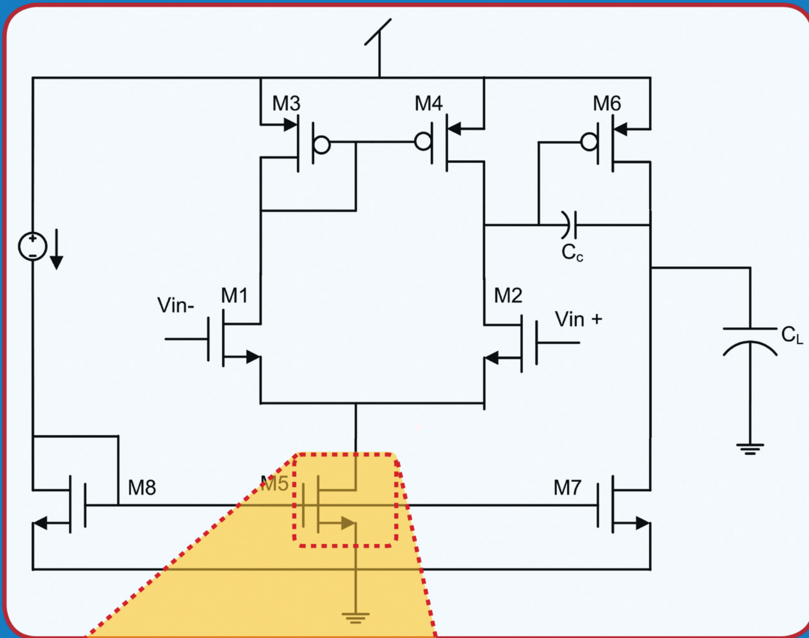
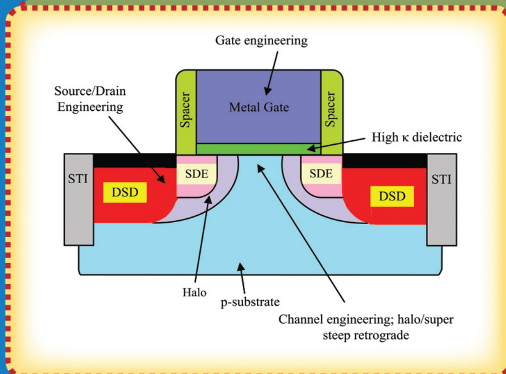
Nano-Scale CMOS Analog Circuits

Models and CAD Techniques for High-Level Design





Soumya Pandit
Chittaranjan Mandal
Amit Patra

CRC Press
Taylor & Francis Group

Nano-Scale CMOS Analog Circuits

Models and CAD Techniques for High-Level Design

Nano-Scale CMOS Analog Circuits

Models and CAD Techniques for High-Level Design

Soumya Pandit
Chittaranjan Mandal
Amit Patra

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

MATLAB® and Simulink® are trademarks of The MathWorks, Inc. and are used with permission. The MathWorks does not warrant the accuracy of the text or exercises in this book. This book's use or discussion of MATLAB® and Simulink® software or related products does not constitute endorsement or sponsorship by The MathWorks of a particular pedagogical approach or particular use of the MATLAB® and Simulink® software.

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2014 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Version Date: 20140107

International Standard Book Number-13: 978-1-4665-6428-2 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Dedicated to our family members

Contents

List of Figures xv

List of Tables xxi

Preface xxiii

About the Authors xxvii

1 Introduction 1

1.1 Introduction 1
1.2 Characterization of Technology Scaling 2
1.2.1 Constant Field Scaling 2
1.2.2 Constant Voltage Scaling 2
1.2.3 Nonscaling Effects 4
1.2.4 Generalized Scaling and Technology Trends 5
1.2.4.1 International Technology Roadmap for Semiconductors 5
1.2.4.2 Predictive Technology Modeling 6
1.2.4.3 Scaling of Geometry Parameters 6
1.2.4.4 Scaling of Channel Doping, Supply Voltage, and Threshold Voltage 8
1.2.4.5 Scaling of Performances 10
1.2.4.6 Scaling of Source-Drain Resistance and Saturation Velocity 12
1.3 Analog Design Challenges in Scaled CMOS Technology . . . 14
1.3.1 Degradation of Output Resistance and Intrinsic Gain 14
1.3.2 Gate Oxide Leakage Current 15
1.3.3 Noise Performance 16
1.3.4 Analog Power Consumption 16
1.3.5 Drain Current Mismatch 18
1.3.6 Transition Frequency 18
1.3.7 Reliability Constraints 18
1.4 Motivation for CAD Techniques 19
1.4.1 Design Productivity Gap 19
1.4.2 Design Creativity Gap 20
1.5 Conventional Design Techniques for Analog IC Design 22
1.5.1 Bottom-Up Design Technique 22

1.5.1.1 Advantages and Limitations 24
1.5.2 Top-Down Design Technique 24
1.5.2.1 Abstraction Levels 25
1.5.2.2 Hierarchical Design Strategy 26
1.5.2.3 Advantages and Limitations 26
1.6 Knowledge-Based CAD Technique for Analog ICs 28
1.6.1 Motivation for Knowledge Extraction and Management 29
1.6.2 Problem Formulations 29
1.6.3 Outline of the Procedure 30
1.6.3.1 Numerical Simulation-Based Evaluation . . . 30
1.6.3.2 Analytical Model-Based Evaluation 32
1.6.4 Salient Features 32
1.7 Summary and Conclusion 33

2 High-Level Modeling and Design Techniques 35
2.1 Introduction 35
2.2 High-Level Model 36
2.2.1 Behavioral Models 36
2.2.2 Performance Models 37
2.2.3 Feasibility Models 37
2.2.4 Characteristics of Good High-Level Models 38
2.3 Behavioral Model Generation Technique 39
2.3.1 Manual Abstraction 39
2.3.2 Model Order Reduction Technique 41
2.3.2.1 MOR for LTI Systems 41
2.3.2.2 MOR for LTV Systems 43
2.3.2.3 MOR for Nonlinear Systems 43
2.3.3 Symbolic Analysis Technique 44
2.3.3.1 Basic Concepts 44
2.3.3.2 Methodology 46
2.3.3.3 Simplification of Expressions 47
2.4 Introduction to Optimization Techniques 49
2.4.1 Optimization Problem Formulation 49
2.4.2 Optimality Criteria 51
2.4.3 Classification of Optimization Algorithms 51
2.4.3.1 Single-Variable Optimization Algorithms . . 51
2.4.3.2 Multi-Variable Optimization Algorithm . . . 51
2.4.3.3 Constrained Optimization Algorithms 52
2.4.3.4 Specialized Optimization Algorithms 53
2.4.3.5 Nontraditional Stochastic Optimization Algorithms 53
2.4.4 Concept of Local Optima and Global Optima 54
2.4.5 Characterization of Optimization Algorithms 55
2.5 Some Important Optimization Algorithms 56
2.5.1 Cauchy's and Newton's Steepest Descent Algorithm . 56

2.5.2 Genetic Algorithm 57
2.5.3 Simulated Annealing 60
2.6 Multi-Objective Optimization Method 61
2.6.1 Pareto Optimal Front 62
2.7 Design Space Exploration 63
2.8 Computational Complexity of a CAD Algorithm 63
2.8.1 Time and Space Complexity 64
2.8.2 Asymptotic Notations 65
2.8.2.1 Big-Oh Notation $O()$ 65
2.8.2.2 Omega Notation $\Omega()$ 66
2.8.2.3 Theta Notation $\Theta()$ 66
2.8.3 Categorization of CAD Problems 67
2.8.4 Complexity Classes for CAD Problems 67
2.9 Technology-Aware Computer Aided IC Design Technique . . 68
2.9.1 Introduction to TCAD 69
2.9.2 Process Simulation through TCAD 70
2.9.3 Device Simulation through TCAD 70
2.9.4 Design for Manufacturability and Yield 71
2.9.5 Process Compact Model Development through TCAD 73
2.9.5.1 Parameter Extraction Technique 74
2.9.6 Design Techniques for Nano-Scale Analog ICs 75
2.10 Commercial Design Tools 75
2.10.1 IC Design 75
2.10.1.1 Cadence® Virtuoso Analog Design Environment 77
2.10.1.2 Synopsys Galaxy Custom Design 79
2.10.1.3 Tanner EDA HiPer Silicon® 80
2.10.1.4 Mentor Graphics Pyxis® Suite 81
2.10.2 TCAD 81
2.10.2.1 Silvaco Tool Suite 81
2.10.2.2 Synopsys Device Simulation Tool Suite . . . 83
2.11 Summary and Conclusion 84

3 Modeling of Scaled MOS Transistor for VLSI Circuit Simulation 85
3.1 Introduction 85
3.2 Device Modeling 86
3.2.1 Categories of Device Models 87
3.3 Compact Models 87
3.3.1 Commercial Compact Models 88
3.4 Long-Channel MOS Transistor 89
3.5 Threshold Voltage Model for Long-Channel Transistor with Uniform Doping 90
3.5.1 Body Effect 91
3.6 SPICE Level 1 Drain Current Model 92

3.6.1 Channel Length Modulation Effect 96
3.7 SPICE Level 3 I-V Model 97
3.8 MOSFET Capacitances 99
3.8.1 Characterization of Intrinsic Capacitances 100
3.8.1.1 Charge Partitioning 103
3.8.2 Characterization of Extrinsic Capacitances 104
3.8.2.1 Overlap and Fringing Capacitances 104
3.8.2.2 Junction Capacitances 105
3.9 Short-Channel MOS Transistor 107
3.10 Threshold Voltage for Short-Channel MOS Transistor 110
3.10.1 Source/Drain Charge Sharing 110
3.10.1.1 Level 2 Compact Model for V_T 111
3.10.2 Drain-Induced Barrier Lowering 112
3.10.2.1 Level 3 Model of DIBL 113
3.10.3 BSIM3/BSIM4 Compact Model for Threshold Voltage 114
3.10.4 Short-Channel Effect Immunity 117
3.11 I-V Model for Short-Channel MOS Transistor 120
3.11.1 Carrier Mobility Degradation 120
3.11.1.1 Surface Mobility 120
3.11.1.2 Mobility Dependence on Gate Field 121
3.11.2 Carrier Velocity Saturation 123
3.11.3 Drain Current in Scaled MOS Transistor 124
3.12 Weak Inversion Characteristics of a Scaled MOS Transistor . 128
3.12.1 Subthreshold Swing 131
3.13 Hot Carrier Effect 134
3.13.1 Spatial Distribution of Lateral Electric Field 134
3.13.2 Substrate Current Due to Hot-Carrier Effects 137
3.13.3 Gate Current Due to Hot-Carrier Effects: Lucky Electron Model 139
3.13.4 Reduction of Drain Field through LDD Structure. . . 141
3.14 Source-Drain Resistance Model 141
3.14.1 Compact Modeling 143
3.14.2 Salicide Technology 144
3.15 Physical Model for Output Resistance 145
3.15.1 Compact Modeling 147
3.16 Poly-Silicon Gate Depletion Effect 150
3.16.1 Electrical Oxide Thickness 152
3.16.2 Reduction of Poly-Gate Depletion 152
3.17 Effective Channel Length and Width 152
3.17.1 Effective Channel Length 154
3.17.1.1 Extraction of the Effective Channel Length 155
3.18 Summary and Conclusion 156

4 Performance and Feasibility Model Generation Using Learning-Based Approach 157
4.1 Introduction . . . 157
4.2 Requirement of Learning-Based Approaches . . . 157
4.3 Regression Problem for Performance Model Generation . . . 158
4.4 Some Related Works . . . 158
4.5 Preliminaries on Artificial Neural Network . . . 160
4.5.1 Basic Components . . . 160
4.5.2 Mathematical Model of Neuron . . . 160
4.5.3 MLP Feed-Forward NN Structure . . . 161
4.5.4 Feed-Forward Computation . . . 162
4.5.5 Success of MLP NN Structures . . . 163
4.5.6 Network Size and Layers . . . 164
4.6 Neural Network Model Development . . . 164
4.6.1 Formulation of Inputs and Outputs . . . 164
4.6.2 Data Range and Sample Distribution . . . 164
4.6.3 Data Collection . . . 167
4.6.4 Data Organization and Data Preprocessing . . . 167
4.6.5 Neural Network Training . . . 168
4.6.6 Quality Measures . . . 170
4.6.7 Generalization Ability, Overlearning, and Underlearning . . . 170
4.7 Case Study 1: Performance Modeling of CMOS Inverter . . . 171
4.8 Case Study 2: Performance Modeling of Spiral Inductor . . . 174
4.9 Dynamic Adaptive Sampling . . . 182
4.9.1 Motivation of the Algorithm . . . 183
4.9.2 Simple Dynamic Sampling Algorithms . . . 183
4.9.3 Dynamic Adaptive Sampling Algorithm . . . 184
4.9.3.1 Initial Sample Size . . . 184
4.9.3.2 Sampling Schedule . . . 184
4.9.3.3 Stopping Criteria . . . 185
4.9.4 Demonstration with CMOS Inverter Problem . . . 186
4.10 Introduction to Least Squares Support Vector Machines . . . 187
4.10.1 Least-Squares Support Vector Regression . . . 187
4.10.2 Least-Squares Support Vector Classification . . . 189
4.10.2.1 Classifier Accuracy . . . 191
4.10.3 Choice of Kernel Functions and Hyperparameter Tuning . . . 192
4.10.3.1 Grid Search Technique . . . 194
4.10.3.2 Genetic Algorithm-Based Technique . . . 196
4.11 Feasible Design Space and Feasibility Model . . . 198
4.12 Case Study 3: Combined Feasibility and Performance Modeling of Two-Stage Operational Amplifier . . . 201
4.12.1 Feasibility Model . . . 202
4.12.2 Performance Model . . . 203

4.13 Case Study 4: Architecture-Level Performance Modeling of Analog Systems . . . 204
4.14 Meet-in-the-Middle Approach for Construction of Architecture-Level Feasible Design Space . . . 208
4.14.1 Application Bounded Space $\mathcal{D}_a$ . . . 208
4.14.2 Circuit Realizable Space $\mathcal{D}_c$ . . . 210
4.14.3 Feasible Design Space Identification . . . 210
4.15 Case Study 5: Construction of Feasibility Model at Architecture Level of an Interface Electronics for MEMS Capacitive Accelerometer System . . . 211
4.16 Summary and Conclusion . . . 215

5 Circuit Sizing and Specification Translation 217
5.1 Introduction . . . 217
5.2 Circuit Sizing as a Design Space Exploration Problem . . . 217
5.2.1 Problem Formulations . . . 217
5.2.2 Solution Techniques . . . 219
5.2.3 Design Flow . . . 220
5.2.3.1 Evaluation of Cost Functions . . . 221
5.3 Particle Swarm Optimization Algorithm (PSO) . . . 221
5.3.1 Dynamics of a Particle in PSO . . . 222
5.3.2 Flow of the Algorithm . . . 223
5.3.3 Selection of Parameters for PSO . . . 223
5.3.3.1 Inertia Weight ω . . . 225
5.3.3.2 Maximum Velocity V_{max} . . . 225
5.3.3.3 Swarm Size S . . . 225
5.3.3.4 Acceleration Coefficient c_1 and c_2 . . . 225
5.4 Case Study 1: Design of a Two-Stage Miller OTA . . . 226
5.5 Case Study 2: Synthesis of on-Chip Spiral Inductors . . . 229
5.6 Case Study 3: Design of a Nano-Scale CMOS Inverter for Symmetric Switching Characteristics . . . 234
5.7 The g_m/I_D Methodology for Low Power Design . . . 237
5.7.1 Study of the g_m/I_D and f_T Parameters for Analog Design . . . 238
5.7.2 g_m/I_D Based Sizing Methodology . . . 241
5.7.3 Case Study 4: Sizing of Low-Power Nano-Scale Miller OTA Using the g_m/I_D Methodology . . . 246
5.8 High-Level Specification Translation . . . 250
5.9 Summary and Conclusion . . . 251

6 Advanced Effects of Scaled MOS Transistors 253
6.1 Introduction . . . 253
6.2 Narrow Width Effect on Threshold Voltage . . . 253
6.2.1 LOCOS Isolated MOS Transistors . . . 254
6.2.2 Shallow Trench Isolated (STI) MOSFETs . . . 255

6.3 Channel Engineering of MOS Transistor 257
6.3.1 Non-Uniform Vertical Doping 257
6.3.1.1 High-to-Low Profile 258
6.3.1.2 Low-to-High Retrograde Profile 260
6.3.1.3 Compact Modeling of Vertical Non-Uniform Doping Effect 260
6.3.2 Pocket (Halo) Implantation 261
6.4 Gate Leakage Current 263
6.4.1 Basic Ideas about Quantum Mechanical Tunneling . . 263
6.4.2 Gate Oxide Tunneling Current 266
6.4.2.1 Energy Band Theory Model 266
6.4.2.2 Fowler–Nordheim Tunneling 266
6.4.2.3 Direct Tunneling 269
6.4.3 Gate Leakage Mechanisms and Leakage Components for MOS Transistors 269
6.4.4 Compact Modeling 272
6.4.5 Effects of Gate Leakage 272
6.5 High-κ Dielectrics and Metal-Gate/High-κ CMOS Technology 275
6.5.1 High-κ Dielectric Materials 275
6.5.2 Metal Gate 278
6.6 Advanced Device Structures of MOS Transistors 279
6.6.1 SOI MOS Transistor 279
6.6.2 Double Gate (DG)-MOS Transistors 280
6.6.3 FinFETs 282
6.7 Noise Characterization of MOS Transistors 284
6.7.1 Fundamental Sources of Noise 284
6.7.2 Characterization of Thermal Noise in MOS Transistors 286
6.7.3 Characterization of Flicker Noise in MOS Transistors . 288
6.7.3.1 Physical Mechanism of Flicker Noise 288
6.7.3.2 Physics-Based Modeling of Flicker Noise . . 291
6.8 Gate Resistance and Substrate Network Model of MOS Transistor for RF Applications 294
6.8.1 Parasitic Components of MOS Transistors 297
6.8.2 Gate Resistance Modeling 298
6.8.2.1 Minimization of Gate Resistance 300
6.8.3 Substrate Network Modeling 301
6.9 Summary and Conclusion 302

7 Process Variability and Reliability of Nano-Scale CMOS Analog Circuits 305
7.1 Introduction 305
7.2 Basic Concepts on Yield and Reliability 306
7.2.1 Yield 306
7.2.2 Design Tolerance and Capability Index 308
7.2.3 Reliability Bathtub Curve 310

7.3 Sources of Variations in Nanometer Scale Technology 313
7.3.1 Process Variations 314
7.3.2 Environmental Variations 316
7.3.3 Aging Variations or Reliability 316
7.4 Systematic Process Variations 316
7.4.1 Optical Proximity Correction 317
7.4.2 Phase Shift Mask 317
7.4.3 Layout-Induced Strain 318
7.4.4 Well Proximity Effect 319
7.5 Random Process Variations 320
7.5.1 Random Discrete Dopants 321
7.5.2 Line Edge Roughness 322
7.5.3 Oxide Thickness Variations 327
7.5.4 High-κ Dielectric Morphology and Metal Gate Granularity 327
7.6 Statistical Modeling 327
7.6.1 Worst Case Corner Analysis 329
7.6.2 Monte Carlo Simulation Technique 331
7.6.3 Statistical Corner Technique 331
7.6.4 Mismatch in Analog Circuits 334
7.7 Physical Phenomena Affecting the Reliability of Scaled MOS Transistor 335
7.7.1 Time Dependent Dielectric Breakdown (TDDB) . . . 337
7.7.1.1 Electrostatics in Dielectrics 337
7.7.1.2 Energy Band Theory of Dielectric Breakdown 338
7.7.1.3 Anode Hole Injection Theory 339
7.7.1.4 Percolation Theory 341
7.7.1.5 Statistics of Gate Oxide Breakdown 341
7.7.1.6 Soft and Hard Breakdown 341
7.7.2 Hot Carrier Injection (HCI) 342
7.7.3 Negative Bias Temperature Instability (NBTI) 342
7.8 Physical Model for MOSFET Degradation Due to HCI . . . 342
7.8.1 Physical Mechanism for Interface Trap Generation . . 343
7.8.1.1 Physical Model 344
7.8.1.2 Application to Analog Circuit Design 345
7.9 Reaction-Diffusion Model for NBTI 347
7.10 Reliability Simulation for Analog Circuits 347
7.11 Summary and Conclusion 349

Bibliography **351**

List of Figures

1.1 Illustration of half-pitch and technology node. 3
1.2 Scaling of physical gate length and effective channel length across technology node. 6
1.3 Scaling of equivalent oxide thickness across technology node. 7
1.4 Scaling of source-drain junction depth across technology node. 8
1.5 Scaling of channel doping across technology node. 9
1.6 Scaling of the power supply voltage across technology node. 9
1.7 Scaling of threshold voltage across technology node. 10
1.8 Scaling of the drain ON current across technology node. . . 11
1.9 Scaling of the drain OFF current across technology node. . . 11
1.10 Scaling of the intrinsic delay across technology node. 12
1.11 Scaling of the static power dissipation across technology node. 13
1.12 Scaling of source-drain junction depth across technology node. 13
1.13 Scaling of the saturation velocity across technology node. . . 14
1.14 Single transistor amplifier. 17
1.15 Driving forces behind research works in analog CAD. 21
1.16 Bottom-up design technique for analog IC. 23
1.17 Abstraction levels in top-down analog IC design technique. . 25
1.18 Hierarchical design strategy in top-down design technique. . 27
1.19 Knowledge-based CAD procedure. 31

2.1 A circuit C consisting of design variables α, with input signal U and output signal Y. 36
2.2 Mapping of feasible design variables to feasible performance parameters. 38
2.3 Manual approach for construction of high-level model using Simulink®. 40
2.4 LTI block. 42
2.5 Active RC filter. 45
2.6 Small signal model of the transistor amplifier circuit. 46
2.7 Flow chart for the steps of formulating an optimization problem. 48
2.8 Strategy for global optimum search. 54
2.9 Illustration of Pareto front for a sample 2D design space. . . 62
2.10 Graphical illustrations of Θ, O and Ω notations. 65
2.11 Relationships among P, NP, and NPC. 68

2.12 Inputs–outputs of a process simulation procedure. 70
2.13 Inputs–outputs of a device simulation procedure. 71
2.14 Outline of TCAD simulation-based DFMY technique. 72
2.15 Process compact model generation using TCAD. 73
2.16 Optimization flow for parameter extraction. 76
2.17 Design technique for nano-scale analog ICs. 77
2.18 General design flow for custom analog ICs. 78
2.19 IC design flow using Tanner EDA tool suite. 80
2.20 IC design flow using mentor graphics EDA tool suite. 82
2.21 Device simulation tools offered by Synopsys TCAD suite. . . 83

3.1 Use of device models in VLSI circuit simulation. 86
3.2 Cross-sectional view of n-type long-channel MOS transistor. 89
3.3 Threshold voltage vs. substrate bias for a long-channel NMOS transistor. 91
3.4 Gate and drain characteristics of an n-channel MOS transistor as obtained from SPICE-Level 1 simulation. 95
3.5 Pinch off phenomenon. 96
3.6 Comparison between the drain characteristics of an n-channel MOS transistor of dimension $W = 5\mu m, L = 1\mu m$, simulated by Level 1 and Level 3 SPICE model. 99
3.7 Various capacitors within an MOS transistor. 100
3.8 The overlap, fringing, and junction capacitances. 104
3.9 Bottom and sidewall component of a junction capacitance. . 106
3.10 Variation of intrinsic capacitances with drain bias as obtained from SPICE simulation. 108
3.11 Source/drain charge sharing leading to threshold voltage reduction. 111
3.12 Illustration of barrier lowering for a short-channel MOS transistor with high drain bias. 112
3.13 Gaussian box for the 2D analysis of V_T roll-off and DIBL. . 113
3.14 SPICE simulation results for short channel effects on threshold voltage, $V_{BS} = 0V$. 118
3.15 Comparison between BSIM simulation results and analytical results for threshold voltage roll off and DIBL characteristics. 119
3.16 Variation of surface mobility of the carriers with gate bias. . 122
3.17 Illustration of carrier velocity saturation. 123
3.18 Variation of the drain saturation voltage with $(V_{GS} - V_T)$ in the presence of velocity saturation with channel length as a parameter. 126
3.19 Comparison of gate characteristics for long-channel ($1\mu m$) and short-channel ($65nm$) MOS transistors. 127
3.20 Equivalent capacitance network in the weak inversion region. 130
3.21 Variation of surface potential with gate bias in weak inversion region. 131

3.22 Variation of weak inversion drain current with gate bias at $V_{DS} = V_{DD}$. 132
3.23 Variation of weak inversion drain current with gate bias for different drain biases as obtained from SPICE simulation results. 133
3.24 Schematic diagram of a MOS transistor in the velocity saturation region and the Gaussian box for computing the spatial distribution of electric field. 134
3.25 Illustration of substrate current due to hot carriers. 138
3.26 Trajectory of lucky electron. 140
3.27 Schematic illustration of the current flow in the source-drain extension and source diffusion region leading to various parasitic components of the source resistance. 142
3.28 MOSFET representation including parasitic source and drain resistances. 143
3.29 Self-aligned source/drain contact. 145
3.30 Schematic illustration of the typical drain current and output resistance 146
3.31 Simulation results showing the variation of output resistance. 148
3.32 Variation of output resistance with drain bias for $V_{GS} = 0.4V$ and $V_{GS} = 0.7V$. 149
3.33 Illustration of poly-silicon gate depletion effect. 150
3.34 Variation of effective gate voltage with applied gate voltage. 153
3.35 Schematic illustrating the definitions of drawn channel length, gate length, and effective channel length. 154
3.36 Channel resistance method for extraction of R_{ds} and ΔL. . 156

4.1 Basic components of an ANN structure. 161
4.2 Mathematical model of a neuron. 162
4.3 Simple MLP NN structure for illustration of the principle of feed forward computation. 163
4.4 Steps for the development of NN model. 165
4.5 Flow chart illustrating the ANN training procedure. 169
4.6 CMOS inverter. 172
4.7 MLP-NN architecture for Case Study 1. 174
4.8 Effect of the value of training goal on test error. 175
4.9 SPICE simulation vs. ANN prediction of rise time and fall time. 176
4.10 SPICE simulation vs. ANN prediction switching point and average power dissipation. 177
4.11 Scatter diagram plot for ANN predicted output for rise time and fall time. 178
4.12 Scatter diagram plot for ANN predicted output for switching point and average power dissipation. 179
4.13 Layout diagram of an on-chip spiral square inductor. 181

4.14 Hypothetical learning curve for ANN model. 183
4.15 Dynamic adaptive sampling algorithm. 185
4.16 Classification problem for linearly separable dataset. 190
4.17 Mapping of the input space to a high dimensional feature space where linear separation of nonseparable data is possible. 192
4.18 Feasible design space and its subspaces. 193
4.19 Outline of GA-based hyperparameter selection procedure. . 195
4.20 n-Channel simple current mirror circuit. 198
4.21 Two-stage CMOS operational amplifier circuit for Case Study 3. 201
4.22 Nonlinear mapping from the sample space to the input and the output space for the construction of architecture level performance model. 204
4.23 Two-stage CMOS OTA circuit for Case Study 4. 205
4.24 Scatter plot of the constructed models. 209
4.25 Meet-in-the-middle way of constructing the feasible design space $\mathcal{D}$. 211
4.26 Considered system for Case Study 5. 212

5.1 General design flow of a circuit sizing/specification translation task. 218
5.2 A CMOS output buffer circuit. 219
5.3 Design space exploration process for circuit sizing/ specification translation task. 220
5.4 Illustration of the dynamics of a particle. 223
5.5 Flow chart of the PSO algorithm. 224
5.6 Schematic diagram of the Miller OTA. 226
5.7 AC simulation results of the synthesized OTA in Case Study 1. 230
5.8 ICMR and CMRR results of the synthesized OTA in Case Study 1. 231
5.9 Flow chart of the on-chip spiral inductor synthesis problem. 232
5.10 Flow chart of the inverter design problem. 235
5.11 Simulation results for the variations of g_m/I_D and f_T with the region of operation and technology nodes. 239
5.12 Simulation results showing the variations of the product of g_m/I_D and f_T with the region of operation and technology node. 240
5.13 Simulation results showing the variations of the transconductance linearity with the region of operation and technology node. 240
5.14 Simulation results for the variations of C_{GS} and C_{GD} with the g_m/I_D. 242
5.15 Simulation results for the g_m/I_D variations. 243

5.16 Simulation results showing the variations of the g_m/I_D with the normalized current I_N. . . . 244
5.17 Simulation results showing the variations of the intrinsic gain and drain current with the region of operation. . . . 245
5.18 AC simulation results of the synthesized OTA in Case Study 4. . . . 248
5.19 CMRR and slew rate of the synthesized OTA in Case Study 4. . . . 249

6.1 Cross-section along the width of a MOS transistor with LOCOS isolation. . . . 254
6.2 Cross-section along the width of a MOS transistor with STI. 255
6.3 Width dependence of threshold voltage for LOCOS and STI MOS transistors. . . . 257
6.4 Step doping and abrupt retrograde channel doping profile. . 258
6.5 Cross-sectional view of a MOS transistor with double halo channel engineering. . . . 261
6.6 Lateral variation of channel doping. . . . 262
6.7 Quantum-mechanical tunneling. . . . 263
6.8 Tunneling of electrons through a MOS capacitor. . . . 267
6.9 Energy band theory for gate oxide leakage tunneling. . . . 268
6.10 Different mechanisms of gate leakage current in MOS transistor. . . . 270
6.11 Schematic diagram showing gate leakage components in n-channel MOS transistor. . . . 271
6.12 Investigation of the gate leakage current of an n-channel MOS transistor. . . . 273
6.13 Variation of gate current for n-channel MOS transistor. . . . 274
6.14 Investigation of f_{gate} of an n-channel MOS transistor. . . . 276
6.15 Limited current gain with scaling of technology. . . . 277
6.16 Requirement of work function of metal gates for n-channel and p-channel MOS transistors. . . . 278
6.17 Thin film SOI MOS transistor. . . . 280
6.18 Schematic diagram of a double gate MOS transistor. . . . 281
6.19 Different architectures of DG-MOSFET: (a) planar DG-MOSFET (b), (c) vertical DG-MOSFETs. . . . 283
6.20 Noise PSD of n-channel and p-channel MOS transistors. . . 289
6.21 Simulated thermal noise PSD for n-channel MOS transistor. 290
6.22 Variation of normalized flicker noise PSD for n-channel MOS transistor with gate bias and drain bias. . . . 295
6.23 Variation of normalized flicker noise PSD for p-channel MOS transistor with gate bias. . . . 296
6.24 Flicker noise spectrum of n-channel MOS transistor with gate length as parameter. . . . 296

6.25 Schematic diagram of a MOS transistor illustrating the parasitic components. 297
6.26 Schematic diagram illustrating the distributed nature of gate and channel resistance. 299
6.27 Illustration of the multi-finger layout concept. 301
6.28 Equivalent circuit of the substrate network. 302

7.1 Statistical spreading of process parameter Γ. 307
7.2 Area under normal distribution. 309
7.3 Schematic illustration of potential capability index. 310
7.4 Schematic illustration of process capability index. 311
7.5 Schematic illustration of the concept of design centering. . . 312
7.6 Reliability bathtub curve. 312
7.7 Classification of the various sources of variations. 314
7.8 Inter-die and intra-die process variations. 315
7.9 Systematic and random variations. 315
7.10 Comparison of the patterns obtained with and without OPC. 318
7.11 Use of alternating phase shift mask. 319
7.12 Illustration of WPE. 320
7.13 Illustration of random distribution of channel dopants. . . . 321
7.14 Effects of random discrete dopants as obtained from SPICE simulation results. 323
7.15 Relation between LER and LWR. 324
7.16 Simplified model for estimating the LER of a gate. 325
7.17 Effects of line edge roughness as obtained from SPICE simulation results. 326
7.18 Effects of oxide thickness variations as obtained from SPICE simulation results. 328
7.19 Contributions of each effect on the statistical variations of threshold voltage and off-current. 329
7.20 Data spread and design corners. 330
7.21 Statistical modeling using Monte Carlo simulation. 332
7.22 Statistical corner modeling using TCAD. 333
7.23 Drain current mismatch. 336
7.24 Polarized dielectric medium. 337
7.25 Schematic illustration of the anode hole injection mechanism. 339
7.26 Percolation of defects and breakdown of ultra-thin gate oxide. 340
7.27 Schematic illustration of the generation of interface traps due to channel hot electrons. 343
7.28 Schematic illustration of the reaction-diffusion model. 346
7.29 Equivalent model of MOS transistor considering all aging mechanisms. 348
7.30 Simplified CAD flow of the simulation of the reliability of an analog circuit. 348

List of Tables

1.1 Scaling Scenarios 7

2.1 Worst Case Time Complexities of Frequently Used Functions 65

3.1 Aspect Ratios for Transistor of Channel Length $L = 65nm$ with Technology Nodes 119
3.2 Summary of OFF Current and Subthreshold Swing as Obtained from SPICE Simulation Results 133

4.1 Range of Circuit Design Parameters 172
4.2 MLP NN Architecture for Case Study 1 173
4.3 ANN Model Accuracy for Case Study 1 180
4.4 ANN Structure for Inductor Modeling Problem 182
4.5 ANN Model Accuracy for the Inductor Modeling Problem . 182
4.6 Comparison of the Total Number of Iterations and CPU Time Required for the Two Dynamic Sampling Methods to Reach Convergence for the Inverter Problem 186
4.7 ANN Model Accuracy 186
4.8 List of Kernel Functions 193
4.9 Transistor Sizes and Feasibility Constraints for Two-Stage OPAMP Circuit 202
4.10 Statistics of the Constructed Feasibility Models 203
4.11 Statistics of the Constructed Performance Models of Case Study 3 203
4.12 Transistor Sizes and Feasibility Constraints for OTA 206
4.13 Grid Search Technique Using Hold-Out Method 206
4.14 Grid Search Technique Using 5-Fold Cross Validation Method 207
4.15 GA Technique Using Hold-Out Method 207
4.16 GA Technique Using 5-Fold Cross Validation 207
4.17 Comparison between GA and Grid Search (GS) Algorithm (Algo) for LS-SVM Construction 208
4.18 Functional Specifications and Design Constraints of Case Study 5 211
4.19 Application Bounded Constraints: Case Study 5 214
4.20 Circuit Realizable Constraints: PA Block of Case Study 5 . 214
4.21 SVM Performances: Case Study 5 214

5.1 Aspect Ratios of Each Transistor of Case Study 1 229
5.2 Comparison between PSO and Simulation Results of Case Study 1 229
5.3 Synthesized Values of Inductor Layout Geometry Parameters 233
5.4 Verification of the Synthesized Inductor Geometry through EM Simulation 234
5.5 Delay Constraints and Design Parameter Bounds 236
5.6 Synthesis Results: $\tau_{in} = 1ns$ 236
5.7 Comparison between PSO Results and SPICE Results: τ_R and τ_F 237
5.8 Comparison between PSO Results and SPICE Results: τ_{PHL} and τ_{PLH} 237
5.9 Aspect Ratios and g_m/I_D Ratio of Each Transistor of Case Study 4 247
5.10 Comparison between Analytical and Simulation Results for Case Study 4 247

6.1 Leakage Current Mechanisms under Different Operating Conditions 269

Preface

With the downscaling of CMOS technology to a sub-90 nm (nano-scale) regime, several major non-idealities related to the physics and fabrication techniques of MOS transistors start playing a significant role in determining the performance of circuits. Analog integrated circuit (IC) designers are consequently faced with many new design challenges at different stages of circuit design. Some of these challenges are degradation of output resistance and intrinsic voltage gain of the transistor, enhanced gate leakage current, reduced overdrive voltage, variations of process technology parameters and ability to withstand enhanced stress over a period of time. On the other hand, the number of functionalities that are to be accommodated has also increased significantly. The application markets for integrated circuits are characterized by short product life cycles and a tight time-to-market constraint. Digital designers are fortunate enough to be able to utilize the benefits of computer-aided design techniques and associated tools to correctly develop large circuits on the first attempt. On the other hand, the majority of the tasks in analog IC design are still hand-crafted. The degree of automation varies from repeated use of SPICE simulations, manual place and route with the assistance of parameterized device generators and post-layout verification. Therefore, the combined problem of achieving both design productivity and creativity really does make the job of a nano-scale analog designer challenging. The development of computer-aided design automation tools for analog circuits has been the subject area of active research for both academia and industries over the last few decades. The design of nano-scale analog circuits requires comprehensive knowledge of the compact models of MOS transistors used for simulation purposes as well as the constraints imposed by process technology and the reliability issues. Otherwise, the verification process of analog circuits in the nano-scale regime becomes too complex. Extraction and management of knowledge, acquired either through repeated experimentation or through analyzing the physics of the processes, become quite involved. Subsequent development of specialized computer-aided design techniques and tools may simplify the task of analog designers to some extent.

