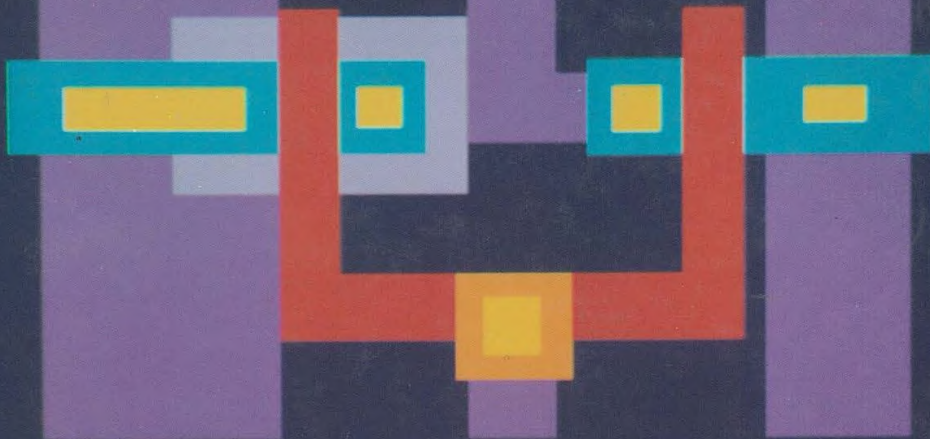




Silicon Processing

for the VLSI Era
Volume 2 - Process Integration



S. Wolf

Silicon Processing

for the VLSI Era

Volume 2
Process
Integration

S. Wolf

TK
7874
.W635
2000

THE COMPANION VOLUME TO THIS BOOK

Silicon Processing

for the VLSI Era

Volume 1 – Process Technology

By STANLEY WOLF and RICHARD N. TAUBER

"Arguably the best available text on the subject"
IEEE Transactions on Education

"The authors of this volume, have succeeded admirably. The book is a uniquely valuable reference for the professional microelectronic engineer.... Volume 2 of this text will be eagerly awaited."
Semiconductor Science and Technology

"Well organized, and presented in a logical and clear fashion. Lucid writing style, with abundant and well-placed illustrations... Overall, an excellent book on silicon processing... I highly recommend it!"
IEEE Circuits and Devices Magazine

TABLE of CONTENTS - Ch. 1 Silicon: Single-Crystal Growth and Wafering; Ch. 2 Crystalline Defects, Thermal Processing; and Gettering; Ch. 3 Vacuum Technology for VLSI Applications; Ch. 4 Basics of Thin Films; Ch. 5 Silicon Epitaxial Film Growth; Ch. 6 Chemical Vapor Deposition of Amorphous and Polycrystalline Films; Ch. 7 Thermal Oxidation of Single-Crystal Silicon; Ch. 8 Diffusion in Silicon; Ch. 9 Ion Implantation for VLSI; Ch. 10 Aluminum Thin Films and Physical Vapor Deposition in VLSI; Ch. 11 Refractory Metals and Their Silicides in VLSI; Ch. 12 Lithography I: Optical Photoresists-Material Properties and Process Technology; Ch. 13 Lithography II: Optical Aligners and Photomasks; Ch. 14 Advanced Lithography: E-Beam and X-Ray; Ch. 15 Wet Processing: Cleaning, Etching and Lift-Off; Ch. 16 Dry Etching for VLSI; Ch. 17 Material Characterization Techniques for VLSI Fabrication; Ch. 18 Structured Approach to Design of Experiments for Process Optimization.

1986 LATTICE PRESS 660 pp. ISBN 0-9616721-3-7

Order using form on last page of this book

SILICON PROCESSING FOR T \$45.00



220002773500 1111 1111 1111

CONTENTS

PREFACE

CHAP. 1 - PROCESS INTEGRATION FOR VLSI AND ULSI 1

1.1 PROCESS INTEGRATION 5

- 1.1.1 Process Sequence Used to Fabricate an Integrated-Circuit MOS Capacitor, 5
- 1.1.2 Specifying a Process Sequence, 6
- 1.1.3 Levels of Process Integration Tasks, 7

1.2 PROCESS-DEVELOPMENT AND PROCESS-INTEGRATION ISSUES 8

REFERENCES 11

CHAP. 2 - ISOLATION TECHNOLOGIES FOR INTEGRATED CIRCUITS 12

2.1 BASIC ISOLATION PROCESSES FOR BIPOLAR ICs 13

- 2.1.1 Junction Isolation, 13
 - 2.1.1.1 *Junction Isolation in the SBC Process*
 - 2.1.1.2 *Collector-Diffusion Isolation*

2.2 BASIC ISOLATION PROCESS FOR MOS ICs (LOCOS ISOLATION) 17

- 2.2.1 Punchthrough Prevention between Adjacent Devices in MOS Circuits, 20
- 2.2.2 Details of the Semirecessed Oxide LOCOS Process, 20
 - 2.2.2.1 *Pad-Oxide Layer.*
 - 2.2.2.2 *CVD of Silicon Nitride Layer.*
 - 2.2.2.3 *Mask and Etch Pad-Oxide/Nitride Layer to Define Active Regions.*

2.2.2.4	<i>Channel-Stop Implant.</i>	
2.2.2.5	<i>Problems Arising from the Channel-Stop Implants.</i>	
2.2.2.6	<i>Grow Field Oxide.</i>	
2.2.2.7	<i>Strip the Masking Nitride/Pad-Oxide Layer.</i>	
2.2.2.8	<i>Regrow Sacrificial Pad Oxide and Strip (Kooi Effect).</i>	
2.2.3	Limitations of Conventional Semi-Recessed Oxide LOCOS for Small-Geometry ICs,	27
2.3	FULLY RECESSED OXIDE LOCOS PROCESSES	28
2.3.1	Modeling the LOCOS Process,	31
2.4	ADVANCED SEMIRECESSED OXIDE LOCOS ISOLATION PROCESSES	31
2.4.1	Etched-Back LOCOS,	31
2.4.2	Polybuffered LOCOS,	32
2.4.3	SILO (Sealed-Interface Local Oxidation),	33
2.4.4	Laterally Sealed LOCOS Isolation,	35
2.4.5	Bird's Beak Suppression in LOCOS by Mask-Stack Engineering,	38
2.4.6	Planarized SILO with High-Energy Channel-Stop Implant,	38
2.5	ADVANCED FULLY RECESSED OXIDE LOCOS ISOLATION PROCESSES	39
2.5.1	SWAMI (Sidewall-Masked Isolation Technique),	39
2.5.2	SPOT (Self-Aligned Planar-Oxidation Technology),	41
2.5.3	FUOX (Fully Recessed Oxide),	41
2.5.4	OSELO II,	43
2.6	NON-LOCOS ISOLATION TECHNOLOGIES I: (TRENCH ETCH AND REFILL)	45
2.6.1	Shallow Trench and Refill Isolation,	45
2.6.1.1	<i>BOX Isolation.</i>	
2.6.1.2	<i>Modifications to Improve BOX Isolation.</i>	
2.6.2	Moderate-Depth Trench and Refill Isolation,	48
2.6.2.1	<i>U-Groove Isolation.</i>	
2.6.2.2	<i>Toshiba Moderate-Depth Trench Isolation for CMOS.</i>	
2.6.3	Deep, Narrow Trench and Refill,	51
2.6.3.1	<i>Reactive Ion Etching of the Substrate.</i>	
2.6.3.2	<i>Refilling the Trench.</i>	
2.6.3.3	<i>Planarization after Refill.</i>	

**2.7 NON-LOCOS ISOLATION TECHNOLOGIES, II: SELECTIVE
EPITAXIAL GROWTH (SEG) 58**

- 2.7.1 Refill by SEG of Windows Cut into Surface Oxide, 59
- 2.7.2 Simultaneous Single-Crystal/Poly Deposition (SSPD), 60
- 2.7.3 Etching of Silicon Trenches and Refilling with SEG to Form Active
Device Regions, 61
- 2.7.4 Selective-Epitaxial-Layer Field Oxidation (SELFOX), 61
- 2.7.5 SEG Refill of Trenches (as an Alternative to Poly Refill), 62
- 2.7.6 Epitaxial Lateral Overgrowth (ELO), 62

**2.8 MISCELLANEOUS NON-LOCOS
ISOLATION TECHNOLOGIES 63**

- 2.8.1 Field-Shield Isolation, 63
- 2.8.2 Buried Insulator between Source/Drain Polysilicon (BIPS), 64

**2.9 SUMMARY: CANDIDATE ISOLATION TECHNOLOGIES FOR
SUBMICRON DEVICES 65**

- 2.9.1 Basic Requirements of VLSI and ULSI Isolation Technologies, 65
- 2.9.2 The Need for Planarity, 65
- 2.9.3 How the Various Isolation Technologies Meet the Requirements, 66

**2.10 SILICON-ON-INSULATOR (SOI) ISOLATION
TECHNOLOGIES 66**

- 2.10.1 Dielectric Isolation, 67
- 2.10.2 Wafer Bonding, 70
- 2.10.3 Silicon-on-Sapphire (SOS), 72
- 2.10.4 Separation by Implanted Oxygen (SIMOX), 72
- 2.10.5 Zone-Melting Recrystallization (ZMR), 75
- 2.10.6 Full Isolation by Porous Oxidized Silicon (FIPOS), 76
- 2.10.7 Novel SOI CMOS Processes with Selective Oxidation and Selective
Epitaxial Growth, 77

REFERENCES 79

CHAP. 3 - CONTACT TECHNOLOGY AND LOCAL INTERCONNECTS FOR VLSI	84
3.1 THE ROLE OF CONTACT STRUCTURES IN DEVICE AND CIRCUIT BEHAVIOR	84
3.1.1 Contact Structures in Planar MOSFETs and Bipolar Transistors,	85
3.2 THEORY OF METAL-SEMICONDUCTOR CONTACTS	87
3.3 EXTRACTING VALUES OF SPECIFIC CONTACT RESISTIVITY FROM MEASUREMENTS	91
3.3.1 Extraction of the Specific Contact Resistivity from an Ideal Contact Structure,	92
3.3.2 Current Flow in Actual Contact Structures,	93
3.3.3 Contact Structures Used to Extract ρ_c ,	94
3.3.4 Procedure for Accurately Extracting ρ_c from CBKR Test Structures,	97
3.3.5 Reported Values of ρ_c for Various Contact Structures,	100
3.3.6 Use of a Simple Contact-Chain Structure to Monitor Contact Resistance,	101
3.4 THE EVOLUTION OF CONVENTIONAL METAL-TO-SILICON CONTACTS	101
3.4.1 The Basic Process Sequence of Conventional Ohmic-Contact Structures to Silicon,	102
3.4.2 Additional Details Concerning the Processing Steps,	103
3.4.2.1 <i>Formation of the Heavily Doped Regions in the Silicon.</i>	
3.4.2.2 <i>Formation of Contact Openings (Etching).</i>	
3.4.2.3 <i>Sidewall Contouring of the Contact Holes by Reflow.</i>	
3.4.2.4 <i>Sidewall Contouring by Etching.</i>	
3.4.2.5 <i>Deposition.</i>	
3.4.2.6 <i>Metal Deposition and Patterning.</i>	
3.4.2.7 <i>Sintering the Contacts.</i>	
3.4.3 Aluminum-Silicon Contact Characteristics,	111
3.4.3.1 <i>The Kinetics of the Al-Si Interface During Sintering.</i>	
3.4.4 Use of Aluminum-Silicon Alloys to Reduce Junction Spiking,	116
3.4.5 Platinum Silicide-to-Silicon Contacts,	117
3.4.5.1 <i>Process Sequence Used to Form PtSi-Si Contacts.</i>	
3.4.5.2 <i>Limitations of the PtSi-Si Contact Structure.</i>	

3.5 DIFFUSION BARRIERS 121

- 3.5.1 Theory of Diffusion Barrier Layers, 121
- 3.5.2 Materials Used as Diffusion Barriers, 124
 - 3.5.2.1 *Sputter-Deposited Titanium-Tungsten (Stuffed Barrier).*
 - 3.5.2.2 *Polysilicon (Sacrificial Barrier).*
 - 3.5.2.3 *Titanium (Sacrificial Barrier).*
 - 3.5.2.4 *Titanium Nitride (Passive Barrier).*
 - 3.5.2.5 *CVD Tungsten.*
 - 3.5.2.6 *Experimental Diffusion Barrier Materials.*

3.6 MULTILAYERED OHMIC-CONTACT STRUCTURES TO SILICON 131

- 3.6.1 Al-Ti:W-PtSi-Si Contacts, 132
- 3.6.2 Al-TiN-Ti-Si Contacts, 132
- 3.6.3 Mo-Ti:W-Si and Mo-Ti-Si Contacts, 134

3.7 SCHOTTKY-BARRIER CONTACTS 134**3.8 THE IMPACT OF THE INTRINSIC SERIES RESISTANCE ON MOS TRANSISTOR PERFORMANCE 137**

- 3.8.1 The Impact of R_s on MOSFET Performance, 137
- 3.8.2 Estimates of R_{sh} , R_{sp} , R_{ac} , and R_{co} , 138
- 3.8.3 Impact of R_s on Device Characteristics, 142
- 3.8.4 Summary of the Impact of Intrinsic Series- Resistance Effects on MOSFET Performance, 142

3.9 ALTERNATIVE (SELF-ALIGNED) CONTACT STRUCTURES FOR ULSI MOS DEVICES 143

- 3.9.1 Self-Aligned Silicide Contacts, 144
 - 3.9.1.1 *Self-Aligned Titanium Silicide Contacts.*
 - 3.9.1.2 *Self-Aligned Cobalt Silicide Contacts.*
 - 3.9.1.3 *Measuring r_c of Self-Aligned Silicide Contacts.*
- 3.9.2 Buried-Oxide MOS Contact Structure (BOMOS), 153

3.10 FORMATION OF SHALLOW JUNCTIONS AND THEIR IMPACT ON CONTACT FABRICATION 154

- 3.10.1 Conventional Shallow-Junction Formation, 154
- 3.10.2 Alternative Approaches to Forming Shallow Junctions, 155
- 3.10.3 Impact of Shallow Junctions on Contact Formation, 160

3.11 BURIED CONTACTS AND LOCAL INTERCONNECTS 160

3.11.1 Butted Contacts and Buried Contacts, 160

3.11.2 Local Interconnects, 162

3.11.2.1 *Selectively Formed TiSi₂.*3.11.2.2 *Ti:W over CoSi₂.*3.11.2.3 *TiN Formed over TiSi₂.*3.11.2.4 *Dual-Doped Polysilicon LI with Diffused Source/Drain Junctions.*3.11.2.5 *CVD W-Clad Polysilicon LI.***REFERENCES****CHAP. 4 - MULTILEVEL INTERCONNECT TECHNOLOGY FOR VLSI AND ULSI 176****4.1 EARLY DEVELOPMENT OF INTERCONNECT TECHNOLOGY FOR INTEGRATED CIRCUITS 176**

4.1.1 Interconnects for Early Bipolar ICs, 176

4.1.2 Interconnects in Silicon-Gate NMOS ICs, 178

4.1.3 Evolution of Interconnects for Bipolar ICs, 179

4.1.4 Evolution of Interconnects for CMOS ICs, 180

4.2 THE NEED FOR MULTILEVEL INTERCONNECT TECHNOLOGIES 180

4.2.1 Interconnect Limitations of VLSI, 181

4.2.1.1 *Functional Density.*4.2.1.2 *Propagation Delay.*4.2.1.3 *Ease of Design and Gate Utilization for ASICs and Wafer Scale Integration.*4.2.1.4 *Cost.*

4.2.2 Problems Associated with Multimetal Interconnect Processes, 187

4.2.3 Terminology of Multilevel Interconnect Structures, 188

4.3 MATERIALS FOR MULTILEVEL INTERCONNECT TECHNOLOGIES 189

4.3.1 Conductor Materials for Multilevel Interconnects, 189

4.3.1.1 *Requirements of Conductor Materials Used for VLSI Interconnects.*4.3.1.2 *Local Interconnect Conductor Materials (Polysilicon, Metal-Silicides, and Polycides).*

4.3.1.3	<i>Aluminum Metallization.</i>	
4.3.1.4	<i>Tungsten and Other Conductor Materials for VLSI Interconnects.</i>	
4.3.2	Dielectric Materials for Multilevel Interconnects,	194
4.3.2.1	<i>Requirements of Dielectric Layers in Multilevel Interconnects.</i>	
4.3.2.2	<i>Poly-Metal Interlevel Dielectric (PMD) Materials.</i>	
4.3.2.3	<i>CVD SiO₂ Films as Intermetal Dielectrics.</i>	
4.3.2.4	<i>Low-Temperature-TEOS SiO₂ Films as Intermetal Dielectrics.</i>	
4.3.2.5	<i>Other Materials and Deposition Processes Used to Form Intermetal Dielectrics.</i>	
4.4	PLANARIZATION OF INTERLEVEL DIELECTRIC LAYERS	199
4.4.1	Terminology of Planarization in Multilevel Interconnects,	199
4.4.1.1	<i>Degree of Planarization.</i>	
4.4.1.2	<i>The Need for Dielectric Planarization.</i>	
4.4.1.3	<i>The Price that Must be Paid as the Degree of Dielectric Planarization is Increased.</i>	
4.4.1.4	<i>Design Rules Related to Intermetal Dielectric-Formation and Planarization Processes.</i>	
4.4.2	Step Height Reduction of Underlying Topography as a Technique to Alleviate the Need for Planarization,	208
4.4.2.1	<i>Provide Substrate Topography that is Completely Planar.</i>	
4.4.2.2	<i>Provide a Planar Surface over Local Interconnect Levels.</i>	
4.4.2.3	<i>Minimize the Thickness of the Metal 1 Layer.</i>	
4.4.2.4	<i>Achieve Smoothing of Steps in DM1 by Sloping the Sidewalls of Metal-1 Lines.</i>	
4.4.3	Deposition of Thick CVD SiO₂ Layers and Etching Back Without a Sacrificial Layer,	211
4.4.4	Oxide Spacers,	212
4.4.5	Polyimides as Intermetal Dielectrics,	214
4.4.6	Planarizing by Use of Bias-Sputtered SiO₂,	217
4.4.7	CVD SiO₂ and Bias-Sputter Etchback,	220
4.4.8	Planarization by Sacrificial Layer Etchback,	222
4.4.8.1	<i>Degree of Planarization Achieved by Sacrificial Etchback.</i>	
4.4.8.2	<i>Advantages of the Sacrificial Etchback Process.</i>	
4.4.8.3	<i>Sacrificial Etchback Process Problems.</i>	
4.4.8.4	<i>Alternative Sacrificial Etchback Processes.</i>	
4.4.9	Spin-On Glass (SOG),	449
4.4.9.1	<i>SOG Process Integration.</i>	
4.4.9.2	<i>The Etchback SOG Process.</i>	
4.4.9.3	<i>The Non-Etchback SOG Process.</i>	
4.4.10	Electron-Cyclotron-Resonance Plasma CVD,	237
4.4.11	Chemical-Mechanical Polishing,	238

4.5. METAL DEPOSITION AND VIA FILLING	240
4.5.1 Conventional Approach to Via Fabrication and Formation of Metal-to-Metal Contacts through the Vias,	240
4.5.1.1 <i>Design Rules of Multilevel Metal Systems which are Impacted by Conventional Via Processing Limitations.</i>	
4.5.2 Advanced Via Processing (Vertical Vias and Complete Filling of Vias by Metal),	244
4.5.2.1 <i>Increases in Packing Density Resulting from Advanced Via Process Technology.</i>	
4.5.3 Processing Techniques which Allow Vertical Vias to be Implemented,	245
4.5.3.1 <i>Required Degree of Via Filling by Plugs.</i>	
4.5.4 CVD W Techniques for Filling Vertical Vias and Contact Holes,	245
4.5.4.1 <i>General Information on the CVD Tungsten Process.</i>	
4.5.4.2 <i>Blanket CVD W and Etchback.</i>	
4.5.4.3 <i>Selective CVD W.</i>	
4.5.5 Other CVD Via Filling Processes,	253
4.5.5.1 <i>Blanket CVD Polysilicon and Etchback for Contact Hole Filling.</i>	
4.5.5.2 <i>Selective Deposition of Poly.</i>	
4.5.5.3 <i>Selectively Formed Silicide Contact Plugs.</i>	
4.5.5.4 <i>CVD Aluminum.</i>	
4.5.6 Alternatives to CVD for Filling of Vias,	254
4.5.6.1 <i>Bias Sputtering of Al to Achieve Complete Filling of Via Holes.</i>	
4.5.6.2 <i>Laser Planarization of Al Films.</i>	
4.5.6.3 <i>Contact Hole and Via Filling by Selective Electroless Metal Deposition.</i>	
4.5.7 Pillar Formation as an Alternative to Filling Contact Holes and Vias,	258
4.6 FILLED GROOVES IN A DIELECTRIC LAYER	259
4.7 MANUFACTURING YIELD AND RELIABILITY ISSUES OF VLSI INTERCONNECTS	260
4.7.1 Factors Which Impact Manufacturing Yield,	261
4.7.2 Multilevel Interconnect-Related Yield Issues,	261
4.7.3 General Reliability Issues Associated with IC Interconnects,	264
4.7.3.1 <i>Electromigration.</i>	
4.7.3.2 <i>Electromigration at the Contacts.</i>	
4.7.3.3 <i>Stress-Induced Metal Cracks and Voids.</i>	
4.7.3.4 <i>Corrosion.</i>	
4.7.4 Reliability Issues Associated with Multilevel Interconnects,	268

4.7.4.1 Hillock Formation and Prevention Measures.

4.7.4.2 Dielectric Void Reliability Problems.

4.8 PASSIVATION LAYERS 273

4.9 SURVEY OF MULTILEVEL METAL SYSTEMS 276

4.9.1 Bipolar Double-Level Metal Systems, 276

4.9.2 CMOS Double-Level-Metal Systems, 277

4.9.2.1 *Non-Planarized DLM (2.0 μm CMOS).*

4.9.2.2 *Non-Planarized DLM: CVD-W Metal (2.0- μm NMOS).*

4.9.2.3 *Resist Etchback, Bias-Sputtered SiO_2 , and SOG DLM for 1.5 μm CMOS.*

4.9.2.4 *Non-Sacrificial Layer Etchback DLM (1.0- μm CMOS).*

4.9.2.5 *Alternative CMOS DLM Process with Ti:W/Mo as Metal 1.*

4.9.2.6 *DLM Processes for Submicron CMOS.*

4.9.3 Three-Level Metal Systems, 283

4.9.4 Four-Level Metal Systems, 285

4.10 SUMMARY OF MULTILEVEL INTERCONNECT TECHNOLOGY REQUIREMENTS FOR VLSI 286

REFERENCES 287

CHAP. 5 - MOS DEVICES AND NMOS PROCESS INTEGRATION 298

5.1 MOS DEVICE PHYSICS 298

5.1.1 The Structure and Device Fundamentals of MOS Transistors, 298

5.1.2 The Threshold Voltage of the MOS Transistor, 301

5.1.3 Impact of Source-Body Bias on V_T (Body Effect), 304

5.1.4 Current-Voltage Characteristics of
MOS Transistors, 305

5.1.5 The Capacitances of MOS Transistors, 307

5.2 MAXIMIZING DEVICE PERFORMANCE THROUGH DEVICE DESIGN AND PROCESSING TECHNOLOGY 307

5.2.1 Output Current (I_D) and Transconductance (g_m), 308

5.2.2 Controlling the Threshold Voltage through Process
and Circuit-Design Techniques, 309

5.2.3 Subthreshold Currents (I_{Dst} when $V_G < IV_T$), 311

- 5.2.4 Switching Speed, 313
- 5.2.5 Junction Breakdown Voltage (Drain-to-Substrate), 313
- 5.2.6 Gate-Oxide Breakdown Voltage, 314
- 5.2.7 High Field-Region Threshold-Voltage Value, 315

5.3 THE EVOLUTION OF MOS TECHNOLOGY (PMOS AND NMOS) 315

- 5.3.1 Aluminum-Gate PMOS, 316
- 5.3.2 Silicon-Gate MOS Technology, 318
- 5.3.3 Reduction of Oxide-Charge Densities, 319
- 5.3.4 Ion Implantation for Adjusting Threshold Voltage, 321
- 5.3.5 Isolation Technology for MOS, 323
- 5.3.6 Short-Channel Devices, 323

5.4 PROCESS SEQUENCE FOR FABRICATING NMOS INVERTERS WITH DEPLETION-MODE LOADS 324

- 5.4.1 Operation of an NMOS Inverter with a Depletion-Mode Load, 324
- 5.4.2 Process Sequence of a Basic E-D NMOS IC Technology, 327
 - 5.4.2.1 *Starting Material.*
 - 5.4.2.2 *Active Region and Field Region Definitions.*
 - 5.4.2.3 *Gate-Oxide Growth and Threshold-Voltage Adjust Implant*
 - 5.4.2.4 *Polysilicon Deposition and Patterning.*
 - 5.4.2.5 *Formation of the Source and Drain Regions.*
 - 5.4.2.6 *Contact Formation.*
 - 5.4.2.7 *Metallization Deposition and Patterning.*
 - 5.4.2.8 *Passivation Layer and Pad Mask.*

5.5 SHORT-CHANNEL EFFECTS AND HOW THEY IMPACT MOS PROCESSING 338

- 5.5.1 Effect of Gate Dimensions on Threshold Voltage, 338
 - 5.5.1.1 *Short Channel Threshold Voltage Effect.*
 - 5.5.1.2 *Narrow Gate-Width Effect on Threshold Voltage.*
- 5.5.2 Short-Channel Effects on Subthreshold Currents (Punchthrough and Drain-Induced Barrier Lowering), 341
- 5.5.3 Short-Channel Effects on I-V Characteristics, 343
- 5.5.4 Summary of Short-Channel Effects on the Fabrication of MOS ICs, 346

5.6 HOT-CARRIER EFFECTS IN MOSFETS 348

- 5.6.1 Substrate Currents Due to Hot Carriers, 349

- 5.6.2 Hot-Carrier Injection into the Gate Oxide, 350
- 5.6.3 Device-Performance Degradation Due to Hot-Carrier Effects, 352
- 5.6.4 Techniques for Reducing Hot-Carrier Degradation, 354
- 5.6.5 Lightly Doped Drains, 354
 - 5.6.5.1 *Drain Engineering for Optimum LDD Structures.*
 - 5.6.5.2 *Asymmetrical Characteristics of LDD MOSFETs.*
- 5.6.6 The Impact of IC Processing
 - on Hot-Carrier Device Degradation, 361
- 5.6.7 Hot-Carrier Effects in PMOS Transistors, 362
- 5.6.8 Gate-Induced Drain-Leakage Current, 363

REFERENCES 363

CHAP. 6 - CMOS PROCESS INTEGRATION 368

6.1 INTRODUCTION TO CMOS TECHNOLOGY 368

- 6.1.1 The Power-Dissipation Crisis of VLSI and How CMOS Came to the Rescue, 368
- 6.1.2 Historical Evolution of CMOS, 370
- 6.1.3 Operation of CMOS Inverters, 373
- 6.1.4 Advantages (and Disadvantages)
 - of Modern CMOS Technologies, 376
 - 6.1.4.1 *Device/Chip Performance Advantages.*
 - 6.1.4.2 *Reliability Advantages of CMOS.*
 - 6.1.4.3 *Circuit Design Advantages of CMOS.*
 - 6.1.4.4 *Cost Analysis of CMOS.*
- 6.1.5 Disadvantages of CMOS, 380

6.2 THE WELL CONTROVERSY IN CMOS 381

- 6.2.1 The Need for Wells in CMOS, 381
- 6.2.2 *p*-Well CMOS, 383
- 6.2.3 *n*-Well CMOS, 384
- 6.2.4 CMOS on Epitaxial Substrates, 385
- 6.2.5 Twin-Well CMOS, 387
- 6.2.6 Retrograde-Well CMOS, 389
- 6.2.7 Summary of CMOS Well-Technology Issues, 392

6.3 *p*-CHANNEL DEVICES IN CMOS 392

- 6.3.1 PMOS Devices with *n*⁺-Polysilicon Gates, 392
 - 6.3.1.1 *Punchthrough Susceptibility.*

- 6.3.2 PMOS Devices with p^+ -Polysilicon Gates, 397
- 6.3.3 Gate Materials having Symmetrical Work Functions (with Respect to both NMOS and PMOS Devices), 398

6.4 LATCHUP IN CMOS 400

- 6.4.1 Parasitic *pnpn* Structures in CMOS Circuits, 400
- 6.4.2 Circuit Behavior of *pnpn* Diodes, 402
- 6.4.3 Device Physics Behavior of *pnpn* Diodes, 403
- 6.4.4 Summary of Conditions That Must Exist in Order for Latchup to Occur, 406
- 6.4.5 Circuit Behavior of Actual *pnpn* Structures in CMOS Circuits, 406
 - 6.4.5.1 Value of β in CMOS Vertical Parasitic Bipolar Transistors.
 - 6.4.5.2 Value of β in CMOS Lateral Parasitic Bipolar Transistors.
- 6.4.6 Circuit and Device Effects that Induce Latchup, 408
 - 6.4.6.1 An external stimulus forward-biases the emitter-base of one transistor, and its collector current then turns-on the second transistor.
 - 6.4.6.2 An external stimulus causes current to flow through both bypass resistors, forward-biasing one or both bipolar transistors.
 - 6.4.6.3 Current is shunted through one of the parasitic transistors by some degradation mechanism, and the resulting collector current flows through the bypass resistor of the second transistor and turns it on.
- 6.4.7 Test Methods for Characterizing Latchup, 410
 - 6.4.7.1 Modelling Latchup in CMOS Technology.
- 6.4.8 Techniques for Reduction or Elimination of Latchup Susceptibility, 413
 - 6.4.8.1 Processing Techniques that Reduce the Current Gains of the Parasitic Bipolar Transistors.
 - 6.4.8.2 Processing Techniques that Reduce R_{sub} and R_w or Eliminate the *pnpn* Structure.
 - 6.4.8.3 Circuit Layout Techniques used to Decouple Parasitic Bipolar Transistors.

6.5 CMOS ISOLATION TECHNOLOGY 419

- 6.5.1 Trench Isolation for CMOS, 425
- 6.5.2 Isolation by Selective-Epitaxial Growth for CMOS, 426

6.6 CMOS PROCESS SEQUENCES 428

- 6.6.1 Basic *n*-Well CMOS Process Sequence, 428
- 6.6.2 Twin-Well CMOS Process Sequence, 431
 - 6.6.2.1 Starting Material.
 - 6.6.2.2 Forming the Wells and Channel Stops.

