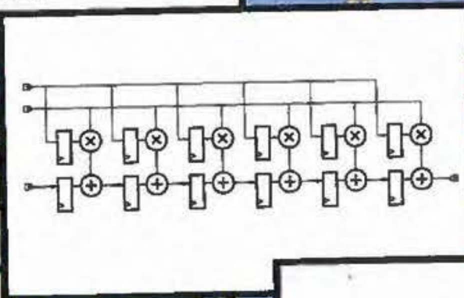
PRINCIPLES OF CMOS VLSI DESIGN

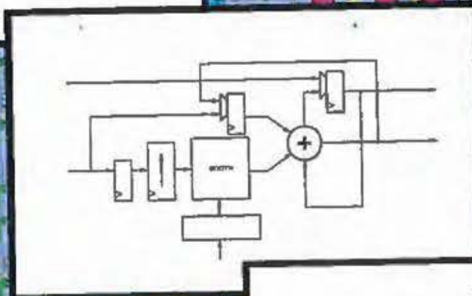
A Systems Perspective

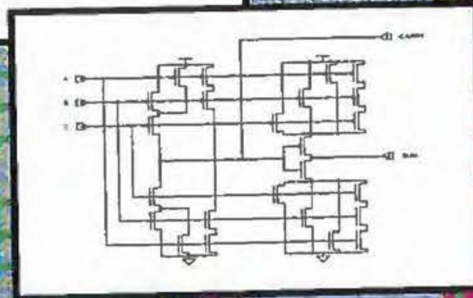
SECOND EDITION

NEIL H. E. WESTE
KAMRAN ESHRAGHIAN

$$H(n) = \sum_{n=0}^{M} k_n x_n$$







This book is in the Addison-Wesley VLSI Systems Series.

Lynn Conway and Charles Seitz, *Consulting Editors*

The VLSI Systems Series

Circuits, Interconnections, and Packaging for VLSI by H. B. Bakoglu

Analog VLSI and Neural Systems by Carver Mead

The CMOS3 Cell Library edited by Dennis Heinbuch

Computer Aids for VLSI Design by Steven Rubin

The Design and Analysis of VLSI Circuits by Lance Glasser and Daniel Dobberpuhl

Principles of CMOS VLSI Design by Neil H. E. Weste and Kamran Eshraghian

Also from Addison-Wesley:
An Introduction to VLSI Systems by Carver Mead and Lynn Conway

LG Display Co., Ltd.
Exhibit 1012
Page 002

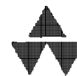
PRINCIPLES OF CMOS VLSI DESIGN

A Systems Perspective

Second Edition

Neil H. E. Weste
TLW, Inc.

Kamran Eshraghian
University of Adelaide



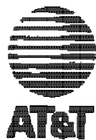
ADDISON-WESLEY PUBLISHING COMPANY
Reading, Massachusetts • Menlo Park, California • New York
Don Mills, Ontario • Wokingham, England • Amsterdam • Bonn
Sydney • Singapore • Tokyo • Madrid • San Juan • Milan • Paris

LG Display Co., Ltd.
Exhibit 1012
Page 003

Sponsoring Editor: Peter S. Gordon
Production Supervisor: Peggy McMahon
Marketing Manager: Bob Donegan
Manufacturing Supervisor: Roy Logan
Cover Designer: Eileen Hoff
Composition Services: Mike Wile
Technical Art Supervisor: Joseph K. Vetere
Technical Art Consultant: Loretta Bailey
Technical Art Coordinator: Alena B. Konecny

Library of Congress Cataloging-in-Publication Data
Weste, Neil H. E.
Principles of CMOS VLSI design : a systems perspective / Neil Weste, Kamran Eshraghian -- 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-201-53376-6
1. Intergrated circuits--Very large scale integration--design and construction 2. Metal oxide semiconductors, Complementary.
I. Eshraghian, Kamran. . II. Title.
TK7874.W46 1992
621.3'95--dc20 92-16564

Cover Photo: Dick Morton
Cover Art: Neil Weste
Photo Credit: Plates 5, 12, and 13, Melgar Photography, Inc., Santa Clara, CA



Copyright © 1993 by AT&T

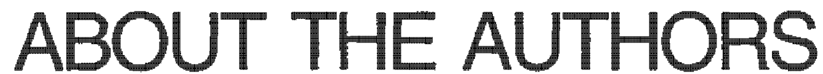
All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher. Printed in the United States of America.

1 2 3 4 5 6 7 8 9 10-MA-96959493

LG Display Co., Ltd.
Exhibit 1012
Page 004

To Avril, Melissa, Tammy and Nicky
and Shohreh, Michelle, Kylie, Natasha and Jason

LG Display Co., Ltd.
Exhibit 1012
Page 005



ABOUT THE AUTHORS

Neil Weste is President of TLW, Inc., a VLSI engineering company in Burlington, Massachusetts. Before cofounding TLW, he was Director of VLSI Systems at Symbolics, Inc., where he led the team that developed the Ivory Lisp microprocessor and the NS design system. Prior to joining Symbolics, Inc., Weste spent six years at AT&T Bell Labs in Holmdel, New Jersey. He worked one year at the Microelectronics Center of North Carolina, with teaching duties at Duke University and the University of North Carolina, Chapel Hill. Weste received his B. Sc., B.E., and Ph.D. from the University of Adelaide, South Australia.

Kamran Eshraghian is an Associate Professor in Electrical Engineering and Director of the Center for Gallium Arsenide VLSI Technology at The University of Adelaide, South Australia. His research interests include very high performance circuits, systems and architectures with applications in digital signal processing. Eshraghian received his B.Tech., B.E., and Ph.D. from the University of Adelaide. Prior to teaching, Eshraghian was with Philips Ltd. as an IC designer.

LG Display Co., Ltd.
Exhibit 1012
Page 006

PREFACE

In the eight years since this book was first published, CMOS technology has steadily moved to occupy a central position in modern electronic system design. Whether digital systems are high speed, high density, low power, or low cost, CMOS technology finds ubiquitous use in the majority of leading-edge commercial applications. CMOS processes have shrunk, and more automated design tools have become commonplace, leading to far more complex chips operating at much higher speeds than a decade ago. While the basic theory of CMOS design remains unchanged, the emphasis and approach to design have changed. With smaller processes and higher speeds comes an increased emphasis on clocking and power distribution, while with complex chip designs and short time-to-market constraints, less emphasis is now placed on die size and the physical details of chip design. The requirement for higher-quality CMOS chips has also increased the need for good approaches to testing.

This edition was updated with these changes in mind. All chapters have undergone extensive revision, and a new chapter on testing replaces one on symbolic layout. Sections on emerging technologies such as BiCMOS, logic synthesis, and parallel scan testing have been added. The overall emphasis has been to include as much as possible of the engineering (and to some extent, the economic) side of CMOS-system design. The artwork has been completely redone and many new figures have been added. All figures were captured on a CMOS VLSI design system. Thus, where possible, diagrams were checked via simulation or net comparison. The tendency has been to include figures where possible ("a picture is worth a thousand words") to trigger the reader's thinking.

LG Display Co., Ltd.
Exhibit 1012
Page 007

As a text, this book provides students with the necessary background to complete CMOS designs and assess which particular design style to use on a given design, from Field Programmable Gate Arrays to full custom design. For the practicing designer, the book provides an extensive source of reference material that covers contemporary CMOS logic, circuit, design, and processing technology.

In common with the first edition, the text is divided into three main sections. The first deals with basic CMOS logic and circuit design and CMOS processing technology. This includes design issues such as speed, power dissipation, and clocking and subsystem design. The second section deals with design approaches and testing. The final section describes three examples of CMOS module/chip designs to provide working examples of the material presented in the first two parts of the book.

In the eighties, designers struggled with tools, circuit techniques, and technology to build CMOS digital systems that could frequently be mastered by one person. The design issues, for example, related to whether a simulation for a circuit could be done and, if so, how accurately. Or perhaps the success of a project depended on a router or a design-rule checker that could deal with large databases. Today, the technology has moved to a point where, to a first order, the technology always works. Failures in design relate to incomplete specifications, inadequate testing, poor communication between designers in a team, or other issues that are somewhat removed from the detailed engineering that still has to take place. That engineering is supported by well-developed design tools. A significant task to be mastered in today's world (once the basics have been learnt) is to take a specification, turn it into a design, enter the design into a CAD system, test it, have it manufactured, and then be able to ship the product.

Increasingly, CMOS VLSI design is being seen as an ideal medium in which to teach the general digital (and analog) system design principles required in such a design process by introducing such issues as structured design and testing. Coupled with education-based Field Programmable Logic Array tools and prototyping kits, courses can be crafted around the basic principles of CMOS design, such as logic design and delay estimation, and coupled with more advanced topics such as simulation, timing analysis, placement and routing, and testing. With reprogrammable hardware, the concept-to-reality delay is reduced to minutes, and the education dynamics of almost-real-time feedback can only help in the education of tomorrow's system designers. The principles used in these laboratory systems are then applicable, with suitable modifications and information, to real-world products, whether such products employ gate-array, standard-cell, or full-custom CMOS design techniques.

Burlington, Mass. N. H. E. W.

LG Display Co., Ltd.
Exhibit 1012
Page 008

Technical Note: The text was revised using Microsoft Word 4/5 on an Apple Mac II (8Mb RAM, 1.2Gb disk) from a scanned OCR'ed version of the first-edition text. The figures were captured by the author using the TLW NS VLSI design software (developed at Symbolics) with custom Lisp code for specialized EPS output and for capturing SPICE simulation results. The NS design system was run under the Genera operating system on the Mac II, using a Symbolics MacIvory 2 board (2.6 Mwords physical memory, 400Mb of paging space), and a Symbolics XL 1200 Lisp machine. All design work (symbolic layout and schematic capture, net comparison, SPICE, timing and switch simulation, compaction, and timing analysis) dealt with in the book was completed on these machines. In fact, an interesting example of "the wheel of reincarnation" applies: the first edition of the book was used in part to create the Ivory Lisp microprocessor, while the processor was used in turn to create the second edition of the text.

ACKNOWLEDGMENTS

Bruce Edwards, Chris Terman, Jud Leonard, and Brian Ogilvie of TLW, Inc. supplied a number of the circuit designs used in this edition. They also provided comments and encouragement during the rewrite. Kurt Keutzer provided material on logic synthesis. Tom Knight, André DeHon, and Thomas Simon provided material on high-performance CMOS pads. Jim Cherry provided material on timing analyzers. Fred Rosenberger helped with metastability and provided corrections to the first edition. Don MacLennan provided feedback on the section on design economics. The author would like to acknowledge the support of Amihai Miron and that of Philips Laboratories for permission to include the ghost canceller system example in Chapter 9. Analog Devices provided the means to fabricate and test the A/D described in Chapter 9. Diane DeCastro was of assistance in the early stages of this rewrite with OCR support for the capture of the previous text. Bryan Ackland and Ismail Eldumiati of AT&T Bell Labs also provided assistance.

A number of reviewers were instrumental in determining the direction of this revision, and their detailed comments on the first drafts were much appreciated. In particular, the author would like to thank Don Trotter for sharing his course notes and his detailed comments on the first draft. Lex A. Akers, Jonathan Allen, Andreas Andreou, James H. Aylor, Prithvraj Banaerjee, Hans van den Biggelaar, Ray Chen, Willhelm Eggimann, Joseph Ku, Ronald K. Lomax, and John Uyemura also provided valuable feedback.

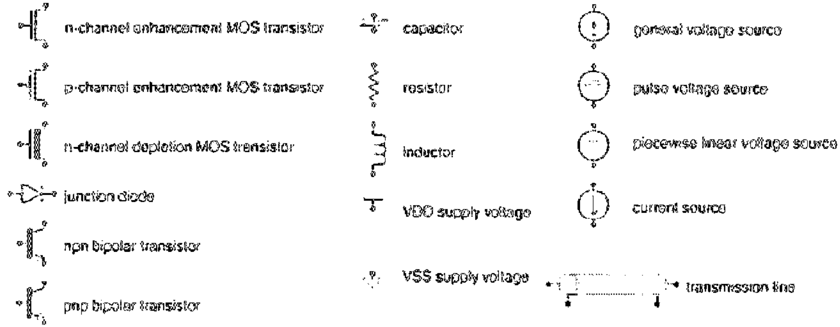
Finally, the author would like to thank his family—Avril, Melissa, Tammy, and Nicky—for their continued support and their work on the book through the long period it took to complete this revision.

Burlington, Mass. N. H. E. W.

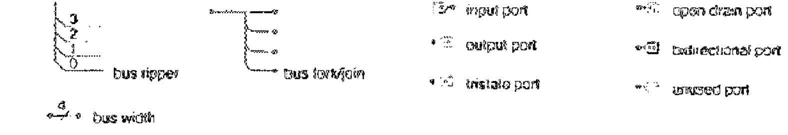
LG Display Co., Ltd.
Exhibit 1012
Page 009

KEY TO SCHEMATICS USED IN THIS BOOK

PRIMITIVES



BUSSES



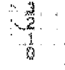
a bus width specifies the width of the bus and the bus ripper or bus fork/join specify which subfields of the bus are extracted from the bus

FOO —4— A,B,C,D

a bus can be named by concatenating names or fields
Here the bus FOO<3:0> is made up of the signals A,B,C and D with FOO<3>=A etc.

a bus ripper can extract arbitrary fields per connection, while a bus fork/join extracts one signal per connection

A —4—



w
x
y
z

A four bit bus with A<0>=z A<1>=y A<2>=x A<3>=w

INST 16

CMD
WR
RA
RB

A 16 bit bus called INST (INST<15:0>) with INST<3:0> = RB<3:0> etc

REPLICATION

replication is indicated by a small x and a number on a schematic icon

DEVICE/GATE SIZES

2/1 an nMOS transistor with Width = 2 and Length = 1
the units are in terms of minimum device width and length
i.e. in a process where Wmin = 2μ and Lmin =0.8μ W=4μ and L=0.8μ

4/2 an invertor with p transistor width = 4*Wmin
and n transistor width = 2*Wmin

LG Display Co., Ltd.
Exhibit 1012
Page 010

CONTENTS

PART 1
INTRODUCTION TO CMOS TECHNOLOGY 1

1
INTRODUCTION TO CMOS CIRCUITS 3

1.1 A Brief History 3
1.2 Book Summary 4
1.3 MOS Transistors 5
1.4 MOS Transistor Switches 7
1.5 CMOS Logic 9
1.5.1 The Inverter 9
1.5.2 Combinational Logic 10
1.5.3 The NAND Gate 11
1.5.4 The NOR Gate 13
1.5.5 Compound Gates 15
1.5.6 Multiplexers 17
1.5.7 Memory—Latches and Registers 19
1.6 Circuit and System Representations 21
1.6.1 Behavioral Representation 22
1.6.2 Structural Representation 24
1.6.3 Physical Representation 28

LG Display Co., Ltd.
Exhibit 1012
Page 011

1.7 An Example 30
1.7.1 Specification 31
1.7.2 Behavioral Description 31
1.7.3 Structural Specification 32
1.7.4 Physical Description 35
1.7.5 Summary 37
1.8 CMOS Scorecard 38
1.9 Summary 39
1.10 References 39

2

MOS TRANSISTOR THEORY 41

2.1 Introduction 41
2.1.1 nMOS Enhancement Transistor 43
2.1.2 pMOS Enhancement Transistor 47
2.1.3 Threshold Voltage 47
2.1.3.1 Threshold Voltage Equations 48
2.1.4 Body Effect 51
2.2 MOS Device Design Equations 51
2.2.1 Basic DC Equations 51
2.2.2 Second Order Effects 53
2.2.2.1 Threshold Voltage—Body Effect 54
2.2.2.2 Subthreshold Region 55
2.2.2.3 Channel-length Modulation 55
2.2.2.4 Mobility Variation 56
2.2.2.5 Fowler-Nordheim Tunneling 57
2.2.2.6 Drain Punchthrough 57
2.2.2.7 Impact Ionization—Hot Electrons 57
2.2.3 MOS Models 58
2.2.4 Small Signal AC Characteristics 59
2.3 The Complementary CMOS Inverter—DC Characteristics 61
2.3.1 β_n/β_p Ratio 68
2.3.2 Noise Margin 69
2.3.3 The CMOS Inverter as an Amplifier 71
2.4 Static Load MOS Inverters 72
2.4.1 The Pseudo-nMOS Inverter 73
2.4.2 Unsaturated Load Inverters 77
2.4.3 Saturated Load Inverters 78
2.4.4 The Cascode Inverter 80
2.4.5 TTL Interface Inverter 80
2.5 The Differential Inverter 81
2.6 The Transmission Gate 86
2.7 The Tristate Inverter 91

LG Display Co., Ltd.
Exhibit 1012
Page 012

2.8 Bipolar Devices 91
2.8.1 Diodes 91
2.8.2 Bipolar Transistors 93
2.8.3 BiCMOS Inverters 96
2.9 Summary 98
2.10 Exercises 98
2.11 Appendix—SPICE Level 3 Model 99
2.12 References 106

3
CMOS PROCESSING TECHNOLOGY 109

3.1 Silicon Semiconductor Technology: An Overview 109
3.1.1 Wafer Processing 110
3.1.2 Oxidation 111
3.1.3 Epitaxy, Deposition, Ion-Implantation, and Diffusion 111
3.1.4 The Silicon Gate Process 113
3.2 Basic CMOS Technology 117
3.2.1 A basic n-well CMOS Process 117
3.2.2 The p-well Process 123
3.2.3 Twin-Tub Processes 124
3.2.4 Silicon On Insulator 125
3.3 CMOS Process Enhancements 130
3.3.1 Interconnect 130
3.3.1.1 Metal Interconnect 130
3.3.1.2 Polysilicon/Refractory Metal Interconnect 132
3.3.1.3 Local Interconnect 133
3.3.2 Circuit Elements 134
3.3.2.1 Resistors 134
3.3.2.2 Capacitors 134
3.3.2.3 Electrically Alterable ROM 136
3.3.2.4 Bipolar Transistors 136
3.3.2.5 Thin-film Transistors 139
3.3.3 3-D CMOS 140
3.3.4 Summary 141
3.4 Layout Design Rules 142
3.4.1 Layer Representations 143
3.4.2 CMOS n-well Rules 144
3.4.3 Design Rule Backgrounder 150
3.4.4 Scribe Line 155
3.4.5 Layer Assignments 155
3.4.6 SOI Rules 156
3.4.7 Design Rules—Summary 156
3.5 Latchup 156

LG Display Co., Ltd.
Exhibit 1012
Page 013

3.5.1 The Physical Origin of Latchup 156
3.5.2 Latchup Triggering 158
3.5.3 Latchup Prevention 160
3.5.4 Internal Latchup Prevention Techniques 161
3.5.5 I/O Latchup Prevention 162
3.6 Technology-related CAD Issues 163
3.6.1 DRC—Spacing and Dimension Checks 164
3.6.2 Circuit Extraction 166
3.7 Summary 167
3.8 Exercises 167
3.9 Appendix—An n-well CMOS Technology Process Flow 168
3.10 References 172

4

CIRCUIT CHARACTERIZATION AND PERFORMANCE ESTIMATION 175

4.1 Introduction 175
4.2 Resistance Estimation 176
4.2.1 Resistance of Nonrectangular Regions 178
4.2.2 Contact and Via Resistance 179
4.3 Capacitance Estimation 180
4.3.1 MOS-Capacitor Characteristics 180
4.3.2 MOS Device Capacitances 183
4.3.3 Diffusion (source/drain) Capacitance 186
4.3.4 SPICE Modeling of MOS Capacitances 188
4.3.5 Routing Capacitance 191
4.3.5.1 Single Wire Capacitance 191
4.3.5.2 Multiple Conductor Capacitances 192
4.3.6 Distributed RC Effects 198
4.3.7 Capacitance Design Guide 202
4.3.8 Wire Length Design Guide 204
4.4 Inductance 205
4.5 Switching Characteristics 207
4.5.1 Analytic Delay Models 208
4.5.1.1 Fall Time 208
4.5.1.2 Rise Time 210
4.5.1.3 Delay Time 211
4.5.2 Empirical Delay Models 213
4.5.3 Gate Delays 214
4.5.4 Further Delay Topics 216
4.5.4.1 Input Waveform Slope 216
4.5.4.2 Input Capacitance 217
4.5.4.3 Switch-Level *RC* Models 218

LG Display Co., Ltd.
Exhibit 1012
Page 014

4.5.4.4 Macromodeling 221
4.5.4.5 Body Effect 223
4.5.5 Summary 225
4.6 CMOS-Gate Transistor Sizing 226
4.6.1 Cascaded Complimentary Inverters 226
4.6.2 Cascaded Pseudo-nMOS Inverters 228
4.6.3 Stage Ratio 229
4.7 Power Dissipation 231
4.7.1 Static Dissipation 231
4.7.2 Dynamic Dissipation 233
4.7.3 Short-Circuit Dissipation 235
4.7.4 Total Power Dissipation 236
4.7.5 Power Economy 237
4.8 Sizing Routing Conductors 238
4.8.1 Power and Ground Bounce 239
4.8.2 Contact Replication 240
4.9 Charge Sharing 240
4.10 Design Margining 243
4.10.1 Temperature 243
4.10.2 Supply Voltage 244
4.10.3 Process Variation 245
4.10.4 Design Corners 246
4.10.5 Packaging Issues 247
4.10.6 Power and Clock Conductor Sizing 248
4.10.7 Summary 248
4.11 Yield 248
4.12 Reliability 250
4.13 Scaling of MOS Transistor Dimensions 250
4.13.1 Scaling Principles 251
4.13.2 Interconnect-Layer Scaling 253
4.13.3 Scaling in Practice 255
4.14 Summary 255
4.15 Exercises 256
4.16 References 257

5

CMOS CIRCUIT AND LOGIC DESIGN 261

5.1 Introduction 261
5.2 CMOS Logic Gate Design 262
5.2.1 Fan-in and Fan-out 264
5.2.2 Typical CMOS NAND and NOR Delays 267
5.2.3 Transistor Sizing 271
5.2.4 Summary 272

LG Display Co., Ltd.
Exhibit 1012
Page 015

5.3 Basic Physical Design of Simple Logic Gates 273
5.3.1 The Inverter 273
5.3.2 NAND and NOR gates 278
5.3.3 Complex Logic Gates Layout 279
5.3.4 CMOS Standard Cell Design 283
5.3.5 Gate Array Layout 285
5.3.6 Sea-of-Gates Layout 286
5.3.7 General CMOS Logic-Gate Layout Guidelines 287
5.3.8 Layout Optimization for Performance 290
5.3.9 Transmission-gate Layout Considerations 291
5.3.10 2-input Multiplexer 294
5.4 CMOS Logic Structures 295
5.4.1 CMOS Complementary Logic 295
5.4.2 BiCMOS Logic 297
5.4.3 Pseudo-nMOS Logic 298
5.4.4 Dynamic CMOS Logic 301
5.4.5 Clocked CMOS Logic (C^2MOS) 302
5.4.6 Pass-transistor Logic 304
5.4.7 CMOS Domino Logic 308
5.4.8 NP Domino Logic (Zipper CMOS) 310
5.4.9 Cascade Voltage Switch Logic (CVSL) 311
5.4.10 SFPL Logic 314
5.4.11 Summary 315
5.5 Clocking Strategies 317
5.5.1 Clocked Systems 317
5.5.2 Latches and Registers 318
5.5.3 System Timing 322
5.5.4 Setup and Hold Time 323
5.5.5 Single-phase Memory Structures 325
5.5.6 Phase Locked Loop Clock Techniques 334
5.5.7 Metastability and Synchronization Failures 337
5.5.8 Single-phase Logic Structures 340
5.5.9 Two-phase Clocking 344
5.5.10 Two-phase Memory Structures 346
5.5.11 Two-phase Logic Structures 350
5.5.12 Four-phase Clocking 351
5.5.13 Four-phase Memory Structures 352
5.5.14 Four-phase Logic Structures 353
5.5.15 Recommended Clocking Approaches 355
5.5.16 Clock Distribution 356
5.6 I/O Structures 357
5.6.1 Overall Organization 357
5.6.2 V_{DD} and V_{SS} Pads 360
5.6.3 Output Pads 360

LG Display Co., Ltd.
Exhibit 1012
Page 016

5.6.4 Input Pads 361
5.6.5 Tristate and Bidirectional Pads 364
5.6.6 Miscellaneous Pads 365
5.6.7 ECL and Low Voltage Swing Pads 367
5.7 Low-power Design 368
5.8 Summary 370
5.9 Exercises 370
5.10 References 372

PART 2
SYSTEMS DESIGN AND DESIGN METHODS 379

6
CMOS DESIGN METHODS 381
6.1 Introduction 381
6.2 Design Strategies 382
6.2.1 Introduction 382
6.2.2 Structured Design Strategies 383
6.2.3 Hierarchy 384
6.2.4 Regularity 387
6.2.5 Modularity 387
6.2.6 Locality 389
6.2.7 Summary 391
6.3 CMOS Chip Design Options 391
6.3.1 Programmable Logic 391
6.3.2 Programmable Logic Structures 392
6.3.3 Programmable Interconnect 395
6.3.4 Reprogrammable Gate Arrays 400
6.3.4.1 The XILINX Programmable Gate Array 400
6.3.4.2 Algotronix 403
6.3.4.3 Concurrent Logic 406
6.3.5 Sea-of-Gate and Gate Array Design 407
6.3.6 Standard-cell Design 413
6.3.6.1 A Typical Standard-cell Library 414
6.3.7 Full-custom Mask Design 417
6.3.8 Symbolic Layout 417
6.3.8.1 Coarse-grid Symbolic Layout 417
6.3.8.2 Gate-matrix Layout 418
6.3.8.3 Sticks Layout and Compaction 420
6.3.8.4 Virtual-grid Symbolic Layout 421
6.3.9 Process Migration—Retargetting Designs 423

LG Display Co., Ltd.
Exhibit 1012
Page 017

6.4 Design Methods 424
6.4.1 Behavioral Synthesis 424
6.4.2 RTL Synthesis 425
6.4.3 Logic Optimization 427
6.4.4 Structural-to-Layout Synthesis 431
6.4.4.1 Placement 431
6.4.4.2 Routing 431
6.4.4.3 An Automatic Placement Example 432
6.4.5 Layout Synthesis 434
6.5 Design-capture Tools 437
6.5.1 HDL Design 437
6.5.2 Schematic Design 438
6.5.3 Layout Design 438
6.5.4 Floorplanning 438
6.5.5 Chip Composition 439
6.6 Design Verification Tools 440
6.6.1 Simulation 441
6.6.1.1 Circuit-level Simulation 441
6.6.1.2 Timing Simulation 442
6.6.1.3 Logic-level Simulation 443
6.6.1.4 Switch-level Simulation 444
6.6.1.5 Mixed-mode Simulators 444
6.6.1.6 Summary 445
6.6.2 Timing Verifiers 445
6.6.3 Network Isomorphism 446
6.6.4 Netlist comparison 447
6.6.5 Layout Extraction 448
6.6.6 Back-Annotation 448
6.6.7 Design-rule Verification 448
6.6.8 Pattern Generation 448
6.7 Design Economics 449
6.7.1 Nonrecurring Engineering Costs (NREs) 450
6.7.1.1 Engineering Costs 450
6.7.1.2 Prototype Manufacturing Costs 451
6.7.2 Recurring Costs 452
6.7.3 Fixed Costs 452
6.7.4 Schedule 453
6.7.5 Personpower 454
6.7.6 An Example—Gate-array Productivity 454
6.8 Data Sheets 456
6.8.1 The Summary 456
6.8.2 Pinout 456
6.8.3 Description of Operation 457
6.8.4 DC Specifications 457

LG Display Co., Ltd.
Exhibit 1012
Page 018

6.8.5 AC Specifications 457
6.8.6 Package Diagram 458
6.9 Summary 458
6.10 Exercises 458
6.11 References 459

7
CMOS TESTING 465
7.1 The Need for Testing 465
7.1.1 Functionality Tests 466
7.1.2 Manufacturing Tests 468
7.1.3 A Walk Through the Test Process 456
7.2 Manufacturing Test Principles 471
7.2.1 Fault models 472
7.2.1.1 Stuck-At Faults 472
7.2.1.2 Short-Circuit and Open-Circuit Faults 473
7.2.2 Observability 474
7.2.3 Controllability 475
7.2.4 Fault Coverage 475
7.2.5 Automatic Test Pattern Generation (ATPG) 476
7.2.6 Fault Grading and Fault Simulation 481
7.2.7 Delay Fault Testing 482
7.2.8 Statistical Fault Analysis 483
7.2.9 Fault Sampling 484
7.3 Design Strategies for Test 485
7.3.1 Design for Testability 485
7.3.2 Ad-Hoc Testing 485
7.3.3 Scan-Based Test Techniques 489
7.3.3.1 Level Sensitive Scan Design (LSSD) 489
7.3.3.1 Serial Scan 490
7.3.3.2 Partial Serial Scan 493
7.3.3.3 Parallel Scan 493
7.3.4 Self-Test Techniques 495
7.3.4.1 Signature Analysis and BILBO 495
7.3.4.2 Memory Self-Test 497
7.3.4.3 Iterative Logic Array Testing 498
7.3.5 IDDQ Testing 498
7.4 Chip-Level Test Techniques 498
7.4.1 Regular Logic Arrays 499
7.4.2 Memories 500
7.4.3 Random logic 500
7.5 System-Level Test Techniques 500
7.5.1 Boundary Scan 500

LG Display Co., Ltd.
Exhibit 1012
Page 019

7.5.1.1 Introduction 500
7.5.1.2 The Test Access Port (TAP) 501
7.5.1.3 The Test Architecture 502
7.5.1.4 The Tap Controller 502
7.5.1.5 The Instruction Register (IR) 503
7.5.1.6 Test-Data Registers (DRs) 504
7.5.1.7 Boundary Scan Registers 504
7.5.3 Summary 506
7.6 Layout Design for Improved Testability 506
7.7 Summary 508
7.8 Exercises 508
7.9 References 508

8

CMOS SUBSYSTEM DESIGN 513

8.1 Introduction 513
8.2 Datpath Operations 513
8.2.1 Addition/Subtraction 515
8.2.1.1 Single-Bit Adders 515
8.2.1.2 Bit-Parallel Adder 517
8.2.1.3 Bit Serial Adders, Carry-save Addition, and Pipelining 520
8.2.1.4 Transmission-Gate Adder 524
8.2.1.5 Carry-Lookahead Adders 526
8.2.1.6 Carry-Select Adder 532
8.2.1.7 Conditional-Sum Adder 532
8.2.1.8 Very Wide Adders 534
8.2.1.9 Summary 536
8.2.2 Parity Generators 537
8.2.3 Comparators 537
8.2.4 Zero/One Detectors 537
8.2.5 Binary Counters 539
8.2.5.1 Asynchronous Counters 539
8.2.5.2 Synchronous Counters 539
8.2.6 Boolean Operations—ALUs 541
8.2.7 Multiplication 542
8.2.7.1 Array Multiplication 545
8.2.7.2 Radix-n Multiplication 547
8.2.7.3 Wallace Tree Multiplication 554
8.2.7.4 Serial Multiplication 557
8.2.8 Shifters 560
8.3 Memory Elements 563
8.3.1 Read/Write Memory 564

LG Display Co., Ltd.
Exhibit 1012
Page 020

8.3.1.1 RAM 564
8.3.1.2 Register Files 580
8.3.1.3 FIFOs, LIFOs, SIPOs 582
8.3.1.4 Serial-Access Memory 583
8.3.2 Read Only Memory 585
8.3.3 Content-Addressable Memory 589
8.4 Control 590
8.4.1 Finite-State Machines 591
8.4.1.1 FSM Design Procedure 591
8.4.2 Control Logic Implementation 595
8.4.2.1 PLA Control Implementation 595
8.4.2.2 ROM Control Implementation 602
8.4.2.3 Multilevel Logic 604
8.4.2.4 An Example of Control-Logic Implementation 604
8.5 Summary 620
8.6 Exercises 621
8.7 References 622

PART 3
CMOS SYSTEM CASE STUDIES 625

9
CMOS SYSTEM DESIGN EXAMPLES 627
9.1 Introduction 627
9.2 A Core RISC Microcontroller 628
9.2.1 Instruction Set 629
9.2.1.1 Address Architecture 629
9.2.1.2 ALU Class Instructions 631
9.2.1.3 Control Transfer Instructions 633
9.2.2 Pipeline Architecture 634
9.2.2.1 Bypassing, Result Forwarding, or Pass-around 637
9.2.2.2 Conditional Branching 638
9.2.2.3 Subroutine Call and Return 639
9.2.2.4 I/O Architecture 639
9.2.3 Major Logic Blocks 640
9.2.3.1 ALU_DP 640
9.2.3.2 Register File 651
9.2.3.3 PC Datapath (PC_DP) 654
9.2.3.4 Instruction Memory 656

LG Display Co., Ltd.
Exhibit 1012
Page 021

9.2.3.5 Instruction Pipe 656
9.2.3.6 Control Logic 658
9.2.4 Layout 663
9.2.4.1 Datapath Floorplans 666
9.2.5 Functional Verification and Testing 669
9.3 A TV Echo Canceller 672
9.3.1 Ghost Cancellation 672
9.3.2 FIR and IIR filters 674
9.3.3 System Architecture 676
9.3.4 Chip Architecture 677
9.3.4.1 Filter Considerations 677
9.3.4.2 Chip Overview 678
9.3.5 Submodules 680
9.3.5.1 Filter Taps 680
9.3.5.2 Delay Lines 685
9.3.5.3 Phase-locked Loop- and Clock-generation 685
9.3.5.4 Peripheral Processing 689
9.3.6 Power Distribution 689
9.3.7 Chip Floorplan 690
9.3.8 Testing and Verification 692
9.3.9 Summary 694
9.4 A 6-bit Flash A/D 694
9.4.1 Introduction 694
9.4.2 Basic Architecture 695
9.4.3 Resistor String 696
9.4.4 The Comparator 696
9.4.5 Thermometer Code Logic 698
9.4.6 Floorplan and Layout 698
9.4.7 Summary 701
9.5 Summary 701
9.6 Exercises 701
9.7 References 702

INDEX 703

LG Display Co., Ltd.
Exhibit 1012
Page 022

MOS TRANSISTOR THEORY

2

2.1 Introduction

In Chapter 1 the MOS transistor was introduced in terms of its operation as an ideal switch. In this chapter we will examine the characteristics of MOS transistors in more detail to lay the foundation for predicting the performance of the switches, which is less than ideal. Figure 2.1 shows some of the symbols that are commonly used for MOS transistors. The symbols in Fig. 2.1(a) will be used where it is necessary only to indicate the switch logic required to build a function. If the substrate connection needs to be shown, the symbols in Fig. 2.1(b) will be used. Figure 2.1(c) shows an example of the many symbols that may be encountered in the literature.

This chapter will concentrate on the static or DC operation of MOS transistors. This is the first design goal that must be satisfied to ensure that logic gates operate as logic gates. All circuits are analog in nature and the digital abstraction only remains an abstraction as long as certain design goals are met. Design for timing constraints is covered in Chapter 4.

An MOS transistor is termed a majority-carrier device, in which the current in a conducting channel between the source and the drain is modulated by a voltage applied to the gate. In an n-type MOS transistor (i.e., nMOS), the majority characters are electrons. A positive voltage applied on the gate with respect to the substrate enhances the number of electrons in the channel

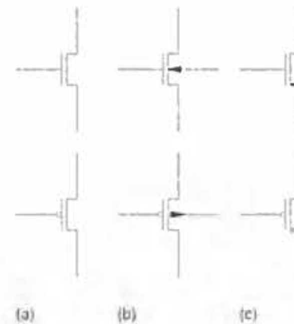


FIGURE 2.1 MOS transistor symbols

LG Display Co., Ltd.
Exhibit 1012
Page 023

(the region immediately under the gate) and hence increases the conductivity of the channel. For gate voltages less than a threshold value denoted by V_t, the channel is cut off, thus causing a very low drain-to-source current. The operation of a p-type transistor (i.e., pMOS) is analogous to the nMOS transistor, with the exception that the majority carriers are holes and the voltages are negative with respect to the substrate.

The first parameter of interest that characterizes the switching behavior of an MOS device is the threshold voltage, V_t. This is defined as the voltage at which an MOS device begins to conduct ("turn on"). We can graph the relative conduction against the difference in gate-to-source voltage in terms of the source-to-drain current (I_{ds}) and the gate-to-source voltage (V_{gs}). These graphs for a fixed drain-source voltage, V_{ds}, are shown in Fig. 2.2. It is possible to make n-devices that conduct when the gate voltage is equal to the source voltage, while others require a positive difference between gate and source voltages to bring about conduction (negative for p-devices). Those devices that are normally cut off (i.e., nonconducting) with zero gate bias (gate voltage–source voltage) are further classed as enhancement-mode devices, whereas those devices that conduct with zero gate bias are called depletion-mode devices. The n-channel transistors and p-channel transistors are the duals of each other; that is, the voltage polarities required for correct operation are the opposite. The threshold voltages for n-channel and p-channel devices are denoted by V_{tn} and V_{tp}, respectively.

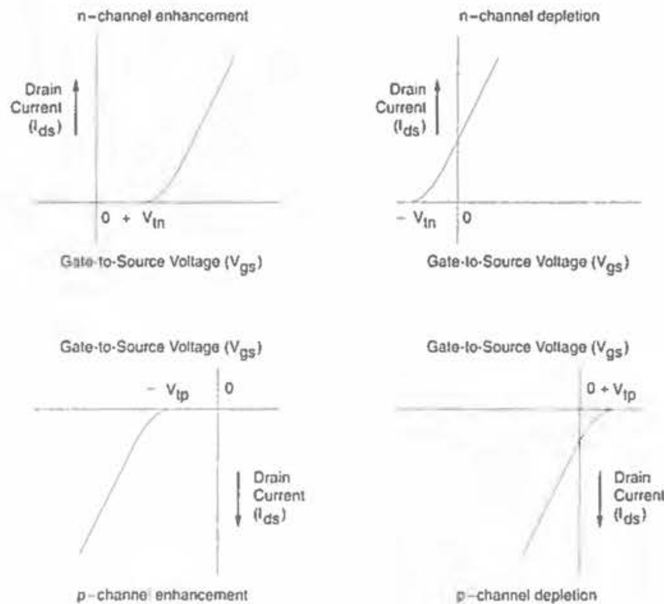


FIGURE 2.2 Conduction characteristics for enhancement and depletion mode MOS transistors (assuming fixed V_{ds})

LG Display Co., Ltd.
Exhibit 1012
Page 024

In CMOS technologies both n-channel and p-channel transistors are fabricated on the same chip. Furthermore, most CMOS integrated circuits, at present, use transistors of the enhancement type.

2.1.1 nMOS Enhancement Transistor

The structure for an n-channel enhancement-type transistor, shown in Fig. 2.3, consists of a moderately doped p-type silicon substrate into which two heavily doped n^+ regions, the *source* and the *drain*, are diffused. Between these two regions there is a narrow region of p-type substrate called the *channel*, which is covered by a thin insulating layer of silicon dioxide (SiO_2) called *gate oxide*. Over this oxide layer is a polycrystalline silicon (polysilicon) electrode, referred to as the *gate*. Polycrystalline silicon is silicon that is not composed of a single crystal. Since the oxide layer is an insulator, the DC current from the gate to channel is essentially zero. Because of the inherent symmetry of the structure, there is no physical distinction between the drain and source regions. Since SiO_2 has relatively low loss and high dielectric strength, the application of high gate fields is feasible.

In operation, a positive voltage is applied between the source and the drain (V_{ds}). With zero gate bias ($V_{gs} = 0$), no current flows from source to drain because they are effectively insulated from each other by the two reversed biased *pn* junctions shown in Fig. 2.3 (indicated by the diode symbols). However, a voltage applied to the gate, which is positive with respect to the source and the substrate, produces an electric field E across the substrate, which attracts electrons toward the gate and repels holes. If the gate voltage is sufficiently large, the region under the gate changes from p-type to n-type (due to accumulation of attracted electrons) and provides a conduction path between the source and the drain. Under such a condition, the surface of the underlying p-type silicon is said to be *inverted*. The term *n-channel* is applied to the structure. This concept is further illustrated by Fig. 2.4(a), which shows the initial distribution of mobile positive holes in a p-type silicon substrate of an MOS structure for a voltage, V_{gs}, much less than a voltage, V_t, which is

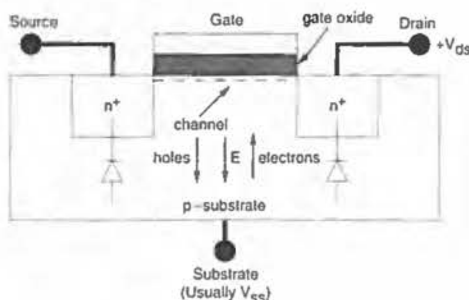


FIGURE 2.3 Physical structure of an nMOS transistor

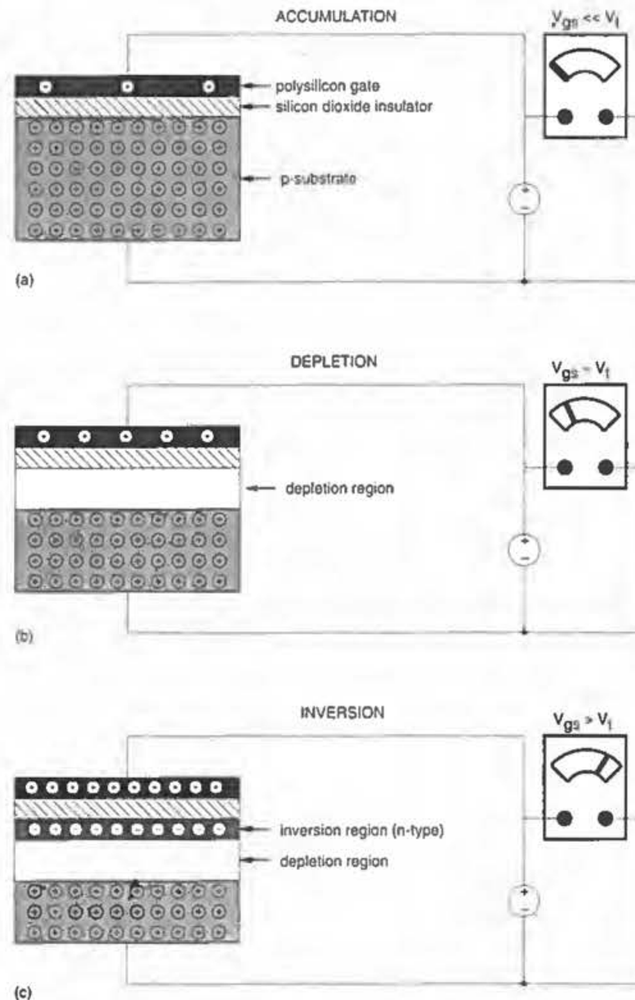


FIGURE 2.4 Accumulation, Depletion and Inversion modes in an MOS structure.

the threshold voltage. This is termed the *accumulation* mode. As V_{gs} is raised above V_t in potential, the holes are repelled causing a depletion region under the gate. Now the structure is in the *depletion* mode (Fig. 2.4b). Raising V_{gs} further above V_t results in electrons being attracted to the region of the substrate under the gate. A conductive layer of electrons in the p substrate gives rise to the name *inversion* mode (Fig. 2.4c).