The motivation for writing this book on models and CAD techniques for nano-scale analog circuit design was the limited availability of satisfactory textbook-based reference material catering to teaching related courses for post-graduate and senior under-graduate students. The authors then planned to write a book to provide comprehensive treatment of modeling the physics of MOS transistors relevant to circuit design, capturing nano-scale challenges

to circuit design and describing CAD techniques starting from the basics. Researchers and post-graduate students are considered to be the primary readers of this book. Analog circuit designers will also find the book extremely beneficial for reviewing several fundamental concepts which dictate the performance of circuits and the associated design tools to a large extent. The objective of this book is to present the links between the physics/technology of scaled MOS transistors on one end and the design and simulation of nano-scale analog circuits on the other.

The book is organized into seven chapters that encompass the area of models and CAD techniques for high-level design of nano-scale CMOS analog circuits. Chapter 1 provides an overview to the term nano-scale CMOS IC technology and the general trends of technology scaling with respect to device geometry, process parameters and supply voltage. The International Technology Roadmap for Semiconductors (ITRS) is also introduced. The critical challenges involved in an analog design process in sub-90 nm process technology are emphasized. The motivation for computer-aided design techniques is presented and the merits and demerits of the various existing design techniques are summarized. The knowledge-based CAD technique for analog circuit design is introduced. Chapter 2 provides a comprehensive description of the various types of high-level models and the useful optimization techniques. For nano-scale circuit designs, technology computer-aided design (TCAD) technique is gaining importance day by day because of its advantage in capturing accurately the various nano-scale effects related to the physics and technology of MOS transistors. This topic is briefly introduced and the advantage of combining this with computer-aided circuit design is emphasized. Various commercial design tools are also mentioned. This chapter, therefore, provides the essential background for the subsequent chapters. Chapter 3 provides a comprehensive overview of compact modeling in the context of scaled MOS transistors for VLSI circuit simulation. The root cause of the design creativity problem faced by the analog designers often lies in improper understanding of the compact models of scaled MOS transistors. Therefore, the present day designers can no longer remain ignorant of this very important topic. An outline of the essential issues related to the compact models has been discussed here. BSIM3 and BSIM4 compact models have been considered as benchmarks in the discussion. Readers are advised to consult the corresponding user guides for the details of these models. Because of the complicated physics involved with MOS transistors and the significant effects of these on circuit performance, designers often use learning-based approaches for construction of high-level performance and feasibility models. Chapter 4 introduces two very important learning-based methods: the artificial neural network (ANN) and the least-squares support vector machine (LS-SVM) method. Several case studies with simulation results have been described demonstrating the practical utilities of these two methods. Circuit sizing and specification translation tasks have been described in Chapter 5 of this book. The particle swarm optimization technique has been introduced to readers and demonstrated through

practical examples of sizing analog circuits. The g_m/I_D methodology, another very important methodology for low-power analog design has also been described in this chapter and is demonstrated through simulation results. The advanced effects of scaled MOS transistors like narrow width effects, vertical and lateral channel engineering and the gate leakage current issue have been discussed in Chapter 6. The various state-of-the-art CMOS technologies such as metal-gate, high-κ technology, silicon-on-insulator technology, double-gate MOS technology and FinFETs have been introduced to readers. These technologies are presently used to design nano-scale analog circuits for system-on-chip purpose. The noise characterization and gate-resistance effects are also briefly described. Finally, Chapter 7 presents an overview of the design challenges that occur due to statistical variations of process technology parameters and reliability constraints. Intra-die process variations have been described in detail and the effects of these on the performance parameters have been demonstrated through simulation results. The reliability constraints related to time-dependent dielectric breakdown, hot carrier injection and negative bias temperature instability have been described. An extensive list of references is provided at the end for more elaborate discussion of the issues and to motivate readers to engage in further research.

The simulation results presented in this book have been carried out at the IC Design Laboratory of the Institute of Radio Physics and Electronics, University of Calcutta, for which one of the authors (SP) acknowledges the financial support provided by the Department of Science of Technology, Govt. of India (under the Fast Track Young Scientist Scheme SR/FTP/ETA-0063/2009). The research results are primarily the outcome of this project. The first author deeply acknowledges his research students Abhijit Dana, Somnath Paul, Kritanjali Das and Sarmista Sengupta for their active technical support. He also expresses his gratitude to his wife. Srabanti Pandit (assistant professor, Electronics and Communication Engineering Department, Heritage Institute of Technology) for fruitful technical discussions on various topics of Chapter 6 of this book. Finally, the authors express their thanks to their family members for putting up with the long working hours maintained by them. All the authors gratefully acknowledge the support received from all staff members of CRC Press who have interacted with them, for their immense patience and responsiveness demonstrated throughout the publishing process of this book.

Soumya Pandit (SP)
Chittaranjan Mandal
Amit Patra

About the Authors

Soumya Pandit earned the B.Sc, degree with physics honors, M.Sc., degree in electronic science from the University of Calcutta in 1998 and 2000 respectively, and the M. Tech degree in radio physics and electronics from the same university in 2002. He earned his PhD degree from Indian Institute of Technology, Kharagpur in information technology in the year 2009. His dissertation contributed an optimization-based methodology for high-level design of analog systems. He led several research projects in the area of CMOS analog circuits and EDA, funded by industries and the government of India during his PhD research. He is currently serving as assistant professor in the Institute of Radio Physics and Electronics, University of Calcutta. His current research activities are on statistical CMOS analog circuit design and optimization, process–device–circuit interaction and soft computing applications. Dr. Pandit is currently the principal investigator of two research projects in the area of nano-scale CMOS analog design optimization, funded by Govt. of India. He is the teacher in charge of the IC Design Laboratory, a research and post-graduate level teaching laboratory at the University of Calcutta. He has to his credit several international journal and conference publications. He is a member of IEEE, USA and an associate member of the Institute of Engineers (India). His name was included in Marquis' *Who's Who in the World*, 2010 Edition.

Chittaranjan Mandal earned his PhD from the Indian Institute of Technology, Kharagpur, India, in 1997. He is currently a professor in the Department of Computer Science and Engineering and also the School of Information Technology at IIT, Kharagpur. His research interests include high-level system design, formal modeling and verification. He has been an Industrial Fellow of Kingston University, UK, since 2000 and was also a recipient of a Royal Society Fellowship for conducting collaborative research in the UK. He has handled sponsored projects from Indian government agencies such as DIT, DST and MHRD and also from private agencies such as Nokia, Natsem and Intel. He has supervised several PhD students and has over 80 international publications, some of which have attracted best paper awards and double digit citations. He also serves as a reviewer for several journals and conferences.

Amit Patra received the B.Tech., M.Tech. and PhD degrees from the Indian Institute of Technology, Kharagpur in 1984, 1986 and 1990, respectively. During 1992–93 and in 2000 he visited the Ruhr-University, Bochum, Germany

as a post-doctoral fellow of the Alexander von Humboldt Foundation. He joined the Department of Electrical Engineering, Indian Institute of Technology, Kharagpur in 1987 as a faculty member, and is currently a professor. He was the professor in-charge, Advanced VLSI Design Lab, at IIT Kharagpur during 2004 – 07. He is currently the dean of Alumni Affairs and International Relations at IIT Kharagpur. His current research interests include power management circuits, mixed-signal VLSI design and automotive embedded control systems.

He has guided 15 doctoral students and published more than 200 research papers in various journals and conferences. He is the co-author of a research monograph entitled *General Hybrid Orthogonal Functions and Their Applications in Systems and Control*, published by the Springer Verlag in 1996. He has carried out more than 40 sponsored projects mostly in the areas of VLSI and power management circuits and control systems. He has been consulted by National Semiconductor Corporation, Infineon Technologies, Freescale Semiconductor and Maxim Corporation in the power management area. In the area of control systems he has collaborated with ISRO, DRDO, ADE, ADA, General Motors and General Electric. As the professor in-charge of the Advanced VLSI Design Laboratory, he also took the lead role in the formation of the AVLSI Consortium at IIT Kharagpur to increase collaboration between industry and the academic community in the area of VLSI.

Dr. Patra received the Young Engineer Award of the Indian National Academy of Engineering in 1996 and the Young Teachers Career Award from the All India Council for Technical Education in 1995. He was a Young Associate of the Indian Academy of Sciences during 1992–97. He is a member of IEEE (USA), the Institution of Engineers (India) and a life member of the Systems Society of India.

1

Introduction

1.1 Introduction

With the advent of nano-scale CMOS technology (sub-90nm), complex System-on-Chips (SoCs) are used in almost all the domains of electronic systems such as telecommunications, multimedia, consumer electronics, instrumentation and defense applications [24]. Despite the trend for replacement of analog functionalities within a SoC by digital signal processing operations, analog circuits are considered to be indispensable for all the applications that interface with the outer world such as interfaces with sensors, microphones, antennas, actuators, loud speakers etc. In addition, there are some mixed signal circuits, such as data converters, which contain analog components [75]. Moreover, high performance (high-speed and low power) digital circuits are often designed in an analog fashion [21].

The design complexity of an integrated circuit (IC) has increased drastically in the nano-scale domain. Several second-order effects related to nano-scale MOS transistors which were hitherto considered to be insignificant now play dominant role in determining circuit performances [66]. On the other hand, for economic reasons, the majority of the SoC application markets are characterized by shortening product life cycles and tightening time-to-market constraints. This pressure leads to the use of efficient computer-aided design (CAD) methodologies and associated design automation tools by the IC designers. In the digital domain, the design automation tools are fairly matured and commercially available. However, for analog circuits, the scenario is not so impressive [8]. Developments in the area of analog CAD tool developments are primarily in the research phase, where most of the CAD tools have been research prototypes tested on a limited set of circuits [65]. Therefore, in a complete mixed-signal integrated circuit, although analog circuits typically occupy only a small fraction of the total die area, their design is often the bottleneck, both in design time and effort [8, 66].

1.2 Characterization of Technology Scaling

The MOS transistor feature size has been subjected to scaling down for the last few decades. The degree of scaling is measured as the half-pitch of the first-level interconnect in DRAM technology [17]. For logic circuits, the smallest feature size refers to the length of the gate of a MOS transistor. These are used to characterize a technology node. The concept of pitch and technology node are illustrated in Fig. 1.1. It is clear that half-pitch = 2λ = technology node. Examples of technology nodes are $0.18\mu m, 0.13\mu m, 0.1\mu m, 90nm, 65nm, 45nm$ and so on. In each technology node, the feature sizes such as the contact holes in the layout of a circuit are reduced by 70% of the corresponding sizes in the earlier technology node. Consequently with each new technology generation, the circuit area is reduced by $50\%, (0.7 \times 0.7 = 0.49)$. The practice of the periodic reduction of the feature size is referred to as technology scaling. It may be noted that with advancement of CMOS device structures, the equality between the half-pitch and the technology node is violated, and today Half-Pitch > Node [91].

The most straightforward benefit of technology scaling is that with introduction of new technology the integration capacity increases twofold and thus the cost per circuit is reduced significantly. This is Gordon Moore's law [14], according to which the complexity of MOS device integration is approximately doubled every eighteen months.

There are two major theoretical models for device scaling (i) constant field scaling and (ii) constant voltage scaling. These are discussed below

1.2.1 Constant Field Scaling

The principle of the constant field scaling is the scaling of the device voltages and dimensions (both vertical and lateral) by the same factor κ so that the electric field remains constant [43]. The doping concentration is increased by the same scaling factor κ in order to keep Poisson's equation invariant with respect to scaling. The constant field scaling ensures that the reliability of the scaled device is not degraded compared to that of the original device.

With the scaling down of the supply voltage and transistor dimensions, a significant effect of the constant electric field scaling is that the circuit speeds up by the same factor κ and the power dissipation is reduced by κ^2. In addition, the power density remains constant in the scaled transistor. The power-delay product improves by a factor κ^3.

1.2.2 Constant Voltage Scaling

In spite of its significant advantages, it has been found that in reality the constant field scaling is not a feasible option. In order to keep the new devices

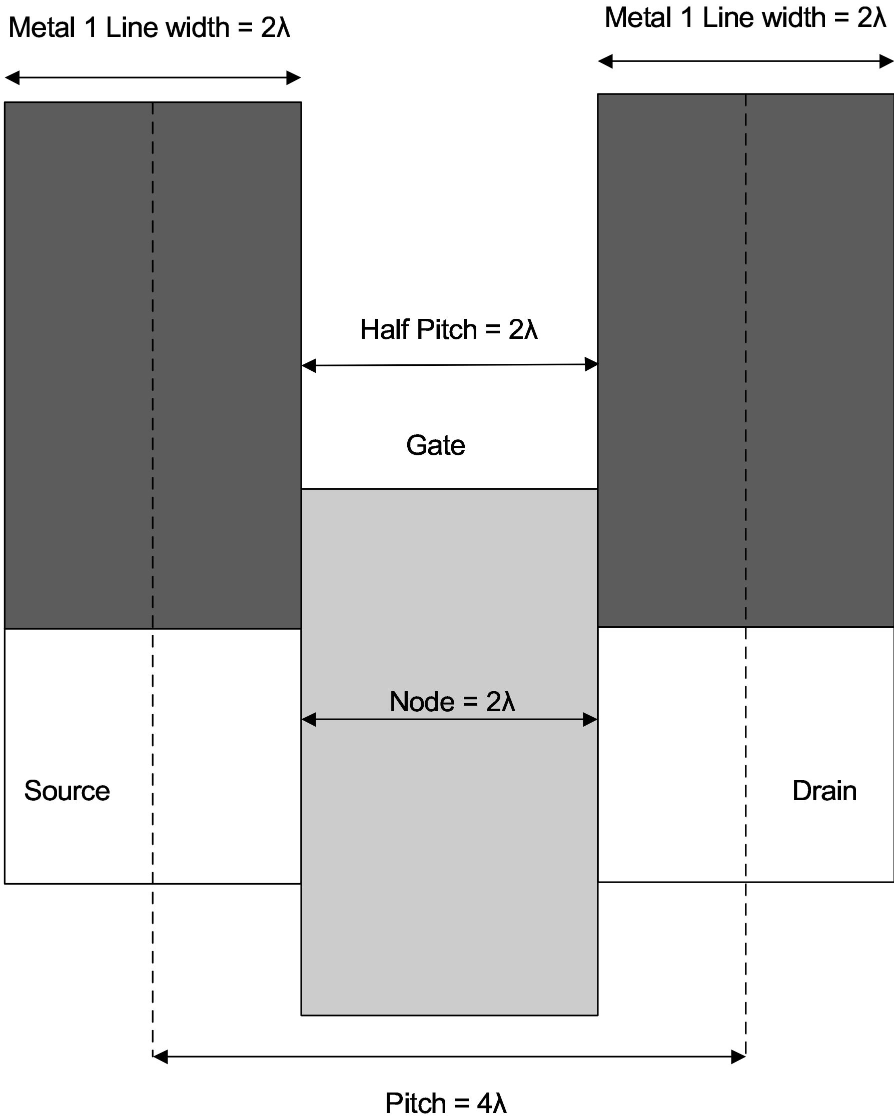


FIGURE 1.1
Illustration of half-pitch and technology node.

compatible with existing components, voltages cannot be scaled arbitrarily. This necessitates multiple supply voltages which leads to considerable increase of the cost of the system. Therefore, in the constant voltage model, the voltages are kept constant [26]. The geometrical dimensions and doping, are however kept as in the case of the constant field scaling. Under the constant voltage model of scaling, the electric field is scaled up by the factor κ and the doping concentration needs to be scaled up by κ^2.

In reality, CMOS technology evolution has followed a suitable combination of the constant field and constant voltage scaling.

1.2.3 Nonscaling Effects

In CMOS technology, the material related parameters such as energy gap, work function etc., do not change with scaling. As a result of this, the threshold voltage of a MOS transistor does not change with technology scaling, in general [192]. This is explained by the basic definition of threshold voltage [192] as follows

$$V_T = V_{FB} + 2\Phi_F + \frac{\sqrt{2\epsilon_{Si} q N_A \left(2\Phi_F + V_{BS}\right)}}{C_{ox}} \tag{1.1}$$

where V_{FB} is the flat band voltage, V_{BS} is the substrate voltage and $2\Phi_F$ is the surface potential at strong inversion. Although the channel doping and the oxide thickness are scaled, they almost mutually nullify each other. The other material-related parameters are independent of technology scaling. Therefore, the threshold voltage of a MOS transistor does not scale with technology scaling.

The OFF current of a MOS transistor is given as follows [192]

$$I_{DS}(V_{GS} = 0, V_{DS} = V_{DD}) = \mu_s C_{ox} \frac{W}{L} (\eta - 1) \left(\frac{kT}{q}\right)^2 \exp\left(-qV_T/\eta kT\right) \tag{1.2}$$

where η is the subthreshold swing factor. It is observed that because of the exponential dependence, the threshold voltage cannot be scaled down without significant increase in the OFF-current. However, even if the threshold voltage is kept constant, the OFF current of a transistor increases because of the increase of the oxide capacitance per unit area with the scaling of the oxide thickness.

The nonscaling trend of threshold voltage restricts the scaling of the supply voltage. This is because of the fact that the intrinsic delay of a MOS transistor increases rapidly with the ratio V_T/V_{DD} when the latter exceeds about 0.3.

The inversion layer thickness remains unchanged with constant field scaling. Considering the fact that the inversion layer capacitance acts in series with the gate capacitance, the total gate capacitance per unit area of a scaled MOS transistor increases by a factor less than κ.

The junction built-in potential and the surface potential do not change significantly with technology scaling. The maximum gate depletion width for

a long channel MOS transistor is given by

$$W_{dm} = \sqrt{\frac{4\epsilon_{Si} kT \ln{(N_A/n_i)}}{q^2 N_A}} \tag{1.3}$$

Here $\ln{(N_A/n_i)}$ is weak function of N_A and can be treated as a constant. Therefore, the depletion width does not scale down as significantly as the other transistor dimensions. In order to scale down the depletion depth, the substrate concentration needs to be scaled up more than that suggested by the constant-field scaling or generalized scaling. If this can be done, the subthreshold swing factor can be kept constant with scaling.

1.2.4 Generalized Scaling and Technology Trends

According to the generalized scaling model [13], the scaling factor need not be the same for the geometrical feature sizes and potentials. The scale factor for the potentials may be $\alpha \neq \kappa$. When $\alpha = \kappa$, the generalized scaling model reduces to the constant field scaling model. On the other hand, when $\alpha = 1$, the generalized model reduces to the constant voltage model.

The technology trend is such that the various device parameters are allowed to be adjusted independently as long as the overall behavior is preserved. The fundamental idea of device scaling is to design the device with scaled technology such that the long-channel behavior of MOS transistors is preserved as far as possible and circuit performance benefits are achieved.

1.2.4.1 International Technology Roadmap for Semiconductors

The International Technology Roadmap for Semiconductors (ITRS) is an annually updated document prepared by semiconductor researchers around the world to generate consensus on the transistor and circuit performances that are required to fulfill the projected markers in the future [90]. The ITRS working group have published many device design targets for comprehensive development of MOS technology. Based on the types of applications, the ITRS working group have classified the technology parameters into two types: high-performance (HP) and Low Power (LP). High-performance (HP) refers to technology required for chips of high performance and high power dissipation such as microprocessor units (MPU) for desktop PCs. The low power technology refers to the technology required for chips for mobile applications where the allowable power dissipation and hence the allowable leakage currents are limited by battery life. The low power technology is sub-divided into two types: low operating power (LOP) and low stand-by power (LSTP). LOP refers to the technology required for chips of relatively high performance mobile applications such as notebook computers where the battery is likely to be of high capacity. On the other hand, the LSTP chips are typically for low power performance consumer type applications with lower battery capacity, e.g., cellular phones. Therefore, the transistors for high performance ICs

are highly scaled and have somewhat higher performances and high leakage current. The transistors for LOP applications are less highly scaled and have somewhat lower performance and much lower leakage current. On the other hand, the transistors for LSTP are scaled the same as those for LOP but the performance and leakage current are lower still, compared to transistors for LOP chips.

1.2.4.2 Predictive Technology Modeling

For early prediction of CMOS devices and circuit performances for future technologies, a research group from Arizona State University has formulated a set of predictive technology models (PTM) based on the standard compact device model framework for circuit simulation purposes [210]. These models are therefore, significantly helpful for the IC designers for prediction purposes with migration of CMOS technologies and the use of new materials.

1.2.4.3 Scaling of Geometry Parameters

The scaling of the physical gate length and the effective channel length with a technology node is demonstrated in Fig. 1.2 as observed from the ITRS as well as a set of PTM models. The scale factors are given in Table 1.1. It is observed that with scaling the physical gate length is reduced by a factor of 0.7246. This is accordance with the principle mentioned earlier.

The scaling of equivalent oxide thickness with a technology node is highlighted in Fig. 1.3. It is observed that the oxide thickness is scaling down as

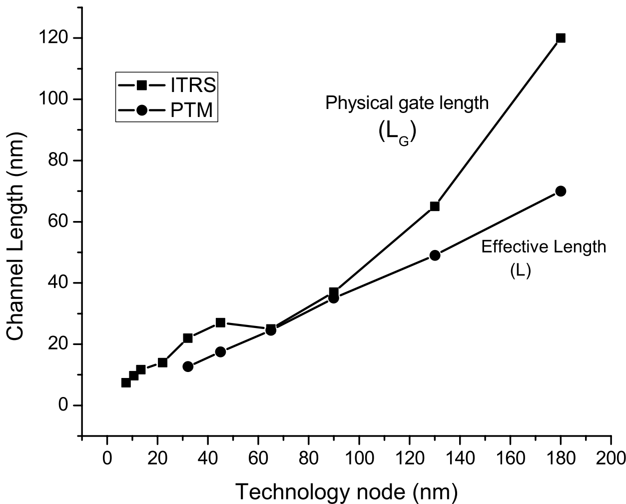


FIGURE 1.2
Scaling of physical gate length and effective channel length across technology node.

TABLE 1.1
Scaling Scenarios

Physical Parameters	Approximate scaling relation $\kappa > 1, \alpha > 1$	HP	LSTP	PTM
Physical gate length, L_G	$L_G/\kappa \downarrow$	1.38	1.38	–
Effective channel length L_{eff}	$L_{eff}/\kappa \downarrow$	–	–	1.40
Equivalent oxide thickness t_{ox}	$t_{ox}/\kappa \downarrow$	1.25	1.19	1.42
Source/Drain junction depth x_j	$x_j/\kappa \downarrow$	1.09	1.13	1.44
Channel doping N_{CH}	$(\kappa^2/\alpha)N_{CH} \uparrow$	1.13	1.13	1.45
Supply voltage V_{DD}	$V_{DD}/\alpha \downarrow$	1.08	1.07	1.11
Long channel threshold voltage V_T	$V_T/\alpha \downarrow$	–	–	1.05
Drain ON current I_{ON}	$\left(\kappa/\alpha^2\right) I_{ON} \uparrow$	1.28	1.25	1.06
Drain OFF current I_{OFF}	$I_{OFF} \uparrow$	2.74	3.08	2.39
Intrinsic delay $D = C_G V_{DD}/I_{ON}$	$(\alpha/\kappa^2)D \downarrow$	0.68	0.60	0.70
Static power dissipation P_S	$P_S \uparrow$	1.81	1.09	–
Parasitic resistance R_{dsw}	$R_{dsw} \downarrow$	1.17	1.26	1.11
Saturation velocity v_{dsat}	$(\kappa/\alpha)v_{dsat} \uparrow$	–	–	1.075

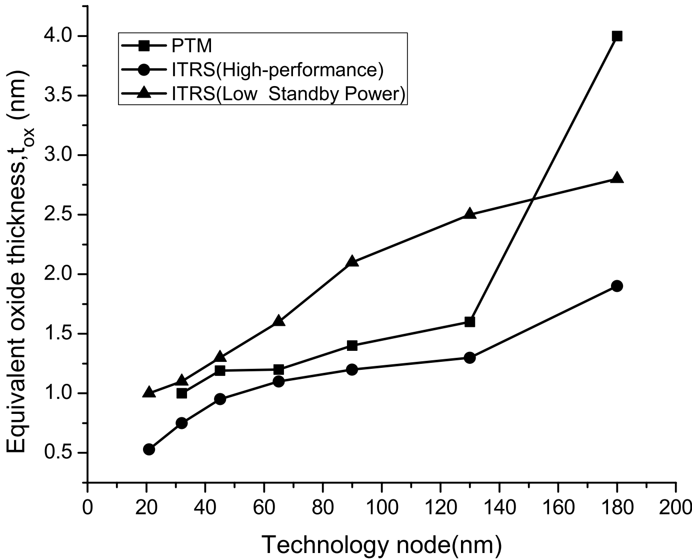


FIGURE 1.3
Scaling of equivalent oxide thickness across technology node.

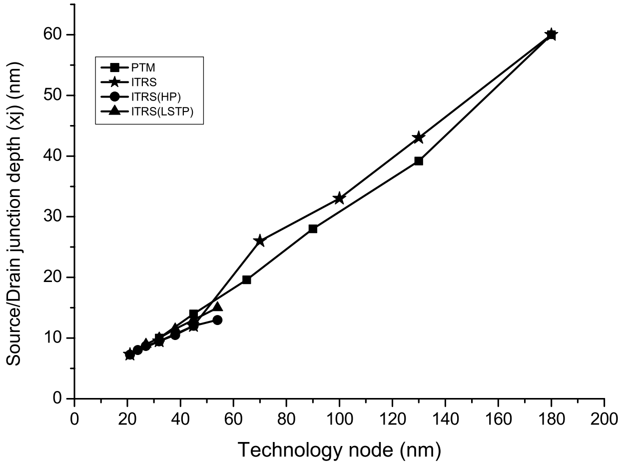


FIGURE 1.4
Scaling of source-drain junction depth across technology node.

the technology is scaled. The scale factors are given in Table 1.1. For the HP logic, the required oxide thickness is smaller compared to that required for the LSTP logic. The reason is that with the scaling of oxide thickness, the intrinsic delay of a transistor reduces, so that the intrinsic speed is increased. However, this comes at the cost of enhanced leakage power. Therefore, for low standby power applications, comparatively thicker oxide thickness is to be used. It is observed that for the HP logic, the general trend of oxide thickness scaling with technology node is approximately 0.02 times the technology node for nodes below 65nm and 0.01 times the technology node for nodes above 65nm.

The variation of the source/drain junction depth with technology node is shown in Fig. 1.4. It is observed that the depth is reducing with technology scaling which is in accordance with the scaling theory. The scale factors are given in Table 1.1. The reduction is essential for short channel effect minimization.

1.2.4.4 Scaling of Channel Doping, Supply Voltage, and Threshold Voltage

The variation of the channel doping with technology node as obtained from the ITRS specifications for HP and LTP logic and various PTM files is demonstrated in Fig. 1.5. The scale factors are shown in Table 1.1. This is in consistency with the scaling principle.

The variation of the supply voltage with technology node is shown in Fig.1.6. It is observed from the graph that the supply voltage changes slightly with technology scaling. The reason is due to the fact that the threshold

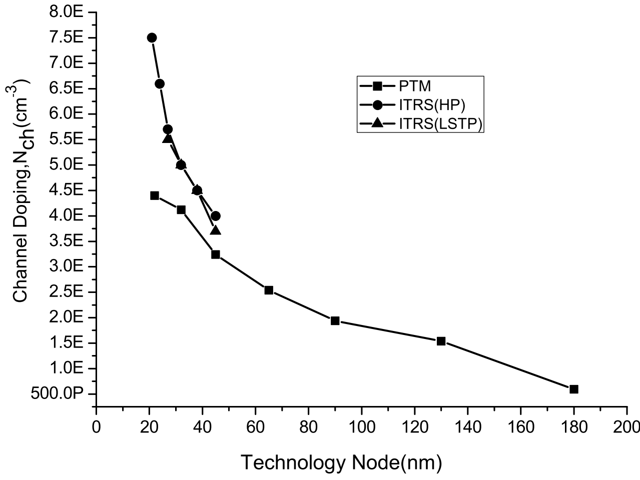


FIGURE 1.5
Scaling of channel doping across technology node.

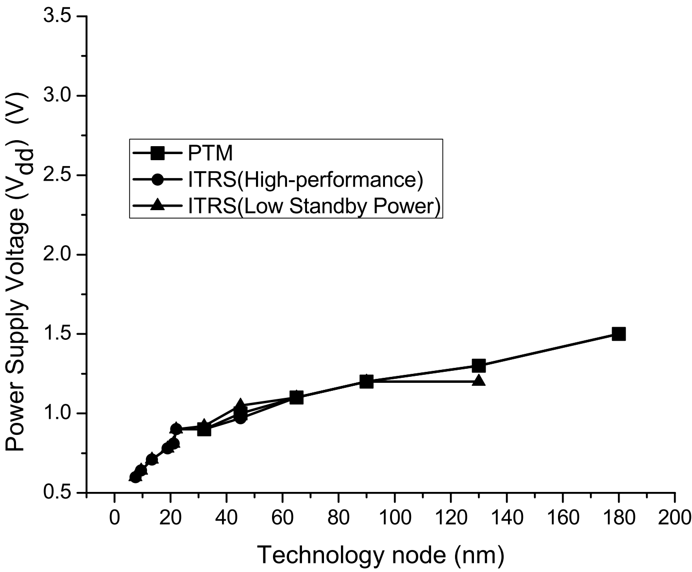


FIGURE 1.6
Scaling of the power supply voltage across technology node.

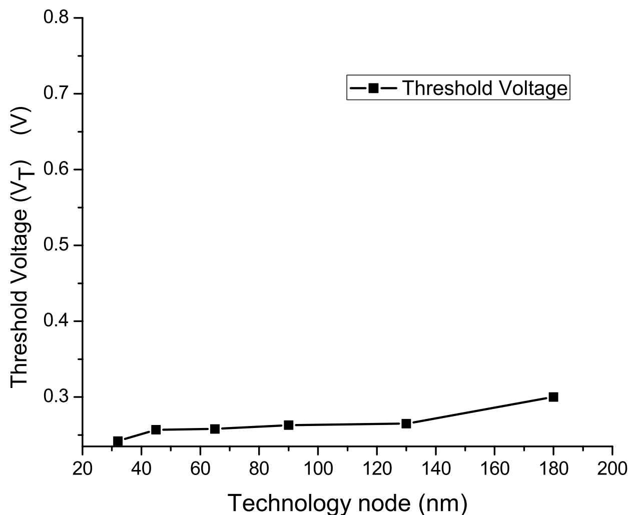


FIGURE 1.7
Scaling of threshold voltage across technology node.

voltage of a MOS transistor does not scale down and the delay of a transistor increases with the ratio V_T/V_{DD}.

The variation of the long-channel threshold voltage of a MOS transistor with technology node as obtained from the various PTM files is shown in Fig. 1.7. The scale factors are shown in Table 1.1. This shows that the threshold voltage remains almost constant with scaling. The theoretical reason for this is discussed as previously.

