current Model Based on the Voltage-Doping Transformation," *IEEE Trans. on Electron Dev.,* **ED-35**, July 1988, p. 1076.

31. C. Duvvury, "A Guide to Short-Channel Effects in MOSFETs," *IEEE Circuits and Systems Magazine,* November 1986, p. 6.

32. J. A. Cooper and D.F. Nelson, "Measurement of the high-field drift velocity of electrons in the inversion layer in silicon," *IEEE Electron Dev. Lett.,* **EDL-2**, July 1983, p. 169.

33. R.H. Crawford. *MOSFETs in Circuit Design*, McGraw-Hill, New York, 1967.

34. D. Frohman-Bentchkowsky, "On the effect of mobility in inversion and accumulation layers on thermally oxidized silicon surfaces," *Proc. IEEE*, **vol. 56**, p. 217, 1968.)

35. A.G. Sabnis and J.T. Clemens, "Characterization of Electron Velocity in the Inverted <100> Si Surface," *Tech. Dig. - IEDM*, p. 19, 1979.

36. M.S. Liang et al., "Inversion layer capacitance and mobility of very thin gate oxide MOSFETs," *IEEE Trans. Electron Dev.,* **ED-33**, March 1986, p. 409.

37. K.K Hung, *Electrical characterization of the Si-SiO_2 interface for thin oxides,* Doctoral Thesis, Hong Kong University, (1987).

38. C.G. Sodini, T. Ekstedt, and J. Moll, "Charge Accumulation and Mobility in Thin Dielectric MOS Transistors," *Solid-State Electronics*, **25**, p. 833, (1982).

39. K. Lee et al., "Physical Understanding of Low Field Carrier Mobility in Silicon Inversion Layer," *IEEE Trans. Electron Devices,* **ED-38**, August 1991, p. 1905.

40. S.R. Hofstein and G. Warfield, "Carrier mobility and current saturation in the MOS transistor," *IEEE Trans. Electron Dev.,* **ED-12**, p. 129, (1965).

41. F.N. Trofimenkoff, "Field-dependent mobility analysis of the field-effect transistor," *Proc. IEEE*, **53**, p. 1765, (1965).

42. D.M. Caughey and R.E. Thomas, "Carrier mobilities in silicon empirically related to doping and field," *Proc. IEEE*, **55**, p. 2192 (1967).

43. C.G. Sodini, P.K. Ko, and J.L. Moll, "The effect of high fields on MOS device and circuit performance," *IEEE Trans. Electron Dev.,* **ED-31**, October 1984, p. 1386.

44. E.A. Talkhan, I.R. Mansour, and A.I. Barboor, "Investigation of the effect of drift-field-dependent mobility on MOSFET characteristics," Parts I and II. *IEEE Trans. Electron Dev.* **ED-19**, August 1972, p. 899.

45. P.K. Ko, "Approaches to Scaling," Ch. 1, in *Advanced MOS Device Physics,* N.G. Einspruch and G. Gildenblatt, Eds. Academic Press, Orlando, FL., 1989, p. 1

46. K.Y. Toh, P.K. Ko, and R.G. Meyer, "An engineering model for short-channel MOS devices," *IEEE J. Solid-State Circuits,* August 1988, p. 950.

47. M.C. Jeng et al., "Performance and reliability of deep-submicron MOSFETs," *Tech. Dig. IEDM,* 1987, p. 710.

48. V.G. Reddi and C.T. Sah, "Source to Drain Resistance Beyond Pinchoff in Metal-Oxide-Semiconductor Transistor (MOST," *IEEE Trans. Electron Dev.*, **ED-12**, p. 108, (1969).

49. D. Frohman-Bentchkowsky and A.S. Grove, *IEEE Trans. Electron Devices* **ED-16**, p. 108, (1969).

50. G. Baum and H. Beneking, "Drift Velocity Saturation in MOS Transistors," *IEEE Trans. Electron Devices*, **ED-17**, June 1972, p. 481.

51. Y.A. Elmansy and A.R. Boothroyd, "A simple two-dimensional model for IGFET operation in the saturation region," *IEEE Trans. Electron Dev.*, **ED-24**(3), p. 254-62, (1977).

52. P.K. Ko, R.S. Muller, and C. Hu, "A Unified Model for Hot-Electron Currents in MOSFETs," *Tech. Dig. IEDM,* p. 600, 1981.

53. P.K. Ko, *Hot-electron effects in MOSFETs,"* Doctoral Thesis, University of California, Berkeley, (1982).

54. R.S. Muller and T.I Kamins, *Device Electronics for Integrated Circuits,* John Wiley & Sons, New York, 1986, p. 491.

55. T.Y. Chan, P.K. Ko, and C. Hu, "Dependence of Channel Electrical Field on Device Scaling," *IEEE Electron Dev. Lett.*, **EDL-6**, October, 1985, p. 531.

56. B. J. Sheu et al., "BSIM: Berkeley Short-Channel IGFET Model for MOS Transistors," *IEEE J. Solid-State Circuits,* Aug. 1987, p. 558.

57. B. J. Sheu, D. L. Scharfetter, and H. C. Poon, "Compact short-channel IGFET Model," Univ. of Calif., Berkeley, CA Memo. UCB/ERL M84/20, March 1984.

58. M.-C. Jeng, *Design and Modeling of Deep-Submicron MOSFETs,* ERL Memo No. ERL M90/90, Electronics Research Laboratory, University of California, Berkeley, October 1990.

59. A.H.-C. Fung, "A Subthreshold Conduction Model for BSIM," Univ. of Calif., Berkeley, Memo, UCB/ERL M65/22, March 1985.

60. B. Eitan and D. Frohman-Bentchkowsky, "Surface Conduction in short-channel MOS Devices as a limitation to VLSI scaling," *IEEE Trans. Electron Dev.* vol. ED-29, p. 254, Feb. 1982.

61. R. H. Dennard, et al., "Design of II MOSFETs with Very Small Physical Dimensions," *IEEE J. Solid-State Circuits,* **SC-9** (5), 256, (1974).

62. G. Bacarani, M. R. Wordeman, and R. H. Dennard, "Generalized Scaling Theory and its Application to a 0.25 μm MOSFET Design," *IEEE Trans. Electron Dev.*, **ED-31**(4), 452 (1984).

63. J.R. Brews et al., "Generalized Guide to MOSFET Miniaturization," *IEEE Electron Dev. Letts.*, **EDL-1**, January 1980, p. 2.

64. F.M. Klaassen, "Design and Performance of Micron-Size Devices," *Solid-State Electronics,* **21,** March 1978, p. 565.

65. J.R. Brews, K.K. Ng, and R.K. Watts, in "The Submicron Silicon MOSFET," Ch. 1, p. 19, *Submicron Integrated Circuits,* Ed. R.K. Watts, Wiley Interscience, New York, 1989.

66. *ibid.* pp. 23- 37.

67. R.A. Haken et al., "BiCMOS Process Technology," Chap. 3, p. 85, in *BiCMOS Technology and Applications,* A. Alvarez Ed. Klewer, Norwell MA, 1989.

68. A. Alvarez et al., "Application of Statistical Design and Response Surface Methods to Computer-Aided VLSI Device Design," *IEEE Trans. Computer-Aided Design,* **CAD-7**, February 1988, p. 272.