The difference between a *pn* junction that exists in a bipolar transistor or diode (or between the source or drain and substrate) and the inversion layer

LG Display Co., Ltd.
Exhibit 1012
Page 026

substrate junction is that in the *pn* junction, the n-type conductivity is brought about by a metallurgical process; that is, the electrons are introduced into the semiconductor by the introduction of donor ions. In an inversion layer substrate junction, the n-type layer is induced by the electric field E applied to the gate. Thus, this junction, instead of being a metallurgical junction, is a *field-induced* junction.

Electrically, an MOS device therefore acts as a voltage-controlled switch that conducts initially when the gate-to-source voltage, V_{gs}, is equal to the threshold voltage, V_t. When a voltage V_{ds} is applied between source and drain, with $V_{gs} = V_t$, the horizontal and vertical components of the electrical field due to the source-drain voltage and gate-to-substrate voltage interact, causing conduction to occur along the channel. The horizontal component of the electric field associated with the drain-to-source voltage (i.e., $V_{ds} > 0$) is responsible for sweeping the electrons in the channel from the source toward the drain. As the voltage from drain to source is increased, the resistive drop along the channel begins to change the shape of the channel characteristic. This behavior is shown in Fig. 2.5. At the source end of the channel, the full gate voltage is effective in inverting the channel. However, at the drain end of the channel, only the difference between the gate and drain voltages is effective. When the effective gate voltage ($V_{gs} - V_t$) is greater than the drain voltage, the channel becomes deeper as V_{gs} is increased. This is termed the "linear," "resistive," "nonsaturated," or "unsaturated" region, where the channel current I_{ds} is a function of both gate and drain voltages. If $V_{ds} > V_{gs} - V_t$, then $V_{gd} < V_t$ (V_{gd} is the gate to drain voltage), and the channel becomes pinched off— the channel no longer reaches the drain. This is illustrated in Fig. 2.5(c). However, in this case, conduction is brought about by a drift mechanism of electrons under the influence of the positive drain voltage. As the electrons leave the channel, they are injected into the drain depletion region and are subsequently accelerated toward the drain. The voltage across the pinched-off channel tends to remain fixed at ($V_{gs} - V_t$). This condition is the "saturated" state in which the channel current is controlled by the gate voltage and is almost independent of the drain voltage. For fixed drain-to-source voltage and fixed gate voltage, the factors that influence the level of drain current, I_{ds}, flowing between source and drain (for a given substrate resistivity) are:

- the distance between source and drain
- the channel width
- the threshold voltage V_t
- the thickness of the gate-insulating oxide layer
- the dielectric constant of the gate insulator
- the carrier (electron or hole) mobility, μ.

LG Display Co., Ltd.
Exhibit 1012
Page 027

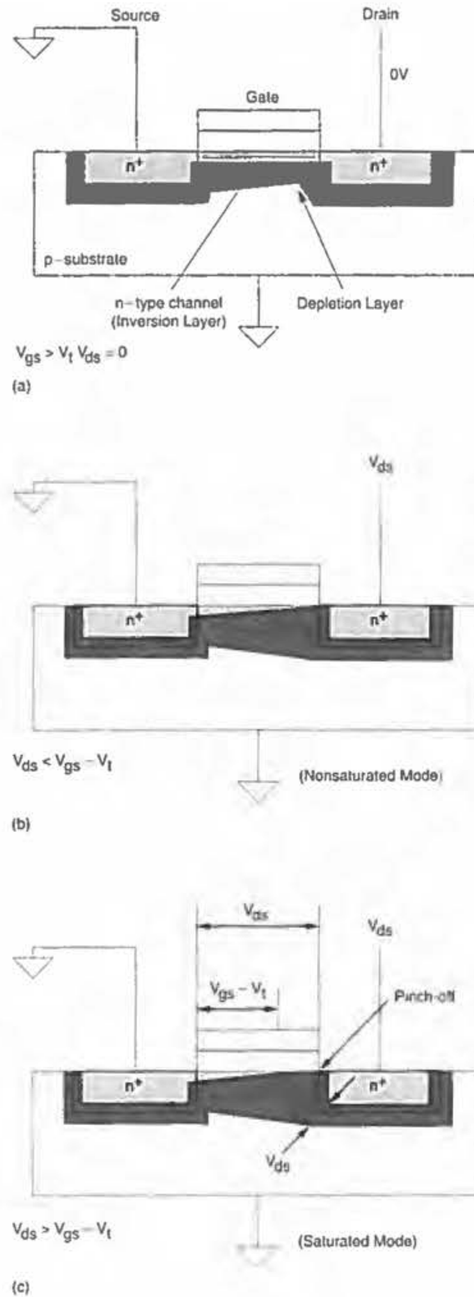


FIGURE 2.5 nMOS device behavior under the influence of different terminal voltages

LG Display Co., Ltd.
Exhibit 1012
Page 028

The normal conduction characteristics of an MOS transistor can be categorized as follows:

- "Cut-off" region: where the current flow is essentially zero (accumulation region).
- "Nonsaturated" region: weak inversion region where the drain current is dependent on the gate and the drain voltage (with respect to the substrate).
- "Saturated" region: channel is strongly inverted and the drain current flow is ideally independent of the drain-source voltage (strong inversion region).

An abnormal conduction condition called avalanche breakdown or punch-through can occur if very high voltages are applied to the drain. Under these circumstances, the gate has no control over the drain current.

2.1.2 pMOS Enhancement Transistor

So far, our discussions have been primarily directed toward nMOS; however, a reversal of n-type and p-type regions yields a p-channel MOS transistor. This is illustrated by Fig. 2.6. Application of a negative gate voltage (w.r.t. source) draws holes into the region below the gate, resulting in the channel changing from n-type to p-type. Thus, similar to nMOS, a conduction path is created between the source and the drain. In this instance, however, conduction results from the movement of holes (versus electrons) in the channel. A negative drain voltage sweeps holes from the source through the channel to the drain.

2.1.3 Threshold Voltage

The threshold voltage, V_t, for an MOS transistor can be defined as the voltage applied between the gate and the source of an MOS device below which the drain-to-source current I_{ds} effectively drops to zero. The word "effec-

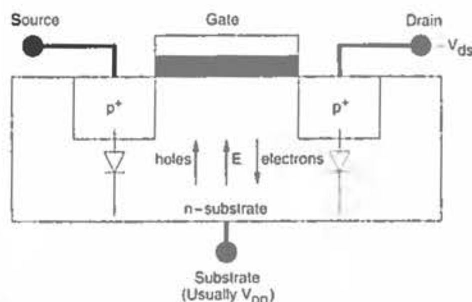


FIGURE 2.6 Physical structure of a pMOS transistor

tively" is used because the drain current never really is zero but drops to a very small value that may be deemed insignificant for the current application (i.e., fast digital CMOS circuits). In general, the threshold voltage is a function of a number of parameters including the following:

- Gate conductor material.
- Gate insulation material.
- Gate insulator thickness–channel doping.
- Impurities at the silicon-insulator interface.
- Voltage between the source and the substrate, V_{sb}.

In addition, the absolute value of the threshold voltage decreases with an increase in temperature. This variation is approximately –4 mV/°C for high substrate doping levels, and –2 mV/°C for low doping levels.[1]

2.1.3.1 Threshold Voltage Equations

Threshold voltage, V_t, may be expressed as

$$V_t = V_{t-mos} + V_{fb} \tag{2.1}$$

where V_{t-mos} is the ideal threshold voltage of an ideal MOS capacitor and V_{fb} is what is termed the flat-band voltage. V_{t-mos} is the threshold where there is no work function difference between the gate and substrate materials.

The MOS threshold voltage, V_{t-mos}, is calculated by considering the MOS capacitor structure that forms the gate of the MOS transistor (see for example[2] or[3]). The ideal threshold voltage may be expressed as

$$V_{t-mos} = 2\phi_b + \frac{Q_b}{C_{ox}} \tag{2.2}$$

where $\phi_b = \frac{kT}{q} ln\left(\frac{N_A}{N_i}\right)$, C_{ox} is the oxide capacitance

and $Q_b = \sqrt{2\varepsilon_{Si} q N_A 2\phi_b}$ which is called the bulk charge term.

The symbol ϕ_b is the bulk potential, a term that accounts for the doping of the substrate. It represents the difference between the Fermi energy level of the doped semiconductor and the Fermi energy level of the intrinsic semiconductor. The intrinsic level is midway between the valence-band edge and the

LG Display Co., Ltd.
Exhibit 1012
Page 030

conduction-band edge of the semiconductor. In a p-type semiconductor the Fermi level is closer to the valence band, while in an n-type semiconductor it is closer to the conduction band. N_A is the density of carriers in the doped semiconductor substrate, and N_i is the carrier concentration in intrinsic (undoped) silicon. N_i is equal to $1.45 \times 10^{10}\ cm^{-3}$ at 300°K. The lowercase k is Boltzmann's constant (1.380×10^{-23} J/°K). T is the temperature (°K) and q is the electronic charge (1.602×10^{-19} Coulomb). The expression kT/q equals .02586 Volts at 300°K. The term ε_{Si} is the permittivity of silicon (1.06×10^{-12} Farads/cm). The term C_{ox} is the gate-oxide capacitance, which is inversely proportional to the gate-oxide thickness (t_{ox}). The threshold voltage, $V_{t\text{-}mos}$, is positive for n-transistors and negative for p-transistors.

The flatband voltage, V_{fb}, is given by

$$V_{fb} = \phi_{ms} - \frac{Q_{fc}}{C_{ox}} \tag{2.3}$$

The term V_{fb} is the flat-band voltage. The term Q_{fc} represents the fixed charge due to surface states that arise due to imperfections in the silicon-oxide interface and doping. The term ϕ_{ms} is the work function difference between the gate material and the silicon substrate ($\phi_{gate} - \phi_{Si}$), which may be calculated for an n^+ gate over a p substrate (the normal way for an n transistor) as follows:[4]

$$\phi_{ms} = -(\frac{Eg}{2} + \phi_b) \approx -0.9V \ \ (N_A = 1\times10^{16}\ cm^{-3}) \tag{2.4a}$$

where

$$E_g = \text{is the band gap energy of silicon} \left(1.16 - .704 \times 10^{-3} \frac{T^2}{T + 1108}\right)^5$$

and T is the temperature (°K). For an n^+ poly gate on an n-substrate (a normal p-transistor)

$$\phi_{ms} = -(\frac{Eg}{2} - \phi_b) = -0.2V \ \ (N_A = 1\times10^{16}\ cm^{-3}) \tag{2.4b}$$

From these equations it may be seen that for a given gate and substrate material the threshold voltage may be varied by changing the doping concentration of the substrate (N_A), the oxide capacitance (C_{ox}), or the surface state charge (Q_{fc}). In addition, the temperature variation mentioned above may be seen.

It is often necessary to adjust the native (original) threshold voltage of an MOS device. Two common techniques used for the adjustment of the threshold voltage entail varying the doping concentration at the silicon-

LG Display Co., Ltd.
Exhibit 1012
Page 031

insulator interface through ion implantation (i.e., affecting Q_{fc}) or using different insulating material for the gate (i.e., affecting C_{ox}). The former approach introduces a small doped region at the oxide/substrate interface that adjusts the flat-band voltage by varying the Q_{fc} term in Eq. (2.3). In the latter approach for instance, a layer of silicon nitride (Si_3N_4) (relative permittivity of 7.5) is combined with a layer of silicon dioxide (relative permittivity of 3.9), resulting in an effective relative permittivity of about 6, which is substantially larger than the dielectric constant of SiO_2. Consequently, for the same thickness as an insulating layer consisting of only silicon dioxide, the dual dielectric process will be electrically equivalent to a thinner layer of SiO_2, leading to a higher C_{ox} value.

In order to prevent the surface of the silicon from inverting in the regions between transistors, the threshold voltage in these field regions is increased by heavily doped diffusions, by implants of the silicon surface, or by making the oxide layer very thick. MOS transistors are self-isolating as long as the surface of the silicon can be inverted under the gate, but not in the regions between devices by normal circuit voltages.

Example

1. Calculate the native threshold voltage for an n-transistor at 300°K for a process with a Si substrate with $N_A = 1.80 \times 10^{-16}$, a SiO_2 gate oxide with thickness 200 Å. (Assume $\phi_{ms} = -0.9V$, $Q_{fc} = 0$.)

$$\phi_b = .02586\ ln\left(\frac{1.8\times10^{16}}{1.45\times10^{10}}\right)$$

$$= .36 \text{ volts}$$

with
$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}$$

$$= \frac{3.9\times8.85\times10^{-14}}{0.2\times10^{-5}}$$

$$= 1.726\times10^{-7}\ Farads/cm^2$$

$$V_t = \phi_{ms} + \frac{\sqrt{2\varepsilon_{Si}qN_A 2\phi_b}}{C_{ox}} + 2\phi_b$$

$$= -0.9 + .384 + .72$$

$$= 0.16 \text{ volts}$$

LG Display Co., Ltd.
Exhibit 1012
Page 032

2.1.4 Body Effect

As we have seen so far, all devices comprising an MOS device are made on a common substrate. As a result, the substrate voltage of all devices is normally equal. (In some analog circuits this may not be true.) However, in arranging the devices to form gating functions it might be necessary to connect several devices in series as shown in Fig. 2.7 (for example, the NAND gate shown in Fig. 1.6). This may result in an increase in source-to-substrate voltage as we proceed vertically along the series chain ($V_{sb1} = 0$, $V_{sb2} \neq 0$).

Under normal conditions—that is, when $V_{gs} > V_t$—the depletion-layer width remains constant and charge carriers are pulled into the channel from the source. However, as the substrate bias V_{sb} ($V_{source} - V_{substrate}$) is increased, the width of the channel-substrate depletion layer also increases, resulting in an increase in the density of the trapped carriers in the depletion layer. For charge neutrality to hold, the channel charge must decrease. The resultant effect is that the substrate voltage, V_{sb}, adds to the channel-substrate junction potential. This increases the gate-channel voltage drop. The overall effect is an increase in the threshold voltage, V_t ($V_{t2} > V_{t1}$).

FIGURE 2.7 The effect of substrate bias on series-connected n-transistors

2.2 MOS Device Design Equations

2.2.1 Basic DC Equations

As stated previously, MOS transistors have three regions of operation:

- Cutoff or subthreshold region.
- Nonsaturation or linear region.
- Saturation region.

The ideal (first order, Shockley) equations[6,7,8] describing the behavior of an nMOS device in the three regions are:

The cutoff region:

$$I_{ds} = 0 \qquad V_{gs} \leq V_t \tag{2.5a}$$

The nonsaturation, linear, or triode region:

$$I_{ds} = \beta\left[(V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2}\right] \qquad 0 < V_{ds} < V_{gs} - V_t \tag{2.5b}$$

[Although this region is commonly called the linear region, I_{ds} varies linearly with V_{gs} and V_{ds} when the quadratic term $V_{ds}^2/2$ is very small (i.e., $V_{ds} << V_{gs} - V_t$).]

LG Display Co., Ltd.
Exhibit 1012
Page 033

The saturation region:

$$I_{ds} = \beta \frac{(V_{gs} - V_t)^2}{2} \qquad 0 < V_{gs} - V_t < V_{ds} \tag{2.5c}$$

where I_{ds} is the drain-to-source current, V_{gs} is the gate-to-source voltage, V_t is the device threshold, and β is the MOS transistor gain factor. The last factor is dependent on both the process parameters and the device geometry, and is given by

$$\beta = \frac{\mu\varepsilon}{t_{ox}}\left(\frac{W}{L}\right) \tag{2.6}$$

where μ is the effective surface mobility of the carriers in the channel, ε is the permittivity of the gate insulator, t_{ox} is the thickness of the gate insulator, W is the width of the channel, and L is the length of the channel. The gain factor β thus consists of a process dependent factor $\mu\varepsilon / t_{ox}$, which contains all the process terms that account for such factors as doping density and gate-oxide thickness and a geometry dependent term (W / L), which depends on the actual layout dimensions of the device. The process dependent factor is sometimes written as μC_{ox}, where $C_{ox} = \varepsilon / t_{ox}$ is the gate oxide capacitance. The geometric terms in Eq. (2.6) are illustrated in Fig. 2.8 in relation to the physical MOS structure.

The voltage-current characteristics of the n- and p-transistors in the nonsaturated and saturated regions are represented in Fig. 2.9 (with the SPICE circuit for obtaining these characteristics for an n-transistor). Note that we use the absolute value of the voltages concerned to plot the characteristics of the p- and n-transistors on the same axes. The boundary between the linear and saturation regions corresponds to the condition $|V_{ds}| = |V_{gs} - V_t|$ and appears as a dashed line in Fig. 2.9. The drain voltage at which the device

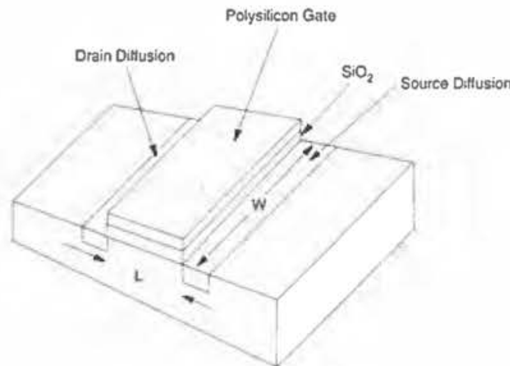


FIGURE 2.8 Geometric terms in the MOS device equation

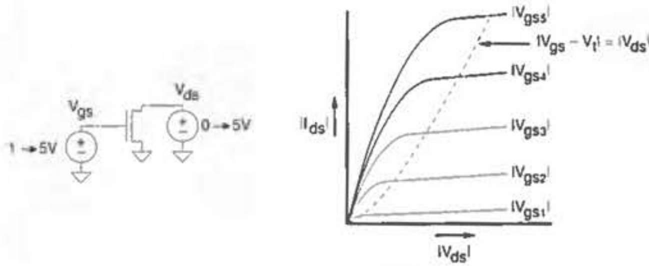


FIGURE 2.9 *VI* characteristics for n- and p-transistors

becomes saturated is called V_{dsat}, or the drain saturation voltage. In the above equations that is equal to $V_{gs} - V_t$.

Example

Typical values (for an n-device) for current (~1μ) processes are as follows:

$\mu_n = 500\ cm^2 / V\text{-}sec$
$\varepsilon = 3.9\varepsilon_0 = 3.9 \times 8.85 \times 10^{-14}\ F/cm$ (permittivity of silicon dioxide, S_iO_2)
$t_{ox} = 200\ \text{Å}$

Hence a typical n-device β would be

$$\frac{500 \times 3.9 \times 8.85 \times 10^{-14}}{.2 \times 10^{-5}} \frac{W}{L} = 88.5 \frac{W}{L} \mu A / V^2$$

On the other hand, p-devices have hole mobilities (μ_p) of about 180 $cm^2/V\text{-}sec$, yielding a β of

$$= 31.9 \frac{W}{L} \mu A / V^2$$

Thus the ratio of n-to-p gain factors in this example is about 2.8. This ratio varies from about 2 to 3 depending on the process.

2.2.2 Second Order Effects

Eq. (2.5) represents the simplest view of the MOS transistor DC voltage current equations. There have been many research papers published on more detailed and accurate models that have been created to fill a variety of requirements, such as accuracy, computational efficiency, and the conservation of charge. The circuit simulation program SPICE[9] and its commercial and proprietary derivations generally use a parameter called LEVEL to spec-

ify which model equations are used. LEVEL 1 models build on those defined in Eq. (2.5) and include some important second-order effects. LEVEL 2 models calculate the currents based on device physics. LEVEL 3 is a semiempirical approach that relies on parameters selected on the basis of matching the equations to real circuits. The MOS device equations in terms of the LEVEL 1 parameters used in SPICE will be covered here; Section 2.10, in this chapter, describes the LEVEL 3 parameters used in the commercially available HSPICE program.

First the term $\mu\varepsilon/t_{ox}$ (μC_{ox}) is defined as the *process gain factor*. In SPICE this is referred to as *KP*. Depending on the vintage of the process and the type of transistor, *KP* may vary from 10–100 $\mu A/V^2$. In addition, it is not unusual to expect a variation of 10%–20% in *KP* within a given process as a result of variations in starting materials and variation in SiO_2 growth.

2.2.2.1 Threshold Voltage–Body Effect

The threshold voltage V_t is not constant with respect to the voltage difference between the substrate and the source of the MOS transistor. This is known as the *substrate-bias effect* or *body effect*. The expression for the threshold voltage may be modified to incorporate V_{sb}, the difference between the source and the substrate.

$$V_t = V_{fb} + 2\phi_b + \frac{\sqrt{2\varepsilon_{Si} q N_A (2\phi_b + |V_{sb}|)}}{C_{ox}}$$

$$V_t = V_{t0} + \gamma\left[\sqrt{(2\phi_b + |V_{sb}|)} + 2\phi_b\right] \tag{2.7}$$

where V_{sb} is the substrate bias, V_{t0} is the threshold voltage for $V_{sb} = 0$ (Eq. 2.1), and γ is the constant that describes the substrate bias effect. The term ϕ_b is defined in Eq. 2.2.

Typical values for γ lie in the range of 0.4 to 1.2. It may be expressed as

$$\gamma = \frac{t_{ox}}{\varepsilon_{ox}}\sqrt{2q\varepsilon_{Si}N_A} = \frac{1}{C_{ox}}\sqrt{2q\varepsilon_{Si}N_A} \tag{2.8}$$

in which q is the charge on an electron, ε_{ox} is the dielectric constant of the silicon dioxide, ε_{Si} is the dielectric constant of the silicon substrate, and N_A is the doping concentration density of the substrate. The term γ is the SPICE parameter called GAMMA. V_{t0} is the parameter *VTO*, N_A is the parameter *NSUB*, and $\phi_s = 2\phi_b$ is *PHI*, the surface potential at the onset of strong inversion.

Example

For with $N_A = 3 \times 10^{16}$ cm^{-3}, $t_{ox} = 200$Å, $\varepsilon_{ox} = 3.9 \times 8.85 \times 10^{-14}$ F/cm, $\varepsilon_{Si} = 11.7 \times 8.85 \times 10^{-14}$ F/cm, and $q = 1.6 \times 10^{-19}$ Coulomb

$$\gamma = \frac{0.2 \times 10^{-5}}{3.9 \times 8.85 \times 10^{-14}} \sqrt{2 \times 1.6 \times 10^{-19} \times 11.7 \times 8.85 \times 10^{-14} \times 3 \times 10^{16}}$$

$$= .57$$

$$\phi_b = .02586\ ln \left(\frac{3 \times 10^{16}}{1.5 \times 10^{10}}\right)$$

$$= .375$$

At a V_{sb} of 2.5 volts, and with

$$V_{t2.5} = V_{t0} + .57\left[\sqrt{.75 + 2.5} - \sqrt{.75}\right]$$

$$= V_{t0} + .53$$

Thus the threshold shifts by approximately half a volt with the source at 2.5 volts for these process parameters.

As we shall learn in Chapter 3, the type of CMOS process can have a large impact on this parameter for both n- and p-transistors. The increase in threshold voltage leads to lower device currents, which in turn leads to slower circuits.

2.2.2.2 Subthreshold Region

The cutoff region described by Eq. (2.5a) is also referred to as the subthreshold region, where I_{ds} increases exponentially with V_{ds} and V_{gs}. Although the value of I_{ds} is very small ($I_{ds} \approx 0$), the finite value of I_{ds} may be used to advantage to construct very low power circuits[10] or it may adversely affect circuits such as dynamic-charge storage nodes. As an approximation, Level 1 SPICE models set the subthreshold current to 0. (See Section 2.11 for the SPICE Level 3 subthreshold equations.)

2.2.2.3 Channel-length Modulation

Simplified equations that describe the behavior of an MOS device assume that the carrier mobility is constant, and do not take into account the variations in channel length due to the changes in drain-to-source voltage, V_{ds}.

For long channel lengths, the influence of channel variation is of little consequence. However, as devices are scaled down, this variation should be taken into account.

When an MOS device is in saturation, the effective channel length actually is decreased such that

$$L_{eff} = L - L_{short} \tag{2.9}$$

where

$$L_{short} = \sqrt{2\frac{\varepsilon_{Si}}{qN_A}(V_{ds} - (V_{gs} - V_t))}$$

The reduction in channel length increases the (W/L) ratio, thereby increasing β as the drain voltage increases. Thus rather than appearing as a constant current source with infinite output impedance, the MOS device has a finite output impedance. An approximation that takes this behavior into account[11] is represented by the following equation:

$$I_{ds} = \frac{k}{2}\frac{W}{L}(V_{gs} - V_t)^2(1 + \lambda V_{ds}) \tag{2.10}$$

where k is the process gain factor $\mu\varepsilon/t_{ox}$ and λ is an empirical *channel-length modulation* factor having a value in the range $0.02V^{-1}$ to $0.005V^{-1}$. In the SPICE level 1 model λ is the parameter *LAMBDA*.

2.2.2.4 Mobility Variation

The mobility, μ, describes the ease with which carriers drift in the substrate material. It is defined by

$$\mu = \frac{\textit{average carrier drift velocity } (V)}{\textit{Electric Field } (E)} \tag{2.11}$$

If the velocity, V, is given in cm/sec, and the electric field, E, in V/cm, the mobility has the dimensions cm^2/V-sec. The mobility may vary in a number of ways. Primarily, mobility varies according to the type of charge carrier. Electrons (negative-charge carriers) in silicon have a much higher mobility than holes (positive-charge carriers), resulting in n-devices having higher current-producing capability than the corresponding p-devices. Mobility decreases with increasing doping-concentration and increasing temperature. The temperature variation becomes less pronounced as the doping density increases. In SPICE μ is specified by the parameter *UO*.

2.2.2.5 Fowler-Nordheim Tunneling

When the gate oxide is very thin, a current can flow from gate to source or drain by electron tunneling through the gate oxide. This current is proportional to the area of the gate of the transistor as follows:[12,13,14]

$$I_{FN} = C_1 WLE_{ox}^2 e^{\frac{-E_0}{E_{ox}}} \tag{2.12}$$

where $E_{ox} = \frac{V_{gs}}{t_{ox}}$ is the electric field across the gate oxide and

E_0 and C_1 are constants.

This effect limits the thickness of the gate oxide as processes are scaled. However, it is of great use in electrically alterable programmable logic devices.

2.2.2.6 Drain Punchthrough

When the drain is at a high enough voltage with respect to the source, the depletion region around the drain may extend to the source, thus causing current to flow irrespective of the gate voltage (i.e., even if it is zero). This is known as a punchthrough condition. Currently, this effect is used in I/O protection circuits to limit the voltages across internal circuit nodes, although it will impact design as devices are scaled down by requiring that internal circuit voltages be reduced to a point where the effect does not occur.

2.2.2.7 Impact Ionization—Hot Electrons

As the length of the gate of an MOS transistor is reduced, the electric field at the drain of a transistor in saturation increases (for a fixed drain voltage). For submicron gate lengths, the field can become so high that electrons are imparted with enough energy to become what is termed "hot." These hot electrons impact the drain, dislodging holes that are then swept toward the negatively charged substrate and appear as a substrate current. This effect is known as *impact ionization*. Moreover, the electrons can penetrate the gate oxide, causing a gate current. Eventually this can lead to degradation of the MOS device parameters (threshold voltage, subthreshold current, and transconductance), which in turn can lead to the failure of circuits.[15,16,17] While the substrate current may be used in a positive manner to estimate the severity of the hot-electron effect, it can lead to poor refresh times in dynamic memories, noise in mixed signal systems, and possibly latchup. Hot holes do not normally present a problem because of their lower mobility.

LG Display Co., Ltd.
Exhibit 1012
Page 039

The presence of hot electrons has guided CMOS device engineering over the last few years. Chapter 3 shows some examples of the process steps that are used to provide long-lifetime submicron devices at 5 volts. Various circuit techniques that aim at reducing the voltage stress at the drains of n-transistors have also been proposed. Hot electrons will eventually push 3-volt and lower power supplies into prominence in CMOS design as the reduction in drain voltage markedly improves device lifetimes and reliability.

As an illustration of the relative magnitude of the substrate current, the following equation is representative[18] (for an $L = 0.8\ \mu$, $t_{ox} = 160\text{Å}$ CMOS process):

$$I_{substrate} = I_{ds}C1\ (V_{ds} - V_{dsat})^{C2} \tag{2.13}$$

where

$$C1 = 2.24 \times 10^{-5} - .1 \times 10^{-5} V_{ds}$$

$$C2 = 6.4$$

$$V_{dsat} = \frac{V_{tm} L_{eff} E_{sat}}{V_{tm} + L_{eff} E_{sat}}$$

with

$$V_{tm} = V_{gs} - V_{tn} - 0.13 V_{bs} - 0.25 V_{gs}$$

$$E_{sat} = 1.10 \times 10^{7} + 0.25 \times 10^{7} V_{gs}$$

L_{eff} is the effective channel length in meters.

2.2.3 MOS Models

In Section 2.2.2 we presented the ideal equations that describe the behavior of MOS transistors. While these incorporate some nonideal effects (channel-length modulation, threshold-voltage variation), they may not accurately model a specific device in a particular process. That is especially true for devices that have very small dimensions (gate lengths, gate widths, oxide thicknesses) as the modeling process becomes increasingly 3D in nature. Researchers have developed and refined a wide range of MOS models in an effort to predict more accurately the performance of MOS devices before they are fabricated for varying design scenarios. For instance, one might predict DC currents very accurately from raw process parameters, thus helping predict the behavior of an as yet untested device. However, because of the complexity

of the model, it might not be appropriate for a fast-execution-time model that might be needed for digital simulation purposes. In that case, a model based on parameters measured from an actual process might be appropriate.

Depending on the particular circuit level simulator that may be available, a wide variety of MOS simulation models may be used. For instance in one commercial circuit simulator there are over 10 different MOS models.[19] Many semiconductor vendors expend a great deal of effort to model the devices they manufacture. Many times these efforts are aimed at internal circuit simulators and proprietary models. Most CMOS digital foundry operations have been standardized on the LEVEL 3 models in SPICE as the level of circuit modeling that is required for CMOS digital system design. Table 2.1 is a summary of the main SPICE DC parameters that are used in Levels 1, 2, and 3 with representative values for a 1μ n-well CMOS process.

SPICE Level 3 model parameters also include process parameters that are used to calculate *VTO*, *KP*, *GAMMA*, *PHI*, and *LAMBDA* if they are not specified. For instance, if *GAMMA* is not specified, *TOX* and *NSUB* may be used to calculate it. Section 2.11 has a full description of the SPICE LEVEL 3 parameters and their use.

Table 2.1 SPICE DC Parameters

Parameter	nMOS	pMOS	Units	Description
VTO	0.7	0.7	volt	Threshold voltage
KP	8×10^{-5}	2.5×10^{-5}	A/V^2	Transconductance coefficient
GAMMA	.4	.5	$V^{0.5}$	Bulk threshold parameter
PHI	.37	.36	volt	Surface potential at strong inversion
LAMBDA	.01	.01	$volt^{-1}$	Channel length modulation parameter
LD	0.1×10^{-6}	0.1×10^{-6}	meter	Lateral diffusion
TOX	2×10^{-8}	2×10^{-8}	meter	Oxide thickness
NSUB	2×10^{16}	4×10^{16}	$1/cm^3$	Substrate doping density

2.2.4 Small Signal AC Characteristics

The MOS transistor can be represented by the simplified ($V_{sb} = 0$) small-signal equivalent model shown in Fig. 2.10 when biased appropriately. Here the MOS transistor is modeled as a voltage-controlled current source (g_m), an output conductance (g_{ds}), and the interelectrode capacitances. These values may be used, for instance, to calculate voltage amplification factors (gain) or bandwidth characteristics when considered along with other circuit elements.

LG Display Co., Ltd.
Exhibit 1012
Page 041

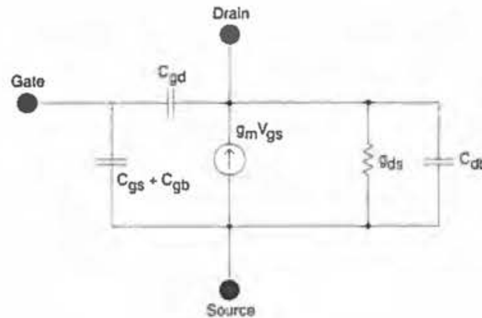


FIGURE 2.10 Small signal model for an MOS transistor

The output conductance (g_{ds}) in the linear region can be obtained by differentiating Eq. (2.5b) with respect to V_{ds}, which results in an output drain-source conductance of

$$g_{ds} = \beta[(V_{gs} - V_t) - 2V_{ds}]$$

$$= V_{ds} \overset{lim}{\rightarrow} 0 = \beta(V_{gs} - V_t) \tag{2.14}$$

Note that consistent with Eq. (2.5b), V_{ds} must be small compared to V_{gs} for the MOS device to be in a linear operating regime.

On rearrangement, the channel resistance R_c is approximated by

$$R_{c(linear)} = \frac{1}{\beta(V_{gs} - V_t)} \tag{2.15}$$

which indicates that it is controlled by the gate-to-source voltage. The relation defined by Eq. (2.15) is valid for gate to source voltages that maintain constant mobility in the channel. In contrast, in saturation [i.e., $V_{ds} \geq (V_{gs} - V_t)$], the MOS device behaves like a current source, the current being almost independent of V_{ds}. This may be verified from Eq. (2.5c) since

$$\frac{dI_{ds}}{dV_{ds}} = \frac{d\left[\frac{\beta}{2}(V_{gs} - V_t)^2\right]}{dV_{ds}} = 0 \tag{2.16}$$

In practice, however, due to channel shortening (Eq. 2.9) and other effects, the drain-current characteristics have some slope. This slope defines the g_{ds} of the transistor. The output conductance can be decreased by lengthening the channel (i.e., L).

The transconductance g_m expresses the relationship between output cur-

rent, I_{ds}, and the input voltage, V_{gs}, and is defined by

$$g_m = \frac{dI_{ds}}{dV_{gs}} \mid V_{ds} = \text{constant} \tag{2.17}$$

It is used to measure the gain of an MOS device. In the linear region g_m is given by

$$g_{m(linear)} = \beta V_{ds} \tag{2.18}$$

and in the saturation region by

$$g_{m(sat)} = \beta (V_{gs} - V_t) . \tag{2.19}$$

Since transconductance must have a positive value, the absolute value is used for voltages applied to p-type devices.

2.3 The Complementary CMOS Inverter–DC Characteristics

A complementary CMOS inverter is realized by the series connection of a p- and an n-device, as shown in Fig. 2.11. In order to derive the DC-transfer characteristics for the inverter (output voltage, V_{out}, as a function of the inverter, V_{in}), we start with Table 2.1, which outlines various regions of operation for the n- and p-transistors. In this table, V_{tn} is the threshold voltage of the n-channel device, and V_{tp} is the threshold voltage of the p-channel

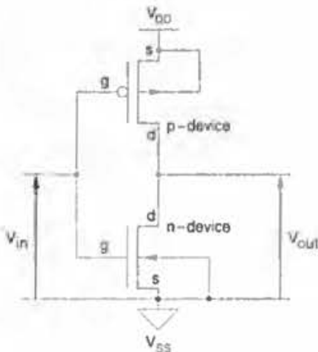


FIGURE 2.11 A CMOS inverter (with substrate connections)

LG Display Co., Ltd.
Exhibit 1012
Page 043

TABLE 2.2 Relations Between Voltages for the Three Regions of Operation of a CMOS Inverter

	CUTOFF	NONSATURATED	SATURATED
p-device	$V_{gsp} > V_{tp}$ $V_{in} > V_{tp} + V_{DD}$	$V_{gsp} < V_{tp}$ $V_{in} < V_{tp} + V_{DD}$ $V_{dsp} > V_{gsp} - V_{tp}$ $V_{out} > V_{in} - V_{tp}$	$V_{gsp} < V_{tp}$ $V_{in} < V_{tp} + V_{DD}$ $V_{dsp} < V_{gsp} - V_{tp}$ $V_{out} < V_{in} - V_{tp}$
n-device	$V_{gsn} < V_{tn}$ $V_{in} < V_{tn}$	$V_{gsn} > V_{tn}$ $V_{in} > V_{tn}$ $V_{dsn} < V_{gs} - V_{tn}$ $V_{out} < V_{in} - V_{tn}$	$V_{gsn} > V_{tn}$ $V_{in} > V_{tn}$ $V_{dsn} > V_{gs} - V_{tn}$ $V_{out} > V_{in} - V_{tn}$

device. The objective is to find the variation in output voltage (V_{out}) for changes in the input voltage (V_{in}).

We begin with the graphical representation of the simple algebraic equations described by Eq. (2.5) for the two inverter transistors shown in Fig. 2.12(a).[20] The absolute value of the p-transistor drain current I_{ds} inverts this characteristic. This allows the VI characteristics for the p-device to be reflected about the x-axis (Fig. 2.12b). This step is followed by taking the absolute value of the p-device, V_{ds}, and superimposing the two characteristics yielding the resultant curves shown in Fig. 2.12(c). The input/output transfer curve may now be determined by the points of common V_{gs} intersection in Fig. 2.12(c). Thus, solving for $V_{inn} = V_{inp}$ and $I_{dsn} = I_{dsp}$ gives the desired transfer characteristics of a CMOS inverter as illustrated in Fig. 2.13. The switching point is typically designed to be 50 percent of the magnitude of the supply voltage: $= V_{DD}/2$. During transition, both transistors in the CMOS inverter are momentarily "ON," resulting in a short pulse of current drawn from the power supply. This is shown by the dotted line in Fig. 2.13.

The operation of the CMOS inverter can be divided into five regions (Fig. 2.13). The behavior of n- and p-devices in each of the regions may be found by using Table 2.2.

Region A. This region is defined by $0 \leq V_{in} \leq V_{tn}$ in which the n-device is cut off ($I_{dsn} = 0$), and the p-device is in the linear region. Since $I_{dsn} = -I_{dsp}$, the drain-to-source current I_{dsp} for the p-device is also zero. But for $V_{dsp} = V_{out} - V_{DD}$, with $V_{dsp} = 0$, the output voltage is

$$V_{out} = V_{DD} \tag{2.20}$$

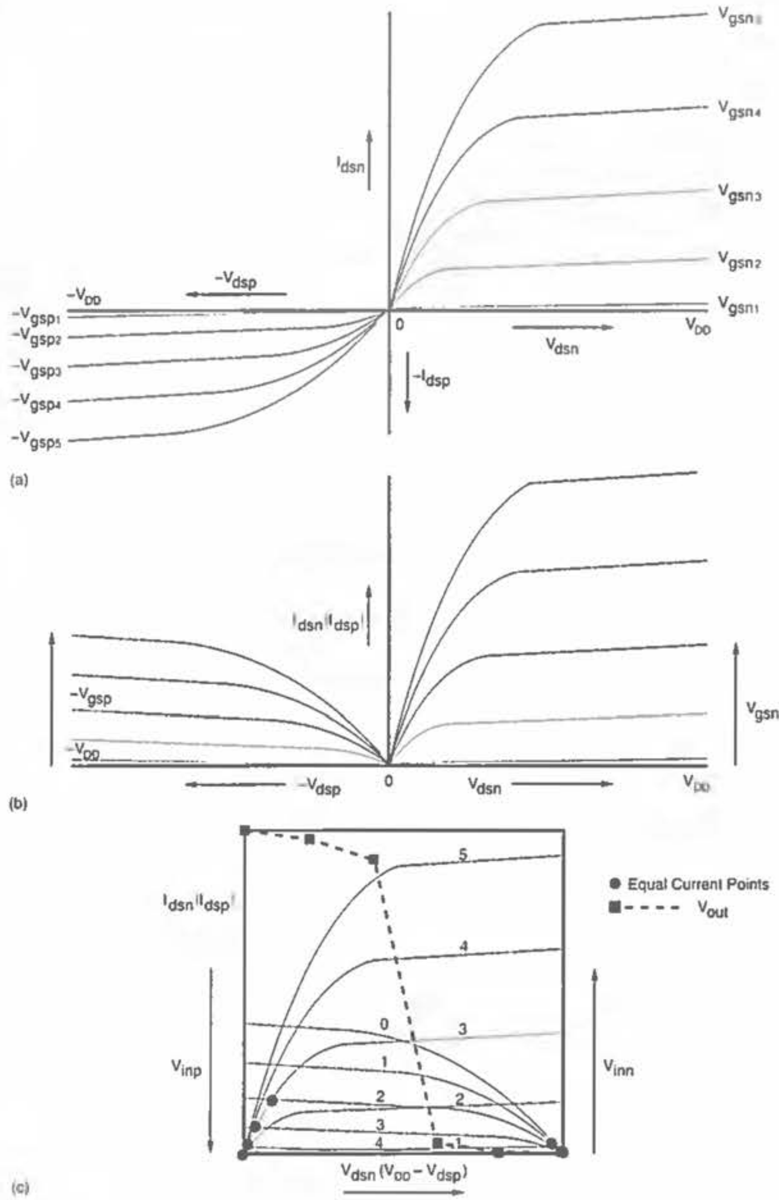


FIGURE 2.12 Graphical derivation of CMOS inverter characteristic

Region B. This region is characterized by $V_{tn} \leq V_{in} < V_{DD}/2$ in which the p-device is in its nonsaturated region ($V_{ds} \neq 0$) while the n-device is in saturation. The equivalent circuit for the inverter in this region can be represented by a resistor for the p-transistor and a current source for the n-

LG Display Co., Ltd.
Exhibit 1012
Page 045

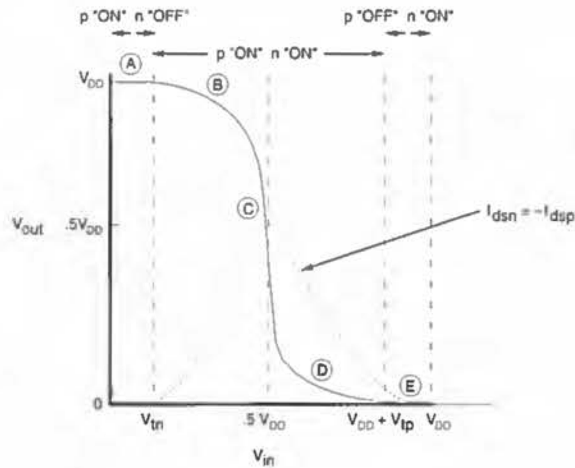


FIGURE 2.13 CMOS inverter DC transfer characteristic and operating regions

transistor as shown by Fig. 2.14(a). The saturation current I_{dsn} for the n-device is obtained by setting $V_{gs} = V_{in}$. This results in

$$I_{dsn} = \beta_n \frac{[V_{in} - V_{tn}]^2}{2} \tag{2.21}$$

where

$$\beta_n = \frac{\mu_n \varepsilon}{t_{ox}} \left(\frac{W_n}{L_n}\right)$$

and

V_{tn} = threshold voltage of n-device
μ_n= mobility of electrons
W_n = channel width of n-device
L_n = channel length of n-device.