- 6.6.2.3 *Active and Field Region Definition.*
- 6.6.2.4 *Gate Oxide Growth and Threshold Voltage Adjustment.*
- 6.6.2.5 *Polysilicon Deposition and Patterning.*
- 6.6.2.6 *Formation of the Source/Drain Regions.*
- 6.6.2.7 *CVD Oxide Deposition and Contact Formation.*
- 6.6.2.8 *Metal 1 Deposition and Patterning.*
- 6.6.2.9 *Intermetal Dielectric Deposition/Planarization and Via Patterning.*
- 6.6.2.10 *Metal 2 Deposition and Patterning.*
- 6.6.2.11 *Passivation Layer Deposition and Patterning.*

6.7 MISCELLANEOUS CMOS TOPICS 441

- 6.7.1 Electrostatic Discharge Protection for CMOS, 441
 - 6.7.1.1 *Diode Protection.*
 - 6.7.1.2 *Node-to-Node Punchthrough.*
 - 6.7.1.3 *Gate-Controlled Breakdown Structure.*
 - 6.7.1.4 *pnpn-Diode ESD Protection for Advanced CMOS Circuits.*
- 6.7.2 Power Supply Voltage Levels for Future CMOS, 446
- 6.7.3 Low-Temperature CMOS, 446
- 6.7.4 Three-Dimensional CMOS, 447

REFERENCES 447

CHAPTER 7 - BIPOLAR AND

BICMOS PROCESS INTEGRATION

453

7.1 BIPOLAR TRANSISTOR STRUCTURES FOR INTEGRATED CIRCUITS 453

- 7.1.1 The Transistor Action 454
 - 7.1.1.1 *Basic Bipolar Transistor Physics.*
 - 7.1.1.2 *Bipolar Transistor Current Gain.*
- 7.1.2 Integrated-Circuit Transistor Structures 458

7.2 DIGITAL CIRCUITS USING BIPOLAR TRANSISTORS 459

- 7.2.1 Basic Bipolar-Transistor Inverter Circuits 459
- 7.2.2 Bipolar Digital-Logic-Circuit Families 460

7.3 MAXIMIZING BIPOLAR TRANSISTOR PERFORMANCE THROUGH DEVICE DESIGN & PROCESSING TECHNOLOGY 464

7.3.1	Current Gain	464
7.3.2	Early Voltage	466
7.3.3	High-Level Injection Effects (Kirk Effect)	467
7.3.4	Operating-Voltage Limits in Bipolar Transistors	468
7.3.4.1	<i>Reachthrough Breakdown.</i>	
7.3.4.2	<i>Punchthrough Breakdown.</i>	
7.3.4.3	<i>Breakdown Voltage and High-Level Injection Limits in Advanced Bipolar Transistors.</i>	
7.3.5	Parasitic Series Resistances in Bipolar Transistors	472
7.3.5.1	<i>Collector Series Resistance, R_C.</i>	
7.3.5.2	<i>Base Series Resistance, R_B.</i>	
7.3.5.3	<i>Base-Spreading Resistance, R_{B2} (and Emitter Current Crowding).</i>	
7.3.5.4	<i>Emitter Series Resistance, R_E.</i>	
7.3.6	Parasitic Junction Capacitances in Bipolar Transistors	475
7.3.6.1	<i>Storage Capacitances in Bipolar Transistors.</i>	
7.3.7	Bipolar Transistor Unity-Gain Frequency, f_T	477
7.3.8	First Order <i>npn</i> Device Design	477
7.3.9	Switching Speed Behavior in Bipolar ICs	478
7.3.9.1	<i>Propagation-Delay Time Calculation in Bipolar Transistors.</i>	
7.3.9.2	<i>Propagation Delay in Digital MOS versus Digital Bipolar Circuits.</i>	
7.3.9.3	<i>General Switching Speed Behavior of Digital Bipolar Circuits.</i>	
7.4	NON-OXIDE-ISOLATED BIPOLAR <i>npn</i> TRANSISTOR STRUCTURES	482
7.4.1	Triple-Diffused (3D) Process	483
7.5	STANDARD-BURIED-COLLECTOR PROCESS	483
7.5.1	Characteristics of <i>npn</i> Transistors Fabricated with the Standard-Buried-Collector (SBC) Process	483
7.5.1.1	<i>Limitations of Junction-Isolated SBC Transistors for VLSI Circuits.</i>	
7.5.2	Standard-Buried-Collector Process Flow	486
7.5.2.1	<i>Starting Material.</i>	
7.5.2.2	<i>Buried Layer Formation.</i>	
7.5.2.3	<i>Epitaxial Growth.</i>	
7.5.2.4	<i>Formation of Isolation Regions.</i>	
7.5.2.5	<i>Deep-Collector Contact Formation (Optional).</i>	
7.5.2.6	<i>Base Region Formation.</i>	
7.5.2.7	<i>Emitter Region Formation.</i>	
7.5.2.8	<i>Contact and Interconnect Layer Formation.</i>	
7.5.2.9	<i>Washed Emitters.</i>	
7.5.2.10	<i>Schottky Contacts.</i>	
7.6	OXIDE-ISOLATED BIPOLAR TRANSISTORS	498

7.7 ADVANCED BIPOLAR TRANSISTOR STRUCTURES FOR VLSI AND ULSI 500**7.8 ADVANCED EMITTER STRUCTURES 501****7.8.1 Polysilicon Emitters 501***7.8.1.1 Models that Describe Polysilicon-Emitter Behavior.**7.8.1.2 Process Technology for Polysilicon-Emitter Fabrication.***7.8.2 Heterojunction Bipolar Transistors (HBTs) 506****7.9 SELF-ALIGNED BIPOLAR STRUCTURES 510****7.9.1 Double-Polysilicon Self-Aligned Structures 510***7.9.1.1 Limitations of Double-Polysilicon SA Structures.**7.9.1.2 Current-Gain Degradation Due to Sidewall Injection in SA Bipolar Structures.**7.9.1.3 Link-Up Region Formation.***7.9.2 Single-Polysilicon Self-Aligned Bipolar Structures 516****7.9.3 Sidewall-Base-Contact Structures (SICOS) 520****7.10 TRENCH-ISOLATED BIPOLAR TRANSISTORS 522****7.11 BICMOS TECHNOLOGY 523****7.11.1 Device and Circuit Advantages of BiCMOS 524***7.11.1.1 Comparison of BiCMOs and CMOS Propagation Delay Times.**7.11.1.2 Power Consumption of BiCMOS versus CMOS Gates.**7.11.1.3 Capability of Providing Either TTL or ECL Outputs From a BiCMOS Chip.**7.11.1.4 Process Complexity Increases Associated with BiCMOS.**7.11.1.5 Extending Process Equipment Life by Fabricating BiCMOS.***7.12 CLASSIFICATION OF BICMOS TECHNOLOGIES 529****7.12.1 Digital BiCMOS Technology 531***7.12.1.1 Low-Cost Digital BiCMOS Technology.**7.12.1.2 High-Performance Digital BiCMOS.**7.12.1.3 Device-Design Issues Related to Optimizing a High-Performance Digital Modified-Twin-Well BiCMOS Process.**7.12.1.4 An Example Process Sequence for Fabricating High-Performance 5-V Digital BiCMOS ICs.***7.12.2 Process Integration of Analog/Digital BiCMOS 543***7.12.2.1 Process-Integration Issues of Medium-Voltage Analog BiCMOS.**7.12.2.2 An Example of an Analog/Digital BiCMOS Process.*

7.12.3 BiCMOS Applications	551
7.12.3.1 <i>Digital Logic Circuits and Gate Arrays.</i>	
7.12.3.2 <i>Interface Driver Circuits.</i>	
7.12.3.3 <i>BiCMOS SRAMs.</i>	
7.12.3.4 <i>Analog/Digital Applications.</i>	
7.13 Trends in BiCMOS Technology	556
7.13 COMPLEMENTARY BIPOLAR (CB) TECHNOLOGY	557
REFERENCES	560
 CHAP. 8 - SEMICONDUCTOR MEMORY PROCESS INTEGRATION	 557
8.1 TERMINOLOGY OF SEMICONDUCTOR MEMORIES	557
8.1.1 Random-Access and Read-Only Memories (RAMs and ROMs)	568
8.1.2 Semiconductor-Memory Architecture	568
8.1.3 Semiconductor-Memory Types	570
8.1.4 Read Access Times and Cycle Times in Memories	571
8.1.5 Recently Introduced On-Chip Peripheral Circuits	571
8.1.6 Logic-Memory Circuits	571
8.2 STATIC RANDOM-ACCESS MEMORIES (SRAMS)	572
8.2.1 MOS SRAMs	575
8.2.1.1 <i>Circuit Operation of MOS SRAM Cells.</i>	
8.2.1.2 <i>SRAM Processing and Cell Layout Issues.</i>	
8.2.1.3 <i>High-Valued Polysilicon Load-Resistors for MOS SRAMs</i>	
8.2.2 Bipolar and BiCMOS SRAMS	584
8.2.2.1 <i>BiCMOS SRAMs.</i>	
8.3 DYNAMIC RANDOM ACCESS MEMORIES (DRAMs)	587
8.3.1 Evolution of DRAM Technology	587
8.3.1.1 <i>One-Transistor DRAM Cell Design.</i>	
8.3.1.2 <i>Operation of the One-Transistor DRAM Cell.</i>	
8.3.1.3 <i>Writing, Reading, and Refreshing DRAM Cells.</i>	
8.3.1.4 <i>Quantity of Charge Stored on DRAM Cells and Their Capacitance.</i>	
8.3.1.5 <i>High-Capacity (Hi-C) DRAM Cells.</i>	
8.3.1.6 <i>CMOS DRAMs.</i>	
8.3.2 Design and Economic Constraints on Advanced DRAM Cells	597

8.3.3	Trench Capacitor DRAM Cells	600
8.3.3.1	<i>Trench Capacitor Processing for DRAMs.</i>	
8.3.3.2	<i>First Generation Trench Capacitor-based DRAM Cells.</i>	
8.3.3.3	<i>Trench Capacitor Structures with the Storage Electrode Inside the Trench (Inverted Trench Cell).</i>	
8.3.3.4	<i>Trench Capacitor Cells with the Access Transistor Stacked Above the Trench Capacitor.</i>	
8.3.4	Stacked Capacitor DRAM Cells	609
8.3.5	Soft-Error Failures in DRAMs	615
8.3.5.1	<i>Techniques Used to Reduce the Soft-Error Rates in DRAMs.</i>	
8.3.6	The DRAM as a Technology Driver	618
8.4	MASKED READ-ONLY MEMORIES (ROMs)	619
8.4.1	Masked ROM Implementation	620
8.5	PROGRAMMABLE ROMS (PROMS)	621
8.6	ERASABLE PROGRAMMABLE READ-ONLY MEMORIES (EPROMS)	623
8.7	ELECTRICALLY-ERASABLE PROMS (EEPROMS)	628
8.7.1	MNOS-Based EEPROMs	628
8.7.2	FLOTOX EEPROMs	629
8.7.3	Textured-Polysilicon EEPROMs	631
8.8	FLASH EEPROMS	632
8.9	NONVOLATILE FERROELECTRIC MOS RAMS	635
	REFERENCES	637

CHAP. 9 - PROCESS SIMULATION 643

9.1	OVERVIEW OF PROCESS SIMULATION	644
9.1.1	Hierarchy of Simulation Tools for IC Development	644
9.1.2	Benefits and Limitations of Process Simulation	645
9.1.3	Overview of Process Simulators	647
9.1.3.1	<i>Simulator Availability.</i>	
9.1.4	General Aspects of Process Simulation	650
9.1.4.1	<i>Analytical and Numerical Methods of Solving the Equations that Describe Processes.</i>	

9.1.4.2 Phenomenological versus Physical Models.

9.1.4.3 Gridding.

9.1.4.4 Interfacing One Simulator with Another.

9.2 ONE-DIMENSIONAL PROCESS SIMULATORS 653

9.2.1 SUPREM III (Stanford University Process Engineering Model III) 655

9.2.1.1 The Basic Operation and Capabilities of SUPREM III.

9.2.1.2 Additional Comments on the Use of SUPREM III.

9.2.2 SUPREM III Models: Ion Implantation 658

9.2.3 SUPREM III Models: Diffusion in Silicon and SiO₂, and Segregation Effects at the Si/SiO₂ Interface 663

9.2.3.1 Diffusion Models Used in SUPREM III.

9.2.3.2 Modeling Low Impurity-Concentration (Intrinsic) Diffusion in Silicon.

9.2.3.3 Modeling High-Impurity Concentration (Extrinsic) Diffusion in Silicon.

9.2.3.4 Oxidation-Enhanced Diffusion Modeling in SUPREM III.

9.2.3.5 Dopant Segregation Effects at the Si-SiO₂ Interface and Diffusion in SiO₂.

9.2.4 SUPREM III Models: Thermal Oxidation of Silicon in One-Dimension 669

9.2.4.1 High Dopant-Concentration Cases.

9.2.4.2 Modeling Other Factors Which Impact the Oxide Growth Rate.

9.2.4.3 Accuracy of Modeling Oxide Growth with SUPREM III.

9.2.5 SUPREM III Models: Epitaxial Growth 674

9.2.6 SUPREM III Models: Deposition, Oxidation, and Material Properties of Polysilicon Films 675

9.2.7 Creating a SUPREM III Input File 677

9.2.8 PREDICT 679

9.3 INTRODUCTION TO 2-DIMENSIONAL PROCESS SIMULATORS 680

9.3.1 Classes of 2-Dimensional Process Simulators 683

9.4 TWO-DIMENSIONAL DOPING-PROFILE AND OXIDATION PROCESS SIMULATORS 684

9.4.1 SUPRA (Stanford University Process Analysis Program) 684

9.4.1.1 SUPRA Ion Implantation Models.

9.4.1.2 SUPRA Diffusion Models.

9.4.1.3 SUPRA Oxidation Models.

9.4.1.4 SUPRA Epitaxial Model.

9.4.1.5 SUPRA Input File.

9.4.2 SUPREM IV 687

9.4.2.1	<i>SUPREM IV Models of Diffusion.</i>	
9.4.2.2	<i>SUPREM IV Models of Oxidation.</i>	
9.4.2.3	<i>SUPREM IV Models of Ion Implantation, Epitaxy, Deposition, and Etching.</i>	
9.4.2.4	<i>SUPREM IV Input File Format.</i>	
9.4.2.5	<i>Comparison of SUPRA and SUPREM IV for 2-D Process Simulation.</i>	
9.4.3	Two-Dimensional Simulation of Thermal Oxidation	690
9.4.3.1	<i>Empirical Models of 2-D Thermal Oxidation.</i>	
9.4.3.2	<i>Physical-Based Models of 2-D Thermal Oxidation.</i>	
9.5	TWO-DIMENSIONAL TOPOGRAPHY SIMULATORS	696
9.6	SAMPLE (SIMULATION AND MODELING OF PROFILES IN LITHOGRAPHY AND ETCHING)	697
9.6.1	Simulating Optical Lithography Processes with SAMPLE	697
9.6.1.1	<i>Optical Imaging Subprogram.</i>	
9.6.1.2	<i>Resist Exposure Subprogram.</i>	
9.6.1.3	<i>Resist Development Subprogram.</i>	
9.6.2	Simulating Etching and Deposition with SAMPLE	706
9.6.3	Creating Input Files for SAMPLE	708
9.7	OTHER 2-D TOPOGRAPHY SIMULATORS	710
9.7.1	PROLITH	710
9.7.2	DEPICT	710
9.7.3	PROFILE	711
9.7.4	SIMBAD	713
9.7.5	SIMPL (Simulated Programs from the Layout)	714
9.7.6	SIMPL-DIX	716
9.7.7	Manufacturing-Based Process Simulators	718
9.8	DEVICE SIMULATORS	718
9.8.1	Simulation of MOS Device Characteristics under Subthreshold and Linear Operation (GEMINI)	719
9.8.2	Simulation of MOS Device Under All dc Operating Conditions (MINIMOS, CADDET, CANDE)	719
9.8.3	Bipolar Device Simulators (SEDAN, BIPOLE)	720
9.8.4	Combined MOS and Bipolar Device Simulators (PICSES, SIFCOD, PADRE, and FIELDAY)	721

9.9 CIRCUIT SIMULATORS AND ELECTRICAL PARAMETER EXTRACTORS	723
9.10 FUTURE CHALLENGES IN PROCESS SIMULATION	723
REFERENCES	724
APPENDIX A IC RESISTOR FABRICATION	731
APPENDIX B PROPERTIES OF SILICON AT 300 °K	737
APPENDIX C PHYSICAL CONSTANTS	738
INDEX	739

LIST OF TECHNICAL REVIEWERS

Each of the chapters was reviewed for technical correctness. The following persons graciously undertook the review task for the chapters indicated:

Chapter 2	Dr. Joseph R. Monkowski Lam Research Corp. - CVD Division Fremont, CA	Dr. Haiping Dun Intel Corp. Santa Clara, CA
Chapter 3	Dr. Robert S. Blewer Sandia National Laboratories Albuquerque, NM	Dr. Stan Swirhun Honeywell SSPL Bloomington, MN
Chapter 4	Dr. Farhad K. Moghadam Intel Corp. Santa Clara, CA	Dr. Terry Herndon MIT - Lincoln Laboratory Lexington, MA
Chapter 5	Mr. Andrew R. Coulson TRW Electronic Systems Group Redondo Beach, CA	
Chapter 6	Dr. John Y. Chen Boeing Electronics Seattle, WA	Dr. Samuel T. Wang International CMOS Technology, Inc. San Jose, CA
Chapter 8	Professor Al F. Tasch, Jr. University of Texas Austin, TX	
Chapter 9	Dr. Michael Kump Technology Modeling Associates, Inc. Palo Alto, CA	

PREFACE

SILICON PROCESSING FOR THE VLSI ERA is a text designed to provide a comprehensive and up-to-date treatment of this important and rapidly changing field. The text will consist of three volumes, of which this book is the second, subtitled, *Process Integration*. Volume 1, subtitled *Process Technology*, was published in 1986. Volume 3, to be subtitled *Assembly, Packaging and Manufacturing Technology* is scheduled for publication in 1993. In Volume 1, the individual processes utilized in the fabrication of silicon VLSI circuits are covered in depth (e.g., epitaxial growth, chemical vapor and physical vapor deposition of amorphous and polycrystalline films, thermal oxidation of silicon, diffusion, ion implantation, microlithography, and etching processes).

In this volume, we undertake to explain how the individual processes described in Volume 1 are combined in various ways to produce silicon integrated circuits. This task is referred to as *process integration*. The first part of the book deals with *sub-process integration*; that is, the effort involved in forming circuit structures that can be implemented into a variety of circuit types. These structures include *isolation structures* (Chap. 2), *metal-silicon contacts* (Chap. 3), and *device-interconnect structures* (Chap. 4).

The second part of the book covers the process integration tasks of full-device-type technologies, including *NMOS* (Chap. 5), *CMOS* (Chap. 6), *bipolar* and *BiCMOS* (Chap. 7), and *semiconductor memories* (Chap. 8). Chapter 9 describes the process simulation tools that are available for aiding in the process integration and development efforts.

Volume 3 will cover process control, VLSI manufacturing issues and facilities, contamination and yield, automation, assembly, packaging, and parametric testing.

The purpose of writing this text was to provide professionals involved in the microelectronics industry with a single source that offers a complete overview of the technology associated with the manufacture of silicon integrated circuits. Other texts on the subject are available only in the form of specialized books (i.e., that treat just a small subset of all of the processes), or in the form of edited volumes (i.e., books in which a group of authors each contributes a small portion of the contents).

Such edited volumes typically suffer from a lack of unity in the presented material from chapter-to-chapter, as well as an unevenness in writing style and level of presentation. In addition, in multi-disciplinary fields, such as microelectronic fabrication, it is difficult for most readers to follow technical arguments in such books,

xxx

especially if the information is presented without defining each technical "buzzword" as it is first introduced.

In our books such drawbacks are avoided by treating the subject of VLSI fabrication from a unified and more pedagogical viewpoint, and by carefully defining technical terms when they are first introduced. The result is intended to be a user friendly book for workers who have come to the semiconductor industry after having been trained in but one of the many traditional technical disciplines.

An important technical breakthrough has occurred in publishing that the author also felt could be exploited in creating a unique book on silicon processing. That is, revolutionary electronic publishing techniques have recently become available, which can cut the time required to produce a published book from a finished manuscript. This task traditionally took 15-18 months, but can be now reduced to less than 3 months. If traditional techniques are used to produce books in such fast-breaking fields as VLSI fabrication, these books automatically possess a built-in obsolescence, even upon being first published.

We have taken advantage of these rapid production techniques, and have been able to successfully meet the reduced production-time schedule. As a result, information contained in technical journals and conferences which was available within three months of the book's publication date has been included.

Written for the professional, the book belongs on the bookshelf of workers in several microelectronic disciplines. Microelectronic fabrication engineers who seek to develop a more complete perspective of the subject, or who are new to the field, will find it invaluable. Integrated circuit designers, test engineers, and integrated circuit equipment designers, who must understand VLSI processing issues to effectively interface with the fabrication environment, will also find it a uniquely useful reference. The book should also be very suitable as a text for graduate-level courses on silicon processing techniques, offered to students of electrical engineering, applied physics, and materials science. It is assumed that such students already possess a basic familiarity with semiconductor device physics. Problems are included at the end of each chapter to assist readers in gauging how well they have assimilated the material in the text.

The book is an outgrowth of several intensive seminars conducted by the author through the Engineering Extension of the University of California, Irvine. Over one thousand engineers and managers from more than 75 companies and government agencies have enrolled in these short courses since they were first presented in 1984.

A book of this length and diversity would not have been possible without the indirect and direct assistance of many other workers. To begin, virtually all of the information presented in this text is based on the research efforts of a countless number 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. 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 authors express heartfelt thanks to those who gave of their time, energy, and intellect.

Each of the chapters was reviewed after the writing was completed. The engineers and scientists who participated in this review were numerous. The main reviewers are

listed on the page before the preface, and we would like to thank them once again at this point for their contributions. That is, the following professionals (each one an expert in the topic covered by the chapter they reviewed), read an entire chapter for technical correctness, and provided appropriate comments and corrections: Robert S. Blewer, John Y. Chen, Andrew R. Coulson, Haiping Dun, Terry Herndon, Michael Kump, Farhad K. Moghadam, Joseph R. Monkowski, Stan Swirhun, Al F. Tasch, Jr., and Samuel T. Wang. In addition, J. B. Price, of Spectrum CVD, Chris A. Mack, of the National Security Agency, and Sidney Marshall, Editor of Solid State Technology, provided other valuable technical and editorial input. Robert Shier, of VTC Corp., and Jim Cable of TRW also kindly provided the author with timely and valuable technical literature.

The copy editing of the book was undertaken by Mary Nadler, and the clarity and grammatical correctness of the prose owes a great deal to her professional efforts. The aesthetically pleasing graphics of the cover were designed by Roy Montibon of Visionary Art Resources, Inc., Santa Ana, CA.

Stanley Wolf

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

Lattice Press
P.O. Box 340-W
Sunset Beach, CA, 90742

An order form, for your convenience, is provided on the final leaf of the book.

CHAPTER 1

PROCESS INTEGRATION

FOR VLSI and ULSI

Since the creation of the first integrated circuit in 1960, the density of devices that can be fabricated on semiconductor substrates has steadily increased. In the late 1970s the number of devices manufactured on a chip exceeded the generally accepted definition of *very large scale integration*, or *VLSI* (i.e., more than 100,000 devices per chip) (Fig. 1-1a). By 1990 this number had grown to more than 32 million devices per chip (16-Mbit DRAMs), and it is generally acknowledged that the era of *ultra-large-scale integration* (*ULSI*) has begun. The increasing device count has been accompanied by a shrinking minimum feature size (Fig. 1-1b), which as of 1990 had decreased to $\sim 0.5 \mu\text{m}$ in the most advanced commercially available chips.

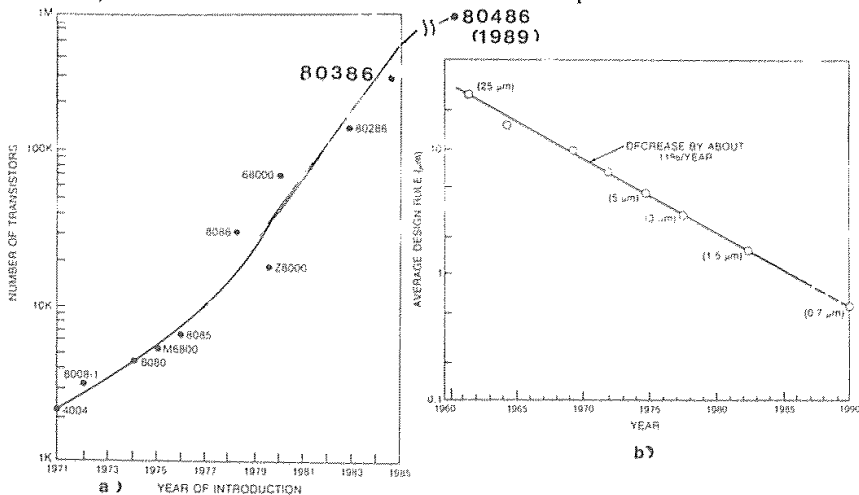


Fig. 1-1 (a) Increase in the number of transistors per microprocessor chip versus year of introduction, for a variety of 8-bit and 16-bit microprocessors, and (b) The decrease in minimum device feature size versus time on integrated circuits.

2 SILICON PROCESSING FOR THE VLSI ERA – VOLUME II

Progress seems likely to continue at a rapid pace, with even further reductions in the unit cost per function and in the power-delay product of integrated circuits projected. Silicon processing has been the dominant technology of IC fabrication and is likely to retain this position for the foreseeable future. The entire adventure of silicon device manufacturing represents a remarkable application of scientific knowledge to the requirements of technology. Our books* are intended to serve as a comprehensive and cohesive report on the state of the art of this technology, as practiced at the time of publication.

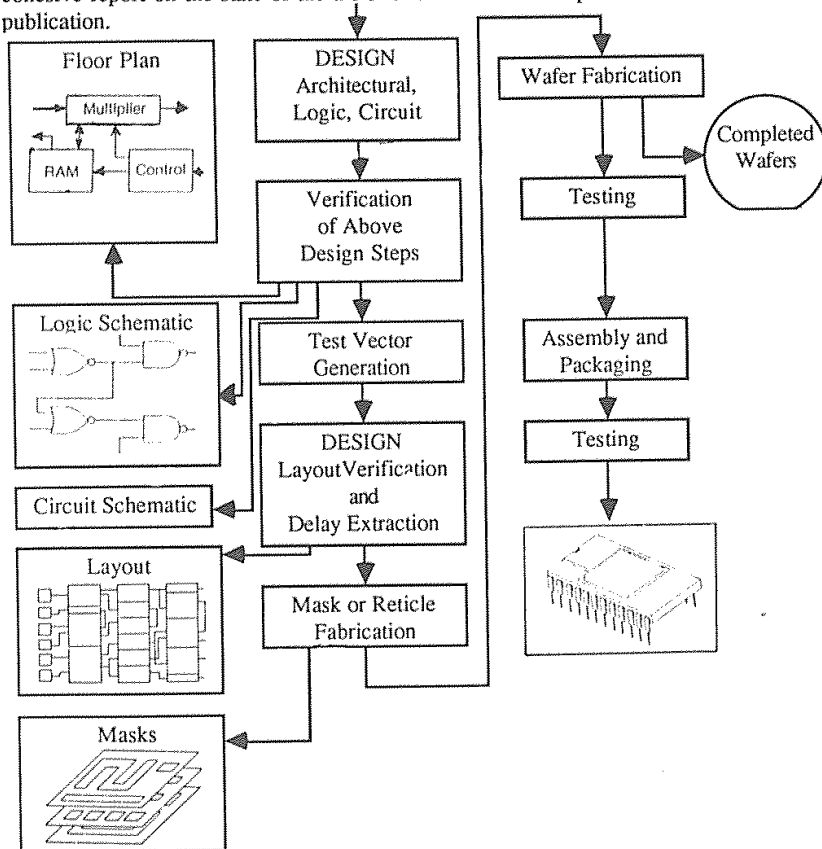


Fig. I-2 Steps required for the manufacture of very large scale integrated circuits (VLSI).

* SILICON PROCESSING FOR THE VLSI ERA - Volume 1: Process Technology; Volume 2 - Process Integration; and Volume 3 - Assembly, Packaging, and Manufacturing Technology (the latter is scheduled to be published *approximately* in 1992)

Figure 1-2 illustrates the sequence of tasks followed in the manufacture of an integrated circuit. These tasks can be grouped into three phases: *design*, *fabrication*, and *testing*. Our books are concerned primarily with the fabrication phase. We describe the IC manufacturing steps that occur from the point at which the circuit design has been completed and the necessary design information has been rendered into the form of a circuit layout. In this form, the layout information is ready to be used to generate a set of masks (or reticles) that will serve as the tools for specifying the circuit patterns on silicon wafers. For VLSI and ULSI circuits, the layout information is stored in a computer.

The details concerning the individual fabrication processes (e.g., those associated with creating patterns, introducing dopants, and depositing films on silicon substrates to form integrated circuits) are the subjects of Volume 1. In this second volume, we

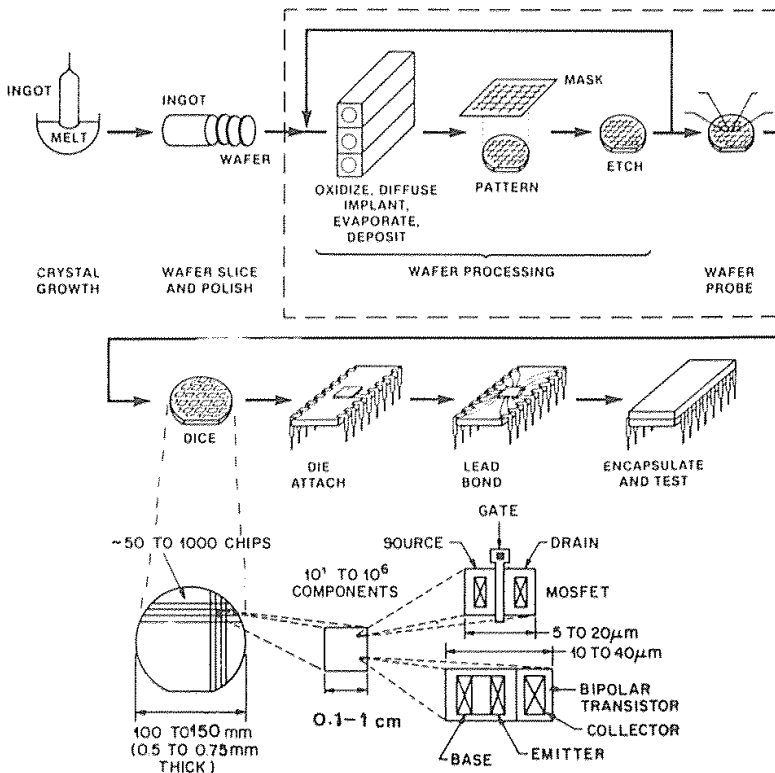


Fig. 1-3 The fabrication process sequence of integrated circuits.

4 SILICON PROCESSING FOR THE VLSI ERA – VOLUME II

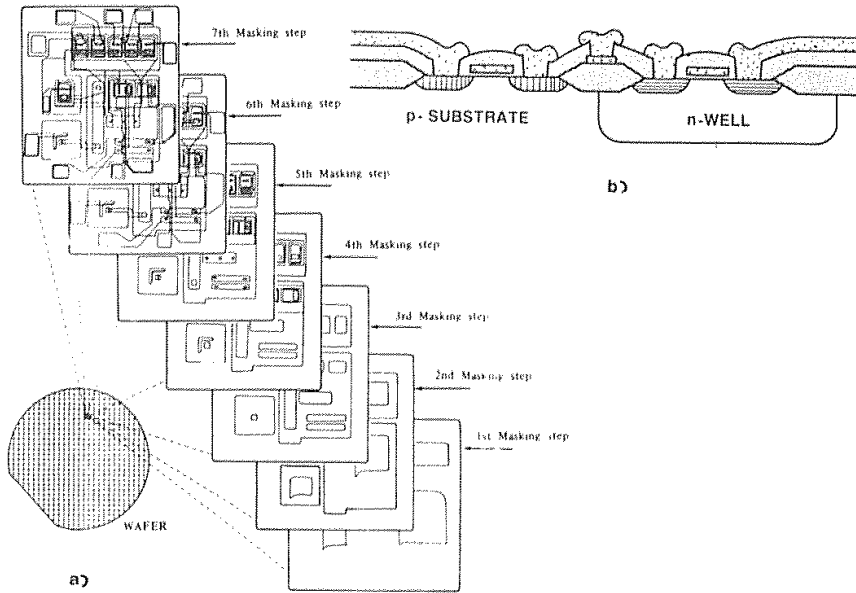


Fig. 1-4 (a) Example of the patterns transferred to a wafer during a seven-mask process sequence, and (b) Cross section of completed devices in a basic CMOS process.

undertake a discussion of the *sequences* of the steps performed to produce the device structures of the circuits. The text describes how the various individual processes are integrated together so that the end result is a completely realized integrated circuit.