1.2.4.5 Scaling of Performances

The variation of the drain current with technology node is shown in Fig. 1.8. It is observed that the drain current is increasing which is in accordance with the constant voltage model. The scale factors are shown in Table 1.1. The drain current is much higher for the HP logic compared to that of the LSTP logic. This is because with high drain current, the intrinsic speed of a MOS transistor increases.

The variation of the OFF current with technology node is shown in Fig. 1.9. The scale factors are shown in Table 1.1. For the LSTP logic, the OFF current is much lower compared to that of the HP logic. Even if the threshold voltage is held unchanged, theoretically the OFF current increases by a factor κ from the C_{ox} contribution when the physical dimensions are scaled down by κ. However, it has been found that the OFF current increases at a rate much higher than κ. This is because of the fact apart from the subthreshold leakage current component several other components, like tunneling leakage current, contribute to the OFF current [154].

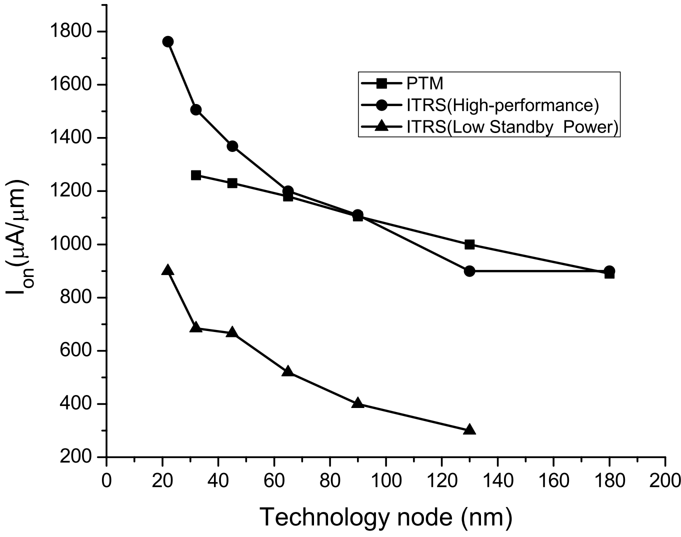


FIGURE 1.8
Scaling of the drain ON current across technology node.

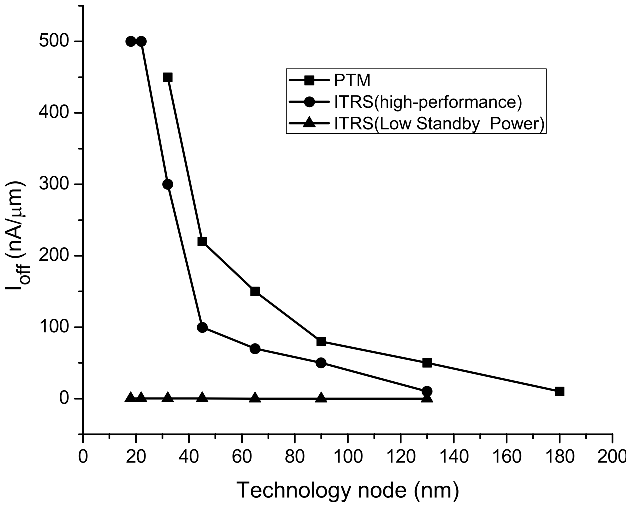


FIGURE 1.9
Scaling of the drain OFF current across technology node.

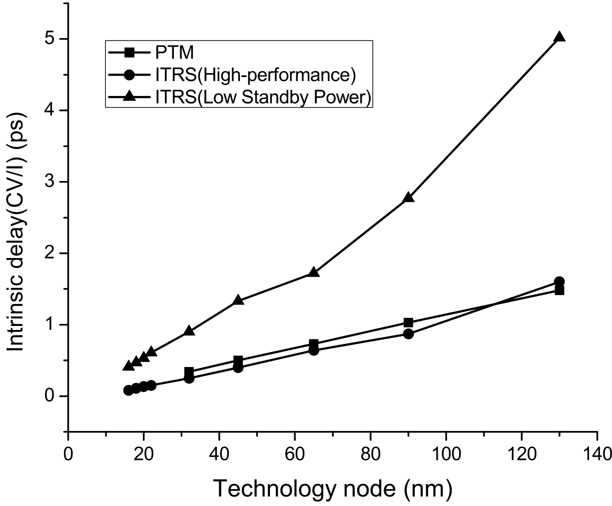


FIGURE 1.10
Scaling of the intrinsic delay across technology node.

The scaling of the intrinsic delay with technology node is shown in Fig. 1.10. The scale factors are shown in Table 1.1. The magnitude of the delay is lower in the HP logic compared to that in the LSTP logic. The scale factors are shown in Table 1.1.

The scaling of the static power dissipation with technology node is shown in Fig. 1.11. The static power dissipation increases with technology scaling. The scale factors are shown in Table 1.1. It is observed that for the HP logic, the static power dissipation is several times that of the LSTP logic. With technology scaling, the increase of static power dissipation is a critical challenge to the IC designers.

1.2.4.6 Scaling of Source-Drain Resistance and Saturation Velocity

The variation of the parasitic source-drain resistance (R_{dsw}) with technology node is shown in Fig. 1.12. The scale factors are shown in Table 1.1. The reduction of R_{dsw} becomes more difficult in short-channel transistors.

The scaling of the saturation velocity with technology node is shown in Fig.1.13. The scale factors are shown in Table 1.1.

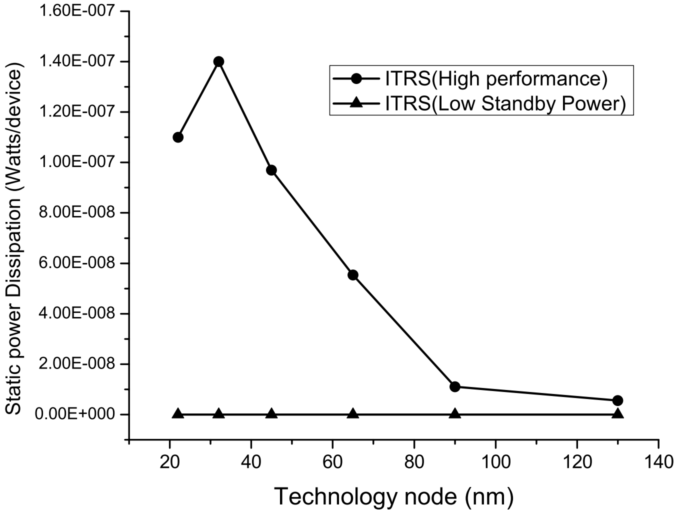


FIGURE 1.11
Scaling of the static power dissipation across technology node.

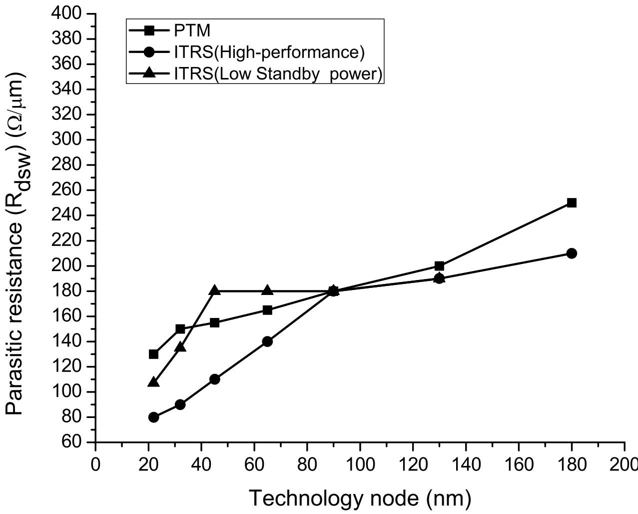


FIGURE 1.12
Scaling of source-drain junction depth across technology node.

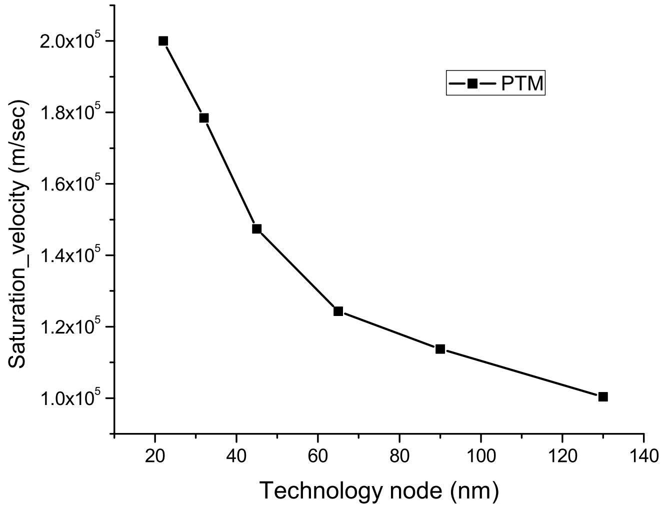


FIGURE 1.13
Scaling of the saturation velocity across technology node.

1.3 Analog Design Challenges in Scaled CMOS Technology

In scaled CMOS technology, several significant effects, referred to as short channel effects of MOS transistors, play a major role in determining the performances of analog circuits [112, 7]. Some important short channel effects of scaled MOS transistors are threshold voltage roll off and drain induced barrier lowering (DIBL), carrier mobility degradation, velocity saturation and gate oxide leakage current [192]. These effects critically affect the performances of nano-scale analog circuits. In addition, with the scaling of supply voltage to 1V, the signal headroom becomes too small to design circuits with sufficient signal integrity at reasonable power consumption levels. Therefore, technology scaling brings about new features but only a few of them are good for analog applications and are true for the designers.

1.3.1 Degradation of Output Resistance and Intrinsic Gain

The output resistance of a MOS transistor is a very important parameter for analog IC design. Ideally the output resistance of a MOS transistor is infinite so that the transistor acts as an ideal current source. However, with the increase of drain bias, the current of a real MOS transistor increases so that the resistance has a finite value. For long channel MOS transistor, the finite output resistance is attributed to channel length modulation effect [192]. For

short channel length MOS transistors, as the drain-source bias is increased, the threshold voltage value reduces, as a result of which the drain current increases. This is referred to as the drain induced barrier lowering (DIBL) effect [192]. The output resistance of a scaled MOS transistor is therefore determined by the combined effects of the channel length modulation effect and the DIBL effect. As a result of this, the output resistance of a scaled MOS transistor is significantly lowered. The deterioration is enhanced with scaling. Consequently the intrinsic voltage gain of a MOS transistor is significantly low. It is considered as one of the major challenges to high-performance analog circuit design in scaled CMOS technology. Suitable measures such as the use of cascodes, cross coupled MOS transistors, bootstrapping etc., need to be taken for enhancement of the circuit gain [164]. In addition, a general practice for analog design is to use transistors with gate lengths $3-5$ times larger than the minimum for a specific technology node.

1.3.2 Gate Oxide Leakage Current

The thin oxide required for scaled CMOS technologies causes tunneling of the carriers such that the gate current becomes non-negligible [192]. It has been observed that at 65nm technology node, the gate leakage current increases by more than six orders of magnitude as compared to that for $0.18\mu m$ technology node. The gate current is caused due to direct tunneling through the thin gate oxide and depends mainly on gate-source voltage bias and gate area [108]. The current gain I_{DS}/I_{GS} is very high for non-scaled technology, such that the MOS transistor acts as a voltage controlled device. However, in scaled CMOS technology, as the gate current increases, the current gain reduces and the MOS transistor operates similar to a bipolar device, with performances controlled by the gate current. An obvious implication of non-zero gate leakage current is that the gate impedance of a MOS transistor is no longer purely capacitive, rather it contains a tunnel conductance in parallel with the traditional capacitance. A characteristic frequency f_{gate} is defined such that [7]

$$f_{gate} = \frac{g_{tunnel}}{2\pi C_{in}} \quad (1.4)$$

$$\approx \beta.10^{16}.v_{GS}^2 \exp\left[t_{ox}\left(v_{GS} - 13.6\right)\right] \quad (1.5)$$

where $\beta \approx 1.5$ for NMOS transistors and ≈ 0.5 for PMOS transistors. For signal frequencies higher than f_{gate}, the input impedance of the MOS transistor can be considered to be capacitive, otherwise, it is resistive and the gate leakage current is dominant. The value of f_{gate} increases with scaling. It has been found that $f_{gate} \approx 1MHz$ at 65nm technology compared to 0.1Hz in 180nm technology [7]. Therefore, leakage current plays significant role for input frequencies lower than $1MHz$ at 65-nm technology.

1.3.3 Noise Performance

The input referred noise of a MOS transistor primarily consists of the thermal noise component and the flicker noise component [70]. The thermal noise component is given as [70]

$$v_{dn}^2 = 4kT\gamma \frac{1}{g_m} B_n \tag{1.6}$$

$$g_m = \sqrt{2\mu_s C_{ox} \frac{W}{L} I_{DS}} \tag{1.7}$$

where k is the Boltzmann constant, T is the absolute temperature, B_n is the noise bandwidth and $\gamma = 2/3$ for long-channel MOS transistor. Keeping the W/L ratio and the drain current I_{DS} constant, with reduction of oxide thickness, the white noise component reduces. However, the transconductance of a MOS transistor is a critical issue in scaled technology because of several physical phenomena like mobility degradation and velocity saturation, so also is the thermal noise component. The flicker noise component is given as [70]

$$v_n^2(f) = \frac{K_F}{C_{ox} W L f} \tag{1.8}$$

where K_F is the flicker noise coefficient with the unit $V^2 - F$. If it is assumed that K_F does not change with scaling, there is an improvement in the noise provided the device area is kept constant. The flicker noise component depends upon the oxide thickness and quality of the deposited oxide layer which improves with the scaling of technology. However, in order to prevent gate leakage current, the gate dielectric is often made up of material with a high dielectric constant, that unfortunately increases the flicker noise component.

1.3.4 Analog Power Consumption

Let us consider a simple single transistor amplifier as shown in Fig. 1.14. The total thermal noise integrated across the band of interest, i.e. the noise bandwidth B_n is given by

$$i_{dn}^2 = 4kT\gamma g_m B_n \tag{1.9}$$

where $\gamma = 2/3$ for a long-channel MOS transistor. The output noise voltage is therefore,

$$v_{dn}^2 = R^2 i_{dn}^2 \tag{1.10}$$

Let us assume that the full scale output voltage across the load capacitor C_L is V_{FS}, Therefore, the maximum rms sine wave voltage is

$$v_s = \frac{V_{FS}}{2\sqrt{2}} \tag{1.11}$$

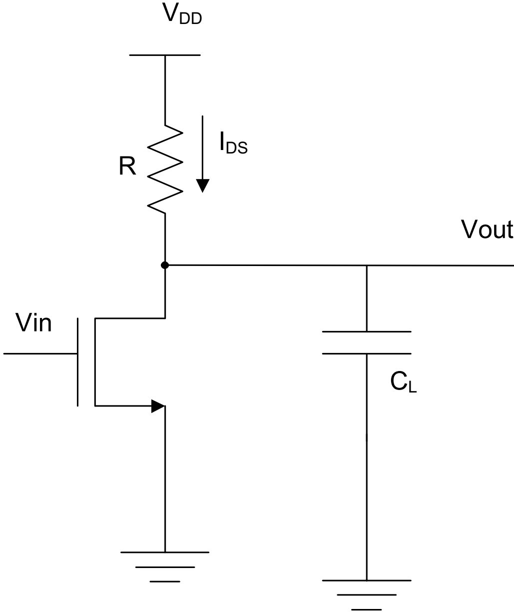


FIGURE 1.14
Single transistor amplifier.

The signal-to-noise ratio is therefore,

$$\text{SNR} = \frac{V_{FS}^2}{32kTR^2\gamma g_m B_n} = \frac{V_{FS}^2 g_m}{32kT\gamma A_v^2 B_n} \tag{1.12}$$

where $A_v = g_m R$ is the dc gain of the amplifier, assuming that the output resistance of the transistor is very high. The transconductance g_m of the MOS transistor is related to the bias current through $I_{DS} = g_m v_{OD}$ where v_{OD} is the equivalent effective overdrive voltage [188]. For a classical long-channel MOS transistor operating in a strong inversion region, $v_{OD} \approx (v_{GS} - V_T)/2$ and in the weak inversion region $v_{OD} = nkT/q$ with $n \approx 1.3$. The power consumption of the circuit is given by [188]

$$P = I_{DS}.V_{FS} = 32kT\gamma A_v^2 B_n \text{SNR} \frac{v_{OD}}{V_{FS}} \tag{1.13}$$

It is thus observed that the power consumption is proportional to the targeted SNR. Therefore, lowering the supply voltage (and hence the full scale output voltage) without reducing v_{OD} increases the power consumption. It is therefore concluded that keeping the performance requirement constant, i.e., SNR and B_n constants, if a design is migrated to scaled technology with reduced supply voltage, the power consumption of the circuit increases [188].

1.3.5 Drain Current Mismatch

Matching between MOS transistors is an important requirement for several analog circuit blocks such as the current mirror, operational amplifiers, data converters etc. The mismatch of the drain current between two adjacent identical MOS transistors is primarily attributed to threshold voltage mismatch caused by a random discrete dopant effect [143]. For long-channel MOS transistors, the dopant fluctuation yields a drain current mismatch. The variance is approximately given by [205].

$$\frac{\sigma^2_{I_{DS}}}{I^2_{DS}} = \frac{D_I}{WL}\frac{4}{(V_{GS} - V_T)^2} \tag{1.14}$$

where

$$D_I = \frac{1}{3}\sqrt{4q^3\epsilon_{Si}\psi_s N_A}\left(\frac{t_{ox}}{\epsilon_{ox}}\right)^2 \tag{1.15}$$

where ψ_s is the surface potential and N_A is the acceptor concentration. It has been found that the mismatch coefficient is directly proportional to the oxide thickness and inversely proportional to square root of the gate area. Therefore, for circuits using short channel length and narrow width transistors, mismatch becomes a critical challenge. The other sources of mismatch due to line edge roughness and oxide thickness variations are becoming significant in scaled CMOS technology [52].

1.3.6 Transition Frequency

The increase of transition frequency, alternatively referred to as the unity gain current frequency, is the most relevant analog benefit. The transition frequency is given as [70]

$$f_T \approx \frac{g_m}{2\pi\left(C_{GS} + C_{GD}\right)} \propto \frac{1}{L_G^2} \tag{1.16}$$

The transition frequency is almost inversely proportional to the square of the gate length. Therefore, with the scaling of technology, the transition frequency of a MOS transistor is significantly improved. This favors the use of CMOS analog circuits for millimeter-wave applications.

1.3.7 Reliability Constraints

With the scaling of CMOS technologies, the yield and the reliability of integrated circuits becomes a critical challenge to the designers [60]. Smaller devices combined with new materials are the cause of the increasing yield and reliability problems [112]. For an ultra-scaled MOS transistor, even with reduced supply voltage of nearly 1V, the strong electric field across the gate oxide can cause damage leading to dielectric breakdown. In addition to this,

with high electric fields, a phenomenon referred to as the hot carrier injection (HCI) occurs. HCI manifests itself mainly as a threshold voltage shift. Degradation of carrier mobility and a change of output resistance is also observed. For p-channel MOS transistors, stressed with negative bias at an elevated temperature, a phenomenon referred to as negative bias temperature instability (NBTI) occurs. NBTI is typically seen as a threshold voltage shift. Degradation of channel carrier mobility is also observed.

1.4 Motivation for CAD Techniques

The two major driving forces which motivate all the research and development activities in the area of computer-aided design techniques for integrated circuit design are (i) an increasing design productivity gap and (ii) a design creativity gap [66, 106]. These two are discussed in the following sub-sections.

1.4.1 Design Productivity Gap

A famous statistical observation is that the number of transistors used in an integrated circuit increases by 58% per year while the capability of the designers to design them increases by only 21% [106]. The fact that the number of available transistors are growing faster than the ability to meaningfully design them is referred to as the design productivity gap. It is illustrated in Fig.1.15(a). The cost of a design is considered to be the greatest threat to the continuation of the progress of semiconductor roadmaps. The primary causes for this continual increase of design productivity gap are (i) increased design complexity in the nano-scale process technology and (ii) reduced time-to-market factor.

The increase of design complexities may be attributed to the following reasons [66].

1. With the scaling of process technology, the number of transistors per chip have increased significantly.
2. Present day applications demand integration of several new functionalities and corresponding system architectures in the same system.
3. The challenges involved within the integrated circuit design procedure using nano-scale CMOS technology have increased several manifold.
4. Statistical variations of process parameters and device leakage current become critical in determining the manufacturability and yield of integrated circuits.

5. The design team becomes large, requiring many kinds of expert knowledge.

The design time eventually increases as the design procedure becomes complex.

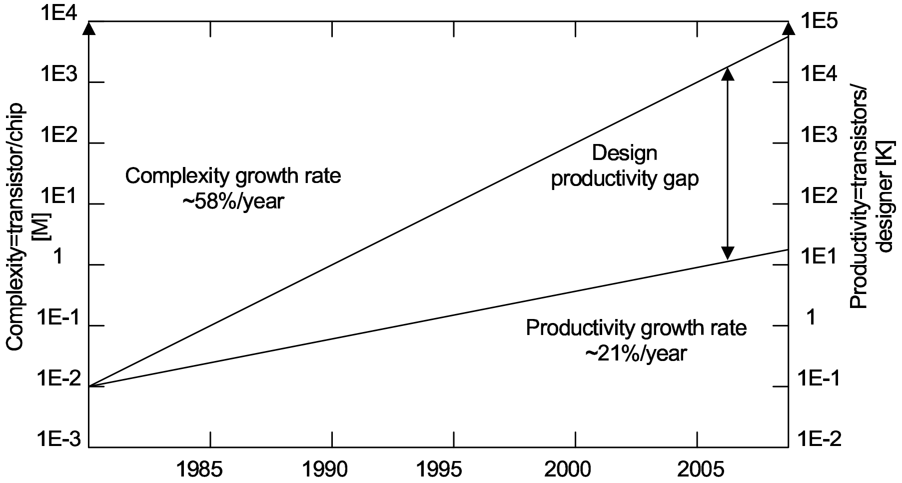
The time-to-market (TTM) factor is defined as the time required to get a product to the market starting from its conceptualization [66]. The TTM factor is very important for SoC markets for economic reasons in the sense that if a vendor misses the initial market window relative to the competition, prices and therefore profit, can be seriously eroded. The smaller the TTM factor, the more revenue a vendor can earn.

The key to managing this increased design complexity while meeting the shortening time-to-market factor is the use of well defined and accurately characterized computer-aided design methodologies and design automation tools [66].

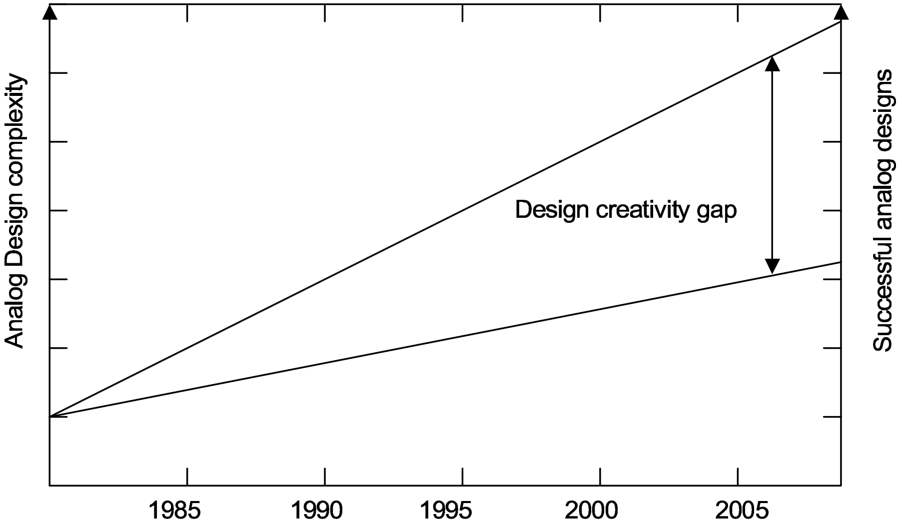
1.4.2 Design Creativity Gap

For analog circuits, it may however, be noted that it is not the size of the design that creates challenge, though it is increasing day by day. Rather, it is the difficulty of achieving the desired performances by the analog circuits designed with scaled CMOS technology. The analog circuits embedded within an SoC are very sensitive to non-idealities and all kinds of higher order effects, including parasitic effects such as crosstalk, substrate noise, supply noise etc. For analog designs, it has not yet been possible to construct a rigorous higher level of abstraction which will efficiently shield out the effects of device-level and process-level details on the circuit performances. The situations become worse in the sub-90nm process technology, where the complex physics of MOS transistors critically affects the performances of analog circuits. This is currently taken care of by the experienced designers using their intuition and creativity, which are virtually impossible to generalize and automate. Therefore, for analog designs, a new type of gap occurs which is the difference between the expected circuit performances and the performances actually achieved using available design automation tools. This is referred to as the design creativity gap. It is illustrated in Fig.1.15(b) It may be noted that the design creativity gap becomes increasingly significant for customized digital circuits too, while using nano-scale CMOS technologies [21].

Development of electronic design automation (EDA) tools alone do not solve the problem of sufficient and quick design of nano-scale analog circuits. It is essential to develop accurate and well-structured computer-aided design techniques. The design techniques suitable for nano-scale IC design must include well documented knowledge of the fundamental physics of nano-scale MOS transistor, standard libraries constructed through accurate characterization of device performances, numerical simulation and optimization techniques with associated tools. It may be noted that the CAD techniques and



(a) Design productivity gap



(b) Design creativity gap

FIGURE 1.15
Driving forces behind research works in analog CAD.

the associated design automation tools are not meant for replacement of the the designers, rather than to aid the designers to be creative, highly productive and develop large circuits with high probability of first-time success.

1.5 Conventional Design Techniques for Analog IC Design

This section presents a comprehensive discussion of the two commonly used design techniques for analog circuits, including their salient features and shortcomings in the context of increased design productivity and the design creativity gap.

1.5.1 Bottom-Up Design Technique

The most commonly used design technique for analog circuit design is the bottom-up design technique. Typically it consists of several steps: (i) defining circuit inputs and outputs (formulations of circuit specifications), (ii) selection of suitable circuit topology and preliminary determination of the transistor dimensions through hand calculations, (iii) simulation of the circuit and fixation of the transistor dimensions, (iv) geometrical layout design, (v) simulations including the geometrical layout parasitics, (vi) fabrication and (vii) testing and verification. A flow chart for the bottom-up design technique is given in Fig. 1.16.

The designer is responsible for all the steps except the fabrication. In the first step, the inputs and the outputs of the circuit are synthesized. The next step is to select a suitable topology which is able to meet the specifications. The quality of the choice is based upon the designer's experience and intuition. The next design step consists of modeling and simulation of the circuit to predict and analyze the performances of the circuit. The designers may need to iterate these steps unless a suitable topology along with the dimensions of all the transistors are obtained for which the specifications are satisfied at the pre-layout stage. Subsequently, the geometrical layout of the selected circuit topology is drawn and the necessary layout parasitics are extracted. The circuit is further simulated taking into considerations the effects of the parasitics on the circuit performances, and the performances of the circuit are checked against the desired specifications. This is also an iterative task and the layout of the circuit is refined. If results are satisfactory, the circuit becomes ready for fabrication. After fabrication, the circuit is tested before it is launched into the market as a packaged product.

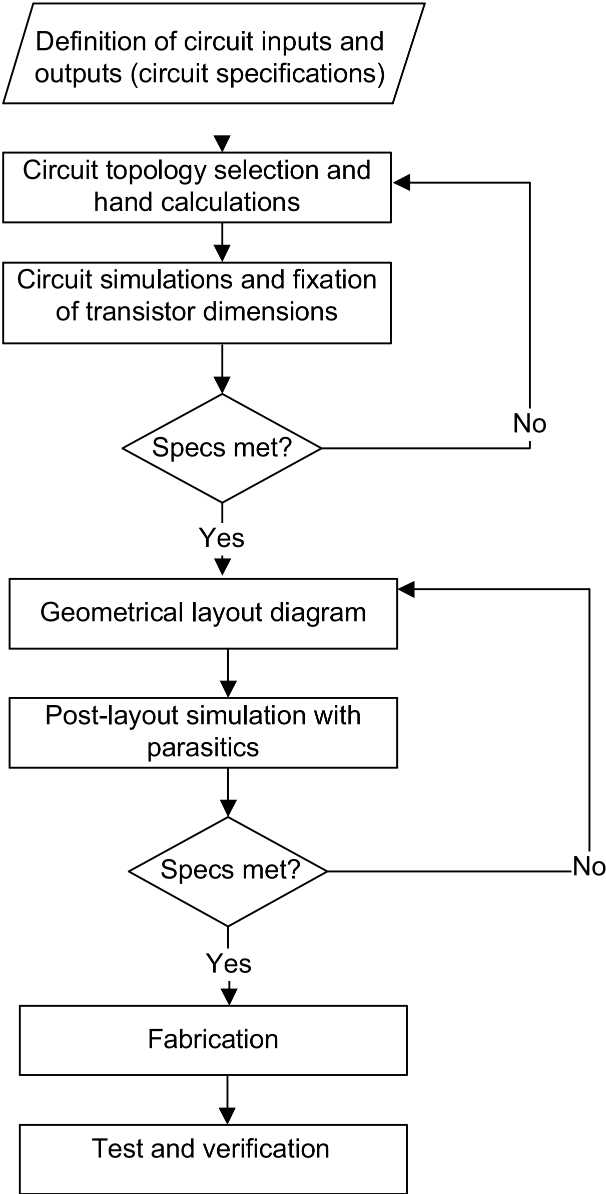


FIGURE 1.16
Bottom-up design technique for analog IC.

1.5.1.1 Advantages and Limitations

While designing a large circuit/system following this technique, the design process starts with the design of individual sub-circuits/blocks, which are then combined to form the complete circuit/system. The significant advantage of this approach is that the device-level details of nano-scale MOS transistors which critically affect the circuit performances are taken care of at the very beginning of the design process through innovative topology design and sizing of the sub-circuits/blocks. This approach is suitable for small designs, however, for large designs, several limitations of this approach have been observed, which are discussed below.

1. For complex system design, the greatest impact on the functionality, performance and cost of the system is observed at the architecture-level. The architecture exploration and architecture optimization is considered to be an important design task. However, there is no provision of this design step in the bottom-up approach. Therefore, such possibilities of improvements are often missed out in this approach.
2. After the individual blocks are designed and combined to form the system, the verification task via a simulation process takes a huge amount of time. Therefore, comprehensive analysis and verification of design in a manageable time frame becomes difficult. The errors arising out of the verification task are expensive to fix because this involves redesign of the block.
3. Communication between the individual block designers is critical for large SoC designs; however, this is performed in an informal way in the traditional design technique. Therefore, any gap arising at the interface of the design of the individual blocks often leads to costly silicon respin.
4. Several important and expensive steps in the traditional design technique are performed serially which increases the design cycle.
5. Because of the lack of algorithmic approach in this technique, it is not suitable for design automation.

1.5.2 Top-Down Design Technique

The two important principles of the top-down design technique are [8, 66, 25, 107] (i) to carry out the design task at several levels of abstraction and (ii) to use a hierarchical design strategy at each level of abstraction. These are discussed below.

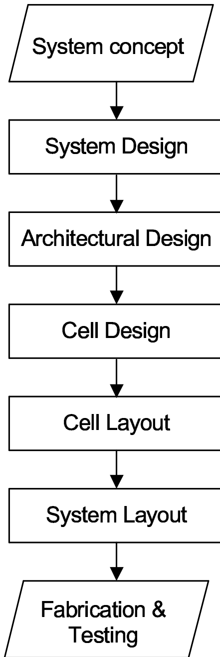


FIGURE 1.17
Abstraction levels in top-down analog IC design technique.

1.5.2.1 Abstraction Levels

The levels of abstraction are shown in Fig. 1.17. In the system concept stage the specifications of a design are collected from the users and the overall product concept is developed. A first-hand idea of the working of the overall system may be demonstrated through specialized tools. Several design and management issues such as selection of process technology, product cost, project management strategies, time-to-market factor etc., are formulated at this stage. The various levels of abstractions are as follows.

1. The first level of the actual design process is the system design level. The overall architecture of the system is designed and partitioned into software and hardware components. The hardware components are specified in suitable hardware description languages (HDLs).
2. The second level is the architectural design level. It consists of high-level decomposition of the hardware part of the system into an architecture consisting of several functional blocks and formulation of specifications for the individual functional blocks. The functional blocks are described with HDLs.
3. The third level is the cell design level. This consists of detailed

transistor level implementation of the individual functional blocks in the chosen process technology such that the specifications of the individual blocks are satisfied.

4. The cell layout level consists of converting the transistor-level implementations of the different blocks into geometrical level representations, i.e., layouts. The parasitics inherent in the cell layouts are extracted and then the circuits are simulated to investigate the effects of layout parasitics on circuit performances and verified against the desired specifications.
5. In the system layout level, the cell layouts are placed and routed to form the system layout. Several issues such as crosstalk, power grid routing, proper shielding etc., are taken care of in this stage. In addition, suitable measures are taken to make the circuit testable.
6. Finally, the entire circuit of the complete system is fabricated and tested. It may be noted that the verification by simulation process is carried out at the individual levels of abstraction and the design task at each level of abstraction often happens to be an iterative procedure.

1.5.2.2 Hierarchical Design Strategy

The design process at each level of abstraction is carried out in a hierarchical way, where the design task at the $(i-1)^{th}$ level of abstraction formulates detailed specifications for the level i, which in turn formulates the specifications for the level $(i+1)$. The design task at the ith level of abstraction consists of three steps: topology selection, specification translation and design verification, as shown in Fig. 1.18. The task of topology selection involves selection of an appropriate topology out of a set of alternatives, which best meet the desired specifications [66]. The specification translation task consists of mapping the specifications at a particular level of abstraction onto the specifications of the various sub-blocks forming the selected topology at that level of abstraction [66]. It may be noted that at the cell design abstraction, the task of specification translation is referred to as circuit sizing, where the transistor dimensions and biases are determined.