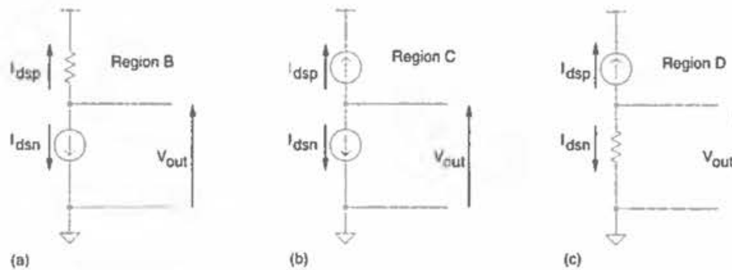


FIGURE 2.14 Equivalent circuits for operating regions of a CMOS inverter

LG Display Co., Ltd.
Exhibit 1012
Page 046

The current for the p-device can be obtained by noting that

$$V_{gs} = (V_{in} - V_{DD})$$

and

$$V_{ds} = (V_{out} - V_{DD})$$

and therefore

$$I_{dsp} = -\beta_p \left[(V_{in} - V_{DD} - V_{tp})(V_{out} - V_{DD}) - \frac{(V_{out} - V_{DD})^2}{2} \right], \quad (2.22)$$

where

$$\beta_p = \frac{\mu_p \varepsilon}{t_{ox}} \left(\frac{W_p}{L_p} \right)$$

and

V_{tp} = threshold voltage of p-device
μ_p = mobility of holes
W_p = channel width of p-device
L_p = channel length of p-device.

Substituting

$$I_{dsp} = -I_{dsn}$$

the output voltage V_{out} can be expressed as

$$V_{out} = (V_{in} - V_{tp}) + \sqrt{(V_{in} - V_{tp})^2 - 2\left(V_{in} - \frac{V_{DD}}{2} - V_{tp}\right) V_{DD} - \frac{\beta_n}{\beta_p}(V_{in} - V_{tn})^2} \quad (2.23)$$

Region C. In this region both the n- and p-devices are in saturation. This is represented by the schematic in Fig. 2.14(b) which shows two current sources in series.

The saturation currents for the two devices are given by

$$I_{dsp} = -\frac{\beta_p}{2}(V_{in} - V_{DD} - V_{tp})^2$$

LG Display Co., Ltd.
Exhibit 1012
Page 047

$$I_{dsn} = \frac{\beta_n}{2}(V_{in} - V_{tn})^2$$

with

$$I_{dsp} = -I_{dsn}.$$

This yields

$$V_{in} = \frac{V_{DD} + V_{tp} + V_{tn}\sqrt{\frac{\beta_n}{\beta_p}}}{1 + \sqrt{\frac{\beta_n}{\beta_p}}}. \quad (2.24)$$

By setting

$$\beta_n = \beta_p \text{ and } V_{tn} = -V_{tp},$$

we obtain

$$V_{in} = \frac{V_{DD}}{2}, \quad (2.25)$$

which implies that region C exists only for one value of V_{in}. The possible values of V_{out} in this region can be deduced as follows:

n-channel: $V_{in} - V_{out} < V_{tn}$
$V_{out} > V_{in} - V_{tn}$
p-channel: $V_{in} - V_{out} > V_{tp}$
$V_{out} < V_{in} - V_{tp}$.

Combining the two inequalities results in

$$V_{in} - V_{tn} < V_{out} < V_{in} - V_{tp}. \quad (2.26)$$

This indicates that with $V_{in} = \frac{V_{DD}}{2}$, V_{out} varies within the range shown. Of course, we have assumed that an MOS device in saturation behaves like an ideal current source with drain-to-source current being independent of V_{ds}. In reality, as V_{ds} increases, I_{ds} also increases slightly; thus region C has a finite slope. The significant factor to be noted is that in region C we have two current sources in series, which is an "unstable" condition. Thus a small

input voltage has a large effect at the output. This makes the output transition very steep, which contrasts with the equivalent nMOS inverter characteristic. (See Section 2.4.) The relation defined by Eq. (2.24) is particularly useful since it provides the basis for defining the gate threshold V_{inv}, which corresponds to the state where $V_{out} = V_{in}$. This region also defines the "gain" of the CMOS inverter when used as a small signal amplifier.

Region D. This region is described by $V_{DD}/2 < V_{in} \leq V_{DD} - V_{tp}$. The p-device is in saturation while the n-device is operating in its nonsaturated region. This condition is represented by the equivalent circuit shown in Fig. 2.14(c). The two currents may be written as

$$I_{dsp} = -\frac{1}{2}\beta_p (V_{in} - V_{DD} - V_{tp})^2$$

and

$$I_{dsn} = \beta_n \left[(V_{in} - V_{tn}) V_{out} - \frac{V_{out}^2}{2} \right]$$

with

$$I_{dsp} = -I_{dsn}$$

The output voltage becomes

$$V_{out} = (V_{in} - V_{tn}) - \sqrt{(V_{in} - V_{tn})^2 - \frac{\beta_p}{\beta_n}(V_{in} - V_{DD} - V_{tp})^2} \quad \textbf{(2.27)}$$

Region E. This region is defined by the input condition $V_{in} \geq V_{DD} - V_{tp}$, in which the p-device is cut off ($I_{dsp} = 0$), and the n-device is in the linear mode. Here, $V_{gsp} = V_{in} - V_{DD}$, which is more positive than V_{tp}. The output in this region is

$$V_{out} = 0. \quad \textbf{(2.28)}$$

From the transfer curve of Fig. 2.13, it may be seen that the transition between the two states is very steep. This characteristic is very desirable because the noise immunity is maximized. This is covered in more detail in Section 2.3.2. For convenience, the characteristics associated with the five regions are summarized in Table 2.3.

LG Display Co., Ltd.
Exhibit 1012
Page 049

TABLE 2.3 Summary of CMOS Inverter Operation

REGION	CONDITION	p-device	n-device	OUTPUT
A	$0 \le V_{in} < V_{tn}$	nonsaturated	cutoff	$V_{out} = V_{DD}$
B	$V_{tn} \le V_{in} < \frac{V_{DD}}{2}$	nonsaturated	saturated	Eq. (2.23)
C	$V_{in} = \frac{V_{DD}}{2}$	saturated	saturated	$V_{out} \ne f(V_{in})$
D	$\frac{V_{DD}}{2} < V_{in} \le V_{DD} - \lvert V_{tp} \rvert$	saturated	nonsaturated	Eq. (2.27)
E	$V_{in} > V_{DD} - \lvert V_{tp} \rvert$	cutoff	nonsaturated	$V_{out} = V_{SS}$

2.3.1 β_n/β_p Ratio

In order to explore the variations of the transfer characteristic as a function of β_n/β_p, the transfer curve for several values of β_n/β_p are plotted in Fig. 2.15(a). Here, we note the gate-threshold voltage, V_{inv}, where $V_{in} = V_{out}$ is

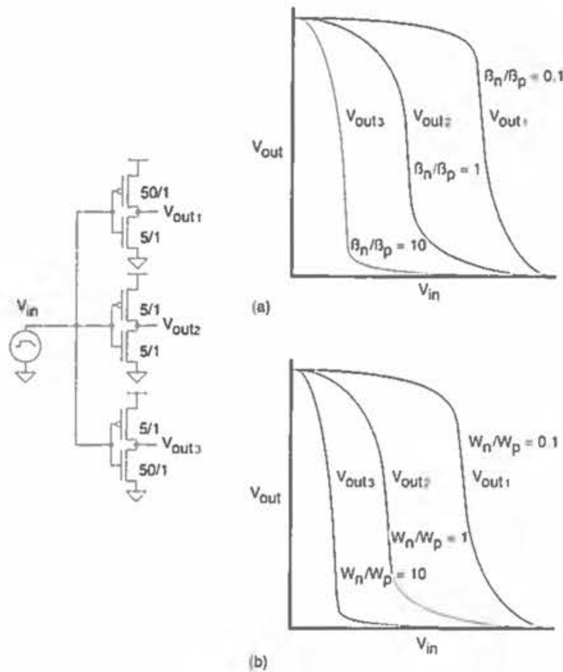


FIGURE 2.15 Influence of $\frac{\beta_n}{\beta_p}$ on inverter DC transfer characteristic

dependent on β_n/β_p. Thus, for a given process, if we want to change β_n/β_p, we need to change the channel dimensions, i.e., channel-length L and channel-width W. From Fig. 2.15(a) it can be seen that as the ratio β_n/β_p is decreased, the transition region shifts from left to right; however, the output voltage transition remains sharp, and hence the switching performance is not affected. For the CMOS inverter a ratio of

$$\frac{\beta_n}{\beta_p} = 1 \tag{2.29}$$

may be desirable since it allows a capacitive load to charge and discharge in equal times by providing equal current-source and -sink capabilities. This will be discussed further in Chapter 4. For interest, the inverter transfer curve is also plotted (Figure 2.15b) for W_n/W_p (the width of the n- and p-transistors). This shows a relative shift to the left compared with the β ratioed case because the p-device has inherently lower gain.

Temperature also has an effect on the transfer characteristic of an inverter.[21] As the temperature of an MOS device is increased, the effective carrier mobility, μ, decreases. This results in a decrease in β, which is related to temperature T by

$$\beta \alpha T^{-1.5} \tag{2.30}$$

Therefore

$$I_{ds} \alpha T^{-1.5} \tag{2.31}$$

Since the voltage transfer characteristics depend on the ratio β_n/β_p, and the mobility of both holes and electrons are similarly affected, this ratio is independent of temperature to a good approximation. Both V_{tn} and V_{tp} decrease slightly as temperature increases, and the extent of region A is reduced while the extent of region E increases. Thus the overall transfer characteristics of Fig. 2.15 shift to the left as temperature increases. Based on the figures given earlier, if the temperature rises by 50°C , the thresholds drop by 200mV each. This would cause a .4 V shift in the input threshold of the inverter.

2.3.2 Noise Margin

Noise margin is a parameter closely related to the input-output voltage characteristics. This parameter allows us to determine the allowable noise voltage on the input of a gate so that the output will not be affected. The

LG Display Co., Ltd.
Exhibit 1012
Page 051

specification most commonly used to specify noise margin (or noise immunity) is in terms of two parameters—the *LOW* noise margin, NM_L, and the *HIGH* noise margin, NM_H. With reference to Fig. 2.16, NM_L is defined as the difference in magnitude between the maximum LOW output voltage of the driving gate and the maximum input LOW voltage recognized by the driven gate. Thus

$$NM_L = |V_{ILmax} - V_{OLmax}| \quad (2.32)$$

The value of NM_H is the difference in magnitude between the minimum HIGH output voltage of the driving gate and the minimum input HIGH voltage recognized by the receiving gate. Thus

$$NM_H = |V_{OHmin} - V_{IHmin}| \quad (2.33)$$

where

V_{IHmin} = minimum HIGH input voltage
V_{ILmax} = maximum LOW input voltage
V_{OHmin} = minimum HIGH output voltage
V_{OLmax} = maximum LOW output voltage.

These definitions are illustrated in Fig. 2.16.

Generally, it is desirable to have $V_{IH} = V_{IL}$ and for this to be a value that is midway in the "logic swing," V_{OL} to V_{OH}. This implies that the transfer characteristic should switch abruptly; that is, there should be high gain in the

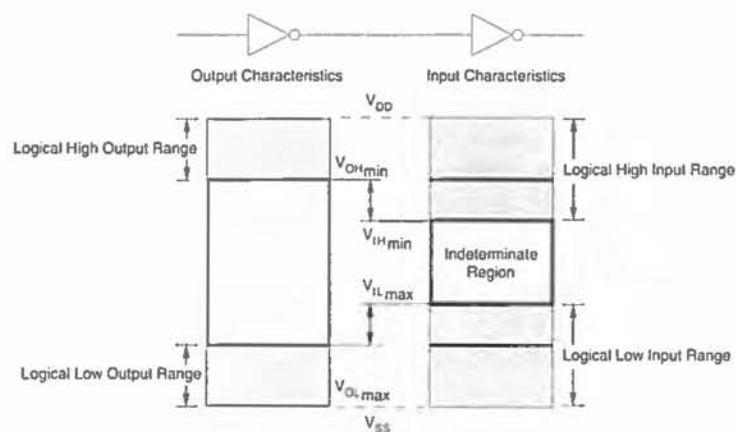


FIGURE 2.16 Noise margin definitions

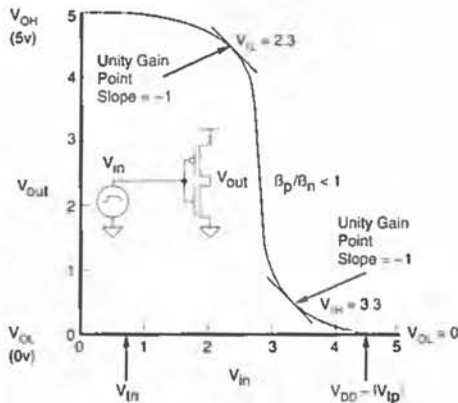


FIGURE 2.17 CMOS inverter noise margins

transition region. For the purpose of calculating noise margins, the transfer characteristic of a typical inverter and the definition of voltage levels V_{IL}, V_{OL}, V_{IH}, V_{OH} are shown in Fig. 2.17. To determine V_{IL}, we note that the inverter is in region B of operation, where the p-device is in its linear region while the n-device is in saturation. The V_{IL} is found by determining the unity gain point in the inverter transfer characteristic where the output transitions from V_{OH}. Similarly, V_{IH} is found by using the unity gain point at the V_{OL} end of the characteristic. For the inverter shown the NM_L is 2.3 volts while the NM_H is 1.7 volts.[22]

Note that if either NM_L or NM_H for a gate are reduced ($\approx 0.1\ V_{DD}$), then the gate may be susceptible to switching noise that may be present on the inputs. Apart from considering a single gate, one must consider the net effect of noise sources and noise margins on cascaded gates in assessing the overall noise immunity of a particular system. This is the reason to keep track of noise margins. Quite often noise margins are compromised to improve speed. Circuit examples later in this book will illustrate this trade-off.

2.3.3 The CMOS Inverter As an Amplifier

It should be noted that the CMOS inverter when used as a logic element is in reality an analog amplifier operated under saturating conditions. In region C in Fig. 2.14, the CMOS inverter acts as an inverting linear amplifier with a characteristic of

$$V_{out} = -AV_{in} \tag{2.34}$$

where A is the stage gain.

LG Display Co., Ltd.
Exhibit 1012
Page 053

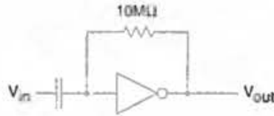


FIGURE 2.18 The CMOS inverter as an amplifier

This region may be further examined with a circuit simulator by using the circuit shown in Fig. 2.18, with a high-value resistor between input and output (10M Ω). The input is DC isolated using a capacitor. The gain of this amplifier is estimated by using the small-signal model of the amplifier shown in Fig. 2.10. This circuit is valid for small signals around the linear operating point of the amplifier. The gain is approximately given by

$$\begin{aligned} A &= g_{mtotal} R_{dseffective} \\ &= (g_{mn} + g_{mp})(r_{dsn} \parallel r_{dsp}) \\ &= g_m r_{ds} \text{ (if } g_{mn} = g_{mp} \text{ and } r_{dsn} = r_{sdp}) \end{aligned} \tag{2.35}$$

This gain is very dependent on the process and transistors used in the circuit but can be in the range from 100 to over 1000. The gain is enhanced by lengthening the transistors to improve the r_{ds} values. This improvement comes at the expense of speed and bandwidth of the amplifier.

2.4 Static Load MOS Inverters

Apart from the CMOS inverter, there are many other forms of MOS inverter that may be used to build logic gates. Figure 2.19(a) shows a generic nMOS inverter that uses either a resistive load or a constant current source. For the resistor case, if we superimpose the resistor-load line on the VI characteristics of the pull-down transistor (Fig. 2.19b), we can see that at a V_{gs} of 5 volts, the output is some small V_{ds} (V_{OL}) (Fig. 2.19c). When $V_{gs} = 0$ volts, V_{ds} rises to 5 volts. As the resistor is made larger, the V_{OL} decreases and the current flowing when the inverter is turned on decreases. Correspondingly, as the load resistor is decreased in value, the V_{OL} rises and the on current rises. Selection of the resistor value would seek a compromise between V_{OL}, the current drawn and the pull-up speed, which vary with the value of the load resistor.

The resistor- and current-source-load inverters shown in Fig. 2.19 are normally implemented using transistors in CMOS processes. In some memory processes, resistors are implemented using highly resistive undoped polysilicon. When transistors are used the inverter is called a saturated load inverter if the load transistor is operated in saturation as a constant current source. If the load transistor is biased for use as a resistor, then it is called an unsaturated load inverter.

In this section we will examine a number of static load inverters that one can implement in CMOS processes. Usually the reason for doing this is to reduce the number of transistors used for a gate to improve density and/or to lower dynamic power consumption.

LG Display Co., Ltd.
Exhibit 1012
Page 054

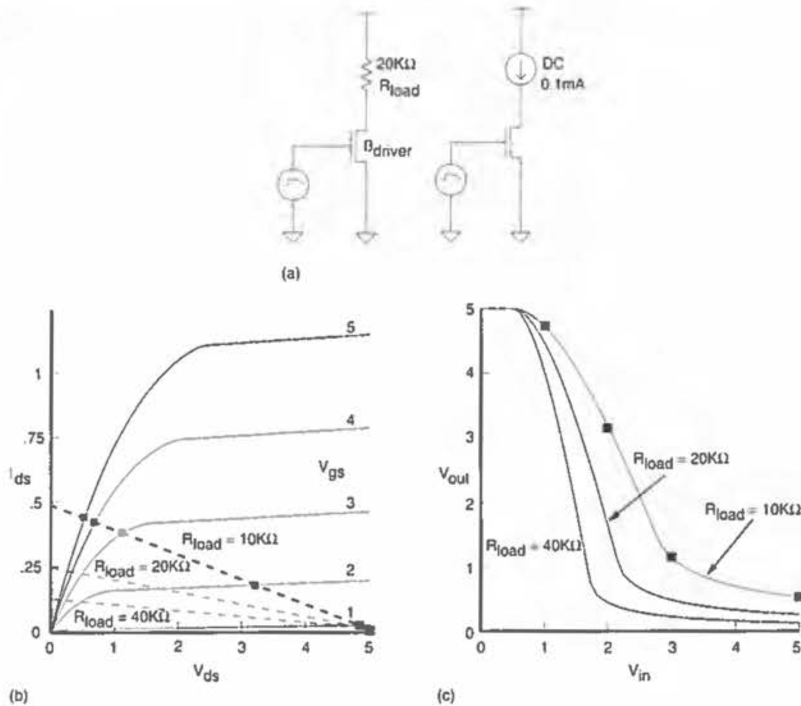


Figure 2.19 A generic static load inverter

2.4.1 The Pseudo-nMOS Inverter

Figure 2.20(a) shows an inverter that uses a p-device pull-up or load that has its gate permanently grounded. An n-device pull-down or driver is driven with the input signal. This is roughly equivalent to the use of a depletion load in nMOS technology (which preceded CMOS technology as a major systems technology) and is thus called "pseudo-nMOS." This circuit is used in a variety of CMOS logic circuits. Similar to the complementary inverter, a graphical solution to the transfer characteristic is shown in Fig. 2.20(b) for various sized p-devices for a particular CMOS process. This shows that the ratio of β_n/β_p affects the shape of the transfer characteristic and the V_{OL} of the inverter (shown in Fig. 2.20c). Figure 2.20(d) shows that when the driver is turned on, a constant DC current flows in the circuit. This is to be contrasted with the CMOS inverter in which no DC current flows when the input is either the terminal high or low state. The importance of whether DC current flows, and hence whether one can use the pseudo-nMOS inverter, depends on the application. CMOS watch circuits rely on the fact that when the circuit is not switching, no current is drawn from the small battery that powers

LG Display Co., Ltd.
Exhibit 1012
Page 055

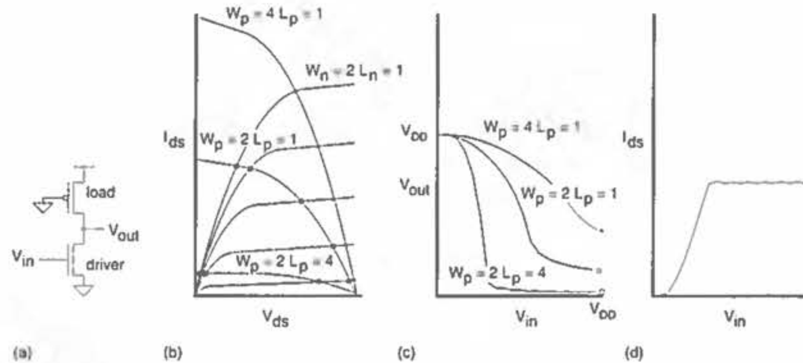


FIGURE 2.20 The pseudo-nMOS inverter and DC transfer characteristics

the watch. In this application, having circuits that consumed DC current would not be advisable. Similarly in circuits which required a power-down mode (as in palmtop or portable computers) one might not want such circuits. Finally, the fact that CMOS complementary circuits do not draw DC current has led some semiconductor manufacturers to have a gross test of CMOS chips that tests the DC current of a chip (IDDQ testing—see Chapter 7). If there is DC current, they assume there is some fault internally and have to do no more testing of that die. Notwithstanding these applications where pseudo-nMOS gates are not applicable, they do find wide application in high-speed circuits and circuits that require large fan-in NOR gates. Even in DC power critical applications, the pseudo-nMOS gate may be used by selectively grounding the gate of the p-device pull-up transistor. (*Note:* The output voltage of a pseudo-nMOS inverter with both driver and load transistors turned off will depend on the subthreshold characteristics of the transistors. This should be rigorously simulated if contemplated, or the output should be clamped to a known voltage.)

For the circuit shown in Fig. 2.20 the current in the n driver transistor is given by

$$I_{dsn} = \frac{\beta_n}{2}(V_{inv} - V_{tn})^2 \quad (V_{out} > V_{in} - V_{tn}).$$

The p-device I_{ds} with $V_{gsp} = -V_{DD}$ is

$$I_{dsp} = \beta_p\left[(-V_{DD} - V_{tp})(V_{out} - V_{DD}) - \frac{(V_{out} - V_{DD})^2}{2}\right].$$

LG Display Co., Ltd.
Exhibit 1012
Page 056

Equating the two currents we obtain

$$\frac{\beta_n}{2}(V_{in}-V_{tn})^2 = \beta_p\left[(-V_{DD}-V_{tp})(V_{out}-V_{DD}) - \frac{(V_{out}-V_{DD})^2}{2}\right].$$

Solving for V_{out},

$$V_{out} = -V_{tp} + \sqrt{(V_{DD}+V_{tp})^2 - C} \qquad (2.36)$$

where $C = k(V_{in}-V_{tn})^2$

and $k = \frac{\beta_n}{\beta_p}$

also

$$\frac{\beta_n}{\beta_p} = \frac{(V_{DD}+V_{tp})^2 - (V_{out}+V_{tp})^2}{(V_{in}-V_{tn})^2} \qquad (2.37)$$

Figure 2.21(a) shows two cascaded pseudo-nMOS inverters. For equal noise margins, the gate-threshold voltage V_{inv} should be set to approximately $0.5V_{DD}$. (Another criteria might set V_{inv} to be halfway between V_{IL} and V_{IH}.) At this operating point, the n-device (pull-down) is in saturation ($0 < V_{gsn} - V_{tn} < V_{dsn}$), and the p-device (pull-up) is in the linear mode of operation ($0 < V_{dsp} < V_{gsp} - V_{tp}$).

With $V_{inv} = 0.5V_{DD}$, $V_{tn} = |V_{tp}| = 0.2V_{DD}$, $V_{DD} = 5$ volts, the following result is obtained

$$\frac{\beta_n}{\beta_p} = 6$$

Recalling that the technology and geometry contributions to β, the ratio of widths of the n-device to the p-device might range between approximately 3/1 for $\mu_n/\mu_p = 2$ and 2/1 where $\mu_n/\mu_p = 3$. Figure 2.21(b) shows some typical transfer characteristics for varying β_n/β_p ratios. The noise margins are as follows:

β_n/β_p	V_{IL}	V_{IH}	V_{OL}	V_{OH}	NM_L	NM_H
2	3.4	4.5	1.4	5	2.0	0.5
4	1.8	3.3	0.6	5	1.2	2.7
6	1.4	2.8	0.35	5	1.05	3.2
8	1.1	2.4	0.24	5	0.86	3.6
100	0.5	1.1	0.00	5	0.5	3.9

LG Display Co., Ltd.
Exhibit 1012
Page 057

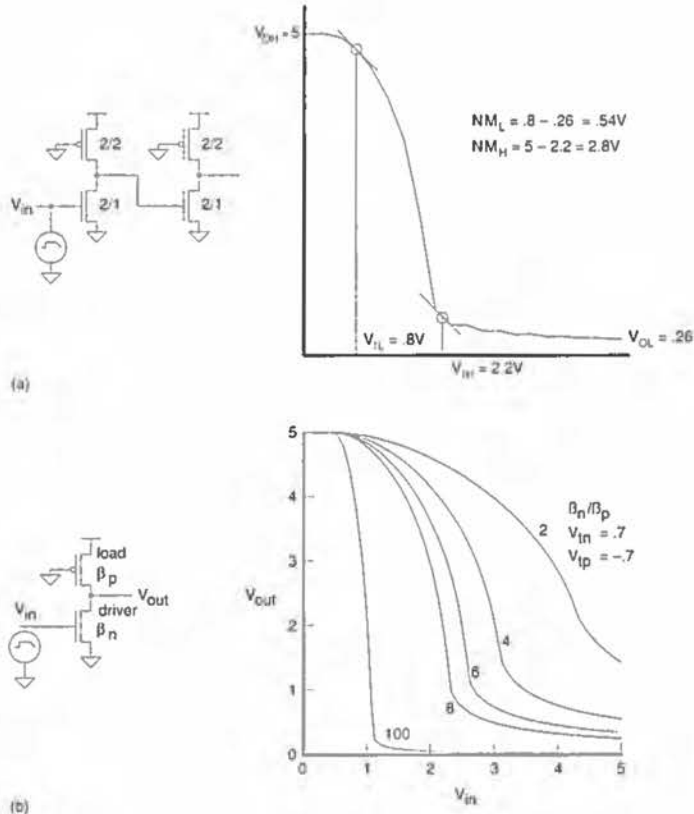


FIGURE 2.21 Cascaded pseudo-nMOS inverters

From this one can see that the low noise margin is considerably worse than the high noise margin. The overall noise margin of a pseudo-nMOS circuit can be enhanced considerably by following such a stage with a CMOS stage ($\beta_n/\beta_p = 1$). In this case for $\beta_n/\beta_p = 6$,

V_{IL}	V_{IH}	V_{OL}	V_{OH}	NM_L	NM_H
2.3	3.3	.35	5	1.95	1.7

This inverter finds widespread use in circuits where an "n-rich" circuit is required and the power dissipation can be tolerated. Typical uses include static ROMs and PLAs. Note that the circuit could use n-load devices and p-active pull-ups, if this were of advantage.

Rather than operate the p-transistor in the linear region it is possible to operate it as a constant current source (saturated load). Figure 2.22(a) shows an inverter with a p-transistor biased to be a constant current source ($V_{out} >$

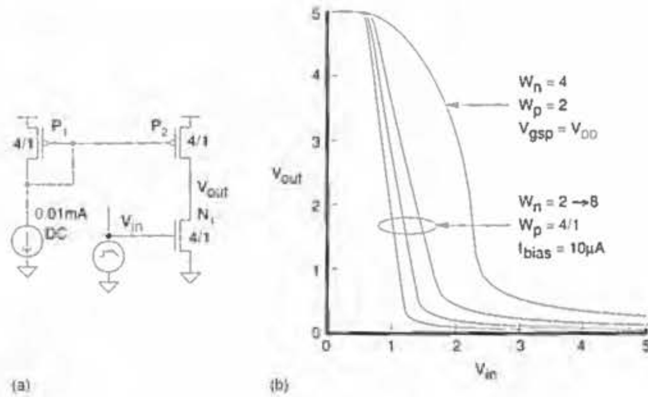


FIGURE 2.22 Constant current source load pseudo-nMOS inverter

$V_{gsp} - V_{tp}$). The constant current p load allows the inverter characteristics to be set to compensate for process changes (see also Fig. 5.27). Figure 2.22(b) shows transfer characteristics for a variety of n-transistor widths. (See also Section 5.4.3.)

2.4.2 Unsaturated Load Inverters

Figure 2.23(a) shows an inverter using an nMOS transistor load. This type of inverter was used in nMOS technologies prior to the availability of nMOS depletion loads and in pMOS technologies prior to the availability of nMOS technologies. It is included here for completeness. The high level is an n threshold down from V_{DD} (but remember that the threshold is modified by

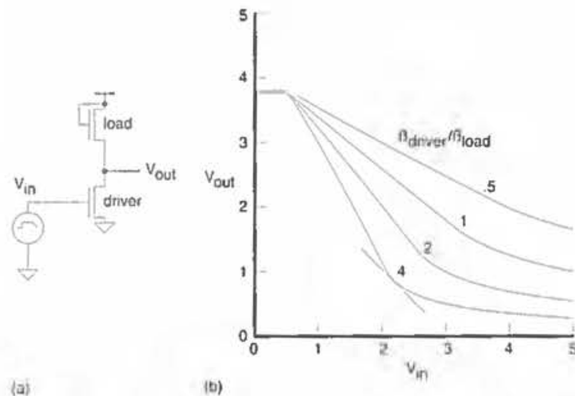


FIGURE 2.23 Unsaturated load inverter

the body effect because the source of the n-load transistor is above V_{SS}). Figure 2.23(b) shows the transfer characteristics for a variety of pull-up to pull-down ratios. For $k = 4$ $V_{OL} = .24$ volts, $V_{IH} = 2.2$ volts, $V_{OH} = 3.8$ volts and $V_{IL} = .56$ volts. Thus the low noise margin is .32 volts and the high noise margin is 1.6 volts for cascaded circuits. The small low noise margin would make this inverter nonoptimal as a conventional logic circuit. However, it might be used in isolated circumstances where p-transistors were not wanted (for instance, in some I/O structures).

2.4.3 Saturated Load Inverters

A number of other "pseudo-nMOS" inverter configurations are possible. Figure 2.24(a) shows a p load with its gate connected to the output. The transfer characteristic is shown in Fig. 2.24(b) for a number of pull-up/pull-down ratios. The output rises to a p threshold down from V_{DD}. In addition as the output voltage approaches $V_{DD} - |V_{tp}|$, the V_{ds} across the pull-up is reduced, thus decreasing the current flowing in the pull-up, which has a detrimental effect on the pull-up speed. While $V_{out} > V_{in} - V_{tn}$ (i.e., for small V_{in} values), the driver transistor is in saturation

$$I_{dsdriver} = \frac{\beta_{driver}}{2}(V_{in} - V_{tn})^2. \tag{2.38}$$

Similarly the load device I_{ds} is permanently in the nonsaturated region

$$I_{dsload} = \frac{\beta_{load}}{2}(V_{out} - V_{DD} - V_{tp})^2. \tag{2.39}$$

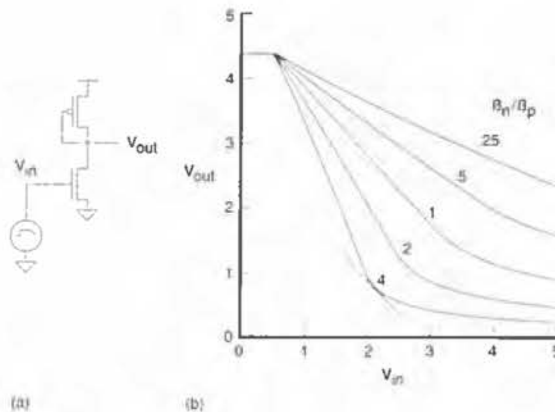


FIGURE 2.24 Saturated load inverter

LG Display Co., Ltd.
Exhibit 1012
Page 060

Equating the two currents we obtain

$$\frac{\beta_{driver}}{2}(V_{in}-V_{tn})^2 = \frac{\beta_{load}}{2}(V_{out}-V_{DD}-V_{tp})^2.$$

Upon rearrangement,

$$V_{out} = V_{DD}+V_{tp}+\sqrt{k}(V_{in}-V_{tn}) \quad (2.40)$$

where $k = \frac{\beta_{driver}}{\beta_{load}}$.

This effectively gives the V_{OH} value ($V_{in}=0$). Similar calculations can yield the V_{OL}. From Fig. 2.24(b), for $k=4$ $V_{OL}=.24$ volts, $V_{IH}=2.1$ volts, $V_{OH}=4.4$ volts, and $V_{IL}=.5$ volts. Thus the low noise margin is .26 volts and the high noise margin is 2.3 volts. The small low-noise-margin makes this inverter unsuitable for cascaded logic use, but it is of use in other circumstances and forms the basis for the differential pair inverter, which we will examine subsequently.

Finally, Fig. 2.25 shows an nMOS depletion load inverter. This inverter relies on the existence of a depletion nMOS transistor to form the load device. That is, the threshold of the depletion transistor is negative. While this is relatively rare in CMOS processes, this inverter formed the basis for the generation of MOS technology that ushered in the VLSI era. By connecting the gate of the load to the output, a constant current load is formed. Unlike the inverter shown in Fig. 2.24, which uses a p-device as a constant current load, the output of this inverter can rise to a full V_{DD} level.

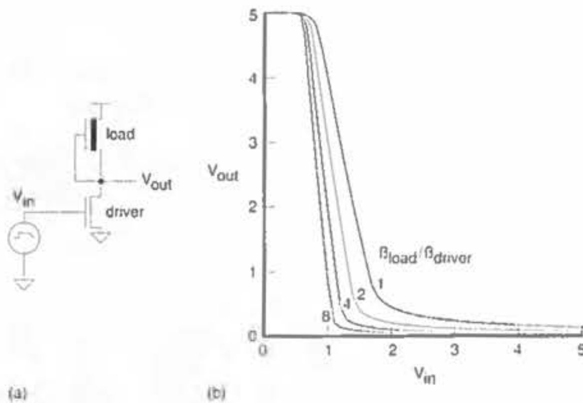


FIGURE 2.25 Depletion load inverter

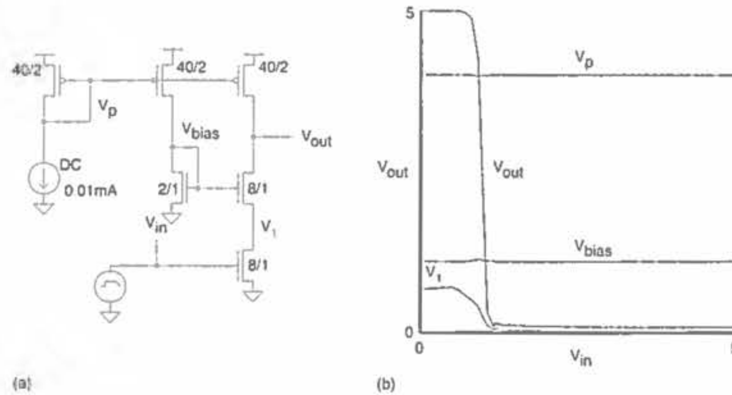


FIGURE 2.26 Cascode inverter

2.4.4 The Cascode Inverter

The cascode inverter is shown in Fig. 2.26. It resembles a pseudo-nMOS inverter but with an n-transistor connected in series with the pull-down n-transistor. If the gate of the series transistor is held at a constant voltage, V_{bias}, the drain of the driver transistor (V_1) will be held to an n threshold below V_{bias}. The output node, V_{out}, swings from V_{DD} to V_{SS}. The series transistor acts as a "common gate" amplifier and in effect isolates the V_1 node from the V_{out} node and keeps the signal swing on V_1 between V_{SS} and $V_{bias} - V_{tn}$. This feature will be used in a logic family discussed in Chapter 5.

2.4.5 TTL Interface Inverter

One final CMOS inverter is shown in Fig. 2.27.[23] This is of use in interfacing to TTL logic systems. The series-p load basically feeds a conventional

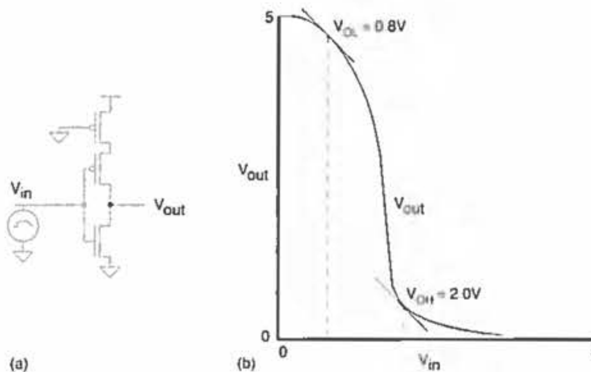


FIGURE 2.27 TTL input inverter

LG Display Co., Ltd.
Exhibit 1012
Page 062

CMOS inverter with a reduced V_{DD} supply. This changes the input threshold to suit a TTL output. ($V_{OL} = 0.8V$ $V_{OH} = 2.0V$).

2.5 The Differential Inverter

All of the inverters that we have examined thus far have been singled-ended; that is, they have a single input signal and produce a single output signal. An inverter that uses two differential inputs and produces two differential outputs is shown in Fig. 2.28(a). Two n-transistors have their sources commoned and fed by a constant current source that is in turn connected to ground. The drains of each n-transistor are connected to resistor loads that are connected to the supply voltage.

If the input voltages V_{left} and V_{right} are set to the same voltage $V_{quiescent}$, then each transistor has a V_{gs} of $V_{quiescent} - V_N$, where V_N is the voltage across the constant current source. Thus the I_{ds} for each transistor is equal and the output voltages V_{out1} and V_{out2} are equal. If the voltages V_{left} and V_{right} are increased equally, then V_N rises to maintain the constant current through the current source. The output voltages, V_{out1} and V_{out2}, will stay at the same value. Applying this common signal to both inputs therefore results in no gain (ideally); this gain is referred to as the Common Mode Gain. If V_{left} is increased by δV, and V_{right} is decreased by δV, then the current in N_1 will increase by δI and the current in N_2 will decrease by δI. V_{out1} will decrease by δIR and V_{out2} will increase by δIR. Thus the differential gain from V_{left} to V_{out1} is

$$A_{diff} = -\frac{2\delta IR}{2\delta V} = -\frac{\delta IR}{\delta V}. \tag{2.41}$$

The term $\delta I/\delta V$ may be recognized as the g_m of the driver transistor. Thus the gain is

$$A_{diff} = -g_m R. \tag{2.42}$$

This is called the Differential Gain because it resulted from applying a differential signal to the inputs. In practical circuits, ideal constant current sources are hard to find so the Common Mode Gain and Differential Gains vary from the ideal. The Common Mode Rejection Ratio ($CMRR$) is defined as

$$CMRR = \frac{Differential\ Gain}{Common\ Mode\ Gain}. \tag{2.43}$$

LG Display Co., Ltd.
Exhibit 1012
Page 063

The value of the load resistor, R_{load}, is a tradeoff between gain (large R) and bandwidth (low R). Also, the value of the current source, I_{source}, represents a balance between power dissipation (low I, small power dissipation) and bandwidth (high I, low R, high bandwidth). As R_{load} is decreased for a given I_{source}, the minimum voltage at the output decreases ($V_{outmin} = V_{DD} - I_{source}R_{load}$). As R_{load} is increased, V_{outmin} decreases usually until a point at which the current source ceases to act as such or some other bias condition prevents the amplifier from operating as such. The size of the driver transistor affects the gain. The larger the transistor the higher the gain, but the larger are the associated parasitic capacitances.