One perspective of such a *process-integration* effort is illustrated in Fig. 1-4. It can be seen that a series of masking steps must be sequentially performed for the desired patterns to be created on the wafer surface. The other processing procedures performed between the masking steps serve to create the desired device structures. An example of the end result of a complete process sequence is shown in Fig. 1-4b as an IC cross-section.

This book is roughly divided into two parts. The first part deals with what we refer to as *subprocess integration* (Chapters 2, 3, and 4), while the second covers *complete process integration* (Chapters 5, 6, 7, and 8). Chapter 9 covers the subject of *process simulation*, an approach applicable to both the subprocess and complete process-integration development efforts.

The function of subprocess integration is to produce device structures that can be used in a variety of integrated-circuit technologies (i.e., isolation, contact, and interconnect structures). The purpose of the process-integration task is to design the process sequences used to manufacture complete IC technologies (i.e., NMOS, CMOS, bipolar, BICMOS, and IC memory devices).

CHAPTER 5

MOS DEVICES AND NMOS PROCESS INTEGRATION

This chapter deals with NMOS process integration. To provide background information for this topic, a review of the physics of long-channel MOS devices (i.e., whose channel lengths are greater than $2\text{ }\mu\text{m}$) and the basics of MOS circuit design is presented. Also included is a discussion of the relationship between desired device performance and process technology. (More rigorous treatments of MOS device physics are found in references 1 - 5.)

The history of NMOS processing is also presented, emphasizing how the obstacles to fabricating reliable NMOS ICs were overcome. A detailed example of the fabrication sequence of a typical NMOS inverter circuit culminates this discussion. The chapter concludes with a description of short-channel and hot-carrier effects in MOS devices, together with processing techniques developed to combat the problems they cause.

5.1 MOS DEVICE PHYSICS

5.1.1 The Structure and Device Fundamentals of MOS Transistors

A perspective view of an *n-channel MOSFET* is shown in Fig. 5-1a, while additional details of its structure and its circuit symbol are given in Figs. 5-1b and 5-1c, respectively. The device has a *gate terminal* (to which the input signal is normally applied), as well as *source* and *drain* terminals (across which the output voltage is developed, and through which the output current flows, i.e., the drain-source current, I_D). The gate terminal is connected to the gate electrode (a conductor), while the remaining terminals are connected to the heavily doped source and drain regions in the semiconductor substrate.

A *channel* region in the semiconductor under the gate electrode separates the source and drain. The channel (of length L and width W) is lightly doped with a dopant type opposite to that of the source and drain. The semiconductor is also physically separated

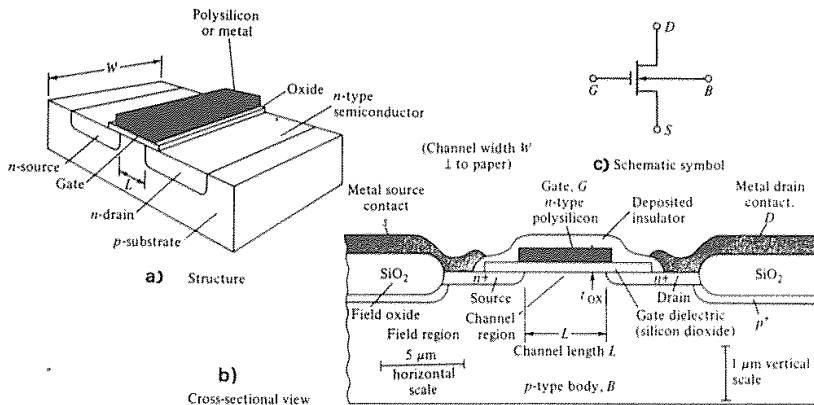


Fig. 5-1 (a) Structure of an MOS device. (b) Cross sectional view. (c) Schematic symbol.⁷
 From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*,
 Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission.

from the gate electrode by an insulating layer (typically, SiO_2), so that no current flows between the gate electrode and the semiconductor.

As shown in Fig. 5-1b, MOS transistors are symmetrical devices, which means that the source and drain are interchangeable. In NMOS circuits, however, *the more positive of these two electrodes is normally defined to be the drain*, and thus the input signal is defined as the voltage between the gate and source terminals ($V_{\text{in}} = V_G$).

In simplest terms, the operation of an MOS transistor involves the application of an input voltage to the gate electrode, which sets up a transverse electric field in the channel region of the device (5-2a). By varying this transverse electric field, it is possible to modulate the longitudinal conductance of the channel region. Since an electric field controls current flow, such devices are termed *field-effect transistors (FETs)*. They are further described as *metal-oxide-semiconductor (MOS) FETs* because of the thin SiO_2 layer that separates the gate and substrate.

The substrate (or body) of the MOS transistor is a silicon wafer; this wafer also provides mechanical support for the finished circuit. In addition, an external electrical connection (or *terminal*) can be made to the body, making the MOS transistor a four-terminal device. (In later sections we will see how the transistor behavior is impacted if a bias is applied between the source and body terminals of the device.)

The top surface of the body consists of *active* or *transistor* regions as well as *passive* or (*field*) regions. The active regions are those in which transistor action occurs; i.e., the channel and the heavily doped source and drain regions. Conduction between separate active regions must be prevented (see chap. 2). A thick oxide layer (0.5 – $1.0 \mu\text{m}$) is normally grown over the field regions as one of the measures to achieve this goal.

If no gate bias is applied, the electrical path between source and drain consists of two back-to-back two pn junctions in series. If a drain bias is applied such that the source-

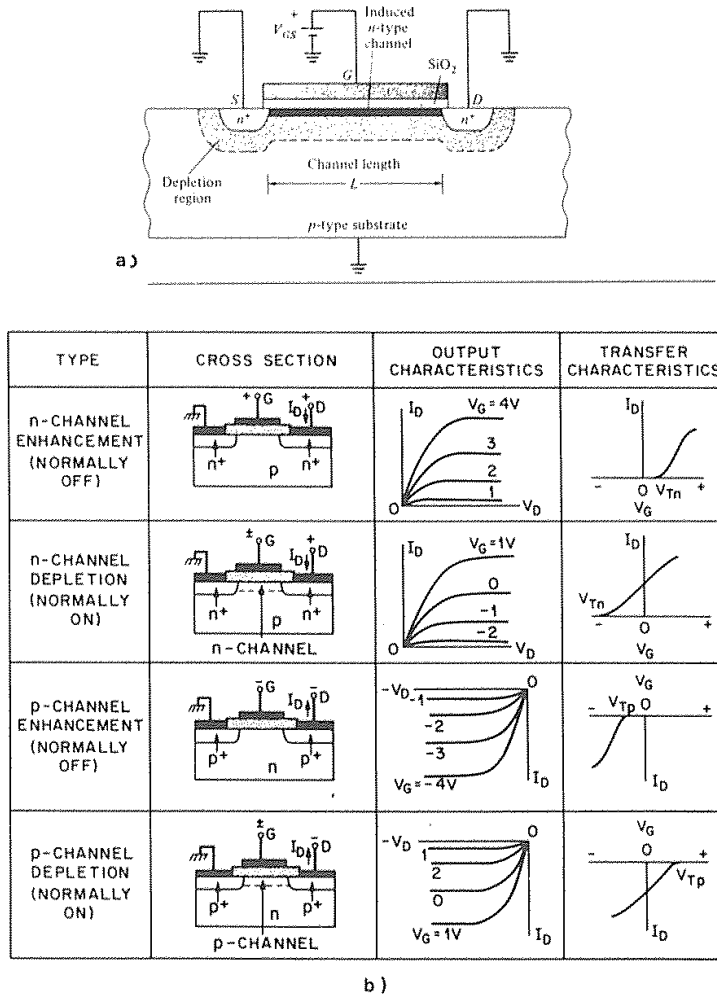


Fig. 5-2 (a) Idealized NMOS cross section with positive V_{GS} applied showing depletion regions and the induced channel.⁷ From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission. (b) Cross sections and output and transfer characteristics of four types of MOSFETs. ⁷ From S. M. Sze, *Semiconductor Devices - Physics and Technology*, Copyright, 1985, John Wiley & Sons. Reprinted with permission.

body and drain-body junctions remain reverse-biased, I_D will consist of only the reverse-bias diode leakage current and hence will be considered negligibly small.

When positive bias is applied to an NMOS transistor gate, electrons will be attracted to the channel region and holes will be repelled. (Holes are the majority carriers in the channel of the p -type body when no gate bias exists.) Once enough electrons have been drawn into the channel by the positive gate voltage to exceed the hole concentration, the region behaves like a n -type semiconductor. Under these circumstances, an n -type channel connects the source and drain regions (Fig. 5-2a). Current will flow if a voltage, V_{DS} , is applied between the source and the drain terminals. 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 $V_G = 0$, are referred to as *enhancement-mode* (or *normally OFF*) transistors (see Fig. 5-2b). With NMOS enhancement-mode transistors, a positive gate voltage, V_G , 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 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 $V_G = 0$ V (see Fig. 5-2b), 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 (that is, the channel is eradicated as long as the bias is applied), such devices are also commonly called *depletion-mode* devices. NMOS depletion-mode devices require a negative gate voltage to be turned *OFF*, while corresponding PMOS devices require positive gate voltages.

5.1.2 The Threshold Voltage of the MOS Transistor

If the source and body of an MOS transistor are both tied to ground ($V_{SB} = 0$), the threshold voltage, V_T , of the transistor can be found from the following equation (note that $V_T = V_{T0}$ when $V_{SB} = 0$):

$$V_{T0} = \phi_{ms} - 2\phi_f - Q_{tot}/C_{ox} - Q_{BO}/C_{ox} \quad (5-1)$$

where ϕ_{ms} is the workfunction difference (in V) between the gate material and the bulk silicon in the channel, ϕ_f is the equilibrium electrostatic potential in a semiconductor (in V), Q_{BO} is the charge stored per unit area (C/cm^2) in the depletion region (when the voltage between source and body is zero), C_{ox} is the gate-oxide capacitance per unit area (F/cm^2), and Q_{tot} is the total positive oxide charge per unit area present at the interface between the oxide and the bulk silicon (see section 5.3.3).

Expressions have been established for these various quantities in terms of doping concentrations in the material, physical constants, device structure dimensions, and temperature. They are:

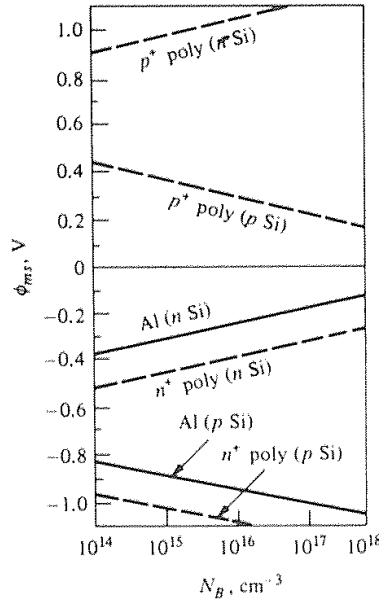


Fig. 5-3 Work-function difference ϕ_{ms} versus doping for degenerate polysilicon and Al electrodes.⁸⁰ Reprinted with permission of Solid-State Electronics.

$$\phi_f = kT/q \ln (n_i / N_A) \quad (p\text{-type semiconductor}) \quad (5 - 2a)$$

$$\phi_f = kT/q \ln (N_D / n_i) \quad (n\text{-type semiconductor}) \quad (5 - 2b)$$

where N_A is the acceptor concentration in a p -type semiconductor (cm^{-3}), N_D is the donor concentration in an n -type semiconductor, and n_i is the intrinsic carrier concentration in the semiconductor. Note that n_i is a strong function of temperature. For silicon at 300 K, however, $n_i = 1.45 \times 10^{10} \text{ cm}^{-3}$. Therefore, at 300 K:

$$\phi_{ms} (\text{metal gate, with Al as gate electrode}) = \phi_f(\text{sub}) - \phi_f(\text{Al}) = \phi_f(\text{sub}) - (+0.6 \text{ V}) \quad (5 - 3a)$$

$$\phi_{ms} (\text{Si gate}) = \phi_f(\text{sub}) - \phi_f(\text{gate}) \quad (5 - 3b)$$

Figure 5-3 shows the value of ϕ_{ms} for various substrate-doping values if an aluminum gate electrode is used. Note also that in silicon-gate NMOS, the doping of the polysilicon gate material is usually n -type, with $N_D \cong 10^{20} / \text{cm}^3$, so that $\phi_f(\text{gate})$ in this case is +0.59 V. Figure 5-4 shows the value of the threshold voltage of n^+ silicon

gate MOS transistors (both PMOS and NMOS) versus substrate doping concentration, assuming either a 65-nm, a 25-nm, or a 15-nm thick gate oxide, and $Q_{\text{tot}} = 0$.

Q_{BO} is found from:

$$Q_{\text{BO}} = -\sqrt{2qN_A\epsilon_{\text{Si}}|-2\phi_f|} \quad \text{NMOS} \quad (5-4a)$$

$$Q_{\text{BO}} = +\sqrt{2qN_D\epsilon_{\text{Si}}|-2\phi_f|} \quad \text{PMOS} \quad (5-4b)$$

and C_{ox} is calculated from:

$$C_{\text{ox}} = \epsilon_{\text{ox}} / t_{\text{ox}} = 3.9 \epsilon_0 / t_{\text{ox}} \quad (5-5)$$

where $\epsilon_{\text{Si}} = 11.8 \epsilon_0$, ϵ_0 is the permittivity of vacuum, and ϵ_{ox} and t_{ox} are the permittivity and thickness of the gate oxide.

It is easy to become confused about the signs of the various terms in the threshold voltage equation. Equation 5-1 gives correct results for NMOS and PMOS if the signs shown in Table 5.1 are used for each of the parameters. When calculating the various terms, the table can be used to insure that the value of each parameter is entered into the equation with the correct sign (e.g., ϕ_{ms} will have a negative value for n^+ Si gate NMOS devices).

Table 5.1 Signs in the Threshold-Voltage Equation⁷

Parameter	NMOS	PMOS
Substrate	<i>p</i> -type	<i>n</i> -type
ϕ_{ms}		
Metal gate	-	-
n^+ Si gate	-	-
p^+ Si gate	+	+
ϕ_f	-	+
Q_{BO}	-	-
Q_{tot}	+	+
γ	+	+
C_{ox}	+	+
Source-body voltage, V_{SB}	+	+

EXAMPLE 5-1: Find the threshold voltage of an NMOS silicon-gate transistor that has substrate doping $N_A = 10^{15}/\text{cm}^2$, gate doping $N_D = 10^{20}/\text{cm}^3$, gate-oxide thickness $t_{\text{ox}} = 100 \text{ nm}$, and $Q_{\text{tot}} = q(1 \times 10^{11}/\text{cm}^2)$. Note that the values for t_{ox} and Q_{tot} are typical of the values that existed in early NMOS transistors.

SOLUTION:

$$\phi_{f(\text{sub})} = kT/q \ln [n_i / N_A] = - 0.026 \ln [10^{15} / 1.4 \times 10^{10}] = - 0.29 \text{ V}$$

$$\phi_{ms} = \phi_{f(\text{sub})} - \phi_{f(\text{gate})} = - 0.29 - kT/q \ln [10^{20} / 1.4 \times 10^{10}] = - 0.88 \text{ V}$$

$$\epsilon_{ox} = 3.9\epsilon_0 = 3.5 \times 10^{-13} \text{ F/cm}; \quad C_{ox} = \epsilon_{ox} / 1 \times 10^{-5} = 35 \times 10^{-9} \text{ F/cm}^2$$

$$Q_{BO} = - [2 \times 1.6 \times 10^{-19} \times 10^{15} \times 1.04 \times 10^{-12} \times | - 0.58 |]^{1/2} = - 1.4 \times 10^{-8} \text{ C/cm}^2$$

$$Q_{BO} / C_{ox} = - 1.4 \times 10^{-8} / 35 \times 10^{-9} = - 0.4 \text{ V}$$

$$Q_{tot} / C_{ox} = 1.6 \times 10^{-19} \times 1 \times 10^{11} / 35 \times 10^{-9} = 0.46 \text{ V}$$

$$V_{T0} = - 0.88 - (- 0.58) - (- 0.4) - (0.46) = - 0.36 \text{ V}$$

Note that a negative value of V_{T0} is yielded, implying that this transistor would be *ON* at $V_G = 0 \text{ V}$ (would therefore behave as a *depletion-mode* device). The parameters used in this calculation of V_T 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 or reliably fabricate enhancement-mode NMOS transistors. We shall later show how this problem was overcome so that enhancement-mode NMOS devices could be manufactured.

5.1.3 Impact of Source-Body Bias on V_T (Body Effect)

As mentioned earlier, the MOS transistor is a four-terminal device, insofar as a contact can also be made to the body region. A bias, V_{SB} , can also be applied between the source and body (e.g., with the source being tied to ground, as shown in Fig. 5-5) and such a bias will have an impact on V_T . If $V_{SB} = 0$, inversion will, of course, occur when the voltage drop across the semiconductor equals $2\phi_f$. If $V_{BS} < 0 \text{ V}$, the semiconductor still attempts to invert when the voltage drop across it reaches $2\phi_f$. However, any inversion-layer carriers that do appear at the semiconductor surface migrate laterally into the source because this region is at a lower potential. Thus, the surface potential must be lowered to $2\phi_f - V_{SB}$ in order for inversion to occur, implying that the threshold voltage required to achieve inversion must be increased.

Hence, for either enhancement-mode or depletion-mode MOSFETs, V_T becomes *more positive* for *n*-channel transistors and *more negative* for *p*-channel transistors as V_{SB} is increased. A simple, quantitative way to view this effect is to assume that biasing changes the inversion point in the semiconductor from $2\phi_f$ to $2\phi_f - V_{SB}$. The threshold-voltage equation (Eq. 5-1), which was applicable when $V_{SB} = 0$, is in turn modified to

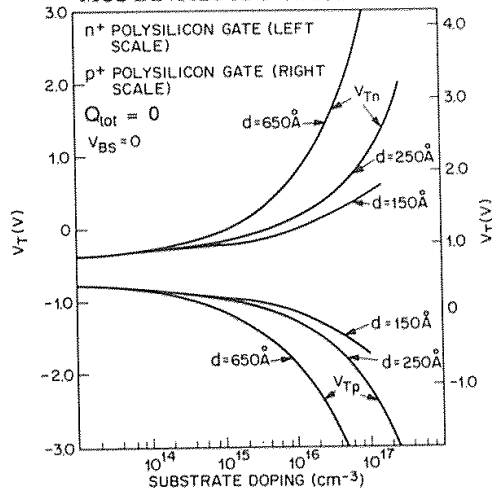


Fig. 5-4 Calculated threshold voltages of n -channel (V_{Tn}) and p -channel (V_{Tp}) transistors as a function of their substrate's doping, assuming n^+ polysilicon gate (left scale) and p^+ polysilicon (right scale). Curves for gate oxide thicknesses d of 150 Å, 250 Å, and 650 Å are shown. From S. M. Sze, Ed., *VLSI Technology*, 2nd. Ed., Chap. 11, "VLSI Process Integration." Copyright, 1988 Bell Telephone Laboratories. Reprinted with permission of McGraw-Hill.

$$V_T = V_{T0} + \gamma \left(\sqrt{|-2\phi_f + V_{SB}|} - \sqrt{2|\phi_f|} \right) \quad (5-6)$$

where the parameter γ (referred to as the *body-effect coefficient* or *body factor*, with units of $V^{1/2}$) is given by

$$\gamma = \frac{\sqrt{2q\epsilon_{si}N_A}}{C_{ox}} \quad \text{NMOS} \quad (5-7a)$$

and

$$\gamma = \frac{\sqrt{2q\epsilon_{si}N_D}}{C_{ox}} \quad \text{PMOS} \quad (5-7b)$$

5.1.4 Current-Voltage Characteristics of MOS Transistors

We will now present the equations that describe the large-signal current-voltage (I - V) characteristics of *long-channel* MOS transistors, assuming an NMOS device with its source grounded and with bias voltages V_{GS} , V_{DS} , and V_{BS} applied as shown in Fig.

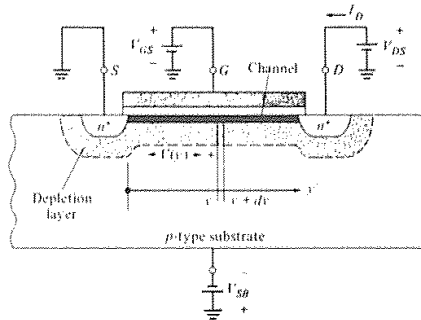


Fig. 5-5 NMOS device with bias voltages V_{GS} , V_{DS} , and V_{BS} applied.⁷ From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission.

5-5. (Any modifications that occur in the I-V characteristics due to short-channel effects will be described section 5.5.)

In the simplest MOS model, if V_G is smaller than V_T , no channel exists, and no current is assumed to flow between the source and drain. If V_G is greater than V_T , a conducting channel is present, and V_{DS} causes a drift current (I_D) to flow from drain to source. For small values of V_{DS} , the drain current I_D is linearly related to V_{DS} . In this so-called *linear region* of operation, the equation that describes I_D is

$$I_D = \frac{k}{2} [2(V_G - V_T) V_{DS} - V_{DS}^2] \quad \begin{matrix} V_G \geq V_T \\ V_{DS} \leq (V_G - V_T) \end{matrix} \quad (5-8)$$

where k (the so-called *device transconductance parameter*) is defined as

$$k = \mu_n C_{ox} (W/L) \quad (5-9)$$

and μ_n is the surface mobility of electrons in the channel.

As the value of V_{DS} is increased, the induced conducting-channel charge decreases near the drain. When V_{DS} equals or exceeds $V_G - V_T$, the channel is said to be pinched off. Increases above this critical voltage produce little change in I_D , and Eq. 5-8 no longer applies. The value of I_D in this region is then given by the following

$$I_D = \frac{k}{2} (V_G - V_T)^2 \quad V_G \geq V_T; \quad V_{DS} > (V_G - V_T) \quad (5-10)$$

This is the so-called *saturation region* of operation.

A plot of I_D versus V_{DS} (with V_G as a parameter) for a long-channel NMOS transistor as described by Eqs. 5-8 and 5-10, is shown in Figure 5-6. If the value of V_G is smaller than V_T , the device is said to be in *cutoff*. In the model given here, I_D is

assumed to be zero in cutoff (i.e., no *subthreshold current*, I_{Dst} , flows). We shall see in a later section that this assumption is not completely correct (even in long-channel devices), and that in short-channel devices such currents can cause severe problems.

5.1.5 The Capacitances of MOS Transistors

It can be shown that the switching speed of MOS digital circuits is not limited by the channel transit time (i.e., the time required for a charge to be transported across the channel), but by the time required to charge and discharge the capacitances that exist between device electrodes and between the interconnecting lines and the substrate. Figure 5-7 shows the significant capacitances between the electrodes of an MOS transistor. The capacitance from gate to other electrodes [C_G] is, to a first approximation, the sum of C_{GB} , C_{GS} , and C_{GD} . Furthermore, since C_{GS} and C_{GD} are small in silicon-gate technology (as the gate and source/drain regions are self-aligned), we can treat C_G as a constant, essentially determined by

$$C_G = W L C_{ox} = W L \epsilon_{ox} / t_{ox} \quad (5-11)$$

The capacitance per unit area of the source/body and drain/body junctions (C_{SB} and C_{DB} , respectively) are calculated using the parallel-plate capacitance formula with a plate spacing of W .¹ The larger the doping of the substrate, the larger the value of these capacitances. As an example, in the case of zero bias and a doping concentration of $10^{16}/\text{cm}^3$, the junction capacitance is approximately $10 \text{ nF}/\text{cm}^2$.

5.2 MAXIMIZING DEVICE PERFORMANCE THROUGH DEVICE DESIGN AND PROCESSING TECHNOLOGY

Having identified the parameters that determine V_T and the I-V characteristics of MOSFETs, we next discuss how to link device design and fabrication procedures in

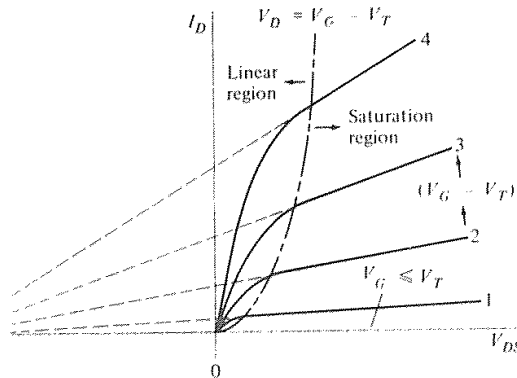


Fig. 5-6 NMOS device with I_D - V_{DS} characteristics.

order to achieve optimum device and circuit performance.

The desired device properties of MOSFETs include the following: high-output current drive (large value of I_D), high and stable transconductance (g_m), predictable and stable threshold voltage (V_T), fast switching speed, very small subthreshold current ($I_{D_{st}}$), high gate-oxide breakdown voltage, high drain/body breakdown voltage, low source/drain-to-body capacitances, high field-region threshold voltage and punchthrough voltage values, and high reliability of device operation. Note that the models used to select the process conditions that provide such optimum device behaviors are Eqs. 5-1, 5-8, and 5-10. In practice, these equations accurately describe the dc circuit behavior of *long-channel MOS devices*. (In section 5.5 we'll see how this behavior is modified in short-channel devices, as well as what device design and processing constraints result.)

5.2.1 Output Current (I_D) and Transconductance (g_m)

Equations 5-8 and 5-10 predict how I_D in MOSFETs can be impacted by various device parameters. Since increasing the *gate width* linearly increases I_D , one option when large drive-currents are needed, is to increase the dimension of the gate width. However, when minimum-sized devices must be used (e.g., for maximum packing density), this option cannot be implemented, and the other parameters that can influence I_D in Eqs. 5-8 and 5-10 must be considered.

From the dependence of I_D on $k' = \mu C_{ox}$, both the mobility of the carriers in the channel and the gate oxide capacitance should be as high as possible. Since electron mobility is greater than hole mobility, circuits using NMOS devices will exhibit higher performance than those built with PMOS devices. In fact, *NMOS transistors of the same width as PMOS transistors will indeed provide roughly 2.5x the current drive*. In addition, the mobility of carriers decreases as the doping concentration of the channel increases. Hence, lightly doped channel regions are also favored.

Because the value of C_{ox} is inversely proportional to t_{ox} , as thin a gate oxide as possible (commensurate with oxide breakdown and reliability considerations) is normally used. I_D is also inversely proportional to channel length, and minimum channel lengths are therefore desirable. On the other hand, if the channel lengths become too short, they adversely impact device operation in other ways, as described in section 5.5.

Although the equations indicate that $(V_G - V_T)$ should be as large as possible, V_G is usually fixed by system specifications and material limitations, and cannot be changed by the device or circuit designer. Similarly, V_T is selected primarily by other circuit considerations (such as adequate noise margin in digital circuits).

The transconductance in saturation, $g_{msat} = dI_D/dV_G$, can be expressed simply by way of the following equation

$$g_{msat} = 2 \mu_n C_{ox} (W/L) [V_G - V_T] \quad (5 - 12)$$

Hence, we can maximize g_{msat} by varying the process and device parameters in the same manner as discussed for I_D .

5.2.2 Controlling the Threshold Voltage through Process and Circuit Design Techniques

In many MOS IC applications, it is critical to be able to establish and maintain a uniform and stable value of V_T (i.e., the value of V_T should not vary with time or with device-operating conditions). An example of the importance of being able to control this parameter involves semiconductor memory devices. In these circuits, charge flows from the memory cells to the sense amplifier. The sense amplifier is a delicately balanced flip-flop whose voltage-sensing capability is directly related to the threshold-voltage variation between the transistors (see chap. 8). Hence, such circuits would not function reliably without the presence of a highly uniform and stable threshold voltage in the circuit devices.

The factors that impact threshold voltage (when $V_{SB} = 0$) are given in Eq. 5-1. Examining each term of this equation will enable us to determine which device parameters can be adjusted to provide practical control of V_T .

The ϕ_{ms} term depends on the work-function difference between the gate, $\phi_f(\text{gate})$, and the semiconductor, $\phi_f(\text{sub})$. For metal and heavily doped silicon gates, $\phi_f(\text{gate})$ is constant while the parameter $\phi_f(\text{sub})$ 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, or ~ 0.06 V ($kT/q = 0.026$ V at 300 K). 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 ($2\phi_f$) 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 $2\phi_f$ term is also not of much use in controlling V_T .

Since every attempt is made to keep Q_{tot} as low as possible through various processing procedures, the Q_{tot}/C_{ox} term is very small in modern MOS devices. Hence, it must also be ruled out as a candidate for controlling V_T .

While it is true that C_{ox} can be varied (primarily by changing the gate oxide thickness), this parameter is not a practical vehicle for controlling V_T in active devices, since the gate oxide is normally made as thin as possible to achieve maximum I_D . In the field regions, however, large V_T values are needed to prevent inversion under the field oxide. A thick field oxide makes C_{ox} small, allowing V_T to be increased. Thus, the C_{ox} is one of the parameters normally used to control V_T in the field regions of the circuits.

This leaves the Q_{BO} term as the remaining candidate for controlling V_T in active devices. Equation 5-4 indicates that the doping concentration can indeed provide a large change in V_T . The signs of Eq. 5-1 indicate that increasing the substrate doping (i.e., N_A in n -channel devices, and N_D in p -channel devices) will increase V_T . Thus, to increase V_T in Example 5-1 in this manner, it would be necessary to increase the substrate doping concentration above $N_A = 10^{15} \text{ cm}^{-3}$. Figure 5-7 shows how V_T can be controlled by changing the substrate doping concentration for various gate-oxide thicknesses (assuming that Q_{tot} is kept small enough that it is not a significant contribution to V_T).

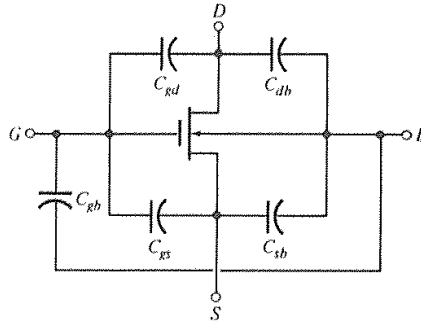


Fig. 5-7 MOS transistor capacitances. $C_G = C_{GS} + C_{GB} + C_{GD}$.⁷ From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission.

Significant increases in substrate-doping concentration give rise to lower junction-breakdown voltages, larger junction capacitances, and lower carrier mobilities, making such substrate doping concentration increases undesirable. 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.