69. *ibid.* ref. 65, p. 24.

70. S.M. Sze, Ed. *VLSI Technology,* 2nd edition, chap. 11, McGraw-Hill, New York, 1988.

71. T. Ohzone et al., *IEEE Trans. Electron Dev.,* **ED-32**, 1789, (1980).

72. G.J. Hu et al., *Tech. Dig. IEDM*, 1982, p. 710.

73. K.M. Cham et al., *Computer-Aided Design and VLSI Device Development,* Boston, Klewer Academic Publishers, 1986, p. 182.

74. J. Zhu, R.A. Martin, and J.Y. Chen,, "Punchthrough Current for Submicrometer MOSFETs in CMOS VLSI," *IEEE Trans. Electron Dev.*, February 1988, p. 145.

75. A. Schmitz and J. Y. Chen, "Design, Modelling, and Fabrication of Submicron CMOS Transistors," *IEEE Trans. Electron Dev.*, **ED-33**, p. 148, (1986).

76. L.C. Parrillo, "CMOS Active and Field Device Fabrication," *Semicond. Internat.,* April 1988, p. 64.

77. R. A. Chapman et al., "A 0.8 μm CMOS Technology," *Tech. Dig. IEDM*, 1987, p. 362.

78. S. Odanaka et al., "A New Half-Micron P-Channel MOSFET with Efficient Punchthrough Stops," *1985 Symp.VLSI Tech., Digest of Tech. Papers*, p. 62.

79. T. Hori et al., "A New P-Channel MOSFET with Large Tilt Angle Implanted Punch-through Stopper (LATIPS)," *IEEE Electron Dev. Letts.,* **EDL-9,** Dec. 1988, p. 641.

80. M. Stinson et al., "Effects of Ion Implantation on Deep-Submicrometer Drain-engineered MOSFET Technologies," *IEEE Trans. Electron Dev.,* **ED-38,** March 1991, p. 487.

81. L. C. Parrillo et al., *Tech Dig. IEDM*, 1984, p. 418.

82. K. M. Cham et al., *IEEE Electron Dev. Lett.*, January 1986, p. 49.

83. C.Y. Wong et al., *Tech. Dig. IEDM*, 1988, p. 238.

84. S. J. Hillenius et al., "A Symmetric Submicron CMOS Technology," Tech. Dig. IEDM, 1986, p. 252.

85. P.-H. Pan, J.G. Ryan, and M.A. Lavoie, *Ext. Abs. of the Electrochem. Soc. Meeting*, Spring 1989, Abs. 138, p. 193.

86. M.L. Chen et al., "Constraints in *p*-Channel Device Engineering for Submicron CMOS Technologies," *Tech. Dig. IEDM*, 1988, p. 390.

87. F.K. Baker, et al., *Tech. Dig. IEDM*, 1989, p. 443.

88. J.M. Sung, *Tech. Dig. IEDM*, 1989, p. 447.

89. Y. Sato, K. Ehara, and K. Saito, *J. Electrochem. Soc.,* June 1989, p. 1777.

90. H. Fang et al., "Low-Temperature Furnace-Grown Reoxidized Nitrided Gate Dielectrics as a Barrier to Boron Penetration," *IEEE Electron Dev. Letts.,* **EDL-13** (4) p. 217 April 1992.

91. M. Nishida and H. Onodera, "An anomalous increase of threshold voltages with shortening of the channel lengths fro deeply boron-implanted *n*-channel MOSFETs," *IEEE Trans. Electron Dev.,* **ED-28**, p. 1101, 1981.

92. M. Orlowski, C. Mazure, and F. Lau, "Submicron short-channel effects due to gate re-oxidation induced lateral diffusion," *IEDM Tech. Dig.* 1987, p. 632.

93. C.Y. Lu and J.M. Sung, "Reverse Short-Channel Effects on Threshold Voltage in Submicron Salicide Devices," *IEEE Electron Dev. Lett.,* **EDL-10**, p. 446, October, 1989.

94. N.D. Arora and M.S. Sharma, "Modeling the Anomalous Threshold Voltage Behavior of Submicrometer MOSFETs," *IEEE Electron Dev. Lett.,* **EDL-13**, p. 92, February 1992.

95. H. Hanafi et al., "A Model for Anomalous Short-Channel Behavior in MOSFETs," *IEEE Electron Dev. Lett.,* **EDL-14**, p. 575, December 1993.

96. D. Sadana et al., "Enhanced Short-Channel Effects in NMOSFETs due to Boron Redistribution Induced by Arsenic Source and Drain Implant," *IEDM Tech. Dig.* 1992, p. 849.

97. H. Jacobs et al., "MOSFET Reverse Short Channel Effect Due to Silicon Interstitial Capture in Gate Oxide," *IEDM Tech. Dig.* 1993, p. 307.

98. C.S. Rafferty et al., "Explanation of Reverse Short Channel Effect by Defect Gradients," *IEDM Tech. Dig. 1993,* p. 311.

99. M.D. Giles, *Appl. Phys. Lett.,* **62(2)**, p. 1940, (1993).

100. J.Y. Chen, "An *n*-well CMOS with self-aligned channel stops," *Tech. Dig. IEDM* 1983, p. 526.

101. T.M. Liu and W.G. Oldham, *IEEE Electron Dev. Letts.,* **EDL-5**, 299 (1984).

102. J.R. Pfiester et al., "A Poly-Framed LDD Sub-half-Micrometer CMOS Technology," *IEEE Electron Dev. Letts.* **EDL-11,** November 1990, p. 529.

103. J.R. Pfiester et al., "An Integrated 0.5 μm CMOS Disposable TiN LDD/Salicide Spacer Technology," *Tech. Dig. IEDM,* 1989, p. 781.

104. L.C. Parrillo, "CMOS Active and Field Device Fabrication," *Semiconductor International,* April 1988, p. 64.

105. S. Meguro et al., *Tech. Dig. IEDM, 1984,* p. 59.

106. J.S.T. Huang, J.W. Schrankler, and J.S. Kueng, *IEEE Trans. Electron Dev.*, **ED-31**, 1889, (1984).

107. S. Ogura et al., *IEEE Trans. Electron Dev.,* **ED-27**, 1359 (1980).

108. Y. Omura and K. Ohwada, *Solid-State Electronics,* **24**, 301 (1981).

109. N. Kotani and S. Kawazu, *Solid-State Electronics,* **22**, p. 63, (1979).

110. T. Kuroi et al., "Novel NICE (Nitrogen Implantation into CMOS Gate Electrode and Source-Drain) Structure for High Reliability and High Performance 0.25 μm Dual Gate CMOS," *Tech. Dig. IEDM, 1993,* p. 325.

111. R.R. Troutman, *IEEE J. Solid-State Circuits*, **SC-9,** p. 55 (April 1974).

112. J.Y. Chen et al., "A Fully Recessed Field Isolation Technology Using Photo-CVD Oxide," *Tech. Dig. IEDM*, 1982, p. 233.

113. T. Shibata et al., "An optimally designed process for submicrometer MOSFETs," *IEEE Trans. Electron Dev.* **ED-29**, p. 531, (1982).