1.5.2.3 Advantages and Limitations

The top-down design technique overcomes the limitations discussed earlier for the bottom-up technique. An architectural design step has been added in the top-down technique. The task of verification is carried out at all levels of abstraction such that the verification of the overall system becomes simplified because the main focus may now be given on the interfaces. Therefore, the significant advantage of top-down design technique is the high probability of fault detection at higher levels of abstraction, and therefore, has a high chance of first-time-success, while obtaining a better overall system design through

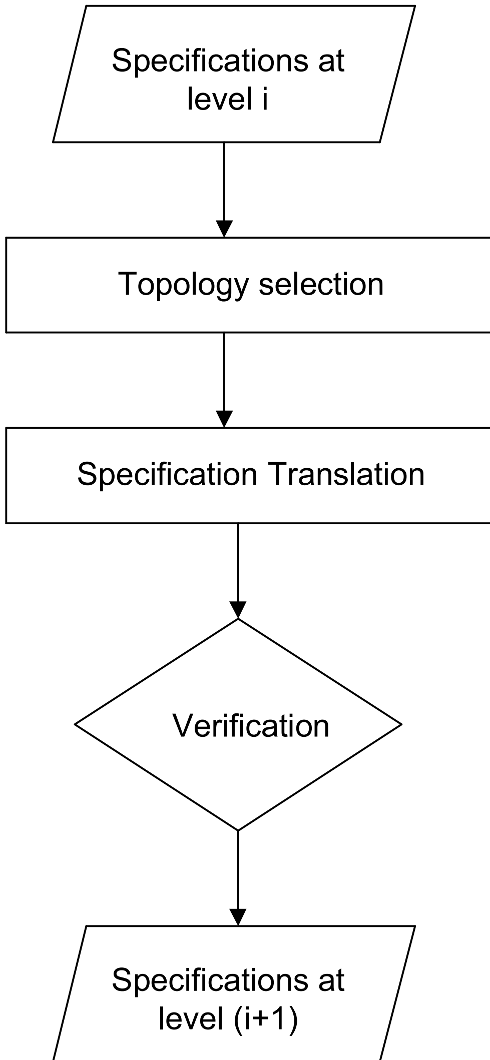


FIGURE 1.18
Hierarchical design strategy in top-down design technique.

architecture exploration. The communication between various designers is better handled in the top-down design technique.

The top-down design technique, although very successful for digital circuit design, suffers from some important limitations when applied to analog circuit designs.

1. The architectural design step is often carried out in somewhat *ad hoc* fashion. The finer circuit-level details are many times ignored at the architectural design level. This makes the resultant architecture often infeasible at the cell design level of abstraction.
2. For analog designs, the design task often cannot be decomposed cleanly so that the boundaries between the various levels of abstractions are not distinct, rather the various sub-tasks are tightly coupled. This is becoming significant with the down scaling of MOS transistors in sub-90nm process technology. Therefore, a serious mismatch occurs between the representations of the circuit used in the architectural design stage and that used in the cell design stage.

The idea of copying down the digital design technique for analog designs has been discarded by the analog designers, who still prefer to rely upon the well tested bottom-up design technique, in spite of its inappropriateness for design automation.

1.6 Knowledge-Based CAD Technique for Analog ICs

The top-down design technique discussed earlier is a structured technique which yields reasonably good results for standard digital circuits. In addition, it is suitable for automation. However, this technique in its present form does not allow the analog designers to inject substantial creativity as an inherent part of the design process. What is missing in this technique is a procedure for extraction and management of design insight/knowledge and the use of this knowledge database in the design process. This is especially significant for analog circuits designed with nano-scale CMOS technology. Apart from this, there must be an algorithmic approach for the process of analysis and design. In nutshell, the major limitation of the conventional design techniques for analog ICs is the lack of algorithmic integration of the designers' knowledge and insight into a computer-aided design framework. This leads to the essential requirement of knowledge-based CAD techniques for analog ICs designed with sub-90nm process technology. The motivation for knowledge extraction and management is discussed in the next sub-section, followed by problem formulation and an outline of the technique.

1.6.1 Motivation for Knowledge Extraction and Management

In sub-90nm CMOS process technology, the short channel effects of MOS transistors as well as other secondary effects play critical role in determining the performances of the circuit and thus makes the task of the designers challenging. Traditionally the designers completely and even in some cases blindly rely upon the SPICE simulation tool which internally takes care of all such effects during the circuit simulation procedure. The SPICE simulation tool uses the BSIM (*B*erkeley *S*hort Channel *I*GFET *M*odel) compact model. The state-of-the art BSIM compact model for sub-90nm design is BSIM4. The situation is then handled through parametric SPICE simulation and sizing of transistor dimensions or even topology utilizing designers' knowledge and intuition. The situation is thus obviously better handled by experienced designers.

Accurate characterizations of the short-channel effects of MOS transistors, statistical process variations and reliability constraints are the two major tasks for successful computer-aided design of nano-scale analog ICs. The characterization data must be well documented and efficiently managed so that the extracted knowledge may be embedded within the design framework.

The task of knowledge extraction and management may be accomplished through development of good high-level mathematical models. A good model must incorporate all relevant information and experience of the circuit to be designed. Good models are considered to be among the cornerstones of an efficient knowledge-based CAD technique. Models of different types are required. These are behavioral, performance and feasibility. In the present text, the models will be limited to the pre-layout simulation stage only, or in other words, the models are high-level models.

1.6.2 Problem Formulations

First some terminologies are defined, followed by formulation of the general design problem.

1. The design variables refer to the specification parameters of the component blocks used in the topology, e.g., gain, bandwidth, etc., of an amplifier at the architecture-level design abstraction, or transistor dimensions at the transistor cell-level design abstraction. The design variable vector is denoted by $\bar{\alpha}$
2. There are two kinds of design objectives: functional objectives ρ_f and performance objectives ρ_p. The functional objectives need to be met by the design in order to be functional. The performance objectives such as power consumption, area etc. need to be minimized.
3. The design variables are constrained within a boundary defined by

the upper limits and lower limits. These form a feasible design space $\mathcal{D}$ within which the optimal design solution must lie.

In a knowledge-based CAD technique, the design problem is translated into a function minimization problem which is solved through numerical optimization algorithms in an iterative manner. This is mathematically written as follows

$$\begin{aligned} \text{Minimize} \quad & \bar{\rho_p}(\bar{\alpha}) \\ \text{subject to} \quad & \bar{\rho_f}(\bar{\alpha}) \geq 0 \;\text{ and }\; \bar{\alpha} \in \mathcal{D} \\ \text{where} \quad & \bar{\alpha} \in \Re^{n_\alpha} \end{aligned} \tag{1.17}$$

1.6.3 Outline of the Procedure

An outline of the procedure is illustrated in Fig. 1.19. The procedure starts with initialization of the design variables. The entire procedure is an iterative process, where the design variables are updated at each iteration, until an equilibrium point is reached. The degree of compliance of the design performances with the optimization goals at each iteration is quantified through a cost function. The two important modules for this procedure are a high-level model and an optimization engine. The implementation of the design methodology is based upon the flow of information between these two modules. The high-level model provides a way to evaluate the optimality of the design with regard to the intended requirements. On the other hand, the optimization engine deals with the cost function and explores the available design space to minimize such a function. The cost function being minimized during the optimization process contains two types of terms: terms related to functional objectives and terms related to performance objectives. The procedure stops when both the desired objectives are optimally satisfied.

There are two different approaches for evaluation of functional and performance objectives: namely the simulation-based approach and the analytical model-based approach. The basic principles of these two approaches are discussed below.

1.6.3.1 Numerical Simulation-Based Evaluation

In this approach, the process of the evaluation of functional and performance objectives inside the optimization loop of Fig. 1.19 is implemented through a numerical simulation technique, e.g., SPICE simulation. The user needs to provide the desired specifications of the component blocks and a simulation plan for each of these specifications. Such plan includes the test set-up (input sources, loads, feedbacks, etc.), the input signals to be applied, simulation commands and the required data processing of the simulation results to obtain the desired objectives. A serious problem with this approach is the required simulation time of a single evaluation procedure. Since global optimization techniques typically take several thousands of iterations, the evaluation time

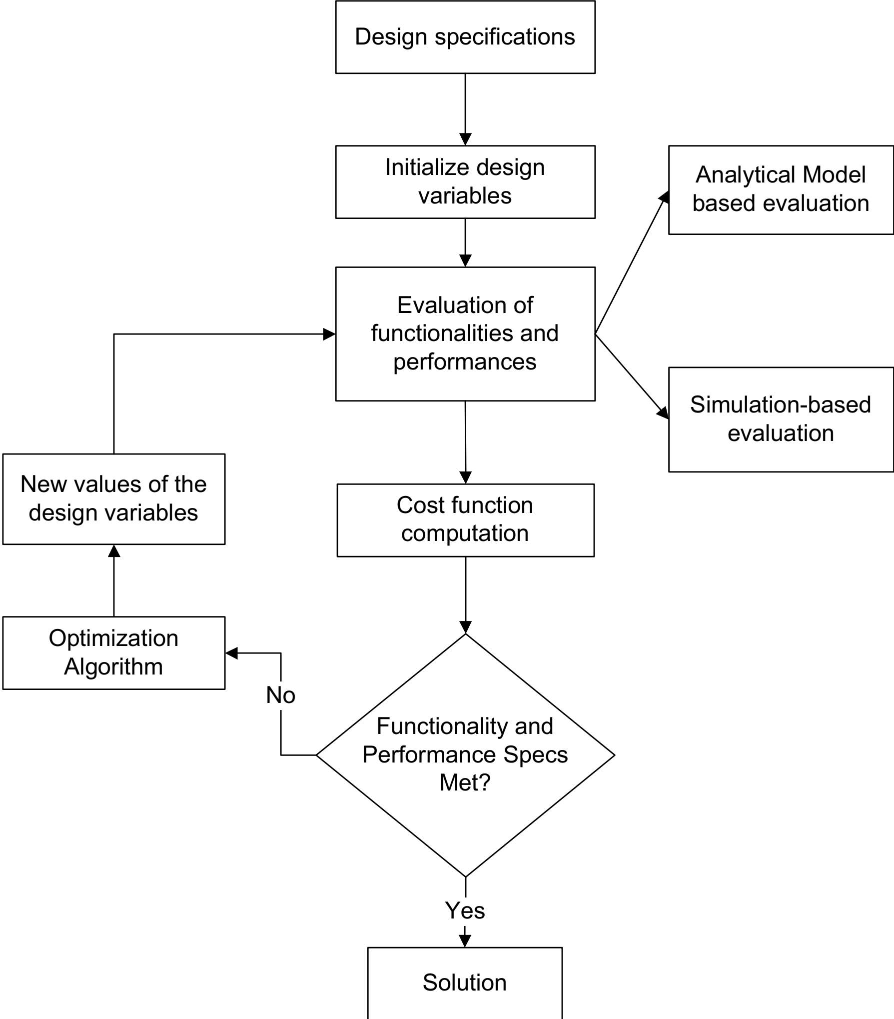


FIGURE 1.19
Knowledge-based CAD procedure.

of one complete run will be unacceptable, if one evaluation takes more than a few seconds. On the other hand, an important advantage of this approach is that the process is accurate and flexible. The user can program a new design problem in a minimum amount of time. A simulation-based technique for architecture-level design and optimization of analog RF receiver front-ends is described in [35]. The design methodology works to evaluate the performance of an RF receiver topology and automatically translates high-level system specifications into a set of specifications for each building block in the topology such that the overall power and/or area consumption of the receiver is minimized. Similar works for a $\Sigma\Delta$ modulator design are reported in [59, 6].

1.6.3.2 Analytical Model-Based Evaluation

This approach uses analytical models for evaluations. The models consist of a set of analytical equations which directly relate the functional and performance specification objectives of a component block to the design variables of the component block. These equations can be derived using automated or handcrafted symbolic analysis [66]. An important advantage of this approach is that the process of evaluation of a set of equations is much faster compared to the simulation process. Therefore, the execution time of a complete optimization process is generally small. The disadvantage is the much larger setup time. The user needs to derive all the design equations, which is a difficult, time consuming and error-prone task. The accuracy of the performance equations compared to circuit-level simulation results is often not good. In addition, several performance characteristics cannot be suitably captured by analytical equations. Furthermore, the design equations are often very specific to the topology of the component block and cannot be reused for other topologies. Symbolic equation-based high-level design procedures have been reported in [48, 49]. This technique has also been applied to the high-level design of $\Sigma\Delta$ modulator in the SD-OPT tool [128].

1.6.4 Salient Features

Some important salient features of this technique are

1. This design technique is to be performed hierarchically at all levels of design abstractions, as summarized in Fig. 1.17. At each level, the design objectives and the design variables change, however.
2. The major limitation of the conventional top-down design approach, i.e. the inability of the designers to inject designers' knowledge into a CAD framework is overcome through the construction of high-level models and use of the same.
3. The nano-scale effects of the MOS transistors can be accurately incorporated within the high-level models so that it is possible to consider such issues carefully.

4. The design task is formulated as an optimization problem which requires a minimum amount of intervention of the designers during execution. Once the high-level models are efficiently constructed these can be embedded within the optimization procedure.

Construction of suitable high-level models considering all sorts of nano-scale effects, and a robust optimization procedure are the two mandatory tasks for the success of the present approach for efficient design of nano-scale analog ICs.

1.7 Summary and Conclusion

This chapter presents a comprehensive overview of the concepts of technology scaling and the effects of technology scaling on CMOS analog circuit design. The two theoretical models of scaling, namely the constant field scaling and the constant voltage scaling have been discussed. The variations of the various important device parameters with technology node have been illustrated graphically based on ITRS specifications and PTM parameters. It has been found that the technology scaling of the various device parameters follows in practice, a combination of the field and the voltage scaling. The critical challenges of nano-scale analog IC design, namely the deterioration of output resistance and intrinsic gain, gate oxide leakage current, drain current mismatch, noise performance etc., are discussed. The power consumption for analog circuits designed with reduced supply voltage is also discussed. The requirements of CAD techniques to cope with the challenges due to increased design complexity and reduced time-to-market factor has also been discussed. The advantages and limitations associated with the conventional design techniques have been mentioned. Finally, the knowledge-based CAD technique has been introduced. This is believed to be the state-of-the art technique required for nano-scale analog IC design. This technique will be discussed in further detail in the subsequent chapters.

2

High-Level Modeling and Design Techniques

2.1 Introduction

The success of the knowledge-based computer-aided design technique discussed in Chapter 1 is dependent on the efficient implementation of two different modules: (i) the high-level models and (ii) the optimization procedure. The term “high-level” as used in the present text means the pre-layout level. The high-level models serve three distinct purposes [160]. First, these are used as alternatives to the conventional approach of SPICE-based design simulation for verification purpose. The models which are specially constructed for this purpose are referred to as behavioral models [156]. Second, high-level models are used for evaluation of functional and performance objectives of a design during the design optimization and exploration procedure. The models which are specially constructed for this purpose are referred to as performance models [101]. The performance models are thus used as alternatives to the SPICE-based simulation procedure for design evaluation. Third, high-level models are often used to judge the feasibility of any design solution during the design optimization procedure. The models which are specially constructed for this purpose are referred to as feasibility models [178]. The optimization procedure is used for design space exploration for the selection of an optimal design. The procedure needs to be accurate and reliable without consuming too much CPU time. The computational complexity of the procedure needs to be good.

For sub-90nm analog IC design, a paradigm shift in the design methodology is required. The conventional CAD technique needs to be complemented with the technology CAD (TCAD) technique for incorporating the enhanced physical effects of MOS transistors, statistical process variabilities and time dependent reliabilities. The objective of this chapter is to present a comprehensive discussion of the fundamental principles of the construction procedure for high-level models and optimization algorithms. The implementation issues are also discussed. The extension of the knowledge-based CAD technique to nano-scale analog IC design through the integration of TCAD technique is also discussed in detail.

2.2 High-Level Model

A high-level model of a component block is defined as a set of mathematical models which expresses the input-output behavior, performance parameters and feasibility of the design as functions of the various design variables of the component block [64]. The high-level models are of three different types: (i) behavioral model, (ii) performance model and (iii) feasibility model. These are discussed in the following sub-sections.

2.2.1 Behavioral Models

Let us consider a circuit, C shown in Fig. 2.1 transforming an input signal, U into an output signal, Y. Suppose the circuit is governed by a vector of design variables $\bar{\alpha} \in \Re^{n_\alpha}$ that influences its behavior. Then

$$Y = \mathcal{B}(U, \bar{\alpha}) \tag{2.1}$$

Here $\mathcal{B}$ is called the behavioral model of the circuit C. The mathematical modeling of the circuit's input-output behavior is called behavioral modeling [123, 152].

The behavioral models are used for design verification. Therefore, these are instance oriented, i.e., these are constructed for a certain circuit only and the device level details of the circuit are fully known during the construction of these models. For the first hand verification of a large circuit consisting of several thousands of devices, the full SPICE-based design verification procedure simply becomes prohibitive in terms of CPU time. For such cases, accurate

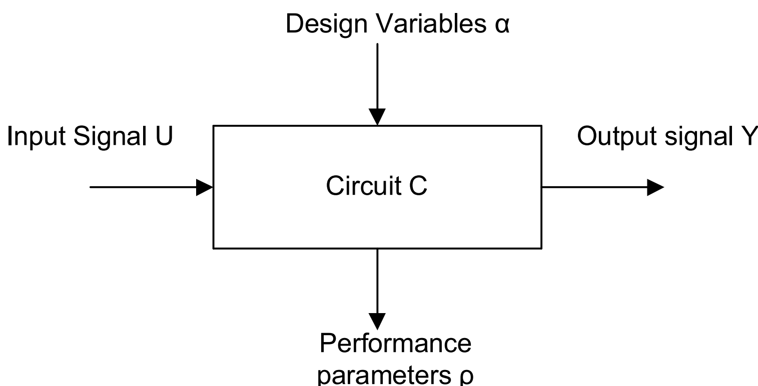


FIGURE 2.1
A circuit C consisting of design variables α, with input signal U and output signal Y.

instance oriented models are used as replacements of the lengthy numerical simulation procedure. The instance oriented models must accurately replicate the input-output functionality of the circuit. Such model-based verification procedure offers the designers much more flexibility and convenience through the design procedure, especially if these are used in a hierarchical fashion at all levels of abstraction [9].

2.2.2 Performance Models

The various performance parameters $\bar{\rho} \in \Re^{n_\rho}$ of the circuit are dependent upon the output signal Y and the design variables $\bar{\alpha}$. Mathematically this is represented through

$$\bar{\rho} = \mathcal{P}(Y, \bar{\alpha}) \tag{2.2}$$

where $\bar{\rho}$ is a vector of all performance parameters, e.g., bandwidth, slew rate for a component block, $\bar{\alpha}$ is a vector of all design variables such as transistor sizes, values of the various resistors and capacitors. Eliminating Y and considering the fact that U is constant for changes of the design variables, it can be written

$$\bar{\rho} = \mathcal{P}(\bar{\alpha}) \tag{2.3}$$

The mathematical modeling of this relationship is called performance modeling [141].

It may be noted that the behavioral models are necessary but not sufficient for use in an automated design optimization and exploration procedure. Parameterized performance models are required for this purpose. These models must predict the performances of the component block as a function of its design variables. Since the device level details of the circuit are often not known during the construction of the parameterized performance models, some loss of model fidelity is often accepted.

2.2.3 Feasibility Models

A specification translation/circuit sizing procedure often yields overambitious values of the various design variables for the component blocks of a system. This happens if the desired functional objectives as well as the implementation related limitations of the underlying circuit components are not taken into account during the sizing process. Well characterized feasibility models are therefore needed, in order to limit the circuit sizing process to determine feasible values of the various design variables of the component blocks while satisfying the desired functional objectives and minimizing the performance objectives [76].

The task of circuit designers is to determine the design variable set $\bar{\alpha}$ either through automated procedure or manually. For all parameters, the designers based on their experience, specify a feasible parameter region

$$\mathcal{D} = \{\bar{\alpha} | \bar{c}(\bar{\alpha})| \geq 0\} \tag{2.4}$$

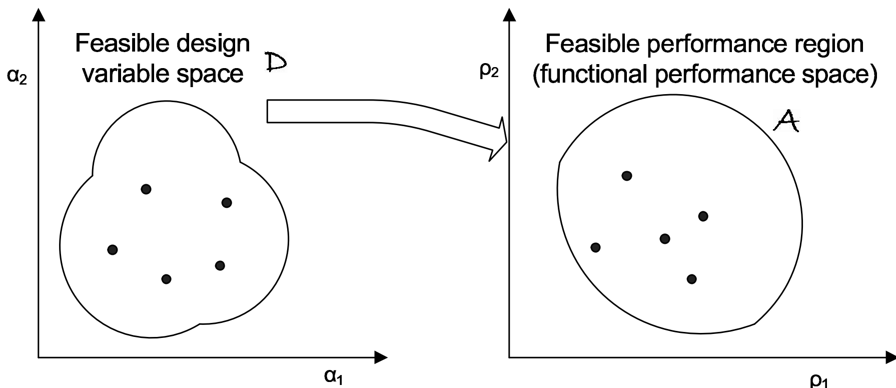


FIGURE 2.2
Mapping of feasible design variables to feasible performance parameters.

Here $\bar{c}(\bar{\alpha})$ is a single nonlinear vector inequality imposed on the design variables. Thus the feasible parameter range is defined as the set of design variable values that satisfy the imposed constraints. The feasible performance region is defined as the set of performance parameters $\bar{\rho}_f$ which are constrained within a boundary such that the overall circuit is functional. The feasible performance parameters result from a region $\mathcal{D}$ of feasible parameter values. Mathematically this is defined as

$$\mathcal{A} = \{\bar{\rho}_f | \bar{\rho} = \mathcal{P}(\bar{\alpha}) \geq 0 \wedge \bar{\alpha} \in \mathcal{D}\} \tag{2.5}$$

The mapping of the feasible design variables to feasible performance parameters is schematically illustrated in Fig.2.2.

2.2.4 Characteristics of Good High-Level Models

The key to success of a CAD technique is the construction of a good high-level model. The model needs to be good in four senses. First, it must accurately represents all practical circuit behavior. For example, a high-level model for a voltage amplifier must capture all of the relevant behavior that characterizes an amplifier's transistor-level implementations. Use of an inaccurate model leads to error in the verification procedure. Second, the model needs to be computationally simple. Simple models are, however, often found to be inaccurate. Thus managing the accuracy and simplicity of the high-level models is the greatest challenge in model generation techniques. Third, the construction time of a high-level model should be low. Fourth, the model needs to be scalable with respect to transistor sizes, biases and process technology. Systematic generation of good high-level models is considered as one of the largest problems in an analog CAD technique [64].

2.3 Behavioral Model Generation Technique

This section presents three important techniques for generation of behavioral models. These are the manual technique, the model order reduction technique and the symbolic analysis technique.

2.3.1 Manual Abstraction

The most prevalent approach toward creating a behavioral model is the manual abstraction. The complete behavior of a component block is split up into two parts—fundamental/ideal behavior and non-idealities. For any analog component block, the ideal behavior is generally a simple mathematical operation such as scaling, integration, multiplication, etc.,. The non-idealities are then modeled in terms of the effects they introduce, e.g., distortion, rather than in terms of the causes, e.g., transistor sizes or particular topologies. Simulation frameworks like Simulink, AMS Designer/Verilog-AMS, etc.,., are suitable for implementing the models.

The technique is illustrated with an example for modeling the transfer function and noise properties of a switch-capacitor (SC) integrator, as shown in Fig. 2.3(a). The z-domain transfer function of the integrator is given by

$$H(z) = \frac{C_s}{C_f} \frac{z^{-1}}{1 - z^{-1}} \tag{2.6}$$

$C_s/C_f = b$ represents the coefficient of the integrator. The most important noise sources affecting the operation of an SC integrator are the thermal noise due to the sampling switches and the intrinsic noise of the operational transconductance amplifier (OTA). The thermal noise associated with the switching is modeled as follows

$$y(t) = b. [x(t) + n_s(t)] \tag{2.7}$$

where

$$n_s(t) = \sqrt{\frac{kT}{C_s}} RN(t) \tag{2.8}$$

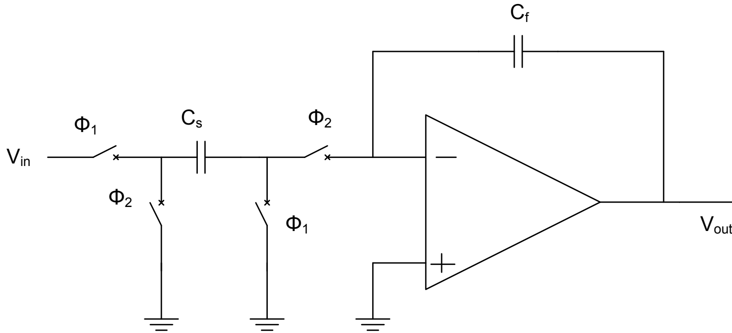
where C_s is the input sampling capacitor, $RN(t)$ is a Gaussian random number with zero mean and unity standard deviation. The behavioral modeling of the switch noise is shown in Fig. 2.3(b)

The input referred thermal noise of the OTA is modeled as

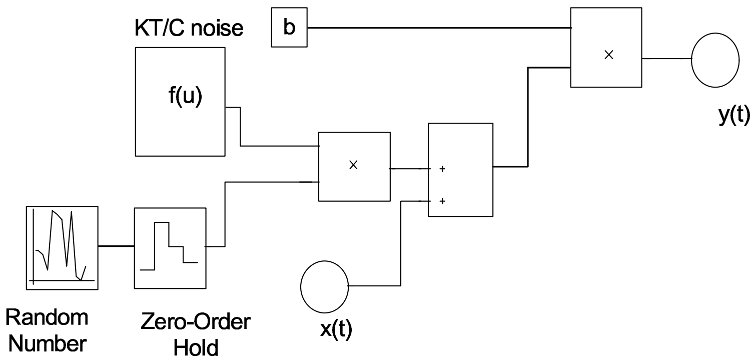
$$z(t) = b.n_{OTA}(t) \tag{2.9}$$

where

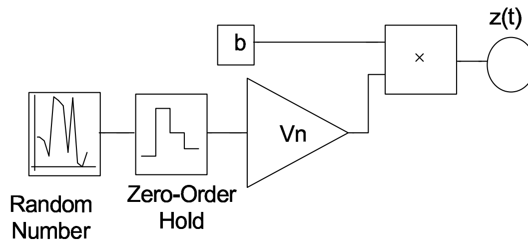
$$n_{OTA}(t) = V_n.RN(t) \tag{2.10}$$



(a) Single-ended SC integrator.



(b) kT/C noise model in Simulink.



(c) OTA noise model in Simulink.

FIGURE 2.3
Manual approach for construction of high-level model using Simulink®.

where b is the integrator gain, V_n represents the *rms* noise voltage of the OTA referred to the integrator input. The behavioral modeling of the switch noise is shown in Fig. 2.3(c)

Several works are available in literature which adopt this technique for behavioral model generation. Behavioral modeling of switched-capacitor $\Sigma\Delta$ modulators following this technique using the Simulink platform is presented in [6, 120, 201]. Behavioral modeling and simulation of pipelined ADC and PLL following this technique are discussed in [159] and [123] respectively.

While the manual approach is the only feasible approach even today for many complex blocks, it has a number of limitations. The numerical simulation technique usually does not provide abstracted parameters of interest such as poles, residues, modulation factors etc. Extracting these manually by processing the simulation results is inconvenient, computationally expensive and error prone. The manual approach often misses some significant nonidealities and interactions which are critical in the design verification process. The situation is getting worse with the down scaling of MOS technology to the sub-90nm regime. Adequate incorporation of nonidealities in the high-level models using the manual approach, if not impossible, is typically complex and laborious. As a result, the potential improvement in the time-to-market factor with the use of a model-based verification procedure can be substantially negated by the time and effort required to first construct the models.

2.3.2 Model Order Reduction Technique

The model order reduction approach is an algorithmic approach for transformation of a large set of mathematical equations to a much smaller one. The technique is general in the sense that as long as the equations of the original circuit are known (may be through SPICE simulations), the internal structural details and operating principles are not required. These reduced order models simulate much more efficiently, while accurately approximating the response of a real circuit. This approach is discussed here for linear time invariant (LTI) and linear time varying (LTV) systems. A fairly complete survey of model order reduction techniques is provided by Roychowdhury in [157, 156], which forms the basis of the following material.

2.3.2.1 MOR for LTI Systems

Let us consider the basic structure of an LTI block as shown in Fig.2.4, described by a set of differential equations as follows

$$\begin{aligned} E\dot{x} &= Ax(t) + Bu(t) \\ y(t) &= C^T x(t) + Du(t) \end{aligned} \tag{2.11}$$

In (2.11), $u(t)$ represents the input to the LTI block, $y(t)$ represents the output and $x(t)$ represents the internal state variables of the block. A, B, C, D and E are constant matrices. This type of state space equation can easily be constructed either from a SPICE netlist or from some AHDL descriptions.

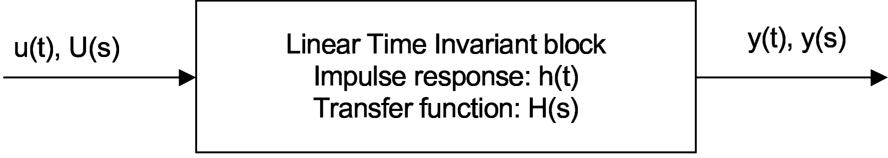


FIGURE 2.4
LTI block.

The dimension n of $x(t)$ is assumed to be very large. The basic idea of model order reduction is to construct an equivalent state space representation of the LTI block, given by

$$\begin{aligned} \hat{E}\dot{\hat{x}} &= \hat{A}\hat{x}(t) + \hat{B}u(t) \\ \hat{y}(t) &= \hat{C}^T\hat{x}(t) + \hat{D}u(t) \end{aligned} \tag{2.12}$$

For meaningful reduction, the size q of the reduced system needs to be much smaller than n, i.e., the size of the original, however, keeping the original behavior identical. In the asymptotic waveform evaluation (AWE) technique, the explicit moments of the transfer function of the original system as well as the reduced system are kept identical. These moments are defined as follows [144]

$$m_1 = \left.\frac{dH(s)}{ds}\right|_{s=s_0}, \quad m_2 = \left.\frac{d^2H(s)}{ds^2}\right|_{s=s_0} \tag{2.13}$$

where s_0 is the frequency point of interest. The explicit moment matching method is useful for interconnect network problem for timing analysis. However, this is observed to become numerically instable as the size of the reduced model becomes greater than 10. To alleviate this problem, the Krylov subspace MOR technique [58] has been proposed, which performs implicit moment matching rather than the computation of the moments of the full system explicitly at any point. In the Krylov subspace MOR technique, two projection matrices $V \in \Re^{n\times q}$ and $W^T \in \Re^{q\times n}$ are constructed, such that the reduced system is given by

$$\begin{aligned} \underbrace{W^TE}_{\hat{E}}\dot{x} &= \underbrace{W^TAV}_{\hat{A}}x(t) + \underbrace{W^TB}_{\hat{B}}u(t) \\ y(t) &= \underbrace{C^TV}_{\hat{C}}x(t) + Du(t) \end{aligned} \tag{2.14}$$

Krylov subspaces are computed using either the Lanczos process or the Arnoldi process [161]. If the Lanczos process is used to compute the Krylov subspaces, then $W^TV \approx I$, which means that the two projection bases are bi-orthogonal. On the other hand, if the Arnoldi process is used then $W = V$ and $W^TV = I$. Krylov subspaces MOR techniques are shown to capture well the dominant poles and residues of the system.

2.3.2.2 MOR for LTV Systems

The MOR technique for LTI systems cannot be directly applied for linear time varying (LTV) systems. The main difference between LTI and LTV systems is that if the input to a LTV system is time shifted it does not necessarily imply that the output will be time shifted by same unit. However, the linearity property remains the same for both the systems. A class of practical nonlinear circuits such as RF mixing, switched capacitor and sampling circuits can be represented by the LTV model. The behavior of a LTV system is described by the following time-varying differential equations

$$\begin{aligned} E(t)\dot{x} &= A(t)x(t) + B(t)u(t) \\ y(t) &= C(t)^T x(t) + D(t)u(t) \end{aligned} \tag{2.15}$$

Time variation in the system is captured by the dependence of A, B, C, D and E on t. For several practical cases, this time variation is periodic, e.g., for mixer circuits the local oscillator input is often a sine or a square wave, switched systems are driven by clock signals. Because of the time varying characteristics of the impulse response and transfer function, LTI MOR techniques cannot be directly used to reduce LTV systems. The LTV system (2.15) first needs to be expressed as an LTI system (2.11) with extra artificial inputs which capture the time variation. The corresponding LTI system is then reduced by employing any LTI-MOR technique. The reduced LTI system is transformed back to LTV form. The use of different LTI MOR techniques, such as the AWE technique, Krylov subspace technique within this framework for reduction of LTV system has been demonstrated in literature [155].

2.3.2.3 MOR for Nonlinear Systems

It may be noted that for completely general nonlinear systems, currently there does not exist any technique that is capable, at least in principle of reducing a large system that conforms to any reasonable fidelity metric. This is because of the fact that nonlinear systems are richly varied with extremely complex dynamical behavior. For practical circuits such as linear amplifiers and mixers, the nonlinear performances such as distortion and intermodulation are handled by limiting these to a very small fraction of the output of the linearized system. Thus a nonlinear system is converted to a weakly nonlinear system. A nonlinear system is described by a set of nonlinear differential algebraic equations as follows

$$\begin{aligned} \dot{q}(x(t)) &= f(x(t)) + bu(t) \\ y(t) &= c^T x(t) \end{aligned} \tag{2.16}$$

In (2.16), $f(.)$ and $q(.)$ are nonlinear vector functions. Assuming weakly nonlinear functions, $f(x)$ and $q(x)$ are approximated by low-order polynomials.