A *circuit approach* known based on applying a bias between the source and body (and hence known as *body biasing*) can also control V_T . When body biasing is used V_T is given by Eq. 5-6. Figure 5-8 gives an example of how the application of V_{SB} can change the V_{T0} value in an NMOS device.¹⁴ The intercept of the channel conductance and the V_G axis when $V_{SB} = 0$ corresponds to V_{T0} , which in this example is +0.99 V.

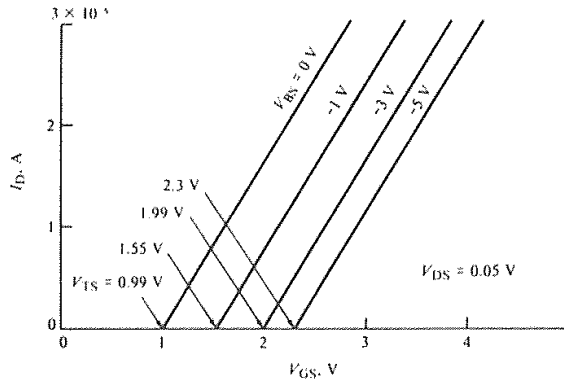


Fig. 5-8 Threshold voltage adjustment using substrate bias.

As V_{SB} is increased from 0 V to -5 V, V_T increases up to +2.3 V. However, body biasing is an added complication, and it is now avoided as a technique for adjusting V_T whenever possible in favor of the newer *ion implantation V_T -adjust method*.

We nevertheless include a discussion of body biasing because it pre-dated ion implantation as a method of controlling V_T . C. T. Sah points out that the availability of the technique allowed IBM to implement MOS memory devices instead of core memory for the first time in the IBM-370/158 mainframe computer in 1973.⁶⁰ The access time of NMOS RAMs at that time was competitive with magnetic core memory ($\sim 1 \mu s$), whereas PMOS RAMs were slower. Body biasing allowed the higher-performance NMOS devices to function as enhancement-mode devices, even though their threshold voltages without body biasing would have caused them to behave like depletion-mode devices.

Another important reason for describing body biasing is that the body effect impacts devices in MOS IC logic circuits even when no deliberate attempt is made to apply an external bias. For example, when a logic gate containing a transistor as the load device has a logic 1 output (see for example Fig. 5-13), the voltage at the source of this device is different from that of the body. Hence, the device is subject to a non-zero V_{SB} , and V_T no longer equals V_{T0} value. The degree to which the change occurs depends on the *body factor*, γ . Since the smallest change in V_T is generally desired, small values of γ are preferred. Equation 5-7 indicates that γ depends on substrate doping, and that lower doping concentrations provide smaller values of γ , providing further impetus for lightly doped substrates in MOS ICs.

The value of γ also decreases with channel length below $1 \mu m$. For example, γ in NMOS devices decreases by about 50% as L_{eff} is decreased from $1 \mu m$ to $0.6 \mu m$, while for PMOS the decrease is about 30%.⁸

5.2.3 Subthreshold Currents (I_{Dst} when $V_G < |V_T|$)

The basic MOS device model neglects all free charges in the channel until the magnitude of the gate voltage exceeds V_T . This is a valid approximation for most subthreshold bias conditions because the free-charge densities in the channel change exponentially with the channel voltage. When this approximation is used, it implies that *no current flows between the drain and source* if $V_G < |V_T|$. As V_G approaches the value of V_T , (corresponding to a condition of *weak inversion*), however, the magnitude of I_D is not well defined by this simple approach. In fact, it is observed that $I_D \neq 0$ if V_G is close to (but still less than) the value of V_T . The current which is observed in such cases is therefore referred to as *subthreshold current*, I_{Dst} .

The values of I_{Dst} can be well predicted in long-channel MOSFETs by modifying the basic MOS model to take into account the fact that the minority carrier concentration at the Si surface is greater than the value at equilibrium, but still less than the bulk doping concentration.¹⁹ The results of this modified model indicate that the magnitude of I_{Dst} is essentially independent of V_{DS} but is exponentially proportional to V_G .

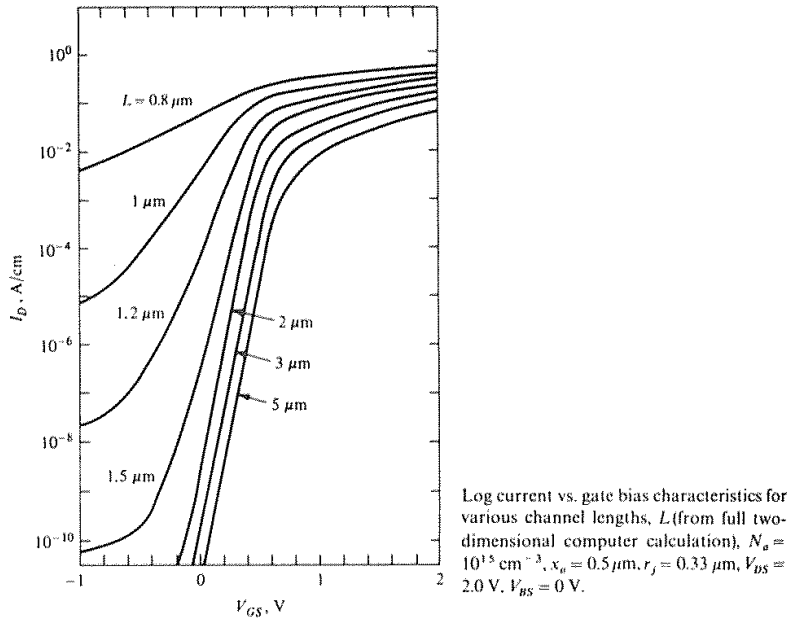


Fig. 5-9 Example of subthreshold MOSFET $I_D - V_{GS}$ curves for various channel lengths.¹⁹ (Reprinted with permission of Solid State Electronics).

Since I_{Dst} can be accurately modeled in long-channel MOSFETs (Fig. 5-9), circuit designers can readily calculate the gate bias required to ensure a given allowable subthreshold leakage current. *Typically, to ensure that I_{Dst} will be negligibly small, the bias applied to the gate should be 0.5 V below V_T .*

Since I_{Dst} is exponentially proportional to V_G , if $\log I_{Dst}$ is plotted versus V_G the result on a semilogarithmic plot will be a straight line for values of V_G below V_T . The slope of the I_{Dst} versus V_G line, when plotted in this manner, is characterized by the *subthreshold swing*, S.S., where

$$\text{S.S.} = \Delta V_G / \Delta \log I_{Dst} \quad (5-13)$$

Hence, S.S. is the change in V_G that produces a decade increase in I_{Dst} . A small value of S.S. is desirable, since it indicates maximum control of the gate over the channel current. Typical values of subthreshold swing in long-channel MOSFETs are around 90 mV/decade. Such values can be achieved by building devices that have a low value of substrate doping concentration (which gives a larger depletion width) and a thin gate oxide. (A more detailed treatment of the derivation of the I_{Dst} dependence on V_G can be found in references 1 and 3.)

Note that when the channel length gets small, the values of I_{Dst} are larger than those predicted by the modified long-channel MOSFET model. This is due to so-called *short-channel effects* (discussed in section 5.5). However, measurements of the subthreshold swing can be used to detect the onset of these short-channel effects (*punchthrough* and *drain-induced barrier lowering*). Since the measurable I_{Dst} is the sum of both the normal subthreshold and the short-channel subthreshold current components, an increase in the value of S.S. will signal the onset of these effects.

5.2.4 Switching Speed

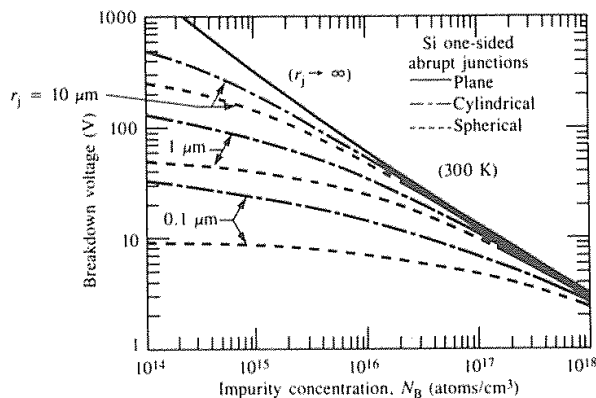
The switching speed of the logic gates in an MOS IC is limited only by the time required to charge and discharge the capacitances between device electrodes and between the interconnecting lines and ground (or other lines). At the circuit level the propagation delay is frequently limited by the interconnection-line capacitances. At the device level, however, the gate delay is determined primarily by the channel transconductance, the MOS gate capacitance (C_G) and the other two MOS parasitic capacitances, C_{DB} and C_{SB} (as defined in the previous section). If these capacitance values can be reduced, the device switching speed will be increased.

The gate capacitance is decreased by decreasing the gate area (although decreasing the gate oxide thickness increases its value). The dominant parasitic capacitance on the device level, however, is that due to C_{SB} and C_{DB} (i.e., junction capacitances). An analytical study has shown that these junction capacitances account for up to 50% of the total capacitance in logic gates.⁹ Therefore, reductions in these capacitances should produce corresponding decreases in the gate delay. In general, *the lower the doping concentration in the body, the lower the junction capacitances.*

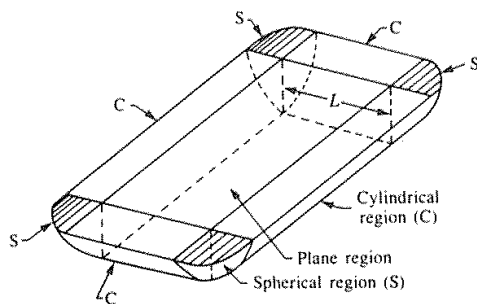
5.2.5 Junction Breakdown Voltage (Drain-to-Substrate)

The source and drain regions are very heavily doped to minimize their resistivities. Thus, the breakdown voltage of the drain-to-substrate junction will be determined by the lighter doping concentration of the body. As seen in Fig. 5-10a (which shows breakdown voltage of a one-sided *pn* junction as a function of the lighter doping concentration), the breakdown voltage decreases as the doping increases. Thus, *lightly doped substrates also yield high junction breakdown voltages.*

Junction curvature enhances the electric field in the curved part of the depletion region (Fig. 5-10b), and this effect reduces the breakdown voltage below that predicted by one-dimensional junction theory. A rectangular source or drain region (formed either by diffusion or ion implantation) has regions with both cylindrical and spherical curvature.^{10,11} Figure 5-10a also shows the effect of junction curvature on the breakdown voltage of a one-sided step junction in silicon. It can be seen that as the junction depths get shallower (making the radius of curvature, r_j , of the spherical and cylindrical structures smaller), the breakdown voltage is significantly reduced, especially for low substrate impurity concentrations.



(a)



(b)

Fig. 5-10 (a) Abrupt pn junction breakdown voltage versus impurity concentration on the lightly doped side of the junction for both cylindrical and spherical structures. r_j is the radius of curvature.¹⁰ (© 1957 IEEE). (b) Formation of cylindrical and spherical regions by diffusion through a rectangular window. From S. M. Sze, *Semiconductor Devices - Physics and Technology*, Copyright, 1985, John Wiley & Sons. Reprinted with permission.

5.2.6 Gate-Oxide Breakdown Voltage

High-quality SiO_2 films will typically break down at electric fields of 5-10 MV/cm (the exact value is a function of oxidation and anneal conditions, oxide charges, surface crystallographic orientation, surface preparation and a number of other factors). This corresponds to 50-100 V across a 100-nm-thick oxide. Present day 5 V processes use gate-oxide thicknesses of 15-100 nm. Below around 5 nm (at less than 3 V), there is a

finite probability that electrons will pass through the gate by means of a quantum-mechanical tunneling effect. For proper device operation, the tunneling current must be small. This effect therefore sets a fundamental lower limit of about 5 nm for the thickness of the gate oxide. A search for alternative gate dielectric materials to mitigate this limitation is being conducted; this issue will be discussed in the CMOS chapter.

Oxide breakdown may also occur at electric-field values smaller than those given above, as a result of process-induced flaws in the gate oxide. Such defects include: metal precipitates on the silicon surface prior to oxide growth (see Vol. 1, chap. 2); high defect density in the silicon lattice at the substrate surface (e.g., stacking faults and dislocations, see Vol. 1, chap. 2); pinholes and weak spots created in the gate oxide by particulates; thinning of the oxide during growth caused by the Kooi effect (see chap. 2); and oxide wearout due to failure mechanisms related to hot-electron injection (this topic will be covered in Vol. 3).

5.2.7 High Field-Region Threshold-Voltage Value

A high value of threshold voltage in the field region is needed to keep the parasitic field channels between adjacent active devices from being turned on. This topic is described in great detail in chapter 2.

5.3 THE EVOLUTION OF MOS TECHNOLOGY (PMOS AND NMOS)

Following the introduction of MOS integrated circuits in the 1960s, MOS technology evolved through several stages. At first, PMOS was the dominant technology, but it was supplanted by NMOS in the early 1970s. NMOS in turn was largely replaced by CMOS in the mid-1980s. (The evolution of CMOS will be described in chap. 6.) In addition, the drive to shrink feature dimensions led to smaller device sizes in all IC technologies. In MOS technology, the main dimension that summarized the shrinkage in each new generation was the *minimum gate length* of the transistors. NMOS overtook PMOS roughly when gate lengths reached $\sim 6 \mu\text{m}$, while CMOS became the dominant technology at gate lengths of $\sim 1.5\text{-}2 \mu\text{m}$.

Although the *physical gate length*, L , of the MOS device is commonly used to identify each generation of technology, this practice can be misleading in that it may convey the impression that this dimension is also the minimum dimension of all other device features fabricated in the technology. In fact, the physical gate length does not necessarily reflect the dimensions of the other design rules (many, if not most, of which are larger than the physical gate length dimension).

In addition, the physical gate length does not accurately represent the electrical or *effective* channel length (L_{eff}), which in fact is given by

$$L_{\text{eff}} = L - 2x_j \quad (5 - 14)$$

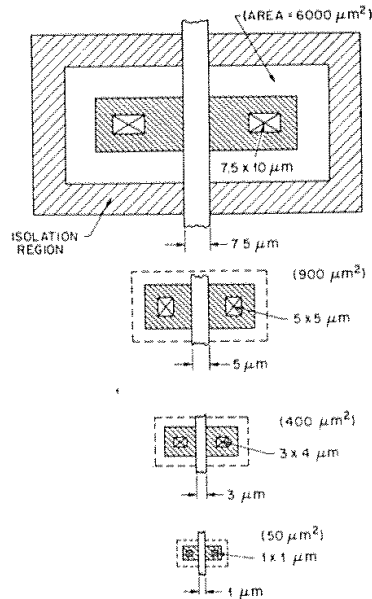


Fig. 5-11 Reduction in the area of the MOSFET as the gate length (minimum feature length) is reduced.¹² Reprinted with permission of Solid State Technology.

where x_{jl} is the lateral distance that the source or drain junction extends under the gate. For example, in a $1.2 \mu\text{m}$ MOS technology, the source and drain junctions both extend $\sim 0.125 \mu\text{m}$ under the gate region, so L_{eff} in this technology is actually $0.95 \mu\text{m}$. Furthermore, in CMOS technologies, the minimum physical length of PMOS devices is generally longer than that of NMOS devices, for reasons that will be discussed later. Figure 5-11 shows the reduction in MOS area as the gate length is reduced.¹² A longer and more detailed history of the evolution of the MOS transistor has been published by C. T. Sah, who describes both the evolution of the field-effect devices themselves and the technology that has been developed to fabricate these devices.⁶⁰

5.3.1 Aluminum-Gate PMOS

The first MOS ICs built in the mid-1960s were implemented with *p*-channel enhancement-mode devices whose threshold voltages were approximately -4 V . (An example of an early PMOS IC was the Intel 256-bit SRAM, introduced in the late-60s with devices having gate lengths of $12 \mu\text{m}$.) Such early MOS integrated circuits were built on silicon wafers of $\langle 111 \rangle$ orientation and used Al as the gate electrode material. These choices grew out of the experience gained in manufacturing bipolar ICs prior to the development of MOS ICs. When MOS IC technology was first being implemented,

semiconductor manufacturers transferred their bipolar fabrication know-how to the newer technology.

Since precise control of dopant diffusion in $\langle 111 \rangle$ Si was a mature bipolar process, $\langle 111 \rangle$ wafers were also the logical choice for building MOS ICs. Similarly, since aluminum metallization was already being implemented, it was natural to adapt Al as the MOS gate electrode material. Unknowingly, these choices worsened the problems which prevented NMOS devices from being used to produce early MOS ICs, and instead, forced the use of PMOS devices in these circuits.

The earliest MOS circuits exhibited two serious limitations, partly as a result of being implemented with p -channel devices. First, the PMOS devices had a V_T of -4 V, which required a power-supply voltage of -12 V for the drain supply (a value that was incompatible with the $+5$ V power-supply voltage used in bipolar digital [TTL] ICs). Second, the circuits were very slow (e.g., a PMOS flip-flop could operate at 500 kHz to 1 MHz, while a bipolar flip-flop could operate at 5-10 MHz). It was known that the latter problem was due to the low surface mobility of holes in the channel and that electron mobility in silicon is nearly three times as large. Therefore, NMOS circuits would have been able to provide significantly improved performance. Nevertheless, the decision to manufacture ICs with PMOS devices was dictated primarily by the existence of large (and quite variable) oxide-charge densities in the early MOS technologies.

These oxide charges were generally positive, and positive voltages on the gate tend to accumulate n -type surfaces (but will deplete or invert p -type surfaces). The oxide charges present in early MOS devices were often large enough to cause inversion in NMOS devices fabricated with reasonably thin gate oxides, even when $V_G = 0$. A thicker gate oxide could have been used to increase V_T , but this would also have degraded g_m to such a degree that circuits built with such devices would have been even slower than those built with PMOS devices. Thus, only *depletion-mode* high performance NMOS devices could be reliably manufactured. Since *enhancement-mode* devices are needed for most applications, this presented a major difficulty.

In addition, because the oxide charges were also capable of inverting p -doped substrate regions under field oxides, it was difficult to reliably isolate n -channel devices. This problem was made worse by the depletion of boron during thermal oxidation, because it reduced the boron concentration at the p -type surface. In summary, at the outset of the MOS era, it was not possible to reliably manufacture integrated circuits with n -channel enhancement-mode MOSFETs.

On the other hand, since positive oxide charges tend to accumulate an n -type surface, they merely increase the negative voltage that is required to turn on a p -channel device. Hence, the manufacture of enhancement-mode PMOS devices was possible despite the presence of the oxide charges. Nevertheless, it was clear that processing innovations were needed if the potential benefits of MOS (i.e., increased packing density, lower power consumption, TTL power-supply voltage compatibility, and process simplicity) were to be fully realized.

5.3.2 Silicon-Gate MOS Technology

One of the key process innovations for MOS ICs was the use of heavily doped polysilicon as the gate electrode in place of Al (see Vol. 1, chap. 6 for more information on polysilicon thin films). The development of this *silicon gate technology* improved the fabrication of MOS ICs in the following ways:

- Since aluminum must be deposited following completion of all high-temperature process steps (including drive-in of the source and drain regions), the gate electrode must be separately aligned to the source and drain. This alignment procedure adversely affects both packing density and parasitic overlap capacitances between the gate and source/drain regions. Since polysilicon has the same high melting point as the silicon substrate, it can be deposited prior to source and drain formation. Furthermore, the gate itself can serve as a mask during formation of the source and drain regions (by either diffusion or ion implantation). The gate thereby becomes nearly perfectly aligned over the channel, with the only overlap of the source and drain being that due to lateral diffusion of the dopant atoms. This *self-alignment* feature simplifies the fabrication sequence, increases packing density, and reduces the gate-source and gate-drain parasitic overlap capacitances.
- The threshold voltage of PMOS devices is reduced by the use of a polysilicon gate, since the ϕ_{ms} is less negative (see Eqs. 5-2). For PMOS devices on <111>-Si, the threshold voltage is reduced from roughly -4 V to -2 V. This smaller threshold voltage value enabled PMOS ICs to become compatible with TTL (bipolar) ICs, allowing MOS to be designed into many digital systems that operated at TTL-defined power supply voltage levels (i.e., 0 V to 5 V).
- The ability of polysilicon to withstand high temperatures also permits it to be completely encapsulated by an SiO₂ layer. This allows the polysilicon film to function as an interconnect path, in addition to serving as the gate electrode. By taking advantage of this new interconnection structure (without having to use a second layer of metal, as was necessary with bipolar ICs), it was possible to give MOS ICs an additional level of interconnection that could be crossed by the usual metal layer, or even by another polysilicon layer. This eased the problem of routing the electrical paths among the devices of an IC, thereby facilitating the layout of compact digital integrated circuits. (Techniques for establishing contact between the polysilicon layer and substrate are described in chap. 3, section 3.11.1). The ability of polysilicon to withstand high temperatures was also exploited to allow the dielectric (e.g. phosphorus-doped SiO₂) that covers it to be flowed, thereby making a significantly smoother surface topography for metallization layers.

The greatest disadvantage of polysilicon as a gate material compared to Al is its significantly higher resistivity. Even when doped at the highest practical

concentrations, a 0.5- μm -thick polysilicon film has a sheet resistance of about 20 Ω/sq (compared to ~ 0.05 Ω/sq for a 0.5- μm -thick Al film). The resulting high values of interconnect line resistance can lead to relatively long RC time constants (i.e., long propagation delays) and severe dc voltage variations within a VLSI circuit. Consequently, the formation of refractory metal silicide layers on top of polysilicon layers (which results in so-called *polycide* films) was developed to reduce the severity of this drawback. Such polycide films can provide sheet resistances of 1 Ω/sq , at the expense of more complex processing (see Vol. 1, chap. 11 for more information on polycides). Despite of the above limitation, the development of silicon-gate technology proved to be the most important contribution to MOS technology during the reign of PMOS.

5.3.3 Reduction of Oxide-Charge Densities

Another set of important advances allowed the magnitude of the positive oxide-charge densities to be reduced. These oxide charge reduction techniques can be summarized as those involving *cleanliness, gettering, annealing, and replacing <111> wafers with <100> wafers*.

Cleanliness and gettering techniques reduced the densities of *mobile ionic charge*, which are due to the incorporation of ionized alkali metal atoms (Na^+ , K^+) in the gate oxide. Na contamination can be controlled through clean gate oxidation processing. Gettering is used to prevent any Na^+ that enters the gate oxide from significantly degrading the V_T (see Vol. 1, chap. 7). Although tests for Na contamination of MOS devices must be routinely performed to ensure that accidental contamination does not occur, it is possible to establish fabrication procedures in which the instability of V_T due to mobile ionic charge is less than 0.1 V.

Another source of positive oxide charge is the *interface trap charge*. Again, as described in Volume 1, chapter 7, two annealing techniques were discovered that reduced this charge to acceptably low levels (i.e., to the low $10^{10} \text{ cm}^{-2} \text{ eV}^{-1}$ range). At these levels the contribution to the threshold voltage of the device is acceptably small for modern MOS devices.

The final source of positive oxide charge is the so-called *fixed oxide charge* which is located in the transition region between the silicon and the SiO_2 (see Vol. 1, chap. 7). It was learned that this charge can be reduced by the use of proper annealing techniques and that the lowest fixed oxide charge densities ($\sim 10^{10} \text{ cm}^{-2}$) are obtained on <100> wafers. As a result, the production of MOS ICs was shifted from <111> to <100> starting material.

In summary, the threshold voltage is negatively shifted by an amount proportional to the sum of these three oxide charge densities, Q_{tot} . When Q_{tot} was quite high this had an extremely important influence on MOS production. Through an enormous effort by the semiconductor community (industry and universities), each of the positive oxide charge density values was reduced. As a result, Q_{tot} can now be kept to less than 5×10^{10} charges/ cm^2 , and in currently used MOS devices, the oxide charge contribution to

threshold voltage is minimal. As an example, the change in threshold voltage (ΔV_T) due to a Q_{tot} of q ($5 \times 10^{10} \text{ cm}^{-2}$) in an MOS device with a 20-nm-thick oxide is:

$$\begin{aligned}\Delta V_T &= Q_{\text{tot}}/C_{\text{ox}} = Q_{\text{tot}} x_o / 3.9 \epsilon_o \\ &= (1.6 \times 10^{-19} \text{ C}) (5 \times 10^{10} \text{ cm}^{-2}) (2 \times 10^{-6} \text{ cm}) / (3.5 \times 10^{-13} \text{ F/cm}) \approx 0.05 \text{ V}.\end{aligned}$$

The primary reason that thermal SiO_2 is used as the gate insulator in almost all MOSFETs is that it exhibits the best interface with silicon (where "best" means that the interface has a very low concentration of interface fixed charges and traps). In fact, if a phenomenon such as the tying up of the dangling silicon bonds at the silicon surface by SiO_2 did not exist, and if an annealing process was not discovered for reducing the remaining bonds and traps to an acceptable level, MOS devices would have remained merely a laboratory curiosity.⁶¹ Experimentation, however, is still being conducted to determine the suitability of other materials as gate insulators; this is discussed in chapter 6 which deals with CMOS.

EXAMPLE 5-2: Recalculate the threshold voltage of the NMOS transistor considered in Example 5-1 of section 5.1.2 when the oxide thickness, t_{ox} , is reduced to 15 nm, and the total oxide-charge density is reduced to $5 \times 10^{10} \text{ cm}^{-2}$.

SOLUTION: $\phi_{f(\text{sub})} = -0.29 \text{ V}; \quad \phi_{\text{ms}} = -0.88 \text{ V};$

$$\epsilon_{\text{ox}} = 3.9\epsilon_o = 3.5 \times 10^{-13} \text{ F/cm}; \quad C_{\text{ox}} = \epsilon_{\text{ox}} / 15 \times 10^{-6} = 2.3 \times 10^{-7} \text{ F/cm}^2$$

$$Q_{\text{BO}} = - [2 \times 1.6 \times 10^{-19} \times 10^{15} \times 1.04 \times 10^{-12} \times |-0.58|]^{1/2} = -1.4 \times 10^{-8} \text{ C/cm}^2$$

$$Q_{\text{BO}}/C_{\text{ox}} = -1.4 \times 10^{-8} / 2.3 \times 10^{-7} = -0.06 \text{ V}$$

$$Q_{\text{tot}}/C_{\text{ox}} = 1.6 \times 10^{-19} \times 5 \times 10^{10} / 2.3 \times 10^{-7} = 0.46 \text{ V}$$

$$V_{\text{TO}} = -0.88 - (-0.58) - (-0.06) - (0.03) = -0.27 \text{ V}$$

Note that the threshold voltage is still negative, and that this device would still be *ON* if the applied gate bias was $V_G = 0 \text{ V}$.

EXAMPLE 5-3: Repeat the threshold-voltage calculation in Example 5-2 for an NMOS transistor whose oxide thickness (t_{ox}) is increased to 500 nm. This oxide thickness is typical of the field-oxide thickness between MOS devices on an IC. Hence, it allows us to calculate V_T for the parasitic NMOS field-region device.

SOLUTION: $\phi_{f(\text{sub})} = -0.29 \text{ V}; \quad \phi_{\text{ms}} = -0.88 \text{ V};$

$$\epsilon_{\text{ox}} = 3.9\epsilon_o = 3.5 \times 10^{-13} \text{ F/cm}; \quad C_{\text{ox}} = \epsilon_{\text{ox}} / 5 \times 10^{-5} = 7 \times 10^{-9} \text{ F/cm}^2$$

$$Q_{BO} = - [2 \times 1.6 \times 10^{-19} \times 10^{15} \times 1.04 \times 10^{-12} \times |-0.58|]^{1/2} = -1.4 \times 10^{-8} \text{ C/cm}^2$$

$$Q_{BO}/C_{ox} = -1.4 \times 10^{-8} / 7 \times 10^{-9} = -2 \text{ V}$$

$$Q_{to}/C_{ox} = 1.6 \times 10^{-19} \times 5 \times 10^{10} / 7 \times 10^{-9} = 1.14 \text{ V}$$

$$V_{TO} = -0.88 - (-0.58) - (-2) - (1.14) = 0.66 \text{ V}$$

This shows that an increase in the oxide thickness in the field will increase the V_T of the NMOS device in Example 5-2 so that it is now an enhancement-mode device. Unfortunately, if $V_G = 5 \text{ V}$, this device would still turn on, and thus the parasitic field device would conduct.

5.3.4 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, because this procedure made it possible to select the substrate-doping value without having to consider its impact on V_T . Substrate doping could now be selected strictly on the basis of optimum device performance, since V_T became separately adjustable by means of ion implantation. In addition, since dopants could be selectively implanted into the field regions, high-performance NMOS circuits could also be reliably fabricated on lightly doped substrates (i.e., without the possibility of inadvertent inversion of the surrounding field regions).

This technique of adjusting V_T involves implantation of boron, phosphorus, or arsenic ions into the regions under the oxide of a MOSFET. The implantation of boron causes a positive shift in the threshold voltage, while phosphorus or arsenic implantation causes a negative shift. For shallow implants, the procedure has essentially the same effect as placing an additional "fixed" charge at the oxide-semiconductor interface. To first order, the threshold-voltage change (ΔV_T) is thereby estimated from¹

$$\Delta V_T = q N_I / C_{ox} \quad (5-15)$$

where N_I is the dose of the implanted ions (atoms/cm²) introduced into the silicon near its surface. For example, Eq. 5-15 predicts that when $N_I = 5 \times 10^{11}$ ions/cm² and $t_{ox} = 25 \text{ nm}$, a shift in V_T of 0.58 V will be produced. Exact modeling can be performed to calculate the actual threshold voltage shift more accurately. Figure 5-12 graphically shows the results of such modeling calculations.^{16,75}

The V_T -adjust implant is usually done through the gate oxide layer. When the correct implant energy for the gate-oxide thickness being used is selected, the peak of the implant will occur at the oxide-silicon interface. 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 distribu-

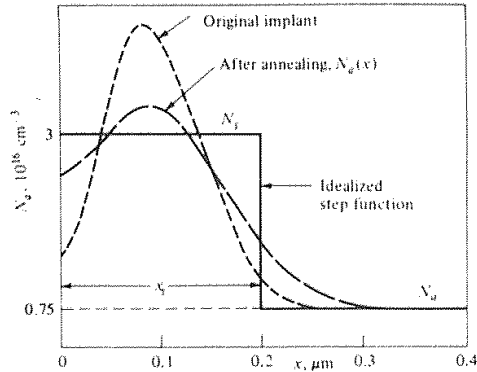


Fig. 5-12 Doping profile of implanted region beneath the gate oxide. The original implant is broadened by thermal annealing. The step doping is used to estimate the threshold voltage shift achieved using ion implantation.⁷⁵ Reprinted with permission of IBM J. of Research and Development.

tion via a "box" distribution (in which the implanted dopant is assumed to have a constant density from the surface to a depth of x_i .) Figure 5-12 shows that V_T does not change greatly as x_i is varied. Thus, the first-order approximation of the threshold-voltage change, ΔV_T (which ignored the depth of the implant as a parameter), will therefore give reasonably good estimates of ΔV_T . This calculation is often accepted in practice.

Ion implantation can also be used to fabricate depletion-mode MOSFETs. Depletion-mode NMOS devices (*i.e.* in which $V_T < 0$ V) are commonly used in NMOS logic circuits (see the following section). In order for the required negative threshold voltage for a depletion-mode NMOS device to be produced, n -type impurities are implanted 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. 5-1 and 5-15.