For instance, in the circuit shown a tail current of $100\mu A$ is chosen. The quiescent conditions required are as follows:

$$V_{left} = V_{right} = 2.5 \text{ volts}$$

$$V_{out1} = V_{out2} = 3.5 \text{ volts}$$

Thus
$$I_{source}R_{load} = V_{DD} - 3.5$$
$$= 1.5$$
$$R_{load} = \frac{1.5}{50\mu A}$$
$$= 30K\Omega$$

Figure 2.28(b) shows the I/O characteristic for the circuit shown in Fig. 2.28(a) for a number of transistor widths. As the transistor width is increased, the gain increases. In addition, as the transistor width is increased,

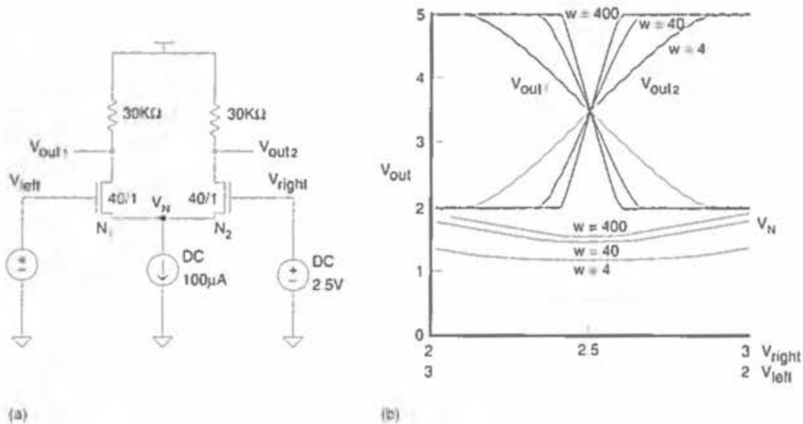


FIGURE 2.28 Basic differential amplifier

the V_N voltage rises as the required V_{gs} to establish the tail current decreases. At the quiescent point the driver transistors are in saturation, and for instance the β for the process is .124mA/V^2 and V_{tn} = .7 volts. Hence,

$$\begin{aligned} g_m &= \beta(V_{gs} - V_t) \\ &= .124 \times 20 \times (2.5 - 1.5 - .7) \\ &= .74mS \text{ (milliSiemens)} \\ A &= g_m R_{load} \\ &= .74 \times 10^{-3} \times 30 \times 10^3 \\ &= 22.3. \end{aligned}$$

From the characteristics in Fig. 2.28(b)

$$A = 22.2,$$

which shows good correspondence.

In Fig. 2.28 we used an ideal current source for the differential pair. An MOS transistor may be used to provide a very good constant current source provided certain operating conditions are met. From the DC operating equations, we know that when a transistor is in the saturation region, the drain current to a first approximation is independent of drain-source voltage. We can improve the characteristics of the MOS constant current source by lengthening the device beyond the minimum dimensions allowed. This reduces the effect of channel-length modulation.

A CMOS differential pair with an nMOS current source and pMOS load resistors is shown in Fig. 2.29. A voltage V_{bias} sets the current in the current source. The constant current source will act as such provided that $V_N > V_{bias} - V_{tn}$. To keep V_{bias} low while providing a reasonable current requires the current source to have a large β.

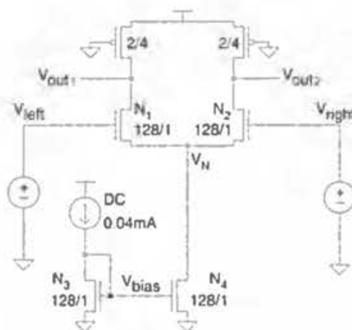


FIGURE 2.29 CMOS differential amplifier

In the circuit, V_{bias} is set by what is termed a current mirror. If a current is forced in N_3, then an identical current will flow in transistor N_4. The reason for this is as follows. With the drain connected to the gate, N_3 is in saturation. Forcing a current I_{s3} in N_3 yields a V_{gs3} of

$$V_{gs3} = \sqrt{\frac{2I_{s3}}{\beta}} + V_t .$$

Now, because N_4 has a $V_{gs} = V_{gs1}$,

$$I_{s4} = \frac{\beta}{2}(V_{gs} - V_t)^2 = I_{s3}.$$

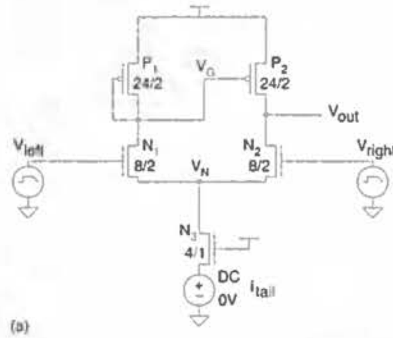
One may cascade current mirrors to provide a variety of current tracking arrangements. If a current multiplication is required, this may be achieved by appropriate ratioing of the current mirror transistors.

Figure 2.30(a) shows a differential amplifier that employs an active current-mirror load structure rather than resistive p-transistors. This structure forms the basis for many RAM sense amplifiers. In this application, the current source is often connected as an unsaturated device. In these circumstances, one has to ensure that the DC conditions are such that the amplifier operates correctly. The active p loads have to be able to source the total current developed by the current source n-transistor. A starting point is to make $\beta_{N_3} = \beta_{P_1} = \beta_{P_2}$. Figure 2.30(b) shows the amplifier characteristic for varying load device sizes. If the p-devices are too small, then when $V_{left} = V_{DD}$, the high value at V_{out} will be lower than possible because P_1 will not be able to source all of the current from N_3. If P_1 and P_2 are made larger with respect to N_3, the low value of the amplifier increases, the gain of the amplifier decreases, and the transition region moves to the left as shown in Fig. 2.30(b). The gain is then determined by the g_m of N_1 and the output conductance of P_2 and N_2. Figure 2.30(c) and Fig. 2.30(d) show the I/O characteristics for the amplifier and the currents that flow in the current source and the two load devices. The small signal gain is given by[24]

$$A = \frac{g_{mn}}{g_o} \tag{2.44}$$

where g_{mn} is the g_m of the driver transistor and g_o is the combined output conductance of the p current load and the n-driver transistor. This is shown in Fig. 2.30(e) for various values of load- and driver-device sizes for a fixed current source. As the length of the devices is increased (r_{ds} increases), the gain of the amplifier increases. Increasing the width of the driver devices

LG Display Co., Ltd.
Exhibit 1012
Page 066



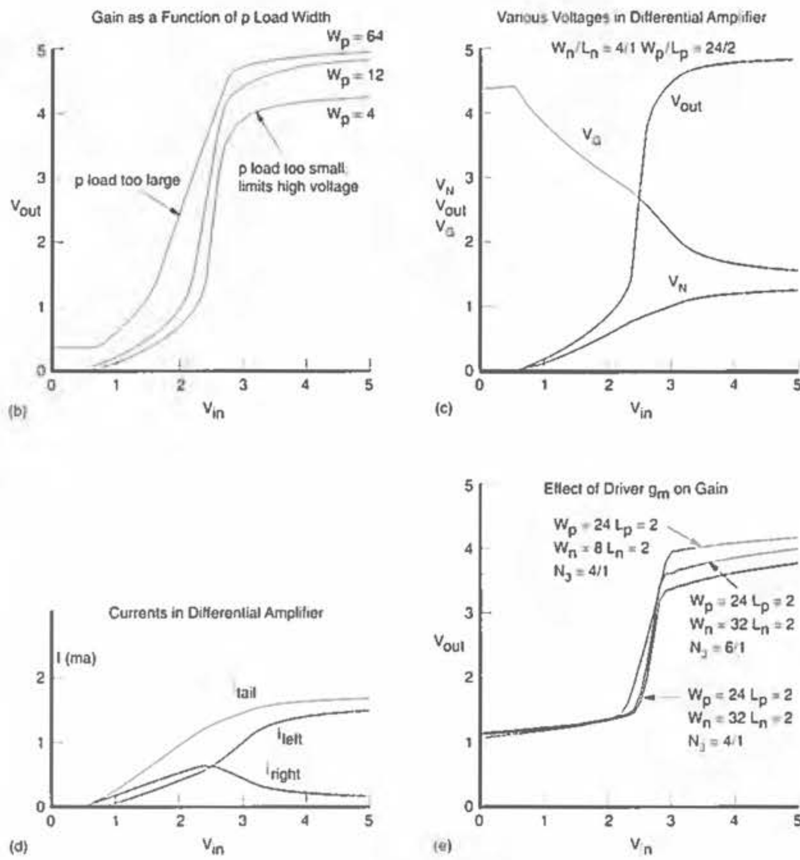


FIGURE 2.30 Active load CMOS differential amplifier

LG Display Co., Ltd.
Exhibit 1012
Page 067

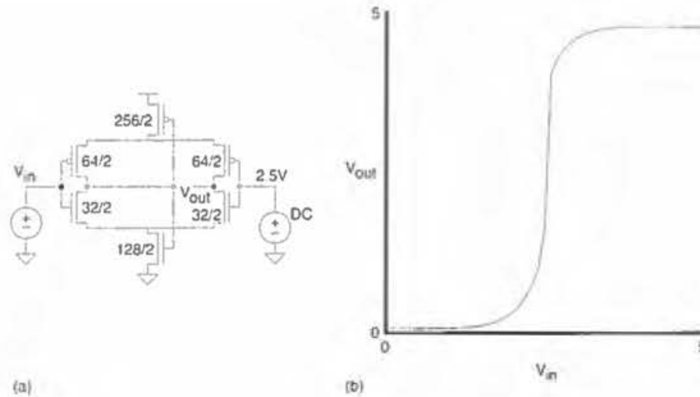


FIGURE 2.31 Self-biased CMOS differential amplifier

does not have as marked an effect on the gain as the $g_m = \beta(V_{gs} - V_t)$. For instance if the β of the driver transistors is quadrupled, then the $(V_{gs} - V_t)$ is halved and the g_m is only doubled.

A further CMOS differential amplifier is shown in Fig. 2.31.[25] It has twice the gain of the amplifier shown in Fig. 2.30 and has the advantage that it is self-biasing. This amplifier is of use in TTL-CMOS input buffers and comparators.

2.6 The Transmission Gate

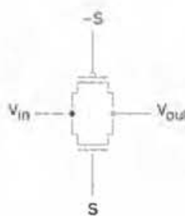


FIGURE 2.32 Transistor connection for CMOS transmission gate

The transistor connection for a complementary switch or transmission gate is reviewed in Fig. 2.32. It consists of an n-channel transistor and a p-channel transistor with separate gate connections and common source and drain connections. The control signal is applied to the gate of the n-device, and its complement is applied to the gate of the p-device. The operation of the transmission gate can be best explained by considering the characteristics of both the n-device and p-device as pass transistors individually. We will address this by treating the charging and discharging of a capacitor via a transmission gate.

nMOS Pass Transistor. Referring to Fig. 2.33(a), the load capacitor C_{load} is initially discharged (i.e., $V_{out} = V_{SS}$). With $S = 0$ (V_{SS}) (i.e., $V_{gs} = 0$ volts), $I_{ds} = 0$, then $V_{out} = V_{SS}$ irrespective of the state of the input V_{in}. When $S = 1$ (V_{DD}), and $V_{in} = 1$, the pass transistor begins to conduct and charges the load capacitor toward V_{DD}, i.e., initially $V_{gs} = V_{DD}$. Since initially V_{in} is at a

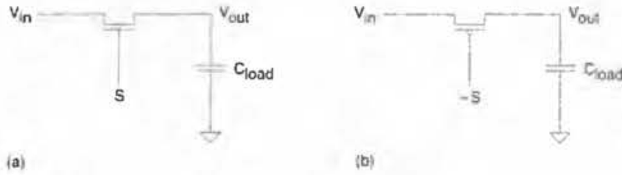


FIGURE 2.33 nMOS and pMOS transistor operation in transmission gate

higher potential than V_{out}, the current flows through the device from left to right. As the output voltage approaches $V_{DD} - V_{tn}$, the n-device begins to turn off. Load capacitor, C_{load}, will remain charged when S is changed back to 0. Therefore the output voltage V_{out} remains at $V_{DD} - V_{tn(V_{dd})}$. $V_{tn(V_{dd})}$ is the n-transistor body affected threshold with the source at V_{DD}. This implies that the transmission of logic one is degraded as it passes through the gate. With $V_{in} = 0$, $S = 1$, and $V_{out} = V_{DD} - V_{tn(V_{dd})}$, the pass transistor begins to conduct and discharge the load capacitor toward V_{SS}, i.e., $V_{gs} = V_{DD}$. Since initially V_{in} is at a lower potential than V_{out} the current flows through the device from right to left. As the output voltage approaches V_{SS}, the n-device current diminishes. Because V_{out} falls to V_{SS}, the transmission of a logic zero is not degraded.

pMOS Pass Transistor. Once again a similar approach can be taken in analyzing the operation of a pMOS pass transistor as shown in Fig. 2.33(b). With $-S = 1$ ($S = 0$), $V_{in} = V_{SS}$, and $V_{out} = V_{SS}$, the load capacitor C_{load} remains uncharged. When $-S = 0$ ($S = 1$), current begins to flow and charges the load capacitor toward V_{DD}. However, when $V_{in} = V_{SS}$ and $V_{out} = V_{DD}$, the load capacitor discharges through the p-device until $V_{out} = V_{tp(V_{ss})}$, at which point the transistor ceases conducting. Thus transmission of a logic zero is somewhat degraded through the p-device.

The resultant behavior of the n-device and p-device are shown in Table 2.4. By combining the two characteristics we can construct a transmission gate that can transmit both a logic one and a logic zero without degradation. As can be deduced from the discussion so far, the operation of the transmission gate requires both the true and the complement version of the control signal.

TABLE 2.4 Transmission Gate Characteristics

DEVICE	TRANSMISSION OF '1'	TRANSMISSION OF '0'
n	poor	good
p	good	poor

LG Display Co., Ltd.
Exhibit 1012
Page 069

The overall behavior can be expressed as:

$$S = 0\,(\sim S = 1):\ \begin{cases} \text{n-device} = \text{off} \\ \text{p-device} = \text{off} \\ V_{in} = V_{SS},\ V_{out} = Z \\ V_{in} = V_{DD},\ V_{out} = Z \end{cases} \tag{2.45}$$

where Z refers to a high impedance state and

$$S = 1\,(\sim S = 0):\ \begin{cases} \text{n-device} = \text{on} \\ \text{p-device} = \text{on} \\ V_{in} = V_{SS},\ V_{out} = V_{SS}. \\ V_{in} = V_{DD},\ V_{out} = V_{DD} \end{cases} \tag{2.46}$$

The transmission gate is a fundamental and ubiquitous component in MOS logic. It finds use as a multiplexing element, a logic structure, a latch element, and an analog switch. The transmission gate acts as a voltage controlled resistor connecting the input and the output.

Figure 2.34(a) shows a typical circuit configuration for a transmission gate in which the output is connected to a capacitor and the input to an inverter. The control input is shown turning the transmission gate on. That is, the gate of the n-channel transmission gate switch is changing from $0 \rightarrow 1$ and the gate of the p-channel is changing from $1 \rightarrow 0$. First consider the case

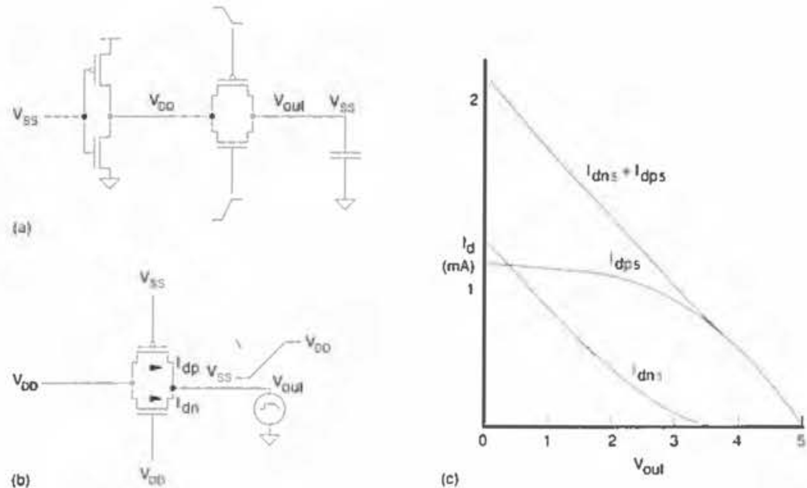


FIGURE 2.34 Transmission gate output characteristic for control input changing

LG Display Co., Ltd.
Exhibit 1012
Page 070

where the control input changes rapidly, the inverter input is low (V_{SS}), the inverter output is high (V_{DD}), and the capacitor on the transmission gate output is discharged (V_{SS}). The currents that flow in this situation may be modeled by the circuit shown in Fig. 2.34(b) in which the input is held at V_{DD} and the output is ramped from V_{SS} to V_{DD}, while the currents in the pass transistors are monitored (in SPICE by using zero-volt voltage sources). In reality, the capacitor discharge would be exponential, but a linear ramp serves to show what happens to the pass transistor currents. As V_{out} rises, the p-transistor current follows a constant V_{gs} of –5 volts (Fig. 2.34c). That is, it starts out in saturation and transitions to the nonsaturated case when $|V_{gsp} - V_{tp}| > |V_{dsp}|$. The n-transistor is always in the nonsaturated region as $V_{dsn} = V_{gsn}$ and $V_{gsn} - V_{tn} < V_{dsn}$. When V_{out} reaches a V_{tn} below V_{DD}, the n-transistor turns off. Thus there are three regions of operation:

Region A. n nonsaturated, p saturated ($V_{out} < V_{tp}$)
Region B. n nonsaturated, p nonsaturated ($V_{tp} < V_{out} < V_{DD} - V_{tn}$)
Region C. n off, p nonsaturated ($V_{DD} - V_{tn} < V_{out}$)

In region A, we can approximate the p-current as a constant current while the n-current varies linearly with V_{out}. Hence the total current is linear with V_{in}. In region B both currents vary linearly with V_{out}. Finally in region C the p-current varies linearly with V_{out}. Thus the transmission gate acts as a resistor, with contributions to its resistance from both n- and p-transistors. This can be seen in Fig. 2.34(c) ($I_{dn5} + I_{dp5}$). Similar simulations may be carried out for $V_{in} = V_{SS}$ and $V_{out} = V_{DD} \rightarrow V_{SS}$.

Another operation mode that the transmission gate encounters in lightly loaded circuits is where the output closely follows the input, such as shown in Fig. 2.35(a). Figure 2.35(b) shows a model of this while Fig. 2.35(c) shows the SPICE circuit used to model this condition including current monitoring voltage sources. Figure 2.35(d) shows the n- and p-pass transistor currents for $V_{out} - V_{in} = -0.1$ volts. It can be seen that again there are three regions of operation:

Region A. n nonsaturated, p off
Region B. n nonsaturated, p nonsaturated
Region C. n off, p unsaturated

The total current decreases in magnitude as V_{in} increases until $V_{in} = V_{tp(body\text{-}effected)}$. Here the p-transistor turns on and in this case slows the decrease of current. When $V_{in} > V_{DD} - V_{tn(body\text{-}effected)}$, the current starts to increase in magnitude as the p current increases. In this simulation the p and n gains were matched. For the region $V_{tp} < V_{in} < V_{DD} - V_{tn}$, the transmission gate will have a roughly constant resistance. The effect of having only one polarity transistor in the transmission gate is also seen. If only an n-transistor is used, the output will rise to an n threshold below V_{DD} as current stops flowing at this point. Similarly, with a single p-transistor, the output would fall to a p threshold above V_{SS}, as current stops flowing in the p-transistor at

LG Display Co., Ltd.
Exhibit 1012
Page 071

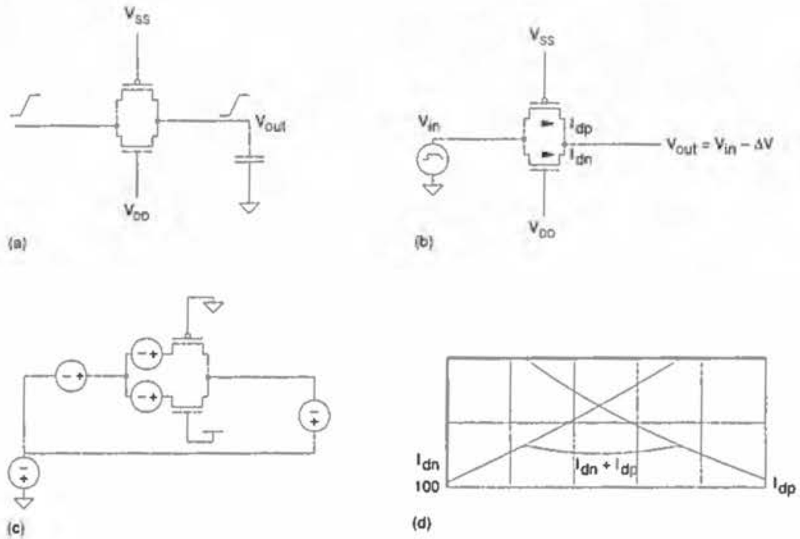


FIGURE 2.35 Transmission gate output characteristic for switched input changing

this point. Note also that as either the p or n current approaches zero, the speed of any circuit would be prejudiced. If the surrounding circuitry can deal with these imperfect high and low values, then single polarity transmission gates may be used. Figure 2.36 shows a plot of the transmission gate "on" resistance for the test circuit shown in Fig. 2.35(c).

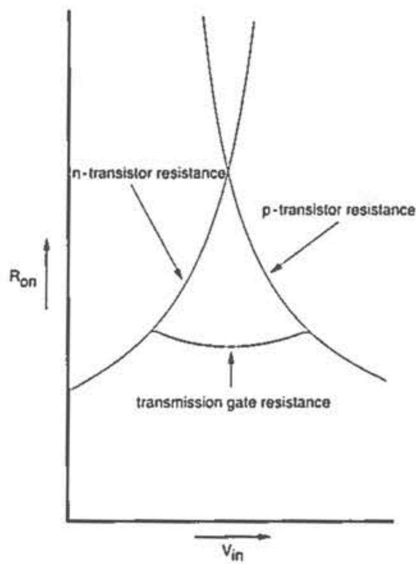


FIGURE 2.36 Resistance of a transmission gate for conditions in Figure 2.35

LG Display Co., Ltd.
Exhibit 1012
Page 072

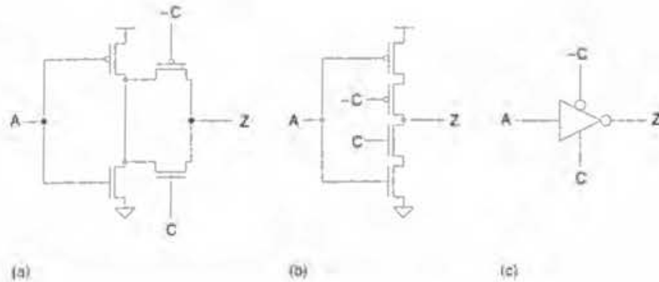


FIGURE 2.37 Tristate inverter

2.7 The Tristate Inverter

By cascading a transmission gate with an inverter the tristate inverter shown in Fig. 2.37(a) is constructed. When $C = 0$ and $-C = 1$, the output of the inverter is in a tristate condition (the Z output is not driven by the A input). When $C = 1$ and $-C = 0$, the output Z is equal to the complement of A. The connection between the n- and p-driver transistors may be omitted (Fig. 2.37b) and the operation remains substantially the same (except for a small speed difference). Figure 2.37(c) shows the schematic icon that represents the tristate inverter. For the same size n- and p-devices, this inverter is approximately half the speed of the inverter shown in Fig. 2.11. This inverter will be discussed in more detail in Chapter 5, because it forms the basis for various types of clocked logic, latches, bus drivers, multiplexers, and I/O structures.

2.8 Bipolar Devices

Thus far we have treated the MOS transistor in isolation as the device of interest. However, there are other semiconductor devices that are fabricated either parasitically or deliberately in a CMOS process. In particular, the junction diode and the bipolar transistor will be examined. The former is of use primarily in digital circuits as a protection device in I/O structures. The latter may be constructed to improve the speed of CMOS in BiCMOS processes. Of concern to all CMOS designers, however, are the parasitic bipolar transistors constructed as a by-product of building the basic nMOS/ pMOS structures in CMOS. These can lead to a circuit debilitating condition known as latchup. This will be covered in detail in Chapter 3.

2.8.1 Diodes

The diode is the most basic of semiconductor devices and is created when a metal and a semiconductor or two semiconductors form a junction When two

diffusions of opposite polarity form a junction, a junction diode is formed. When a metal and semiconductor merge either an ohmic contact is made or a Schottky diode is created. In most CMOS processes only ohmic contacts are formed where metal contacts diffusions.

For instance, in an nMOS (or pMOS) transistor, the source and drain terminals form np (or pn) junction diodes to the substrate (or well). The schematic symbol for a junction diode is shown in Fig. 2.38(a). The two terminals are designated the anode and cathode. The *VI* characteristics of a diode are shown in Fig. 2.38(b). The current in a diode is given by[26]

$$I = A_d I_s \left(e^{\frac{qV}{kmt}} - 1 \right) \tag{2.47}$$

where

A_d = area of the diode
I_s = the saturation current/unit area
q = electronic charge
k = Boltzmann's constant
t = Temperature
m = a constant between 1 and 2 to account for various nonlinearities
(m ~ 2 for pn junction diodes and m ~ 1.2 for Schottky diodes).

There are a number of characteristics of interest. When a positive voltage is applied to the cathode with respect to the anode, electrons are attracted to the supply and holes are repelled, leading to a "reverse-biased" condition

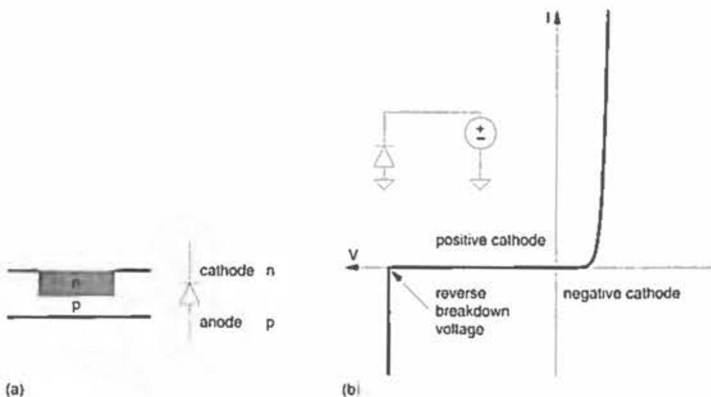


FIGURE 2.38 Diode *VI* characteristics

in which a very small reverse current flows. This results in a depletion region similar to that in the MOS transistor when it is in the depletion regime before inversion. In the above equation the exponential term is reduced in importance and the current is approximated by ($A_d = 1$)

$$I_{reverse} = -I_s \,(\sim 1 \times 10^{-15} A) \tag{2.48}$$

This condition applies until the voltage exceeds the reverse breakdown voltage of the junction, at which point the current increases rapidly due to avalanche multiplication. This occurs when electrons accelerated by the high field across the junction impact silicon atoms, thereby producing electron-hole pairs. When a negative voltage is applied to the cathode, the diode becomes forward biased. The current is approximated by ($A_d = 1$)

$$I_{forward} = I_s e^{\frac{qV}{kmt}} \tag{2.49}$$

As Fig. 2.38(b) shows, the current rapidly increases when the cathode-anode voltage is less than −0.6 volts. The x axis is reflected.

2.8.2 Bipolar Transistors

By building an NPN diffusion sandwich, as shown in Fig. 2.39(a), an NPN bipolar transistor may be constructed. Similarly a PNP transistor may be constructed by sandwiching an n diffusion between two p diffusions. The terminals of a bipolar transistor are called the collector, base, and emitter. The behavior of a transistor may be modeled (and is in the SPICE simulation program) by the structure shown in Fig. 2.39(b) for an NPN transistor. If V_{BE}, the base-emitter voltage, is set at around .7 volts and V_{CE} the collector-to-emitter voltage, is positive, the base emitter diode is forward biased and

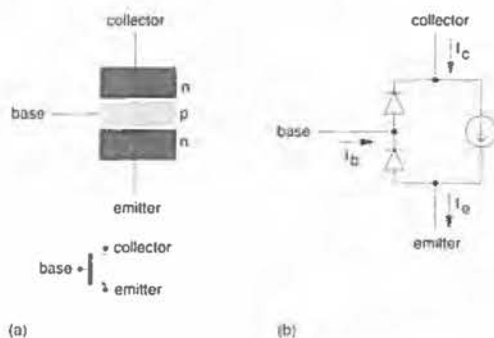


FIGURE 2.39 Structure and model of an NPN bipolar transistor

LG Display Co., Ltd.
Exhibit 1012
Page 075

the collector base diode is reverse biased. By using the Ebers-Moll model,[27] the collector current may be calculated as

$$I_C = I_s\left(e^{\frac{qV_{BE}}{mkt}} - 1\right)\left(1 + \frac{V_{CE}}{V_A}\right). \tag{2.50}$$

While the emitter current is given by

$$I_E = I_C\left(1 + \frac{1}{\beta\left(1 + \frac{V_{CE}}{V_A}\right)}\right) \tag{2.51}$$

where $kT/q = .026$ (at 300°K)

V_{CE} = the collector-emitter voltage
V_{BE} = the base-emitter voltage
m = a constant between 1 and 2
V_A = the Early voltage (an approximation to allow for nonideal phenomena that result in finite output conductance)
β = forward current gain
I_S = the junction saturation current.

The forward current gain, β, (not to be confused with MOS β's) typically ranges from 20–500.

The VI characteristics of a typical NPN transistor are shown in Fig. 2.40.

The basic design equations for use with digital bipolar circuits are described in association with the inverter shown in Fig. 2.41. Here, the collector of an NPN transistor is connected to a positive supply via resistor R_c. The base is connected via resistor R_b to an input voltage V_{in}. The base current I_b is given by

$$I_b = \frac{V_{in} - V_{be}}{R_b} \tag{2.52}$$

where V_{be} = the base emitter voltage (~0.7 volts)
and V_{in} = the input voltage.

The collector current is given by

$$I_c = \beta I_b$$

LG Display Co., Ltd.
Exhibit 1012
Page 076

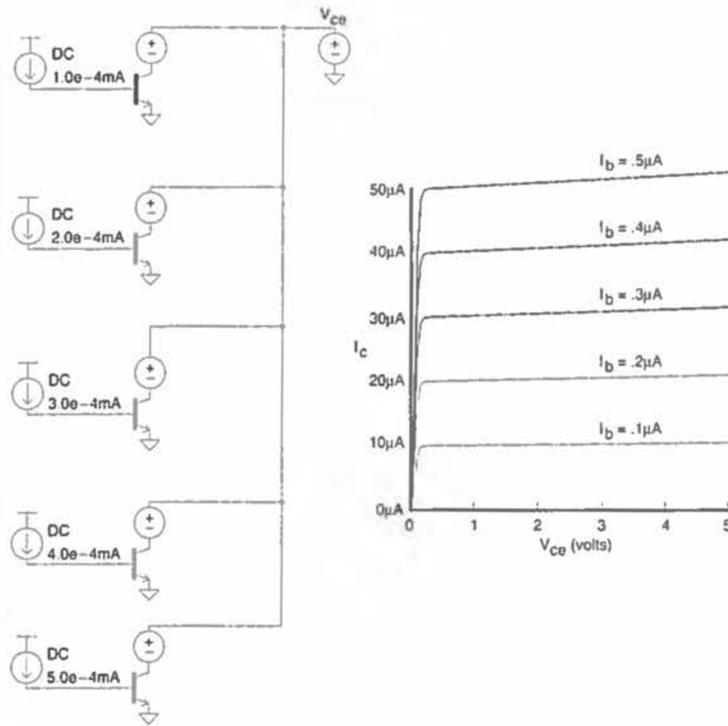


FIGURE 2.40 NPN transistor *VI* characteristics

and hence the collector voltage is given by

$$V_{out} = V_{DD} - I_c R_c$$

$$V_{out} = V_{DD} - \beta \frac{V_{in} - V_{be}}{R_b} R_c.$$

The gain, A, is given by

$$\frac{dV_{out}}{dV_{in}} = \frac{\beta R_c}{R_b}. \qquad (2.53)$$

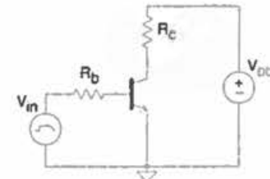


FIGURE 2.41 Inverter using an NPN transistor

An n-well CMOS process inherently has a PNP transistor that is created between the substrate (collector), well (base), and source/drain diffusions (emitter). This PNP transistor is not that useful except for application as a current reference. This transistor is a vertical PNP because the transistor is formed by the vertical stacking of junctions.

Extra processing steps must be added to CMOS processes to build more useful NPN transistors. These steps result in what is termed a BiCMOS process (for Bipolar and CMOS). Similar to the case with p- and n-channel transistors in CMOS, NPN bipolar transistors have much higher gain and better high-frequency response than PNP transistors. Thus BiCMOS processes concentrate on adding a high-performance NPN transistor.

2.8.3 BiCMOS Inverters

The availability of an NPN transistor can markedly improve the output drive capability of a conventional CMOS inverter due to the high current gain of the NPN transistor.[28,29] Figure 2.42 shows one version of a BiCMOS inverter. When the input is low, P_1 is turned on and supplies base current to NPN_1 and sets the base voltage to V_{DD}. N_3 is turned on and clamps the base of NPN_2 to V_{SS}. Thus NPN_2 is off and the output rises to a V_{be} below V_{DD}. When the input is high, the base of NPN_1 is clamped to V_{SS} by N_1 and N_2 supplies the base current for NPN_2. The output falls to a small voltage above V_{SS}. This voltage is called V_{CEsat}, the collector emitter voltage with the transistor in saturation. This is due to the finite "on resistance" of the transistor and may be reduced by increasing I_b. It is normally in the range of $0.1 \rightarrow 0.3$ volts. Thus this inverter has an output swing between $V_{DD} - V_{be}$ and $V_{SS} + V_{CEsat}$. The V_{be} drop causes DC dissipation in any following CMOS gates, a problem which is not improved as the supply voltage is reduced.

A second BiCMOS inverter is shown in Fig. 2.43. Transistors P_2 and N_2 are used as resistors to bias the NPN transistors. When the input is low, P_1 feeds base current to NPN_1 and P_2 serves to pull the output high. The value of P_2 is a compromise to achieve high speed pull-up without bypass of the base current to NPN_1. When the input is high, N_2 feeds the base of NPN_2

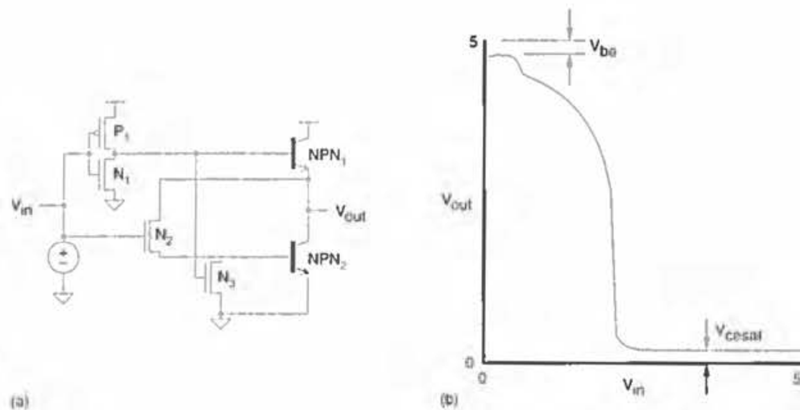


FIGURE 2.42 Basic BiCMOS inverter

LG Display Co., Ltd.
Exhibit 1012
Page 078

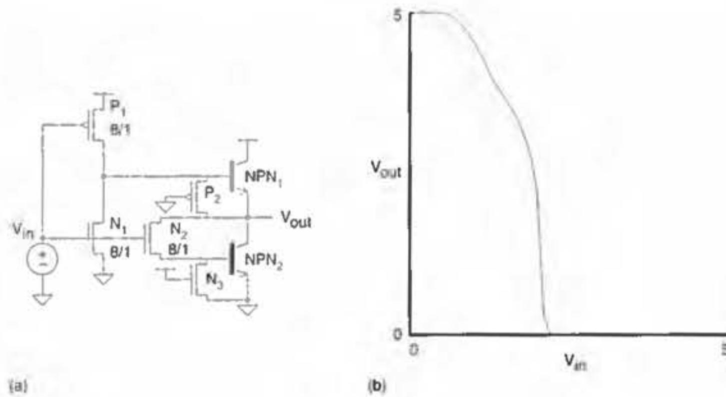


FIGURE 2.43 BiCMOS inverter using MOS transistors as resistors

while N_3 serves to pull the output to V_{SS}. The primary advantage of this implementation is that the output falls to V_{SS} and rises to V_{DD}.

A third BiCMOS inverter is shown in Fig. 2.44. In this inverter a feedback inverter is added to control the "resistor" transistors. When the input is low and the output is high, the feedback inverter places a zero on P_2 and N_2, thereby pulling the output high. When the input is high, the output becomes low and the feedback inverter places a high on N_2 and P_2, which pulls the output low.

A final BiCMOS inverter is shown in Fig. 2.45[30] which only uses a pull-up NPN transistor. When the input is low, P_1, P_2, NPN_1, and the feedback inverter combine to pull the output high. When the input is high, N_3 pulls the output low. This inverter is of particular use for 3.3 Volt supply circuits. The technique of using an nMOS transistor as the sole pull-down element can be used for the BiCMOS inverters shown in Figs. 2.42 and 2.43. Section 5.4.2 has an extended reference list of research on BiCMOS.

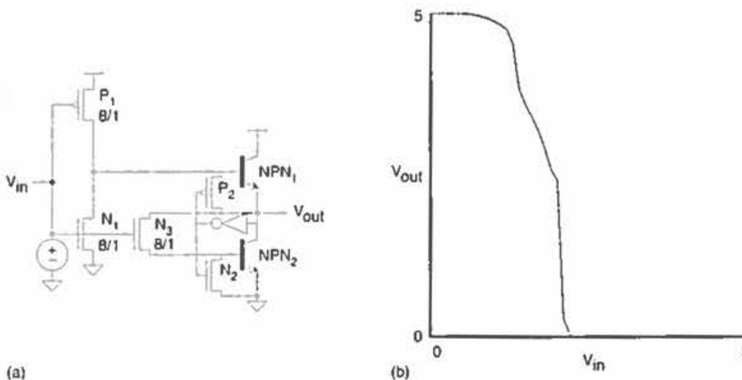


FIGURE 2.44 BiCMOS inverter with feedback inverter.

LG Display Co., Ltd.
Exhibit 1012
Page 079

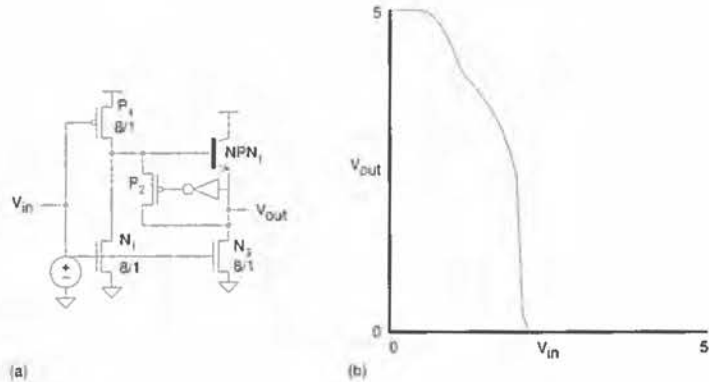


FIGURE 2.45 BiCMOS inverter with nMOS pulldown

2.9 Summary

This chapter has examined the DC characteristics of MOS transistors, diodes, bipolar transistors and CMOS inverters. In addition the operation of the CMOS transmission gate was reviewed. Finally, the circuit configurations of some BiCMOS inverters were surveyed. The circuits treated in this chapter are the basis for the majority of logic and memory circuits used in CMOS digital system design. Ensuring their correct DC operation is the first step in constructing a correctly functioning circuit. The second step, satisfying temporal (or timing) constraints requires one to be able to estimate the speed of a circuit. This will be treated in Chapter 4.

2.10 Exercises

1. Calculate the noise margin for the BiCMOS inverter shown in Fig. 2.42
   a. BiCMOS → CMOS ($\beta_n = \beta_p$) and
   b. BiCMOS → pseudo-nMOS ($\beta_n/\beta_p = 4$)
   ($V_{be} = .7$ volts)
2. Calculate the noise margin for a CMOS inverter operating at 3.3V with $V_{tn} = 0.7$V, $V_{tp} = -0.7$V, $\beta_p = \beta_n$. What would you do to the transistor characteristics to improve the noise margin?
3. Derive the V_{OH} and V_{OL} for the inverter shown in Fig. 2.23.
4. Derive the VI equations that predict the V_{OL} for the inverter shown in Fig. 2.24.

LG Display Co., Ltd.
Exhibit 1012
Page 080

5. Design an input buffer that may be used to interface with a TTL driver ($V_{DD} = 5V$, $V_{OL} = 0.8V$, $V_{OH} = 2.0V$). Show full derivations of DC conditions.

6. Design a buffer that interfaces internal 3.3V logic to CMOS I/O logic operating at 5V.

7. Does the body effect of a process limit the number of transistors that can be placed in series in a CMOS gate at low frequencies?

8. Sometimes the substrate is connected to a voltage called the substrate bias to alter the threshold of the n-transistors. If the threshold of an n-transistor is to be raised, explain to what polarity the voltage substrate would be connected. Draw a circuit diagram showing V_{DD}, V_{SS}, and substrate supplies connected to an inverter.

9. Under what voltage conditions is a p-transistor with a source connected to $V_{DD} = 5V$ ($V_{tp} = -0.7$V) a good current source?

10. Using switched current mirrors, show how you would construct a current sourced digital to analog (D/A) converter with eight distinct current level outputs.

11. Calculate the threshold implant necessary to increase the threshold voltage to 0.6V for the example in Section 2.1.3.1.