114. G.J. Hu and R.H. Bruce, "Design Tradeoffs Between Surface and Buried-Channel FETs," *IEEE Trans. Electron Dev.,* **ED-32**, March 1985, p. 584.

115. K.M. Cham and S.Y. Chiang, "Device design for the Submicrometer *p*-channel FET with n^+ polysilicon gate," *IEEE Trans. Electron Dev.,* **ED-31**, p. 964 (1984).

116. K.M. Cham et al., "Characterization and modeling of the trnch surface inversion problem for the trench isolated CMOS technology," in *IEDM Tech. Dig.,* p. 23, 1983.

117. T. Skotnicki, G. Merckel, and T. Pedron, "Analytical Study of Punchthrough in Buried-Channel PMOSFETs," *IEEE Trans. E.ectron Dev.*, April 1989, p. 690.

118. Y. Hiruta et al., "Impact of Hot Electron Trapping of Half-Micron PMOSFETs with p^+ Poly -Si Gate," *Tech. Dig. IEDM,* 1986, p. 718.

119. J. Cooper and D. Nelson, "High-Field Drift Velocity of Electrons at the Si-SiO_2 Interface as Determined by the Time-of-Flight Technique," *J. Appl. Phys.* **54**, 1445 (1983).

120. T.E. Hendrikson, "A Simplified Model for Subpinchoff Condition in Depletion Mode IGFETs," *IEEE Trans. Electron Dev.*, **ED-25**, p. 435 (1978).

PROBLEMS

1. Calculate the threshold voltage (a) in 1 Ω-cm *p*-type silicon and (b) in 1 Ω-cm *p*-type silicon. The MOS devices for each case have : (i) n^+-polysilicon gate; (ii) t_{ox} = 100 nm; and (iii) the oxide is free of charge except for a surface density of $Q_f = 5x10^{10}$ cm^{-2}.

2. For exercise **1**, what are the threshold voltages if p^+ polysilicon is used for both the *n*- and *p*-channel devices.

3. Describe punchthrough current and subthreshold current and explain the difference between them.

4. An *n*-channel MOSFET has a tox of 20 nm, an L_{eff} = 1 μm, r_j = 0.2 μm, and V_T = 0.6V. If the device is biased at V_{GS} = 3V and V_{DS} = 5V, calculate the saturation voltage V_{DSsat} and the maximum electric field.

5. Use one-dimensional junction theory to estimate the punchthrough voltage of a MOSFET with a channel length of 1 μm. Assume a substrate doping of $3x10^{16}/cm^3$, a substrate bias of 0V, and a uniformly doped substrate (no V_T-adjust implant).

6. Using Eq. 5-97, calculate the buried-channel S_t value for x_1 = 500Å, t_{ox} = 250Å, $x_2 << x_3$ = 3000Å and compare it with the S_t value calculated for a surface-channel MOSFET with the same t_{ox} and a channel depletion width of $d = x_3$.

7. How does a MOSFET's threshold change when constant field scaling rules are applied? Does it scale proportionately?

8. Consider a long-channel MOSFET with L = 3 μm, Z = 21 μm, $N_A = 5x10^{15}$ cm^{-3}, $C_{ox} = 1.5x10^{-7}$ F/cm^2, and V_T = 1.5V. Calculate V_{DSsat} for the case when V_{GS} = 4V. If a constant scaling factor is used to reduce the channel length to 1 μm, find the following scaled-down parameters: Z, C_{ox}, and I_{Dsat}.

9. A MOSFET with gate oxide thickness of 20 nm is to be fabricated on a *p*-type substrate of doping $N_A = 10^{16}$ cm^{-3}. The source/drain regions have a depth of 0.2 μm and $N_D = 10^{19}$ cm^{-3}. The device is to be operated with V_{BS} = 0V, but V_{DS} could be as large as 3V. From Eq. 5-84 estimate the minimum channel length this MOSFET can have while still exhibiting electrically long-channel behavior. (The gate length is the channel length plus twice the junction depth.) A smaller value of minimum channel length could be obtained if a larger N_A was used. Suggest some undesirable side effects that such a choice could produce.

10. Qualitatively describe the differences between a surface-channel and a buried-channel MOSFET.

11. We have stated in this chapter that the saturation of electron velocity can limit the drain current in a MOSFET with a channel length of 1 μm or less operating from a 5-V supply. Examine this effect by using SPICE to generate the I_D-V_{DS} characteristic of an NMOSFET with L = 1 μm formed on a substrate of doping $N_A = 10^{16}$ cm^{-3} with a 50 nm thick gate oxide. Set V_{GS} = 5V and sweep V_{DS} from 0 to 5V using the Level 2 model. First run a simulation in which VMAX is not specified, and then repeat the simulation with VMAX = $8x10^6$ cm sec^{-1}, which corresponds roughly to the scatter-limited velocity of electrons in a MOSFET channel.

12. Qualitatively summarize why a surface channel NMOSFET is prone to suffering punchthrough only after the punchthrough voltage is exceeded, but that a buried-channel PMOSFET exhibits punchthrough current behavior even when drain voltages are very small.

13. Why does the subthreshold slope of the I_D-V_{GS} curve appear to increase at the onset of subsurface punchthrough?

CHAPTER 6

ISOLATION STRUCTURES IN SUBMICRON CMOS ICs

Implementing electric circuits involves connecting isolated devices through specific conducting paths. Thus, to fabricate monolithic-silicon ICs, devices isolated from one another must first be created in the silicon substrate. Only later are they interconnected to form the desired circuit function. Isolation of devices in the substrate of an IC is also important for other reasons. For example, the state (*ON* or *OFF*) and conductance of individual MOSFETs can only be controlled if proper isolation exists among devices. If not, leakage currents may occur, causing dc power dissipation, noise-margin degradation, and voltage shift on dynamic nodes. Crosstalk among transistors can also destroy the logic state of a gate. In CMOS circuits, leakage current in the isolation region can escalate latchup. From these perspectives, it is evident that device isolation technology is one of the critical aspects of IC fabrication.

Establishing effective isolation in submicron ICs in the face of increased transistor counts and decreased isolation space is a challenging task. In ULSI chips, a tiny amount of leakage per device can induce significant power dissipation for the entire circuit. As an example, if a chip has ten million transistors, and each device exhibits leakage current in the 10-20 nA range, the total power dissipation for the entire chip is increased by several tenths of a watt. Ensuring that isolation leakage currents in submicron ICs are limited to much smaller values (e.g., $\leq$10 pA/μm) is major challenge.

Besides negligibly small leakage currents, the other desired characteristics of IC isolation schemes include: (1) the spacing between active areas should be as small as possible; (2) the process of fabricating the isolation structure should not adversely impact the characteristics of the active devices; (3) the isolation structure should produce as planar a surface topography as possible; and (4) the overall process sequence used to fabricate the isolation structures should be simple to implement and control.

A variety of techniques have been developed to isolate devices in integrated circuits. One reason for such diversity is that different IC types (e.g., NMOS, CMOS, and bipolar) have somewhat different isolation requirements. Furthermore, the various isolation technologies exhibit differing characteristics with respect to minimum isolation spacing, surface planarity, process complexity, and density of defects generated during fabrication of the isolation structure. Tradeoffs can be made among these attributes when choosing an isolation technology for a particular circuit application. A

survey of the evolution of isolation technologies from the early 1960's down to those used in 1.0 μm IC generations is provided in Vol. 2, chap. 2. In this chapter we limit the discussion to the chief isolation technologies adopted for submicron CMOS and BiCMOS ICs (1.0 μm to 0.5 μm). The problem of *latchup* (which represents a form of isolation breakdown between the power and ground rails in CMOS and BiCMOS ICs) is not discussed here, but is covered instead in Vol. 2, chap. 6.