EXAMPLE 5-4: If the NMOS transistor of Example 5-2 is to be used in an application that requires a $V_T = 1.0$ V, calculate the boron dose needed to adjust V_T to this value.

SOLUTION: The threshold voltage of the device in Example 5-2 is -0.27 V. We wish to have a V_T of 1.0 V. Thus V_T must be shifted (ΔV_T) by 1.27 V. Using Eq. 5-15, we see that the boron dose needed to cause this ΔV_T is

$$\begin{aligned} N_I &= \Delta V_T C_{ox}/q = [1.27 \text{ V} \times 2.3 \times 10^{-7} \text{ F/cm}^2] / 1.6 \times 10^{-19} \text{ C} \\ &= 1.8 \times 10^{12} \text{ boron atoms/cm}^2. \end{aligned}$$

5.3.5 Isolation Technology for MOS

Although MOS transistors are inherently self-isolating devices it is still necessary to prevent the formation of spurious channels between MOS devices (see chap. 2). This can be accomplished with the combination of a thick field oxide between the devices and a high surface concentration under the field oxide.

Prior to about 1970 the process for obtaining thick oxide regions in the field involved growing an oxide to the desired thickness on the wafer surface and then etching windows into it. This approach caused severe steps in the wafer topography, which were difficult to cover with subsequent metal layers. The introduction of LOCOS isolation in 1970 substantially overcame this problem (see chap. 2). The smoothly tapered step from the edge of the active region to the top of the field oxide in LOCOS permits overlying conductors to be easily deposited on such steps without the occurrence of significant thinning. In addition, with the development of the threshold-voltage adjustment process via ion implantation, high surface concentrations of boron could also be selectively placed under the field oxide regions. These two advances made it feasible to reliably isolate devices in NMOS circuits.

EXAMPLE 5-5: If the parasitic field NMOS transistor of Example 5-3 is implanted with the boron implant dose calculated in Example 5-4 before the field oxide is grown, determine V_T of this device (assume that no boron is lost because of segregation effects during the oxide growth).

SOLUTION: The threshold voltage of the device in Example 5-3 is 0.66V. Using Eq. 5-15, we see that a boron dose of 1.8×10^{12} atoms/cm² would cause a ΔV_T of

$$\begin{aligned}\Delta V_T &= q N_I / C_{ox} = 1.6 \times 10^{-19} \text{ C} \times 1.8 \times 10^{12} \text{ cm}^{-2} / 7 \times 10^{-9} \text{ F/cm}^2 \\ &= 41 \text{ V}\end{aligned}$$

Thus, the V_T of the parasitic field device would be (0.66 V + 41 V), or almost 42 V. This example shows that a combination of a thicker field oxide and a channel-stop implant dose can increase the threshold voltage of the parasitic field device to sufficiently large values.

5.3.6 Short-Channel Devices

As MOS channel lengths got smaller than about 3 μm , so-called short-channel effects began to become increasingly significant. As a result, device design and, consequently, process technology had to be modified to take these effects into account so that optimum device performance could continue to be obtained. Short-channel effects and their impact on processing in will be discussed in section 5.5.

5.4 PROCESS SEQUENCE FOR FABRICATING NMOS INVERTERS WITH NMOS DEPLETION-MODE LOADS

This section will describe the process sequence used to fabricate silicon-gate NMOS digital integrated circuits. A simple inverter circuit with enhancement-mode pull-down and depletion-mode pull-up NMOS transistors is used in the example. Many other logic circuits can also be implemented with this *enhancement-depletion* (E-D) NMOS process sequence.

A brief outline of how the inverter functions from a circuit point of view is given at the outset to help define some of the characteristics of MOS IC circuits. The process flow described in this section represents a relatively simple NMOS technology. In fact, many techniques have been developed to improve the performance and packing density of MOS circuits beyond those that can be produced by this process sequence. These include alternatives to LOCOS isolation structures; thin gate oxides (and alternative gate oxide materials); shallow source/drain junctions; spacers (for forming lightly-doped drain structures and salicides); punchthrough prevention implants; double polysilicon; 2-level (or more) metallization; and self-aligned contact structures. While most of these techniques are discussed in this chapter; others are discussed elsewhere in the book (e.g., isolation technology in chap. 2, salicides and self-aligned contact structures in chap. 3, and multilevel metallization technology in chap. 4).

5.4.1 Operation of an NMOS Inverter with a Depletion-Mode Load

The most common applications of MOS transistors are in integrated circuit digital logic gates and memory arrays. Several types of circuits have been developed to implement logic gates in MOS ICs, with each circuit type characterized by the type of *load device* it utilizes. The class of logic-gate circuits that has become standard in most NMOS digital ICs is based on enhancement-depletion (E-D) NMOS technology, and such E-D logic gates are the basis for most NMOS microprocessors, microprocessor peripheral devices, and static NMOS memories.

The inverter is the fundamental logic gate. E-D NMOS inverters are composed of two transistors; an enhancement mode MOSFET called the *driver*, which is switched *ON* and *OFF* by the input signal; and a depletion-mode MOSFET, called the *load*. The circuit diagram of this inverter is shown in Fig. 5-13a, and an example of a layout (as it would appear on the completed wafer) is shown in Fig. 5-13b. The cross-sectional view of this structure is shown in Fig. 5-13c.

The load connects the power supply voltage V_{DD} and the output of the inverter. The gate of the load transistor is electrically connected to its source region, so that V_G in the load transistor always equals zero. Since depletion-mode devices are always *ON* when $V_G = 0$, tying together the gate and drain ensures that the device is always *ON*. The driver transistor has its source connected to ground, while its drain region is electrically connected to the source region of the load transistor and the output. The threshold voltage of the driver transistor, V_{TE} , is selected to be between 0 and V_{DD} ,

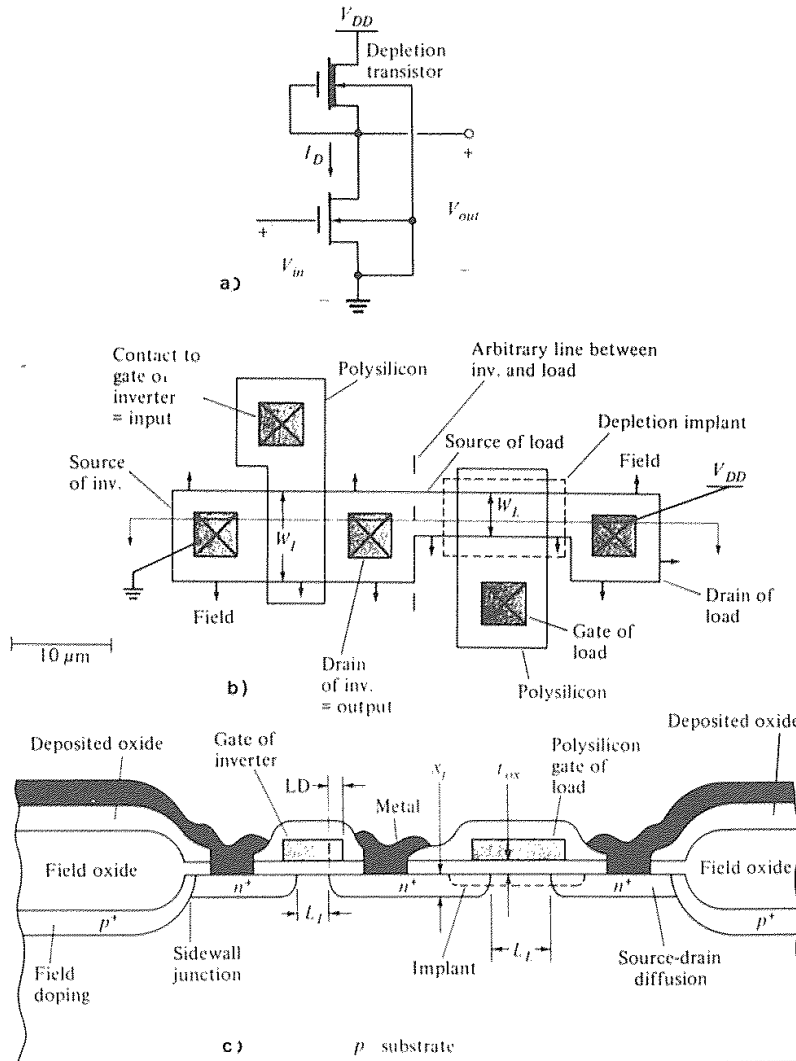


Fig. 5-13 NMOS inverter, depletion MOSFET load. (a) Schematic representation. (b) Layout. (c) Cross sectional view.⁷ From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission.

while the threshold voltage of the load transistor (since this is a depletion-mode device), V_{TD} , is selected to have a negative voltage value.

The input signal to the inverter (a voltage, V_{in}) is fed to the gate of the driver transistor, and the output signal is the voltage level at the output node, V_o . When there is a logic 0 input signal (low voltage at V_{in}), V_G at the driver transistor is $<V_{TE}$. In this case, there is no conducting channel between source and drain, so the impedance between the output and ground is very high. Since the depletion transistor is *ON*, however, the output is electrically connected to V_{DD} , and V_o rises (or is *pulled up*) to logic 1 (close to V_{DD}).

When the input voltage to the inverter is logic 1 (close to V_{DD}), the gate voltage applied to the driver is greater than V_{TE} , thus turning the driver transistor *ON*. A low-impedance path then exists between the output node and ground. Hence, the driver transistor can conduct a large current with a small voltage drop across it, allowing the output to go to logic 0. This demonstrates how the logic level at the input is inverted at the output of the circuit.

The desired characteristics of digital logic gates as IC elements are the following: fast switching speed (i.e., small propagation-delay time); low power dissipation; small size (i.e., minimum silicon area); and high noise margin. We'll now see how the NMOS inverter just described is designed to meet these requirements. (Note: the circuit design considerations used to arrive at the configuration presented here can be found in texts dealing with IC design.^{7,17,18} Since our interest here is in how these design choices impact the processing of the device, readers interested in their justification are advised to consult the references mentioned.)

In order to obtain maximum packing density (i.e., most logic gates per area of substrate) with minimum power consumption* an E-D NMOS inverter would be designed in the following manner:¹⁷ The driver transistor would have a gate area whose dimensions would be the minimum that could be fabricated in that generation of technology. The depletion load transistor would have an effective *channel length* (L_{effD}) that would be four times the effective channel length of the driver device (L_{effE}) and a minimum gate width.

As a simple example, let us assume an inverter is to be fabricated with E-D NMOS technology, and that the minimum manufacturable feature dimension is $5\text{ }\mu\text{m}$. Let us also assume that a $1\text{-}\mu\text{m}$ lateral diffusion occurs under each side of the gate. In this case, L_{effE} is $3\text{ }\mu\text{m}$ and thus L_{effD} should be $12\text{ }\mu\text{m}$. This means that the drawn gate lengths of the driver and load devices would be $5\text{ }\mu\text{m}$ and $14\text{ }\mu\text{m}$ respectively. The drawn gate widths of both devices would be $5\text{ }\mu\text{m}$ (since these are the dimensions between the walls of the field oxide).

If a 5 V power supply is assumed, the threshold voltages of the driver and depletion load transistor (V_{TE} , and V_{TD}) would be selected to be approximately 1.0 V and -3.0 V, respectively. (These two V_T values are chosen in order to give a high noise margin without severely impacting switching speed.) To conserve space, *buried contacts* between the polysilicon gate and the silicon substrate would be employed. (The procedure used for fabricating such contacts is presented in chap. 3.)

* But as a result, with less than maximum switching speed,

While the E-D NMOS technology has some important advantages over other IC technologies (particularly high packing density), it does have some drawbacks that become extremely serious as the number of devices on the chip gets very large. The most important of these is the high total power that is consumed. The origin of this power consumption arises from the operation of the NMOS inverter. When the inverter (and similarly other logic gates) has a low output state, both driver and load are *ON*, allowing current to flow from V_{DD} to ground. The power consumed by each inverter in a low-output state is the product of this current and V_{DD} . Thus, if a series of inverters are connected together, 50% of them will be drawing power at all times. When the devices get small enough, the power density on the chip becomes so large that it becomes necessary to replace NMOS with CMOS, since this technology consumes much less power per logic gate.

5.4.2 Process Sequence of a Basic E-D NMOS IC Technology

Figure 5-14 is a flow-chart representation of the sequence of steps that were used to fabricate typical E-D NMOS digital integrated circuits for gate lengths down to about $3\text{ }\mu\text{m}$.⁷⁶ Figures 5-15a through 5-15j show what occurs on the wafer as this sequence of process steps is followed. The process being illustrated is a seven mask process (including the passivation pad mask, even though this final pad mask is not shown). The E-D NMOS inverter described in the previous section is used here as a vehicle for showing how device features on the wafer surface are created during the course of the process flow.

5.4.1.1 Starting Material. The starting material is a lightly doped ($\sim 5 \times 10^{14}$ – 10^{15} atoms/cm²) *p*-type <100> silicon wafer (substrate). As described earlier, the lightly doped substrate is chosen to provide low source/drain-to-substrate capacitance, high source/drain-to-substrate breakdown voltage, high carrier mobility, and low sensitivity to source-substrate bias effects. A backside gettering process, such as implanting with Ar, to create crystalline-damaged regions that will trap mobile impurities during subsequent heat steps during the process may be used prior to the next step (see Vol. 1, chap 2).

5.4.1.2 Active Region and Field Region Definitions. The first task in this processing sequence is to define the active device and field regions on the wafer surface. This is done by selectively oxidizing the field regions so that they are covered with a thick field oxide, using the LOCOS process. The steps involved in this task are those of boxes 3-9 in Fig. 5-14, and are illustrated in Figs. 5-15a and 5-15b.

A thin pad oxide (20-60 nm thick) is first thermally grown or CVD-deposited on the wafer surface as a stress-relief layer. This is followed by the deposition of a CVD nitride layer (100-200 nm thick). *Mask #1* is then used to expose a resist film that was spun on after the nitride deposition (Fig. 5-15a2). After exposure and development, the resist layer remains behind only in the regions that will be the active device regions

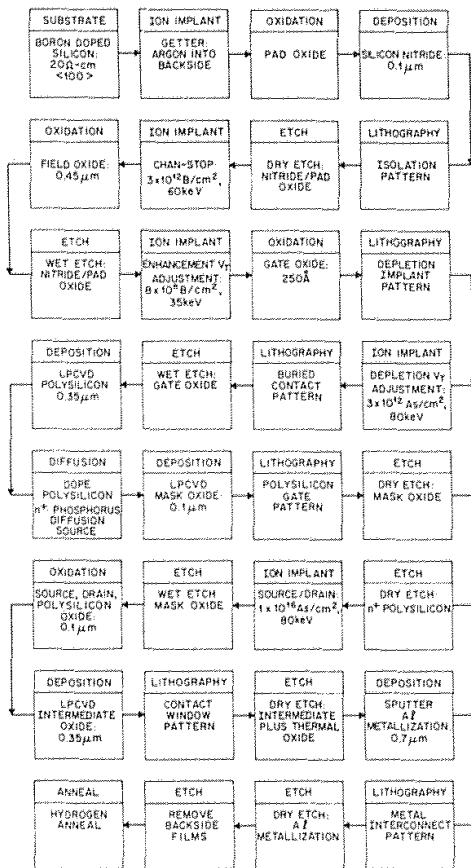


Fig. 5-14 Main steps in an *n*-channel, polysilicon-gate MOS IC process flow.⁷⁶ (© 1980 IEEE).

(Fig. 5-15b1). Next the nitride and pad oxide are anisotropically dry-etched away in the regions not covered by the resist (field regions). Thus, after the removal of the resist, the active areas are covered with the nitride/pad-oxide layer (Fig. 5-15b3).

In the next step, a boron implant (10^{12} - 10^{13} atoms/cm², 40-80 keV) is performed to create *channel stops* in the field regions. The nitride/pad oxide layer now acts as a mask (Fig. 5-15c1) to prevent the boron from penetrating the silicon in the active areas. (Note that in some processes the resist is not removed until after the channel-stop

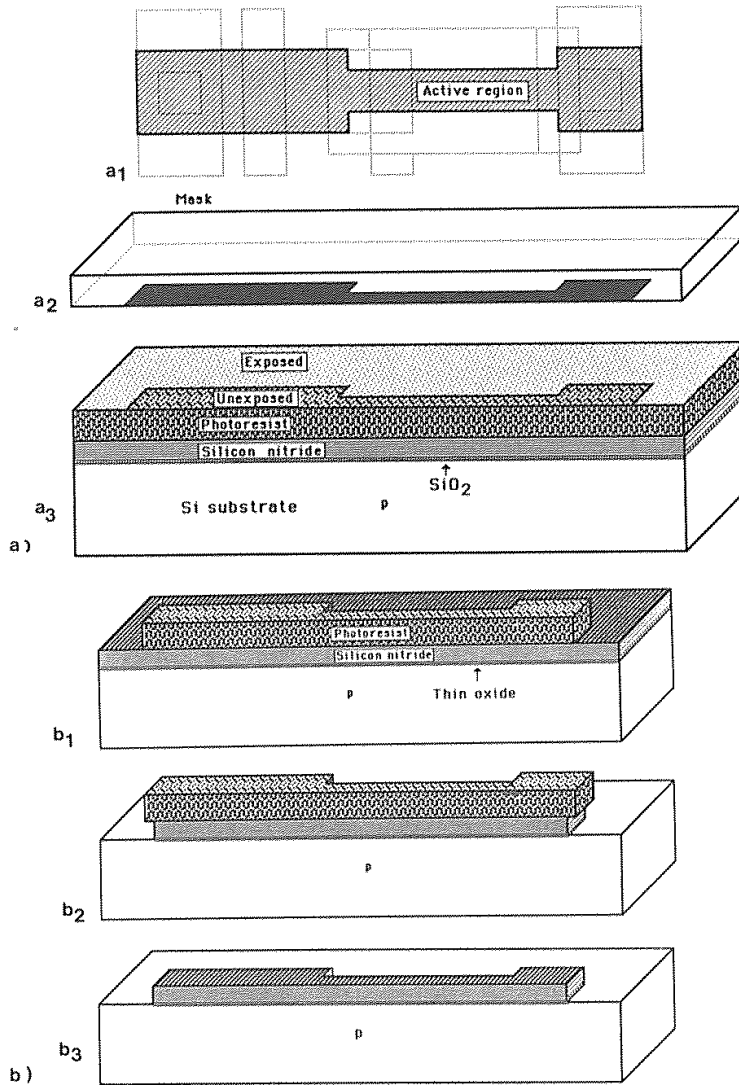


Fig. 5-15 NMOS E-D inverter fabrication sequence. (a) Patterning of the active region. (b) Patterning the silicon nitride-pad oxide layers. From W. Maly, *Atlas of IC Technologies*, Copyright 1987 by the Benjamin/Cummings Publishing Company. Reprinted with permission.

330 SILICON PROCESSING FOR THE VLSI ERA – VOLUME II

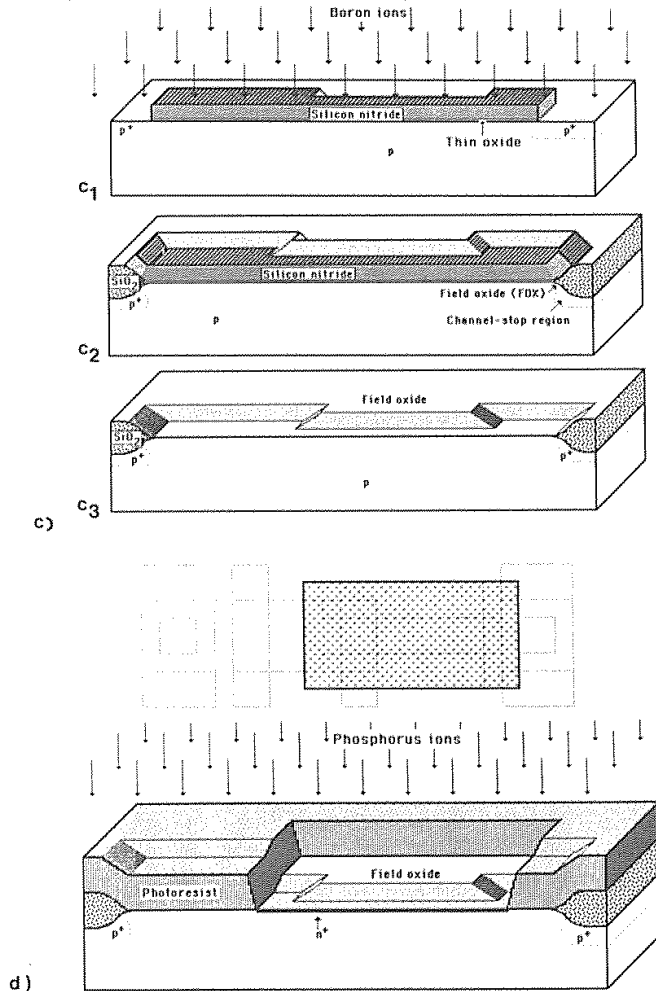


Fig. 5-15 (c) Active region formation and channel-stop implant. (d) Implantation of the channel of the depletion-mode transistor. From Maly, *Atlas of IC Technologies*.

implant, as the nitride/pad oxide layer may be too thin to act as an effective implant mask. In fact, with the patterned resist still in place, the channel stop may be implanted through the nitride/pad oxide layer, which in this case would not be etched until the implant is performed.)

A thermal-oxidation step is then performed to grow a thick ($0.5\text{--}1.0\text{ }\mu\text{m}$) field oxide over the regions where the nitride has been etched away. In this process, the field oxide is self-aligned to the channel stops. During the oxide growth, some lateral oxidation also occurs under the nitride edges, forming the *bird's beak* structure (see chap. 2 for more details concerning this effect, as well as other problems that arise in connection with the field-oxide growth step). After the field oxide has been grown, the remaining nitride and pad oxide are stripped, leaving the active areas with exposed silicon surfaces for further processing (Fig. 5-15c3).

5.4.1.3 Gate-Oxide Growth and Threshold-Voltage Adjust Implant

In the next major step the gate oxide is grown, and the threshold voltages of the enhancement-mode and depletion-mode transistors of the inverter are adjusted through ion implantation. The growth of the gate oxide is a critical step, as a defect-free, very thin ($15\text{--}100\text{-nm}$), high-quality oxide without contamination is essential for proper device operation. The gate oxide is grown only in the exposed active region (the field-oxide thickness is actually increased slightly as a result of this oxide-growth step). As noted earlier, the drain current in an MOS transistor is inversely proportional to the gate-oxide thickness (for a given set of terminal voltages). As a result, the gate oxide is normally made as thin as possible, commensurate with oxide breakdown and reliability considerations.

In order for a high-quality gate oxide to be obtained, the surface of the active area is wet-etched to remove any residual oxide. A sacrificial oxide is often deliberately grown on the exposed active areas after field oxidation to remove any dry-etch induced damage or unwanted nitride (due to the Kooi effect, see chap. 2).⁶⁵ After such oxides have been stripped, the gate oxide is grown slowly and carefully, usually through dry oxidation in a chlorine ambient (see Vol. 1, chap. 7).

The threshold-voltage adjust implant of the enhancement-mode devices is performed next. In this step, boron is implanted through the gate oxide ($10^{12}\text{--}10^{13}\text{ atoms/cm}^2$, $50\text{--}100\text{ keV}$), but the ions are not given enough energy to penetrate the field oxide. No mask is used in this step. (Note that in many processes, another pre-gate oxide is grown, through which this implant is performed. It is again stripped off following the implant, and the gate oxide is then grown.)

Next, the depletion-mode devices of the circuit are given their threshold-voltage implant dose (Fig. 5-15d). The areas of the depletion-mode transistor channels are implanted with phosphorus or arsenic ions ($\sim 10^{12}\text{ atoms/cm}^2$, 100 keV) to give a threshold voltage of about -3.0 V . The implant dose is adjusted so that it overcompensates for the previous boron threshold-voltage-adjust implant, thus making the surfaces *n*-type. A negative threshold voltage is thus yielded, as required to establish a depletion-mode device. Photoresist (patterned by the use of *Mask #2*) is used to selectively allow the depletion-mode transistor channels to be implanted. The ions cannot penetrate the resist to reach active areas below. Likewise, the ions cannot penetrate any field oxide that is exposed by the resist opening. Hence, the location of the depletion transistor channel is defined by the intersection of the *Mask #2* window and the active region.

Buried contacts are then opened in the gate oxide using *Mask #3* (Fig. 5-15e). This opening in the gate oxide must be provided wherever it is desired to have polysilicon electrically contact the active silicon area (details of buried-contact formation are described in chap. 3). Since the polysilicon is deposited on the gate oxide, it will remain isolated from the substrate below unless a special opening is cut in the gate oxide. With *Mask #3*, resist covers the entire wafer except in those areas where the buried contact is desired. The gate oxide can then be etched from these regions, uncovering the silicon below.

5.4.1.4 Polysilicon Deposition and Patterning. A layer of polysilicon (typically $0.4\text{--}0.5\ \mu\text{m}$ thick) is next deposited by CVD over the whole wafer (see Vol. 1, chap. 6 for more information on the properties of polysilicon and its deposition process). Either ion implantation or diffusion with phosphorus is then used

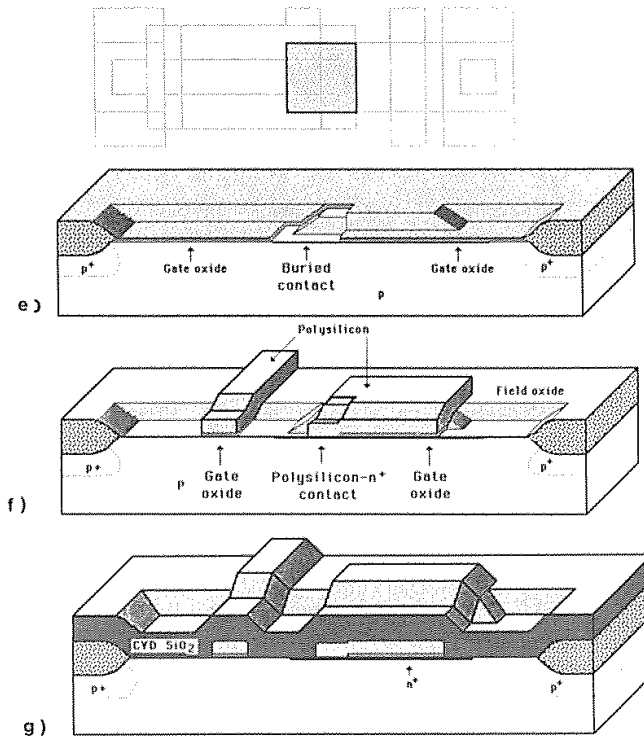


Fig. 5-15 (e) Buried contact etching. (f) Patterning of the polysilicon layer followed by gate oxide etching. (g) Deposition of the CVD SiO_2 layer followed by the diffusion of the drain and source regions. From Maly, *Atlas of IC Technologies*.

to dope the polysilicon to a sheet resistance of 20-30 Ω/sq . This resistance is adequate for MOS circuits with gate lengths $\geq 3 \mu\text{m}$. For smaller devices, polycide layers (i.e., composite layers of refractory metal silicides and polysilicon) can be used to reduce the sheet resistance to $\sim 1 \Omega/\text{sq}$ (see Vol. 1, chap. 11). Using a polycide gives us the benefits of both silicon-gate and metal-gate technologies.

The gate structure and polysilicon interconnect structures are then patterned using *Mask #4* (Fig. 5-15f). Following exposure and development of the resist, the polysilicon film is etched (in current technology this is done by means of a dry-etch process). This is a critical etch step for several reasons. First, the channel length of the device depends on the gate length, because of the self-aligned nature of the silicon gate technology. Hence, the gate-length dimension must be precisely maintained across the entire wafer, and from wafer to wafer. Second, the profile of the etched poly gate structure should be vertical; this will prevent variation of channel lengths by the penetration of the ions of the thinner regions of the gate sidewalls during formation of the source/drain regions by ion implantation. Third, to achieve the above goals, an anisotropic polysilicon etch process must be employed. This type of process, however, requires overetching to remove the locally thicker regions of polysilicon that exist wherever it crosses steps on the wafer surface. During the overetch time, areas of the thin gate oxide are exposed to the etchants. Thus, it is necessary to use a polysilicon etch process that is highly selective with respect to SiO_2 .

5.4.1.5 Formation of the Source and Drain Regions. Once the gate has been fabricated, the source and drain regions can be formed. This is normally done by ion implantation without the use of a lithography step (Fig. 5-15g). The gate and the field oxide act as masks to prevent the ion implantation from penetrating to the silicon substrate below. Therefore, only the active regions covered by the gate oxide (and no gate polysilicon), are implanted. An n^+ implant is used, with an energy that is insufficient to penetrate the gate-poly or field-oxide layers (arsenic is typically used, with a dose of $\sim 10^{16}$ atoms/ cm^2 and an energy of 30-50 keV). As noted earlier, the source and drain are thereby *self-aligned* to the gate, and the dimension of the polysilicon gate thus plays a major role in the defining of the MOS gate length.

Following the source/drain implant, an anneal (or drive-in) step is performed to activate the implanted atoms and to position the source/drain junctions as desired. During this step, some of the phosphorus doping of the polysilicon outdiffuses into the silicon substrate wherever a buried contact opening the gate oxide has been cut. This diffusion (which occurs both vertically and laterally into the silicon below) forms a heavily doped n^+ region under the polysilicon in the buried-contact exposed region. The lateral diffusion of the implanted source/drain dopant thereby becomes electrically connected to the n^+ region under the polysilicon buried-layer region. In this manner, an electrical connection between the polysilicon and the silicon is established at the buried contact locations. In some processes, the junction formed by the buried-contact dopant outdiffusion from the polysilicon is deeper than the source/drain junctions, while in others it is not as deep.

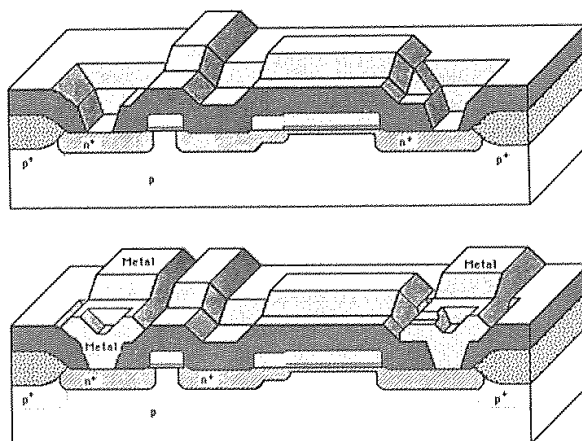


Fig. 5-15 (h) Contact cuts of the E-D inverter. Metallization of the E-D inverter. (i) Completed NMOS E-D inverter structure. Figures 15a through 15i from W. Maly, *Atlas of IC Technologies*, Copyright 1987 by the Benjamin/Cummings Publishing Company. Reprinted with permission.

The source/drain drive-in step also plays a part in determining the effective channel length (L_{eff}). That is, if the lateral junction depth is x_{jl} (which is primarily determined by the lateral diffusion during the drive-in step, because the lateral straggle of arsenic at 30 keV is only ~ 5 nm, see Vol. 1, chap. 9), L_{eff} will be decreased by $2x_{jl}$ from the gate length at the mask level. Note that the channel *width* is also reduced by the bird's beak encroachment into the active area (see chap. 2). Thus, the actual width, W_w , of an MOS device is $W_w = W - \Delta W$, where W is the width at the mask level and ΔW is the channel-width shrinkage during processing.