This is obtained by retaining the first few terms of the Taylor series expansion of the nonlinear functions, as follows

$$f(x) = f(x_0) + A_1(x - x_0) + A_2(x - x_0)^2 + ... \tag{2.17}$$

Here x_0 is typically the DC solution. With this approximation and assuming $q(x) = x$ for simplicity, (2.16) can be written as

$$\begin{aligned} \dot{x}(t) &= f(x_0) + A_1(x - x_0) + A_2(x - x_0)^2 + ... + bu(t) \\ y(t) &= c^T x(t) \end{aligned} \tag{2.18}$$

The advantage of this approach is that several existing techniques such as the Volterra series theory [167] and weakly nonlinear perturbation techniques can be used to compute the response of the system. Weakly nonlinear systems are well represented by second order polynomials. The polynomials, in general are however, known to be extremely poor global approximators because of their oscillatory nature. Another useful method to approximate nonlinear function is to use a piecewise linear (PWL) approximation. The idea is to split the state space into a number of disjoint regions, and within each region, a linear function is used to approximately match the nonlinear function. This may even be extended to a piecewise polynomial (PWP) method which combines weakly nonlinear MOR techniques with the piecewise idea, by approximating the nonlinear function in each piecewise region by a polynomial rather than a purely linear one [50]. PWP technique is found to be extremely useful for generation of behavioral models for practical circuits such as operational amplifiers in which strong and weak nonlinearities both play important functional roles.

It needs to be appreciated that the task of generation of a behavioral model for practical analog circuits in an automated way is very difficult. Formulation of a generalized technique for construction of behavioral models of analog circuits is extremely complicated, if not impossible. This is because of the very diverse behavior of the analog circuits. The PWL method is specialized. The PWP method, although heuristic, is broadly applicable. Despite the progress made so far, still more research in the area of automatic generation of reduced order models is certainly needed.

2.3.3 Symbolic Analysis Technique

2.3.3.1 Basic Concepts

Symbolic analysis at the circuit level is defined as the formal technique for determining the behavior or characteristic of a circuit with the independent variable (time or frequency), the dependent variables (voltages and currents), and (some or all of) the circuit components denoted by symbols. An excellent tutorial introduction to symbolic analysis technique has been provided in [61, 63, 151]. The technique is complementary to numerical analysis and qualitative

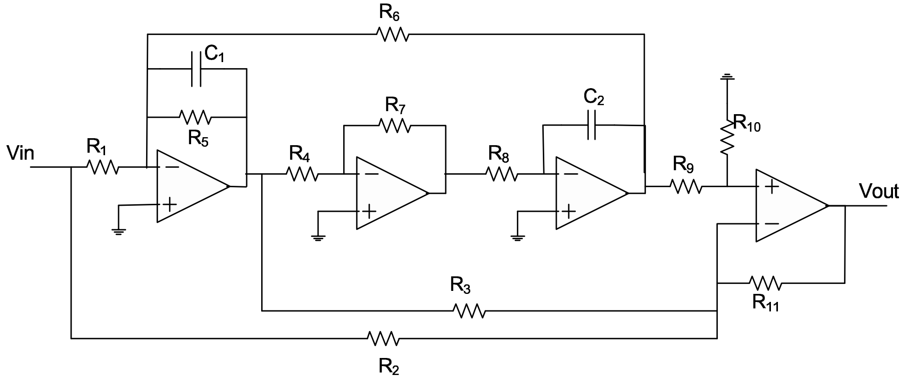


FIGURE 2.5
Active RC filter.

analysis. It has been possible to automate the symbolic analysis technique to a large extent and several computer programs have been developed which receive the circuit description as input, automatically carry out the symbolic analysis and thus generate the symbolic expression for the desired circuit characteristic. Analog Insydes is a well-known commercial tool for symbolic analysis of analog circuits.

For lumped LTI circuits, the symbolic analysis technique yields symbolic network functions in the complex frequency variable σ (s in the continuous time domain and z in the discrete time domain). Typically this is as follows.

$$H\left(\sigma\right) = \frac{N(\sigma, p_1, p_2, ..., p_m)}{D(\sigma, p_1, p_2, ..., p_m)} = \frac{\sum_i \sigma^i . a_i(p_1, p_2, ..., p_m)}{\sum_i \sigma^i . b_i(p_1, p_2, ..., p_m)} \tag{2.19}$$

where in the partially expanded form on the right, the coefficient $a_i(...)$ and $b_i(...)$ of each power of σ for both the numerator and denominator polynomial are symbolic polynomial functions in the circuit elements p_j. These polynomials can be in a nested format or expanded into the sum-of-product form. This is illustrated with the example of an active RC filter, the schematic of which is shown in Fig. 2.5. The symbolic transfer function of this circuit is [63]

$$\begin{aligned}\{-G_4G_8(G_1G_2G_9 + G_1G_3G_9 + G_1G_9G_{11} + G_2G_6G_9 + G_2G_6G_{10}) + \\ sG_7C_2(G_1G_3G_9 + G_1G_3G_{10} - G_2G_5G_9 - G_2G_5G_{10}) - s^2G_2G_7C_1C_2(G_9 + G_{10}) \\ \}/\{G_{11}(G_9 + G_{10})(G_4G_6G_8 + sG_5G_7C_2 + s^2G_7C_1C_2)\}\end{aligned} \tag{2.20}$$

Here G_X is the conductance corresponding to resistor R_X. In the derivation, the operational amplifier has been considered to be ideal. More realistic models could have been taken.

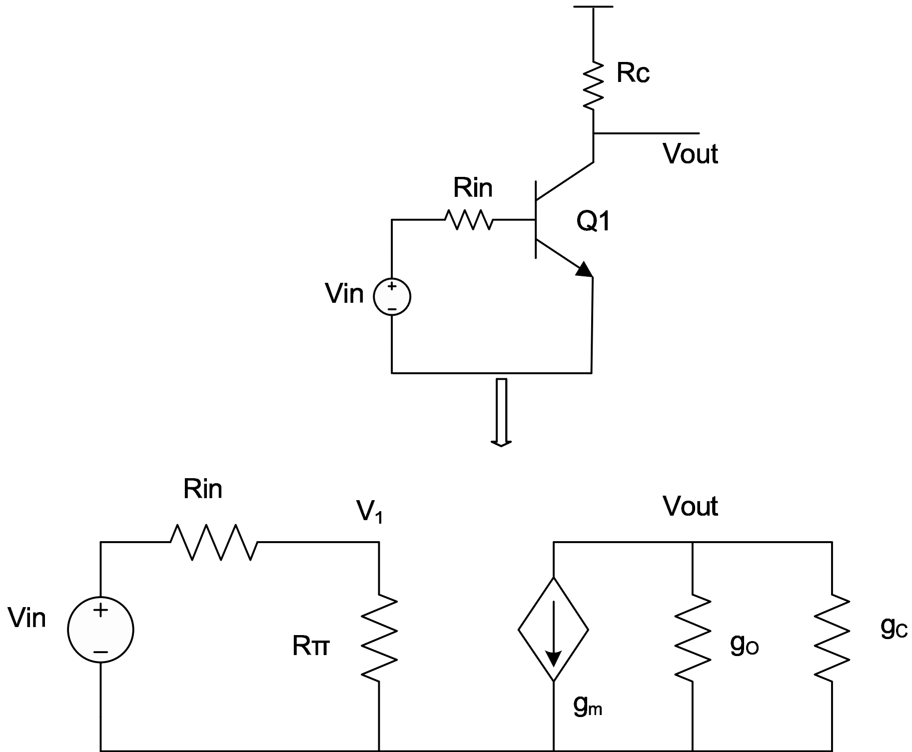


FIGURE 2.6
Small signal model of the transistor amplifier circuit.

2.3.3.2 Methodology

In order to explain the basic methodology of the symbolic analysis, let us consider a bipolar single-transistor amplifier as shown in Fig. 2.6. The input is a SPICE netlist. A linearized small-signal equivalent model of the circuit is generated. The behavior of the linearized circuit is described by a set of equations with symbolic coefficients, which is shown in matrix form below.

$$\begin{bmatrix} g_{in} + g_{\pi} & 0 \\ g_m & g_0 + g_L \end{bmatrix} \begin{bmatrix} V_1 \\ V_{out} \end{bmatrix} = \begin{bmatrix} g_{in} \\ 0 \end{bmatrix} V_{in} \tag{2.21}$$

After simplification, the simplified form of the network function is

$$\frac{V_{out}}{V_{in}} = -\frac{g_m}{g_c} \tag{2.22}$$

The network function is obtained by algebraic operations on this set of equations. This method of obtaining a symbolic network function is referred to as the algebraic (matrix or determinant) method. An alternative technique

is to represent the behavior of the linearized circuit with a graph containing symbolic branch weights. The network function is obtained by suitable operations on this graph, such as the enumeration of loops or a spanning tree.

2.3.3.3 Simplification of Expressions

Whatever technique is used, the symbolic network function so generated can be further post-processed in order to make the function compact [152]. This is desired to reduce the CPU time and memory consumption required for large circuits. With this, an approximate symbolic expression $h(\bar{p})$ for the original symbolic circuit characteristic $g(\bar{p})$ is obtained such that

$$\left\| \frac{g(\bar{p}) - h(\bar{p})}{g(\bar{p})} \right\| \leq \epsilon_{\max} \tag{2.23}$$

where ϵ_{max} is a predefined error. The approximation technique consists of discarding less significant terms from an expression. Consider, for example, the expression $g_{m1} + g_{01} + g_{02} + s(C_L + C_{db1})$. If it can be assumed that $g_{m1} >> g_{01}, g_{02}$ and $C_L >> C_{db1}$, then this expression can be simplified to $g_{m1} + sC_L$. The simplified expression distinctly reveals the dominant contributions at the penalty of a small error. Thus symbolic expression approximation is a trade-off between the expression accuracy (error) and the expression simplicity (complexity).

There are several strategies for simplification of a symbolic function. These are (i) simplification after generation (SAG), (ii) simplification during generation (SDG) and (iii) simplification before generation (SBG). The SAG technique is the conventional technique in which the exact symbolic solution is first calculated and then the least significant terms are eliminated. Only a few terms of the generated expression are usually kept, so that a lot of resources required to generate the pruned terms are wasted. The SDG techniques aim to calculate directly an approximated solution, which contains only the dominant terms. The idea is to build the wanted simplified expression by generating the terms one by one on decreasing order of dominance, until the approximation error is below the maximum user-supplied value. The SDG technique offers two advantages over the conventional SAG technique. First, the analysis procedure is fast as there is no wastage of time in generating insignificant terms, and second, the complexity of the circuit that can be analyzed is increased due to smaller memory consumption. In order to further extend the capabilities of the symbolic analysis technique to a larger circuit, that is still used as a big, flat circuit, the SBG technique is used. This consists of *a priori* removal of unimportant elements or shortening of nodes. Partial removal of elements in order to make the circuit small is also allowed. For example, biasing circuitry may be replaced by a single voltage or current source, because it will not affect the input-output transfer function of the circuit. But it makes the symbolic calculations complex. The added value of SBG is that it quantifies the impact of every simplification.

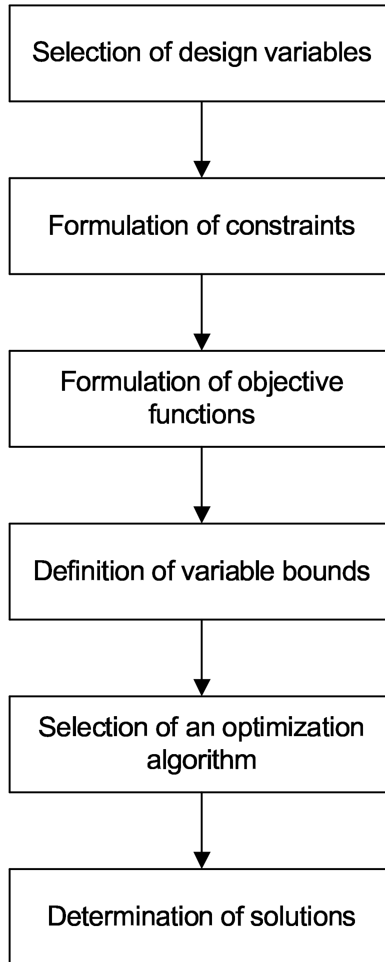


FIGURE 2.7
Flow chart for the steps of formulating an optimization problem.

2.4 Introduction to Optimization Techniques

The optimization procedure is the second major component of the knowledge-based CAD techniques discussed in Chapter 1. An optimization algorithm is procedure that is executed iteratively by comparing various candidate solutions within a defined search space till the optimum or a satisfactory solution is obtained [42, 20]. This section provides a comprehensive discussion about the formulation of a general optimization problem and the commonly used solution techniques.

2.4.1 Optimization Problem Formulation

A flow chart illustrating the tasks for formulating an optimization problem is provided in Fig. 2.7. The task of formulating an optimization problem begins with the task of identification of the design variables. The design variables are those variables which are primarily varied through an iterative procedure in order to get an optimal solution. While the final solution depends upon several design variables, only those parameters which have dominant first order effect on the solution are generally considered as design variables in an optimization problem. The efficiency and speed of optimization algorithms critically depend upon the number of selected design variables. The first thumb rule of formulating an optimization problem is to select a small set of design variables at the beginning. Depending upon the performance of the algorithm in terms of finding an optimal solution and computational complexity, the size of the set can be increased. In a circuit sizing problem, the transistor size is the design variable considered to critically affect the performance which is selected to be optimized.

The second task is the formulation of constraints associated with the optimization problem. The constraints represent some functional relationships among the design variables, which are as significant as the function to be optimized. In fact, a solution which appears to be the best in optimizing an objective function if not satisfying the constraints will not be considered as a feasible solution, and hence, will be discarded. For example, in a circuit sizing problem, if the problem is minimization of power dissipation, then constraints may be the satisfaction of desired gain and bandwidth. There are usually two types constraints used in an optimization problem: the inequality constraint and the equality constraint. Inequality constraints signify that the functional relationship among the design variables are either greater than, smaller than, or equal to a desired value. On the other hand, equality constraints signify that the functional relationships should exactly match a desired value. The latter type of constraints are usually difficult to meet and need to be avoided, wherever possible. Equality constraints are sometime helpful in eliminating the number of design variables and hence reduce the number of design

variables of an optimization problem. For example, in a circuit sizing problem for determining the sizes of a differential amplifier, the transistor dimensions of the input MOS transistors are equal. Therefore, either of these can be considered as the design variable, instead of selecting both of these.

The third task is to formulate the objective function in terms of the design variables. A majority of the optimization problems involve minimization of the objective function. However, some problems involve maximization, which in turn can be converted to a minimization problem by multiplying the objective function with -1 and vice versa. In several problems, there could be more than one objective that the designer may want to optimize simultaneously. Such classes of optimization problems are referred to as multi-objective optimization problems. The multi-objective optimization algorithms are complex and computationally expensive. Hence, if possible, these may be avoided. A well known approach to avoid multi-objective optimization problem is to consider the most important objective as the objective function of the optimization problem, and treat the other objectives as constraints by restricting their values within a certain range.

The final task of the optimization problem formulation is to set the minimum and the maximum bounds on each design variable. The task of determination of the variable bounds is not an easy one and requires detailed knowledge of the problem under consideration. A useful approach to select the variable bounds is to make an initial guess about the optimal solution and select the variable bounds based on it. Then these may be improved by actually solving the problem. If the design variables corresponding to the optimal solution are found to lie on or near the minimum or the maximum bound, the chosen bound may not be correct.

After completion of all the tasks, the optimization problem can be mathematically written in a special format, known as the nonlinear programming (NLP) format, which is as follows [42]

$$\begin{array}{lll} & \text{Minimize} \quad f(\bar{\alpha}) & \\ \text{subject to} & & \\ & g_j(\bar{\alpha}) \geq 0, & j = 1, 2, ..., J; \\ & h_k(\bar{\alpha}) = 0, & k = 1, 2, ..., K; \\ & \alpha_i^{(L)} \leq \alpha_i \leq \alpha_i^{(U)}, & i = 1, 2, ..., N; \end{array} \tag{2.24}$$

The constraints are written so that the right-side of the inequality or equality sign is zero. It may be noted that the four tasks mentioned above are not independent of each other. Some constraints may be included or deleted by the designers while formulating the objective function. Even the selection of design variables in some cases is fine tuned after some iterations. Whatever updating of these is made, depends upon the nature of the problem and the kind of optimization algorithms that are selected. Therefore, it is evident that it is almost impossible to apply a single formulation procedure to all design problems.

2.4.2 Optimality Criteria

The following three types of optimal points are first defined [149, 42].

1. A point or solution α^* is said to be a local optimal point if there does not exist any point in the neighborhood which has a function value smaller than $f(\alpha^*)$
2. A point or solution α^{**} is said to be a global optimal point if no point in the entire design space has a function value smaller than $f(\alpha^{**})$
3. A point or solution α^* is said to be an inflection point if the function value increases locally as α^* increases and decreases locally as α^* reduces, or if the function value decreases locally as α^* increases and increases locally as α^* decreases.

Without presenting the mathematical derivations, the sufficient conditions of optimality are given as follows[42]:

Let at the point α^*, the first derivative of the function be zero and the first nonzero higher derivative be denoted by θ; then

- *If θ is odd, α^* is an inflection point*
- *If θ is even, α^* is a local optimum.*

  1. *If the derivative is positive, α^* is a local minimum.*
  2. *If the derivative is negative, α^* is a local maximum.*

2.4.3 Classification of Optimization Algorithms

The optimization algorithms are classified into a number of types, which are now briefly discussed.

2.4.3.1 Single-Variable Optimization Algorithms

Single-variable functions involve only one variable. Therefore, single-variable optimization algorithms are very simple. These algorithms are again classified into two types: direct techniques and gradient-based techniques. Direct techniques do not require any derivative information of the objective function; only objective function values are used to guide the search process. On the other hand, in the gradient-based techniques, the first and/or second-order derivatives of the objective function are used to guide the search process.

2.4.3.2 Multi-Variable Optimization Algorithm

Here the objective function consists of more than one design variable. In a multi-variable function, the gradient of a function is a vector quantity. Assuming that the objective function is a function of N variables represented by

$\alpha_1, \alpha_2, ..., \alpha_N$, the gradient vector at any point $\bar{\alpha}^t$ is represented by $\nabla f(\bar{\alpha}^t)$ and is defined as follows [42, 149]:

$$\nabla f(\bar{\alpha}^t) = \left(\frac{\partial f}{\partial \alpha_1}, \frac{\partial f}{\partial \alpha_2}, ..., \frac{\partial f}{\partial \alpha_N} \right)^T \Bigg|_{\bar{\alpha}^t} \tag{2.25}$$

The first-order partial derivatives are computed numerically. The second-order derivatives for a multi-variable function form a matrix, which is known as the Hessian matrix and is defined as follows:

$$\nabla^2 f(\bar{\alpha}^t) = \begin{bmatrix} \frac{\partial^2 f}{\partial \alpha_1^2} & \frac{\partial^2 f}{\partial \alpha_1 \partial \alpha_2} & .... & \frac{\partial^2 f}{\partial \alpha_1 \partial \alpha_N} \\ \frac{\partial^2 f}{\partial \alpha_1 \partial \alpha_2} & \frac{\partial^2 f}{\partial \alpha_2^2} & .... & \frac{\partial^2 f}{\partial \alpha_2 \partial \alpha_N} \\ . & . & .... & . \\ . & . & .... & . \\ . & . & .... & . \\ \frac{\partial^2 f}{\partial \alpha_N \partial \alpha_1} & \frac{\partial^2 f}{\partial \alpha_N \partial \alpha_2} & .... & \frac{\partial^2 f}{\partial \alpha_N^2} \end{bmatrix}_{\bar{\alpha}^t} \tag{2.26}$$

The first and second-order derivatives are computed numerically through the central difference technique. The optimality criteria is

A point $\bar{\alpha}^t$ is a stationary point if $\nabla f(\bar{\alpha}^t) = 0$. Furthermore, the point is a minimum, a maximum, or an inflection point if $\nabla^2 f(\bar{\alpha}^t)$ is positive-definite, negative-definite, or otherwise. A commonly used way to identify whether a matrix is positive-definite or not is to evaluate the eigenvalues of the matrix, If all the eigenvalues are positive the matrix is positive-definite [183]. The alternative approach is to calculate the principal determinants of the matrix and if all the principal determinants are positive, the matrix is positive-definite.

2.4.3.3 Constrained Optimization Algorithms

Constrained optimization algorithms locate the final solution point within a feasible search space. The constrained optimization algorithms are grouped into direct and gradient-based methods. Some of these algorithms employ single-variable and multi-variable unconstrained optimization algorithms. Constrained algorithms are mostly used in engineering design problems. A general constrained optimization problem is defined in (2.24). The Lagrange multiplier method is an elegant formulation to obtain the solution to a constrained problem [198]. It allows the transformation of a constrained problem into an unconstrained problem. Using this technique, the inequality and equality constraints are added to the objective function to form an unconstrained problem. The following Kuhn–Tucker conditions are obtained by satisfying the first-order optimality condition of the resultant unconstrained problem.

$$\nabla f(\bar{\alpha}) - \sum_{j=1}^{J} \beta_j \nabla g_j(\bar{\alpha}) - \sum_{k=1}^{K} \lambda_k \nabla h_k(\bar{\alpha}) = 0 \tag{2.27}$$

$$g_j(\bar{\alpha}) \geq 0, \quad j = 1, 2, ..., J; \tag{2.28}$$

$$h_k(\bar{\alpha}) = 0, \quad k = 1, 2, ..., K; \tag{2.29}$$

$$\beta_j g_j(\bar{\alpha}) = 0, \quad j = 1, 2, ..., J; \tag{2.30}$$

$$\beta_j \geq 0, \quad j = 1, 2, ..., J; \tag{2.31}$$

The Lagrange multiplier β_j corresponds to the j-th inequality constraint and the multiplier λ_k corresponds to the k-th equality constraint. In the K-T condition, the first equation is due to the optimality condition of the unconstrained Lagrangian function. The second and third equations are required for satisfying the constraints. The fourth equation arises only for inequality constraints. This is defined such that if an inequality constraint is inactive at a point $\bar{\alpha}$, i.e., $g_j(\bar{\alpha}) > 0$, $\beta_j = 0$ and if the constraint is active, i.e., $g_j(\bar{\alpha}) = 0$, then $\beta_j g_j = 0$. The final inequality condition suggests that in the case of active constraints, the corresponding Lagrange multiplier must be positive. A point $\bar{\alpha}^t$ and two vectors $\bar{\beta}$ and $\bar{\lambda}$ that satisfy all the above conditions are referred to as the Kuhn–Tucker points. There are a total of $(N + 3J + K)$ Kuhn–Tucker conditions. If there exists at least one set of $\bar{\beta}$ and $\bar{\lambda}$ vectors, which satisfies all K–T conditions, the point is said to be a K–T point.

2.4.3.4 Specialized Optimization Algorithms

There are several algorithms which are applicable for only a certain class of optimization problems. Two commonly used such algorithms are integer programming and geometric programming (GP). The GP technique is designed to solve NLP problems which can only be expressed in posynomial form. Several MOS transistor performances can be expressed in posynomial so that this method is widely used in analog circuit sizing.

2.4.3.5 Nontraditional Stochastic Optimization Algorithms

There are two important limitations of the traditional deterministic technique. First, it is often not possible to determine the gradient of the cost function. Second, the optimization process in many cases is quickly trapped in a local optimum of the cost function. Another problem is the rapid increase of the execution time with the increase in the number of design variables and design space. These techniques are used primarily for the fine tuning of sub-optimal sizings. On the other hand, in nontraditional stochastic techniques, the algorithm moves from one solution point to another with probabilistic transition rules, and the design variables are varied randomly. The derivatives of the cost function are not required. Greedy stochastic algorithms only accept a new set of variables if it reduces the cost function value. The main advantage of the stochastic methods over the deterministic ones is the capability to escape from local optimum and hence a higher probability to reach a global optimum. Because of this, these algorithms are very popular to solve engineering design problems. Examples include simulated annealing, genetic algorithms, particle swarm optimization algorithms, ant colony optimization algorithms, etc.

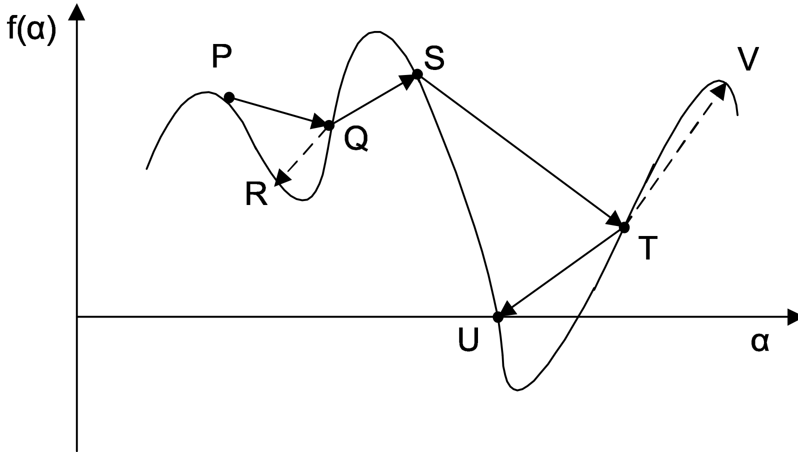


FIGURE 2.8
Strategy for global optimum search.

2.4.4 Concept of Local Optima and Global Optima

The majority of the optimization problems used in analog circuit CAD contain an objective function consisting of a number of optima of which one (or more) is the global optimum. Other optima have worse function values compared to that of the global optima. Therefore, it is essential for the designers to identify the global optimum point which corresponds to the best function value. For example, the Himmelblau function: $f(\alpha_1, \alpha_2) = \left(x_1^2 + x_2 - 11\right)^2 + \left(x_1 + x_2^2 - 7\right)^2$ consists of four minima, viz., $(3,2)^T$, $(-2.805, 3.131)^T$, $(-3.779, -3.283)^T$ and $(3.584, -1.848)^T$. The global optimum cannot really be characterized as different from the local optimum. The standard Kuhn-Tucker conditions identify the local optimum only. It is very difficult to identify the existence of a global optimum point within a problem. The KT conditions lead the solution to be trapped in any one of the local optima, depending on the location of the starting point for the iterations. At present, only continuous convex problems are guaranteed to have a global solution. If an optimization problem can be shown to be convex, then the local minimum will also be the global minimum.

The strategy for a global optimum search is schematically illustrated in Fig. 2.8. Let us suppose that the iteration starts from the point P. A search direction to point Q is found through standard approach. If a local optimization algorithm is used, the search direction leading to point R will be accepted and any search direction toward point S will be rejected. But if point S is found somehow, it could move on to T and then U and continue to the global optimum. The central idea of finding the global minimum is therefore to encourage solution points that are not trapped within the local minimum through conventional optimality criterion. Therefore, it may happen that from point T,

the search direction moves towards V. Therefore, it is observed that with this strategy although it is possible to escape the local minimum trap, theoretically it takes a large amount of time to actually locate the global minimum point. Therefore, sometimes a local search algorithm is used once point T is located.

2.4.5 Characterization of Optimization Algorithms

The various optimization algorithms are characterized by four important metrics, which are discussed below:

1. *Convergence.* An ordered set of real numbers $a_1, a_2, a_3, ..., a_n$ is called a sequence and is denoted by (a_n). If the number of terms in this set is unlimited, then the sequence is said to be an infinite sequence and a_n is its general term. A sequence is said to tend to a limit l, if for every $\epsilon > 0$, a value N of n can be found such that $|a_n - l| < \epsilon$ for $n \geq N$. If the sequence (a_n) has a finite limit, it is called a convergent sequence, otherwise it is said to be a divergent [183]. The sequence (a_n) is said to be bounded, if there exists a number k such that $a_n < k$ for every n. The sequence (a_n) increases steadily or decreases steadily according as $a_{n+1} \geq a_n$ or $a_{n+1} \leq a_n$, for all values of n. Both increasing and decreasing sequences are called monotonic sequences. Therefore, a sequence which is monotonic and bounded is convergent [183]. In an optimization algorithm, the set of dependent values evaluated through the objective function forms a sequence. The convergence of an optimization algorithm is assessed by observing the magnitude of the change over the last few terms of the sequence. A commonly used criterion for stopping the search process of an algorithm is when the difference between the successive values of the function is sufficiently small. Another useful criterion is to stop the search when the distance moved in the parameter space during the last few iterations is less than the location accuracy required.

2. *Reliability.* The reliability of an optimization technique is its ability to find the global minimum and therefore to avoid local minima. The reliability is characterized by percentage of success out of several numbers (50 say) of run.

3. *Accuracy.* The accuracy is measured by the distance of a solution point from the global minima and is thus characterized through several error metrics like mean square error, etc.,. For error minimization algorithms, the stopping criterion is often expressed in terms of the calculated error, when the error is less than a predefined value, the algorithm stops. Since the exact global optimum value is generally not known *a priori*, the accuracy is measured in terms of the optimum achievable value averaged over several runs.

4. *Computation time.* The total CPU computation time of an algorithm is determined by the number of evaluations of the objective function and the time taken for evaluating once the objective function. Since the latter is dependent upon the chosen problem, the computation time of an optimization algorithm is characterized by the number executing the objective function.

2.5 Some Important Optimization Algorithms

Some important optimization algorithms from the view point of CAD techniques for analog ICs are described below. It may be noted that due to the growing complexity of the design problem, the designer can no longer afford to rely on a particular algorithm. Two algorithms, usually one global search algorithm and another local search algorithm are often conjoined to implement a design space exploration procedure.

2.5.1 Cauchy's and Newton's Steepest Descent Algorithm

The optimization algorithm is framed as a search algorithm. The gradient of a function at a point is the direction of the most rapid increase in the value of the function at that point. The descent direction is obtained by multiplying the gradient by -1. Thus the search direction at any point is $\bar{\alpha}^t$ is $\bar{s}^t = -\nabla f(\bar{\alpha}^t)$. The algorithm starts with an initial guess $\bar{\alpha}^0$ about the solution point. At each iteration, the next minimum point from the current point is obtained through a unidirectional search along $\bar{s}^t$. The minimum point becomes the current point and the search is continued from that point. The algorithm continues until a point having a small enough gradient vector is obtained. The algorithm guarantees improvement in the function value at every iteration. The step-by-step procedure is [42]:

1. Let M = maximum number of iterations to be performed, $\bar{\alpha}^0$ = initial point, ϵ_1, ϵ_2 be the two termination parameters and set $t = 0$
2. Calculate $\nabla f(\bar{\alpha}^t)$
3. If $\|\nabla f(\bar{\alpha}^t)\| \leq \epsilon_1$, **Terminate**;
   Else if $t \geq M$; **Terminate**;
   Else go to Step 4.
4. Next point is $\bar{\alpha}^{(t+1)} = \bar{\alpha}^t + \bar{d}^t.\bar{s}^t$. Do a unidirectional search to find $\bar{d}^t$ such that $f(\bar{\alpha}^{(t+1)}) = f(\bar{\alpha}^t + \bar{d}^t.\bar{s}^t)$ is minimum. Termination criterion is $\left|\nabla f(\bar{\alpha}^{(t+1)}).\nabla f(\bar{\alpha}^t)\right| \leq \epsilon_2$.
5. Check whether $\frac{\|\bar{\alpha}^{(t+1)} - \bar{\alpha}^t\|}{\|\bar{\alpha}^t\|} \leq \epsilon_1$?If yes; **Terminate**;
   Else set $t = t + 1$ and go to Step 2.

The unidirectional search is performed through golden section search technique [42, 198]. It may be noted that Cauchy's technique is well suited for the case when $\bar{\alpha}^0$ is far away from the optimum point $\bar{\alpha}^*$. When the current point is very close to the optimum point, the change in gradient vector is very small. Therefore, the next point will also be close to the current point. Therefore, the algorithm is slowed down near the true minimum. The convergence can be made faster by using a second derivative, which is done in Newton's method. The steepest descent is $s^t = -\left[\nabla^2 f(\bar{\alpha}^t)\right]^{-1} \nabla f(\bar{\alpha}^t)$. Newton's technique is identical to Cauchy's steepest descent technique except that the next point is calculated through the second derivative method. This technique is suitable and efficient when the initial point is close to the optimum point. Combining the advantage of each algorithm, Marquardt's technique is formulated, where Cauchy's technique is initially followed and thereafter, Newton's technique is used. The transition from Cauchy's technique to Newton's technique is adaptive and depends on the history of the obtained intermediate solutions.

2.5.2 Genetic Algorithm

A Genetic Algorithm (GA) is a search based optimization method that draws inspiration from the concept of natural selection and survival of the fittest in the biological world [67]. GA falls into the wider category of search methods known as the Evolutionary Algorithms (EAs). The GA starts with an initial population whose elements are called *chromosomes*. A chromosome consists of a fixed number of variables, which are called *genes*. In order to evaluate and rank the chromosomes in a population, a *fitness function* based on the objective function is defined. A set of three operators are specified to construct the complete structure of a GA procedure. These are *selection/reproduction*, *crossover* and *mutation* operators. The selection operator selects an intermediate population from the current one in order to be used by the other operators; crossover and mutation. In this selection process, the chromosomes with higher fitness function values have a greater chance to be chosen than those with lower fitness function values. The crossover operator defines how the selected chromosomes (parents) are recombined to create new structures (offspring) for possible inclusion in the population. Mutation is a random modification of a randomly selected chromosome. Its function is to guarantee the possibility of exploring the space of solutions for any initial population and to permit the escape from a zone of local minimum. The GA operators; selection, crossover and mutation have been extensively studied. Several implementation techniques of these operators have been proposed to fit a wide variety of problems. More details about the GA elements are discussed below before stating a standard GA procedure.