We begin with a description of how isolation in MOS ICs is characterized, including some test structures used for this purpose. The less difficult problem of isolation among devices having identical polarity (i.e., among devices which are exclusively PMOS or NMOS) is treated first. The additional problems arising when isolation must be established among devices having different polarities - e.g., between PMOS and NMOS devices - are considered later as part of the CMOS isolation discussion.

Next, the basic *LOCOS isolation* technique (for LOCal Oxidation of Silicon) is reviewed, including the reasons why basic LOCOS becomes inadequate for submicron MOSFETs. This sets the stage for the discussion of the two most widely used isolation structures used to replace basic LOCOS: (1) *polybuffered LOCOS*; and (2) *shallow-trench and refill isolation*. These are examined in depth. Although *deep-trench isolation* is also important, it is more commonly used in bipolar ICs, and thus the discussion of this approach (as well as that of silicon-on-insulator [SOI] isolation techniques), is left to Vol. 2, chap. 2.

6.1 CHARACTERIZING MOS ISOLATION

MOS transistors are self-isolated as long as: (1) the source-substrate and drain-substrate *pn* junctions are held at reverse bias, and (2) spurious channels are prevented from forming among adjacent devices. Since the reverse-bias current of properly-fabricated Si *pn* junctions at 300K is negligibly small, the component of drain current due to reverse-bias drain-substrate leakage should also be insignificantly small. In addition, in the *OFF* state, very little drain-source current should flow (it is also assumed that such mechanisms as punchthrough and GIDL do not produce any significant leakage). Thus, only if the device is turned *ON* by the establishment of a channel under the gate should a significant drain current exist. Similarly, leakage currents between *adjacent* MOS devices should also be negligible provided spurious channels are prevented from forming between them (Fig. 6-1).* *The primary task of isolation in MOS ICs is therefore to ensure that spurious channels do not arise among the MOSFETs in the IC.*

Having stated that spurious channels between MOS devices must be prevented, let us first identify the source of their origin - namely, parasitic MOS structures. Such parasitic devices arise in planar ICs because wiring paths are established with metal or

* Such MOS self-isolation also implies that the buried *pn* junction needed to isolate bipolar transistors is unnecessary in MOS ICs. This represents a substantial area savings for NMOS (and PMOS) circuits as compared to junction-isolated bipolar circuits, allowing much higher density MOS circuits to be fabricated.

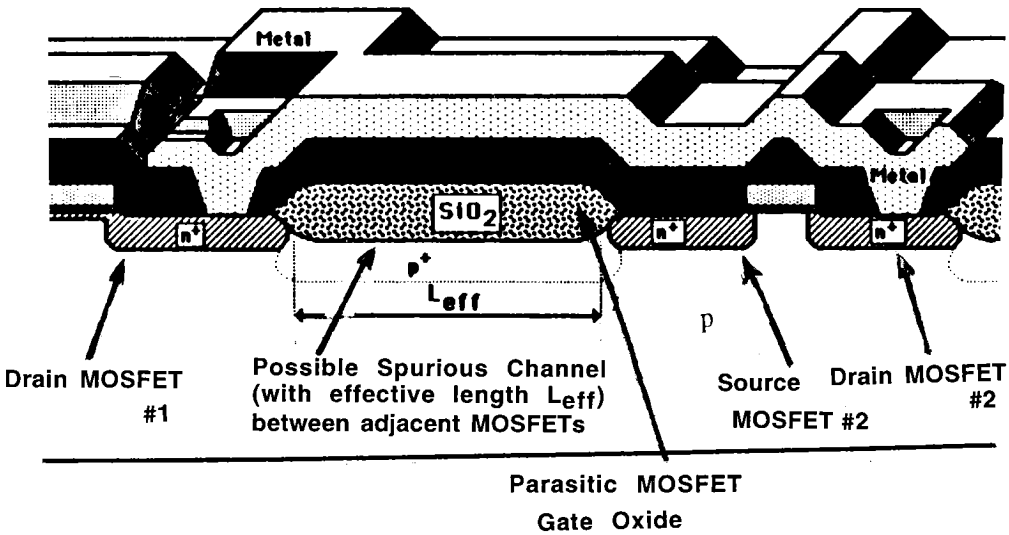


Fig. 6-1 No current should flow between source and drain regions of adjacent MOSFETs. This figure shows the parasitic field transistor with a possible channel under the field oxide if the substrate under this oxide is inverted. From W. Maly, *Atlas of IC Technologies,* Copyright 1987 by the Benjamin/Cummings Publishing Co. Reprinted with permission.

poly-silicon lines adherent to the oxide in the field regions (Fig. 6-1). These conducting lines behave as the gate electrodes of the parasitic MOSFETs, with the field oxide beneath them acting as the gate oxide, and the diffused regions (Drain MOSFET #1 and Source MOSFET #2) in Fig. 6-1 serving as the parasitic MOSFET source and drain, respectively. The channel region of such parasitic devices exists between the source and drain as is also shown in Fig. 6-1. If a voltage exceeding the threshold voltage of the parasitic MOSFET is applied to these conducting lines, inversion of this parasitic channel region occurs, establishing a leakage current path between adjacent devices. Consequently, to avoid the loss of isolation among neighboring MOS devices, the threshold voltage of such parasitic field transistors (V_{TF}) must be greater than any possible operating voltage.

Two methods for increasing V_{TF} are: (1) making the field-oxide thicker, and; (2) increasing the doping beneath the field oxide. V_{TF} could be made adequately high by forming a sufficiently thick oxide on the wafer surface, but the large resulting oxide steps would give rise to severe step-coverage problems.(Note that thick field oxides were used as the isolation structures in early MOSFETs. Such grow-field-oxide and-etch structures are illustrated in Figs. 6-6b and 6-40a.)Thus, practical approaches generally combine a more moderate increase in field oxide thickness with an increase in doping under the field oxide.

The field oxide is nevertheless much thicker than the gate oxide in the active regions. The thicker oxide (typically 0.5 to 1.0 μm thick) reduces the parasitic capacitance between the interconnect runners and the substrate, which also improves the speed characteristics of the circuits. Normally, ion implantation is used to increase the doping

combination of thick oxide and channel-stop implant can provide adequate isolation for PMOS and NMOS (and for oxide-isolated bipolar) ICs. Additional isolation issues, however, arise in CMOS circuits, as will be discussed in later sections.

As mentioned in the introduction, the maximum allowable leakage current between neighboring devices is extremely small in many IC applications (e.g., ≤10 pA/μm). In practice, the *nominal* V_{TF} needs to be 8-10V above the supply voltage (V_{DD}) to ensure that less than 1 pA/μm of current flows between isolated MOS devices when V_{DD} is applied to the parasitic gate electrode.[1] This is because the *nominal* constant-current-V_{TF} is typically specified in the same way as the constant-current-V_T of active MOSFETs (i.e., nominal V_{TF} is typically defined as the gate voltage that must be applied to cause a drain current of 1 μA per micron of device width [I_D = 1 μA/μm] in the parasitic MOSFET). Hence, for 5V circuit operation, *nominal* V_{TF} values usually exceed 15V, at least an order of magnitude higher than V_T's of active MOSFETs.