The depth of the source and drain is thus a critical dimension, but the doping concentration is not as important. (A discussion of shallow source/drain junction formation techniques is presented in chap. 3, section 3.10.) To a first approximation, the device characteristics will not depend on the doping concentration value, provided it is sufficiently heavy.

A diffusion step may be used to dope the source/drain regions. In some of these cases the dopant source of the diffusion is the CVD oxide layer that is deposited after the gate has been defined (see next section).

5.4.1.6 Contact Formation. After the source/drain regions have been formed, a CVD process is used to deposit a layer of doped SiO_2 (glass), about $1 \mu\text{m}$ thick, onto the wafers (see Vol. 1, chap. 6). The dopant in the SiO_2 is either phosphorus (in which

case the material is referred to as *phosphosilicate glass*), or both phosphorus and boron (making it a layer of *borophosphosilicate glass*). In some processes a thin thermal oxide is grown on the polysilicon prior to deposition of the glass layer. Nevertheless, a thick layer of SiO_2 cannot be thermally grown because of the excessive redistribution of the impurities that would take place during such growth. Hence, a lower-temperature CVD process must be used to get a sufficiently thick oxide.

The doped CVD glass layer plays several roles in the fabrication and operating aspects of the circuit. First, it acts as an insulating layer between the polysilicon and the metal to be deposited. Second, it reduces the parasitic capacitance of the interconnect metallization layer. Third, the addition of the phosphorus to the glass makes the layer an excellent getter of Na ions (recall that contamination by Na can cause instabilities in the V_T of the MOS devices). The phosphorus-doped glass immobilizes the otherwise mobile Na atoms within the CVD layer, preventing them from reaching the gate oxide and altering the threshold voltage. Finally, the dopants in the glass make it viscous at elevated temperatures (1000-1100°C for PSG, and 800-950°C for BPSG, see Vol. 1, chap. 6), allowing the layer of doped glass to be flowed after it is deposited. Through this procedure, a rounding of the contours of the glass and a smoothing out of any sharp steps is achieved. This produces better step coverage of the metal (which is deposited next) over the otherwise severe wafer topography. The high-temperature glass-flowing step also serves to activate the source/drain implanted junctions and drive them to their desired positions.

Contact openings are next created by a lithography-and-etch step (Fig. 5-15h). *Mask #5* is used to define contact opening patterns in a photoresist film, and a dry-etch process is then used to open the contact windows through the CVD SiO_2 to the underlying polysilicon and the n^+ regions in the silicon.

The contact-opening step can be critical, as the contact size and alignment limit the minimum size of the device. The source and drain regions must be large enough for the contact to fit, with allowance for alignment tolerance. If the contact opening exposes a part of the substrate, the drain or source will be shorted to the substrate. Likewise, any overlap of the source/drain contact opening and the gate will cause the gate to be shorted to the source or drain.

To keep the transistor as small as possible, the contact window is usually made at the minimum size achievable with the given process. In some processes, the exposed silicon in the contact is redoped to prevent shorting between the source/drain areas and the substrate (see chap. 2). Also note that the gate contact (i.e., Al to polysilicon) is often made outside the active device area to avoid possible damage to the thin gate oxide.

After the contact etch is completed and the resist is stripped, the doped CVD glass is again subjected to another flowing step. This procedure rounds the corners of the top of the oxide windows so that metal step coverage into the contact windows is improved. The process is called *reflow* of the contact windows (see chap. 3 and Vol. 1, chap. 6).

5.4.1.7 Metallization Deposition and Patterning. After the contacts have been opened, the metallization layer is deposited ($\sim 1 \mu\text{m}$ thick). Because the metal

layer is highly conductive, it is used whenever possible to interconnect circuit elements and to carry large amounts of supply current. The metal interconnect lines that are fabricated must have sufficient thickness, width, and step coverage to keep the current density in each line below the value that could produce electromigration failure (see chap. 4). In addition, the spacing between adjacent metal lines must be kept large enough that the lines will never touch, even under worst-case process variations.

Although evaporation was the method employed to deposit Al in the early days of MOS, it has generally been replaced by sputtering. To a great extent, the change was made because Al alloys with tightly controlled compositions became the materials of choice for the metal layer. Sputtering allows alloys to be deposited with much better compositional fidelity (see Vol. 1, chap. 10).

The metal alloy that was eventually chosen for NMOS is Al:1wt% Si. The silicon is added to the aluminum film to prevent spiking of the contacts during subsequent annealing steps (see chap. 3). In CMOS, such Al:Si alloys are being phased out as the metallization material for reasons that are discussed in the CMOS chapter and in chapters 3 and 4. Either wet-chemical etching or dry etching is used to pattern the Al film, using *Mask #6* (Fig. 5-15i and 5-15j).

Following the patterning of the metal, the Al-silicon contacts are alloyed. This step brings the Al and the n^+ silicon into intimate contact, since it allows the thin native SiO_2 layer that is likely to exist at the Al-Si interface to be reduced by the Al (see chap. 3). Such intimate contact between Al and n^+ Si establishes a low-resistance ohmic contact. The anneal process exposes the wafer to a 375-500°C temperature in an H_2 or

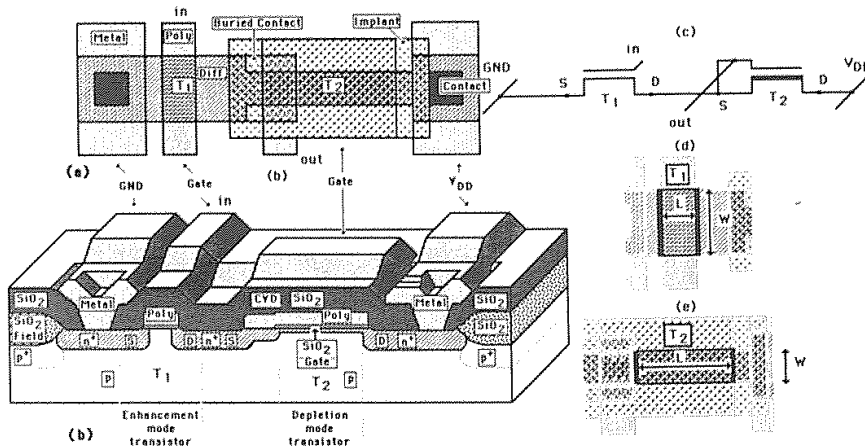


Fig. 5-15 (j) E-D inverter. 1) Composite drawing of the layout. 2) Cross section of complete structure. 3) Electrical diagram. 4) The enhancement transistor. 5) The depletion transistor. From Maly, *Atlas of IC Technologies*.

$N_2 + H_2$ (5%) ambient for about 30 minutes. As a result, this step may also be used as the annealing process for reducing the interface trap density in the gate oxide that was introduced by earlier processing steps (see Vol. 1, chap. 7).

5.4.1.8 Passivation Layer and Pad Mask. Finally, a *passivation* (or *overcoat*) layer, such as CVD PSG or plasma-enhanced CVD silicon nitride, is put down onto the wafer surface. This layer seals the device structures on the wafer from contaminants and moisture, and also serves as a scratch protection layer.

Openings are etched into this layer so that a set of special metallization patterns under the passivation layer is exposed. These metal patterns are normally located in the periphery of the circuit and are called *bonding pads* (Fig. 5-16). Bonding pads are typically about $100 \times 100 \mu m$ in size and are separated by a space of 50 to $100 \mu m$. Wires are connected (bonded) to the metal of the bonding pads and are then bonded to the chip package. In this way connections are established from the chip to the package leads.

The bonding-pad openings are created by patterning the passivation layer with *Mask #7*. If a PSG layer is used, the phosphorus (2-6 wt%) in the glass not only causes the PSG to act as a getter for Na but also prevents the glass film from cracking. Care must be taken to ensure that not more than 6% phosphorus is incorporated into the PSG, as this can cause corrosion of the underlying metal if moisture enters the circuit package (see Vol. 1, chap. 10). When silicon nitride is used, care must be taken to ensure that the deposited nitride film exhibits low stress (either tensile or compressive), so that it will not crack, since cracking would compromise the sealing capability of the film.

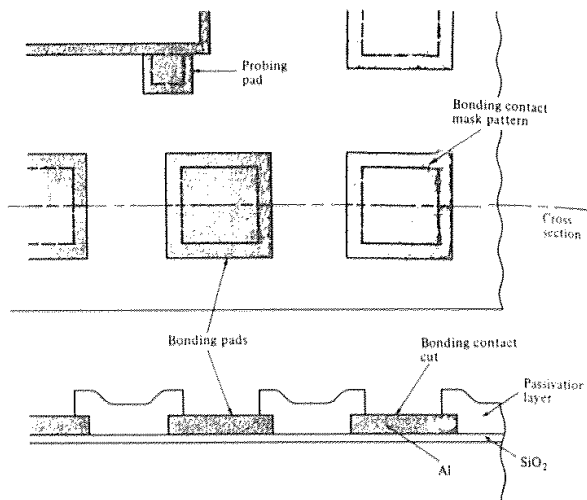


Fig. 5-16 Passivation layer and bonding pad openings. (Note, cross-section not to scale.)

5.5 SHORT-CHANNEL EFFECTS AND HOW THEY IMPACT MOS PROCESSING

The device characteristics of MOSFETs (such as threshold voltage, subthreshold currents, and I-V characteristics beyond threshold) are well predicted by Eqs. 5-1 through 5-10, if the channel lengths of the transistors are "long" (i.e., if they exceed $2\ \mu\text{m}$ in length). A guide to the design of such long-channel MOSFETs is given in reference 70.

For shorter channel devices, however, a series of effects arise that result in significant deviations from the values predicted by the long-channel models. Such *short-channel effects* become a dominant part of MOS device behavior when channel lengths decrease below $2\ \mu\text{m}$. The effects are briefly described here to provide readers with a basis for understanding the processing steps that have been developed to mitigate their adverse impact on device performance. So-called *hot-carrier effects* will be described separately in the following section, even though these have only been observed in short-channel devices. More details on short-channel effects in *p*-channel CMOS devices are also given in chapter 6. A guide to the design of submicron channel MOSFETs (those with channel lengths $\leq 1\ \mu\text{m}$), is presented in reference 62.

Short-channel effects can be divided into the following categories: (a) those that impact V_T ; (b) those that impact subthreshold currents; and (c) those that impact I-V behavior beyond threshold.

5.5.1 Effect of Gate Dimensions on Threshold Voltage

5.5.1.1 Short Channel Threshold Voltage Effect. V_T becomes less well predicted by Eq. 5-6 as the dimensions of the gate are reduced, and the error becomes significant when the dimensions are reduced to less than $2\ \mu\text{m}$. To get good agreement with measured data, a term $|\Delta V_T|$ must be subtracted from the V_T value obtained from Eq. 5-6. Thus, as the device length is reduced, the measured value of V_T of *n*-channel enhancement-mode devices becomes *less positive* than that given by Eq. 5-6, while for *n*-channel depletion-mode devices, V_T becomes *more negative*. For *p*-channel (enhancement-mode) devices, V_T becomes *less negative*.

The discrepancy arises because the equations for V_T given earlier in the chapter are based on one-dimensional theory.²⁰ It is assumed that the space charge *under* the gate is a function of only the vertical electric field, E_x (and is thereby influenced only by the charge *on the gate*). If the channel length is long, this is a reasonable assumption, as the influence of the drain and source junctions on the quantity of charge in the channel can be neglected. However, as the channel length approaches the dimensions of the widths of the depletion regions of the source and drain junctions, these depletion regions become a greater part of the *channel-depletion region* (Fig. 5-17a). Thus, some of the channel-depletion region charge is actually linked to the charge in the depletion region within the source and drain structures, rather than being linked to the gate charge. Hence, some of the channel region is partially depleted without any influence of the gate voltage. (In the extreme, if the two built-in depletion regions spanned the entire channel length when no voltage existed between source and drain, they could deplete all

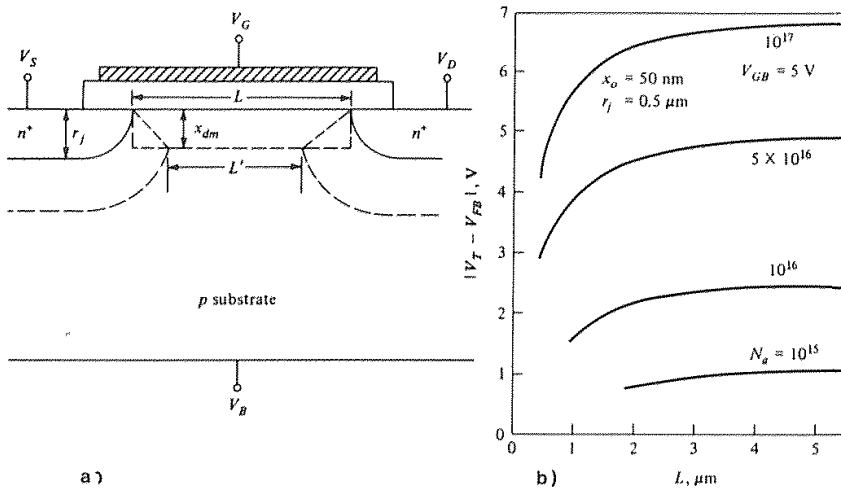


Fig. 5-17 (a) Yau's model of charge sharing.⁷³ (b) Theoretical threshold voltage as a function of channel length for various substrate doping concentrations.⁷³ Reprinted with permission of Solid-State Electronics.

of the channel region at the Si-SiO₂ interface.) Since some of the channel is depleted without the need to apply a gate bias, *less gate charge is required to invert the channel in short-channel devices than in a long-channel device with comparable substrate doping.*

A relatively simple equation that predicts the threshold lowering (ΔV_T) in terms of the device parameters is

$$\Delta V_T = \frac{q N_A W_{\max} x_j \left(\sqrt{1 + \frac{2 W_{\max}}{x_j}} - 1 \right)}{C_{ox} L} \quad (5-16)$$

where $W_{\max}$ is the maximum depth of the depletion region in the channel, and x_j is the junction depth of the source drain regions.⁷⁴

This effect is important because in order to be able to establish slightly positive V_T values (e.g., $< +1$ V) in long-channel NMOS transistors with lightly doped channels, it is necessary to increase the doping concentration at the surface of the channel. Consequently, in order to allow short-channel enhancement-mode NMOS devices to be fabricated with the same V_T value, the substrate doping concentration must be further increased. Since the magnitude of this effect increases as the device length is reduced (as is predicted by Eq. 5-16 and illustrated in Fig. 5-17b), it will be necessary to

progressively increase the substrate doping concentrations as devices are made smaller, in order to maintain suitably positive values of V_T .

Application of a drain voltage causes the drain depletion region to extend into the channel region, where it acts as an additional substrate bias, and reduces V_T .²¹

5.5.1.2 Narrow Gate-Width Effect on Threshold Voltage. In contrast to the short-channel-length effect, devices with narrow channel-widths require that such a positive-value correction term be *added* to Eq. 5-6 to give good agreement with the calculated values.²² This effect is primarily due to the encroachment of the channel-stop dopants under the edges of the sides of the gate (see chap. 2). This has the effect of doping the channel at these edges more heavily than at the center (Fig. 5-18a). Thus, it requires more charge on the gate to invert the channel than if such encroachment did not occur, and this causes a shift in V_T from its predicted value (Fig. 5-18b).²³ On the other hand, if the voltage on the gate is held constant, the edges of the channel will have a higher V_T than the center of the channel. Because $(V_G - V_T)$ is smaller, narrow-channel devices will thus conduct less drain current. Several schemes are also described in chapter 2 for reducing the channel-stop encroachment and thus reducing the narrow-width effect.

Narrow-width effects are still observed even if there are no channel-stop implants. These arise because the relatively thinner depletion region under the field-oxide device distorts the depletion region under the gate oxide and prevents the formation of an

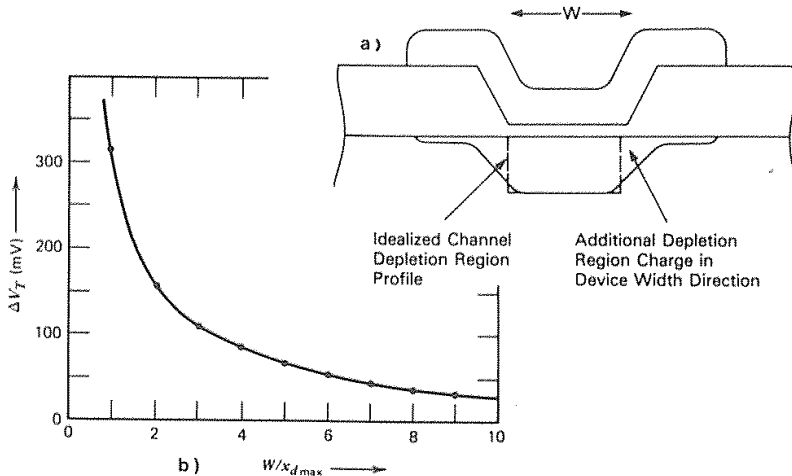


Fig. 5-18 (a) Schematic depiction of the narrow-channel effect. (b) Threshold voltage shift ΔV_T caused by the narrow-channel effect for a MOSFET with $N_A = 10^{15} \text{ cm}^{-3}$, and $t_{ox} = 50 \text{ nm}$.²²

inversion layer at the two edges. Although this leads to a slightly higher V_T , the effect is much less severe than that observed in devices with heavy channel-stop implants.

5.5.2 Short-Channel Effects on Subthreshold Currents (Punchthrough and Drain-Induced Barrier Lowering)

In section 5.2.3 we described the nature of subthreshold current flow (I_{Dst}) in MOSFETs, noting that a specific value of the *subthreshold-swing* parameter (S.S.) can be attributed to such "normal" I_{Dst} currents in long-channel devices. In short-channel MOSFETs, however, larger I_{Dst} values are observed at lower voltages than predicted by long-channel device models: one manifestation is an increase in the value of S.S. Note that even relatively small values of I_{Dst} can limit the transistor's ability to isolate nodes in a dynamic circuit or can allow excess current in static inverters. Hence, care must be taken to minimize I_{Dst} . Two of the primary causes of increased I_{Dst} are *punchthrough* and *drain-induced barrier lowering* (DIBL).

Punchthrough is normally observed when the gate voltage is well below V_T . It occurs as a result of the widening of the drain depletion region when the reverse-bias voltage on the drain is increased. The electric field of the drain may eventually penetrate into the source region and thereby reduce the potential energy barrier of the source-to-body junction (Fig. 5-19).²³ When this occurs, more majority carriers in the source region have enough energy to overcome the barrier, and an increased current then flows from source to body. Some of this current is collected by the drain, thereby increasing I_{Dst} . In general, punchthrough current begins to dominate I_{Dst} when the drain and

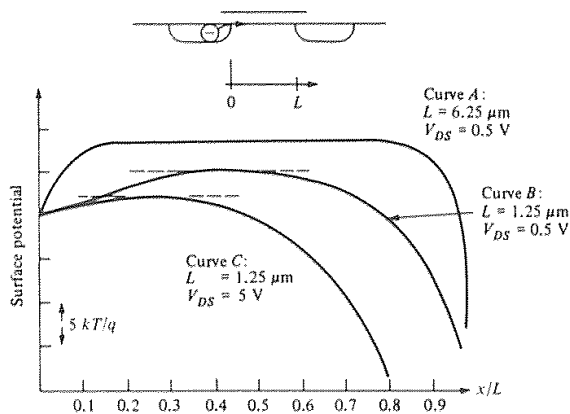


Fig. 5-19 Surface potential in the channel for devices with different channel lengths.²³ (© 1979 IEEE).

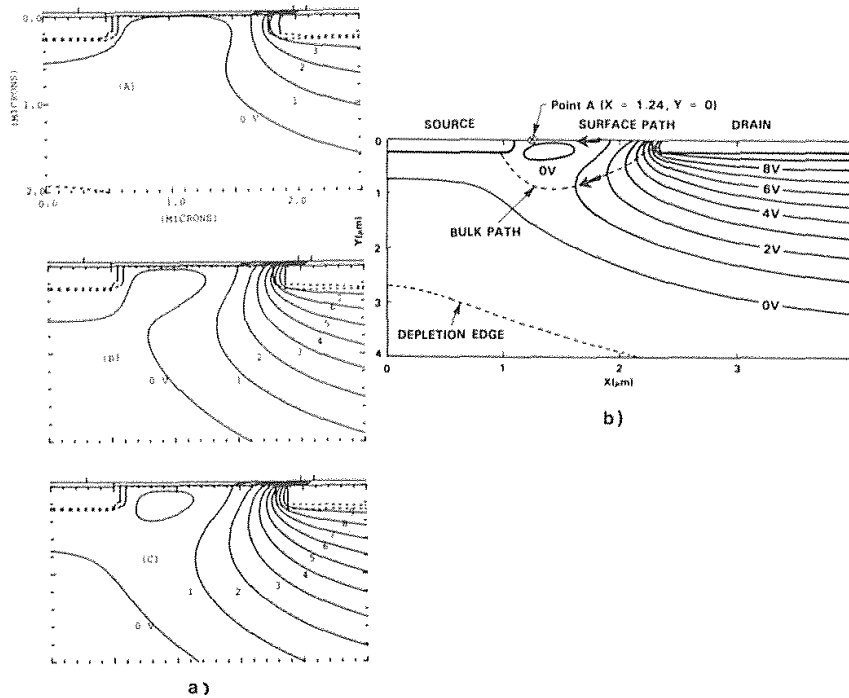


Fig. 5-20 (a) 2-D potential profile of an n -channel MOSFET with a drain bias of: 1) 3 V; 2) 7 V; 3) 9 V. Channel length = $1\ \mu\text{m}$. (b) Simulation of the potential profile of an n -channel MOSFET with a gate and drain bias of 0 and 9 V respectively. The surface DIBL and bulk punchthrough paths are indicated. From K. M. Cham, *et al.*, *Computer Aided Design and VLSI Device Development*. Copyright 1986 Kluwer Academic Publishers. Reprinted with permission.

source depletion regions meet, and it can be suppressed by keeping the total width of the two depletion regions smaller than the channel length.²⁴

Calculations of the potential in the bulk channel region in devices that use ion implantation to adjust V_T indicate that the barrier is lowest away from the Si-SiO₂ interface (usually at almost the same depth as the source/drain junction depths). That is, the V_T -adjust implant increases the doping concentration near the surface of the channel, causing the drain depletion region to be wider in the bulk than it is near the Si-SiO₂ interface. As a result, punchthrough current flows *below* the surface (Fig. 5-20). Consequently, the gate voltage has less control over the subthreshold current (i.e., even with sufficient gate voltage to turn off the channel, I_{Dst} can still flow in such devices).

An enhancement-mode device which is not turned off when $V_G = 0$ loses its ability to function as a switch.

Similarly, the application of a drain voltage in short-channel devices can also cause drain-induced barrier lowering (DIBL). That is, the drain voltage can cause the *surface potential* to be lowered (Fig. 5-21).^{23, 71} As a result, the potential energy barrier at the *surface* will be lowered, and the subthreshold current in the channel region at the Si-SiO₂ interface can be increased (5-21b). This implies that I_{Dst} at the surface due to DIBL is expected to become larger as the gate voltage approaches V_T .

These two effects illustrate the complexity involved in modeling the overall subthreshold I-V behavior of short-channel MOSFETs. That is, both punchthrough current (in the bulk), as well as DIBL-induced current (at the surface), may simultaneously contribute to I_{Dst} .

To prevent punchthrough current in short-channel devices, the substrate doping can be increased to decrease the depletion-layer widths. These widths can be estimated using the formula for the width of a one-sided step junction:

$$W = \sqrt{\frac{2 K_s \epsilon_o (|V_A| + V_{bi})}{q N_B}} \quad (5-17)$$

where the built-in voltage, V_{bi} , given by

$$V_{bi} = 0.56 + (kT/q) \ln(N_B/n_i) \quad (5-18)$$

and where V_A is the total applied voltage and N_B is the doping concentration of the body. Figure 5-22 gives the depletion-layer width of *pn* junctions as a function of doping and applied voltage.

However, increasing the substrate doping also increases the source-to-body and drain-

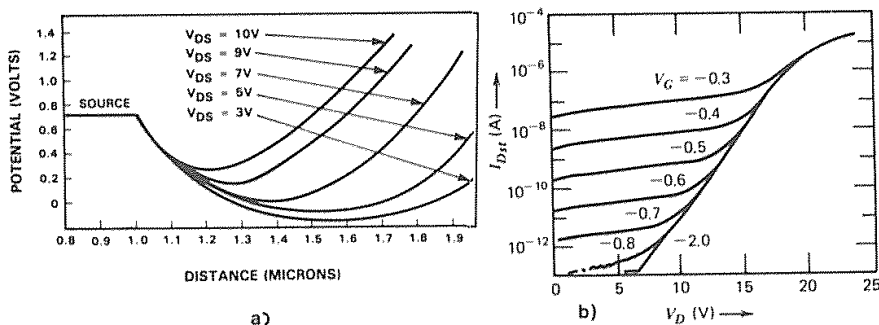


Fig. 5-21 (a) DIBL versus drain bias for short-channel MOSFET. (b) Experimental low-current characteristics for a MOSFET with $L = 2.1 \mu\text{m}$, $V_{SB} = 0.81$ V. © 1974 IEEE.

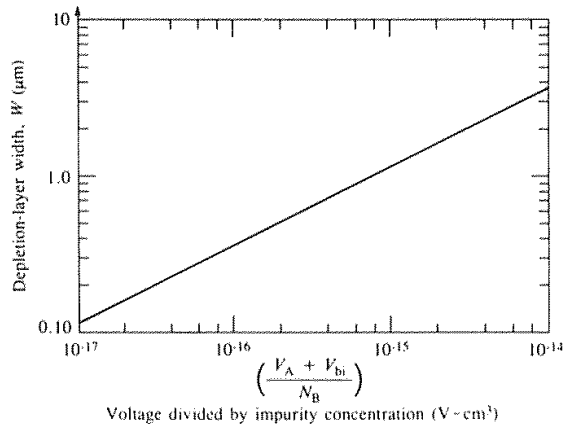


Fig. 5-22 Depletion-layer width of a one-sided step function as a function of doping and applied voltage calculated from Eqs. 5-17 and 5-18.

to-body junction capacitances, as well as the body factor. In addition, it reduces the breakdown voltages of the source/drain junctions. To avoid these drawbacks, an additional boron implant (whose peak concentration is located at a depth near the bottom of the source-drain regions) can be performed. This additional doping 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 normal I_{Dst} current of the device.

For example, in Fig. 5-23, a $1.2\text{-}\mu\text{m}$ device with a body doping of $1.9 \times 10^{15} \text{ cm}^{-3}$ without such a punchthrough-stopping implant shows a large value of I_{Dst} even when $V_G = 0 \text{ V}$ (curve A). This indicates that the device is already exhibiting punchthrough.²⁶ Implants of boron with a dose of $8 \times 10^{11} \text{ atoms/cm}^2$ and different energies are then performed in an attempt to reduce I_{Dst} to the values exhibited by a long-channel device (curve B). If the implant is too shallow, the extra implant has the effect of shifting the V_T of the device to well beyond the desired value. When the energy is increased so that the implant is sufficiently deep, the value of I_{Dst} drops to that exhibited by the long-channel device. At the same time, the surface concentration remains essentially unchanged, so that V_T is not appreciably shifted.

In another example, the S.S. of a device without a punchthrough-prevention implant is measured as its length is varied (Fig. 5-24a). At an L_{eff} of $\sim 0.85 \text{ }\mu\text{m}$ the S.S. starts to increase, indicating that punchthrough current begins to dominate I_{Dst} . By adding an implant step that places boron atoms in the dashed subsurface region shown in Fig. 5-24b, the punchthrough component of I_{Dst} is suppressed so that it is not observed until L_{eff} becomes nearly as small as $0.5 \text{ }\mu\text{m}$.

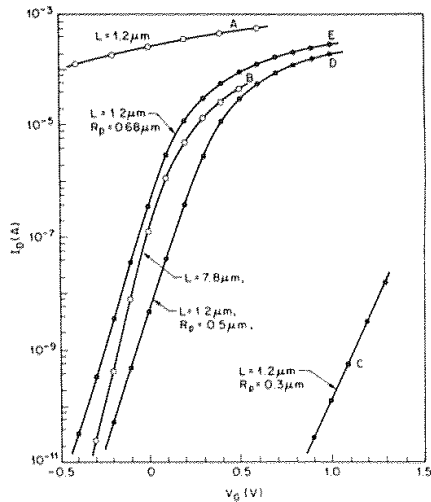


Fig. 5-23 Drain current versus gate voltage for n -channel devices with a substrate doping of 1.9×10^{15} atoms/cm³, source/drain junctions $0.47\text{-}\mu\text{m}$ deep, $575\text{ \AA}$ gate oxide, drain voltage of 5 V , and V_{BS} of 0 V . Devices A and B have no channel implant, and devices C and E have a boron channel implant of 8×10^{11} atoms/cm² at various energies.²⁶ (© 1978 IEEE).

5.5.3 Short-Channel Effects on I-V Characteristics

The I-V characteristics of short-channel devices are significantly altered in three ways. First, the combined effects of reduced gate length and gate width produce a change in V_T . Second, the channel length is modulated by the drain voltage when the device is in saturation (i.e., $V_{DS} > [V_G - V_T]$), causing an increase in device gain over that predicted in an ideal long-channel device (*channel-modulation effect*). Third, the mobility of the carriers in the channel is reduced by two effects, which also reduces I_D . (The two effects are the *mobility-degradation factor*, due to the gate field, and the *velocity-saturation factor*).

Figure 5-25 shows the I-V characteristics of an MOS device.²⁸ The curves in Fig. 5-25a are those of an ideal long-channel MOS device, while those in Fig. 5-25b show the effect of adding the channel modulation factor. Figure 5-25c shows the combined effect of adding the mobility degradation factor to those of Fig. 5-25b. The velocity saturation factor also has the effect of making the both I_D and g_m independent of channel length in silicon MOS transistors for $L_{eff} \leq 1.25\text{ }\mu\text{m}$. More details on these effects are provided in suitable device physics texts.^{1,3,28}

A general guide to the design of short-channel MOSFETs is given in references 62 and 69. In addition, a simple *engineering model* for short-channel devices has also been developed.⁶⁸ Its purpose is to provide a simple picture of the essential electrical behaviors of the short-channel MOSFETs from the perspective of a circuit designer. That is, this engineering model relates the terminal voltages to the drain current, much as Eqs. 5-8 and 5-10 yield the I-V characteristics for long-channel MOS devices. Consequently, device designers who need to relate device and process parameters to circuit parameters should also find this model useful.