1. *Fitness Function:* A fitness function F is a designed function that measures the goodness of a solution. It is designed in such a way that better solutions have a higher fitness function value than worse

solutions. The following fitness function is often used

$$F(\bar{\alpha}) = \frac{1}{1 + f(\bar{\alpha})} \tag{2.32}$$

where $f(\bar{\alpha})$ is the objective function. The fitness function plays a major role in the selection process.

2. *Coding:* Coding in GA defines the forms in which chromosomes and genes are expressed. There are mainly two types of coding; binary and real. Binary GA requires the solutions to be coded as finite-length binary strings of 1's and 0's. This is naturally suited to combinatorial optimization problems with discrete search spaces. In real-parameter GA, the solutions are represented as direct real numbers. Binary GA presents a number of difficulties like Hamming cliffs and inability to achieve any arbitrary precision when applied to problems with continuous search spaces. To avoid these limitations, the real-parameter GAs are developed.

3. *Selection:* Genetic Algorithm is modeled on Darwin's evolution theory of the survival of the fittest. Thus, in any generation of solutions, the best ones survive with higher probability and create offspring. There exists a number of selection operators for reproduction in the GA literatures but, the essential idea in all of them is that solutions are selected from the current population and their multiple copies are inserted in the mating pool in a probabilistic manner. The various methods of selecting chromosomes from the pool of parent solutions are: proportionate selection, tournament selection, and rank selection etc. The proportionate selection is the most commonly used selection method and is usually implemented with a roulette-wheel simulation method. Every solution is assigned a fitness value F_i, and has a roulette-wheel slot sized in proportion to its fitness. In order to create a new population, the roulette-wheel is spun n times, each time selecting an instance of the solution chosen by the roulette wheel pointer. Thus, the probability p_i of selecting the ith solution is given by

$$p_i = \frac{F_i}{\sum_{i=1}^{n} F_i} \tag{2.33}$$

4. *Crossover:* A crossover operator aims to interchange the information and genes between chromosomes. Therefore, crossover operators combine two or more parents to reproduce new children. One of these children possibly collects all the good features that exist in his parents. A crossover operator is applied with probability p_c. The uniform crossover technique is a commonly used crossover technique. Two arbitrary chromosomes (parents) are randomly selected from the population and their genes are rearranged at several

crossover points, which are determined randomly in order to generate two new chromosomes (children).

5. *Mutation:* The mutation operator is used with a low probability p_m to alter the solutions locally to possibly create better solutions. The need for mutation is to maintain a good diversity of the population. Although this operator performs a random change in the solution chosen for mutation, the low mutation probability ensures that the process creates only a few such solutions in the search space and the evolution does not become random.
6. *Elite-Preserving Operator:* In order to ensure that the statistics of the population-best solutions do not degrade with generations, the elite-preserving operator is often used in GAs. Typically, the best $\alpha\%$ of the population from the current population is directly copied to the next generation. The rest of the new population is created by the usual genetic operations applied on the entire current population. Thus, the best solutions of the current population not only get passed from one generation to another, but they also participate with other members of the population in creating other population members.

With this background on GA operators, a simple GA procedure utilizing these operators is presented below, based upon [42].

1. Select an appropriate coding scheme to represent the design variables, a selection operator, a crossover operator and a mutation operator. Select a population size n, crossover probability p_c, and mutation probability p_m. Initialize a random population of chromosomes of size l. Choose a maximum allowable generation number t_{max}. Set $t = 0$.
2. Evaluate each chromosome in the population.
3. If $t > t_{max}$ or other termination criteria is satisfied, Terminate.
4. Perform reproduction operation on the population.
5. Perform crossover operation on random pairs of chromosomes.
6. Perform mutation operation on every chromosome.
7. Evaluate chromosomes in the new population. Set $t = t + 1$ and go to step 3.

The algorithm is straightforward with repeated application of the three operators discussed earlier to a population of points. This algorithm is widely used for the sizing of analog circuits [175, 40]

2.5.3 Simulated Annealing

A simulated annealing (SA) procedure simulates an annealing process to achieve the minimum function value in a minimization problem. The SA algorithm successively generates a trial point in a neighborhood of the current solution and determines whether or not the current solution is to be replaced by the trial point based on a probability, depending on the difference between their function values [42, 198]. Suppose at any instant the current point is $\bar{\alpha}^t$ and the function value is $E(t) = f(\bar{\alpha}^t)$. Using the Metropolis algorithm, the probability of the next point $\bar{\alpha}^{t+1}$ depends on the difference in the function value at these points, i.e., on $\Delta E = E(t+1) - E(t)$. This is given by the following Boltzmann probability distribution

$$P(E(t+1)) = \min\left[1, \exp(-\Delta E/kT)\right] \tag{2.34}$$

If $\Delta E \leq 0$, the probability is unity and the point is always accepted. There is nothing special about it. But it is interesting to note that even if $\Delta E > 0$, according to Metropolis algorithm, there is some finite probability of selecting the point $\bar{\alpha}^{t+1}$. This is in line with the search strategy discussed earlier for global optima point. The probability of acceptance however, depends on the relative magnitude of ΔE and T values. If the value of T is large, which generally is the case for the initial iterations, the probability of accepting such points is high. On the other hand, for low value of T, the probability of accepting such new points is low. The algorithm begins with an initial point and a high temperature T. The next point is created in the neighborhood of the first and is accepted depending upon the difference of the function values and the value of T. This completes one iteration of the SA procedure. A number of points is usually tested at a particular temperature before reducing the value of T. The SA algorithm is stated as follows, based upon [42]:

1. Select an initial point $\bar{\alpha}^1$, a termination criteria ϵ. Set T a sufficiently large value, the number of iterations performed at a particular temperature is n, and set t = 0.
2. Calculate a neighboring point $\bar{\alpha}^2$. Generally, a random point in the neighborhood is created.
3. If $\Delta E = E(\bar{\alpha}^2) - E(\bar{\alpha}^1) < 0$, set $t = t + 1$;
   Else create a random number r in the range $(0, 1)$. If $r \leq \exp(-\Delta E/T)$ set $t = t + 1$;
   Else go to step 2.
4. If $\left|\bar{\alpha}^2 - x^1\right| < \epsilon$ and T is small, terminate
   Else if $(t \bmod n) = 0$ then reduce T according to a cooling schedule.
   Go to step 2;
5. Else go to step 2.

One of the most powerful features of SA is its ability to escape easily from being trapped in local minima by accepting up-hill moves through a probabilistic

procedure, especially in the earlier stages of the search. On the other hand, the main drawbacks that have been noticed on SA are its suffering from slow convergence and its wandering around the optimal solution if high accuracy is needed.

Some of the key analog CAD tools using SA as the optimization tool are OPTIMAN [62], ASTRX/OBLX [137], ORCA [35], SD-Opt [128] and so on. On the other hand, some of the key analog CAD tools using genetic algorithm are ANTIGONE [127], Watson [174] etc.,.

2.6 Multi-Objective Optimization Method

In virtually every engineering problem a trade-off exists between two or more competing objectives, i.e., improving one forces the other(s) to worsen. In analog design optimization problems, multi-objective optimization problems often need to be solved. This is because the analog performance parameters are often tightly coupled and competitive in nature.

Traditional single-objective optimization algorithms provide only one solution (sometimes a set of candidate solutions) to such problems which minimizes/maximizes an overall objective function obtained by mixing individual targets through appropriate weightage factors. The use of the single-objective optimization technique for solving trade-off problems leads to inferior results for several reasons.

- The results of such an optimization process are values of design variables for which global cost function is minimized. However, no information is available on how far this design point is from the optimal value for each of the (possibly conflicting) individual objectives.

- With this technique, it is not possible to know the specific design variable which is driving the optimizer toward the solution obtained.

- There is no formal procedure for the choice of individual weights through which the individual cost functions are combined.

A multi-objective optimization problem is formally defined as follows [41]:

$$
\begin{array}{lrll}
\text{Minimize} & \bar{f}(\bar{\alpha}), & & \\
\text{subject to} & g_j(\bar{\alpha}) \geq 0, & j = 1, 2, ..., J; & \\
 & h_k(\bar{\alpha}) = 0, & k = 1, 2, ..., K; & \\
 & \alpha_i^L \leq \alpha_i \leq \alpha_i^U, & i = 1, 2, ..., n. &
\end{array}
\tag{2.35}
$$

A solution $\bar{\alpha}$ is a vector of n design variables: $\bar{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_n]^T$ and $\bar{f}(\bar{\alpha})$ is a vector of objective functions.

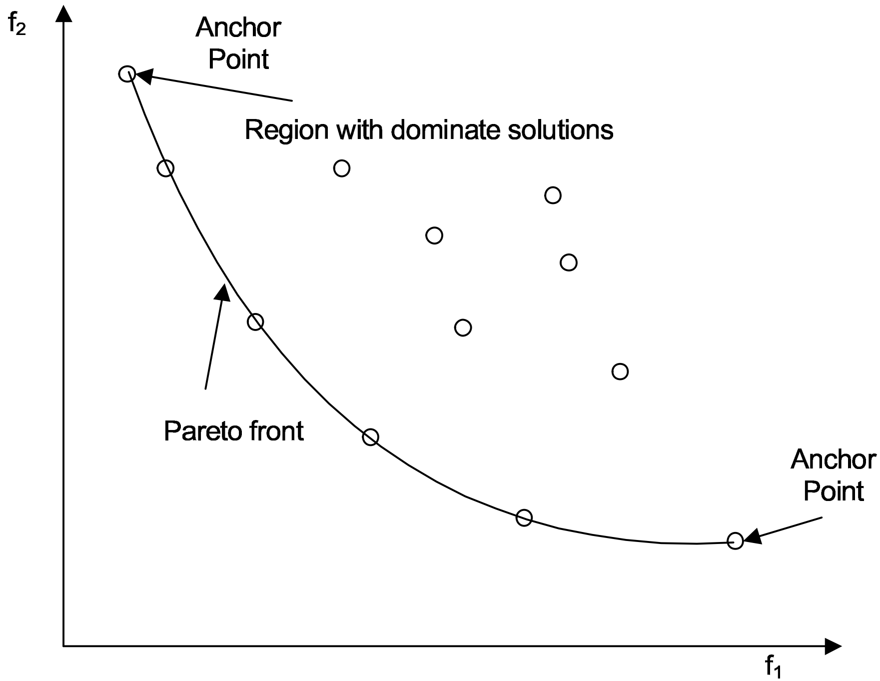


FIGURE 2.9
Illustration of Pareto front for a sample 2D design space.

2.6.1 Pareto Optimal Front

For multi-objective optimization problems, it is not always possible to conclude that one solution is better than the other as it may be better in one objective but worse in others. However, some cases arise where one solution outperforms another in all given specifications. In such cases, that particular solution is considered to dominate others. The set of solutions of a design that is not dominated by any of the other solutions is called the Pareto optimal set, which represents all the trade-offs involved in the design procedure. A solution point $\bar{\alpha_i} \in \mathcal{D}$ is said to be Pareto optimal if and only if there does not exist another point $\bar{\alpha_j} \in \mathcal{D}$, such that $\bar{f}(\bar{\alpha_j}) \leq \bar{f}(\bar{\alpha_i})$ and $f_m(\bar{\alpha_j}) \leq f_m(\bar{\alpha_i})$ for at least one function. Figure 2.9 illustrates an example of Pareto front in a 2D optimization problem for the minimization of both cost functions. The anchor points are the extreme optimal results of the specifications, while the other points in the Pareto front are usually trade-offs between several specifications. A and B are two specific points among other. Improvement in one objective f_1, i.e., $f_{1A} < f_{1B}$ requires a degradation in the other $f_{2A} > f_{2B}$ objective. The concept of Pareto optimality is important to analog sizing process because it usually attempts the minimization of several objectives simultaneously. The Pareto optimal points are interesting to the designers because they charac-

terize all optimum performance values achievable by a component block and the trade-offs that are involved [179, 180, 181]. Therefore, with the knowledge of the Pareto optimal front, the designers can select the whole set of equally optimal candidate solutions.

The aim of a multi-objective optimization algorithm is to find samples of the Pareto front with respect to the complete set of feasible design vectors. The genetic algorithm is often used to solve multi-objective optimization problems.

2.7 Design Space Exploration

Design space exploration (DSE) refers to the task of discovering and evaluating design alternatives during the design development process. This needs to be carried out very carefully because of the sheer number of design alternatives to be explored. For sizing of complex circuits, the number of design alternatives could be several thousands, which however, includes both feasible and infeasible. A manual, ad-hoc approach to DSE is tedious, error-prone, and does not scale. The three essential components of a DSE procedure are (i) design space representation, (ii) design performance estimation and (iii) exploration strategy.

The task of design space representation establishes the dimensions of the design space, which usually contains multiple dimensions. A practical design space is a bounded design space, which is defined by the design variable bounds. The representation should be formal, so that it can be subject to automated analysis and exploration techniques. In addition, the representation should be expressive enough to capture the various feasibility constraints. Estimation techniques consist of quantitative measures used to assess the various candidate designs. The estimation techniques must also be able to tackle the challenge of solving a large number of complex constraints at reasonable computational costs. The exploration strategy is the approach of visiting different design alternatives within the design space. It may be noted that in the parlance of design space exploration, a design is represented by a particular value for each design variable and the optimal design in terms of various optimization criteria is the Pareto design.

2.8 Computational Complexity of a CAD Algorithm

An algorithm is a sequence of well-defined steps/instructions to be executed for completing a task or solving a problem. A major criterion for a good algorithm is its efficiency, which is measured by the amount of time and memory required

to solve a particular problem. In real units, these are expressed in seconds and megabytes. However, these depend on the computing power of the specific machine and on the specific data set. In order to standardize the measurement of the efficiency of an algorithm, the computational complexity theory was developed. This theory allows one to estimate and express the efficiency of an algorithm as a mathematical function of its input size [33, 81]. The input size of an algorithm, in general refers to the number of items in the input data size. For example, when sorting n words, the input size is n.

2.8.1 Time and Space Complexity

Computational complexity is divided into (i) time complexity and (ii) space complexity. These estimate the time and memory requirements of an algorithm respectively. The time complexity of an algorithm is loosely considered to be the amount of computer time it needs to run to completion. The space complexity of an algorithm is the amount of memory it needs to run to completion. The former is considered more important compared to the later, because the memory requirements of the majority of algorithms is lower than the capacity of the current machines. Therefore, when the term complexity is used alone, it unambiguously mean time complexity.

The time complexity of an algorithm is calculated on the basis of the number of required elementary computational steps taken by the algorithm to compute the function it was developed for. The number of steps are interpreted as a function of the input size. However, it may be noted that most of the time, the total number of elementary computational steps varies from input to input because of the presence of conditional statements such as an if-else statement. Therefore, average-case complexity is considered to be a more meaningful characterization of the algorithm. However, accurate determination of the average-case complexity of an algorithm is not an easy task, which necessitates the use of the worst-case complexity metric. The worst-case complexity of an algorithm is the complexity with respect to the worst possible inputs, which gives an upper bound on the average-case complexity.

The running time of an algorithm increases with the size of the input. The rate of growth or the order of growth of the running time of an algorithm is the parameter that differentiates the efficiency between two algorithms. While comparing the goodness of two algorithms, the algorithm for which the order of growth is higher is considered to be inferior compared to the other. If the computational complexity of an algorithm can be expressed as an equation in terms of the input size n, then only the order of the dominating term needs to be considered, because other lower order terms are relatively insignificant for a large n. For example, if the complexity of any algorithm is expressed as $\lg n - 1 + 3/n$, then it can be simplified to only $\lg n$ leaving out the terms -1 and $3/n$. In addition, the constant coefficient of the dominating term can also be neglected while computing the complexity. Under this criterion, the complexities n and $n/2$ are virtually equal. In other words, they are said to

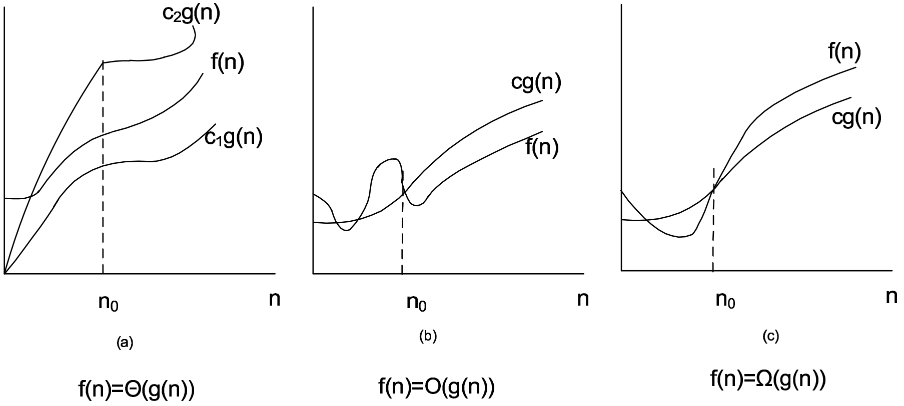


FIGURE 2.10
Graphical illustrations of Θ, O and Ω notations.

TABLE 2.1
Worst Case Time Complexities of Frequently Used Functions

Order	Name	n=100	n=10,000
$O(1)$	Constant time	1E-5 s	1E-5s
$O(\lg n)$	Logarithmic time	0.000008s	0.000013s
$O(n)$	Linear time	1E-3 s	0.01s
$O(n \lg n)$		0.00065s	0.13s
$O(n^2)$	Quadratic time	0.01s	100s
$O(n^3)$	Cubic time	1s	278 hours
$O(2^n)$	Exponential time	10^{14} centuries	10^{2995} centuries

have asymptotically equal complexity for larger n and are usually represented with several asymptotic notations.

2.8.2 Asymptotic Notations

Asymptotic notations capture how the running time of an algorithm grows with the size of the input. There are three types of asymptotic notations that are commonly used in the computational complexity theory. These are illustrated graphically in Fig. 2.10 and are discussed as follows.

2.8.2.1 Big-Oh Notation $O()$

It denotes the asymptotic upper bound of the complexity function. Let us consider a complexity function $f(n)$, and $g(n)$ be the asymptotic upper bound of $f(n)$. We denote by $O(g(n))$ (pronounced 'big-oh of g of n') the set of

functions [33]

$$O(g(n)) = \{f(n) : \exists \text{constants}\, c > 0, n_0 > 0 \text{ s.t. } 0 \leq f(n) \leq cg(n) \forall n \geq n_0\} \quad (2.36)$$

A non-negative function $f(n)$ belongs to the set of functions $O(g(n))$ if there exists positive constant c that makes $f(n) \leq cg(n)$ for a sufficiently large n. It is appropriate to write $f(n) \in O(g(n))$ because $O(g(n))$ is a set, but it is conventionally written as $f(n) = O(g(n))$. This is to be carefully noted that the equality sign denotes set memberships in all kinds of asymptotic notations.

Some examples: $3n + 2 = O(n)$ as $3n + 2 \leq 4n, \forall n \geq 2$, $10n^2 + 4n + 2 = O(n^2)$ as $10n^2 + 4n + 2 \leq 11n^2, \forall n \geq 5$. The statement $f(n) = O(g(n))$ states only that $g(n)$ is an upper bound on the value of $f(n), \forall n, n \geq n_0$. However, it does not say anything about how good this bound is. Therefore, for this statement to be informative, $g(n)$ should be as small a function of n as is possible for which $f(n) = O(g(n))$.

Table 2.1 [86] shows the most frequently used O-notations, their names and the comparisons of actual running times with different values of n. The actual running time on a million instructions per second machine is reported. The constant time complexity is designated as $O(1)$. It signifies that the running time of the algorithm is independent of the input size and is the most efficient. The other O notations are listed in their rank order of efficiency. If the time complexity of an algorithm can be expressed with or is asymptotically bounded by a polynomial function, it has polynomial time complexity. Otherwise, it has exponential time complexity.

2.8.2.2 Omega Notation $\Omega()$

Just as O-notation provides an asymptotic upper bound on a function, Ω-notation provides an asymptotic lower bound. Let us consider a complexity function $f(n)$, and $g(n)$ as the asymptotic lower bound of $f(n)$. We denote by $\Omega(g(n))$ (pronounced 'big-omega of g of n') the set of functions [33]

$$\Omega(g(n)) = \{f(n) : \exists \text{constants}\, c > 0, n_0 > 0 \text{ s.t. } 0 \leq cg(n) \leq f(n) \forall n \geq n_0\} \quad (2.37)$$

Some examples: $3n + 2 = \Omega(n)$ as $3n + 2 \geq 3n$ for $n \geq 1$, $10n^2 + 4n + 2 = \Omega(n^2)$ as $10n^2 + 4n + 2 \geq n^2, \forall n \geq 1$.

2.8.2.3 Theta Notation $\Theta()$

$\Theta()$ gives an asymptotic tight bound on a function, $g(n)$ is an asymptotically tight bound for $f(n)$. We denoted by $\Theta(g(n))$ (pronounced 'big-theta of g of n') the set of functions

$$\Theta(g(n)) = \{f(n) : \exists \text{constants}\, c_1 > 0, c_2 > 0, n_0 > 0 \text{ s.t. } 0 \leq c_1 g(n) \leq f(n) \leq c_2 g(n) \forall n \geq n_0\} \quad (2.38)$$

The theta notation is more precise than both the big oh and omega notations. The function $f(n) = \Theta(g(n))$ iff $g(n)$ is both an upper bound and lower bound on $f(n)$. However, O-notation is favored over Θ-notation for the following two reasons: (i) upper bounds are considered sufficient for characterizing the complexity of an algorithm and (ii) it is often much more difficult to prove a tight bound that it is to prove an upper bound.

2.8.3 Categorization of CAD Problems

The problems that are faced within a CAD technique are categorized into two groups according to the forms of their answers: decision problems and optimization problems. The answers to decision problems are either yes or no (or more formally 1 or 0). On the other hand, an optimization problem seeks an optimized value of a target variable. The feasibility checking problem for design variables, for example, is a decision problem, because the answer could only be whether the variables are feasible or not. On the other hand, finding the transistor dimensions of a topology for which certain performance is minimum is an optimization problem. An optimization problem is "hard" and the decision problem is in a sense "easier" or at least "no harder."

2.8.4 Complexity Classes for CAD Problems

Many of the CAD algorithms are polynomial time algorithms on inputs of size n, for which the worst case complexity is $O(n^k)$ for some constant k. However, all CAD problems cannot be solved in polynomial time. Several CAD problems are *NP* complete problems, for which no polynomial-time algorithm has yet been discovered, nor has anyone yet been able to prove that no polynomial-time algorithm can exist for any one of them. There are three classes of CAD problems: *P*, *NP* and *NPC*, the latter class being the *NP*-complete problems. An informal definition is provided here [33].

The class *P* consists of those problems that are solvable in polynomial time. The worst-case complexity of an algorithm meant for solving this problem is $O(n^k)$ for an input of size n. The class *NP* consists of those problems that are "verifiable" in polynomial time. This means that if somehow a candidate solution is provided, it can be verified for correctness in polynomial time. *NP* problems can be solved in polynomial time on a nondeterministic computer. A problem in *P* is also in *NP*, since if the problem can be solved in polynomial time, then it can definitely be verified. Therefore, it can be written that $P \subseteq NP$. The open question is whether or not *P* is a proper subset of *NP*. A problem *P1* belongs to the class *NPC*, if it is an *NP* problem, i.e., can be solved in polynomial time on a nondeterministic computer and a problem already known to the *NP* complete is polynomially reducible to *P1*. The relationships among *P*, *NP* and *NPC* is shown in Fig. 2.11. Both *P* and *NPC* are wholly contained with *NP* and $P \cap NPC = \emptyset$.

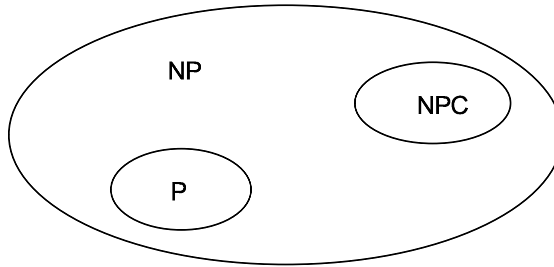


FIGURE 2.11
Relationships among *P*, *NP*, and *NPC*.

2.9 Technology-Aware Computer Aided IC Design Technique

Design challenges related to manufacturability and yield of integrated circuits are forcing the IC designers and EDA tool developers to understand the IC manufacturing. In particular, process and device related information needs to be incorporated into the IC design technique, in order to make the designs correct at the first attempt, and robust. The increasing gap between the feature size of the transistor and the wavelength of light used for the lithography technique leads to device-to-device variation of the transistor performances. The random fluctuations in the number of dopant atoms in the channel region of a MOS transistor critically affects the transistor performances. Circuit simulation tools presently lack the capability to predict the effect of several reliability stress effects, such as gate insulator time-dependent dielectric breakdown (TDDB), hot carrier injection (HCI), negative bias temperature instability (NBTI), and junction breakdown as a function of device terminal voltages. Therefore, for sub-90nm process technologies, designers need to include device/process technology aware simulation tools in the design procedure. The major problem of incorporating these into the design technique is the lack of proper models. Ideally such models need to be developed from actual silicon data through proper characterization. The alternative approach is to include technology computer-aided design (TCAD) techniques into the mainstream IC design technique. Since overall circuit performances are greatly affected by the increased effects of semiconductor process technology at the 65 nm node and below, TCAD can help IC designers to optimize performances and yield around the expected process variability. This offers significant savings in time and money to semiconductor vendors. TCAD simulations, which complement experimental silicon, provide a more comprehensive way to characterize technologies and optimize their performance, thereby reducing the number of re-spins and delivering high-quality products sooner. According to the

International Technology Roadmap for Semiconductors, product development costs can be reduced as much as 40 percent by using TCAD.

2.9.1 Introduction to TCAD

TCAD is an electronic design automation technique that models IC fabrication and device operation. TCAD is the art of abstracting IC electrical behavior by critical analysis and detailed understanding of process, device, and circuit simulation data. Utilizing TCAD technique it is possible to model and simulate all the steps involved from circuit simulation to device/circuit fabrication[162, 56]. The various modeling and simulation procedures that can be carried out through TCAD tools are described below:

1. *Lithography modeling.* Imaging of the mask through lithography machines, study of photoresist characteristics and processing.
2. *Frontend process modeling.* Simulation of the physical effects of various fabrication steps up to metallization.
3. *Compact device modeling and interconnect modeling.* TCAD allows to develop compact device models for the various active and passive components of the circuit, in addition to modeling of interconnects. These models are closely related to actual silicon results and therefore are very much accurate in predicting the various second order effects on circuit performances, which are becoming very significant in nano-scale domain. No other technique except TCAD can provide faithful results for these.
4. *Reliability Modeling.* Simulation of the reliability and related effects on process, device, and circuit level performances of integrated circuits.
5. *Equipment Modeling.* Simulation of the local influence of the equipment on each point of the wafer, especially in deposition, etc., etching, and chemical-mechanical polishing (CMP) processes.
6. *Modeling for design robustness, manufacturing and yield.* TCAD offers provision for accurate modeling and simulation of the impact of process variability and dopant fluctuation on IC performance and determines design specifications for manufacturability and yield of ICs.
7. *Package and Materials Modeling.* It is possible to model and simulate the various electrical, mechanical, and thermal effects of chip packages. The effects of the use of various materials on the physical and electrical properties of devices and integrated circuits can also be modeled and simulated.

The two major components of TCAD are: process simulation and device simulation. These are discussed below.

2.9.2 Process Simulation through TCAD

Process simulation refers to numerical simulation of the physical effects of IC fabrication steps up to metallization. The IC fabrication processing steps which can be simulated through process CAD are ion implantation, diffusion, oxidation, physical etching and deposition, lithography, stress formation and silicidation. The process CAD tool generates input data files for device simulator as realistically as possible based on microscopic information. The various inputs and outputs of a process simulation procedure are shown in Fig. 2.12. For MOSFET devices, threshold voltage and other device parameters are directly related to the distribution of the channel doping profile within the device structure. An accurate description of a channel doping profile can be generated using process simulation. Therefore, process simulation is critical to reproduce the doping distributions within the structure for accurate device simulation and compact model parameter extraction for circuit analysis.

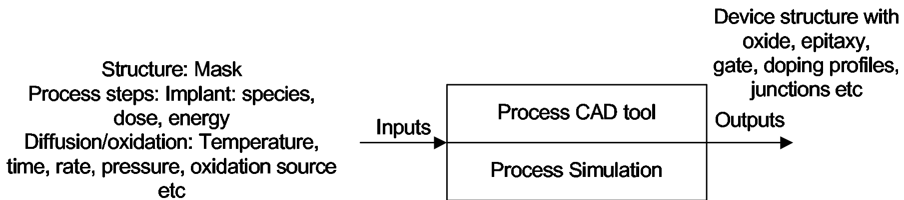


FIGURE 2.12
Inputs–outputs of a process simulation procedure.

2.9.3 Device Simulation through TCAD

Device simulation refers to numerical simulation of IC device operation. The motivation for the use of device simulation in IC design technique is the optimization of device performances for specific applications. For generation of an accurate device-level compact model like BSIM, device simulation is a necessary step. The device models used by device CAD tools range from the simple drift diffusion, which solves Poisson and continuity equations, to more complex and computationally challenging models such as the energy balance, which solves some higher moment simplification of the Boltzmann Transport Equation (BTE). Therefore, the ability of device CAD to accurately model today's device performance and predict tomorrow's device limitations is of utmost importance. The various inputs and outputs of a device simulation procedure are shown in Fig. 2.13

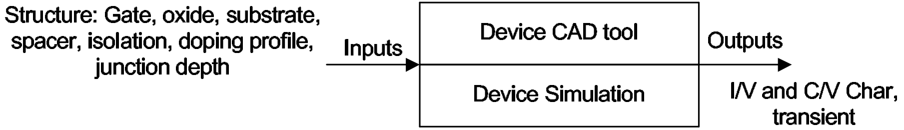


FIGURE 2.13
Inputs–outputs of a device simulation procedure.

2.9.4 Design for Manufacturability and Yield

With the transfer of a circuit layout to the fabrication process, a whole new set of challenges arises. With shrinking feature dimensions, parametric variability has become a dominant yield loss component. Parametric sensitivity to process variability is expected to worsen at 65 nm and below. Design engineers till today rely on two key pieces of information for designing chips: design rules and SPICE models. The design rules define the spacing between edges of polygons in the same mask layer or between mask layers such as poly line spacing or metal overlaps of contacts. These rules are set by lithography limits including feature resolution and layer alignment or by electrical limits such as leakage or breakdown of the electrically-active layers in the devices. The SPICE models on the other hand, predict the current or charge between the device terminals as a function of the applied voltage between the terminals. The SPICE models thus encapsulate the internal physics of the MOS transistors. Both the design rules and SPICE models are simplified representation of the process and device characteristics intended for transferring only as much manufacturing information to the design tools as is needed. With the scaling of process technology to nano-scale regime, some of the simplifying assumptions regarding the process and device characteristics are no longer valid and more exact physics of lithography/process effects and device effects have to be incorporated in the design technique. Thus the design for manufacturability and yield (DFMY) refers to the design and verification technique employed to ensure that the production silicon meets the performance objectives of the original circuit design. This therefore, reduces the design creativity gap.

SPICE models with various corner cases are traditionally the only link between manufacturing and design. The standard practice of studying the effects of process variations on circuit performances is to impose somewhat artificial statistical or systematic distributions on the SPICE model parameters. There are several limitations to this practice in the context of nano-scale CMOS technology. First, the SPICE model parameters in several cases deviate far from their underlying physics, as they often end up fitting parameters to silicon data. Second, the majority of the device characteristics such as threshold voltage and sub-threshold leakage current are correlated. Variations in major manufacturing steps such as halo implant and annealing temperature cause global changes in device properties. Such correlations are often are not taken care of in the SPICE models. Finally, except for very few, the SPICE param-

eters cannot be directly linked to any one specific process parameter. In other words, there is no common platform to communicate between manufacturing and design. However, successful DFMY requires that manufacturing information be accessible to design. By combining calibrated TCAD simulation results with compact model development procedure, it is possible to develop self-consistent process-dependent compact SPICE models, with process parameter variations as explicit variables. Such models then can be included in any circuit simulation procedure to evaluate the effects of process and device parameters on circuit performances. This is illustrated with a block diagram as shown in Fig.2.14. TCAD simulations of process, device and interconnect are used to create models, for which the device parameters and interconnect design variables are considered as inputs. The constructed models can be included in IC design tools such as static timing analyzer and circuit optimization procedure. In addition, layout related effects such as stress and well proximity effects can also be included for performance optimization.

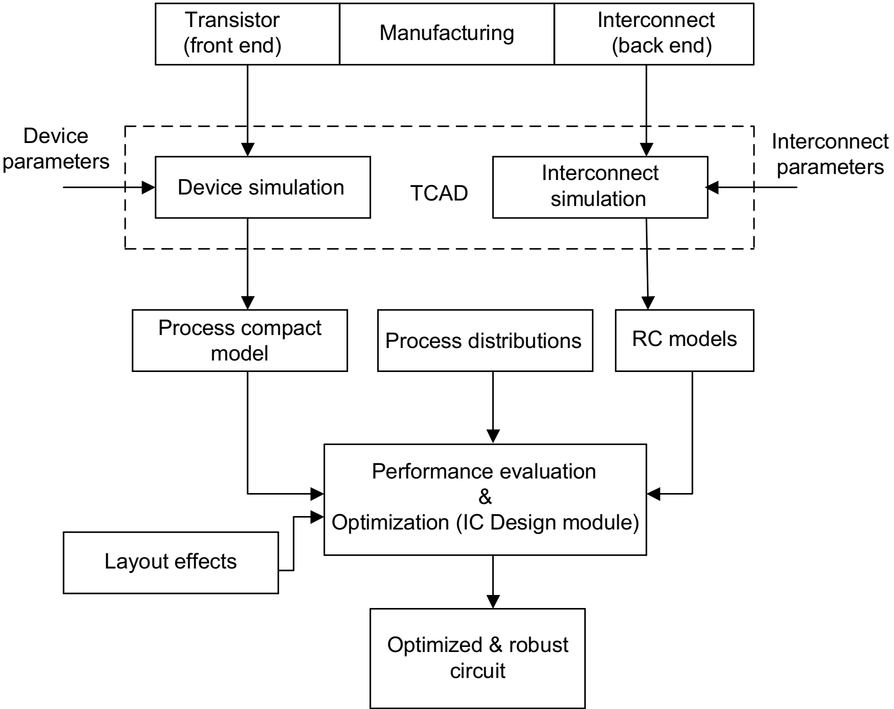


FIGURE 2.14
Outline of TCAD simulation-based DFMY technique.