12. For the values given in Section 2.2.2.7 ($L_{eff} = 0.6$ µm $V_{tn} = 0.6V$) calculate the worst case substrate current as a percentage of I_{ds} for a 2-input NAND gate operated at 5V and 3.3V.

2.11 Appendix—SPICE Level 3 Model

The following is a summary of the Level 3 MOS model parameters used in the HSPICE program from Meta-Software, Inc. These parameters are consistent with most SPICE implementations that are widely available. The following is reproduced in large part from the HSPICE User's Manual with the kind permission of Meta-Software. This model uses empirical values determined from processed test devices as a basis for the model equations. The basic model parameters—*LEVEL*, *COX*, *KAPPA*, *KP*, *TOX*, and *VMAX*—are reviewed in Table 2.5 in terms of a 0.5 – 1µ n-well process. (*Note:* These values should be used only as a guide—check with your CMOS manufacturer for accurate model parameters.)

The Level 3 model parameters for modeling the effective width and length are given in Table 2.6.

The Level 3 model also uses some parameters that vary the threshold voltage. These are as given in Table 2.7.

The parameters related to mobility are given in Table 2.8.

LG Display Co., Ltd.
Exhibit 1012
Page 081

TABLE 2.5 Basic Model Parameters

NAME	UNITS	TYPICAL 1μm CMOS VALUE	DESCRIPTION
LEVEL		3.0	DC model selector.
COX	F/m^2	35 – 17E–4 (100 – 200 Å)	The oxide capacitance per unit gate area. If *COX* is not specified, then it will be calculated from *TOX*.
KAPPA	1/V	0.01 – .02	Saturation field factor, used in channel-length modulation equation.
KP	amp/V^2	2.0E-5	The intrinsic transconductance parameter. If not specified, then *KP* is calculated as *KP* = *UO.COX*.
TOX	*m*	1 – 2E–8	Gate oxide thickness.
VMAX	*m/s*	1.5 – 2E5	Maximum drift velocity of carriers; 0.0 indicates an infinite value.

TABLE 2.6 Effective Width and Length Parameters

NAME	UNITS	TYPICAL 1μm CMOS VALUE	DESCRIPTION
DEL	*m*	0.0	Channel-length reduction on each side.
LD	*m*	.01 – .1E-6	Lateral diffusion into channel from source and drain diffusion. If *LD* is unspecified, but *XJ* is specified, then *LD* = 0.75 *XJ*.
LREF	*m*	0.0	Channel-length reference.
LMLT	*m*	1.0	Length shrink factor.
WD	*m*	.05 – .1E-6	Lateral diffusion into channel from bulk along width.
WMLT		0.0	Diffusion layer and width shrink factor.
WREF	*m*	1.0	Channel-width reference.
XJ	*m*	.1 – .7E-6	Metallurgical junction depth.
XL	*m*	0.0	Accounts for masking and etching effects.
XW	*m*	0.0	Accounts for masking and etching effects.

TABLE 2.7 Threshold Voltage Parameters

NAME	UNITS	TYPICAL 1μm CMOS VALUE	DESCRIPTION
DELTA		1.0 – 1.5	Narrow width factor for determining threshold.
ETA		.05 – .15	Static feedback factor for adjusting threshold.
GAMMA	$V^{0.5}$	.2 – .6	Body effect factor. If *GAMMA* is not specified it is calculated from $GAMMA = \frac{\sqrt{2q\varepsilon_{Si}NSUB}}{COX}$
ND	$1/V$	1.0	Drain subthreshold factor.
N0		1.0	Gate subthreshold factor.
LND	$\mu m/V$	0.0	*ND* length sensitivity.
LN0	μm	0.0	*N0* length sensitivity.
NFS	$cm^{-2}V^{-1}$	7.5*E*11	Fast surface state density.
NSUB	cm^{-3}	2*E*16	Bulk surface doping. If not specified, calculated from *GAMMA*.
PHI	V	.74	Surface inversion potential. If not specified it is calculated from NSUB as $PHI = 2\frac{kT}{q}ln(\frac{NSUB}{Ni})$.
VTO	V	0.5 → 0.7 (N) -0.5 → -0.7 (P)	Zero-bias threshold voltage. If not specified it will be calculated from other parameters.
WIC		0.0	Subthreshold model selector.
WND	$\mu m/V$	0.0	*ND* width sensitivity.
WN0	μm	0.0	*N0* width sensitivity.

TABLE 2.8 Mobility Parameters

NAME	UNITS	TYPICAL 1μm CMOS VALUE	DESCRIPTION
THETA	$1/V$	0.05 – 0.15	Mobility degradation factor.
UO	$cm^2/V.s$	600 (N) 250 (P)	Low field bulk mobility.

The drain current is calculated as follows in the Level 3 model.

Cutoff Region $V_{gs} \le V_t$

$$i_{ds} = 0$$

On region, $V_{gs} > V_t$

$$i_{ds} = \beta \left(V_{gs} - V_t - \frac{(1+fb)}{2} V_{de}\right) V_{de}$$

where

$$\beta = KP \frac{w_{eff}}{l_{eff}}$$

$$= u_{eff} COX \frac{w_{eff}}{l_{eff}}$$

$$V_{de} = min(V_{ds}, V_{dsat})$$

and

$$fb = fn + \frac{GAMMA\ fs}{4\sqrt{PHI + V_{sb}}}.$$

(The 4 in this equation should be 2 but HSPICE emulates the original SPICE program and uses 2.)

The narrow width effect is included through the fn parameter,

$$fn = \frac{DELTA}{w_{eff}}.$$

The term fs expresses the effect of the short channel and is determined as

$$fs = 1 - \frac{XJ_{scaled}}{l_{eff}} \left\{ \frac{LD_{scaled} + wc}{XJ_{scaled}} \sqrt{1 - \left(\frac{wp}{XJ_{scaled} + wp}\right)^2} - \frac{LD_{scaled}}{XJ_{scaled}} \right\}$$

$$wp = xd\sqrt{(PHI + V_{sb})}$$

$$xd = \sqrt{\frac{2\varepsilon_{Si}}{qNSUB}}$$

$wc =$

$$XJ_{scaled}\left[0.0831353 + 0.8013929\left(\frac{wp}{XJ_{scaled}}\right) - 0.0111077\left(\frac{wp}{XJ_{scaled}}\right)^2\right]$$

LG Display Co., Ltd.
Exhibit 1012
Page 084

$$XJ_{scaled} = XJ \cdot SCALM$$
$$LD_{scaled} = LD \cdot SCALM.$$

SCALM is a global scaling factor applied to all MOS models in a given HSPICE run. The effective channel length and width in the Level 3 model is determined as follows:

$$l_{eff} = L_{scaled} LMLT + XL_{scaled} - 2(LD_{scaled} + DEL_{scaled})$$

where

$$L_{scaled} = L \cdot SCALM$$
$$XL_{scaled} = XL \cdot SCALM$$
$$DEL_{scaled} = DEL \cdot SCALM$$

LMLT is a scaling factor applied on a model by model basis.

$$w_{eff} = M(W_{scaled} WMLT + XW_{scaled} - 2WD_{scaled})$$

where

$$W_{scaled} = W \cdot SCALM$$
$$XW_{scaled} = XW \cdot SCALM$$
$$WD_{scaled} = WD \cdot SCALM$$

M is a parameter that allows for multiple parallel devices. The default value is 1.

$$LREF_{scaled} = LREF_{scaled} LMLT + XL_{scaled} - 2(LD_{scaled} + DEL_{scaled})$$
$$WREF_{scaled} = M(WREF_{scaled} MLT + XW_{scaled} - 2WD_{scaled})$$

Similar to *LMLT*, *WMLT* is a model scaling factor.
The threshold voltage is calculated as follows:

$$V_{th} = V_{bi} - \frac{8.14 \times 10^{-22}}{COXl_{eff}^{3}} V_{ds} + GAMMAfs\sqrt{PHI + V_{sb}} + fn(PHI + V_{sb})$$

with

$$V_{bi} = V_{fb} + PHI$$

or

$$V_{bi} = VTO - GAMMA\sqrt{PHI}$$

The saturation voltage V_{dsat} is calculated as

$$V_{dsat} = \frac{V_{gs} - V_{th}}{1 + fb}$$

$$V_{dsat} = V_{sat} + V_c - \sqrt{V_{sat}^2 + V_c^2}$$

where

$$V_c = \frac{VMAXl_{eff}}{u_s}$$

If the model parameter *VMAX* is not specified, then

$$V_{dsat} = V_{sat}$$

The parameter μ_s is the normal field mobility. It is calculated as

$$u_s = \frac{UO}{1 + THETA\,(V_{gs} - V_{th})} \qquad V_{gs} > V_{th}$$

The degradation of mobility due to the lateral field and the carrier velocity saturation is determined if *VMAX* is specified.

$$u_{eff} = \frac{u_s}{1 + \frac{V_{de}}{V_c}}$$

The effects of channel length modulation are calculated as follows:

$$\Delta l = xd\sqrt{KAPPA.\,(V_{ds} - V_{dsat})} \qquad VMAX = 0$$

$$\Delta l = \frac{ep.xd^2}{2} + \sqrt{\left(\frac{ep.xd^2}{2}\right)^2 + KAPPAxd^2\,(V_{ds} - V_{dsat})}$$

where *ep* is the lateral electric field at the pinch off point. Its value is approximated by:

$$ep = \frac{V_c\,(V_c + V_{dsat})}{l_{eff}V_{dsat}}$$

LG Display Co., Ltd.
Exhibit 1012
Page 086

The current in saturation is computed as

$$I_{ds} = \frac{I_{ds}}{1 - \frac{\Delta l}{l_{eff}}}.$$

In order to prevent the denominator from going to zero, HSPICE limits the Δl as follows:

$$\text{if } \Delta l > \frac{l_{eff}}{2}$$

$$\text{else } \Delta l = l_{eff} - \frac{\left(\frac{l_{eff}}{2}\right)^2}{\Delta l}$$

In the subthreshold region the current is characterized by the model parameter for fast surface states, *NFS*. The modified threshold voltage, V_{on}, is determined as follows:

$$V_{on} = V_{th} + fast \qquad NFS > 0$$

where

$$fast = \frac{kt}{q}\left[1 + \frac{qNFS}{COX} + \frac{GAMMAfs\sqrt{(PHI + V_{sb})} + fn(PHI + V_{sb})}{2(PHI + V_{sb})}\right]$$

The current I_{ds} is given by

$$I_{ds} = I_{ds}(V_{on}, V_{de}, V_{sb})\, e^{\frac{V_{gs} - V_{on}}{fast}} \qquad V_{gs} < V_{on}$$

$$I_{ds} = I_{ds}(V_{gs}, V_{de}, V_{sb}) \qquad V_{gs} \geq V_{on}$$

The modified threshold voltage is not used in strong inversion.

2.12 References

1. L. Vadasz and A. S. Grove, "Temperature of MOS Transistor Characteristics Below Saturation," *IEEE Trans. on Electron Devices*, vol. ED-13, no. 13, 1966, pp. 190–192.
2. J. Mavor, M. A., Jack Denyer and P. B. Denyer, *Introduction to MOS LSI Design*, Reading, Mass.: Addison-Wesley, 1983, pp. 18–61 (footnote).
3. John Y. Chen, *CMOS Devices and Technology for VLSI*, Englewood Cliffs, N.J.: Prentice Hall, 1990, pp. 5–37.
4. John Y. Chen, *op. cit.*, p. 211.
5. *HSPICE User's Manual*, Campbell, Calif.: Meta-Software, 1990, pp. 7–34.
6. W. Shockley, "A unipolar field effect transistor," *Proc. IRE.*, vol. 40, Nov. 1952, pp. 1365–1376.
7. R. S. Cobbold, *Theory and Application of Field Transistors*, New York: Wiley Interscience, 1970, pp. 239–267.
8. C. T. Sah, "Characteristics of the Metal-Oxide-Semiconductor Transistor," *IEEE Trans. Ed*, ED-11, Jul. 1964, pp. 324–345.
9. L. W. Nagel, "SPICE2: A Computer Program to Simulate Semiconductor Circuits," Memo ERL-M520, Berkeley, Calif.: University of California, May 9, 1975.
10. Eric A. Vittoz, "MicroPower Techniques," in *Design of VLSI Circuits for Telecommunications*, edited by Y. Tsividis and P. Antognetti, Englewood Cliffs, N.J.: Prentice-Hall, 1985.
11. Paul R. Gray and Robert G. Meyer, *Analysis and Design of Analog Integrated Circuits, Second Edition*, New York: Wiley and Sons, 1984.
12. M. Lenzlinger and E. H. Snow, "Fowler-Nordheim tunneling into thermally grown SiO_2," *Journal of Applied Physics*, vol. 40, 1969, pp. 278–281.
13. John Y. Chen, *op. cit.*, pp. 174–232.
14. S. M. Sze, *Physics of Semiconductor Devices, Second Edition*, New York: Wiley and Sons, 1981, pp. 496–504.
15. John Y. Chen, *op. cit.*, pp. 187–199.
16. Chenming Hu, "IC Reliability Simulation," *IEEE JSSC*, vol. 27, no. 3, Mar. 1992, pp. 241–246.
17. Wen-Jay Hsu, Bing J. Sheu, Sudhir M. Gowda, and Chang-Gyu Hwang, "Advanced Integrated-Circuit Reliability Simulation Including Dynamic Stress Effects," *IEEE JSSC*, vol. 27, no. 3, Mar. 1992, pp. 247–257.
18. Takayusu Sakurai, Kazutaka Nogami, Masakazu Kakumu, and Tetsuya Iizuka, "Hot-Carrier Generation in Submicrometer VLSI Environment," *IEEE JSSC*, vol. SC-21, no. 1, Feb. 1986, pp. 187–192.
19. HSPICE User's Manual, *op. cit.*, pp. 7–34.
20. W. N. Carr and J. P. Mize, *MOS/LSI Design and Application*, New York: McGraw-Hill, 1972.
21. R. S. C. Cobbold, "Temperature Effects on MOS Transistors," *Electronic Letters*, vol. 2, no. 6, June 1966, pp. 190–192.
22. N. Weste and K. Eshraghian, *Principles of CMOS VLSI Design*, Reading, Mass.: Addison-Wesley, 1984, Edition 1, Appendix A.

LG Display Co., Ltd.
Exhibit 1012
Page 088

23. Don Trotter, "CMOS Course Notes," Mississippi State, Miss.: Electrical and Computer Engineering Dept., Mississippi State University, 1991.
24. Adel S. Sedra and Kenneth C. Smith, *Microelectronic Circuits, Second Edition*, New York: Holt, Rinehart and Winston, 1987.
25. Mel Bazes, "Two novel fully complementary self-biased CMOS differential amplifiers," *IEEE JSSC*, vol. 26, no. 2, Feb. 1991, pp. 165–168.
26. Douglas J. Hamilton and William G. Howard, *Basic Integrated Circuit Engineering*, New York: McGraw-Hill, 1975.
27. Douglas J. Hamilton and William G. Howard, *op. cit.*, pp. 212–241.
28. Hyun J. Shin, "Performance Comparison of Driver Configurations and Full Swing Techniques for BiCMOS Logic Circuits," *IEEE JSSC*, vol. 25, no. 3, June 1990, pp. 863–865.
29. Larry Wissel and Elliot L. Gould, "Optimal Usage of CMOS within a BiCMOS Technology," *IEEE JSSC*, vol. 27, no. 3, Mar. 1992, pp. 300–306.
30. Hiroyuki Hara, Takayasu Sakurai, Makato Noda, Tetsu Nagamatsu, Katsuhiro Seta, Hiroshi Momose, Youichirou Niitsu, Hiroyuki Miyakawa and Yoshinori Watanabe, "0.5μm 2M-Transistor BiPNMOS Channelless Gate Array," *IEEE Journal of Solid State Circuits*, Vol. 26, No. 11, Nov. 1991, pp. 1615–1620.

LG Display Co., Ltd.
Exhibit 1012
Page 089

CMOS PROCESSING TECHNOLOGY

3

The purpose of this chapter is to introduce the CMOS designer to the technology that is responsible for the semiconductor devices that might be designed. This is of importance in understanding the potential and limitations of a given technology. It also gives some background for the geometric design rules that are the interface medium between designer and fabricator.

The basics of semiconductor manufacturing are first introduced. Following this, a basic n-well CMOS process is described showing the process steps and how they relate to the design description passed from the designer to the fabrication engineer. Following this, a number of enhancements to the basic CMOS technology are described. Many of these are now required by mainstream CMOS logic and memory designers. The next section introduces the reader to layout design rules that prescribe how to manufacture the CMOS chip. The nature of CMOS latchup and the solutions to this problem are then covered. Finally, some CAD issues as they relate to process technology are covered. An appendix, Section 3.9, outlines the actual steps used in a CMOS process for those who want to get down to that level of detail.

3.1 Silicon Semiconductor Technology: An Overview

Silicon in its pure or *intrinsic* state is a semiconductor, having a bulk electrical resistance somewhere between that of a conductor and an insulator. The conductivity of silicon can be varied over several orders of magnitude by

LG Display Co., Ltd.
Exhibit 1012
Page 090

introducing *impurity* atoms into the silicon crystal lattice. These *dopants* may either supply free electrons or holes. Impurity elements that use electrons are referred to as *acceptors* since they accept some of the electrons already in the silicon, leaving vacancies or holes. Similarly, *donor* elements provide electrons. Silicon that contains a majority of donors is known as *n-type* and that which contains a majority of acceptors is known as *p-type*. When n-type and p-type materials are brought together, the region where the silicon changes from n-type to p-type is called a *junction*. By arranging junctions in certain physical structures and combining these with other physical structures, various semiconductor devices may be constructed. Over the years, silicon semiconductor processing has evolved sophisticated techniques for building these junctions and other structures having special properties.

3.1.1 Wafer Processing

The basic raw material used in modern semiconductor plants is a *wafer* or disk of silicon, which varies from 75 mm to 230 mm in diameter and is less than 1 mm thick. Wafers are cut from ingots of single-crystal silicon that have been pulled from a crucible melt of pure molten polycrystalline silicon. This is known as the 'Czochralski,' method (Fig. 3.1) and is currently the most common method for producing single-crystal material. Controlled amounts of impurities are added to the melt to provide the crystal with the

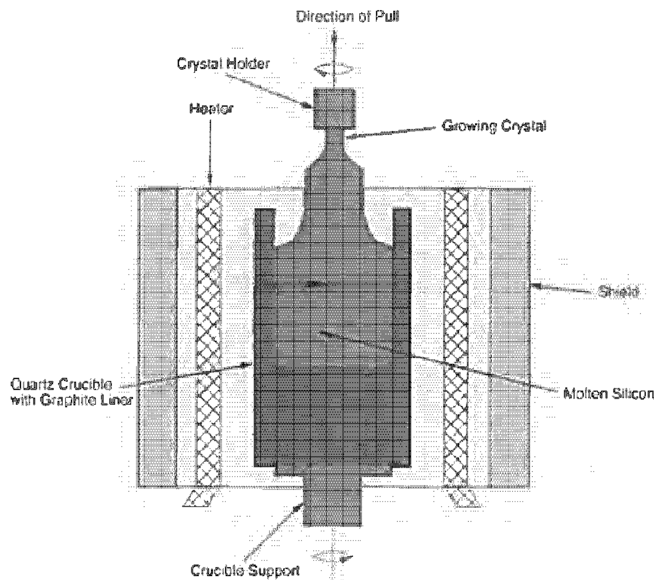


FIGURE 3.1 Czochralski method for manufacturing silicon ingots

required electrical properties. The crystal orientation is determined by a seed crystal that is dipped into the melt to initiate single-crystal growth. The melt is contained in a quartz crucible, which is surrounded by a graphite radiator. The graphite is heated by radio frequency induction and the temperature is maintained a few degrees above the melting point of silicon (≈1425°C). The atmosphere above the melt is typically helium or argon.

After the seed is dipped into the melt, the seed is gradually withdrawn vertically from the melt while simultaneously being rotated. The molten polycrystalline silicon melts the tip of the seed, and as it is withdrawn, refreezing occurs. As the melt freezes, it assumes the single crystal form of the seed. This process is continued until the melt is consumed. The diameter of the ingot is determined by the seed withdrawal rate and the seed rotation rate. Growth rates range from 30 to 180 mm/hour.

Slicing into wafers is usually carried out using internal cutting-edge diamond blades. Wafers are usually between 0.25 mm and 1.0 mm thick, depending on their diameter. Following this operation, at least one face is polished to a flat, scratch-free mirror finish.

3.1.2 Oxidation

Many of the structures and manufacturing techniques used to make silicon integrated circuits rely on the properties of the oxide of silicon, namely, silicon dioxide (SiO_2). Therefore the reliable manufacture of SiO_2 is extremely important.

Oxidation of silicon is achieved by heating silicon wafers in an oxidizing atmosphere such as oxygen or water vapor. The two common approaches are:

- Wet oxidation: when the oxidizing atmosphere contains water vapor. The temperature is usually between 900°C and 1000°C. This is a rapid process.
- Dry oxidation: when the oxidizing atmosphere is pure oxygen. Temperatures are in the region of 1200°C, to achieve an acceptable growth rate.

The oxidation process consumes silicon. Since SiO_2 has approximately twice the volume of silicon, the SiO_2 layer grows almost equally in both vertical directions. This effect is shown in Fig. 3.2 for an n-channel MOS device in which the SiO_2 (field oxide) projects above and below the unoxidized silicon surface.

3.1.3 Epitaxy, Deposition, Ion-Implantation, and Diffusion

To build various semiconductor devices, silicon containing varying proportions of donor or acceptor impurities is required. This may be achieved using epitaxy, deposition, or implantation. Epitaxy involves growing a single-crys-

LG Display Co., Ltd.
Exhibit 1012
Page 092

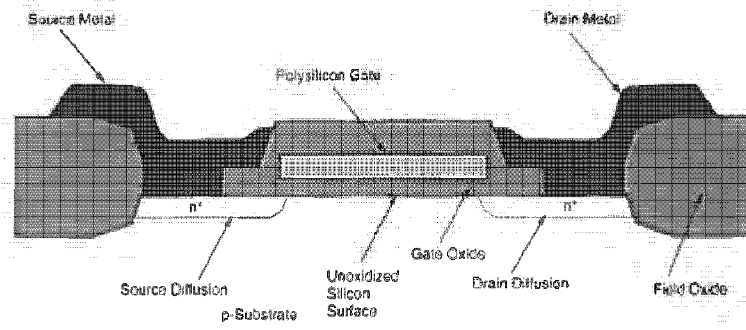


FIGURE 3.2 An nMOS transistor showing the growth of field oxide below the silicon surface

tal film on the silicon surface (which is already a single crystal) by subjecting the silicon wafer surface to elevated temperature and a source of dopant material. Deposition might involve evaporating dopant material onto the silicon surface followed by a thermal cycle, which is used to drive the impurities from the surface of the silicon into the bulk. Ion implantation involves subjecting the silicon substrate to highly energized donor or acceptor atoms. When these atoms impinge on the silicon surface, they travel below the surface of the silicon, forming regions with varying doping concentrations. At any elevated temperature (> 800°C) diffusion will occur between any silicon that has differing densities of impurities, with impurities tending to diffuse from areas of high concentration to areas of low concentration. Hence it is important once the doped areas have been put in place to keep the remaining process steps at as low a temperature as possible.

Construction of transistors and other structures of interest depends on the ability to control where and how many and what type of impurities are introduced into the silicon surface. What type of impurities are introduced is controlled by the dopant source. Boron is frequently used for creating acceptor silicon, while arsenic and phosphorous are commonly used to create donor silicon. How much is used is determined by the energy and time of the ion-implantation or the time and temperature of the deposition and diffusion step. Where it is used is determined by using special materials as masks. In places covered by the mask ion implantation does not occur or the dopant does not contact the silicon surface. In areas where the mask is absent the implantation occurs, or the predeposited material is allowed to diffuse into the silicon. The common materials used as masks include

- photoresist.
- polysilicon (polycrystalline silicon).
- silicon dioxide (SiO_2).
- silicon nitride (SiN).

LG Display Co., Ltd.
Exhibit 1012
Page 093

The ability of these materials to act as a barrier against doping impurities is a vital factor in this process, called *selective diffusion*. Thus selective diffusion entails

- patterning *windows* in a mask material on the surface of the wafer.
- subjecting exposed areas to a dopant source.
- removing any unrequired mask material.

In the case of an oxide mask, the process used for selectively removing the oxide involves covering the surface of the oxide with an acid resistant coating, except where oxide windows are needed. The SiO_2 is removed using an etching technique. The acid resistant coating is normally a photosensitive organic material called *photoresist* (PR), which can be polymerized by ultraviolet (UV) light. If the UV light is passed through a mask containing the desired pattern, the coating can be polymerized where the pattern is to appear. The polymerized areas may be removed with an organic solvent. Etching of exposed SiO_2 then may proceed. This is called a positive resist. There are also negative resists where the unexposed PR is dissolved by the solvent. This process is illustrated in Fig. 3.3. In established processes using PRs in conjunction with UV light sources, diffraction around the edges of the mask patterns and alignment tolerances limit line widths to around 0.8 µm. During recent years, electron beam lithography (EBL) has emerged as a contender for pattern generation and imaging where line widths of the order of 0.5 µm with good definition are achievable. The main advantages of EBL pattern generation are as follows:

- Patterns are derived directly from digital data.
- There are no intermediate hardware images such as rectICLes or masks: that is, the process can be direct.
- Different patterns may be accommodated in different sections of the wafer without difficulty.
- Changes to patterns can be implemented quickly.

The main disadvantage that has precluded the use of this technique in commercial fabrication lines is the cost of the equipment and the large amount of time required to access all points on the wafer.

3.1.4 The Silicon Gate Process

So far we have touched on the single-crystal form of silicon used in the manufacture of wafers and the oxide used in the manufacture and operation of circuits. Silicon may also be formed in a *polycrystalline* form (not having a single-crystalline structure) called *polysilicon*. This is used as an intercon-

LG Display Co., Ltd.
Exhibit 1012
Page 094

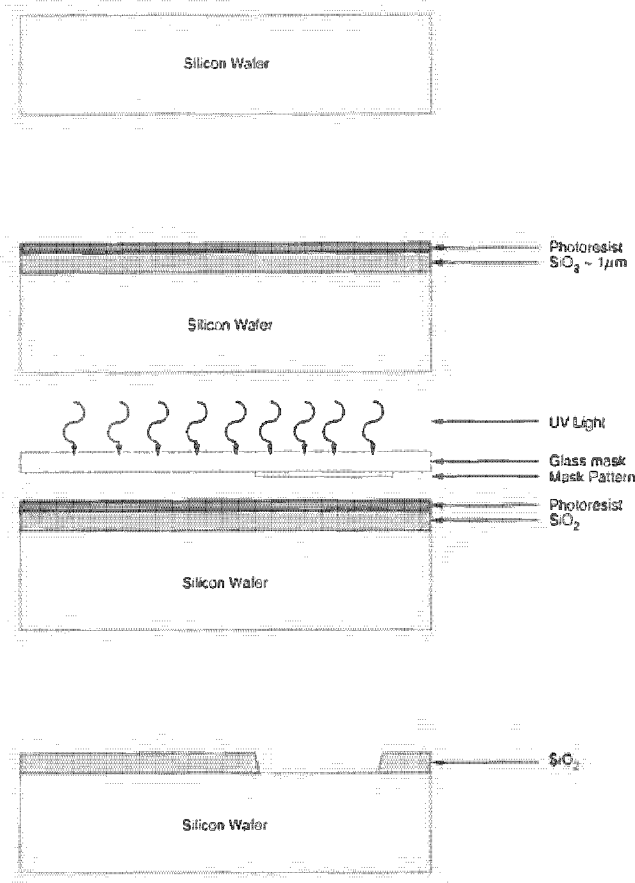


FIGURE 3.3 Simplified steps involved in the patterning of SiO_2: (a) Bare silicon wafer; (b) Wafer with SiO_2 and resist; (c) Exposing resist to UV light; (d) Final etched SiO_2

nect in silicon ICs and as the gate electrode on MOS transistors. The most significant aspect of using polysilicon as the gate electrode is its ability to be used as a further mask to allow precise definition of source and drain electrodes. This is achieved with minimum gate-to-source/drain overlap, which, we will learn, improves circuit performance. Polysilicon is formed when silicon is deposited on SiO_2 or other surfaces. In the case of an MOS transistor gate electrode, undoped polysilicon is deposited on the gate insulator. Polysilicon and source/drain regions are then normally doped at the same time. Undoped polysilicon has high resistivity. This characteristic is used to provide high-value resistors in static memories. The resistivity of polysilicon may be reduced by combining it with a refractory metal (see Section 3.2.4).

The steps involved in a typical silicon gate process entail photomasking and oxide etching, which are repeated a number of times during the processing sequence. Figure 3.4 shows the processing steps after the initial pattern-

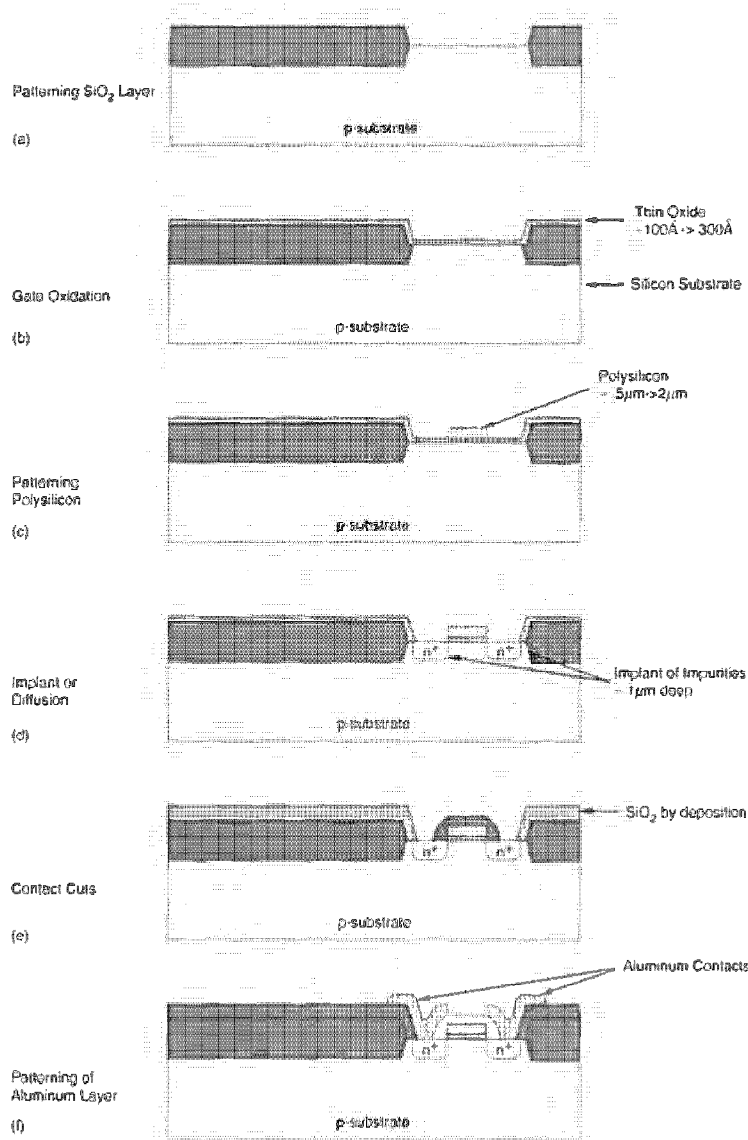


Figure 3.4 Fabrication steps for a silicon gate nMOS transistor

LG Display Co., Ltd.
Exhibit 1012
Page 096

ing of the SiO_2, which was shown in Fig. 3.3. The wafer is covered with SiO_2 with at least two different thicknesses (Fig. 3.4b). A thin, highly controlled layer of SiO_2 is required where active transistors are desired. This is called the gate-oxide or thinox. A thick layer of SiO_2 is required elsewhere to isolate the individual transistors. This is normally called the field oxide. We will examine a variety of methods of achieving these two oxide thicknesses in Section 3.2.1.

Polysilicon is then deposited over the wafer surface and etched to form interconnections and transistor gates. Figure 3.4(c) shows the result of an etched polysilicon gate. The exposed gate oxide (not covered by polysilicon) is then etched away. The complete wafer is then exposed to a dopant source or is ion-implanted, resulting in two actions (Fig. 3.4d). Diffusion junctions are formed in the substrate and the polysilicon is doped with the particular type of dopant. This also reduces the resistivity of the polysilicon. Note that the diffusion junctions form the drain and source of the MOS transistor. They are formed only in regions where the polysilicon gate does not shadow the underlying substrate. This is referred to as a *self-aligned* process because the source and drain do not extend under the gate. Finally, the complete structure is covered with SiO_2 and contact holes are etched to make contact with underlying layers (Fig. 3.4e). Aluminum or other metallic interconnect is evaporated and etched to complete the final connection of elements (Fig. 3.4f). Further oxide layers, contact holes and metallization layers are normally added for extra interconnect.

Note that parasitic MOS transistors exist between unrelated transistors, as shown in Fig. 3.5. Here the source and drain of the parasitic transistor are existing source/drains and the gate is a metal or polysilicon interconnect overlapping the two source/drain regions. The "gate-oxide" is in fact the thick field oxide. The threshold voltage of this transistor is much higher than that of a regular transistor (this device is commonly called a field device) (Eq. 2.1). The high threshold voltage is usually ensured by making the field oxide thick enough and introducing a "channel-stop" diffusion, which raises

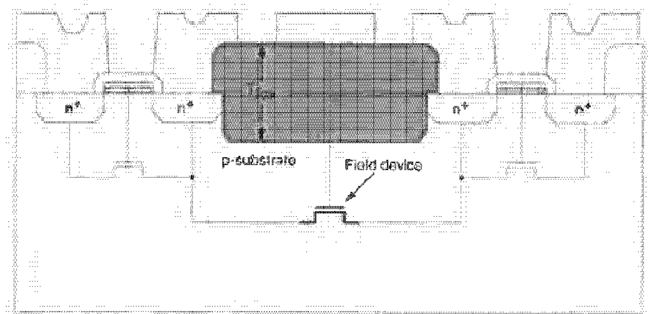


FIGURE 3.5 A parasitic MOS transistor or field device

the impurity concentration in the substrate in areas where transistors are not required, thus further increasing the threshold voltage (Section 2.1.3.1). These devices do have some useful purposes where the fact that they turn on at voltages higher than normal operating voltages may be used to protect other circuitry.

3.2 Basic CMOS Technology

CMOS (Complementary Metal Oxide Silicon) technology is recognized as the leading VLSI systems technology. CMOS provides an inherently low power static circuit technology that has the capability of providing a lower power-delay product than comparable design-rule bipolar, nMOS, or GaAs technologies. In this section we provide an overview of four dominant CMOS technologies, with a simplified treatment of the process steps. This is included primarily as a guide for better appreciation of the layout styles that may be used to implement CMOS gates.

The four main CMOS technologies are:

- n-well process.
- p-well process.
- twin-tub process.
- silicon on insulator.

In addition, by adding bipolar transistors a range of BiCMOS processes are possible.

During the discussion of CMOS technologies, process cross-sections and layouts will be presented. Figure 3.6 summarizes the drawing conventions.

3.2.1 A Basic n-well CMOS Process

A common approach to n-well CMOS fabrication has been to start with a lightly doped p-type substrate (wafer), create the n-type well for the p-channel devices, and build the n-channel transistor in the *native* p-substrate. Although the processing steps are somewhat complex and depend on the fabrication line, Fig. 3.7 illustrates the major steps involved in a typical n-well CMOS process. The mask that is used in each process step is shown in addition to a sample cross-section through an *n*-device and a p-device. Although we have shown a polysilicon gate process, it is of historical significance to note that CMOS was originally implemented with metal (aluminum) gates. This technology (in p-well form) formed the basis for the majority of low

LG Display Co., Ltd.
Exhibit 1012
Page 098

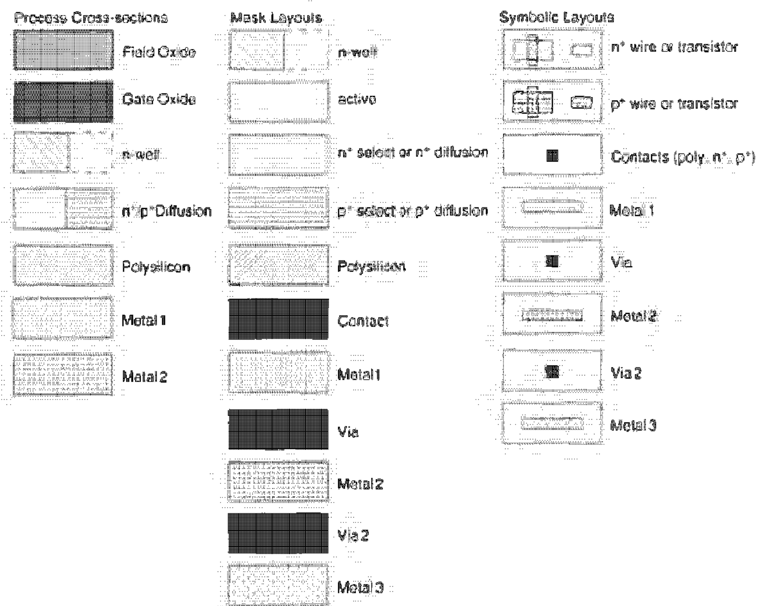


FIGURE 3.6 CMOS process and layout drawing conventions

power CMOS circuits implemented in the 1970s. The technology is robust and still in use. As can be seen from Fig. 3.7, the mask levels are not organized by component function. Rather they reflect the processing steps.

- The first mask defines the n-well (or n-tub); p-channel transistors will be fabricated in this well. Ion implantation or deposition and diffusion is used to produce the n-well (Fig. 3.7a). The former tends to produce shallower wells which are compatible with fine dimension processes. As the diffusion process occurs in all directions, the deeper a diffusion is the more it spreads laterally. This lateral spread affects how near to other structures wells can be placed. Hence, for closely spaced structures a shallow well is required. From a patterned well shape, the final well will extend outside the patterned dimension by the lateral diffusion.
- The next mask is called the "active" mask, because it defines where areas of thin oxide are needed to implement transistor gates and allow implantation to form p- or n-type diffusions for transistor source/drain regions (Fig. 3.7b). Other terms for this mask include *thinox*, *thin-oxide*, *island*, and *mesa*. A thin layer of SiO_2 is grown and covered with SiN. This is used as a masking layer for the following two steps.

LG Display Co., Ltd.
Exhibit 1012
Page 099

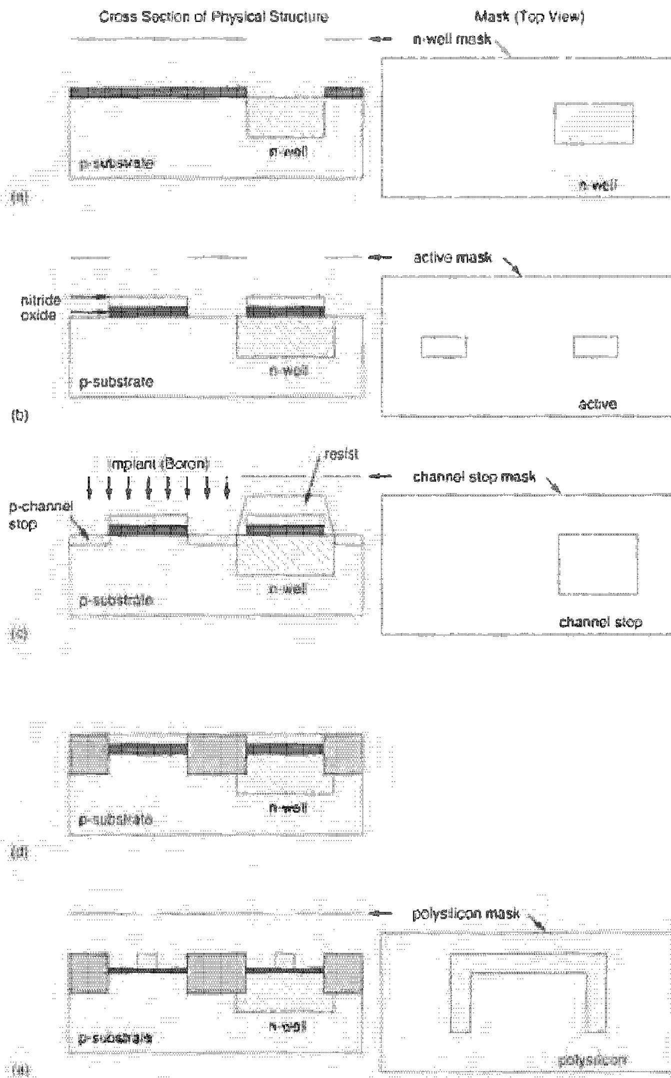


FIGURE 3.7 A typical n-well CMOS process

- The channel-stop implant is usually then completed. This uses the p-well mask (the complement of the n-well mask). It dopes the p-substrate in areas where there are no n-transistors p^+ using a photoresist mask (Fig. 3.7c). This, in conjunction with the thick field oxide that will cover these areas, aids in preventing conduction between unrelated transistor source/drains.