Operationally, parasitic field-oxide MOSFETs must exhibit such large *nominal* V_{TF} values for several reasons. First, allowance must be made for supply voltage fluctuations (normally 10% of the specified value), as well as for possible conditions of overshoot on interconnect lines. Second, as the operating temperature of a MOSFET rises, the threshold voltage decreases (mainly due to changes in φ_B and φ_{ms}). Reductions of as much as 2V in V_{TF} values have been observed as the temperature is increased from 25°C to 125°C.[2] Finally, the thick field-oxide and relatively heavy channel-stop doping result in (parasitic) devices with very poor subthreshold swing S_t, and S_t increases even more as T is raised (e.g., by ~40% at T = 100°C - see section 4.71 for a background discussion of S_t). Because the S_t effects have the largest impact on isolation leakage (and hence, on the nominal V_{TF} required), we examine them more thoroughly for the case of parasitic field MOSFETs.

In an *ideal* MOSFET, I_D exhibits an abrupt drop for small decreases in V_{GS} below V_T, implying that S_t in such devices is also very small. Therefore, in an *actual* MOSFET, the smaller the S_t value, the more closely it approaches ideal behavior. Now, as discussed in chap. 4, section 4.7.1, if a MOSFET exhibits an S_t value of ~0.1 V/dec - typical of S_t an ordinary active MOSFET - the value of I_{Dst} will drop from 1 μA/μm to 1 pA/μm for a 0.6V decrease in V_{GS} below V_T. But, the S_t values of parasitic field-MOSFETs is much larger (in the range of 0.5-1.5 V/dec), due to their much greater gate-oxide thickness. (Equation 4-91 indicates that S_t will increase if t_{ox} is increased.) Hence, a significant current still flows in the parasitic MOSFET if V_{GS} is decreased by only 0.6 V below the nominal V_{TF}. Thus, to keep the isolation leakage current acceptably small (e.g. ≤10 pA/μm) when $V_{GS} = V_{DD}$ = 5V, it is necessary to ensure that the nominal V_{TF} is least 5 or 6 V above the worst-case parasitic gate bias. In fact, the minimum nominal V_{TF} of 15V is arrived at by assuming a maximum operating temperature of 100°C, and then allowing for the 2V drop in nominal V_{TF} at the elevated temperature, and for the increase in S_t with increasing T (from 0.6V/dec at 25°C to ~1.0V/dec at 100°C). With such a nominal V_{TF}, I_D will remain <10 pA/μm when V_{DD} = 5V at 100°C.

If the circuits are to be operated in a radiation environment, an additional threshold-voltage margin must be allowed. That is, radiation-induced oxide charges will produce a

threshold-voltage shift, and the shift will be proportional to the gate oxide thickness to a power of between 2 and 3. For such rad-hard circuits, the nominal V_{TF}s of the parasitic field-oxide MOSFETs may thus be specified at ~20V.

6.1.1 Test Structures for NMOS Isolation Structures

As noted in the previous section, the traditional method for determining the *nominal* V_{TF} value has been to measure V_{GS} when $I_D = 1\ \mu A$ per μm of device width, and then call this the nominal V_{TF}. However, since 1 $\mu A/\mu m$ would generally represent unacceptably high leakage in parasitic field-oxide MOS transistors, a more appropriate way to characterize isolation structures would be to measure the gate voltage V_{TFiso} at much smaller drain currents (e.g. at 1 pA-1 nA) *and* at drain voltage equal to the power supply voltage, e.g., at $V_{DD} = 5V$, and this approach is now being used more frequently. (V_{TFiso} is measured at $V_{DS} = V_{DD}$ because this represents the worst-case DIBL-induced-leakage condition in a short-channel device.) Furthermore, when it is necessary to fully characterize the efficacy of isolation in a PMOS or NMOS process, V_{TFiso} versus n^+ to n^+ (or p^+ to p^+) spacing is typically measured over some range of drain current values. From such data it can be determined whether V_{TF} exceeds V_{DD} at the maximum acceptable leakage current (e.g., 1 or 10 pA/μm) and at the minimum isolation spacing. This section describes the isolation test structures used to obtain such data. The gate voltage at which the maximum-allowed leakage current arises in the isolation test structure is defined as V_{TFiso} for that value of I_{Dst}.

Figure 6-2 shows a layout (top-view) and cross-sectional view of test structures used to measure the leakage current between adjacent NMOS devices.[1] Figs. 6-2a and 6-2b depict an IC structure with an aluminum interconnect line running between two active *n*-channel MOSFETs, and Figs. 6-2c and 6-2d show the case of a polysilicon line running between two active *n*-channel MOSFETs. This allows testing of both types of field transistor types: those in which the aluminum and line acts as a gate on the thick field oxide, and those in which polysilicon acts as the gate (the two n^+ diffusions adjacent to the respective gate film act as source and drain, forming the parasitic FET). Since the source and drain in both of these types of parasitic FETs are identical, they are referred to as symmetric parasitic FETs. In Fig. 6-3 we depict two layouts: in Fig. 6-3a two poly-gate FETs with a poly interconnect line running between them are shown, and in Fig. 6-3b with a poly line serving as both the gate and the common interconnect line between them. Both conditions occur in actual ICs, and the test structure of Figs. 6-2c and d is used to characterize the isolation behavior of both of these layouts.

Notice a subtle but important difference between the two test structures of Fig. 6-2. In the latter structure (the poly gate FET) the n^+ regions adjacent to the polysilicon gate do not abut the field oxide. This exact structure may not exist in an actual IC layout because polysilicon lines normally do not extend over the active (thin oxide) area (but instead are layed out as shown in Fig. 6-3a) unless they are also used as the gate electrode of active transistors (see Fig. 6-3b, in which the polysilicon line serves as a gate electrode as well as the interconnect between the two electrodes). In Figs. 6-2c and d the silicon surface under the thin oxide regions is normally inverted at a very low gate voltage (<1V), resulting in ultra shallow "n^+"