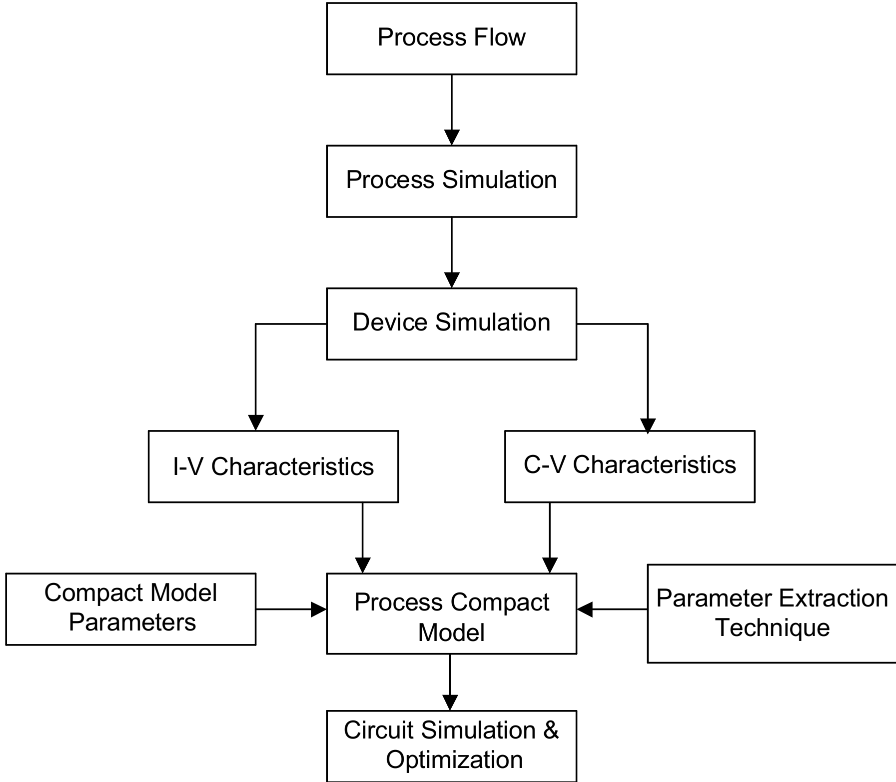


FIGURE 2.15
Process compact model generation using TCAD.

2.9.5 Process Compact Model Development through TCAD

A process compact model captures the functional relationship between the space of process variables and device characteristics through a response surface model. For, example, for the case of threshold voltage and low field mobility of a transistor, this can be written as

$$V_{TH0} = f_1(t_{ox}, \mathrm{Ch}_{dose}, \mathrm{Ha}_{dose}, \mathrm{Spike}_T) \tag{2.39}$$

$$U0 = f_2(t_{ox}, \mathrm{Ch}_{dose}, \mathrm{Ha}_{dose}, \mathrm{Spike}_T) \tag{2.40}$$

The process compact models (PCMs) are used by the design engineers for high-level design to ensure that the fabricated design meets the performance objectives of the original circuit design.

The technique for developing PCM is illustrated in Fig. 2.15. The user input is a description of the steps of the actual fabrication procedure. The process simulation generates a set of data with the structural information like device geometry, doping distribution, and so on. The output of the process

simulation procedure is fed to the device simulation procedure, which generates the I-V and C-V characteristics data set. The TCAD tools are calibrated with actual silicon results prior to these. The simulated characteristics are fed to a regression model development procedure. The model development procedure also considers a set of parameterized template consisting of the process parameters. The unknown parameter values are determined through a parameter extraction procedure.

2.9.5.1 Parameter Extraction Technique

Parameter extraction technique is an important step of compact model development procedure. The basic idea of parameter extraction technique is to find out the values of the unknown parameters of a parameterized model template such that the errors between the model generated results and available experimental/TCAD simulation data are minimized. Several different techniques are available; however, the appropriate technique depends on the model and on the applications of the model.

Both global optimization and local optimization techniques can be used for parameter extraction. Through global optimization technique it is possible to extract parameters such that the model generated results fit the experimental data in all the operating regions. However, in many cases, the extracted values are far from the actual physical values. On the other hand, in local optimization each parameter is extracted in a certain optimization region where the device behavior is dominant. Thus parameters optimized locally may not perfectly fit the experimental data in all the operating regions, but they closely resemble their actual physical value.

Also, there are two different strategies for model parameter extraction: the single device extraction strategy and the group extraction strategy. In the single device extraction strategy, experimental data from a single device are used to extract the model parameter. This strategy fits one device well, but usually does not fit other devices with different geometries. On the other hand, in the group device extraction strategy, parameters are extracted from experimental data of multiple devices having different geometries. This strategy, therefore, may not fit one device extremely well, but can fit many devices with different geometries reasonably well.

A simple parameter extraction technique which is used widely and even in BSIM parameter extraction procedure is to combine Newton-Raphson iteration and a linear-least square fit with either one, two or three variables. The model equation is arranged in a form suitable for Newton-Raphson's iteration as shown below

$$\begin{aligned}\phi_{exp}(\alpha_1,\alpha_2,\alpha_3)-\phi_{model}(\alpha_1^i,\alpha_2^i,\alpha_3^i) &= \frac{\partial\phi_{model}}{\partial\alpha_1}\Delta\alpha_1^i+\frac{\partial\phi_{model}}{\partial\alpha_2}\Delta\alpha_2^i\\ &+ \frac{\partial\phi_{model}}{\partial\alpha_3}\Delta\alpha_3^i \qquad (2.41)\end{aligned}$$

where $\phi_{model}()$ is the parameterized model template, the values of whose

unknown coefficients $\alpha_1, \alpha_2, \alpha_3$ are to be determined, $\alpha_1^i, \alpha_2^i, \alpha_3^i$ represent the parameter values after i^{th} iteration. (2.41) can be written in a fashion $y = a + bx_1 + cx_2$ by dividing (2.41) by $\partial\phi_{model}/\partial\alpha_1$. Here y, b and c are known, through fitting technique, a, x_1 and x_2 are determined. It may be noted if it is not possible to calculate the derivatives analytically, then these need to be calculated numerically, e.g., through the central difference technique. The parameter values for the $(i+1)^{th}$ iteration are given by

$$\alpha_m^{i+1} = \alpha_m^i + \Delta\alpha_i^m \quad m = 1, 2, 3 \tag{2.42}$$

The iteration continues until the increments are smaller than some predetermined values. The flow chart of the optimization technique for parameter extraction is shown in Fig.2.16.

2.9.6 Design Techniques for Nano-Scale Analog ICs

The existing design techniques in general do not provide any platform for interaction between the IC designers and TCAD designers. This leads to the causes of several design failures at the first attempt and thereby increases the design creativity gap. With the increased complexity of nano-scale analog ICs, the high-level design procedure for nano-scale ICs needs to incorporate the exact device and process related knowledge within the design flow. This necessitates a paradigm shift in the design technique. TCAD flow has to be integrated with IC design flow. This is illustrated through the flow diagram shown in Fig. 2.17. Process compact models form the bridge between the two flows. Knowledge-based CAD technique, which has been discussed in depth in the Chapter 1 has to be calibrated with TCAD results (which in turn are assumed to be calibrated with characterized silicon results) and make a compact technology aware CAD framework for high-level design of nano-scale analog ICs.

2.10 Commercial Design Tools

This section provides a brief introduction to some of the important commercial CAD/EDA tools available for analog IC design and TCAD purpose. The readers are suggested to carefully browse the websites of the individual vendors for more detailed and updated information.

2.10.1 IC Design

The generic flow of the traditional IC design procedure is shown in Fig.2.18. The procedure starts with a set of specifications. The schematics of the transistor-based circuit topology are drawn via a schematic editor. The circuit

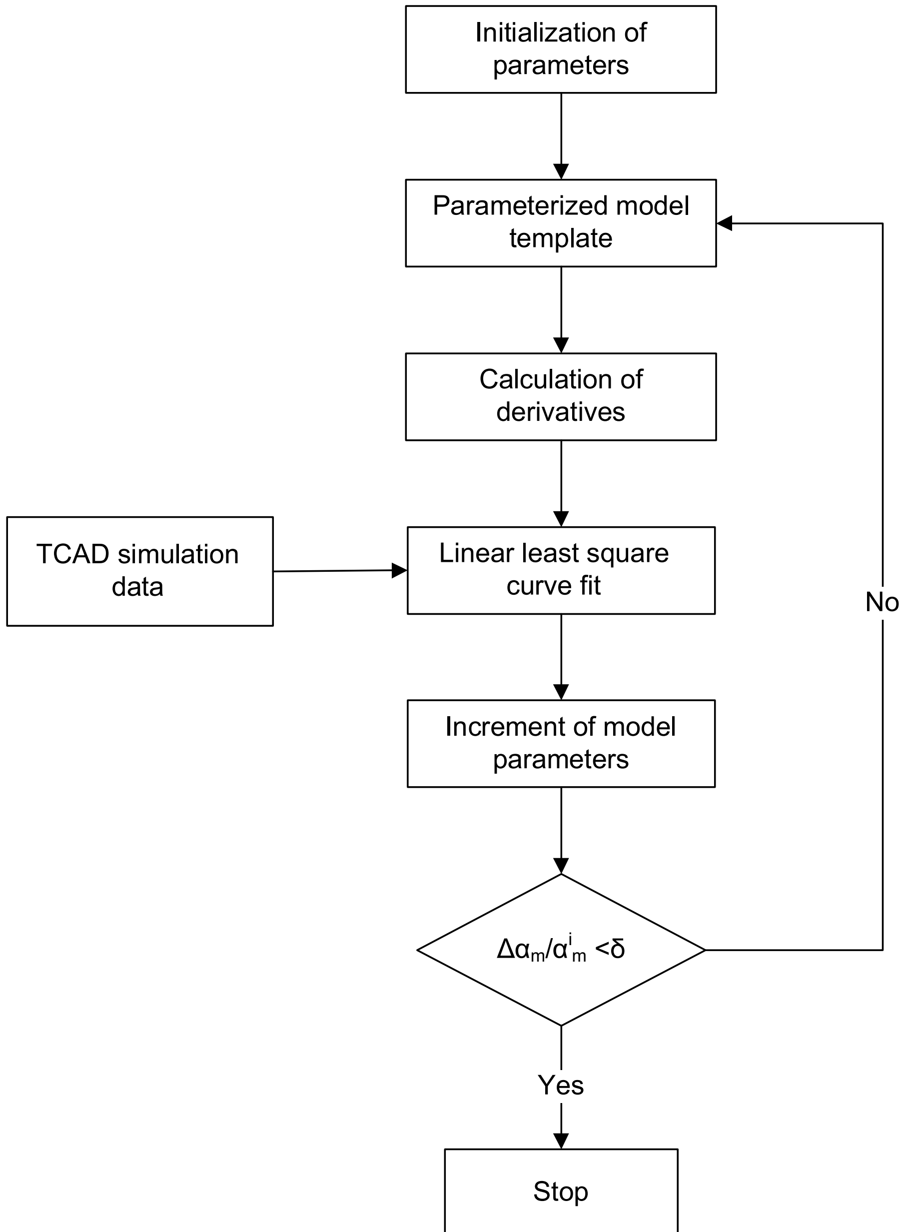


FIGURE 2.16
Optimization flow for parameter extraction.

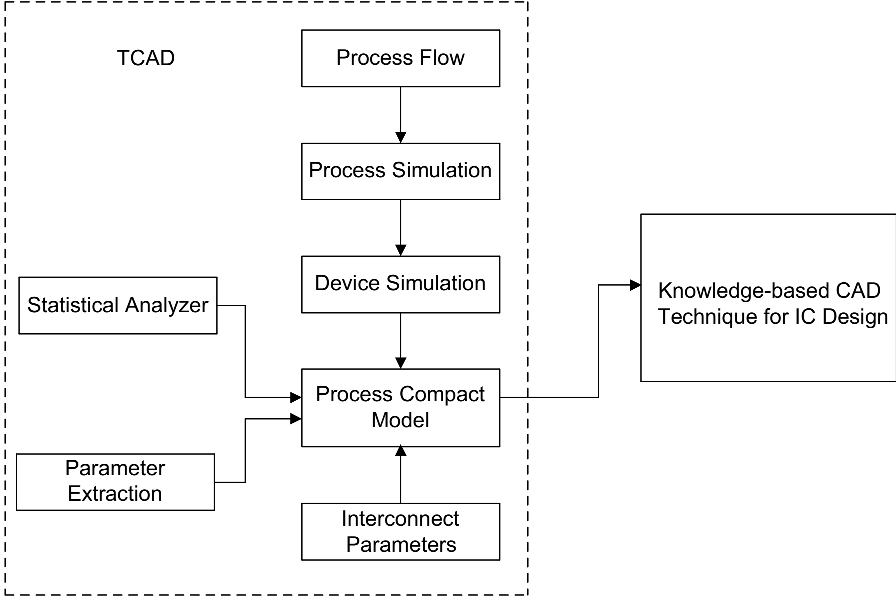


FIGURE 2.17
Design technique for nano-scale analog ICs.

is then simulated through some SPICE package. This is referred to as the pre-layout simulation. The design tasks carried up to this stage is are collectively known as the front-end design.

Subsequently the layout of the circuit is drawn and verified for any design rule violations. This is referred to as the design rule check (DRC) action. The various parasitics involved with the interconnects are extracted through a RC extraction tool. A SPICE-compatible netlist is extracted from the layout. The layout is subsequently verified with the netlist. This is referred to as the Layout versus Schematic (LVS) comparison. The extracted netlist is simulated with the SPICE tool. This is referred to as the post-layout simulation. With the satisfaction of the desired specifications, the circuit is ready for tape-out. The design tasks performed for physical design and verification are collectively known as the back-end design.

The generic flow shown in Fig.2.18 is followed by almost all the EDA tools. The commonly used EDA tool suites used for custom analog IC design are from the vendors like Cadence®, Synopsys®, Mentor Graphics® and Tanner EDA®. The tool suite available from these vendors are mentioned below.

2.10.1.1 Cadence® Virtuoso Analog Design Environment

Cadence Virtuoso Analog Design Environment allows different applications and tools to integrate into a single framework thus allowing all the stages of

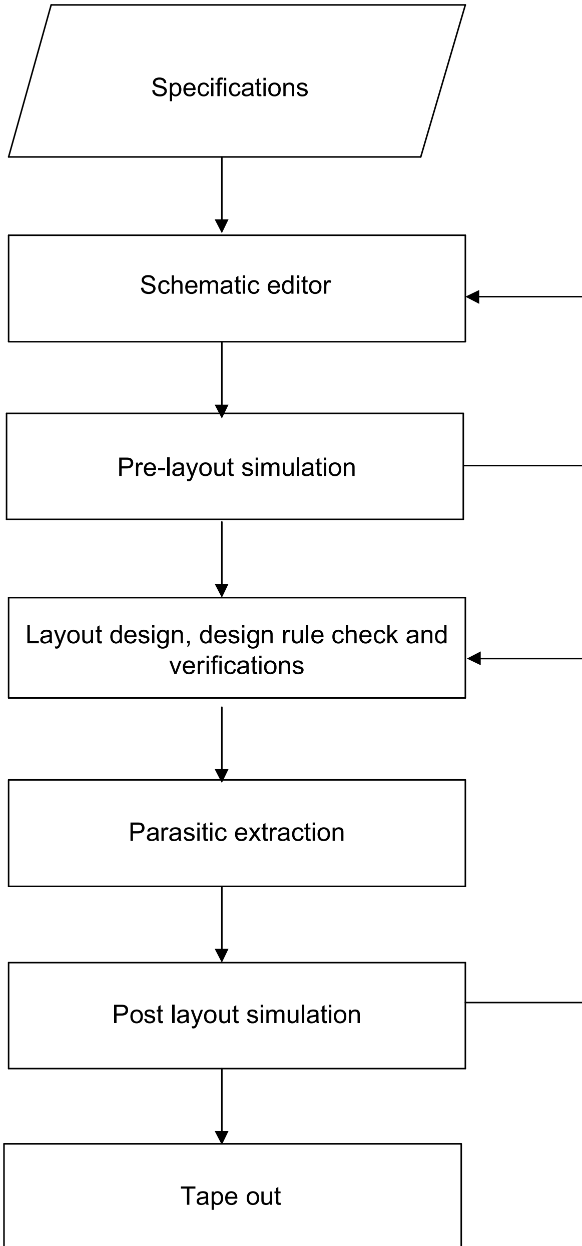


FIGURE 2.18
General design flow for custom analog ICs.

IC design and verification to be performed from a single environment. The basic flow is similar to that shown in Fig.2.18. Starting from architectural description using hardware description languages to final structural schematic implementations at the transistor level, the Virtuoso Schematic Editor family helps the designers to implement each stage in their designs, while capturing and ensuring consistency of design intent with constraints. The Virtuoso Schematic Editor product family is integrated with the Virtuoso Analog Design Environment.

Virtuoso Spectre® Circuit Simulator provides a high-precision SPICE simulation of pre- and post-layout analog/RF designs with a comprehensive set of analyses for faster convergence. This simulator is also integrated with the Analog Design Environment. The simulator uses silicon-accurate device models that are universally supported by all foundry process design kits (PDKs).

Once the circuit specifications are fulfilled in simulation, the circuit layout is created using the Virtuoso Layout Editor. Virtuoso layout suite supports custom analog, digital, RF, and mixed-signal designs at the device, cell, block, and chip level. Virtuoso Layout Suite L enables users to open multiple cells or blocks in a single editing session, or to open different views of the same design, ensuring consistency in complex designs.

For the purpose of DRC and LVS, the Assura® Physical Verification tool is used. Assura® Physical Verification uses hierarchical processing and multiprocessor techniques to increase performance and capacity.

Finally, a netlist including all layout parasitics is extracted, and a final simulation of this netlist is made. This is called a PostLayout simulation, and is performed with the same Cadence simulation tools. The parasitics may be extracted through QRC Extraction. Once the layout functionality is verified, the final layout is converted to a certain standard file format depending on the foundry (GDSII, CIF, etc.,) using the Cadence conversion tools.

2.10.1.2 Synopsys Galaxy Custom Design

The tool used for schematic entry is the Galaxy Custom Designer SE. As with all custom designer tools, the schematic editing tasks are accomplished with few clicks and quick menu access. HSPICE® and NanoSim® are the two simulation packages available from Synopsy® for circuit simulation. HSPICE® offers foundry-certified MOS device models with state-of-the-art simulation and analysis algorithms and is therefore used by many designers for accurate circuit simulation. NanoSim is an advanced circuit simulator for analog, high performance digital and mixed-signal circuit simulation. It provides a combination of timing and power diagnostic function needed to efficiently analyze today's nanometer IC design. Galaxy custom designer LE is the layout entry and editing tool offered by Synopsys®. Since this tool is integrated with the full Custom Designer system, it provides transistor-level layout and editing capabilities in a unified platform. Galaxy custom designer schematic-driven layout (SDL) offers the advantage of drawing layouts directly from the

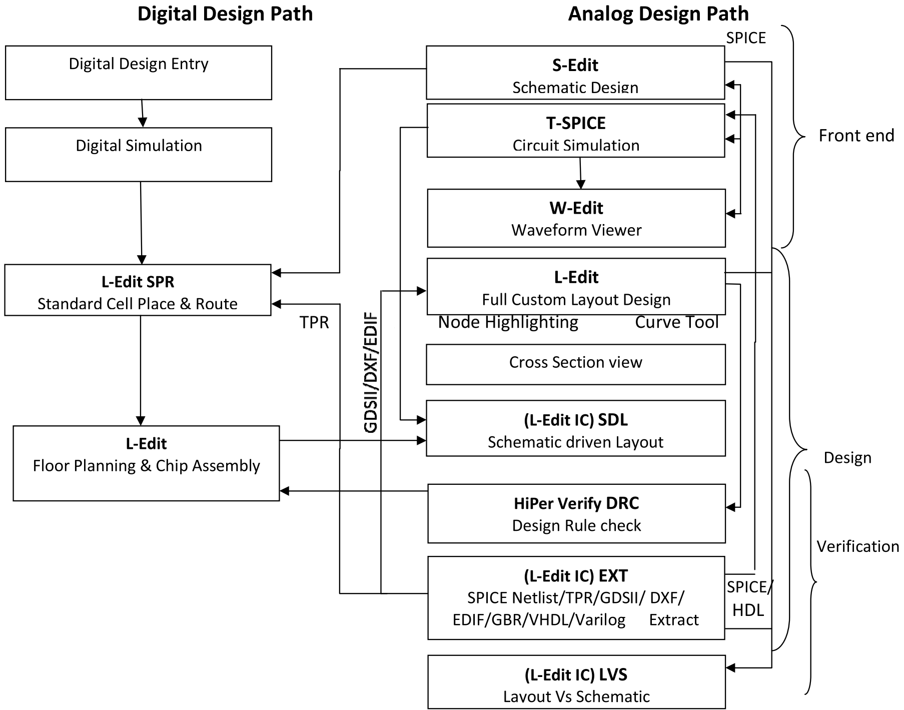


FIGURE 2.19
IC design flow using Tanner EDA tool suite.

schematics. This therefore, considerably boost the designers productivity and layout throughput. However, the resultant layout no doubt needs considerable attention of the designers for fine tuning. Intelligent layout automation in Custom Designer LE results in DRC and LVS correct cell and macro layouts. Star-RCXT® is the tool used for parasitic extraction. PrimeRail® is a full-chip power network analysis solution for low power and high-performance designs at 90-nm and below. It offers gate-level and transistor level static and dynamic voltage drop and eletromigration analysis during implementation and sign-off.

2.10.1.3 Tanner EDA HiPer Silicon®

HiPer Silicon gives the designer a complete analog design flow from schematic capture, circuit simulation and waveform probing to physical layout and verification, which is ideal for analog, mixed-signal and high frequency IC design. The tool used for schematic capture is the S-Edit®. This tool is tightly integrated with SPICE simulation and therefore allows the designers the flexibility to view the operating point results directly on the schematic and perform waveform cross-probing to view node voltages and device terminal currents or

charges. This tool allows for checking of common errors such as undriven nets, unconnected pins and nets driven by multiple outputs so that errors can be caught early before running simulation. T-SPICE® is the simulation package offered by Tanner. T-SPICE offers HSPICE® and PSPICE® compatible syntax and supports the latest industry models such as PSP, BSIM4.4 etc.,. It also supports Verilog-A for analog behavioral modeling. L-Edit® is the hierarchical physical layout editor offered by Tanner. L-Edit supports parameterized cells allowing the designers to create automatic custom layout generators or use DevGen to easily setup layout generators for commonly used devices such as MOS transistors, resistors or capacitors. The tool supports interactive DRC checking. Schematic drive layout facility is also supported by Tanner and is integrated with the L-Edit layout editor. For physical verification and parasitic extraction, the tool offered by Tanner is the HiPer Verify® and PX®. HiPer Verify® is a comprehensive solution for analog/mixed signal IC design rule checking and netlist extraction. On the other HiPer PX® is a high performance parasitic extraction tool that is integrated with Tanner's L-Edit layout editor for easy and rapid extraction of parasitics. The custom design flow using the Tanner tool suite is shown in Fig. 2.19.

2.10.1.4 Mentor Graphics Pyxis® Suite

The analog/custom IC design flow using the Mentor Graphics tool suite is shown in Fig. 2.20. The IC Nanometer Design package provides a complete environment for the design, capture, layout and verification of analog, digital and mixed-signal integrated circuits. The Pyxis suite of IC design tools includes tools for

- Schematic capture, netlisting, simulation setup and results viewing
- Physical layout
- Editing, schematic-driven layout, and top-level floorplanning and routing

Eldo® and Eldo RF are the SPICE simulation packages offered by Mentor Graphics for simulation purpose. The Calibre® and Calibre xRC are the tools for physical verification and parasitic extraction.

2.10.2 TCAD

The two major vendors offering TCAD tool suite are Silvaco® and Synopsys®.

2.10.2.1 Silvaco Tool Suite

The journey of the modern commercial TCAD tools started with the development of two famous general-purpose simulation software programs. SUPREM (Stanford University Process Engineering Models) and PISCES (Poisson and Continuity Equation Solver) came as an outgrowth of the research done at Stanford University. SUPREM3 is a one-dimensional process simulator,

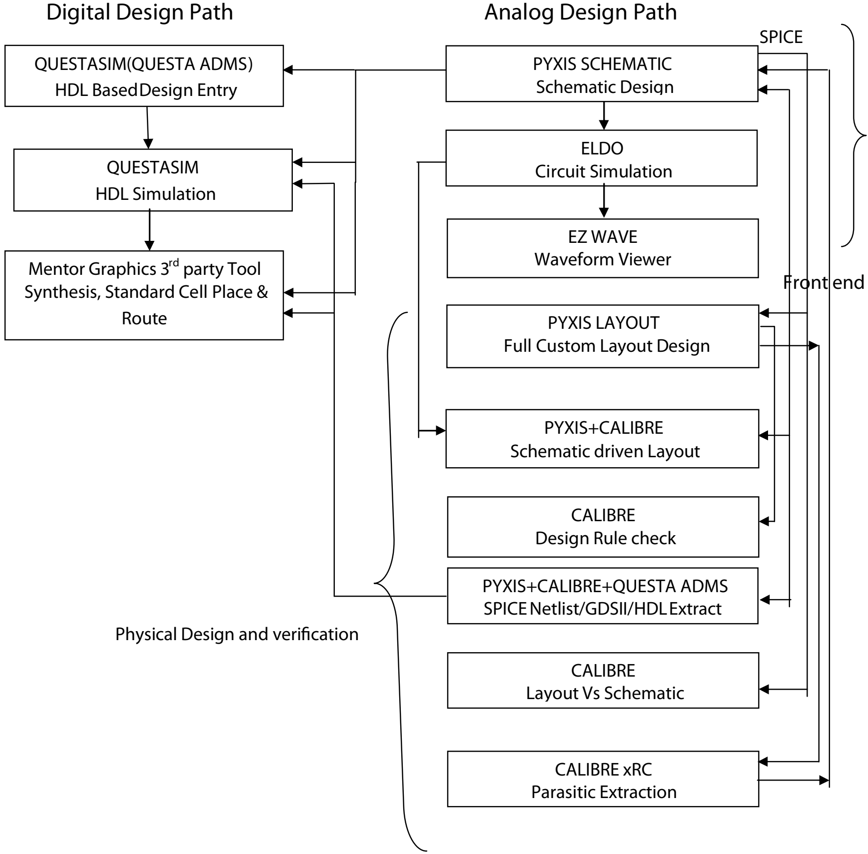


FIGURE 2.20
IC design flow using mentor graphics EDA tool suite.

while SUPREM4 is a two-dimensional process simulator. PISCES is the two-dimensional device simulator. ATHENA® and ATLAS® are the commercial equivalent alternatives of these programs, as Silvaco later licensed these programs from Stanford University. The Athena framework supports some tools like S-SUPREM4® (2D process simulator), ELITE® (physical etching and deposition simulator, OPTILITH® (2D optical lithography simulator), MC DEPOSIT/Etch (2D Monte Carlo deposition and etch simulator), MC Implant (Advanced Monte Carlo Implantation Simulator) S-SUPREM3 (1D process simulator). The ATLAS® framework supports S-PISCES (2D device simulator), Device 3D (3D device simulator), BLAZE 2D (2D device simulators for advanced materials), Mixed Mode (2D circuit simulator for advanced devices), Quantum 2D (simulation models for quantum confinement effects, MC Device 2D (Monte Carlo device simulator), NOISE 2D (2D small signal

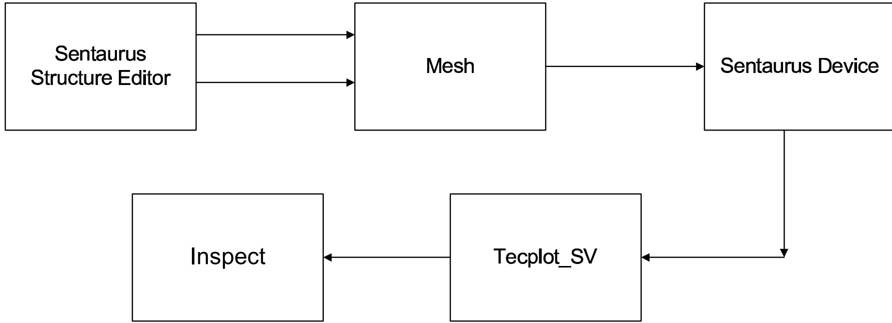


FIGURE 2.21
Device simulation tools offered by Synopsys TCAD suite.

noise simulator). In addition, there are tools for device schematic diagrams and waveform viewing. These are DevEdit 2d and DevEdit 3D for structure drawing and Tonyplot for waveform viewing.

There are several other tools like SPAYN for statistical parameters analysis and UTMOST-III and UTMOST-IV for model parameter extraction including that of BSIM.

2.10.2.2 Synopsys Device Simulation Tool Suite

Synopsys TCAD offers a comprehensive suite of products that includes industry leading process and device simulation tools, as well as a powerful GUI-driven simulation environment for managing the simulation tasks and analyzing the simulation results. For device simulation purposes the input device structure typically comes from process simulation steps using tools like Sentaurus Process or Taurus TSUPREM-4. Some important tools required for device simulation purposes are listed below (see Fig. 2.21).

1. Sentaurus Structure Editor (SSE): This tool is used for device structure creation. The structures are generated or edited interactively using the graphical user interface (GUI). In addition, it can be used for a 3D process emulator based on CAD technology.
2. Mesh: This engine helps to mesh the structure created with SSE. The tool produces finite-element meshes for use in semiconductor device simulation.
3. Sentaurus Device: This tool is used to simulate the electrical characteristics of the device.
4. Tecplot SV: This tool has extensive 2D and 3D capabilities and is used for scientific visualization and plotting of simulated data.
5. Inspect: The electrical characteristics are plotted with the help of

this tool. It is basically a curve display and analysis program. The curves are specified at discrete points.

2.11 Summary and Conclusion

This chapter introduced the two important components of the knowledge-based CAD technique-(i) high-level models and (ii) optimization techniques. A comprehensive overview of the three important techniques, namely manual abstraction technique, model order reduction technique and symbolic analysis technique for construction of behavioral models has been provided. The optimization procedures required for analog CAD design including problem formulation and optimality criteria have been described. The concepts of local optima and global optima have been introduced. The computational complexity of algorithms including asymptotic notations and complexity classes have been described. This chapter also introduces the technology computer-aided design. Thereafter, it has been suggested that a combination of the TCAD and knowledge-based CAD procedures is essential to cope with the recent challenges of the nano-scale analog IC design. Finally, a brief description of the commercial CAD/EDA tools has been presented.

3

Modeling of Scaled MOS Transistor for VLSI Circuit Simulation

3.1 Introduction

The modern very large scale integrated (VLSI) circuits consist of several billions of transistors with the MOS transistor forming the basic building block. The successful design of a complete circuit requires extensive circuit simulation where an accurate and faithful mathematical description of the MOS transistor, referred to as a compact device model is an essential prerequisite. The device models act as the link between the physical world (technology, manufacturing, ...) and the design world (circuit simulation, timing analysis) of the semiconductor industry [1]. The models are incorporated within the SPICE simulator for circuit simulation purposes. In the conventional design practice, the circuit designers remain unaware of the details of the compact device models and rely completely on the SPICE simulator. Consequently, for most of the design failures they do not have any option other than to blame the models. However, in the present scenario of tremendous market pressure and complex nano-scale process technology, the IC designers cannot afford to remain unaware of compact models. They need to know the fundamentals of the characteristic properties of scaled MOS transistors and basic issues related to the modeling of the same. This will equip them in anticipating several possible design failures after fabrication of the designs. This chapter attempts to present the essential features of compact device models for the scaled MOS transistor at an introductory level. The compact model developed by the University of Berkeley, which is coined as the BSIM model, has been considered while pointing out the commercial use of the various modeling aspects. It may be noted that this chapter is in no way a replacement for the official documentation for the BSIM compact model.