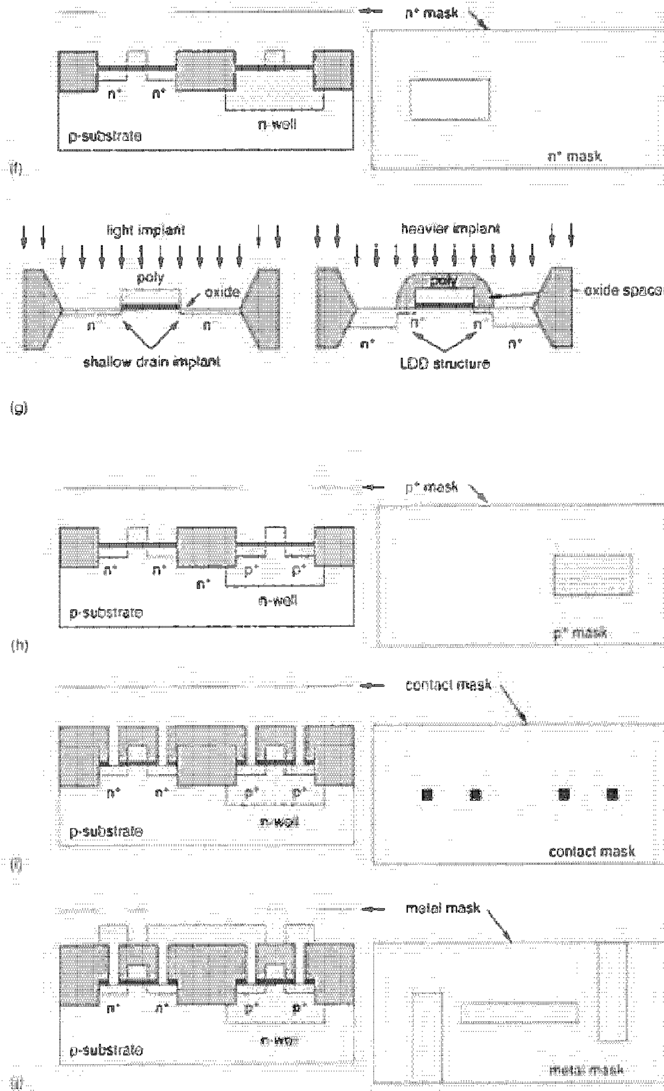


FIGURE 3.7 *(continued)*

- Following the channel-stop implant, the photoresist mask is stripped, leaving the previously masked SiO_2/SiN sandwich defining the active regions. The thick field oxide is then grown. This grows in areas where the SiN layer is absent. The oxide grows in both directions ver-

tically and also laterally under the SiO_2/SiN sandwich (Fig. 3.7d). This lateral movement results in what is called a "bird's beak" because of the shape of the oxide encroachment under the gate oxide mask. This general oxide construction technique is called LOCOS for LOcal Oxidation Of Silicon. The oxide encroachment results in an active area that is smaller than patterned. In particular, the width dimension of a transistor will be reduced from what might be expected from the photolithography. Other techniques such as SWAMI (Side-WAll Masked Isolation)[1,2] have been developed to reduce the effect of the bird's beak. Of additional concern is the final planarity of the field oxide/gate oxide interface. If the difference in height is too great, the subsequent conductors may have "step coverage" problems in which a conductor thins and can even break as it crosses a thick to thin oxide boundary. To counter this, many planarization techniques have been developed. One such technique is to pre-etch the silicon in areas where the field oxide is to be grown by around half the final required field oxide thickness. The LOCOS oxide is then grown and the final field oxide/gate oxide interface is very planar.

- An n-transistor threshold voltage adjust step might then be performed using a p-well photoresist mask. In current fabrication processes the polysilicon is normally doped n^+. With normal doping concentrations suitable for small dimension processes, this results in threshold voltage for n-devices of around 0.5–0.7 volts. However, the p-device threshold is around −1.5 to −2.0 volts. Thus the p-device has to have its threshold voltage adjusted more than the n-device. This is done by introducing an additional negatively charged layer at the silicon/oxide interface. This moves the channel from the silicon/oxide interface further into the silicon, creating a "buried channel" device.[3] Following these two steps the gate oxide is grown.
- Polysilicon gate definition is then completed. This involves covering the surface with polysilicon and then etching the required pattern (in this case an inverted "U"). As noted previously, the "poly" gate regions lead to "self-aligned" source-drain regions (Fig. 3.7e).
- An n-plus (n^+) mask is then used to indicate those thin-oxide areas (and polysilicon) that are to be implanted n^+. Hence a thin-oxide area exposed by the n-plus mask will become an n^+ diffusion area (Fig. 3.7f). If the n-plus area is in the p-substrate, then an n-channel transistor or n-type wire may be constructed. If the n-plus area is in the n-well (not shown), then an *ohmic* contact to the n-well may be constructed. An ohmic contact is one which is only resistive in nature and is not rectifying (as in the case of a diode). In other words, there is no junction (n-type and p-type silicon abutting). Current can flow in both directions in an ohmic contact. This type of mask is sometimes

called the *select* mask because it *selects* those transistor regions that are to be n-type. In modern small dimension processes, to reduce hot carrier effects, considerable effort may go into what is termed "drain engineering."[4] Rather than using one single diffusion or implantation step and mask to produce the source/drain regions, quite complicated structures are constructed. Typical of these structures is the LDD or Lightly Doped Drain structure, which is illustrated in Fig. 3.7(g). This consists of a shallow n-LDD implant that covers the source/drain region where there is no poly (i.e., the normal source/drain region). A spacer oxide is then grown over the polysilicon gate. An n^+ implant is then used to produce n^+ implants that are spaced from the edge of the original poly gate edges. The spacer is then removed, resulting in a structure that is more resistant to hot-electron effects. Current 0.25μm processes revert to a simpler self-aligned structure presumably because of the complexity of the LDD structure.

- The next step usually uses the complement of the n-plus mask, although an extra mask is normally not needed. The "absence" of an n-plus region over a thin-oxide area indicates that the area will be a p^+ diffusion or p-active. P-active in the n-well defines possible p-transistors and wires (Fig. 3.7h). A p^+ diffusion in the p-substrate allows an ohmic contact to be made. Following this step, the surface of the chip is covered with a layer of SiO_2. The LDD step is not necessarily done for p-transistors because their hot-carrier susceptibility is much less than that of n-transistors. For this reason, the drawn length dimension of p-transistors might be larger than that of the n-transistors.
- Contact cuts are then defined. This involves etching any SiO_2 down to the surface to be contacted (Fig. 3.7i). These allow metal (next step) to contact diffusion regions or polysilicon regions.
- Metallization is then applied to the surface and selectively etched (Fig. 3.7j) to produce circuit interconnections.
- As a final step (not shown), the wafer is passivated and openings to the bond pads are etched to allow for wire bonding. Passivation protects the silicon surface against the ingress of contaminants that can modify circuit behavior in deleterious ways.

The cross-section of the finished n-well process is shown in Fig. 3.8(c). The layout of the n-well CMOS transistors corresponding to this cross-section is illustrated in Fig. 3.8(b). The corresponding schematic (for an inverter) is shown in Fig. 3.8(a). From Fig. 3.8 it is evident that the p-type substrate accommodates n-channel devices, whereas the n-well accommodates p-channel devices. (Figure 3.8 also appears in color as Plate 1.)

In an n-well process, the p-type substrate is normally connected to the negative supply (V_{SS}) through what are termed V_{SS} substrate contacts, while

LG Display Co., Ltd.
Exhibit 1012
Page 103

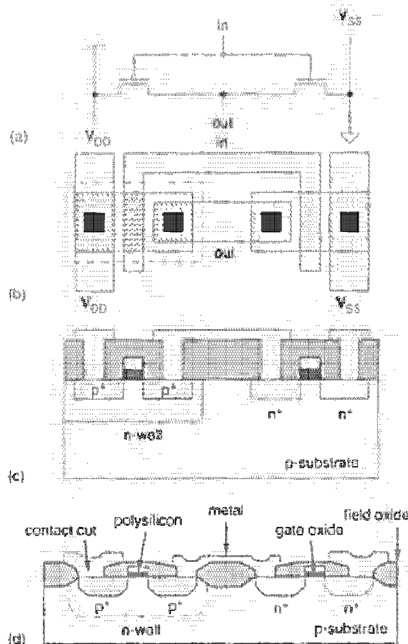


FIGURE 3.8 Cross section of a CMOS inverter in an n-well process

the well has to be connected to the positive supply (V_{DD}) through V_{DD} substrate (or well or tub) contacts. As the substrate is accessible at the top of the wafer and the bottom, connecting the substrate may be accomplished from the backside of the wafer. Topside connection is preferred because it reduces parasitic resistances that could cause latchup (see later). Substrate connections that are formed by placing n^+ regions in the n-well (V_{DD} contacts) and p^+ in the p-type substrate (V_{SS} contacts) are illustrated by Fig. 3.9(a). The corresponding layout is shown in Fig. 3.9(b). Other terminology for these contacts include "well contacts," "body ties," or "tub ties" for the V_{DD} substrate connection. We will use the term "substrate contact" for both V_{SS} and V_{DD} contacts, because this terminology can be commonly used for most bulk CMOS processes. It should be noted that these contacts are formed during the implants used for the p-channel and n-channel transistor formation.

3.2.2 The p-well Process

N-well processes have emerged in popularity in recent years. Prior to this, p-well processes were one of the most commonly available forms of CMOS. Typical p-well fabrication steps are similar to an n-well process, except that a p-well is implanted rather than an n-well. The first masking step defines the p-well regions. This is followed by a low-dose phosphorous implant driven

LG Display Co., Ltd.
Exhibit 1012
Page 104

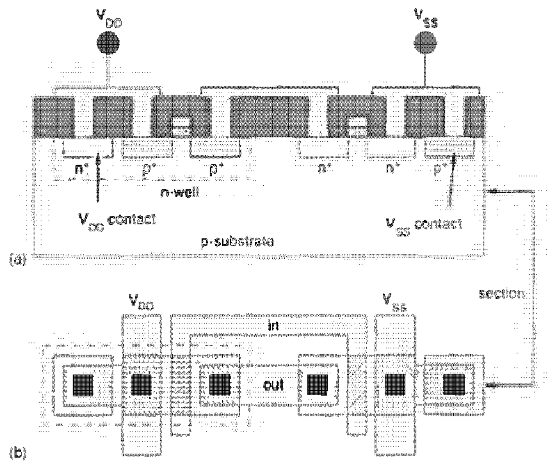


FIGURE 3.9 Substrate and well contacts in an n-well process

in by a high-temperature step for the formation of the p-well. The well depth is optimized to ensure against n-substrate to n^+ diffusion breakdown, without compromising p-well to p^+ separation. The next steps are to define the devices and other diffusions; to grow field oxide; contact cuts; and metallization. A p-well mask is used to define p-well regions, as opposed to an n-well mask in an n-well process. A p-plus (p^+) mask may be used to define the p-channel transistors and V_{SS} contacts. Alternatively, we could use an n-plus mask to define the n-channel transistors, because the masks usually are the complement of each other.

P-well processes are preferred in circumstances where the characteristics of the n- and p-transistors are required to be more balanced than that achievable in an n-well process. Because the transistor that resides in the native substrate tends to have better characteristics, the p-well process has better p devices than an n-well process. Because p-devices are inherently lower gain than n devices, the n-well process exacerbates this difference while a p-well process moderates the difference.

3.2.3 Twin-Tub Processes

Twin-tub CMOS technology provides the basis for separate optimization of the p-type and n-type transistors, thus making it possible for threshold voltage, body effect, and the gain associated with n- and p-devices to be independently optimized.[5,6] Generally, the starting material is either an n^+ or p^+ substrate with a lightly doped *epitaxial* or *epi* layer, which is used for protection against latchup (see Section 3.5). The aim of *epitaxy* (which means "arranged upon") is to grow high-purity silicon layers of controlled thick-

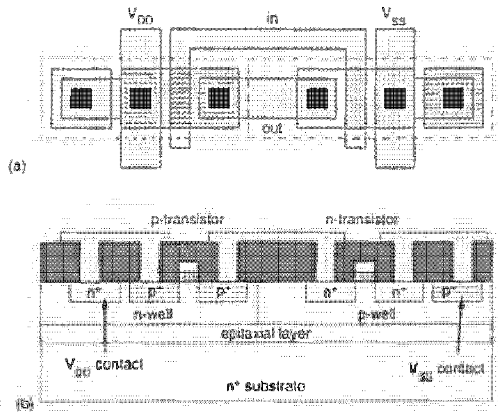


FIGURE 3.10 Twin-well CMOS process cross section

ness with accurately determined dopant concentrations distributed homogeneously throughout the layer. The electrical properties of this layer are determined by the dopant and its concentration in the silicon. The process sequence, which is similar to the n-well process apart from the tub formation where both p-well and n-well are utilized, entails the following steps:

- Tub formation.
- Thin-oxide construction.
- Source and drain implantations.
- Contact cut definition.
- Metallization.

Since this process provides separately optimized wells, balanced performance n-transistors and p-transistors may be constructed. Note that the use of threshold adjust steps is included in this process. These masks are derived from the active and n-plus masks. The cross-section of a typical twin-tub structure is shown in Fig. 3.10. The substrate contacts (both of which are required) are also included.

3.2.4 Silicon On Insulator

Rather than using silicon as the substrate, technologists have sought to use an insulating substrate to improve process characteristics such as latchup and speed. Hence the emergence of Silicon On Insulator (SOI) technologies. SOI CMOS processes have several potential advantages over the traditional CMOS technologies.[7] These include closer packing of p- and n-transistors, absence of latchup problems, and lower parasitic substrate capacitances. In

LG Display Co., Ltd.
Exhibit 1012
Page 106

the SOI process a thin layer of single-crystal silicon film is epitaxially grown on an insulator such as sapphire or magnesium aluminate spinel.[8] Alternatively, the silicon may be grown on SiO_2 that has been in turn grown on silicon. This option has proved more popular in recent years due to the compatibility of the starting material with conventional silicon CMOS fabrication. Various masking and doping techniques (Fig. 3.11) are then used to form p-channel and n-channel devices. Unlike the more conventional CMOS approaches, the extra steps in well formation do not exist in this technology. The steps used in typical SOI CMOS processes are as follows:

- A thin film (7–8 μm) of very lightly-doped n-type Si is grown over an insulator. Sapphire or SiO_2 is a commonly used insulator (Fig. 3.11a).
- An anisotropic etch is used to etch away the Si except where a diffusion area (n or p) will be needed. The etch must be anisotropic since the thickness of the Si is much greater than the spacings desired between the Si "islands" (Fig. 3.11b, 3.11c).
- The p-islands are formed next by masking the n-islands with a photoresist. A p-type dopant, boron, for example—is then implanted. It is masked by the photoresist, but forms p-islands at the unmasked islands. The p-islands will become the n-channel devices (Fig. 3.11d).
- The p-islands are then covered with a photoresist and an n-type dopant—phosphorus, for example—is implanted to form the n-islands. The n-islands will become the p-channel devices (Fig. 3.11e).
- A thin gate oxide (around 100–250 Å) is grown over all of the Si structures. This is normally done by thermal oxidation.
- A polysilicon film is deposited over the oxide. Often the polysilicon is doped with phosphorus to reduce its resistivity (Fig. 3.11f).
- The polysilicon is then patterned by photomasking and is etched. This defines the polysilicon layer in the structure (Fig. 3.11g).
- The next step is to form the n-doped source and drain of the n-channel devices in the p-islands. The n-islands are covered with a photoresist and an n-type dopant, normally phosphorus, is implanted. The dopant will be blocked at the n-islands by the photoresist, and it will be blocked from the gate region of the p-islands by the polysilicon. After this step the n-channel devices are complete (Fig. 3.11h).
- The p-channel devices are formed next by masking the p-islands and implanting a p-type dopant such as boron. The polysilicon over the gate of the n-islands will block the dopant from the gate, thus forming the p-channel devices (Fig. 3.11i).
- A layer of phosphorus glass or some other insulator such as silicon dioxide is then deposited over the entire structure.

LG Display Co., Ltd.
Exhibit 1012
Page 107

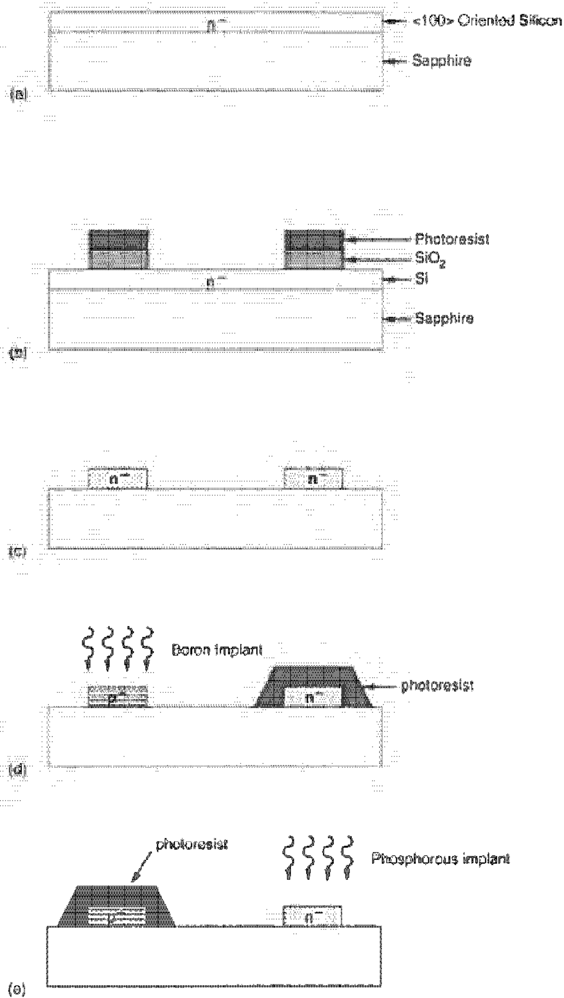


FIGURE 3.11 SOI process flow

- The glass is etched at contact-cut locations. The metallization layer is formed next by evaporating aluminum over the entire surface and etching it to leave only the desired metal wires. The aluminum will flow through the contact cuts to make contact with the diffusion or polysilicon regions (Fig. 3.11j)
- A final passivation layer of phosphorus glass is deposited and etched over bonding pad locations (not shown).

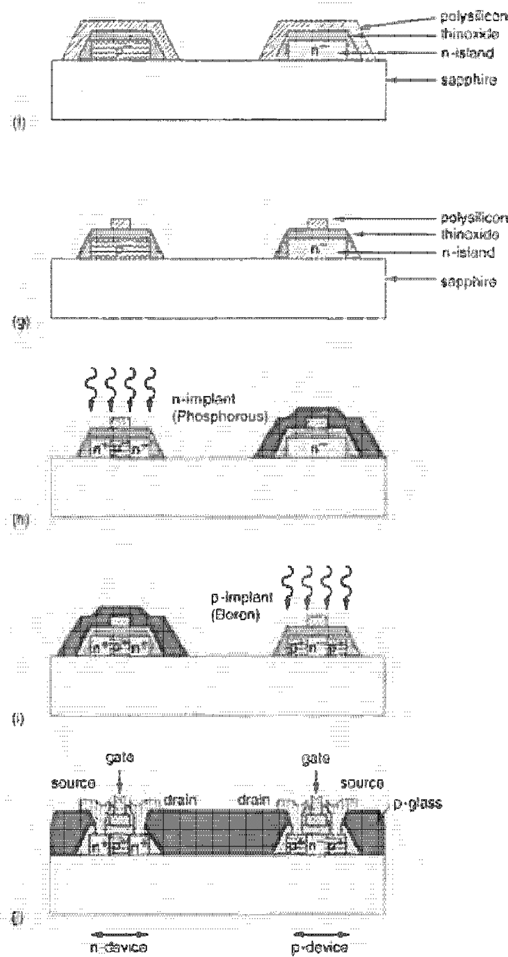


FIGURE 3.11 *(continued)*

Because the diffusion regions extend down to the insulating substrate, only "sidewall" areas associated with source and drain diffusions contribute to the parasitic junction capacitance. Since sapphire and SiO_2 are extremely good insulators, leakage currents between transistors and substrate and adjacent devices are almost eliminated.

In order to improve the yield, some processes use "preferential etch," in which the island edges are tapered. Thus aluminum or poly runners can enter and leave the islands with a minimum step height. This is contrasted to "fully anisotropic etch," in which the undercut is brought to zero, as shown in Fig. 3.12. An "isotropic etch" is also shown in the same diagram for comparison.

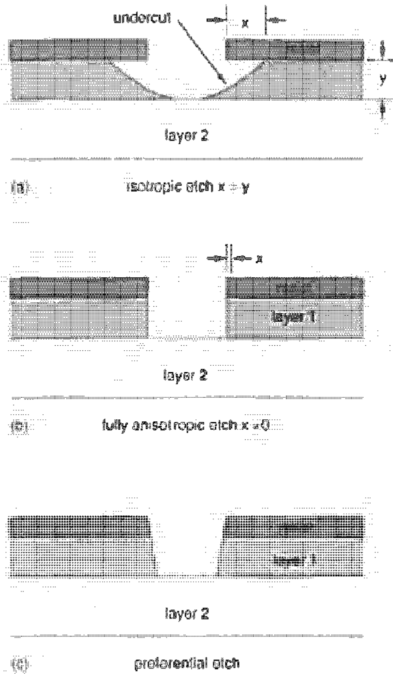


FIGURE 3.12 Classification of etching processes

The advantages of SOI technology are as follows:

- Due to the absence of wells, transistor structures denser than bulk silicon are feasible. Also direct n-to-p connections may be made.
- Lower substrate capacitances provide the possibility for faster circuits.
- No field-inversion problems exist (insulating substrate).
- There is no latchup because of the isolation of the n- and p-transistors by the insulating substrate.
- Because there is no conducting substrate, there are no body-effect problems. However, the absence of a backside substrate contact could lead to odd device characteristics, such as the "kink" effect in which the drain current increases abruptly at around 2 to 3 volts.[9]
- There is enhanced radiation tolerance (in fact, this is almost the sole reason the technology has been justified to date).

However, on the negative side, due to absence of substrate diodes, the inputs are somewhat more difficult to protect. Because device gains are lower, I/O structures have to be larger. Although parasitic capacitances to the

LG Display Co., Ltd.
Exhibit 1012
Page 110

substrate are reduced, the coupling capacitance between wires still exists so that the actual reduction in stray load capacitance is less than one would hope (see Chapter 4). The density advantage of SOI is not particularly important, because the density of contemporary digital processes is determined by the number and density of the metal interconnection layers. Single-crystal sapphire, spinel substrates, and silicon on SiO_2 are considerably more expensive than silicon substrates, and their processing techniques tend to be less developed than bulk silicon techniques. Recently, companies have started to produce SOI substrates that can be used interchangeably with silicon substrates in bulk CMOS fabrication lines. As the barrier to using insulating substrates is reduced, more use of them might be seen in day-to-day circuits, where the possible performance increase justifies the increase in processing cost and complexity.

3.3 CMOS Process Enhancements

A number of enhancements may be added to the CMOS processes, primarily to increase routability of circuits, provide high-quality capacitors for analog circuits and memories, or provide resistors of variable characteristics.

These enhancements include

- double- or triple- or quadruple-level metal (or more).
- double- or triple-level poly (or more).
- combinations of the above.

We will examine these additions in terms of the additional functionality that they bring to a basic CMOS process.

3.3.1 Interconnect

Probably the most important additions for CMOS logic processes are additional signal- and power-routing layers. This eases the routing (especially automated routing) of logic signals between modules and improves the power and clock distribution to modules. Improved routability is achieved through additional layers of metal or by improving the existing polysilicon interconnection layer.

3.3.1.1 Metal Interconnect

A second level of metal is almost mandatory for modern CMOS digital design. A third layer is becoming common and is certainly required for leading-edge high-density, high-speed chips. Normally, aluminum is used for the

LG Display Co., Ltd.
Exhibit 1012
Page 111

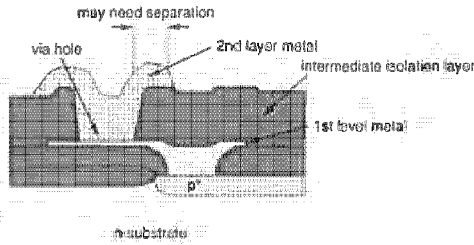


FIGURE 3.13 Two-level metal process cross section

metal layers. If some form of planarization is employed the second-level metal pitch can be the same as the first. As the vertical topology becomes more varied, the width and spacing of metal conductors has to increase so that the conductors do not thin and hence break at vertical topology jumps (*step coverage*).

Contacting the second-layer metal to the first-layer metal is achieved by a *via*, as shown in Fig. 3.13. If further contact to diffusion or polysilicon is required, a separation between the via and the contact cut is usually required. This requires a first-level metal *tab* to bridge between metal2 and the lower-level conductor. It is important to realize that in contemporary processes first-level metal must be involved in any contact to underlying areas. A number of contact geometries are shown in Fig. 3.14. Processes usually require metal borders around the via on both levels of metal although some pro-

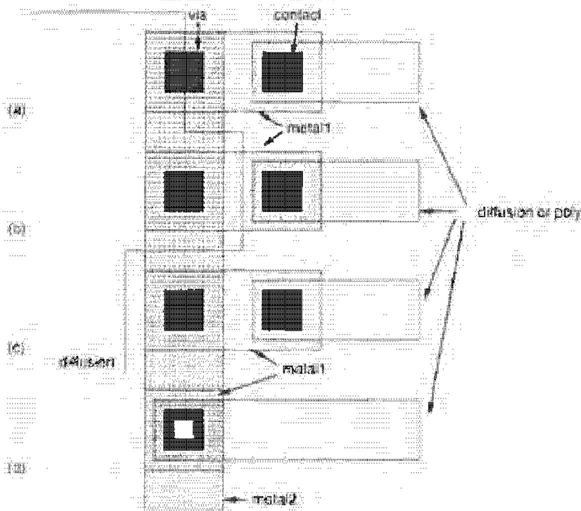


FIGURE 3.14 Two-level metal via/contact geometries

LG Display Co., Ltd.
Exhibit 1012
Page 112

cesses require none. Processes may have no restrictions on the placement of the via with respect to underlying layers (Fig. 3.14a) or they may have to be placed inside (Fig. 3.14b) or outside (Fig. 3.14c) the underlying polysilicon or diffusion areas. Aggressive processes allow the stacking of vias on top of contacts, as shown in Fig. 3.14(d). Consistent with the relatively large thickness of the intermediate isolation layer, the vias might be larger than contact cuts and second-layer metal may need to be thicker and require a larger via overlap although modern processes strive for uniform pitches on metal1 and metal2.

The process steps for a two-metal process are briefly as follows:

- The oxide below the first-metal layer is deposited by atmospheric chemical vapor deposition (CVD).
- The second oxide layer between the two metal layers is applied in a similar manner.
- Depending on the process, removal of the oxide is accomplished using a plasma etcher designed to have a high rate of vertical ion bombardment. This allows fast and uniform etch rates. The structure of a via etched using such a method is shown in Fig. 3.13.

3.3.1.2 Polysilicon/Refractory Metal Interconnect

The polysilicon layer used for the gates of transistors is commonly used as an interconnect layer. However, the sheet resistance of doped polysilicon is between 20 and 40 Ω/square. If used as a long distance conductor, a polysilicon wire can represent a significant delay (see Chapter 4).

One method to improve this that requires no extra mask levels is to reduce the polysilicon resistance by combining it with a refractory metal. Three such approaches are illustrated in Fig. 3.15.[10] In Fig. 3.15(a), a *silicide* (e.g., silicon and tantalum) is used as the gate material. Sheet resistances of the order of 1 to 5 Ω/square may be obtained. This is called the silicide gate approach. Silicides are mechanically strong and may be dry etched in plasma reactors. Tantalum silicide is stable throughout standard processing and has the advantage that it may be retrofitted into existing process lines. Figure 3.15(b) uses a sandwich of silicide upon polysilicon,

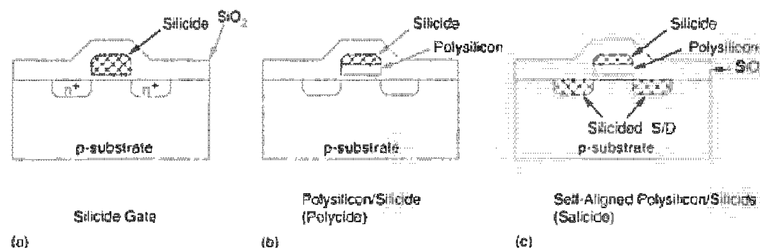


FIGURE 3.15 Refractory metal interconnect

LG Display Co., Ltd.
Exhibit 1012
Page 113

which is commonly called the *polycide* approach. Finally, the silicide/polysilicon approach may be extended to include the formation of source and drain regions using the silicide. This is called the *salicide* process (Self ALigned SILICIDE) (Fig. 3.15c). The effect of all of these processes is to reduce the "second layer" interconnect resistance, allowing the gate material to be used as a moderate long-distance interconnect. This is achieved by minimum perturbation of an existing process. An increasing trend in processes is to use the salicide approach to reduce the resistance of both gate and source/drain conductors.

3.3.1.3 Local Interconnect

The silicide itself may be used as a "local interconnect" layer for connection within cells.[11] As an example TiN[12] is used. Local interconnect allows a direct connection between polysilicon and diffusion, thus alleviating the need for area-intensive contacts and metal. Figure 3.16 shows a portion (p-devices only) of a six transistor SRAM cell that uses local interconnect. The local interconnect has been used to make the polysilicon-to-diffusion connections within the cell, thereby alleviating the need to use metal (and contacts). Metal2 (not shown) bit lines run over the cell vertically. Use of local interconnect in this RAM reduced the cell area by 25%. In general, local

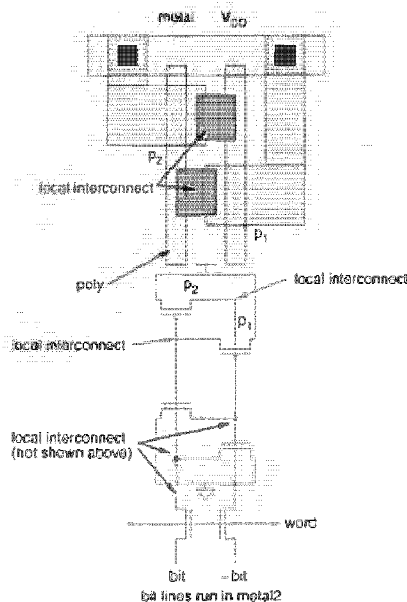


FIGURE 3.16 Local interconnect as used in a RAM cell

LG Display Co., Ltd.
Exhibit 1012
Page 114

interconnect if available can be used to complete intracell routing, leaving the remaining metal layers for global wiring.

3.3.2 Circuit Elements

3.3.2.1 Resistors

Polysilicon, if left undoped, is highly resistive. This property is used to build resistors that are used in static memory cells. The process step is achieved by preventing the resistor areas from being implanted during normal processing. Resistors in the tera-Ω (10^{12} Ω) region are used.[13] A value of 3 TΩ, results in a standby current of 2μA for a 1 Mbit memory.

For mixed signal CMOS (analog and digital), a resistive metal such as nichrome may be added to produce high-value, high-quality resistors. The resistor accuracy might be further improved by laser trimming the resulting resistors on each chip to some predetermined test specification. In this process a high-powered laser vaporizes areas of the metal resistor until it meets a measurement constraint. Sheet resistance values in the KΩ/square are normal. The resistors have excellent temperature stability and long-term reliability.

3.3.2.2 Capacitors

Good-quality capacitors are required for switched-capacitor analog circuits while small high-value/area capacitors are required for dynamic memory cells. Both types of capacitors are usually added by using at least one extra layer of polysilicon, although the process techniques are very different.

Polysilicon capacitors for analog applications are the most straightforward. A second thin-oxide layer is required in order to have an oxide sandwich between the two polysilicon layers yielding a high-capacitance/unit area. Figure 3.17 shows a typical polysilicon capacitor. The presence of this second oxide can also be used to fabricate transistors. These may differ in characteristics from the primary gate oxide devices.

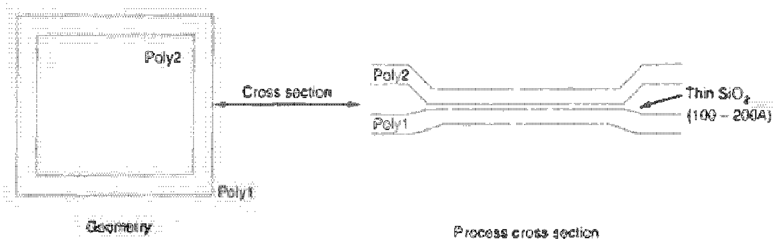


FIGURE 3.17 Polysilicon capacitor

LG Display Co., Ltd.
Exhibit 1012
Page 115

For memory capacitors, recent processes have used three dimensions to increase the capacitance/area. One popular structure is the trench capacitor, which has evolved considerably over the years to push memory densities to 64Mbits and beyond.[14] A typical trench structure is shown in Fig. 3.18(a).[15] The sides of the trench are doped n^+ and coated with a thin 10nm oxide. Sometimes oxynitride is used because its high dielectric constant increases the capacitance. The trench is filled with a polysilicon plug, which forms the bottom plate of the cell storage capacitor. This is held at $V_{DD}/2$ via a metal connection at the edge of the array. The sidewall n^+ forms the other side of the capacitor and one side of the pass transistor that is used to enable data onto the bit lines. The bottom of the trench has a p^+ plug that forms a channel-stop region to isolate adjacent capacitors. The trench is 4μm deep and has a capacitance of 90fF. Rather than building a trench, Fig. 3.18(b) shows a fin-type capacitor used in a 64-Mb DRAM.[16,17] The storage capacitance is 20 to 30fF. The fins have the additional advantage of reducing the bit capacitance by shielding the bit lines. The fabrication of 3D-process structures such as these is a constant reminder of the skill, perseverance, and ingenuity of the process engineer.

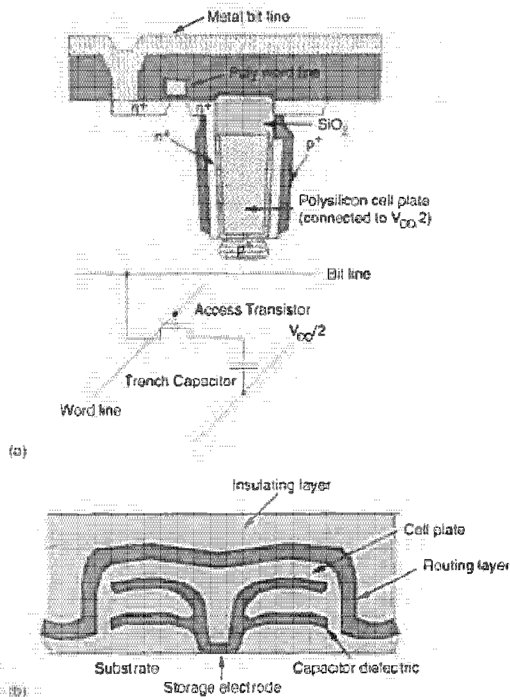


FIGURE 3.18 Dynamic memory capacitors; © IEEE 1988, © IEEE 1991.

LG Display Co., Ltd.
Exhibit 1012
Page 116

3.3.2.3 *Electrically Alterable ROM*

Frequently, electrically alterable/erasable ROM (EAROM/EEROM) is added to CMOS processes to yield permanent but reprogrammable storage to a process. This is usually added by adding a polysilicon layer. Figure 3.19 shows a typical memory structure, which consists of a stacked-gate structure.[18,19] The normal gate is left floating, while a control gate is placed above the floating gate. A very thin oxide called the tunnel oxide separates the floating gate from the source, drain, and substrate. This is usually about 10 nm thick. Another thin oxide separates the control gate from the floating gate. By controlling the control-gate, source, and drain voltages, the very thin tunnel oxide between the floating gate and the drain of the device is used to allow electrons to "tunnel" to or from the floating gate to turn the cell off or on, respectively, using Fowler-Nordheim tunneling (Section 2.2.2.5). Alternatively, by setting the appropriate voltages on the terminals, "hot electrons" can be induced to charge the floating gate, thereby programming the transistor. In non-electrically alterable versions of the technology, the process can be reversed by illuminating the gate with UV light. In these cases the chips are usually housed in glass-lidded packages. (See also Section 6.3.2).

3.3.2.4 *Bipolar Transistors*

The addition of the bipolar transistor to the device repertoire forms the basis for BiCMOS processes. Adding an npn-transistor can markedly aid in reducing the delay times of highly loaded signals, such as memory word lines and microprocessor busses. Additionally, for analog applications bipolar transistors may be used to provide better performance analog functions that MOS alone.

To get merged bipolar/CMOS functionality, MOS transistors can be added to a bipolar process or vice versa. In past days, MOS processes always had to have excellent gate oxides while bipolar processes had to have precisely controlled diffusions. A BiCMOS process has to have both.

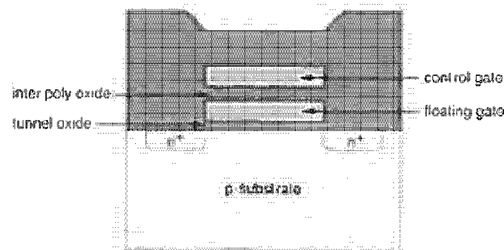


FIGURE 3.19 EEPROM technology

A mixed signal BiCMOS process[20] cross section is shown in Fig. 3.20. This process features both npn- and pnp-transistors in addition to pMOS and nMOS transistors. The major processing steps are summarized in Fig. 3.21, showing the particular device to which they correspond. The base layers of the process are similar to the process shown in Fig. 3.7. The starting material is a lightly-doped p-type substrate into which antimony or arsenic are diffused to form an n^+ buried layer. Boron is diffused to form a buried p^+ layer. An n-type epitaxial layer 4.0 μm thick is then grown. N-wells and p-wells are then diffused so that they join in the middle of the epitaxial layer. This epitaxial layer isolates the pnp-transistor in the horizontal direction, while the buried n^+ layer isolates it vertically. The npn-transistor is junction-isolated. The base for the pnp is then ion-implanted using phosphorous. A diffusion step follows this to get the right doping profile. The npn-collector is formed by depositing phosphorus before LOCOS. Field oxidation is carried out and the gate oxide is grown. Boron is then used to form the p-type base of the npn-transistor. Following the threshold adjustment of the pMOS transistors, the polysilicon gates are defined. The emitters of the npn-transistors employ polysilicon rather than a diffusion. These are formed by opening windows and depositing polysilicon. The n^+ and p^+ source/drain implants are then completed. This step also dopes the npn-emitter and the extrinsic bases of the npn- and pnp-transistors (extrinsic because this is the part of the base that is not directly between collector and emitter). Following the deposition of PSG, the normal two-layer metallization steps are completed. (*Note:* Generating the diffusions may require two distinct steps, the first being to get the impurities to the area where a diffusion is required and the second to drive the diffusion into the substrate to gain an acceptable impurity profile. These profiles have a major impact on the performance of the bipolar transistors.)

Representative of a high-density digital BiCMOS process is that represented by the cross section shown in Fig. 3.22.[21] The buried-layer–epitaxial-layer–well structure is very similar to the previous structure. However,

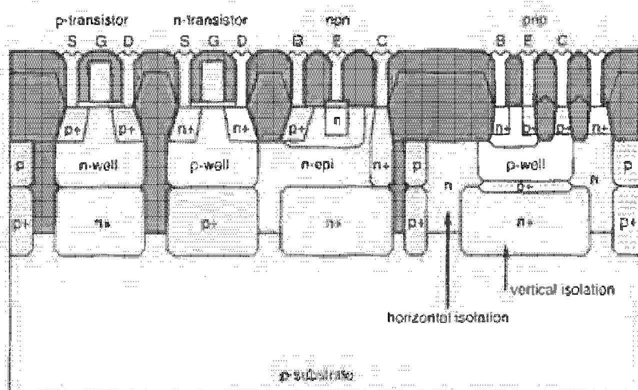


FIGURE 3.20 Typical mixed signal BiCMOS process cross section; © IEEE 1990.

LG Display Co., Ltd.
Exhibit 1012

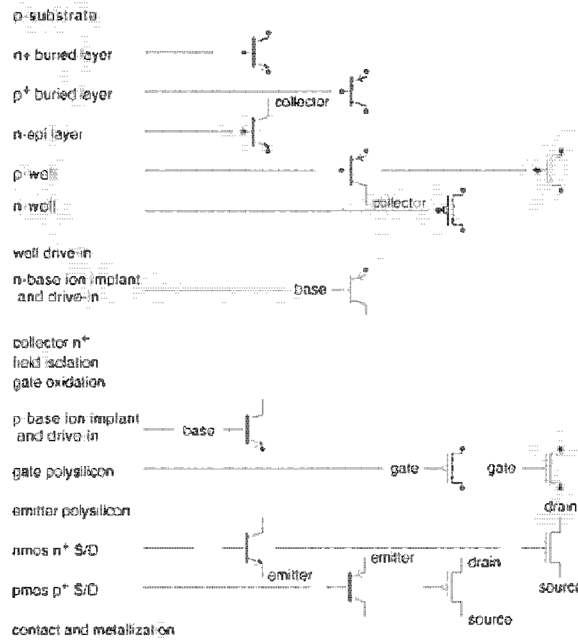


FIGURE 3.21 BiCMOS process steps for the cross section shown in Figure 3.20

because this is a 0.8μm process, LDD structures must be constructed for the p-transistors and the n-transistors. The npn is formed by a double-diffused sequence in which both base and emitter are formed by impurities that diffuse out of a covering layer of polysilicon. This process, intended for logic applications, has only an npn-transistor. The collector of the npn is connected to the n-well, which is in turn connected to the V_{DD} supply. Thus all npn-collectors are commoned. A typical npn-transistor with a 0.8μm-square emitter has a current gain of 90 and an f_t of 15 GHz.

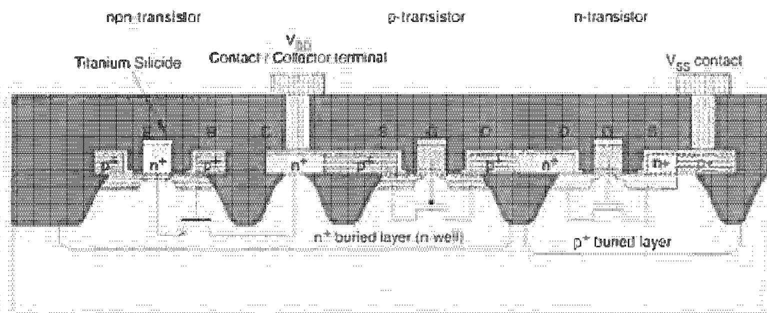


FIGURE 3.22 Digital BiCMOS process cross section; © IEEE 1991.