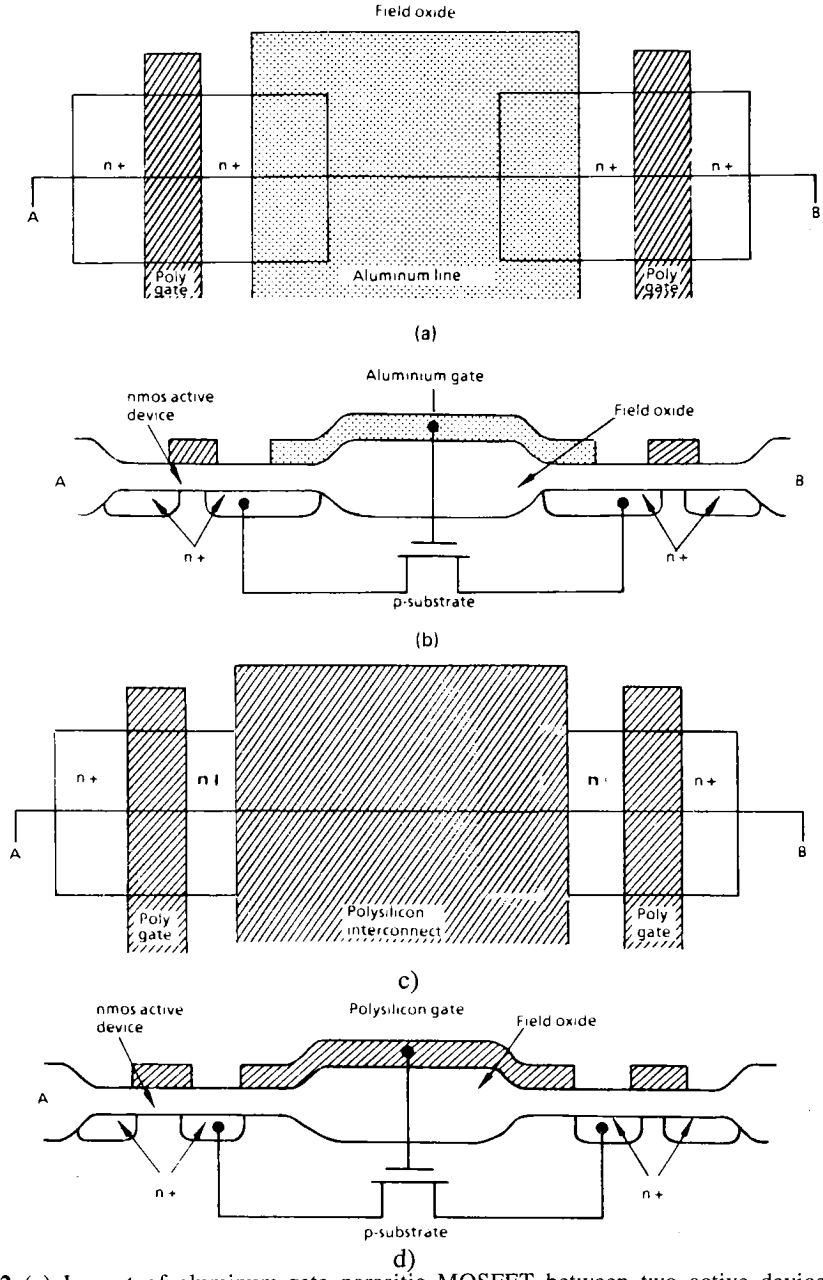


Fig. 6-2 (a) Layout of aluminum-gate parasitic MOSFET between two active devices in NMOS technology; (b) Cross-sectional view of (a); (c) Layout of polysilicon-gate parasitic MOSFET between two active devices in NMOS technology; (d) Cross-sectional view of (c).[1] From J.Y. Chen, *CMOS Devices and Technology for VLSI*, p. 235-6, Copyright Prentice-Hall, 1990. Reprinted with permission.

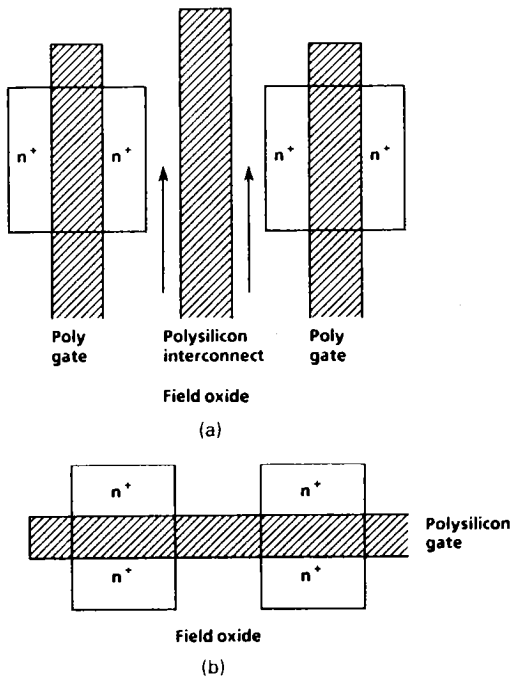


Fig. 6-3 (a) Layout of polysilicon lines as transistor gate in real NMOS IC; (b) Layout of interconnect in real NMOS ICs.[1] From J.Y. Chen, *CMOS Devices and Technology for VLSI*, p. 237, Copyright Prentice-Hall, 1990. Reprinted with permission.

regions, which then abut the field oxide. If the silicon surface under the field oxide is then also inverted, the two transistors are basically shorted and isolation between the two transistors is completely broken down. Thus, the test structure shown in Figs. 6-2c and 6-2d is a good test vehicle for the case shown in Fig. 6-3b, because leakage current measured from the parasitic FET indicates the onset of surface inversion under the field oxide.

The test structure of Figs. 6-2c and d also represents a stringent test structure for the layout shown in Fig. 6-3a (which is often seen in a real circuits). That is, since polysilicon in the test structure completely overlaps the field oxide, this structure tests the worst-case condition of the layout of Fig. 6-3a, in which the poly does not completely overlap the field oxide. (That is, it is harder to cause inversion beneath regions of field oxide that are not covered with a poly line, namely the areas pinpointed by arrows in Fig. 6-3a). The configuration in Fig. 6-3a is thus unlikely to exhibit isolation problems if the test structure of Figs. 6-2c and d passes the isolation test.

Another difference between the aluminum and polysilicon gate parasitic FETs is that there is a thicker oxide in the former devices due to an additional oxide layer deposited to

separate polysilicon and aluminum interconnects. The total oxide under the aluminum is typically 1-2 μm, while in the polysilicon device the gate oxide is composed of thermally grown oxide only, which is about 0.5 - 1.0 μm thick. In both cases, the thick field oxide and the doped silicon surface underneath inhibit surface inversion unless the gate voltage is raised several volts above the 5V supply.

Figure 6-4 shows the V_{TFiso} values measured at drain currents of 10 pA, 1 nA, and 0.1 μA in a 20 μm wide *n*-channel FET, using the test structures shown in Figs. 6-2a & b, and 6-2c & d.[3] The LOCOS field-oxide thicknesses for polysilicon- and aluminum-gate FETs are about 0.5 and 1.0 μm, respectively in this example. A boron field implant of 4×10^{13} cm^{-2} at 25 keV is used for the channel-stop formation before the field oxide growth. The higher current curves (0.1-1.0 µA) corresponds to a threshold similar to the threshold defined at the onset of strong inversion. As noted earlier, this definition is suitable for the active MOS transistors but not particularly useful for