5.5.4 Summary of Short-Channel Effects on the Fabrication of MOS ICs

In the first section of this chapter, we showed that the use of lightly doped substrates generally produced optimum device behavior in long-channel MOS transistors. In this

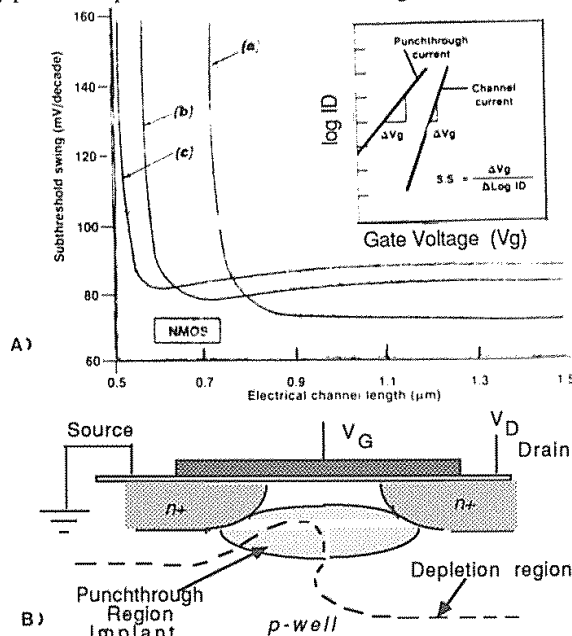


Fig. 5-24 (A) Subthreshold slope versus electrical channel length for NMOS devices ($V_{Tn} = 0.7$ V), having a common threshold adjustment implant and punchthrough implant doses of: (a) zero; (b) $2 \times 10^{11} \text{ cm}^{-2}$; and (c) $3 \times 10^{11} \text{ cm}^{-2}$.²⁷ Reprinted with permission of Semiconductor International. (B) NMOS cross-section with implant placed into punchthrough region.

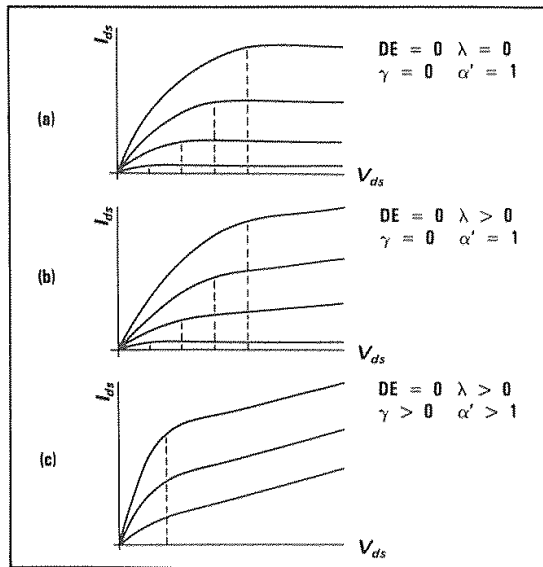


Fig. 5-25 The I-V curves of an MOS device showing the effects of progressively increasing short-channel behavior. (a) Long-channel behavior. (b) With channel-length modulation. (c) Addition of velocity saturation.²⁸ (© 1986 IEEE).

section we noted that higher substrate doping is needed to overcome some of the detrimental impacts of short-channel effects. Thus, trade offs need to be made in selecting the proper substrate doping-concentration values to achieve optimum short-channel MOS device performance. Some of these trade offs are discussed by Kakumu,²⁹ who points out that a higher substrate doping concentration produces decreased ring-oscillator gate delay in submicron CMOS because of increased junction capacitances and decreased carrier mobility (due to increased impurity scattering).

5.6 HOT-CARRIER EFFECTS IN MOSFETs

If device dimensions are reduced and the supply voltage remains constant (e.g., 5 V), the lateral electric field generated in MOS devices increases. If the electric field becomes strong enough, it can give rise to so-called *hot-carrier* effects in MOS devices. This has indeed become a significant problem in NMOS devices with channel lengths smaller than $1.5\ \mu\text{m}$ (and in PMOS devices with submicron channel lengths).³⁰ Hot-electron effects are more severe than hot-hole effects because of the higher electron mobility. Therefore, we begin our discussion by considering hot-electron effects in *n*-channel devices. At the end of this section we will also discuss the impact of hot-carrier effects on *p*-channel devices.

The maximum electric field, E_M , in a MOSFET occurs near the drain during saturated operation. A rigorous calculation of the field near the drain is a complex procedure, requiring a computer-aided solution of the two-dimensional Poisson equation, with the results of one such analysis being shown in Fig. 5-26. Nevertheless, the value of E_M can be estimated from³²

$$E_M = (V_{DS} - V_{Dsat}) / m \quad (5-19)$$

where

$$m = 0.22\ t_{ox}^{1/3} / x_j'^{1/2} \quad (5-19a)$$

and t_{ox} is the gate oxide thickness and x_j' approximately corresponds to the source/drain junction depth. Although V_{Dsat} depends on L_{eff} , the dependence is weaker than a linear

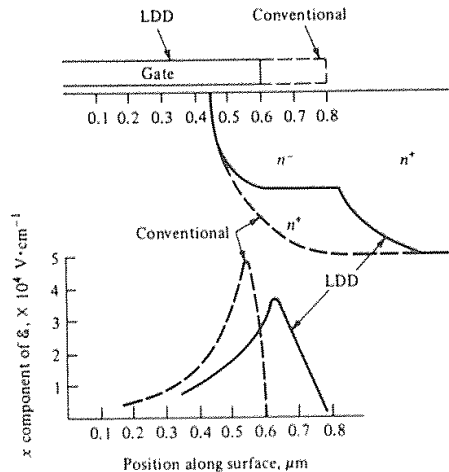


Fig. 5-26 Magnitude of the electric field at the Si-SiO₂ interface as a function of distance: $L = 1.2\ \mu\text{m}$, $V_{DS} = 8.5\ \text{V}$, $V_{GS} = V_T$.³⁸ (© 1980 IEEE).

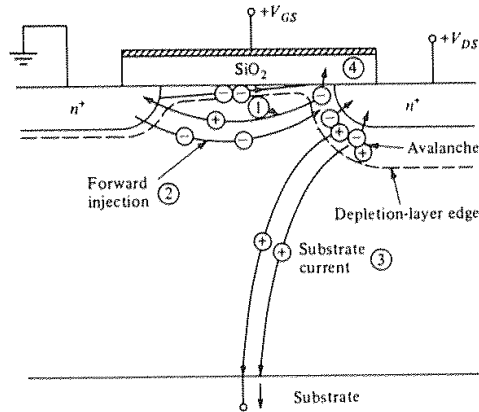


Fig. 5-27 Hot-carrier generation and current components. ①- Holes reaching the source. ②- Electron injection from the source. ③- Substrate hole current. ④- Electron injection into the oxide.

relationship (especially if $L_{eff} > 1 \mu m$). Thus, we can see that E_M increases as device dimensions shrink, but this is due to thinner gate oxides and shallower junctions, as well as to the reduction in L_{eff} .

Regardless of the factors that increase their magnitude, such high electric fields cause the electrons in the channel to gain kinetic energy and become "hot" (i.e., their energy distribution is shifted to a much higher value than that of electrons which are in thermal equilibrium with the lattice). Such hot electrons (which become hot near the drain edge of the channel, since that is where E_M exists) can cause several effects in the device. First, those electrons that acquire ≥ 1.5 eV of energy can lose it via *impact ionization*, which generates electron-hole pairs. The total number of electron-hole pairs generated by impact ionization is *exponentially dependent on the reciprocal of the electric field*, $\sim 1/E_M$. In the extreme, this electron-hole pair generation can lead to a form of *avalanche breakdown* (Process 1, shown in Fig. 5-27). Second, the hot holes and electrons can overcome the potential energy barrier between the silicon and the SiO_2 (~ 3.1 eV), thereby causing *hot carriers to become injected into the gate oxide*. Each of these events brings about its own set of repercussions.

5.6.1 Substrate Currents Due to Hot Carriers

When electron-hole pairs are created by impact ionization, the electrons are normally attracted to the drain, and they add to the drain current. The *holes*, on the other hand,

enter the substrate and constitute a part of the parasitic substrate current (*Process 3* in Fig. 5-27). This substrate current, I_{sub} , can itself produce several problems:

- If a substrate bias-generator circuit is included on-chip, the output of the bias generator will be less negative with increasing substrate current.
- If some of these holes are collected by the source (instead of by the body contact), *and* this collected hole current causes a voltage drop in the substrate material on the order of 0.6 V, the substrate-source *pn* junction will conduct significantly. Electrons will then be injected from the source to the substrate, just like electrons injected from emitter to base of an *npn* transistor (the *forward injection* shown in Fig. 5-27). These electrons can, in turn, gain sufficient energy as they travel toward the drain to cause additional impact ionization and create new electron-hole pairs. A positive-feedback mechanism thus exists, one that can sustain itself if the drain voltage exceeds a certain value. This is observed externally as a form of breakdown, referred to as a *snapback breakdown*. A particularly clear explanation of this effect, including the reason for the negative-resistance, or "snapback," portion of the curves, is given in reference 3.
- As some of the holes are accelerated through the drain-substrate depletion region, they may acquire enough energy to cause secondary impact ionization. This will create electrons far from the drain region. Some these electrons, instead of being collected by the drain, may escape the drain field and instead travel (sometimes hundreds of microns) to other nodes on the chip. This may lead to a reduction of the storage time of dynamic circuit nodes in DRAMs (i.e., manifested as a degradation of the refresh time).³³ This excess electron current is reported to be around 10^{-4} times smaller than the substrate ionization impact current itself.
- I_{sub} may induce latchup in CMOS circuits.

The magnitude of the substrate current depends exponentially on the value of E_M . An example of how I_{sub} depends on decreasing channel length is seen in Fig. 5-28.³⁴ This shows the maximum I_{sub} generated at a voltage of 5 V versus the effective channel length for MOS transistors processed with the same technology. The magnitude of I_{sub} would increase even more rapidly with shrinking L_{eff} if the oxide thickness and junction depth were scaled together with the channel length.

5.6.2 Hot-Carrier Injection into the Gate Oxide

The hot holes and electrons that are injected into the gate oxide cause another set of problems. First, some of these carriers pass to the gate electrode (mostly hot holes) and thus constitute a gate current, I_G , typically in the fA (10^{-15} A) range. For higher, but still nondestructive, biases, the gate current can grow rapidly to become several pA (10^{-12} A).

However, some fraction of the hot carriers injected into the gate oxide do not reach the gate electrode. This is because the gate oxide contains empty electron states (also known as *traps*), which can be filled by the injected hot carriers. Such occupied traps generally become electron traps, even if they are initially filled with holes. As a result, there is a negative charge density caused by the trapping of the hot carriers in the oxide. This *trapped charge* acts like a contribution to the fixed oxide charge term in the expression for the device threshold voltage. Furthermore, the trapped charge accumulates with time. Due to the polarity of the trapped charge, the resulting shift in the *n*-channel device threshold voltage is *positive* (i.e., V_T increases in NMOS devices). If I_G becomes of the order of pA, the trapping of a fraction of the electrons injected into the oxide can become a significant effect (the fraction that gets trapped is ~ 1 in 10^6 injected electrons). As a result, this increase in *negative* stored charge can lead to a permanent change in V_T in the MOSFET. (The means by which charge in the oxide impacts V_T was described earlier.)

EXAMPLE 5-6: V_T Shift due to Hot-Electron Effects. Given a short-channel NMOS device with $t_{ox} = 20$ nm, a gate *width* of $5\text{ }\mu\text{m}$, and a gate current of 1 nA that is momentarily caused by hot carrier injection. The current flows through a region of the gate oxide near the drain end which is $5\text{ }\mu\text{m}$ wide, and $0.2\text{ }\mu\text{m}$ long. Assume that 1 in 10^6 electrons of the gate current becomes trapped at an average distance of $0.1\text{ }t_{ox}$ from the Si-SiO₂ interface. How long does the gate current have to flow to change V_T by 0.1 V in the region where the injection is occurring?

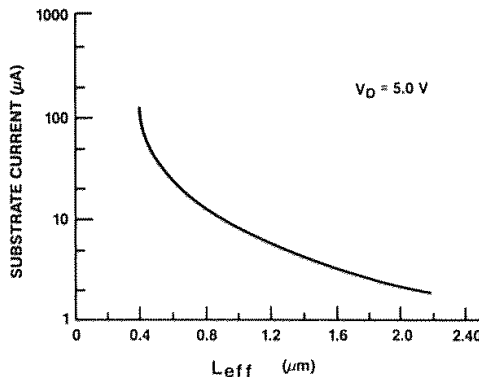


Fig. 5-28 The maximum substrate current due to impact ionization produced at a drain voltage of 5 V vs effective channel length for $250\text{\AA}$ gate oxide transistors.³⁴ (© 1986 IEEE).

SOLUTION: The gate current of 1 nA produces a current density across the injection area of

$$10^{-9} \text{ nA} / 1 \times 10^{-8} \text{ cm}^2 = 0.1 \text{ A cm}^{-2}$$

Since 1 in 10^6 electrons become trapped, the rate of charge trapping, J_{ot} , in the gate oxide is $0.1 \times 10^{-6} = 1 \times 10^{-7} \text{ C sec}^{-1} \text{ cm}^{-2}$. The shift in V_T due to the trapped charge in the oxide can be found from

$$\Delta V_T = (1/C_{ox}) (0.9 t_{ox}/t_{ox}) \Delta Q_{tot}$$

or, solving for ΔQ_{tot} ,

$$\begin{aligned} \Delta Q_{tot} &= C_{ox} \Delta V_T / 0.9 = [1.7 \times 10^{-7} \times 0.1] / 0.9 \\ &= 1.89 \times 10^{-8} \text{ C/cm}^2 \end{aligned}$$

The time, t , necessary to trap this quantity of charge is

$$t = \Delta Q_{tot} / J_{ot} = 1.89 \times 10^{-8} / 1 \times 10^{-7} = 0.19 \text{ sec}$$

5.6.3 Device-Performance Degradation Due to Hot-Carrier Effects

The increase that occurs in V_T leads to other changes in the MOS characteristics. First, the saturation current decreases because $(V_G - V_T)$ becomes smaller (Fig. 5-29a).⁷⁷ Second, as the substrate current increases, the transconductance decreases. Finally, since the trapped charge accumulates with time, the device performance will become unacceptable for a given application after a certain time of device operation.

A device lifetime, τ , can therefore be defined by selecting the maximum percentage of allowed degradation in the critical device parameter. It has been found, however, that this lifetime (regardless of whether V_T or g_m is monitored) can be related to I_{sub} by a power-law relationship (i.e., τ is observed to be inversely proportional to I_{sub} when plotted on a log-log plot). Figures 5-29b and c show τ as a function of I_{sub} for: (a) τ is defined as a 10-mV shift in V_T ; and (b) τ is defined as a 10% degradation in g_m .

This has led to accelerated testing techniques that stress the devices with higher drain voltages than would be used in normal operating conditions.³⁵ A larger I_{sub} is thereby produced, leading to shorter times to device failure (i.e., possibly defined by a 10-mV shift in V_T or a 10% degradation in g_m). The time to failure under normal operating drain voltages (τ) is then extrapolated from these data using curves as shown in Figs. 5-29b and 5-29c.^{36,37} Device design can then be modified to yield a desired lifetime (typically, 10 years) under normal operating conditions.

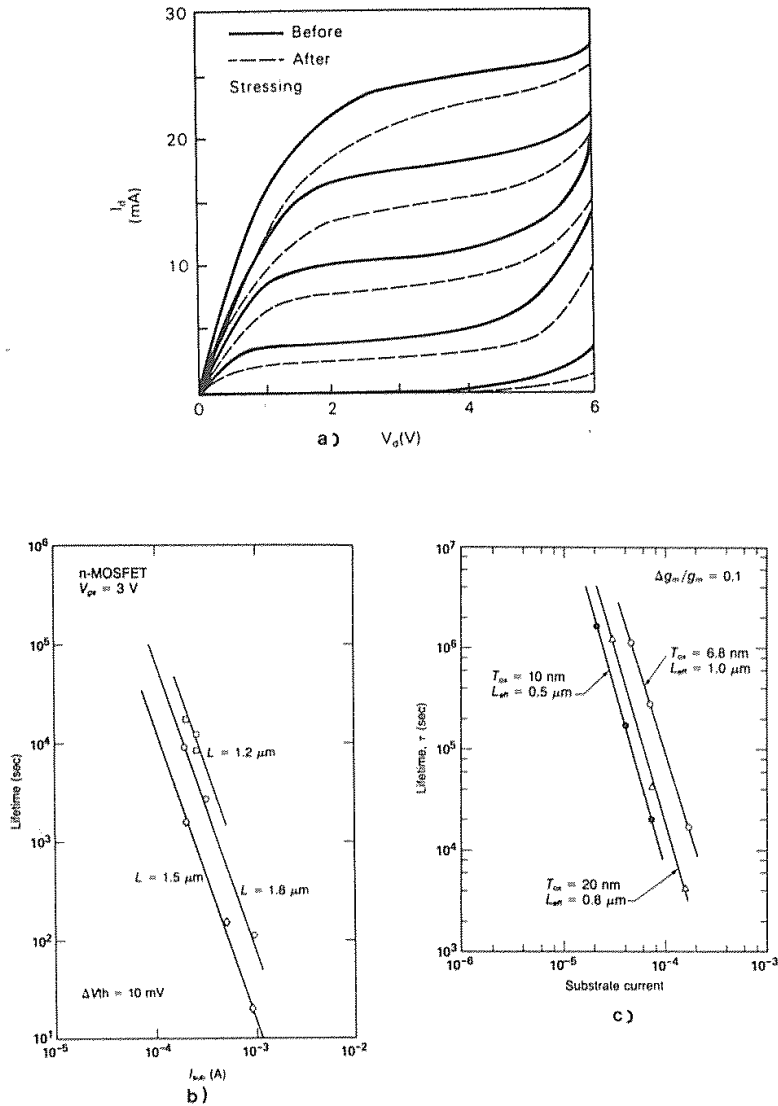


Fig. 5-29 (a) Degradation of I_{DSsat} after stressing.⁷⁷ (© 1984 IEEE). Degradation of other device-performance parameters as a function of substrate current (I_{sub}): (b) Lifetime defined as a 10-mV shift in V_T ; (c) Lifetime defined as a 10-percent degradation in g_m .

5.6.4 Techniques for Reducing Hot-Carrier Degradation

Device performance degradation from hot electron effects can be reduced by a number of techniques. These include the following:

- The voltages applied to the device can be decreased (e.g., by lowering the power-supply voltage from 5 V to 3.3 V). The decision to implement this reduction, however, is not in the hands of the device designer or fabricator. The issue of reduced supply voltages for submicron MOS technologies is covered in chapter 6, section 6.7.2.
- The time the device is under the voltage stress can be shortened (e.g., by using a lower duty cycle and clocked logic).
- Appropriate drain engineering design techniques, which results in special drain structures that reduce hot electron effects in MOS devices (i.e., double-diffused drains and lightly-doped drains), can be implemented.
- The density of trapping sites in the gate oxide can be reduced through the use of special processing techniques.
- Thin, lightly doped epitaxial-layers on low-resistance substrates can be employed to shunt away substrate current and help eliminate the problem of impact-ionization-induced latchup.

We will next discuss the details of the latter three approaches since their implementation is the task of device and process engineers.

5.6.5 Lightly Doped Drains (LDD)

It has been determined that hot-carrier effects will cause unacceptable performance degradation in NMOS devices built with conventional drain structures if their channel lengths are less than $2\text{ }\mu\text{m}$. To overcome this problem, such alternative drain structures as *double-diffused drains* and *lightly doped drains* (LDDs) must be used. The purpose of both types of structures is the same – namely, to absorb some of the potential into the drain and thus reduce E_M . Double-diffused drains, however, are less effective for short-channel devices (i.e., $\leq 1.25\text{ }\mu\text{m}$) because they cause deeper junctions and more overlap capacitance. We will thus restrict our discussion to LDDs. Table 5-2 shows the evolution in AT&T's Twin-Tub CMOS technology, with respect to such drain structures.³⁰

In the LDD structure, the drain is formed by two implants (Fig. 5-30). One of these is self-aligned to the gate electrode, and the other is self-aligned to the gate electrode on which two oxide *sidewall spacers* have been formed. (An extensive report on the details of sidewall spacer technology for both MOS and bipolar devices can be found in ref. 66.) The purpose of the lighter first dose is to form a lightly doped section of the

Table 5.2 Evolution of Device Structures in AT&T's Twin-Tub CMOS Technology Development³⁰ (Twin-Tub VI announced in 1989)⁸³

	Twin-Tub I	Twin-Tub II	Twin-Tub III	Twin-Tub IV	Twin-Tub V	Twin-Tub VI
Design Rule	3.5 μm	2.5 μm	1.75 μm	1.25 μm	0.9 μm	0.6 μm
L_{eff}	2 μm	1.5 μm	1.3 μm	1.0 μm	0.75 μm	0.4 μm
t_{ox}	600 Å	350 Å	250 Å	200 Å	150 Å	125 Å
Device Structure	Conventional	Conventional	DDD	LDD	N&P LDD	N&P LDD

drain at the edge near the channel. In NMOS devices, this dose is normally $1\text{--}2 \times 10^{13}$ atoms/cm² of phosphorus.

The value of E_M is reduced by this structure because the voltage drop is shared by the drain *and* the channel (in contrast to a conventional drain structure, in which almost

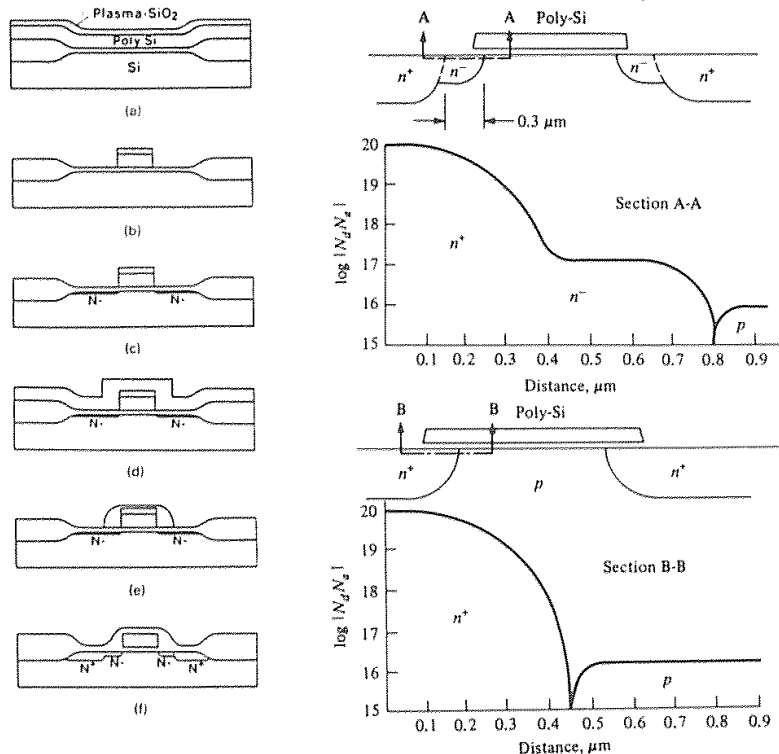


Fig. 5-30 (a) Process sequence used to form lightly doped drain (LDD) structures. (b) Doping profile in an LDD structure taken through section A – A. (c) Doping profile in a conventional drain structure taken through section B – B.

the entire voltage drop occurs across the lightly doped *channel* region). Figure 5-25 shows the electric-field profile at the drain of a MOSFET, both with and without an LDD structure.³⁸ We see that the electric field can be reduced by about 30-40%. Since the hot-electron-induced gate currents are exponentially dependent on E_M , this is sufficient to reduce currents by many orders of magnitude. As a result, the stability of the device is greatly increased. To ensure a high quality interface under the sidewall-spacer oxide, it is important that the thermally grown gate oxide remain in place after the polysilicon gate etch. This requires a polysilicon etch process with high selectivity to oxide, since the gate oxide is typically less than 20 nm.

The heavier, second dose forms a low resistivity region of the drain region, which is also merged with the lightly doped region. (In NMOS devices this implant is typically arsenic at a dose of about 1×10^{15} atoms/cm².) Since it is further away from the channel than would be the case in a conventional drain structure, the depth of the heavily-doped region of the drain can be made somewhat greater without adversely impacting the device operation (e.g., 0.3 μm deep versus 0.18 μm deep). The increased junction depth lowers both the sheet resistance and the contact resistance of the drain (see chap. 3). Deeper junctions also provide better protection against junction spiking.

The disadvantages of LDD structures are their increased fabrication complexity compared to conventional drain structures and the increased parasitic resistance of the source and drain regions caused by the lightly doped regions of the drain. This increase in parasitic channel resistance results in devices that dissipate more power for a constant applied voltage.

The effect on the I_D - V_{DS} curves of the MOSFET due to the additional voltage drop across the two lightly doped regions is that I_D does not saturate until a higher value of V_{DS} is applied. Due to the extra series resistance, the total channel conductance is appreciably lower in the linear region, but only slightly lower in the saturation region. This is because the channel region in saturation already has a high channel resistance, while in the linear region the resistance is much lower. The additional series resistance of the LDD therefore does not significantly increase the total resistance of the MOSFET in saturation, and I_D remains less affected. Consequently, the drain current is reduced more significantly in the linear region than in the saturation region.)

The I-V curves for LDD devices are consequently recognizable by their characteristically round shape, as shown in Fig. 5-31, curve 6. These curves also illustrate progressively the other short-channel effects described in section 5.5, beginning with the I-V curve of an ideal long-channel MOS device (Curve 1). Thus, curve 6 shows the I-V curve of a short-channel MOSFET with an LDD.

5.6.5.1 Drain Engineering for Optimum LDD Structures. A proper LDD design should provide adequate hot-carrier protection while not introducing excessive source/drain resistance, which would degrade the device performance. The design and fabrication effort undertaken to achieve such an optimum LDD structure is referred to as *drain engineering*; this as we shall see, is a relatively complex task. A thorough approach to designing adequate hot-electron protection would entail addressing all of the following objectives:

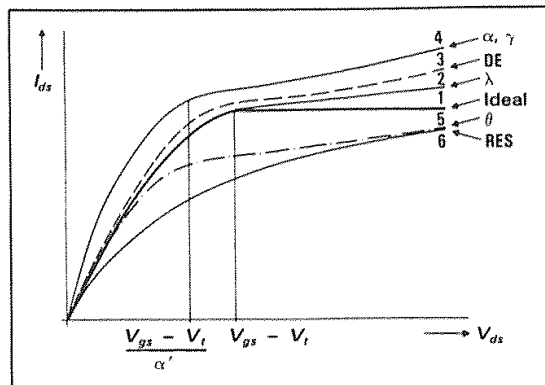


Fig. 5-31 Progressive influence of short-channel effects and LDD on the characteristics of an ideal, long channel MOSFET (Curve 1). Curve 6 shows the I-V curve of a short channel MOSFET that is also fabricated with an LDD, while Curve 5 shows the characteristic of the same device without an LDD.²⁸ (© 1986 IEEE).

1. Reducing the maximum electric field in the silicon as much as possible.
2. Ensuring that the injection position (i.e., the E_M point) is located under the gate edge.
3. Ensuring that the impact ionization region is pushed far below silicon surface to reduce the possibility of hot carriers reaching the Si-SiO₂ interface.
4. Separating the point where the electric field in the silicon is a maximum from the point of maximum current flow in the channel.
5. Minimizing the increase in the parasitic resistance due to the LDD structure.

The degree by which E_M is reduced (*Objective #1*) depends primarily on the doping value of the lightly doped extension of the drain, as well as on its length. If the doping level is too high, the value of E_M is not sufficiently reduced, and hot-electron protection is not provided. If the level is too low, the drive-current capability of the device is severely reduced, and the surface will easily be depleted by the hot carriers that do get trapped in the gate oxide above it (i.e., the device will again be vulnerable to hot-electron degradation). A model which calculates the electric field in LDD structures has been published.⁷² It concludes that the lightly doped region should be long enough to attenuate the electric field to a value that is below the critical ionization field, but should still be short enough to keep the series resistance from becoming excessive. In early LDD structures in NMOS this meant that the primary parameters selected were the n^- length and its doping concentration.³⁹

A process for forming LDD structures in *either* or *both* PMOS and NMOS structures in CMOS with only two photomasks was reported by Parrillo, et al.⁴⁰ An extension of this process also uses removable TiN gate sidewall spacers and incorporates self-alignment of the lightly doped regions to their respective gates without overcompensation (Fig. 5-32). Another example LDD structure in 0.8- μm CMOS is given in reference 41.

To keep the location of the E_M point under the gate (*Objective #2*) and yet maintain high drivability requires the proper combination of gate-drain (g/d) overlap length and spacer length. The g/d overlap length should be long enough to let the E_M point lie under the gate, and the spacer length should be short enough to let the n^+ region reach under the gate. In general, for process controllability, if a 0.3- μm -thick polysilicon film is used for the gate, a 0.2- μm -wide spacer is selected.

To attain *Objective #3*, a metal-coated LDD structure has been developed.⁴² It uses a deeper n^- phosphorus profile than the n^+ source-drain (arsenic), in order to steer the maximum current path away from E_M (Fig. 5-33b).

A buried LDD structure has also been proposed to reach *Objectives #3 and #4*. This structure uses a retrograde LDD profile, with the peak concentration below the Si surface (Fig. 5-33c). Besides a reduction in substrate current similar to the metal-coated LDD, this profile also suppresses hot-carrier injection by driving the current away from the surface and shifting the avalanche region further into the bulk silicon.

The buried LDD approach together with an additional shallow arsenic implant (Fig. 5-34d), may allow NMOS transistors of 0.6 μm to be built with adequate hot-electron protection and with minimal the increase in parasitic resistance due to the LDD structure (*Objective #5*).^{30,43,44} The idea is to form an abrupt profile at the silicon surface underneath the spacer to reduce the series resistance. The buried LDD profile is

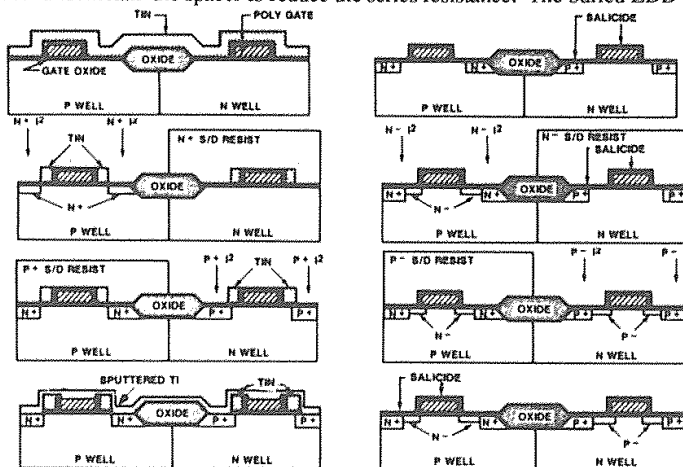


Fig. 5-32 Process sequence showing the TiN removable spacer process.⁸² (© 1989 IEEE).

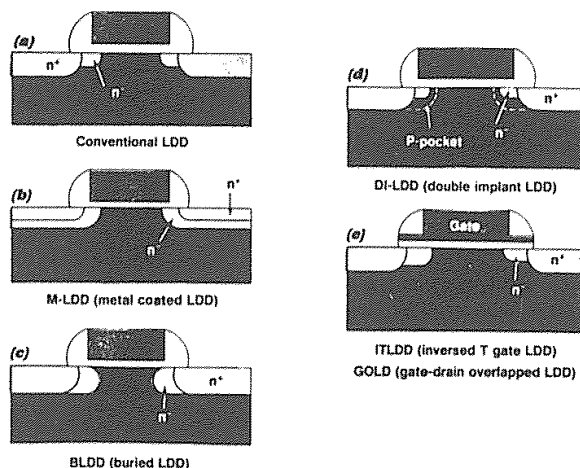


Fig. 5-33 Conventional LDD drain engineering and some of its variations – all designed to reduce hot-carrier effects – are shown.³⁰ (© 1986 IEEE).

used to force the main current path and the impact-ionization region deep into the silicon. There are two peaks in the lateral electric-field distribution, and the maximum current path is through the saddle point of these two high field regions. Thus, this device structure improves both the current drive and the hot-carrier resistance. A further modification of the buried LDD is the *sloped-junction LDD (SJLDD)*, shown in Fig. 5-33e.⁴⁵ In this structure a 165-keV phosphorus implant is used to form the lightly doped drain region; this provides improved device lifetime under high field stress. The main reason for the improvement is that hot-carrier generation is driven further away from the Si-SiO₂ interface by the junction of the SJLDD.