LG Display Co., Ltd.
Exhibit 1012
Page 119

3.3.2.5 Thin-film Transistors

A thin-film transistor has source/drain and channel regions constructed from deposited thin films of semiconductor material. Apart from SOI processes, thin-film transistors are currently used in high-density memories and in flat-panel displays, although they have been around since the early 1960s.[22,23] Those used in memories are examples of TFTs that are added to existing CMOS processes.[24]

Representative of those transistors used in memories is the p-transistor, which is shown in Fig. 3.23(a), which is used as a load transistor in a static memory cell in a high-density SRAM.[25] In this device, third-level poly forms the gate of the device, while fourth-level poly 40nm thick forms the source, drain, and channel. The channel is separated from the gate by a 40nm

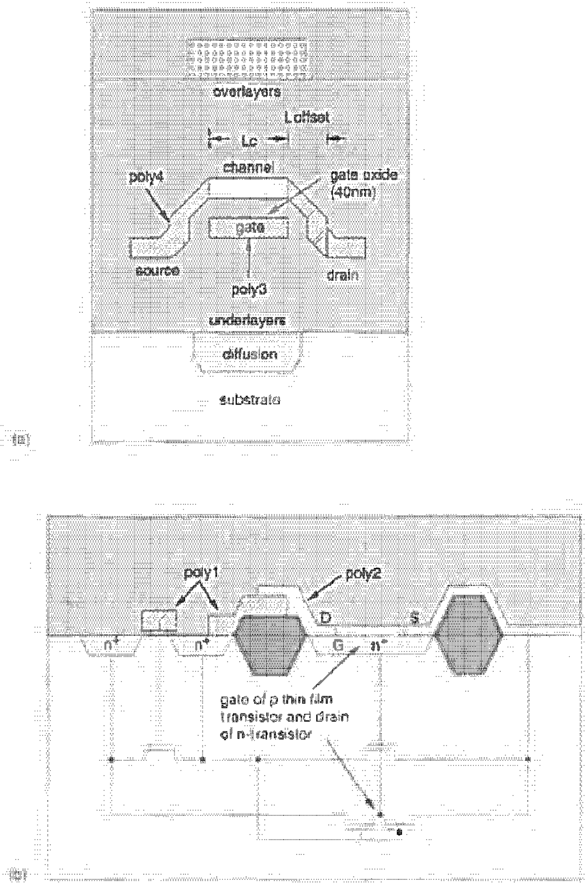


FIGURE 3.23 Examples of thin-film pMOS transistors as used in memories; © IEEE 1990.

oxide. In addition the drain is offset from the gate by the distance L_{offset}. As shown the transistor is called an "inverted staggered" thin-film transistor. The advantage of a pMOS load in memories is that the off current is of the order of 100fA compared to about 3pA for a polysilicon resistor load. For a 4Mb SRAM this results in a standby current of 0.2μA. In addition, the on current is around 10pA, which is high enough to counter any leakage current that would corrupt the data.

Another thin-film pMOS load transistor is shown[26] in Fig. 3.23(b). It is constructed from a thin film of amorphous (noncrystalline) silicon (α-silicon), 100nm thick. This film regrows crystal grains when heated to about 600°C. The larger the crystals (1–2 μm), the better the on and off characteristics of the transistor. The gate of the pMOS transistor in this instance is the source diffusion of the nMOS memory transistor. A thin gate-oxide film 40nm thick separates the "substrate" of the thin-film transistor from the gate. The pMOS transistor is 0.6μ wide and 1.4μ long. This 3D structure reduces the size of the memory cell quite considerably. As processes mature, it is highly likely that more use will be made of three-dimensional structures similar to these pMOS loads.

Thin-film CMOS transistors are also used in active-matrix LCD displays.[27] These devices have thresholds in the 4-volt range and mobilities of around 120 cm^2/Vs for the p-channel and 140 cm^2/Vs for the n-channel transistors.

3.3.3 3-D CMOS

The addition of thin-film transistors in memories effectively uses the third (vertical) dimension available on a chip. More general 3-D logic structures have been proposed and fabricated in CMOS.

One such example is shown in the process crossection in Fig. 3.24(a).[28] The substrate is an n^+ substrate upon which a p epitaxial layer is grown. Standard n-transistors are built on this epi layer with the exception of a "sinker" layer, which allows the sources of n-transistors to be down-connected to the n^+ substrate, which forms a ground plane and the V_{SS} connection. This eliminates half of the metal power wiring because V_{SS} is fed via the backside connection. A second gate oxide is grown over the n-transistor. A "seed" opening at the n-transistor drain, which allows high-quality silicon to be grown vertically and laterally, is opened. This is planarized, and a third oxide is grown on top of this epitaxially grown silicon. A polysilicon gate, used to implant self-aligned p^+ source/drains, which extend to the bottom of the epitaxially grown layer, is added on top of this structure. Planarization and metallization then are completed.

The final structure has a p- and an n-transistor with a common gate, while the p-transistor has an extra parallel gate which controls it. This basic structure allows an inverter and a 2-input multiplexer to be constructed (Fig. 3.24b). Note that there are actually two p-transistors in parallel, created

LG Display Co., Ltd.
Exhibit 1012
Page 121

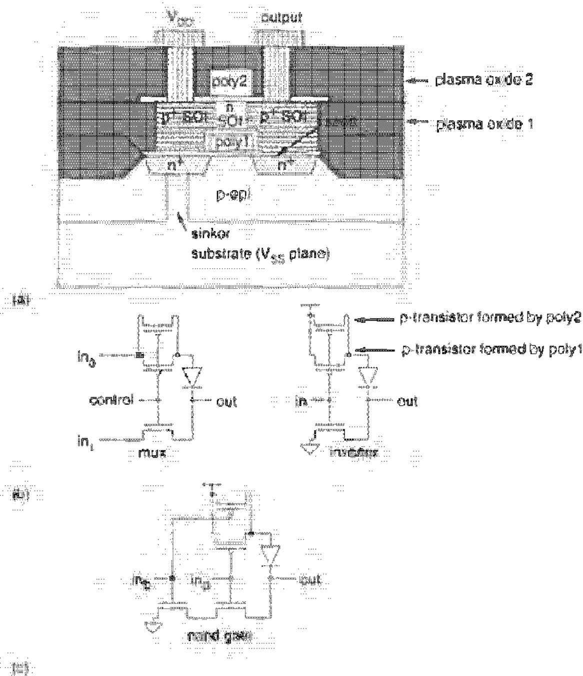


FIGURE 3.24 3-D CMOS logic technology; © IEEE 1992.

by poly1 and poly2. They both act on a common n SOI channel. By connecting the two gates together the resulting p-transistor has almost the same β as the n-transistor, thereby equalizing signal delays. By adding another series n-transistor, a 2-input NAND gate may be built (Fig. 3.24c). In the case of the inverter the p-source is connected to V_{DD}, the poly gates are commoned, and the n-source is "sinkered" to the substrate. For the 2-input multiplexer (or selector) the n- and p-sources are connected to the mux inputs while the gates are commoned to form the select line. For a NAND gate the poly gates are separately driven, as shown in Fig. 3.24(c). The diodes shown in the circuits in Fig. 3.24(b) and 3.24(c) are due to the abutting p^+ SOI drain of the p-transistor and the n^+ drain of the n-transistor. Using this novel technology, the inventors were able to design circuits that were up to 33% smaller than comparable 2-D structures.

3.3.4 Summary

This concludes the discussion of some relevant CMOS technology. Processes are constantly under development with new structures and new techniques being introduced to yield smaller, higher speed, less costly, and more

LG Display Co., Ltd.
Exhibit 1012
Page 122

reliable ICs. As a designer, you should keep abreast of CMOS technology directions because they often make previously impossible systems or ideas possible. A good forum is the annual IEEE International Electron Devices Meeting (IEDM).

3.4 Layout Design Rules

Layout rules, also referred to as *design rules,* can be considered as a prescription for preparing the photomasks used in the fabrication of integrated circuits. The rules provide a necessary communication link between circuit designer and process engineer during the manufacturing phase. The main objective associated with layout rules is to obtain a circuit with optimum yield (functional circuits versus nonfunctional circuits) in as small an area as possible without compromising reliability of the circuit.

In general, design rules represent the best possible compromise between performance and yield. The more conservative the rules are, the more likely it is that the circuit will function. However, the more aggressive the rules are, the greater the probability of improvements in circuit performance. This improvement may be at the expense of yield.

Design rules specify to the designer certain geometric constraints on the layout artwork so that the patterns on the processed wafer will preserve the topology and geometry of the designs. It is important to note that design rules do not represent some hard boundary between correct and incorrect fabrication. Rather, they represent a tolerance that ensures very high probability of correct fabrication and subsequent operation. For example, one may find that a layout that violates design rules may still function correctly, and vice versa. Nevertheless, any significant or frequent departure (*design-rule waiver*) from design rules will seriously prejudice the success of a design.

Two sets of design-rule constraints in a process relate to line widths and interlayer registration. If the line widths are made too small, it is possible for the line to become discontinuous, thus leading to an open circuit wire. On the other hand, if the wires are placed too close to one another, it is possible for them to merge together; that is, shorts can occur between two independent circuit nets. Furthermore, the spacing between two independent layers may be affected by the vertical topology of a process.

The design rules primarily address two issues: (1) the geometrical reproduction of features that can be reproduced by the mask-making and lithographical process and (2) the interactions between different layers.

There are several approaches that can be taken in describing the design rules. These include 'micron' rules stated at some micron resolution, and lambda(λ)-based rules. Micron design rules are usually given as a list of minimum feature sizes and spacings for all the masks required in a given process.

LG Display Co., Ltd.
Exhibit 1012
Page 123

For example, the minimum active width might be specified as 1 μm. This is the normal style for industry. The lambda-based design rules popularized by Mead and Conway[29] are based on a single parameter, λ, which characterizes the linear feature—the resolution of the complete wafer implementation process—and permits first-order scaling. As a rule, they can be expressed on a single page. While these rules have been successfully used for 4–1.2 μm processes, they will probably not suffice for submicron processes.

Normally, there is some minimum grid dimension in terms of which the design rules are expressed. This is a result of the economic reality that eventually the mask has to be built and the higher the lithographic tolerance, the higher the cost of the mask. Also, historically, some mask making systems had digital accuracy limitations (i.e., 16 bits of precision). At the 1.25μ–2μ level, a minimum grid unit of .2–.25μ was adequate. In submicron processes a value of .05–.1μ is more common. In this text, we will use the λ rules to illustrate principles. Normal industry practice is to deal with the micron dimensions to ensure that the circuits built are as small as possible. Contemporary CAD tools now allow designs to migrate between compatible CMOS processes without having to resort to the linear scaling that λ rules impose.

3.4.1 Layer Representations

The advances in the CMOS processes are generally complex and somewhat inhibit the visualization of all the mask levels that are used in the actual fabrication process. Nevertheless the design process can be abstracted to a manageable number of conceptual layout levels that represent the physical features observed in the final silicon wafer. At a sufficiently high conceptual level all CMOS processes use the following features:

- Two different substrates.
- Doped regions of both p- and n-transistor-forming material.
- Transistor gate electrodes.
- Interconnection paths.
- Interlayer contacts.

The layers for typical CMOS processes are represented in various figures in terms of:

- a color scheme proposed by JPL based on the Mead-Conway colors.
- other color schemes designed to differentiate CMOS structures (e.g., the colors as used on the front cover of this book)
- varying stipple patterns.
- varying line styles.

TABLE 3.1 Layer Representations for the n-well CMOS process

LAYER	COLOR	SYMBOLIC	COMMENTS
N-well	Brown		Inside brown is n-well, outside is p-type substrate.
Thin-oxide	Green	n-transistor	Thinox may not cross a well boundary.
Poly	Red	Polysilicon	Generally n^+.
p^+	Yellow	p-transistor	Inside is p^+.
Metal1	Light blue	Metal1	
Metal2	Tan	Metal2	
Contact-cut, via	Black	Contact	
Metal3	Grey	Metal3	
Overglass			

Some of these representations are shown in Table 3.1. Where diagrams are presented, a legend will be used to indicate any different layer assignments from these defaults. At the mask level, some layers may be omitted for clarity. At the symbolic level only n- and p-transistors will be shown (i.e., no wells or select layers). The symbolic representations should be viewed as translating to the appropriate set of masks for whatever process is being considered.

The p-well and twin-tub bulk CMOS processes as well as the SOI process can be represented in a similar manner. For example, in p-well bulk CMOS the only difference in the resulting wafer structure is the reversal of the role of the well and the original substrate. Different process lines may use different combinations of the n^+, p^+, n-well, or p-well masks to define the process. It is very important to intimately understand what set of masks a particular process line uses if you are responsible for generating interface formats. For instance, an n^+ mask, which is the reverse of a p^+ mask, may be used. Thus n^+ active area denotes n-transistors, and so on. Conceptually, the mask levels in a silicon-on-insulator process are probably the simplest. The levels and visible geometry in this process correspond directly to the features that a designer has to deal with conceptually (i.e., n-regions and p-regions). Perhaps the most significant difference between SOI and bulk CMOS processes, from the designer's point of view, is the absence of wells.

3.4.2 CMOS n-well Rules

In this section we describe a version of n-well rules based on the MOSIS CMOS Scalable Rules and compare those with the rules for a hypothetical (but realistic) commercial 1μ CMOS process (Table 3.2). The MOSIS rules are expressed in terms of λ. These rules allow some degree of scaling

LG Display Co., Ltd.
Exhibit 1012
Page 125

TABLE 3.2 CMOS Layout Rules

	λ RULE	λ/μ RULE (0.5μ)	μ RULE
A. N-well layer			
A.1 Minimum size	10λ	5μ	2μ
A.2 Minimum spacing (wells at same potential)	6λ	3μ	2μ
A.3 Minimum spacing (wells at different potentials)	8λ	4μ	2μ
B. Active Area			
B.1 Minimum size	3λ	1.5μ	1μ
B.2 Minimum spacing	3λ	1.5μ	1μ
B.3 N-well overlap of p^+	5λ	2.5μ	1μ
B.4 N-well overlap of n^+	3λ	1.5μ	1μ
B.5 N-well space to n^+	5λ	2.5μ	5μ
B.6 N-well space to p^+	3λ	1.5μ	3μ
C. Poly 1			
C.1 Minimum size	2λ	1μ	1μ
C.2 Minimum spacing	2λ	1μ	1μ
C.3 Spacing to Active	1λ	0.5μ	0.5μ
C.4 Gate Extension	2λ	1μ	1μ
D. p-plus/n-plus (p^+, n^+ for short)			
D.1 Minimum overlap of Active	2λ	1μ	1μ
D.2 Minimum size	7λ	3.5μ	3μ
D.3 Minimum overlap of Active in abutting contact (see Fig. 3.2.7)	1λ	0.5μ	2μ
D.4 Spacing of p^+/n^+ to n^+/p^+ gate	3λ	1.5μ	1.5μ
E. Contact			
E.1 Minimum size	2λ	1μ	0.75μ
E.2 Minimum spacing (Poly)	2λ	1μ	1μ
E.3 Minimum spacing (Active)	2λ	1μ	0.75μ
E.4 Minimum overlap of Active	2λ	1μ	0.5μ
E.5 Minimum overlap of Poly	2λ	1μ	0.5μ
E.6 Minimum overlap of Metal1	1λ	0.5μ	0.5μ
E.7 Minimum spacing to Gate	2λ	1μ	1μ
F. Metal1			
F.1 Minimum size	3λ	1.5μ	1μ
F.2 Minimum spacing	3λ	1.5μ	1μ

(continued)

TABLE 3.2 ***(continued)***

	λ RULE	λ/μ RULE (0.5μ)	μ RULE
G. Via			
G.1 Minimum size	2λ	1μ	0.75μ
G.2 Minimum spacing	3λ	1.5μ	1.5μ
G.3 Minimum Metal1 overlap	1λ	0.5μ	0.5μ
G.4 Minimum Metal2 overlap	1λ	0.5μ	0.5μ
H. Metal2			
H.1 Minimum size	3λ	1.5μ	1μ
H.2 Minimum spacing	4λ	2μ	1μ
I. Via2			
I.1 Minimum size	2λ	1μ	1μ
I.2 Minimum spacing	3λ	1.5μ	1.5μ
J. Metal3			
J.1 Minimum size	8λ	4μ	4μ
J.2 Minimum spacing	5λ	2.5μ	2.5μ
J.3 Minimum Metal2 overlap	2λ	1μ	1μ
J.4 Minimum Metal3 overlap	2λ	1μ	1μ
K. Passivation			
K.1 Minimum opening		100μ	100μ
K.2 Minimum spacing		150μ	150μ

between processes as, in principal, we only need to reduce the value of λ and the designs will be valid in the next process down in size. Unfortunately, history has shown that processes rarely shrink uniformly. Thus industry usually uses the actual micron-design rules and codes designs in terms of these dimensions, or uses symbolic layout systems to target the design rules exactly. At this time, the amount of polygon pushing is usually constrained to a number of frequently used standard cells or memories, where the effort expended is amortized over many designs. Alternatively, the designs are done symbolically, thus relieving the designer of having to deal directly with the actual design rules.

The rules are defined in terms of:

- feature sizes,
- separations and overlaps.

In addition to the rules stated above, there are various spacing rules for the periphery of the chip which frequently depend on the vendor (e.g., spacing of all layers to die boundary is 20–50μ).

LG Display Co., Ltd.
Exhibit 1012
Page 127

For each mask required in a process one needs to know whether it is "light field" or "dark field," whether light will pass through the mask to expose a photolithographic pattern or whether light will be blocked by the mask. In addition, biases are added or subtracted from the drawn dimensions of the mask to allow for varying types of processing. For instance, the active mask might be bloated to take into account the encroachment of field oxide during LOCOS. Contacts might be shrunk as etching tends to make them larger during processing. The rules in Table 3.2 are illustrated in Fig. 3.25 (and in Plate 2). The comparison between the lambda rules and micron rules reveal differences that are accentuated as process line-widths are reduced below the 1μm level. In particular, the metal widths and spacings and contact overlaps yield different pitches. For instance, the metal1 contacted pitch (contact to contact) is 4.5μ for $\lambda = 0.5\mu$ but 2.75μ for the equivalent micron rules. Thus the micron rules result in a 50% size reduction. The metal2 rules differ by 5μ to 2.75μ—almost a factor of 2. As many circuits are dominated by routing, this can translate almost directly to the final density of the circuit. On the other hand, the transistor pitch is generally determined by the contact-poly-contact pitch, which is 4μ for the λ rules and 3.25μ for the micron rules, which can also lead to significant layout density differences.

TABLE 3.3 Submicron CMOS Process Dimensions

LAYER		NEC[30]	HITACHI[31]	TOSHIBA[32]	HITACHI[33]	IBM[34]
Gate Oxide		15nm	13.5nm	11nm		7nm
Poly1	Width	.55μ (.65μ for p)	.6μ	.5μ	.3μ	.4μ
	Space	.55μ	.6μ	.6μ		
Poly2	Width	.55μ	.6μ	.5μ		
	Space	.55μ	.6μ	.6μ		
Poly3	Width	.55μ	.6μ	.8μ		
	Space	.55μ	.6μ	.7μ		
Poly4	Width		.6μ			
	Space		.6μ			
Contact	Size		.6μ	.6μ		
Metal1	Width	.9μ	.7μ	1.4μ	.3μ	
	Space	.55μ	.6μ	.7μ	.4μ	
Via	Size		.6μ	1.2μ		
Metal2	Width	.9μ	.7μ	1.4μ	.45μ	
	Space	.55μ	.6μ	1.2μ	.65μ	
Metal3	Width				.55μ	
	Space				.75μ	

LG Display Co., Ltd.
Exhibit 1012
Page 128

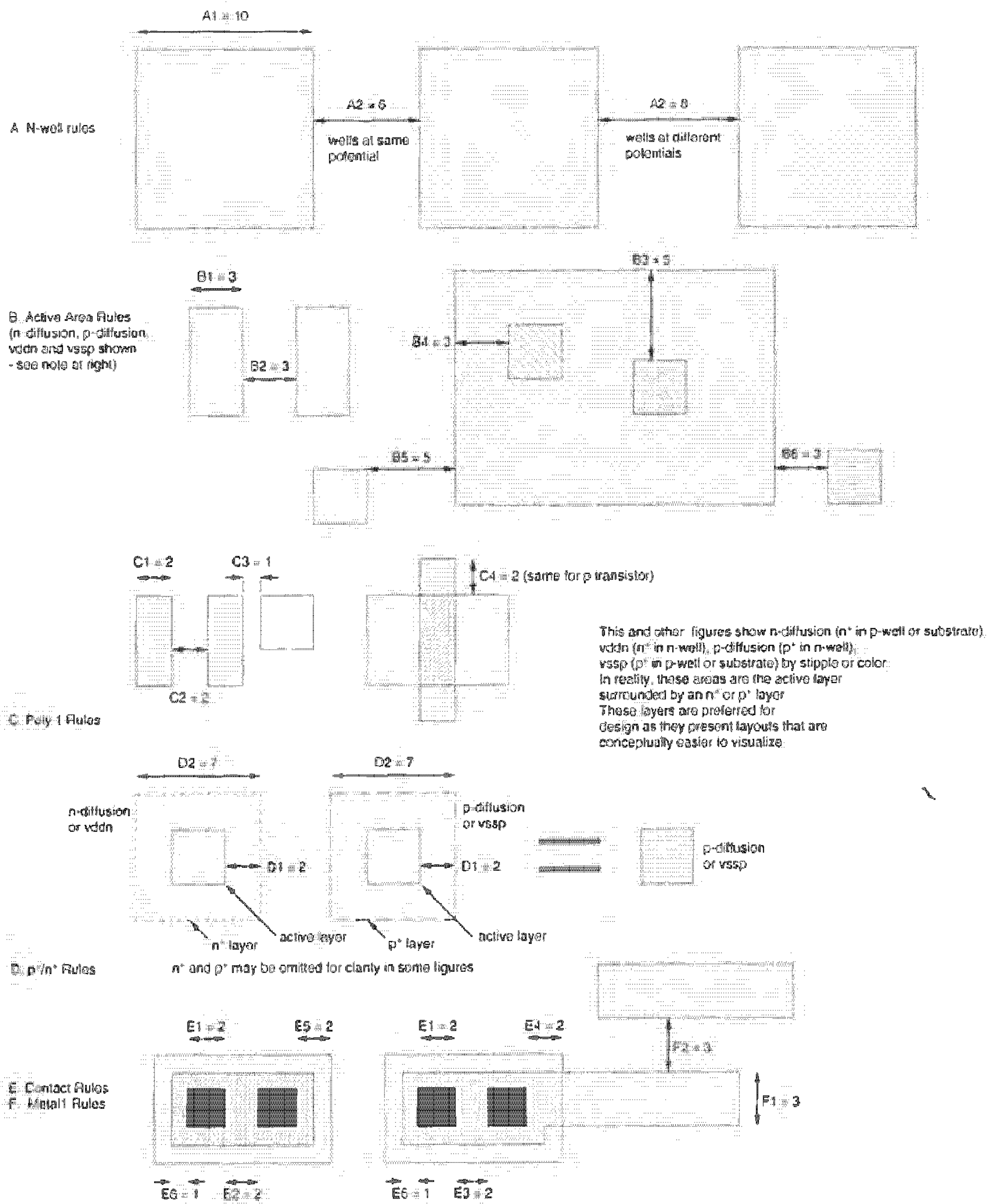


FIGURE 3.25 n-well CMOS design rules

LG Display Co., Ltd.
Exhibit 1012
Page 129

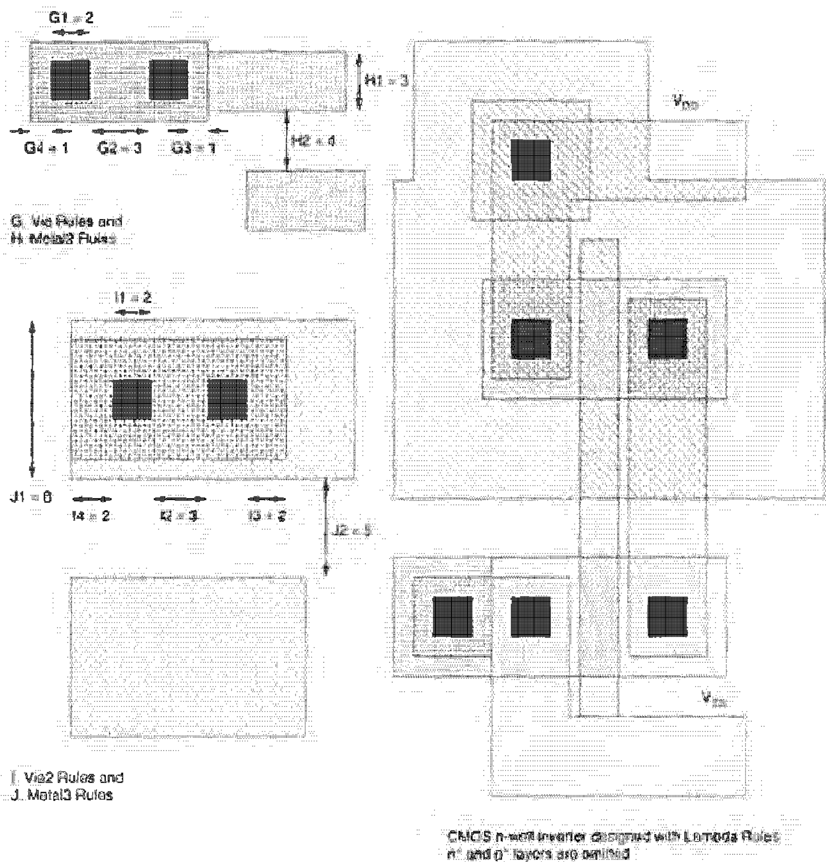


FIGURE 3.25 *(continued)*

Representative of processes in the 0.25–0.6μ range, the previous table (Table 3.3) summarizes the basic dimensions from published papers describing 4Mb static CMOS SRAMs and high-speed microprocessors. The RAM processes tend to have more poly layers (between 2 and 4) to enable small, dense memory cells to be constructed and the logic processes tend to have more metal layers (2 to 4) to improve routability.

These can be used as a guide to estimate sub 1μ technology rules. In particular the paper describing the IBM 0.25μ process in Table 3.3 provides a good overview of the considerations that go into a .25μm process.

3.4.3 Design Rule Backgrounder

In this section we will examine some of the reasons for the design rules listed above.

Well Rules: The n-well is usually a deeper implant compared with the transistor source/drain implants, therefore it is necessary for the outside dimension to provide sufficient clearance between the n-well edges and the adjacent n^+ diffusions. The inside clearance is determined by the transition of the field oxide across the well boundary. Some processes may permit zero inside clearance, but problems such as the 'birds-beaks' effect usually prevent this. A further point to be noted is that to avoid a shorted condition, active is not permitted to cross a well boundary. Since the n-well sheet resistance can be several KΩs per square, it is necessary to thoroughly ground the well. This will prevent excessive voltage drops due to substrate currents. Thus the rule to follow in grounding the n-well would be to put a substrate contact wherever space is available consistent with the rules outlined in Section 3.5.

Transistor rules: Where poly crosses active, the source and drain diffusion is masked by the poly region. The source, drain, and channel are thereby self-aligned to the gate. It is essential for the poly to completely cross active, otherwise the transistor that has been created will be shorted by a diffused path between source and drain. To ensure this condition is satisfied, poly is required to extend beyond the edges of the diffusion region. This is often termed the "gate extension." This effect is shown in Fig. 3.26(a) where the diffusion has increased in size and the poly has been overetched, resulting in a short. The thin oxide must extend beyond the poly gate so that diffused regions exist to carry charge into and out of the channel (Fig. 3.26b). Poly and active regions that do not meet intentionally to form a transistor should be kept separated. Both types of transistors have an active region (diffusion or implant) and a polysilicon region. A p-device has an n-well region surrounding it, whereas an n-device has an n^+ (n-plus) region surrounding it. Thin oxide areas that are not covered by n are p^+ and hence are p-devices or wires (within the n-well). Therefore a transistor is n-channel if it is inside an n^+ region; otherwise it is a p-channel device. From the above discussion it can be noted that there are two types of implant/diffusion used to form the p- and n-transistors. What is important to note is that n^+ diffusion is obtained by "logical anding" of active and n^+ (n-plus) masks, whereas p^+ diffusion is derived by "logical anding" of active and (NOT n^+) masks. Frequently, in order to simplify design the n-plus and/or p-plus masks are ignored during design and inserted automatically. A problem can occur if the orthogonal distance of n^+ (n-plus) to p^+ (p-plus) is used (Rule B.3 + B.5 for instance or B.4 + B.6). While the select layers may be added without problems for orthogonally spaced structures, diagonally positioned

LG Display Co., Ltd.
Exhibit 1012
Page 131

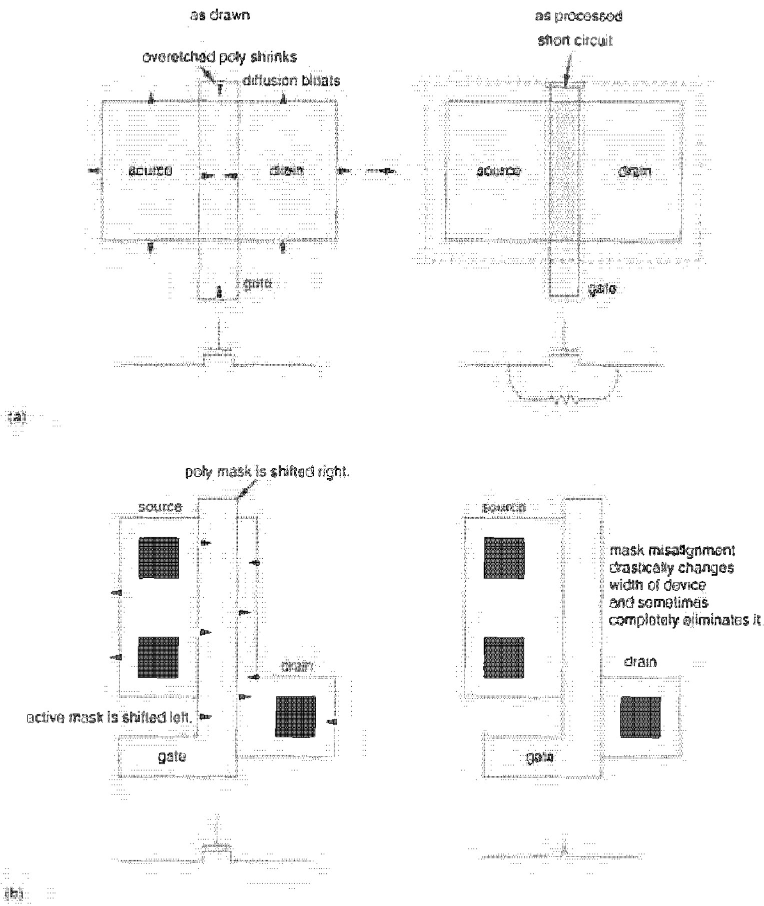


FIGURE 3.26 Effects of insufficient gate extension and source-drain extension

diffusions may violate the n-plus–p-plus spacing rules. In symbolic layout systems this frequently leads to a second set of spacings that describe diagonal constraints.

Contact Rules: There are several generally available contacts:

- Metal to p-active (p-diffusion).
- Metal to n-active (n-diffusion).
- Metal to polysilicon.
- V_{DD} and V_{SS} substrate contacts.
- Split (substrate contacts).

LG Display Co., Ltd.
Exhibit 1012
Page 132

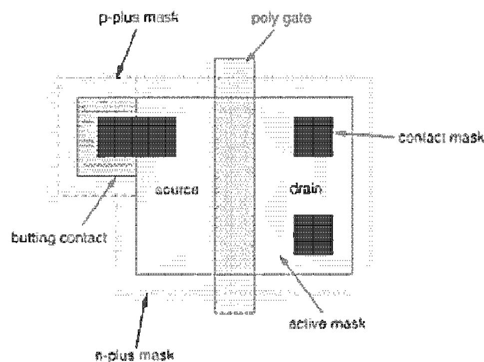


FIGURE 3.27 Structure of a merged or abutting substrate contact

Depending on the process, other contacts such as "buried" polysilicon-active contacts may be allowed. This contact allows direct connection between polysilicon and the active transistor region. Sometimes this type of contact is allowed to only one type of active area.

Because the substrate is divided into "well" regions, each isolated well must be "tied" to the appropriate supply voltage; that is, the n-well must be tied to V_{DD} and the substrate (what amounts to a p-well) must be tied to V_{SS}. This is achieved by the use of well or substrate contacts. One needs to note that every p-device must be surrounded by an n-well and that the n-well must be connected to V_{DD} via a V_{DD} contact. Furthermore, every n-device must have access to a V_{SS} contact. The split or merged contact is equivalent to two separate metal-diffusion contacts that are strapped together with metal (Fig. 3.27). This structure is used to tie transistor sources to either the substrate or the n-well. A version is also shown at the source of the n-transistor in the inverter in Fig. 3.25. Separate contacts are shown; this is consistent with modern processes, which usually require uniform contact sizes to achieve well-defined etching characteristics. Merged contact structures in older processes may have used an elongated contact rectangle (Fig. 3.27). The V_{SS} or V_{DD} merged contacts may be inset into the source of the corresponding n-transistor where wide transistors are employed. An alternative separated contact structure is shown for the V_{DD} contact for the p-transistor in Fig. 3.25. Here the n^+ well contact is separated from the p^+ source/drain diffusion.

Guard Rings: Guard rings that are p^+ diffusions in the p-substrate and n^+ diffusions in the n-well are used to collect injected minority carriers. If they are implemented in a structure, then n^+ guard rings must be tied to V_{DD}, while p^+ guard rings must be tied to V_{SS}. A p^+ diffusion with n^+ guard ring is shown in Fig. 3.28(a), while an n^+ diffusion with p^+ guard ring is shown in

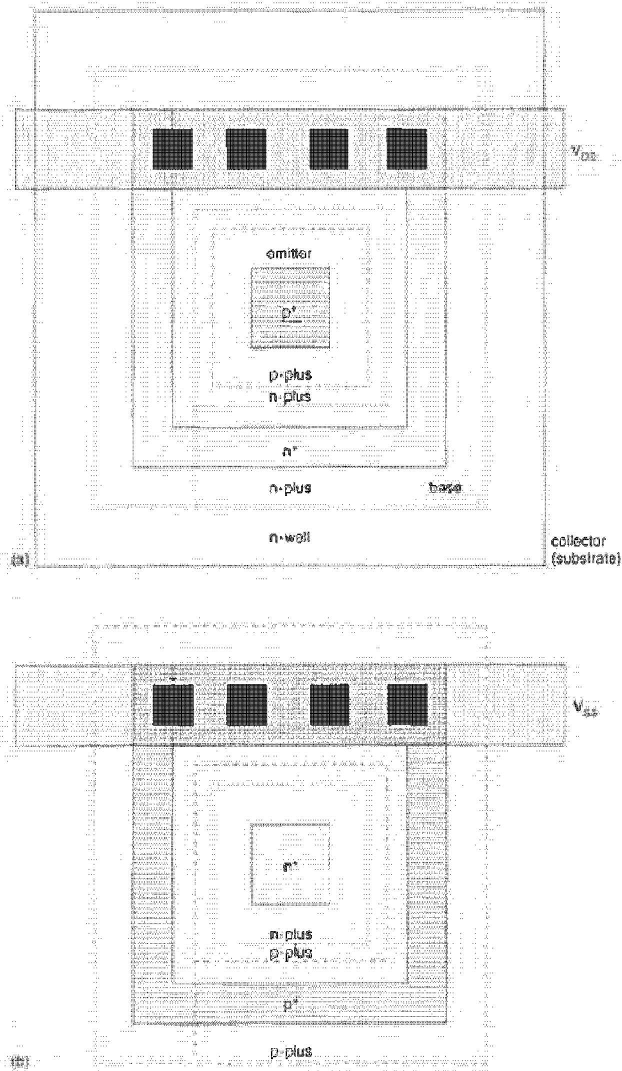


FIGURE 3.28 Guard rings

Fig. 3.28(b). Different well-enclosure rules may apply for guard-ring structures. The reason for guard rings will become more clear in Section 3.5. Incidentally, the structure shown in Fig. 3.28(a) is also that for a pnp transistor if one was required. The transistor terminals have been marked. The area of the center p^+ region is the area of the emitter. The base is the n-well and is connected via the n^+ ring. The collector is the substrate.

LG Display Co., Ltd.
Exhibit 1012
Page 134

Metal Rules: Metal spacings may vary with the width of the metal line (so-called *fat-metal* rules). That is, at some width, the metal spacing may be increased. This is due to etch characteristics of small versus large metal wires. There may also be maximum-metal-width rules. Additionally, there may be rules that are applied to long closely spaced parallel metal lines. Some processes require a certain proportion of the chip area to be covered with metal, and in such cases metal might have to be added to chip "white space" (assuming there is some!). These rules usually relate to constraints imposed by manufacturability requirements.

Via Rules: Processes may vary in whether they allow vias to be placed over polysilicon and diffusion regions. Some processes allow vias to be placed within these areas but do not allow the vias to straddle the boundary of polysilicon or diffusion. This results from the sudden vertical topology variations that occur at sublayer boundaries.

Metal2 Rules: The possible increase in width and separation of second-level metal are conservative rules to ensure against broken conductors or shorts between adjoining wires due to the vertical topology. Modern processes frequently have the metal1 and metal2 pitches identical.

Via2 Rules: Similarly to first vias, the rules for placement of via2 may vary with process.

Metal3 Rules: These rules usually but not always increase in width and separation over metal2. Metal3 is generally used primarily for power-supply connections and clock distribution.

Some additional rules that might be present in some processes are as follows:

- Extension of polysilicon in the direction that metal wires exit a contact.
- Differing p- and n-transistor gate lengths.
- Differing gate poly extensions, depending on the device length or the device construction.

Whereas earlier processes tended to be process driven and frequently have long and involved design rules, increasingly more processes have become "designer friendly" or more specifically computer friendly because most of the mask geometries for designs are algorithmically produced. Also, system companies have created "generic" rules that span a number of different CMOS foundries that they might use. Some processes have design guidelines that feature structures to be avoided to ensure good yields. In general

LG Display Co., Ltd.
Exhibit 1012
Page 135

though, at this time, process technology is so well developed, features so small, and time to market so short that the traditional yield improvement cycle is only done for the highest volume parts. Frequently, the technology changes so fast that it is better to reimplement the circuit in the new smaller technology than worry about improving the yield on the older larger process. Of course at some time, a limit will come to how small technologies can be made and then a return to classical yield optimization will probably resurface.

Passivation or Overglass: This is a protective glass layer that covers the final chip. Openings are required at pads and any internal test points.

3.4.4 Scribe Line

The scribe line is a specifically designed structure that surrounds the completed chip and is the point at which the chip is cut with a diamond saw. The construction of the scribe line varies from manufacturer to manufacturer.

3.4.5 Layer Assignments

Table 3.4 lists the MOSIS Scalable CMOS Design-rule layer assignments for the Caltech Intermediate Form (CIF) language and Calma stream format.

TABLE 3.4 MOSIS Scalable CMOS Design-rule Layer Assignments

LAYER	CIF LAYER NAME	CALMA NUMBER
Well	CWG	14
N-well	CWN	1
P-well	CWP	2
Active	CAA	3
Select	CSG	15
P-select	CSP	8
N-select	CSN	7
Poly	CPG	4
Poly Contact	CCP	45
Poly 2 (Electrode)	CEL	5
Electrode Contact	CCE	55
Active Contact	CCA	35
Metal1	CMF	10
Via	CVA	11
Metal2	CMS	12
Via2*	CVB	65
Metal3*	CMT	14
Overglass	COG	13

*Author's assignment

LG Display Co., Ltd.
Exhibit 1012
Page 136

3.4.6 SOI Rules

Usually SOI rules closely follow bulk CMOS rules except that n^+ and p^+ regions can abut. This allows some interesting multiplexer and latch circuits. A spacing rule between island edge and unrelated poly is used to ensure against shorts between the poly and island edges. This can be caused by thin or faulty oxide covering over the islands.

3.4.7 Design Rules—Summary

In commercial designs, λ rules are rarely sufficient to describe high-density, high-performance circuits. While all of these rules can be worst-cased, very inefficient designs result. A better approach is to implement systems that synthesize the correct geometry from an intermediate form. Therefore, symbolic styles of design provide a solution for creating generic CMOS circuits that can be implemented with a wide range of fabrication processes.

3.5 Latchup

If every silver lining has a cloud, then the cloud that has plagued CMOS is a parasitic circuit effect called "latchup." The result of this effect is the shorting of the V_{DD} and V_{SS} lines, usually resulting in chip self-destruction or at least system failure with the requirement to power down. This effect was a critical factor in the lack of acceptance of early CMOS processes, but in current processes it is controlled by process innovations and well-understood circuit techniques.