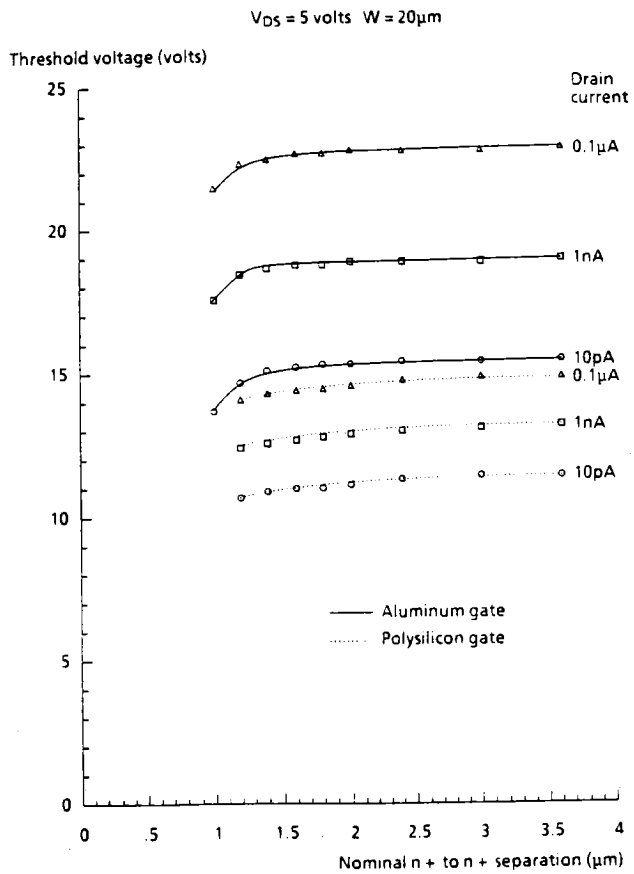


Fig. 6-4 *n*-Channel field threshold voltages V_{TFiso} for symmetrical parasitic MOSFETs between two active devices.[1] From J.Y. Chen, *CMOS Devices and Technology for VLSI*, p. 253, Copyright Prentice-Hall, 1990. Reprinted with permission.

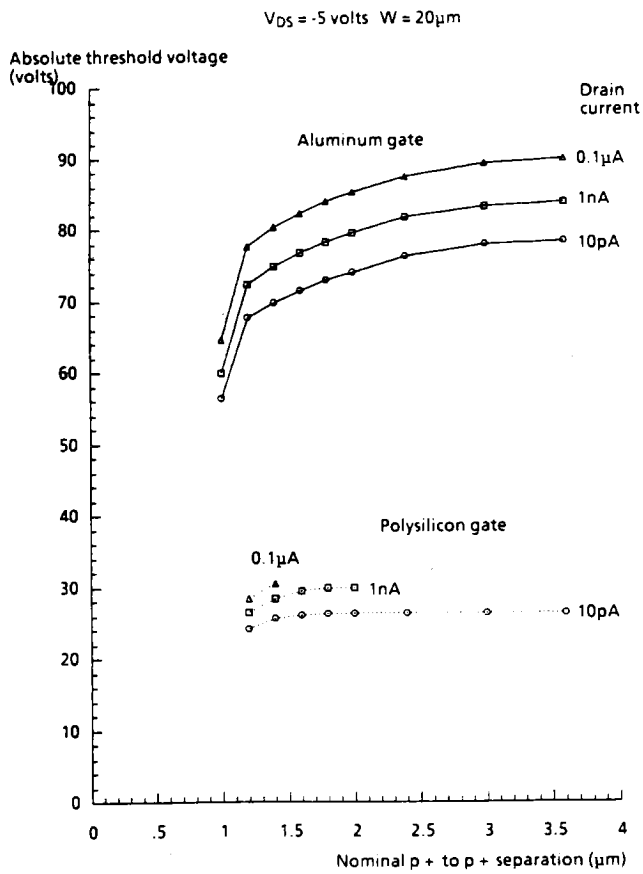


Fig. 6-5 *p*-Channel field threshold voltages V_{TFiso} for symmetrical parasitic MOSFETs between active devices.[1] From J.Y. Chen, *CMOS Devices and Technology for VLSI*, p. 254, Copyright Prentice-Hall, 1990. Reprinted with permission.

parasitic FETS in which only very low level of leakage currents (10 pA - 1 nA) can be tolerated.

Notice that the V_{TFiso}s are higher for the aluminum gate because of the associated thicker field oxide. The thicker oxide also gives poorer subthreshold slope, causing V_{TFiso} curves to separate further. The V_{TFiso} decreases as the channel length (n^+-to-n^+ separation) is reduced in a way that is analogous to the short-channel effect commonly observed in active MOSFETs. The effect is less significant in the polysilicon gate structure because the short-channel effect is smaller in device with thinner gate oxides and shallower S/D junctions. (The S/D junction in the poly-gate parasitic structure is the "n^+" [inversion layer charge]/substrate junction.) This effect should be taken into account when setting n^+-to-n^+ layout design rules. For the example shown in Fig. 6-4, V_{TFiso} values are 10 V or above even at 1 μm separation between two *n*-channel active

devices. Leakage is also possible in the parasitic FETs if punchthrough between n^+ regions occurs. Due to the relatively high doping concentration in the field region, punchthrough usually does not take place unless the n^+-to-n^+ separation is scaled below 1 μm.

Parasitic FETs among *p*-channel active devices are identical to those shown in Figs. 6-2a, 6-2b, 6-2c, and 6-2d except that the doping types are reversed. Symmetric parasitic PMOSFETs are, in general, harder to turn on because the net oxide charge density is commonly positive, making the V_{TF} even more negative. As shown in Fig. 6-5, the absolute V_{TFiso} values are well above 5 V.[3] The test devices used here are fabricated from the same CMOS process which made the devices used for obtaining the data in Fig. 6-4. The channel stop is formed by a shallow phosphorus implant inside the *n*-well. The very high V_{TFiso} values in the polysilicon gate devices are not always measurable due to dielectric breakdown in the thin oxide region between the polysilicon gate and the source. Notice that the aluminum gate devices exhibit greater short-channel V_T roll-off effects, due to the thicker oxide. But, in summary, isolation among *p*-channel active devices is normally not a problem.

6.1.2 Punchthrough Prevention between Adjacent Devices in MOS Circuits

Parasitic conduction between adjacent devices due to punchthrough (Fig. 6-1) must also be prevented (see chap. 5 for additional information on punchthrough). In MOS circuits, the source and drain regions of each transistor must be kept far enough from the source and drain regions of any neighboring devices so that the depletion regions do not merge (i.e., this distance must be roughly greater than twice the maximum depletion-region width). Substrate doping and channel-stop doping concentrations must also be considered, because lighter substrate doping allows wider depletion-region widths, and punchthrough is more likely to occur at lower voltages in lightly doped regions. It is reported that at isolation spacings of <0.8 μm even advanced LOCOS processes begin to exhibit excessive leakage due to punchthrough.[4]

6.2 THE BASIC LOCOS ISOLATION PROCESS

The most widely used method for creating isolation in NMOS and PMOS ICs is the basic LOCOS structure (*LOC*al *O*xidation of *S*ilicon), introduced by Appels and Kooi in 1970.[5,21] (Note that Kooi has also recently written a book *The Invention of LOCOS* giving a first hand description of the development of LOCOS.[6]) In the LOCOS approach, the oxide is selectively grown over the field regions of the IC. This is done by covering the active regions with a thin layer of silicon nitride. When the field oxide is grown, the active regions remain covered by nitride, which prevents oxidation of the silicon beneath. In the field regions the surface of the Si substrate is exposed prior to field oxidation by etching away the nitride layer in these areas. In addition, the silicon in the field regions is also selectively implanted at this point with the channel-stop dopant. Thus, the channel-stop regions become self-aligned to the field oxide.