5.6.5.2 Asymmetrical Characteristics of LDD MOSFETs. Formation of the lightly doped regions of LDDs is accomplished by means of ion implantation, with the polysilicon gate used as the implant mask. This can cause asymmetric doping of the source and drain regions,^{46,64} with such asymmetry occurring as a result of implanting off-axis (typically, at an angle of 7°) to avoid channeling (see Vol. 1, chap. 9). This produces lateral shadowing (S in Fig. 5-34) of the substrate on one side of the poly gate, and penetration of dopants through the leading corner of the poly on the other side. The problem is compounded by etch processes that produce a re-entrant angle in the poly gate sidewall (Fig. 5-34 - poly etch process A.) The impact of this effect on the device characteristics can be examined by considering the implants to the lightly doped and the heavily doped regions separately.

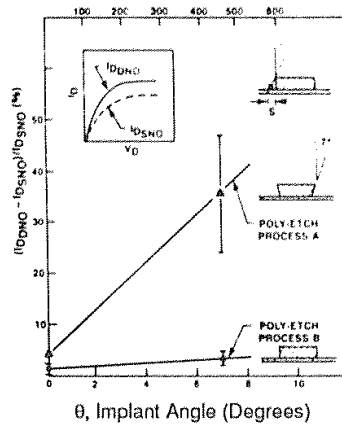


Fig. 5-34 Percent asymmetry in saturated drain currents ($V_{DS} = V_{GS} = 5$ V) versus source/drain implant angle for LDD NMOS devices having reentrant- and vertical-wall polysilicon gates. Insert schematically shows source-non-overlap (SNO) and drain-overlap (DNO) I-V behavior.⁴⁰ (© 1986 IEEE).

If the implant that forms the lightly doped region of the drain (usually phosphorus in NMOS devices) is done off-axis, the as-implanted region in the shadow of the gate will have its edge displaced from the gate edge. As a result, there will be *no overlap* of the implanted region and the gate edge. Even after a drive-in step, the overlap of this edge with the gate will be less than the overlap of the nonshadowed implant region edge and the gate. If the less-overlapped side is used as the source, the extra resistance introduced will reduce the drain current ($I_{D_{SNO}}$ in Fig. 5-34), thus degrading circuit speed. However, if the drain side has less overlap, the extra resistance will not affect I_D (curve $I_{D_{DNO}}$ in Fig. 5-34). The asymmetry in drive current when the same device is connected in these two ways can be as much as 40%; this can be catastrophic in circuits that require closely matched device characteristics.

The problems arising from such n^- implant-shadowing can be alleviated by one of the following measures:

- Insuring that the poly-etch process produces vertical sidewalls in the poly gates.
- Avoiding excessive reoxidation of the polysilicon after its definition (note that this also prevents a gap from being established between the drain region edge and the edge of the gap). A dry-oxidation step is often employed to reduce the accelerated oxidation rate on n^+ polysilicon.
- Implanting with a vertical implant through a screen oxide²⁷ (which reduces the channeling that would otherwise occur when a vertical implant is performed).

- Using a poly-etch process that produces a slight positive bevel to the poly sidewall profile, and combining this etch with a smaller angle (3° vs 7°) off-axis implant.⁴⁶

Let us next examine how the implantation of the heavily doped region impacts the structure of the LDD. If an off-axis implant is again used with the heavily doped implant, asymmetric implant effects can again alter the LDD structure. That is, on the side of the gate that is *not* shadowed, the arsenic atoms penetrate the leading corner of the spacer and thereby enter the substrate under the spacer to some extent. If the spacer is too narrow, this can wipe out the lightly doped drain-extension region. When this happens, the device can lose the hot-carrier protection that was to be provided by the LDD structure, and it will become vulnerable to degradation by hot-carrier effects. To avoid this, the spacer must be sufficiently wide. Some guidelines for choosing the proper spacer width to deal with this problem are given in reference 47.

The general problem involving the gate-to-drain overlap with respect to its impact on the MOSFET characteristics was reviewed by Ko, et al.⁴⁸ It was observed that the critical dimensions in drain structures having weak overlap (WO) to the gate were only tens of nm in devices in which $L_{\text{eff}} = 1 \mu\text{m}$. In general, the drain current decreases as the overlap is weakened, a double hump appears in the substrate current of asymmetrical WO devices, and the reliability of the WO devices can be lower than devices having adequate overlap. To avoid random degradation of device performance by WO effects that arise as a result of process variations, stringent process-control measures must be exercised when submicron devices are fabricated.

5.6.6 The Impact of IC Processing on Hot-Carrier Device Degradation

The lifetime of a device in which the hot-electron degradation impacts device performance can be increased by keeping the number of trapping centers in the gate oxide to a minimum. In essence, this reduces the density of states that the hot electrons injected into the oxide can occupy. Maintaining a minimum number of trapping centers can be achieved in several ways during the device-fabrication process sequence. First, the gate oxide should be grown by a process that produces low interface-trap densities (such processes are described in Vol. 1, chap. 7). It has also recently been reported that the incorporation of optimized amounts of fluorine during the gate-oxide growth process suppresses interface-state generation during injection.⁴⁹ Second, the degradation of the oxide caused by damage during process steps carried out in a plasma environment should be minimized (damaging processes include plasma-enhanced CVD of oxides and nitrides, RIE, and sputter deposition of metal). Finally, the amount of hydrogen that is incorporated at the Si-SiO₂ interface should be reduced. The latter two topics will be considered in more detail here.

In typical CMOS processing, most of the damage created by RIE or ion implantation is annealed out at a high temperature (above 800°C) before metallization. Once the first layer of aluminum is deposited, however, the maximum annealing temperature is

limited to $\sim 450^\circ\text{C}$. It has therefore been reported that the RIE of the second layer of metal in multilevel metal processes deteriorates the device-aging characteristics. It has been found that while a subsequent anneal performed at 375°C is ineffective in annealing out this damage,⁵⁰ an anneal at 450°C improves device-aging characteristics.

Excess hydrogen at the Si-SiO₂ interface is also reported to be a culprit in increasing the density of the interface states. In most cases, hydrogen is used to fill the dangling bond, forming Si-H at the Si-SiO₂ interface. However, the Si-H bond can be easily broken by injected hot electrons. Furthermore, excess hydrogen introduced during processing can diffuse to the interface and lead to enhanced bond-breaking behavior. Such excess hydrogen can arrive during the final sinter step before passivation (i.e., in H₂ or N₂ + 5% H₂), or during the deposition of a conventional silicon nitride passivation layer. It has therefore been reported that sintering in pure N₂ produces devices with a lower device degradation rate.⁵¹ Similarly, it has been reported that when a fluorinated nitride (F-SiN) is used as a passivation layer,^{50,52} devices exhibited slower degradation rates than those seen in devices with conventional SiN passivation layers. F-SiN films can be produced by incorporating NF₃ in the deposition process to form an Si-F bonding structure instead of Si-H in this film.

The presence of fluorine in the gate oxide (possibly inadvertently originating from a BF₂⁺ source/drain implant) has recently been reported to increase the hot electron resistance of devices fabricated with such oxides.⁷⁴ Other reports confirm that a deliberate incorporation of fluorine into the gate oxide (from implanting fluorine into a polysilicon film, and then diffusing it into the gate oxide) produces a more hot-electron resistant interface.^{78,79}

5.6.7 Hot-Carrier Effects in PMOS Transistors

Hot-carrier effects are not significant in PMOS transistors at channel lengths greater than $1\ \mu\text{m}$ because the impact-ionization rate of holes is 3 to 4 orders of magnitude lower than that of electrons at a given electric field.⁵³ At submicron channel lengths, however, hot-electron effects in PMOS start to become important. It has been reported that two hot-carrier effects predominate in such PMOS devices. Both are caused by hot electrons that are generated by impact ionization and are then injected into the oxide, becoming trapped. Hot holes do not appear to cause significant effects unless a device is stressed at large magnitudes of V_G . In PMOS devices, the gate-bias polarity favors electron injection (which is opposite to that in NMOS) but retards the injection of holes.

In the first degradation effect, these electrons are trapped near the drain and shorten the effective channel length. This reduction in L_{eff} is even more severe in the buried-channel PMOS devices (used in CMOS technologies that have an n^+ polysilicon gate), because their buried-channel nature makes them more vulnerable to source-drain punchthrough,⁵⁴ and consequently increases subthreshold leakage. This effect is called *hot-electron-induced punchthrough* (HEIP).

The second effect was observed in PMOS devices with p^+ , as well as n^+ , polysilicon gates. In these devices, the injected and trapped hot electrons also produce V_T shifts

(which reduce the magnitude of V_T) and increases in the transconductance, g_m . Nevertheless, contrary to what occurs in NMOS, in PMOS these effects tend to saturate. This is explained by the fact that the electrons trapped near the drain reduce the electric field present there in PMOS devices.⁵⁵ As a result, this effect does not appear to limit PMOS devices fabricated with p^+ polysilicon gates as long as L_{eff} is $\geq 0.6 \mu\text{m}$.⁵⁶ Still, LDD structures appear to be needed for p -channel devices fabricated with n^+ polysilicon gates when L_{eff} gets to be $0.8 \mu\text{m}$ or less.^{54,57}

The LDD structure also helps to reduce subthreshold leakage in PMOS devices, where excessive junction depths caused by lateral diffusion have been a problem. The use of LDD basically increases L_{eff} , hence compensating for the L_{eff} reduction caused by hot-electron injection. A new LDD structure for PMOS devices, called a *halo LDD*, is described in reference 63. In this structure, a deeper phosphorus implant is placed below the lightly doped drain-extension p -type implant. The punchthrough resistance of the PMOS device is reported to be significantly improved by this LDD structure.

5.6.8 Gate-Induced Drain-Leakage Current

Another type of leakage current between drain and substrate in thin gate-oxide (12-28 nm) MOS devices has been observed at drain voltages much lower than the breakdown voltage.^{58,59} This is not really a hot-electron-induced effect, but since it uses LDD structures to reduce its magnitude, we describe it in this section. The basis of this current is band-to-band tunneling that occurs through the gate oxide into the deep depletion layer in the gate-to-drain overlap region. It has been reported that in order for this leakage current to be limited to less than $0.1 \text{ pA}/\mu\text{m}^2$, the oxide field in the gate-to-drain overlap region must be limited to 1.9 MV/cm . LDD structures are effective in reducing this leakage current, but only if the doping in the n^- extensions is less than $10^{18}/\text{cm}^3$. Buried LDD structures with the peak of the n concentration at several hundred angstroms beneath the Si-SiO₂ interface appear to be even more effective than conventional LDD structures in reducing the effect. This effect is predicted to become more of a problem as devices continue to be scaled - that is, if a 10-nm-thick gate oxide is used, the E-field will be less than 1.9 MV/cm in the oxide only if a 3.1 V power supply is used.

REFERENCES

1. R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 2nd Ed., New York, John Wiley & Sons, 1986.
2. R. F. Pierret, *Field-Effect Devices*, Vol. IV in the Modular Series on Solid-State Devices, Reading, MA, Addison-Wesley, 1983.
3. D. K. Schroder, *Advanced MOS Devices*, Vol. VII in the Modular Series on Solid-State Devices, Reading, MA, Addison-Wesley, 1987.
4. Y. P. Tsividis, *The Operation and Modeling of the MOS Transistor*, New York, McGraw-Hill Book Co., 1987.

5. E. H. Nicollian and J.R. Brews, *MOS Physics and Technology*, Wiley-Interscience, New York, 1982.
6. S. M. Sze, *Physics of Semiconductor Devices*, 2nd Ed. Wiley, New York, 1981.
7. D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, McGraw-Hill Book Co., New York, 1983.
8. R. A. Chapman et al., *Tech. Dig. IEDM*, 1987, p. 362.
9. R. J. Bayruns et al., *IEEE J. Solid-State Circuits*, SC-19, 1984, p. 755.
10. H. L. Armstrong, "A Theory of Voltage Breakdown of Cylindrical P-N Junctions," *IRE Trans. Electron Dev.*, ED-4, 15 (1957).
11. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967, Chap. 6
12. E. C. Douglas, "Advanced Process Technology for VLSI Circuits," *Solid State Tech.*, 24, May 1981, p. 65.
13. C. M. Osborn and E. Bassous, *J. Electrochem. Soc.*, 122, 89, (1975).
14. S. M. Sze, *Semiconductor Devices: Physics & Tech.*, Wiley, New York, 1985.
15. N. Kotani and S. Kawazu, *Solid-State Electronics*, 22, p. 63, (Jan. 1979).
16. T. Yamaguchi et al., *IEEE Trans. Electron Dev.*, ED-31, 1984, p. 205.
17. C. Mead and L. Conway (Eds.), *Introduction to VLSI Systems*, Reading, MA, Addison-Wesley, 1980.
18. A. Mukherjee, *NMOS and CMOS VLSI Systems Design*, Prentice-Hall, Englewood Cliffs, N.J., 1986.
19. N. Kotani and S. Kawazu, *Solid-State Electronics*, 22, 63 (January 1979).
20. L. D. Yau, *Solid-State Electronics*, 17, 1059, (1974).
21. L. M. Dang, *IEEE J. Solid-State Circuits*, SC-14, 358 (1975).
22. G. Merkel, *Process and Device Modelling for IC Design* (F. Van de Wiele, W. L. Engle, and P.G. Jespers, Eds.), Noordhoff, Leyden (1977), p. 705.
23. R. R. Troutman, "VLSI limitations from Drain-Induced Barrier Lowering," *IEEE Trans. Electron Dev.*, ED-26, p.461, 1979.
24. J. Zhu et al., "Punchthrough Current for Submicron CMOS VLSI," *IEEE Trans. Electron Dev.*, Feb. 1988, p. 145.
25. K. M. Cham, et al., *Computer Aided Design and VLSI Device Development*, Kluwer Academic Publishers, Boston, 1986, p. 194.
26. H. Nihara, et al., "Anomalous Drain Current in NMOS and its Suppression by Deep Ion Implantation," *IEDM Tech. Dig.*, p. 487 (1978).
27. L. C. Parrillo, "CMOS Active and Field Device Fabrication," *Semiconductor International*, April 1988, p. 64.
28. C. Duvvury, "A Guide to Short-Channel Effects in MOSFETs," *IEEE Circuits and Systems Magazine*, Nov. 1986, p. 6.
29. M. Kakamu et al., "Power Supply Voltage for Future CMOS VLSI in Half- and Sub Micrometer," *Tech. Dig. IEDM*, 1986, p. 399.
30. M. L. Chen, "CMOS Hot-Carrier Protection with LDD," *Semiconductor Internat.*, April 1988, p. 78
31. P. K. Ko, "Hot Electron Effects in MOSFETs," Doctoral Thesis, Dept. of EECS, University of California, Berkeley, June 1972.

32. T. Y. Chan, P.K. Ho, and C. Hu, *IEEE Electron Dev. Letts.*, EDL-6, p. 551, 1985.
33. P. K. Chatterjee, *Tech. Dig. IEDM*, 1979, p. 14.
34. M. L. Woods, "MOS VLSI Reliability and Yield Trends," *Proceedings IEEE*, Dec. 1986, p. 1715.
35. P. Yang and S. Aur, "Modelling of Device Lifetime due to Hot Carrier Effects," *Proc. of 1985 Internat. Symp. on VLSI Tech.*, p. 227.
36. C. Hu et al., *IEEE J. Solid-State Circuits*, SC-20, 1985, p. 295.
37. E. Takeda and N. Suzuki, "An Empirical Model for Device Degradation due to Hot Carrier Injection," *IEEE Trans. Electron Dev. Letts.*, EDL-4, p. 111, 1983.
38. S. Ogura et al., *IEEE Trans. Electron Dev.*, ED-27, 1980, p. 1359.
39. H. Mikoshiba, "Comparison of Drain Structures in n-Channel MOSFETs," *IEEE Trans. Electron Dev.*, Jan. 1986, p. 140.
40. L. C. Parrillo et al., *IEDM Tech. Dig.*, 1986, p. 244.
41. R.A. Chapman et al., *IEDM Tech. Digest*, 1987, p. 362.
42. Y. Tsunashima et al., *Proc. of VLSI Symposium*, 1985, p. 114.
43. T. Toyoshima et al., *Proc. of VLSI Symposium*, 1985, p. 362.
44. T. Noguchi, *Tech. Dig. IEDM*, 1986, p. 730.
45. S. Jain et al., *IEEE Electron Dev. Lett.*, Oct 1988, p. 539.
46. T. Y. Chan et al., *Electron Dev. Lett.*, Jan. 1986, p. 16.
47. J. R. Pfister and F.K. Baker, "Asymmetrical High Field Effects in Submicron MOSFETs," *IEDM Tech. Dig.*, 1987, p. 51.
48. P. K. Ko et al., *Tech Dig. IEDM*, 1986, p. 292.
49. E. F. Da Silva, *Tech. Dig. IEDM*, 1987, p. 848.
50. M. -L. Chen, "Hot Carrier Aging in Two-Level Metal Processing," *Tech. Dig. IEDM*, 1987, p. 55.
51. F. C. Hsu, *Proc. of VLSI Symp.*, 1985, p. 108.
52. S. Fujita et al., *Tech. Dig. IEDM*, 1985, p. 64.
53. J. Y. Chen, "CMOS-The Emerging VLSI Technology," *IEEE Circuits and Devices Magazine*, March, 1986, p. 16.
54. M. Koyanagi et al., *IEEE Trans. Electron Dev.*, ED-34, 1987, p. 871.
55. C. Hu et al., *IEEE J. Solid-State Circuits*, SC-20, 1985, p. 295.
56. Y. Hiruta et al., *Tech. Dig. IEDM*, 1986, p. 718.
57. T. Mizuno et al., *Tech. Dig. IEDM*, p.726.
58. T. Y. Chan, J. Chen, P.K. Ko, and C. Hu, *IEDM Tech. Dig.* 1987, p. 718.
59. C. Chang and J. Lien, *IEDM Tech. Dig.* 1987, p. 714.
60. C. T. Sah, "The Evolution of the MOS Transistor," *Proceedings of the IEEE*, Oct. 1988, p. 1280.
61. P. Balk, "Effects of Hydrogen Annealing on Silicon Surfaces," presented at the *Electrochem. Soc. Spring Meeting*, May, 1965. Ext Abs. No. 109, p. 237.
62. M.-C. Jeng et al., "Design Guidelines for Deep-Submicrometer MOSFETs," *Tech. Dig. IEDM*, 1988, p. 386.
63. M. L. Chen et al., *Tech. Dig. IEDM*, 1988, p. 390.
64. R. W. Gregor, "Consequences of Ion Beam Shadowing in CMOS Source/Drain Formation," *IEEE Electron Dev. Letts.*, Dec. 1986, p. 677.

65. I. Ahmed, H. Naguib, and C. Gomez, *Ext. Abs. ECS Meeting*, Fall 1988, Abs. No. 282, p. 405.
66. S. H. Dhong and E. J. Petrillo, *J. Electrochem. Soc.*, Feb. 1986, p. 389.
67. F. Hsu and K. -Y. Chiu, *IEEE Electron Dev. Letts.*, EDL-5, p. 162, 1984.
68. K. - Y. Toh, P.-K. Ko, and R. G. Meyer, *IEEE J. Solid-State Ckts*, Aug. 1988, p. 950.
69. J. W. Brews et al., "Generalized Guide for MOSFET Miniaturization," *IEEE Electron Dev. Lett.*, EDL-1, p. 2, 1980.
70. R. H. Dennard et al., "Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions," *IEEE J. Solid-State Circuits*, SC-9, Oct. 1974, p. 256.
71. S. G. Chamberlain and S. Ramanan, *IEEE Trans. Electron Dev.*, Nov. 1986, p. 1745.
72. K. Marayam, J. C. Lee, and C. Hu, *IEEE Trans. Electron Dev.*, July 1987, p. 1509.
73. L. D. Yau, "A Simple Theory to Predict the Threshold Voltage in Short-Channel IGFETs," *Solid-State Electronics*, vol. 17, p. 1059, 1974.
74. P. Wright et al., *IEEE Electron Device Letts.*, August 1989, p. 347.
75. 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).
76. R. Siquesch et al., *Tech. Dig. IEDM*, 1980, p. 429.
77. F. H. Hsu and K. Y. Chiu, *Tech. Dig. IEDM*, p. 96, 1984.
78. P. Wright et al., "Hot-Electron Immunity of SiO₂ Dielectrics with Fluorine Incorporation," *IEEE Electron Device Letters*, August 1989, p. 347.
79. Y. Nishioka et al., "Hot-Electron Hardened Si-Gate MOSFET Utilizing F Implantation," *IEEE Electron Device Letters*, April 1989, p. 141.
80. W. M. Werner, *Solid-State Electronics*, 17, 769, (1974).
81. R. R. Troutman, *IEEE J. Solid-State Circuits*, SC-9, 55 (April 1974).
82. J. R. Pfister et al., *Tech. Dig. IEDM*, 1989, p. 781.
83. C.-Y. Lu et al., *IEEE Trans. Electron Dev.*, November, 1989, p. 2530.

PROBLEMS

- 5.1 (a) A *p*-channel MOS transistor is fabricated on an *n* substrate doped with 10^{15} phosphorus atoms per cm³, and a gate oxide which is 1000 Å thick. Calculate the threshold voltage if $\phi_{ms} = -0.6$ eV and $Q_{tot} = 5 \times 10^{11}$ cm⁻². (b) The threshold voltage of the MOSFET in part (a) is reduced by using ion implantation of boron. What is the required boron dose in order to obtain a threshold voltage of -1.5 V?
- 5.2 In the E-D NMOS process, depletion-mode NMOSFETs are used for load devices. This requires a negative threshold, which can be obtained by implanting a shallow arsenic or phosphorus dose into the channel region. Calculate the arsenic dose needed to achieve a -3 V threshold in a *n*⁺ polysilicon-gate NMOS transistor which has a substrate doping of 3×10^{16} /cm³ and a gate oxide thickness of 50 nm?
- 5.3 (a) Why is <100>-orientation preferred in NMOS fabrication? (b) What are the disadvantages if too thin a field oxide is used in NMOS devices? (c) How is a self-aligned gate obtained and what are its advantages? (d) List three functions of the PSG overglass layer?
- 5.4 In NMOS processing, the starting material is a *p*-type 10-Ω-cm, <100>-oriented silicon wafer. The source and drain are formed by arsenic implantation of 10^{16} ions/cm² at

80 keV: the channel is implanted with 8×10^{11} boron ions/cm² at 30 keV through a gate oxide of 250 Å thickness. (a) Estimate the threshold voltage change caused by the boron ion implantation step. (b) Draw the doping profile along a coordinate perpendicular to the surface and passing through (b1) the channel region, and (b2) the source region.

5.5 Describe punchthrough current and subthreshold current and explain the differences between them.

5.6 Design a submicron MOSFET with a gate length of 0.75 μm. (The gate length is the channel length plus twice the junction depth.) If the junction depth is 0.2 μm, the gate oxide is 200 Å, and the maximum drain voltage is limited to 2.5 V, find the required channel doping so that the MOSFET can maintain its long-channel characteristics.

5.7 An *n*-channel MOSFET has a gate oxide 200 Å thick, an $L_{\text{eff}} = 1.0$ μm, source/drain junction depths of 0.2 μm, and a threshold voltage of 0.6 V. If the device is biased at $V_{\text{GS}} = 3$ V and $V_{\text{DS}} = 5$ V, calculate the saturation voltage V_{Dsat} , and the maximum electric field, E_{max} .

5.8 Explain the difference between the gate-threshold voltage, V_T , and the field-threshold voltage, V_{TF} , of a MOS device.

5.9 (a) Sketch a set of masks for fabricating an MOS transistor using Si-gate technology. (b) Repeat this exercise for an E-D NMOS inverter circuit with LOCOS isolation, buried contacts, and a depletion-mode load transistor.

5.10 An NMOS structure consists of an *n*-type substrate with $N_D = 5 \times 10^{15}$ cm⁻³, a gate oxide of 100 nm thickness, and an Al contact. The measured threshold voltage is -2.5 V. Calculate the surface charge density.

5.11 If a circuit designer wants to keep the off-state leakage current to less than 1 pA and also wants the threshold voltage value to be 0.9 V (where V_T is defined as the gate voltage that provides $I_{\text{DS}} = 1$ μA), what value of subthreshold swing (S.S.) will this device exhibit.

CHAPTER 6

CMOS PROCESS INTEGRATION

Complementary MOS (CMOS) is so-named because it uses both *p*- and *n*-type (complementary) MOS transistors in its circuits. (Figure 6-1 depicts a CMOS inverter.) Since CMOS technology is significantly more complex than NMOS with respect to the device physics and fabrication issues, the discussion here will include a thorough introduction to these subjects. (Note that an excellent comprehensive text on CMOS, dealing with both circuit and design issues, has recently been published.)¹

6.1 INTRODUCTION TO CMOS TECHNOLOGY

6.1.1 The Power-Dissipation Crisis of VLSI, and How CMOS Came to the Rescue

NMOS remained the dominant MOS technology as long as the integration level of devices on a chip was sufficiently low. It was inexpensive to fabricate, very functionally dense, and faster than PMOS. The earliest NMOS technologies required only 5 masking steps (including the pad mask).^{*} On the other hand, NMOS logic-gates (e.g., inverters) draw dc power during one of the inverter states. Therefore, an NMOS integrated circuit will draw a steady current even when being operated in the standby mode (i.e., even when no signal is being propagated through the circuit). Consequently, as the number of logic gates on the chip grows, the current being drawn (and hence the power being dissipated) also increases. Although this was always a limitation of NMOS for such applications as space-borne or portable electronic systems, it did not represent a drawback for most other applications when the number of devices on a chip was rela-

^{*} Early NMOS logic gates used butted contacts as well as enhancement-mode transistors for both the driver and the load. Depletion-mode loads and buried contacts came later. Hence, of the seven masks of the E-D NMOS process described in chapter 5, only five were required in early NMOS technology. This resulted not only in lower manufacturing costs per wafer, but also in higher yields. Thus, the cost of manufacturing NMOS circuits was brought even lower.

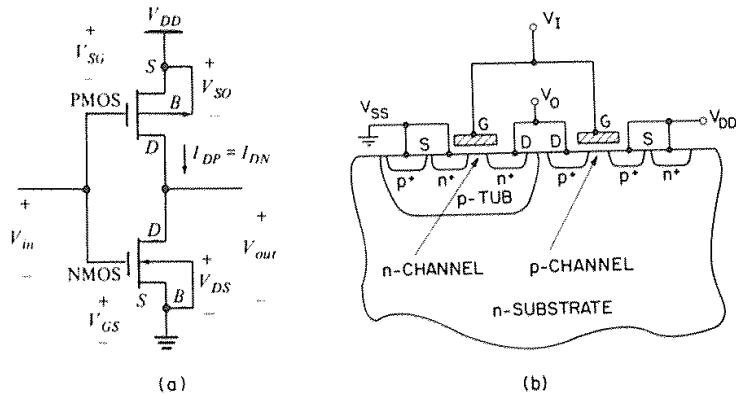


Fig. 6-1 CMOS inverter. (a) Circuit schematic. (b) Device cross section.

tively small. Such was the situation at the level of device integration that existed up to the mid-1970s.

With the dawning of the VLSI era, however, power consumption in NMOS circuits began to exceed tolerable limits. A lower-power technology was needed to exploit the VLSI fabrication techniques. CMOS represented just such a technology.

From a quantitative perspective, the ascendancy of CMOS was the inevitable result of the two-hundred-fold increase in functional density, and the twenty-fold increase in speed of integrated circuits between 1968 and 1987. For example, in 1969, 256-bit SRAM circuits (e.g., the Intel 1101) used 12- μm design rules to create the six-transistor SRAM cells; each cell occupied 20,600 μm^2 . In 1987, however, SRAM cells using 1.2- μm design rules occupied only 150 μm^2 . (The memory access-time in these respective memory circuits decreased from 1 μs to 46 ns.) By 1987 the decreased size and attendant increase in chip size allowed 256-kbit SRAMs to be built.

To take another example: The Intel 4004 4-bit microprocessor, introduced in 1971, had 2300 devices and was built in PMOS. The 8086 model, a 16-bit microprocessor introduced in 1978, had 29,000 devices and was built in NMOS, and the 80386, introduced in 1985, had 275,000 devices and was built in CMOS.

Chips can dissipate a maximum of about 5 W of power and still be used in conventional, but expensive, IC ceramic packages. In order for the much less expensive plastic packages to be used, however, the maximum power dissipation is limited to about 1 W. The Intel 8086 dissipated around 1.5 W of power when operated at 8 MHz. Thus, by the late 1970s it was already possible to manufacture NMOS chips whose power dissipation approached unacceptably high values. (Note that when the 8086 was later reissued in CMOS technology under the model number 80C86, its power consumption dropped to about 250 mW.)

In a CMOS inverter (unlike in an NMOS inverter) only one of the two transistors is driven at any one time. This means that when a CMOS inverter is not switching from one state to the other, a high-impedance path exists from the supply voltage to ground, regardless of the state the inverter is in. Hence, virtually no current flows, and almost no dc power is dissipated. CMOS thus allows the manufacture of circuits that need only several microwatts of standby power (Fig. 6-2).

The problem of power dissipation can also be considered from both a *chip* perspective and a *system* perspective. From the chip perspective, if microprocessors of the 32-bit generation were built in NMOS, they would dissipate 5 to 6 W of power. This would lead to severe heating and reliability concerns. In addition, expensive ceramic packages would be needed to house such chips. When such microprocessors are designed in CMOS, the power dissipation decreases to about 1 W.

From the system perspective, let's consider memory chips. Although a 1-Mbit DRAM may consume only 120 mW of power in NMOS, it consumes even less (~50 mW) in CMOS. Since there may be thousands of memory chips in a system (versus only a few microprocessors), the ramifications of lower power-dissipation are significant. Smaller power-supplies and smaller cooling fans are but two of these ramifications.

6.1.2 Historical Evolution of CMOS

Although CMOS is now the dominant integrated-circuit technology, for much of its 25-year history it was considered to be only a runner-up for the design of MOS ICs. The pairing of complementary *n*- and *p*-channel transistors to form low-power ICs was originally proposed by Sah and Wanlass in 1963.² The first CMOS ICs were fabricated

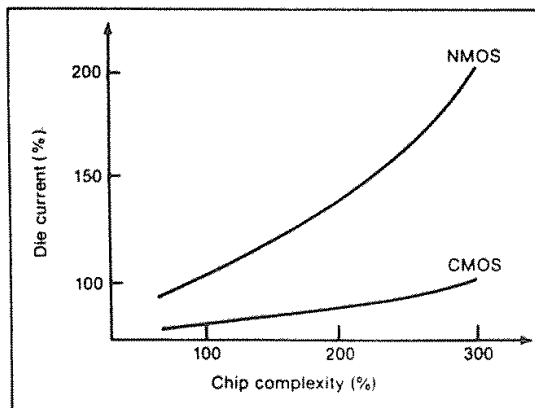


Fig. 6-2 As circuit complexity increases, NMOS power consumption rises to levels that eventually prevent further growth. In contrast, CMOS power consumption increases only slightly as the device count on a chip rises.