3.5.1 The Physical Origin of Latchup

The source of the latchup effect[35,36,37] may be explained by examining the process cross section of a CMOS inverter, shown in Fig. 3.29(a), on which is overlaid an equivalent circuit. The schematic depicts, in addition to the expected nMOS and pMOS transistors, a circuit composed of an npn-transistor, a pnp-transistor, and two resistors connected between the power and ground rails (Fig. 3.29b). Under the right conditions, this parasitic circuit has the VI characteristic shown in Fig. 3.29(c), which indicates that above some critical voltage (known as the trigger point) the circuit "snaps" and draws a large current while maintaining a low voltage across the terminals (known as the holding voltage). This is, in effect, a short circuit. As mentioned, the bipolar devices and resistors shown in Fig. 3.29(b) are parasitic, that is, an unwanted byproduct of producing pMOS and nMOS transistors. Further

LG Display Co., Ltd.
Exhibit 1012
Page 137

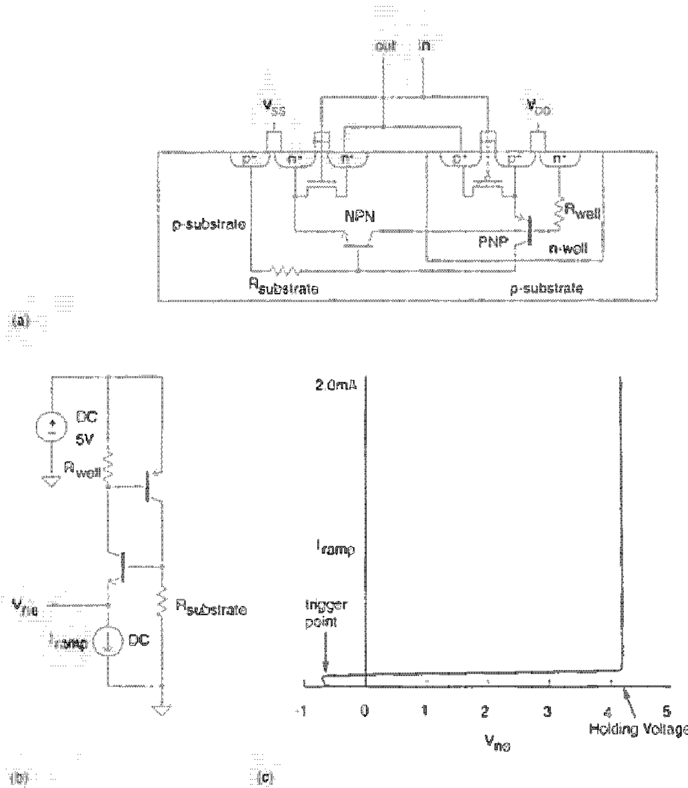


FIGURE 3.29 The origin, model, and VI characteristics of CMOS Latchup

examination of Fig. 3.29(a) reveals how these devices are constructed. The figure shows a cross-sectional view of a typical (n-well) CMOS process. The (vertical) pnp-transistor has its emitter formed by the p^+ source/drain implant used in the pMOS transistors. Note that either the drain or source may act as the emitter although the source is the only terminal that can maintain the latchup condition. The base is formed by the n-well, while the collector is the p-substrate. The emitter of the (lateral) npn-transistor is the n^+ source/drain implant, while the base is the p-substrate and the collector is the n-well. In addition, substrate resistance $R_{substrate}$ and well resistance R_{well} are due to the resistivity of the semiconductors involved.

Consider the circuit shown in Fig. 3.29(b). If a current is drawn from the npn-emitter, the emitter voltage becomes negative with respect to the base until the base emitter voltage is approximately 0.7 volts. At this point the npn-transistor turns on and a current flows in the well resistor due to common emitter current amplification of the npn-transistor. This raises the base

emitter voltage of the pnp-transistor, which turns on when the pnp $V_{be} = -0.7$ volts. This in turn raises the npn base voltage causing a positive feedback condition, which has the characteristic shown in Fig. 3.29(c). At a certain npn-base-emitter voltage, called the *trigger point*, the emitter voltage suddenly "snaps back" and enters a stable state called the ON state. This state will persist as long as the voltage across the two transistors is greater than the holding voltage shown in the figure. As the emitter of the npn is the source/drains of the n-transistor, these terminals are now at roughly 4 volts. Thus there is about 1 volt across the CMOS inverter, which will most likely cause it to cease operating correctly. The current drawn is usually destructive to metal lines supplying the latched up circuitry.

3.5.2 Latchup Triggering

For latchup to occur, the parasitic npn-pnp circuit has to be triggered and the holding state has to be maintained. Latchup can be triggered by transient currents or voltages that may occur internally to a chip during power-up or externally due to voltages or currents beyond normal operating ranges. Radiation pulses can also cause latchup. Two distinct methods of triggering are possible, lateral triggering and vertical triggering.

Lateral triggering occurs when a current flows in the emitter of the lateral npn-transistor. The static trigger point is set by[38]

$$I_{ntrigger} \approx \frac{V_{pnp\text{-}on}}{\alpha_{npn} R_{well}}, \tag{3.1}$$

where

$V_{pnp\text{-}on} \sim 0.7$ volts—the turn-on voltage of the vertical pnp-transistor
α_{npn} = common base gain of the lateral npn-transistor
R_{well} = well resistance.

Vertical triggering occurs when a sufficient current is injected into the emitter of the vertical-pnp transistor. Similar to the lateral case, this current is multiplied by the common-base-current gain, which causes a voltage drop across the emitter base junction of the npn transistor due to the resistance, $R_{substrate}$. When the holding or sustaining point is entered, it represents a stable operating point provided the current required to stay in the state can be maintained.

Current has to be injected into either the npn- or pnp-emitter to initiate latchup. During normal circuit operation in internal circuitry this may occur due to supply voltage transients, but this is unlikely. However, these condi-

LG Display Co., Ltd.
Exhibit 1012
Page 139

tions may occur at the I/O circuits employed on a CMOS chip, where the internal circuit voltages meet the external world and large currents can flow. Therefore extra precautions need to be taken with peripheral CMOS circuits. Figure 3.30(a) illustrates an example where the source of an nMOS output transistor experiences undershoot with respect to V_{SS} due to some external circuitry. When the output dips below V_{SS} by more than 0.7V, the drain of the nMOS output driver is forward biased, which initiates latchup. The complementary case is shown in Fig. 3.30(b) where the pMOS output transistor experiences an overshoot more than 0.7V beyond V_{DD}. Whether or not in these cases latchup occurs depends on the pulse widths and speed of the parasitic transistors.[39]

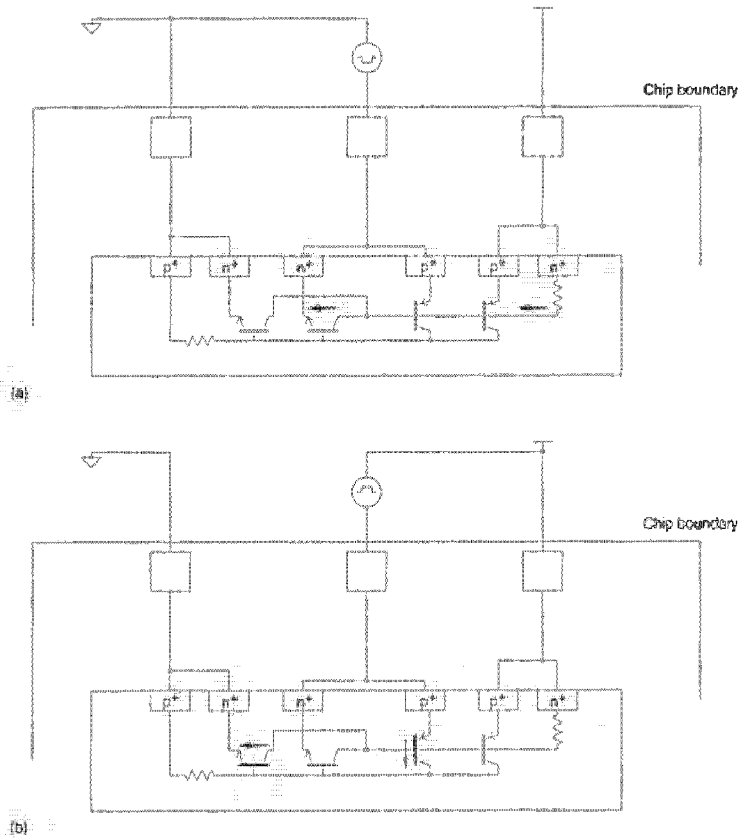


FIGURE 3.30 Externally induced latchup

3.5.3 Latchup Prevention

For latchup to occur an analysis of the circuit in Fig. 3.29(b) finds the following inequality has to be true[40]:

$$\beta_{npn}\beta_{pnp} > 1 + \frac{(\beta_{npn} + 1)\,(I_{Rsubstrate} + I_{Rwell}\beta_{pnp})}{I_{DD} - I_{Rsubstrate}}, \tag{3.2}$$

where

$$I_{Rsubstrate} = \frac{V_{be\,npn}}{R_{substrate}}$$

$$I_{Rwell} = \frac{V_{be\,pnp}}{R_{well}}$$

I_{DD} = total supply current.

This equation yields the keys to reducing latchup to the point where it should never occur under normal circuit conditions. Thus, reducing the resistor values and reducing the gain of the parasitic transistors are the basis for eliminating latchup.

Latchup may be prevented in two basic ways:

- Latchup resistant CMOS processes.
- Layout techniques.

A popular process option that reduces the gain of the parasitic transistors is the use of silicon starting-material with a thin epitaxial layer on top of a highly doped substrate. This decreases the value of the substrate resistor and also provides a sink for collector current of the vertical pnp-transistor. As the epi layer is thinned, the latchup performance improves until a point where the up-diffusion of the substrate and the down-diffusion of any diffusions in subsequent high-temperature procession steps thwart required device doping profiles. The so-called retrograde well structure is also used. This well has a highly doped area at the bottom of the well, whereas the top of the well is more lightly doped. This preserves good characteristics for the pMOS (or nMOS in p-well) transistors but reduces the well resistance deep in the well. A technique linked to these two approaches is to increase the holding voltage above the V_{DD} supply. This guarantees that latchup will not occur.

It is hard to reduce the betas of the bipolar transistors to meet the condition set above. Nominally, for a 1μ n-well process, the vertical pnp has a

beta of 10–100, depending on the technology. The lateral npn-current-gain, which is a function of n^+ drain to n-well spacing, is between 2 and 5.[41] (These values are illustrative—they should be checked for the particular process being used.)

Apart from the inherent resistance to latchup of a particular process, there are a number of well-proven techniques to design CMOS layouts that are latchup resistant.

3.5.4 Internal Latchup Prevention Techniques

From Fig. 3.29(b) it may be seen that the emitter of the npn-transistor has to an nMOS transistor source returned to V_{SS}. The substrate resistor occurs between this emitter and the supply represented by a substrate contact. Clearly, if the n-transistor source is shorted to the p^+ substrate contact, much has been done to reduce $R_{substrate}$. Conversely, the well resistor occurs between the p^+ source nominally to V_{DD} and the n^+ well contact. Thus a key technique to reduce latchup is to make good use of substrate and well contacts.

In most current processes the possibility of latchup occurring in internal circuitry has been reduced to the point where a designer need not worry about the effect as long as *liberal* substrate contacts are used. The definition of "liberal" is usually acquired from designers who have completed successful designs through a given process. Modeling the parasitics is possible,[42] but the actual switching transients existent in the circuit have a great effect on any possible latchup condition. A few rules may be followed that reduce the possibility of internal latchup to a very small likelihood:

- Every well must have a substrate contact of the appropriate type.
- Every substrate contact should be connected to metal directly to a supply pad (i.e., no diffusion or polysilicon underpasses in the supply rails).
- Place substrate contacts as close as possible to the source connection of transistors connected to the supply rails (i.e., V_{SS} n-devices, V_{DD} p-devices). This reduces the value of $R_{substrate}$ and R_{well}. A very conservative rule would place one substrate contact for every supply (V_{SS} or V_{DD}) connection.
- Otherwise a less conservative rule is place a substrate contact for every 5–10 transistors or every 25–100μ.
- Lay out n- and p-transistors with packing of n-devices toward V_{SS} and packing of p-devices toward V_{DD} (see layout styles in Chapter 5). Avoid "convoluted" structures that intertwine n- and p-devices in checkerboard styles (unless you are designing in SOI which is latchup free).

LG Display Co., Ltd.
Exhibit 1012
Page 142

3.5.5 I/O Latchup Prevention

Reducing the gain of the parasitic transistors is achieved through the use of guard rings (first encountered in Fig. 3.28). A p^+ guard ring is shown in Fig. 3.31(a) for an n^+ source/drain, while Fig. 3.31(b) shows an n^+ guard ring for a p^+ source/drain. As shown in the figures, these guard bands act as "dummy-collectors" and spoil the gain of the parasitic transistors by collecting minority carriers and preventing them from being injected into the respective bases. Carriers can still flow underneath these structures which leads sometimes to double guard banding which is illustrated in Fig. 3.31(c). While these techniques can be used on internal structures, the area penalty is usually too high except for applications such as space-borne electronics where radiation induced latchup must be avoided at all costs.

Luckily enough, as has been observed, the most likely place for latchup to occur is in I/O structures where large currents flow, large parasitics may be present, and abnormal circuit voltages may be encountered. Here the area penalty of guard rings is not at all significant. In these structures two options can be taken. The first is to use proven I/O structures designed by experts who understand the process at a detailed level. Second, rules may be applied to the design of these structures that minimize the possibility of latchup. Typical rules (n-well process) include:

- Physically separate the n- and p-driver transistors (i.e., with the bonding pad).

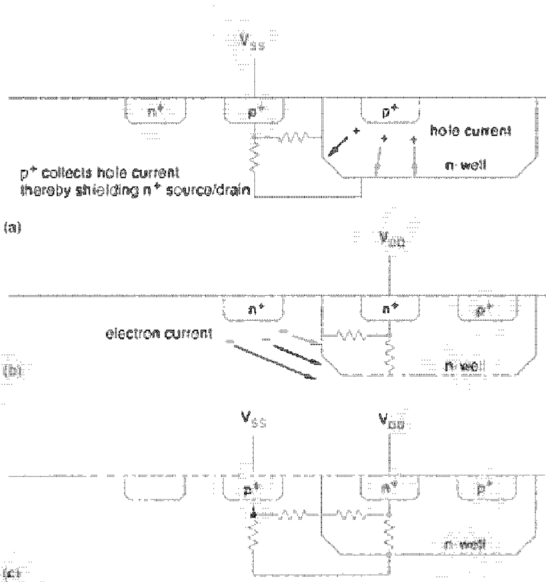


FIGURE 3.31 The use of dummy collectors to reduce latchup

LG Display Co., Ltd.
Exhibit 1012
Page 143

- Include p^+ guard rings connected to V_{SS} around n-transistors.
- Include n^+ guard rings connected to V_{DD} around p-transistors.
- Source diffusion regions of the n-transistors should be placed so that they lie along equipotential lines when current flows between V_{SS} and the p-wells; that is, source fingers should be perpendicular to the dominant direction of current flow rather than parallel to it. This reduces the possibility of latchup through the n-transistor source, due to an effect called "field aiding."[43]
- Shorting n-transistor source regions to the substrate and the p-transistor source regions to the n-well with metallization along their entire lengths will aid in preventing either of these diodes from becoming forward-biased, and hence reduces the contribution to latchup from these components.
- The n-well should be hard-wired (via n^+) to power so that any injected charge is diverted to V_{DD} via a low-resistance path. The n-well has a relatively high sheet-resistance and is susceptible to charge injection.
- The spacing between the n-well n^+ and the p-transistor source contact should be kept to a minimum. This allows minority carriers near the parasitic pnp-transistor emitter-base junction to be collected, and reduces R_{well}. The rules for the 1μ process suggest one contact for every 10μ–50μ.
- The separation between the substrate p^+ and the n-transistor source contact should be minimized. This results in reduced minority carrier concentration near the npn-emitter-base junction. Similar spacings to those suggested above apply for processes in the 1μ range.

More details on layout and design techniques for I/O circuitry may be found in Chapter 5.

3.6 Technology-related CAD Issues

The mask database is the interface between the semiconductor manufacturer and the chip designer. Two basic checks have to be completed to ensure that this description can be turned into a working chip. First, the specified geometric design rules must be obeyed. Second, the interrelationship of the masks must, on passing through the manufacturing process, produce the correct interconnected set of circuit elements. To check these two requirements, two basic CAD tools are required, namely a Design Rule Check (DRC) program and a mask circuit-extraction program. The most common approach to implementing these tools is to provide a set of subprograms that perform

LG Display Co., Ltd.
Exhibit 1012
Page 144

general geometry operations. A particular set of DRC rules or extraction rules for a given CMOS process (or any semiconductor process) is then specified by a specification of the operations that must be performed on each mask and the intermask checks that must be completed. Accompanied by a written specification, these *run-sets* are usually the defining specification for a process.

In this section we will examine a hypothetical DRC and extraction system to illustrate the nature of these run-sets.

3.6.1 DRC—Spacing and Dimension Checks

Although we might design the physical layout of a certain set of mask layers, the actual masks used in fabrication are derived from the original specification. Similarly, when we want a program to determine what we have designed by examining the interrelationship of the various mask layers, it may be necessary to determine various logical combinations between masks.

To examine these concepts, let us posit the existence of the following functions (loosely based on the CADENCE DRACULA DRC program[44]), which we will apply to a geometric database (i.e., rectangles, polygons, paths):

- **AND** `layer1 layer2 -> layer3`
  ANDs `layer1` and `layer2` together to produce `layer3` (i.e., the intersection of the two input mask descriptions)
- **OR** `layer1 layer2 -> layer3`
  ORs `layer1` and `layer2` together to produce `layer3` (i.e., the union of the two input mask descriptions)
- **NOT** `layer1 layer2 -> layer3`
  Subtracts `layer2` from `layer1` to produce `layer3` (i.e., the difference of the two input mask descriptions)
- **WIDTH** `layer > dimension -> layer3`
  Checks that all geometry on `layer` is larger than `dimension`. Any that is not is placed in `layer3`
- **SPACE** `layer > dimension -> layer3`
  Checks that all geometry on `layer` is spaced further than `dimension`. Any that is not is placed in `layer3`

The following layers will be assumed as input:

```
nwell
active
pplus
nplus
```

```
poly
poly-contact
active-contact
metal
```

Typically, useful sublayers are first generated. First, the four kinds of active area are isolated. The set of rules to accomplish this is as follows:

```
NOT all nwell -> substrate
AND nwell active -> nwell-active
NOT active nwell -> pwell-active
AND nwell-active pplus -> pdiff
AND nwell-active nplus -> vddn
AND pwell-active nplus -> ndiff
AND pwell-active pplus -> vssp
```

In the above specification a number of new layers have been specified. For instance, the first rule states that wherever `nwell` is absent, a layer called `substrate` exists. The second rule states that all active areas within the nwell are `nwell-active`. A combination of `nwell-active` and `pplus` or `nplus` yields `pdiff` (p diffusion) or `vddn` (well tie).

To find the transistors, the following set of rules is used:

```
AND poly ndiff -> ngates
AND poly pdiff -> pgates
```

The first rule states that the combination of polysilicon and n diffusion yields the `ngates` region—all of the n-transistor gates.

Typical design rule checks (DRC) might include the following :

```
WIDTH metal > 1.25 -> metal-width-error
SPACE metal > 1.0 -> metal-space-error
```

For instance the first rule determines if any metal is narrower than 1.25μ and places the errors in the `metal-width-error` layer. This layer might then later be used with the original and an interactive mask editor to identify the errors.

A bloat command changes the dimensions of a layer.

- **`BLOAT`** `layer1 dimension -> layer2`
  Expand or contract `layer1` by `dimension` to produce `layer2`.

For instance

```
BLOAT metal 0.5 metal-exp
```

would create a layer `metal-exp` in which all metal geometries were increased in size peripherally by 0.5μ. Bloats and shrinks may be used to derive other required layers. For instance, if the gates of all p-transistors had

LG Display Co., Ltd.
Exhibit 1012
Page 146

to be increased in length by 0.5μ, the following sequence might be used:

```
BLOAT pgates 0.25 pgates-bloat
```

The following sequence produces the `nplus` layer from an original specification containing only `ndiff` (n-transistors) and `vddn` (V_{DD} substrate ties).

```
AND ndiff vddn all-ndiff
BLOAT all-ndiff 2 nplus
```

3.6.2 Circuit Extraction

Now imagine that we wish to determine the electrical connectivity of a mask database. The following commands are required:

- **`CONNECT`** `layer1 layer2`
  Electrically connect `layer1` and `layer2`
- **`MOS`** `name drain-layer gate-layer source-layer substrate-layer`
  Define an MOS transistor in terms of the component terminal layers. (This is admittedly, a little bit of magic.)

The connections between layers may be specified as follows:

```
CONNECT active-contact pdiff
CONNECT active-contact ndiff
CONNECT active-contact vddn
CONNECT active-contact vssp
CONNECT active-contact metal
CONNECT vssp substrate
CONNECT vddn nwell
CONNECT poly-contact poly
CONNECT poly-contact metal
```

The connections between the diffusions and the metal are specified by the first seven statements. The last two statements specify how the metal is connected to the poly.

Finally, the active devices are specified in terms of the layers that we have derived.

```
MOS nmos ndiff ngates ndiff substrate
MOS pmos pdiff pgates pdiff nwell
```

An output statement might then be used to output the extracted transistors in some netlist format (i.e., SPICE format). This is then used as an interface to a program that compares the connectivity that we have derived from the mask with that of, say, a circuit diagram.

LG Display Co., Ltd.
Exhibit 1012
Page 147

It is important to realize that the above run set is manually generated. The data extracted from such a program is only as good as the input. For instance, if parasitic routing capacitances are required, then each and every layer interaction must be coded. If parasitic resistance is important in determining circuit performance, it too must be specifically included in the extraction run set. Many different coding styles exist that define the abstract layers in which the designer conceives the layout. For instance, if there are different rules that specify a well overlap for a guard structure compared with an internal structure, then a special guard layer might have to be coded in the mask database. Similar decisions have to be made concerning structures, such as resistors, that are constructed from diffusion or polysilicon.

3.7 Summary

This chapter has covered some of the more common CMOS technologies that are in current use. A representative set of n-well design rules have been introduced. These form the interface between the designer and the manufacturer. A range of process options were discussed to enhance the basic CMOS process. The important condition known as latchup has been introduced along with necessary design rules to avoid this condition in CMOS chips. Finally, some of the CAD/process interface issues were surveyed.

3.8 Exercises

1. A p-well process has the following layers:

   p-well
   active
   n-plus
   p-plus
   poly
   contact
   metal

   Draw the mask combinations for the following:

   a p-transistor
   an n-transistor
   a V_{SS} contact

LG Display Co., Ltd.
Exhibit 1012
Page 148

a V_{DD} contact
a contact to an n-transistor source/drain
a single guard-ringed n-transistor
a double guard-ringed p-transistor

Use the design rules from Table 3.2 as appropriate.

2. Write a program that can generate a single metal CMOS inverter in an n-well technology that parametizes the widths of the p/n transistors. Use the design rules in Table 3.2.
3. Explain how the parasitic channel, which couples unrelated nMOS transistors in an n-well process, is reduced.
4. How might you use a field transistor to prevent overvoltage in a CMOS chip?
5. Explain why substrate and well contacts are important in CMOS.
6. How does a "dummy collector" prevent latchup?
7. A pad requires a pull-up resistor, which is implemented as a p-transistor that has the source connected to V_{SS}. Does this structure require any latchup protection? What about an n pull-down ($D = input$, $G = V_{DD}$, $S = V_{SS}$)?
8. A CMOS process has unequal n- and p-transistor lengths ($L_N = 0.8\mu$, $L_P = 1.0\mu$). However, a design is desired that uses the same length for each device (1.0μ). Construct a DRC run-set using the commands outlined in Section 3.6.1 that will correctly shrink all the n-transistor gates ($1.0\mu \rightarrow 0.8\mu$), and output data for the final polysilicon mask, assuming that the overall mask has to be bloated by 0.1μ.
9. Most DRC systems deal with merged "canonical" databases, where the rectangles, polygons, etc., in the geometric database are merged before geometric operations are commenced. What could happen to abutting geometric shapes if the source geometry were sized then canonicalized?

3.9 Appendix—An n-well CMOS Technology Process Flow

This section covers in gory detail the processing steps in a now old but representative n-well process developed at the University of California at Berkeley. It is described in terms of a Process Input Description Language (PIDL),[35] which can be used by a software process emulator to predict the

topologies of the final structures. The steps are representative of those taken in processes today, albeit somewhat less complicated. The overall process flow gives an idea of the many steps required to produce even a simple CMOS chip.

The commands in the PIDL language are as follows:

- SUBSTRATE <NAME> (*TYPE=[P,N] IMPURITY=[ ])
  Specifies the substrate name, type, and impurity level.
- OXIDE <NAME> THICKNESS = [ ]
  Specifies oxide layer and thickness.
- DEPOSITION <NAME> (*) THICKNESS=[ ]
  Specifies a layer and thickness of a deposited layer. The (*) is followed by TYPE=[ ] IMPURITY=[ ] if the deposited layer is silicon.
- ETCH <NAME> DEPTH=[ ]
  Specifies a material and an etch depth.
- DOPE TYPE=[P, N] PEAK=[ ] DEPTH=[ ] DELTA=[ ] BLOCK=[ ]
  Specifies parameters necessary to define a diffusion step.
- MASK <RESIST NAME> <EXPOSED NAME> <MASK NAME> <POLARITY OF MASK>
  Specifies a resist layer and associated information.

The complete process input file is as follows (with abbreviations) (© IEEE 1983)[46]

```
1.  LEVEL 1
2.  SUBS SILICON TYPE=P IMPU=1e13; the substrate type and
    impurity is specified
```

Initial oxidation:

```
3.  OXIDE OX1 THICK=0.1; this grows an oxide on the silicon
    surface
```

N-well definition:

```
4.  DEPO NTRD THICK=0.5; nitride is deposited over the oxide
5.  DEPO RST THICK=0.5; resist is deposited
6.  MASK RST DRST MNNL POSI; the resist is positive-masked
    (n-well)
7.  ETCH DRST DEPTH=0.6; the exposed resist is etched
8.  ETCH NTRD DEPTH=0.6; the nitride is etched
9.  ETCH RST DEPTH=0.6; the remaining resist is etched
10. OXIDE OX2 THICK=0.5; oxide is regrown
11. ETCH NTRD DEPTH=0.6
```

LG Display Co., Ltd.
Exhibit 1012
Page 150

```
12. DOPE TYPE=N PEAK=1.5e15 DEPTH=0.0 DELTA=1 5 BLOCK=0.2
    ; well diffusion
13. ETCH OX DEPTH=0.7; oxide etched
14. OXIDE OX3 THICK=0.1; oxide regrown
```

All active area definition:

```
15. DEPO NTRD THICK=0.5; nitride deposited
16. DEPO RST THICK=0.5; resist deposited
17. MASK RST DRST MAA POSI; the resist is positive masked
    (active)
18. ETCH DRST DEPTH=0.6; the exposed resist is removed
19. ETCH NTRD DEPTH=0.6; the nitride thus exposed is etched
20. ETCH RST DEPTH=0.6; the remaining resist is removed
```

Field dope for n-channel:

```
21. DEPO RST THICK=1.0; deposit resist
22. MASK RST DRST MNWL POSI; mask
23. ETCH DRST DEPTH=1.1; etch exposed resist
24. DOPE TYPE=P PEAK=1e21 DEPTH=0.05 DELTA=0. 15
    BLOCK=0.2; diffusion step
25. ETCH RST DEPTH=1.1; remove resist
26. OXIDE OX4 THICK=0.7; grow oxide
27. ETCH NTRD DEPTH=0.6
```

Threshold adjust dope:

```
28. DOPE TYPE=P PEAK=1E20 DEPTH=0.0 DELTA=0.05 BLOCK=0.2
    ; diffusion
```

Regrow gate oxide:

```
29. ETCH OX DEPTH=0.1; remove oxide
30. OXIDE OX5 THICK=0.1; regrow oxide
```

Poly gate definition:

```
31. DEPO POLY THICK=0.30; deposit polysilicon
32. DEPO RST THICK=0.5; deposit resist
33. MASK RST DRST MSI POSI; mask resist with poly mask
34. ETCH DRST DEPTH=0.6; remove exposed resist
35. ETCH POLY DEPTH=0.6; etch exposed polysilicon
36. ETCH RST DEPTH=0.6; remove remaining resist
```

Arsenic dope for n-channel source and drain:

```
37. DEPO RST THICK=1.0; deposit resist
38. MASK RST DRST MIIN POSI; mask for n+
```

LG Display Co., Ltd.
Exhibit 1012
Page 151

```
39. ETCH DRST DEPTH=1.1; remove exposed resist
40. DOPE TYPE=N PEAK=1e22 DEPTH=0.0 DELTA=0.2 BLOCK=0.2
    ; diffusion (or implant)
41. ETCH RST DEPTH=1.1; remove resist
```

Boron dope for p-channel source and drain:

```
42. DEPO RST THICK=1.0; deposit resist
43. MASK RST DRST MTIN NEGA; mask for p+
44. ETCH DRST DEPTH=1.1; remove exposted resist
45. DOPE TYPE=P PEAK=1e22 DEPTH=0.0 DELTA=0.2 BLOCK=0.2
    ; diffusion
46. ETCH RST DEPTH=1.1; remove remaining resist
```

LPCVD oxide (Liquid Phase Chemical Vapor Deposition Oxide):

```
47. DEPO OX6 THICK=0.5; deposit oxide
```

Contact definition:

```
48. DEPO RST THICK=1.0; deposit resist
49. MASK RST DRST MCC NEGA; mask with contact mask
50. ETCH DRST DEPTH=1.1; etch exposed resist
51. ETCH OX DEPTH=1.1; etch oxide down to diffusion
52. ETCH RST DEPTH=1.1; remove resist
```

Metallization:

```
53. DEPO METL THICK=1.0; deposit metal
54. DEPO RST THICK=1.0; deposit resist
55. MASK RST DRST MME POSI; mask with metal mask
56. ETCH DRST DEPTH=1.1; remove exposed resist
57. ETCH METL DEPTH=1.1; remove exposed metal
58. ETCH RST DEPTH=1.1; remove resist
```

Some of the abbreviations are as follows:

```
NTRD  Nitride
RST  Resist
METL  Metal (Aluminum)
NEGA  Negative
POSI  Positive
MNWL  N-well mask
MAA  Thin-oxide mask
MSI  Polysilicon mask
MTIN  NPlus mask
MCC  Contact mask
MME  Metal mask
```

Using the abbreviations and language definitions, the sequence in processing may be traced. For instance, steps 31–36 deposit and etch the polysilicon layer. Step 31 deposits .3μ of polysilicon. Step 32 deposits .5μ of resist called RST. Step 33 masks this resist with a positive polysilicon mask and calls the exposed resist DRST. Step 34 etches DRST to a depth of .6μ. The exposed polysilicon is then etched to a depth of .6μ in step 35. Finally, resist RST is etched away, leaving the final polysilicon pattern. Cross sections may be generated automatically from this process file using the SIMPL-1 program.[47]

3.10 References

1. John Y. Chen, *CMOS Devices and Technology VLSI*, Englewood Cliffs, N.J.: Prentice-Hall, 1990, pp. 233–284.
2. K. Y. Ciu, J. L. Moll, and J. Manoliu, "A bird's beak free local oxidation technology for VLSI," *IEEE Trans. on Electron Devices*, ED-29, pp. 536–540.
3. John Y. Chen, *op. cit.*, pp. 5, 37, and 174–232.
4. John Y. Chen, *op. cit.*, pp. 174–232.
5. L. C. Parrillo *et al.*, "Twin-tub CMOS—a technology for VLSI circuits," *IEEE Int. Electron Devices Meeting Technical Digest*, 1980, Washington, D.C., pp. 752–755.
6. J. Agraz-Guerera, W. Bertram, R. Melin, R. Sun, and J. J. Clemens, "Twin-tub III—a third generation CMOS technology," *IEEE Int. Electron Devices Meeting Technical Digest*, 1984, Washington, D.C., p. 63.
7. H. M. Manasevit and W. I. Simpson, "Single crystal silicon on a sapphire substrate," *J. Appl. Phys.*, vol. 35, 1964, pp. 1349–1351.
8. Yasuaki Hokari, Masao Mikami, Koji Egami, Hideki Tsuya, and Masaru Kanamori, "Characteristics of MOSFET prepared on Si/Mg)Al_2O_3/SiO_2/Si structure," *IEEE JSSC*, vol. 20, no. 1, Feb. 1985, pp. 173–177.
9. Koichi Kato, Tetsunori Wada, and Kenji Taniguchi, "Analysis of kink characteristics in silicon-on insulator MOSFET's using two-carrier modeling," *IEEE JSSC*, vol. SC-20, no. 1, Feb. 1985, pp. 378–382.
10. T. P. Chow, "A review of refractory gates for MOS VLSI," *IEEE Electron Devices Meeting Technical Digest*, Dec. 1983, Washington, D.C., pp. 513–517.
11. T. Tang *et al.*, "Titanium nitride local interconnect technology for VLSI," *IEEE Trans. Electron Devices*, vol. ED-34, Mar. 1987, pp. 682–688.
12. Hiep Van Tran, David B. Scott, Pak Kuen Fung, Robert H. Haverman, Robert H. Eklund, Thomas E. Ham, Roger A. Haken, and Ashwin H. Shah, "An 8-ns 256K ECL SRAM with CMOS memory array and battery backup capability," *IEEE JSSC*, vol. 23, no. 5, Oct. 1988, pp. 1041–1047.
13. Tomohisa Wada, Toshihiko Hirose, Hirofumi Shinohara, Yuji Kawai, Kojiro Yuzuriha, Yoshio Kohno, and Shimpei Kayano, "A 34-ns 1-Mbit CMOS SRAM using triple polysilicon," *IEEE JSSC*, vol. SC-2, no. 5, Oct. 1987, pp. 727–732.
14. Koichiro Mashiko, Masao Nagatomo, Kazutami Arimoto, Yoshio Matsuda, Kiyohiro Furutani, Takayuki Matsukawa, Michihiro Yamada, Tsutomu Yoshihara, and Takao Nakano, "A 4-Mbit DRAM with folded-bit-line adaptive side-

wall-isolated capacitor (FASIC) cell," *IEEE JSSC*, vol. SC-22, no. 5, Oct. 1987, pp. 643–650.

15. Toshio Yamada, Hisakazu Kotani, Junko Matsushima, and Michihiro Inoue, "A 4-Mbit DRAM with 16-bit concurrent ECC," *IEEE JSSC*, vol. 23, no. 1, Feb. 1988, pp. 20–26.
16. Shigeru Mori, Hiroshi Miyamoto, Yoshikazu Morooka, Shigeru Kikuda, Makoto Suwa, Mitsuya Kinoshita, Atsushi Hachisuka, Hideaki Arima, Michihiro Yamada, Tsutomu Yoshihara, and Shimpei Kayano, "A 45-ns 64-Mb DRAM with a merged match-line test architecture," *IEEE JSSC*, vol. 26, no. 11, Nov. 1991, pp. 1486–1492.
17. Masao Taguchi, Hiroyoshi Tomita, Toshiya Uchida, Yasunhiro Ohnishi, Kimiaki Sato, Taiji Ema, Masaaki Higashitani, and Takashi Yabu, "A 40-ns 64-Mb DRAM with 64-b parallel data bus architecture," *IEEE JSSC*, vol. 26, no. 11, Nov. 1991, pp. 1493–1497.
18. Richard D. Jolly, Rod Tesch, Ken J. Campbell, David L. Tennant, Jay F. Olund, Robert B. Lefferts, Brendan T. Cremen, and Philip A. Andrews, "A 35-ns 64K EEPROM," *IEEE JSSC*, vol. SC-20, no. 5, Oct. 1985, pp. 971–978.
19. Koichi Seki, Hitoshi Kume, Yuzuru Ohji, Takashi Kobayashi, Atsushi Hiraiwa, Takashi Nishida, Takeshi Wada, Kazuhiro Komori, Kazuto Izawa, Toshiaki Nishimoto, Yasuroh Kubota, and Kazuyoshi Shohji, "An 80-ns 1-Mb flash memory with on-chip erase/erase-verify controller," *IEEE JSSC*, vol. 25, no. 5, Oct. 1990, pp. 1147–1152.
20. Katsumoto Soejima, Akira Shida, Hiroshi Koga, Junnichi Ukai, Hiroshi Sata, and Masaki Hirata, "A BiCMOS technology with 660MHz vertical p-n-p transistor for analog/digital ASIC's," *IEEE JSSC*, vol. 25, no. 2, Apr. 1990, pp. 410–416.
21. Ali A. Iranmanesh, Vida Ilderem, Madan Biswal, and Bami Bastani, "A 0.8mm advanced single-poly BiCMOS technology for high-density and high-performance applications," *IEEE JSSC*, vol. 26, no. 3, Mar. 1991, pp. 422–423.
22. P. K. Weimer, "The Insulated-Gate Thin-Film Transistor," in *Physics of Thin Films, Vol. 2*, New York: Academic Press, 1963, pp. 147–192.
23. Richard S. C. Cobbold, *Theory and Applications of Field-Effect Transistors*, New York: Wiley Interscience, 1970, pp. 54–64.
24. Satwinder D. S. Malhi, Hisashi Shichijo, Sanjay K. Banerjee, Ravishankar Sundaresan, Mostafa Elahy, Gordan P. Pollack, William F. Richardson, Ashwin H. Shah, Larry R. Hite, Richard H. Womack, Pallab K. Chatterjee, and Hon Wai Lam, "Characteristics and three-dimensional integration of MOSFET's in small grain LPCVD polycrystalline silicon," *IEEE JSSC*, vol. SC-20, no. 1, Feb. 1985, pp. 178–201.
25. Katsuro Sasaki, Koichiro Ishibashi, Katsuhiro Shimohigashi, Toshiaki Yamanaka, Nobuyuki Moriwaki, Shigeru Honjo, Shuji Ikeda, Atsuyoshi Koike, Satoshi Meguro, and Osamu Minato, "A 23-ns 4-Mb CMOS SRAM with 0.2mm standby current," *IEEE JSSC*, vol. 25, no. 5, Oct. 1990, pp. 1075–1081.
26. Takayuki Ootani, Shigeyuki Hayakawa, Masakazu Kakumu, Akira Aono, Masaaki Kinugawa, Hideki Takeuchi, Kazuhiro Noguchi, Tomoaki Yabe, Katsuhiko Sato, Kneji Maeguchi, and Kiyofumi Ochi, "A 4-Mb CMOS SRAM with a PMOS thin-film-transistor load cell," *IEEE JSSC*, vol. 25, no. 5, Oct. 1990, pp. 1082–1092.
27. Yutaka Takafuji, Toshihiro Yamashita, Yasunobu Akebi, Tomoaki Toichi, Takayuki Shimada, and Katsunobu Awane, "A poly-Si TFT monolithic LC data driver with redundancy," *IEEE, Proceedings of ISSCC*, Feb. 1992, San Francisco, Calif., pp. 118–119.

LG Display Co., Ltd.
Exhibit 1012
Page 154

28. Gerhard Roos and Bernd Hoefflinger, "Complex 3D CMOS circuits based on a triple-decker cell," *IEEE JSSC*, vol. 27, no. 7, Jul. 1992, pp. 1067–1072.
29. C. A. Mead and L. A. Conway, *Introduction to VLSI Systems*, Reading, Mass.: Addison-Wesley, 1980.
30. Shingo Aizaki, Toshiyuki Shimizu, Masayoshi Ohkawa, Kazuhiko Abe, Akane Aizaki, Manabu Ando, Osamu Kudoh, and Isao Sasaki, "A 15nS 4-Mb CMOS SRAM," *IEEE JSSC*, vol. 25, no. 5, Oct. 1990, pp. 1063–1067.
31. Katsuro Sasaki, Koichiro Ishibashi, Katsuhiro Shimohigashi, Toshiaki Yamanaka, Nobuyuki Moriwaki, Shigeru Honjo, Shuji Ikeda, Atsuyoshi Koike, Satoshi Meguro, and Osamu Minato, "A 23-ns 4-Mb CMOS SRAM with 0.2mm standby current," *IEEE JSSC*, vol. 25, no. 5, Oct. 1990, pp. 1075–1081.
32. Takayuki Ootani, *et al., op. cit.*
33. Osamu Nishii, Makoto Hanawa, Tadahiko Nishimukai, Makoto Susuki, Kazuo Yano, Mitsuru Hiraki, Shohji Shukuri, and Takashi Nishida, "A 1,000 MIPS BiCMOS microprocessor with superscalar architecture," IEEE Proceedings of ISSCC, Feb. 1992, San Francisco, Calif., pp. 114–115.
34. W. S. Chang, B. Davari, M. R. Wordeman, Y. Taur, C. C. H. Hsu and M. D. Rodriquez, "A High-Performance 0.25µm CMOS Technology I—Design and Characterization," *IEEE Transactions on Electron Devices*, vol. 39, no. 4, April 1992, pp. 959–966, *and* B. Davari, W. H. Chang, K. E. Petrillo, C. Y. Wong, D. Moy, Y. Taur, M. R. Wordeman, J. Y. C. Sun, C. C. H. Hsu and M. R. Polcari, "A High-Performance 0.25µm CMOS Technology II—Technology," *IEEE Transactions on Electron Devices*, vol. 39, no. 4, Apr. 1992, pp. 967–975.
35. D. B. Estreich, "The physics and modeling of latch-up in CMOS integrated circuits," *Tech. Report No. G-2-1-9*, Integrated Circuits Laboratory, Stanford Electronics Lab., Stanford University, Nov. 1980.
36. D. B. Estreich and R. W. Dutton, "Modeling latch-up in CMOS integrated circuits and systems," *IEEE Transactions on CAD*, vol. CAD-1, no. 4, Oct. 1982, pp. 347–354.
37. R. R. Troutman, *Latch-Up in CMOS Technology: The Problem and Its Cure*, Boston, Mass.: Kluwer Academic Publishers, 1986.
38. William M. Coughran, Mark R. Pinto, and R. Kent Smith, "Computation of steady-state CMOS latchup characteristics," *IEEE Transactions on CAD*, vol. 7, no. 2, Feb. 1988, pp. 307–323.
39. John Y. Chen, *op. cit.*, pp. 285–322.
40. John Y. Chen, *op. cit.*, pp. 286–288.
41. John Y. Chen, *op. cit.*, pp. 289–290.
42. William M. Coughran *et al., op. cit.*
43. D. B. Estreich and R. W. Dutton, *op. cit.*
44. "DRACULA III," Design Rule Check Program CADENCE, Design Systems, Inc., San Jose, Calif.
45. M. A. Grimm, K. Lee, and A. R. Neureuther, "SIMPL-1 (SIMulated Profiles from the layout-version 1)," *Proc. IEDM 1983*, Dec. 1983, pp. 255–258.
46. M. A. Grimm *et al., op. cit.*
47. M. A. Grimm *et al., op. cit.*

LG Display Co., Ltd.
Exhibit 1012
Page 155