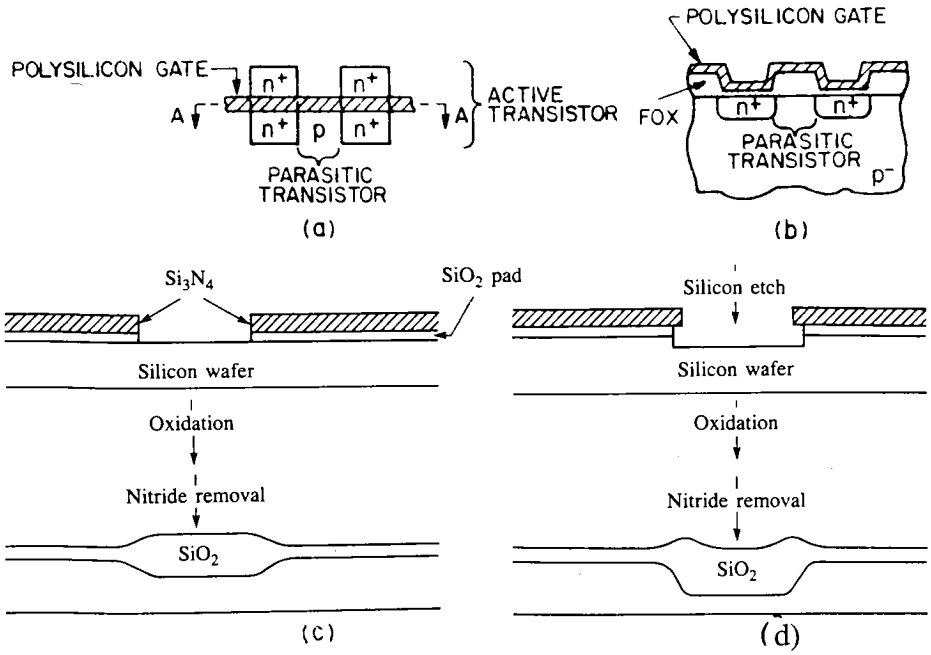


Fig. 6-6 MOSFET isolation. (a) Top view of adjacent MOS transistors with common polysilicon gate illustrating active and parasitic transistor conduction paths. Cross section through A - A for: (b) Grow-oxide-and-etch isolation structure; (c) Semirecessed LOCOS isolation structure; and (d) Fully recessed LOCOS structure.

If the field oxide is selectively grown without etching the silicon (Fig. 6-6c) the resulting field oxide will be *partially (or semi) recessed.* On the other hand, if the silicon is etched (to a depth of about half the desired field oxide thickness) after the oxide-preventing layer is patterned (Fig. 6-6d), the field oxide can be grown until it forms a planar surface with the silicon substrate. (Note that this occurs because the growing oxide film is about twice as thick as the silicon layer it consumes; see Vol. 1, chap. 6). This is known as a *fully recessed isolation oxide process.* In the semirecessed process, the height of the oxide protruding above the level of the active region surface is larger than in the fully recessed process, but it is smaller than in the grow-oxide-and-etch process. In addition, the semirecessed oxide step has a gentle slope that is more easily covered by subsequent polysilicon and metal layers.

The semirecessed LOCOS structure has become the workhorse isolation technology for MOS devices down to about 1.5-μm geometries. The advantages of semirecessed LOCOS over fully recessed LOCOS are that fabrication is less complex and fewer process-induced defects are created in the silicon substrate. Fully recessed LOCOS is primarily used for isolating digital bipolar ICs, where a thicker field oxide is needed than in MOS. A semirecessed oxide with such a larger thickness would produce intolerably high oxide steps, while the use of the fully recessed approach mitigates this problem. Details describing fully recessed LOCOS are given in Vol. 2, chap. 2.

6.2.1 Details of the Basic Semi-recessed LOCOS Process

Details of each of the steps used to fabricate the basic, conventional semirecessed LOCOS structures are presented in this section (Fig. 6-7).

6.2.1.1 Pad-Oxide Layer. A wafer with a bare silicon surface is cleaned, and a 40-50 nm SiO_2 layer (called a *pad* or *buffer* oxide) is thermally grown on the surface. Its function is to cushion the transition of stresses between the silicon substrate and the subsequently deposited nitride. For example, assume that a nitride film is deposited directly on a silicon wafer. When the field oxide is subsequently grown, stresses occur in the nitride due to various effects, including: (1) mismatch between the thermal coefficient expansion of the nitride film and the silicon substrate; (2) a tendency of the

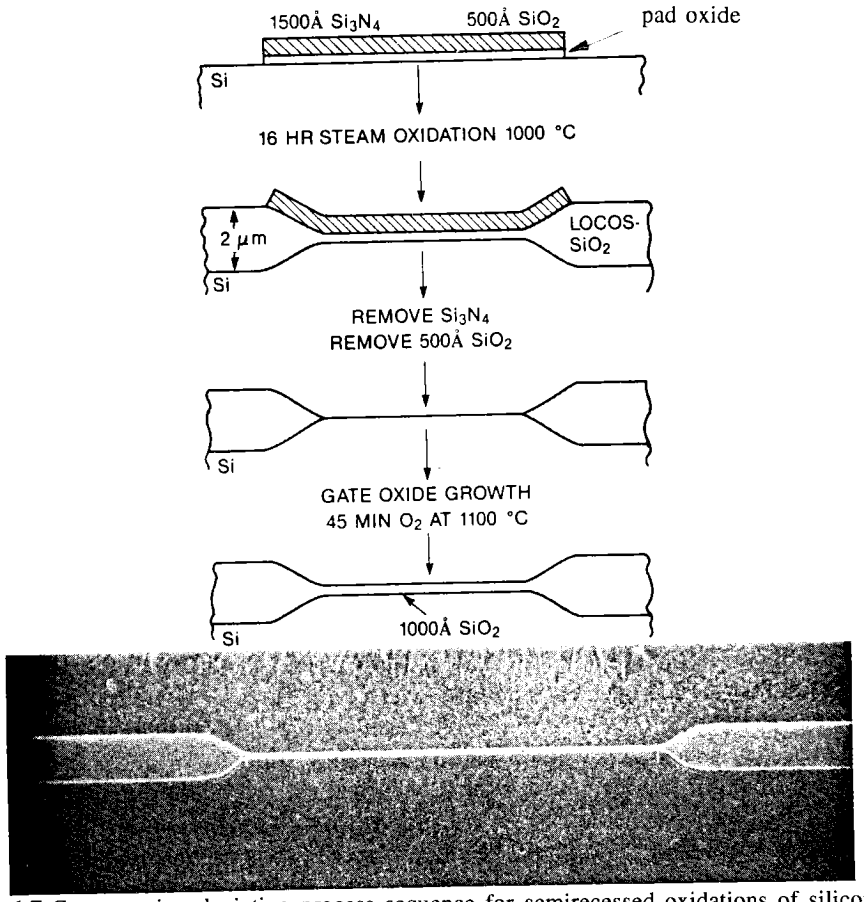


Fig. 6-7 Cross section depicting process sequence for semirecessed oxidations of silicon.[21] This paper was originally presented at the Spring 1973 Meeting of the Electrochemical Society. Reprinted with permission.