3.2 Device Modeling

The use of device models in the VLSI simulation process is illustrated in Fig. 3.1. It is observed that the VLSI circuit simulation process essentially consists of simulation of the circuit netlist through a SPICE simulation engine utilizing a set of device models. Various kinds of device models corresponding to all possible electrical devices are available to the SPICE simulator. These are invoked when required during the simulation procedure. The accuracy of the circuit simulation results depends on the accuracy of the device models that

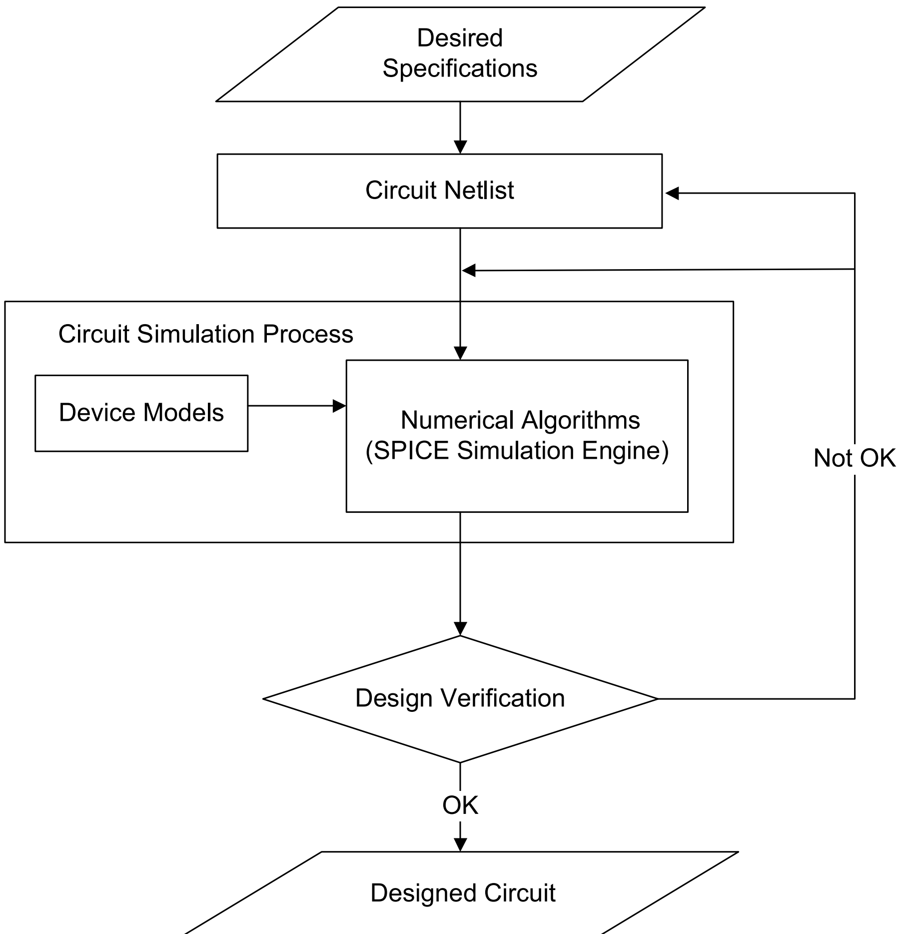


FIGURE 3.1
Use of device models in VLSI circuit simulation.

are being used by the SPICE simulator, since the numerical algorithms used by the engine are fairly mature.

3.2.1 Categories of Device Models

There are three categories of device models: (i) numerical models, (ii) table lookup models, and (iii) analytical (or compact) models. The numerical models are based upon numerical solutions of carrier transport equations, device geometry and doping profile related equations. This provides very accurate estimation of the device characteristics. However, these are computationally intensive and hence are not suitable for simulation of large circuits. The technology computer-aided design (TCAD) simulation of device characteristics belongs to this type of modeling. This may be used for exploration of novel device structures, conventional devices with new materials and associated performances. The look-up table approach, on the other hand, uses measured device current and capacitances (and in some cases small signal parameters) as functions of bias voltages and device sizes for characterizing the device performances which are subsequently used for circuit simulation. The look-up table approach is especially useful when good physical models are not available, e.g., for advanced device structures such as double gate MOS transistors, FinFETs, channel engineered structures etc. In addition, sometimes this approach is used for fast circuit simulation. The third category is the analytical or compact model. These are based on device physics. The compact device models for MOS transistor is one of the subjects of discussion in the present chapter and is introduced in the next section.

3.3 Compact Models

A compact MOSFET model is a set of mathematical equations whose parameters are used as inputs to a SPICE-like circuit simulator [1]. The compact model is expected to able to reproduce faithfully the transistor characteristics for various dimensions, range of temperature and process variations. In addition, the description must be valid under a variety of operating conditions. The compact model equations are often functions of transistor dimensions and thus are scalable. However, these are often long and complex, in order to describe the device characteristics accurately in all the operating regimes. Fitting parameters are introduced in many occasions to improve the accuracy of the model. The commercial circuit simulators use compact device models for circuit simulation purposes.

3.3.1 Commercial Compact Models

The commercial compact models are classified into three categories: (i) threshold voltage V_T based models, (ii) Inversion charge Q_n based models and (iii) surface potential ψ_s based models.

The journey of the V_T based models start with the MOS1 model which is a very simple MOSFET model, suitable for long-channel and uniformly doped MOS transistors. Because of its simplicity this model is still used by the circuit designers for hand calculations and preliminary circuit simulation. MOS2 and MOS3 models are based upon the MOS1 model with the introduction of many empirical parameters to account for the short-channel effects. However, these models fail to incorporate the exact physics of the scaled MOS transistor within the model framework. BSIM1 (Berkeley Short-Channel IGFET Model 1) and subsequently BSIM2 were developed for $1\mu m$ technology. These models incorporated some improved physics of the scaled MOS transistors. However, these also used several empirical parameters for each short-channel phenomenon in order to match with experimental silicon results with reasonable accuracy. One set of parameters cannot be used for a wide range of device dimensions. Statistical modeling could not be performed with these models. BSIM3, which is the third generation BSIM model, brought a major breakthrough in compact models. It was developed from a coherent quasi-two-dimensional analysis of a MOS transistor; both computational robustness and physical basis form the guiding philosophy of the development of BSIM3 [30]. The minimum channel length of the MOS transistor which can be supported through BSIM3v3 is $0.15\mu m$. BSIM4 is the last model so far from Berkeley that belongs to the V_T based category. For simulation of transistors with channel length of $0.13\mu m$ and below, BSIM4 model has to be used.

The inversion charge based models are based upon the calculation of the drain current in terms of the inversion charge density at the source and drain ends of the channel. This category of compact models is based upon the physics of MOS transistors with minimum dependence on empirical fitting parameters. The parameter extraction procedure is therefore, relatively simple. The compact models such as EKV (Enz-Krummenacher-Vittoz), ACM (Advanced Compact Model) belong to this category.

The surface potential based models are based upon the calculation of drain current in terms of the surface potential at the source and drain ends of the channel. Commercial compact models such as HiSIM (Hiroshima University STARC IGFET Model) and PSP (Pennsylvania State University Surface Potential) models belong to this category.

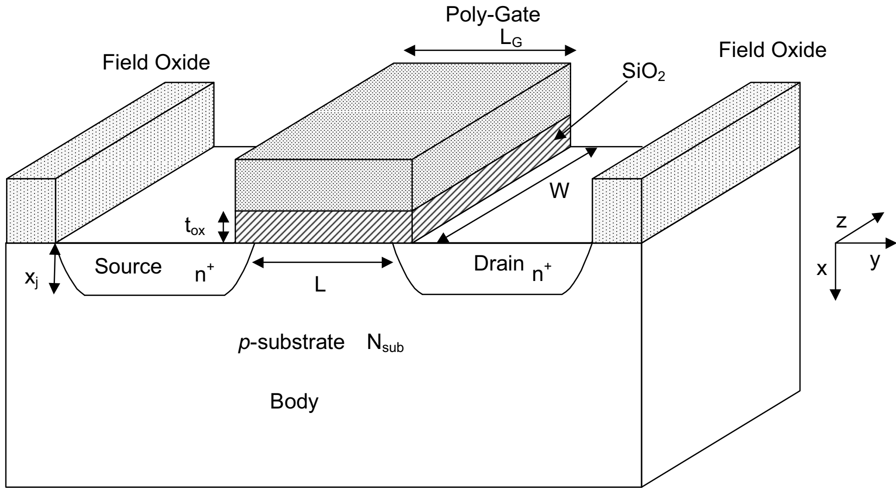


FIGURE 3.2
Cross-sectional view of n-type long-channel MOS transistor.

3.4 Long-Channel MOS Transistor

The schematic diagram of the cross section of a Metal-Oxide Semiconductor (MOS) transistor is shown in Fig.3.2. The MOS transistor is a four terminal device. The four terminals are called Drain (D), Gate (G), Source (S) and Body (B). For an n-channel MOS transistor, the substrate/body is of p-type semiconductor into which two n^+ regions, the source and the drain, are formed usually by ion implantation. The $\mathrm{SiO_2}$ gate dielectric is formed by thermal oxidation of Si for a high quality $\mathrm{SiO_2} - \mathrm{Si}$ interface. The gate contact over the oxide layer is heavily doped polysilicon or a combination of silicide and polysilicon. The length of the channel between the source and the drain regions is called the channel length L. The width of the channel, in the direction normal to the channel length is called as the channel width W. The other device parameters are the oxide layer thickness t_{ox}, source-drain junction depth x_j and the substrate concentration N_{sub}.

The two types of MOS transistors that are mostly used in VLSI circuits are n-channel enhancement mode MOS transistor (for which the conducting carriers are the electrons) and p-channel enhancement mode MOS transistor (for which the conducting carriers are the holes). For the former, sufficiently large positive gate voltage is required to create conducting channel between the source and the drain regions. On the other hand, for the latter, sufficiently large negative gate voltage is required for the same. The following discussion is concentrated on the n-channel enhancement mode MOS transistors. The source terminal is the conventional voltage reference; the reason is attributed

to digital circuits. The presented theories, however, are also applicable for p-channel MOS transistor with suitable changes in the signs of appropriate bias parameters.

3.5 Threshold Voltage Model for Long-Channel Transistor with Uniform Doping

The gate-to-source voltage V_{GS} required to produce an inversion layer in the channel (i.e., the portion of the semiconductor between the source and the drain regions and underneath the interface becomes of opposite polarity compared to that in the substrate) is called the threshold voltage V_T [189]. When the applied gate-to-source voltage V_{GS} is greater than the threshold voltage, the transistor operates in strong inversion region and when V_{GS} is lower than the threshold voltage, the transistor operates in the weak inversion region. The threshold voltage for a long-channel MOS transistor is calculated as follows [189].

This voltage consists of several components:

1. work function difference Φ_{MS} between the gate metal and silicon,
2. voltage $-Q_b/C_{ox}$ across the oxide layer to sustain the bulk depletion charge Q_b. It is to be noted that Q_b is negative for n-channel MOS transistors and positive for p-channel transistors and C_{ox} is the gate oxide capacitance per unit area,
3. voltage equal to $(\psi_s = 2\Phi_F)$ (surface potential at strong inversion) to induce inversion charge sheet in the channel region. It is to be noted that this quantity is positive for n-channel MOS transistors and negative for p-channel MOS transistors,
4. positive charge density Q_i exists in the oxide at the silicon interface which needs to be compensated by a gate voltage equal to $-Q_i/C_{ox}$.

With this, the threshold voltage is thus written as

$$V_T = \Phi_{MS} + 2\Phi_F - \frac{Q_b}{C_{ox}} - \frac{Q_i}{C_{ox}} \tag{3.1}$$

The quantity $\Phi_{MS} - Q_i/C_{ox}$ is known as the flat band voltage V_{FB} so that the threshold voltage is more commonly written as

$$V_T = V_{FB} + 2\Phi_F - \frac{Q_b}{C_{ox}} \tag{3.2}$$

In (3.2), the bulk depletion charge Q_b is defined as

$$Q_b = \sqrt{2q\epsilon_{Si}N_A\psi_s} \tag{3.3}$$

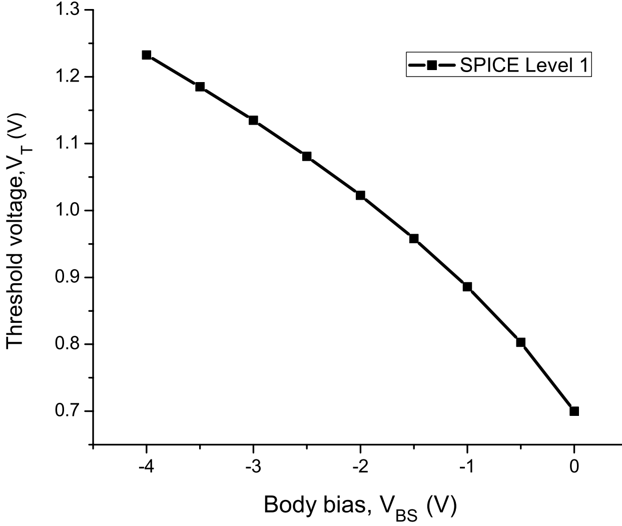


FIGURE 3.3
Threshold voltage vs. substrate bias for a long-channel NMOS transistor.

where $\psi_s = 2\Phi_F = 2\frac{kT}{q}\ln\left(\frac{N_A}{n_i}\right)$ is the surface potential at strong inversion and $N_A = N_{sub}$ is the uniform substrate concentration.

3.5.1 Body Effect

The body of a MOS transistor is usually connected to a constant power supply voltage or ac ground. However, the source voltage often changes, so that the body-to-source voltage V_{BS} is often non-zero. When multiple transistors are connected in series in a circuit, they share a common body, i.e., the silicon substrate but their sources do not have the same voltage. With a reverse bias V_{BS} applied between the substrate/body and source (V_{BS} is negative for n-channel transistor and V_{BS} is positive for p-channel transistor), the depletion region is widened. The bulk depletion charge Q_b is enhanced and the threshold voltage required to achieve inversion is increased to accommodate the larger Q_b. This is known as the body effect. Thus the body/substrate plays the role of second gate. The enhanced depletion charge density is given as

$$Q_b = \sqrt{2q\epsilon_{Si}N_A(2\Phi_F - V_{BS})} \tag{3.4}$$

Substituting this (3.2), we get

$$\begin{aligned} V_T &= V_{FB} + 2\Phi_F - \frac{\sqrt{2q\epsilon_{Si}N_A(2\Phi_F - V_{BS})}}{C_{ox}} \\ V_T &= V_{T0} + \gamma\left(\sqrt{2\Phi_F - V_{BS}} - \sqrt{2\Phi_F}\right) \end{aligned} \tag{3.5}$$

In (3.5) γ is known as the body-effect parameter and V_{T0} is referred to as the zero substrate bias long-channel threshold voltage of a MOS transistor. The parameter γ is defined as

$$\gamma = \frac{\sqrt{2q\epsilon_{Si}N_A}}{C_{ox}} \tag{3.6}$$

It may be noted that the body-effect parameter γ is negative for n-channel MOS transistors and is positive for p-channel MOS transistors. The gate oxide capacitance per unit area is defined as

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} \tag{3.7}$$

where ϵ_{ox} and t_{ox} are the permittivity and the thickness of the oxide respectively. A typical value of γ is $0.5V^{1/2}$ and $C_{ox} = 3.45fF/\mu m^2$ for $t_{ox} = 10nm$.

The variation of the threshold voltage of an n-channel MOS transistor as obtained from SPICE simulation is shown in Fig. 3.3. The body-effect parameter γ as obtained from the graph is $0.344V^{1/2}$. With $KP = 110 \times 10^{-6} A/V^2, 2\Phi_F = 0.7V, \mu_0 = 0.0634m^2/V - s$, the value of γ obtained theoretically is $0.336V^{1/2}$. It is therefore, observed that V_T is a sublinear function of V_{BS}.

The substrate bias sensitivity is defined as

$$\frac{dV_T}{dV_{BS}} = -\frac{\gamma}{2\sqrt{2\Phi_F - V_{BS}}} \tag{3.8}$$

At zero substrate bias, substituting (3.6), the substrate bias sensitivity is written as

$$\frac{dV_T}{dV_{BS}} = -\frac{1}{C_{ox}}\frac{\epsilon_{Si}}{W_{dm}} = -\frac{C_{dm}}{C_{ox}} = -(m-1) \tag{3.9}$$

In (3.9), W_{dm} is the width of the depletion region, C_{dm} is the depletion capacitance and m is referred to as the body-effect coefficient. The depletion depth is defined as

$$W_{dm} = \sqrt{\frac{4\epsilon_{Si}\Phi_F}{qN_A}} \tag{3.10}$$

This model is derived based on the assumption that the transistor is of long-channel length and large width and the substrate doping concentration N_A is uniform.

3.6 SPICE Level 1 Drain Current Model

In strong inversion, the motion of the channel electrons is primarily due to drift motion. The drain-to-source current is

$$I_{DS} = WQ_n(y)v_d(y) \tag{3.11}$$

where $-Q_n$ is the inversion charge per unit area at a position y in the channel and $v_d(y)$ is the drift velocity of carriers at that position. The chosen coordinate system is shown in Fig. 3.2, $y = 0$ means source end and $y = L$ means the drain end. With low drain bias, the drift velocity is given by

$$v_d(y) = \mu_{ns}\xi(y) \tag{3.12}$$

where μ_{ns} is the mobility of the electrons in the channel and $\xi(y) = -\dfrac{\partial V_{CS}(y)}{\partial y}$. Here $V_{CS}(y)$ is the quasi-Fermi potential of the electrons in the channel, also known as the channel potential defined with respect to source. With these, substituting (3.12) in (3.11), the drain current is given as

$$I_{DS} = -WQ_n(y)\mu_n\frac{\partial V(y)}{\partial y} \tag{3.13}$$

Before proceeding further, several assumptions are made [132]. These are

1. The mobility μ_{ns} of the carriers in the channel region is constant.
2. The variation of the electric field in the y-direction (along the channel) is much less than the corresponding variation in the x-direction (perpendicular to the channel). This is referred to as the gradual channel approximation [192, 189], as a result of which the inversion charge density is controlled by the vertical electric field only. It is this assumption which reduces the general 2-D Poisson's equation to the 1-D form (x-component only). This assumption is valid in the entire channel region except beyond the pinch-off point (to be discussed later). It may be noted that the gradual channel approximation is not valid for scaled MOS transistors.
3. The threshold voltage V_T is not a function of the position y along the channel.

With these assumptions, the inversion charge density is given as

$$Q_n(y) = -C_{ox}\left[V_{GS} - V_T - V_{CS}(y)\right] \tag{3.14}$$

Substituting (3.14) in (3.13) and integrating along the channel from the source end $(y = 0, V_{CS} = 0)$ to the drain end $(y = L, V_{CS} = V_{DS})$, we have

$$\int_0^L I_{DS}dy = \mu_{ns}WC_{ox}\int_0^{V_{DS}}\left[V_{GS} - V_T - V_{CS}(y)\right] \tag{3.15}$$

From this, the drain current for a long-channel MOS transistor is written as

$$I_{DS} = \mu_{ns}C_{ox}\frac{W}{L}\left[\left(V_{GS} - V_T - \frac{1}{2}V_{DS}\right)V_{DS}\right] \tag{3.16}$$

It is interesting to appreciate a simple physical implication of (3.16), which can be written as

$$I_{DS} = WC_{ox}\left(V_{GS} - V_T - \frac{1}{2}V_{DS}\right)\mu_{ns}\frac{V_{DS}}{L} \tag{3.17}$$

Here $C_{ox}\left(V_{GS} - V_T - \frac{1}{2}V_{DS}\right)$ may be interpreted as the average inversion charge density Q_n in the channel, V_{DS}/L is the average electric field in the channel. Thus the drain current is due to an average inversion charge density that drifts under the influence of a constant electric field.

It is observed from (3.14) that as the drain bias increases, the average inversion charge density decreases and dI_{DS}/dV_{DS} decreases. By differentiating (3.16) with respect to V_{DS}, it can be shown that dI_{DS}/dV_{DS} becomes zero at a definite V_{DS}, which is written as

$$\frac{dI_{DS}}{dV_{DS}} = 0 = \frac{W}{L}\mu_{ns}C_{ox}\left(V_{GS} - V_T - V_{DS}\right) \tag{3.18}$$

This yields

$$V_{DSsat} = \left(V_{GS} - V_T\right) \tag{3.19}$$

V_{DSsat} is referred to as the drain-to-source saturation voltage. When the applied drain bias is less than V_{DSsat}, the MOS transistor operates in the linear region of its I-V characteristics and when the applied drain bias is greater than V_{DSsat}, the MOS transistor operates in the saturation region of its I-V characteristics. In the linear region, the MOS transistor simply acts like a resistor with a sheet resistivity

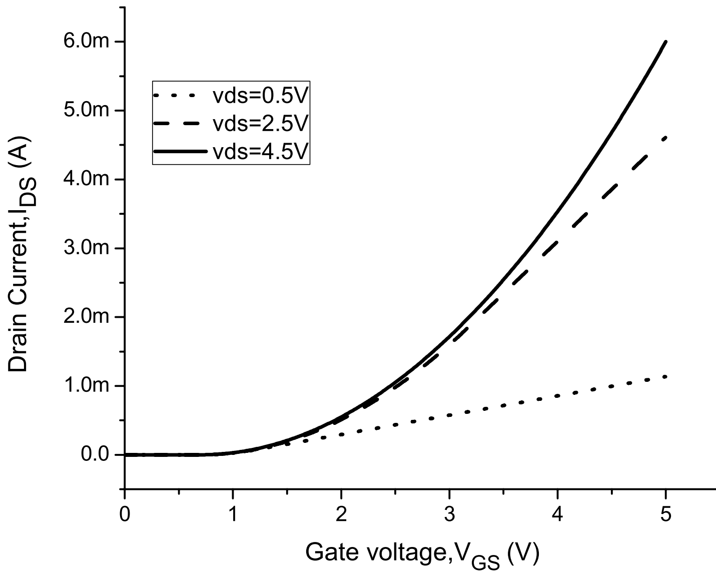
$$\rho_{sh} = \frac{1}{\mu_{ns}}C_{ox}\left(V_{GS} - V_T\right) \tag{3.20}$$

modulated by the gate voltage. Beyond V_{DSsat}, I_{DS} stays constant at I_{DSsat}, independent of V_{DS}. This is given as from (3.19) and (3.16)

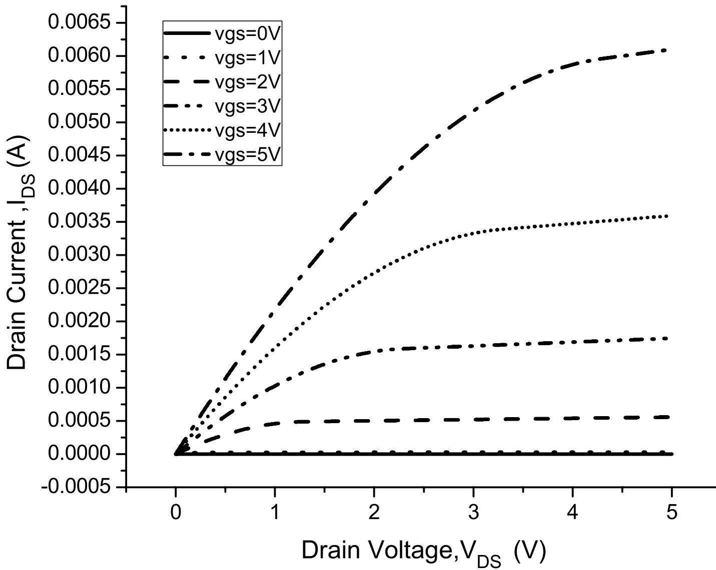
$$I_{DSsat} = \mu_{ns}C_{ox}\frac{W}{2L}\left(V_{GS} - V_T\right)^2 \tag{3.21}$$

The saturation of drain current is understood from the inversion charge density. The surface channel vanishes at the drain end of the channel when saturation occurs. This phenomenon is referred to as pinch-off of the channel. The electrons on reaching the pinch-off region in the channel are swept across due to high drift velocity caused due to the drain potential. Thus the pinch-off region does not present a barrier to the current flow. The drain current given by (3.16) and (3.21) defines the SPICE-Level 1 I-V model.

The gate characteristics as obtained from SPICE Level-1 simulation is shown in Fig. 3.4(a). The corresponding drain characteristics are shown in Fig.3.4(b). From the gate characteristics, the threshold voltage is obtained as 0.7V. It is observed from the drain characteristics that the saturation voltage increases with the increase in the drain voltage.



(a) Gate characteristics



(b) Drain characteristics

FIGURE 3.4
Gate and drain characteristics of an n-channel MOS transistor as obtained from SPICE-Level 1 simulation.

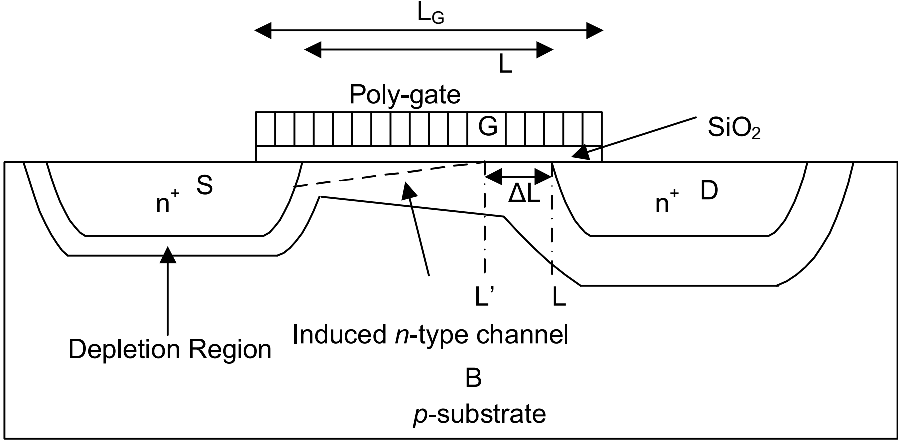


FIGURE 3.5
Pinch off phenomenon.

3.6.1 Channel Length Modulation Effect

The saturation drain current model derived earlier predicts that the drain current is independent of V_{DS} in the pinch-off region. However, in practice, the drain current in the pinch-off region varies slightly as the drain voltage is varied. This is because of the presence of the depletion region between the physical pinch-off point in the channel (the location in the channel where the inversion charge becomes zero) and the physical drain region. As V_{DS} exceeds V_{DSsat}, the pinch-off region begins to move slightly towards the source so that the effective channel length reduces and thus the drain current increases. This is known as channel length modulation effect [132].

Considering Fig. 3.5, let the depletion layer width between the pinch-off point and the drain region be ΔL; then the reduced channel length is given by

$$L' = L - \Delta L \tag{3.22}$$

Substituting L by L' in the saturation drain current model (3.21), a more accurate model for current in the pinch-off region is

$$I_{DSsat} = \mu_{ns} C_{ox} \frac{W}{2L'} \left(V_{GS} - V_T\right)^2 \tag{3.23}$$

Because ΔL and thus L' are functions of V_{DS} in the pinch-off region, the drain current I_{DS} varies with V_{DS}. From (3.22) and (3.23), we obtain

$$\frac{\partial I_{DS}}{\partial V_{DS}} = \frac{\partial I_{DS}}{\partial L'} \frac{\partial L'}{\partial V_{DS}} = -\frac{\mu_{ns} C_{ox} W}{2L'^2} \left(V_{GS} - V_T\right)^2 \frac{\partial L'}{\partial V_{DS}} = \frac{I_{DS}}{L'} \frac{\partial \Delta L}{\partial V_{Ds}} \tag{3.24}$$

The Early voltage is defined as

$$V_A = \frac{I_{DS}}{\left(\frac{\partial I_{DS}}{\partial V_{DS}}\right)} = L' \left(\frac{\partial \Delta L}{\partial V_{DS}}\right)^{-1} \tag{3.25}$$

The channel length modulation effect is usually characterized by the reciprocal of the Early voltage, $\lambda = 1/V_A$, where λ is known as the channel length modulation parameter. With this, the saturation drain current model is modified as follows

$$\begin{aligned} I_{DS} &= \mu_{ns} C_{ox} \frac{W}{2L} (V_{GS} - V_T)^2 \left(1 + \frac{V_{DS} - V_{DSsat}}{V_A}\right) \\ &= \mu_{ns} C_{ox} \frac{W}{2L} (V_{GS} - V_T)^2 \left[1 + \lambda (V_{DS} - V_{Dssat})\right] \end{aligned} \tag{3.26}$$

The channel length modulation parameter λ is inversely proportional to the effective channel length and the value reduces with increase in doping level in the channel. Typical values of λ are in the range $0.05V^{-1}$ to $0.005V^{-1}$. The channel length modulation effect is demonstrated through the drain characteristics in Fig. 3.4(b). The value of the saturation drain current I_{DSsat} and λ can be calculated by fitting any of the drain characteristics with a straight line and noting the values of the intercept and slope. The value of λ as observed from Fig.3.4(b) is $0.04V^{-1}$, which exactly matches with that provided in the simulation model file.

3.7 SPICE Level 3 I-V Model

The derivation of the SPICE Level 1 drain current model is based on the assumption that the bulk depletion charge Q_b due to the ionized acceptors in the depletion region near the surface remains constant throughout the channel. However, this is not a valid assumption. The channel voltage $V_{CS}(y)$ increases from source to drain, therefore the width of the depletion region increases from source to drain. Consequently the bulk depletion charge increases along the channel direction y. The surface potential is pinned at $\psi_s = 2\Phi_F + V_{CS}(y)$. This is referred to as the bulk charge effect [132]. The bulk depletion charge density is written as

$$Q_b(y) = -qN_{sub}W_{dm} = -\sqrt{2qN_{sub}\epsilon_{Si}\left[2\Phi_F + V_{CS}(y) - V_{BS}\right]} \tag{3.27}$$

The (-)ve sign in (3.27) is considering n-channel MOS transistor and $N_{sub} = N_A$. The total charge density in silicon is given by

$$Q_s = -C_{ox}(V_{GS} - V_{FB} - \psi_s) = -C_{ox}(V_{GS} - V_{FB} - 2\Phi_F - V_{CS}(y)) \tag{3.28}$$

The inversion charge density is then given by the difference of (3.28) and (3.27) and is given by

$$Q_n = -C_{ox}\left(V_{GS} - V_{FB} - 2\Phi_F - V_{CS}(y)\right) + \sqrt{2\epsilon_{Si} q N_{sub}\left(2\Phi_F + V_{CS}(y)\right)} \tag{3.29}$$

considering $V_{BS} = 0$. Substituting this in (3.13) and performing the integration as done earlier in (3.15), the drain current as a function of gate and drain voltages is given by

$$I_{DS} = \mu_{ns} C_{ox} \frac{W}{L}\left[V'_{GS} V_{DS} - \frac{2\sqrt{2\epsilon_{Si} q N_{sub}}}{3C_{ox}}\left[\left(2\Phi_F + V_{DS}\right)^{3/2} - \left(2\Phi_F\right)^{3/2}\right]\right] \tag{3.30}$$

where $V'_{GS} = \left(V_{GS} - V_{FB} - 2\Phi_F - \frac{1}{2}V_{DS}\right)$. This is the SPICE-Level 3 model. However, it is computationally difficult to evaluate (3.30) because of the mixed square and 3/2 power law dependence for I_{DS} on V_{DS}.

Due to the bulk charge effect, the threshold voltage of a MOS transistor varies along the channel and is expressed as

$$V_T(y) = V_T(0) + \gamma\left(\sqrt{2\Phi_F - V_{BS} + V_{CS}(y)} - \sqrt{2\Phi_F - V_{BS}}\right) \tag{3.31}$$

In (3.31), $V_T(0)$ means the threshold voltage at the source. Using Taylor expansion, a linear expression can be found to describe the bulk charge effect due to V_{DS},

$$V_T(y) = V_T(0) - \alpha V_{CS}(y) \tag{3.32}$$

where the bulk-charge factor α is defined as

$$\alpha = 1 + \frac{g\gamma}{2\sqrt{2\Phi_F - V_{BS}}} \tag{3.33}$$

and

$$g = 1 - \frac{1}{1.744 + 0.836\left(2\Phi_F - V_{BS}\right)} \tag{3.34}$$

From (3.9) and (3.33), it is observed that the bulk-charge factor α is very closely related to the body-effect coefficient m and some authors use them interchangeably. With the introduction of the bulk charge effect, the inversion charge density (3.14) is given by

$$Q_n(y) = -C_{ox}\left[V_{GS} - V_T - \alpha V_{CS}(y)\right] \tag{3.35}$$

The drain current for a long-channel MOS transistor (3.16) is thus given by

$$I_{DS} = \mu_{ns} C_{ox} \frac{W}{L}\left[\left(V_{GS} - V_T - \frac{\alpha}{2}V_{DS}\right) V_{DS}\right] \tag{3.36}$$

The value of the bulk-charge factor α is greater than 1 (a typical value is 1.5) and thus the drain current is reduced from the value predicted without

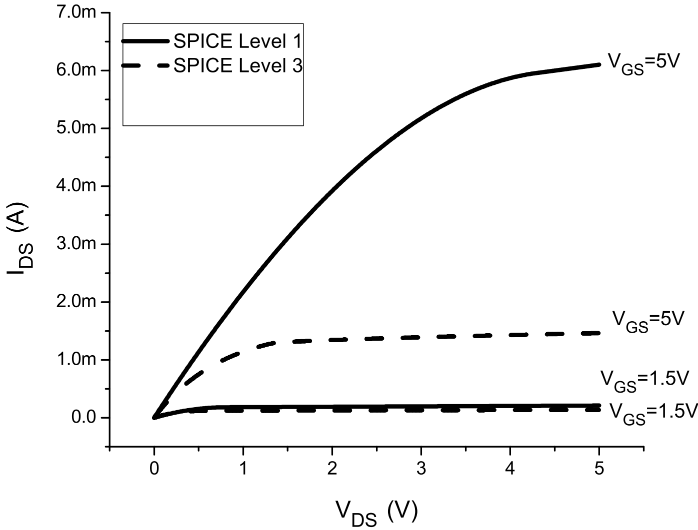


FIGURE 3.6
Comparison between the drain characteristics of an n-channel MOS transistor of dimension $W = 5\mu m, L = 1\mu m$, simulated by Level 1 and Level 3 SPICE model.

considering the bulk-charge factor. The drain-to-source saturation voltage is thus given by

$$V_{DSsat} = \frac{V_{GS} - V_T}{\alpha} \tag{3.37}$$

The drain saturation current is thus given by

$$I_{DSsat} = \mu_{ns} C_{ox} \frac{W}{2L} \frac{(V_{GS} - V_T)^2}{\alpha} \tag{3.38}$$

A comparison between the drain characteristics of an n-channel MOS transistor of same dimension, simulated with SPICE Level 1 model and Level 3 model is shown in Fig. 3.6. It is observed that the SPICE Level 1 model overestimates the current. This is because of the assumption involved that the threshold voltage variation along the channel is constant, which in turn, overestimates the channel inversion charge.

3.8 MOSFET Capacitances

The various capacitors present within an n-channel MOS transistor are identified in Fig. 3.7. The MOS transistor capacitances are categorically divided into

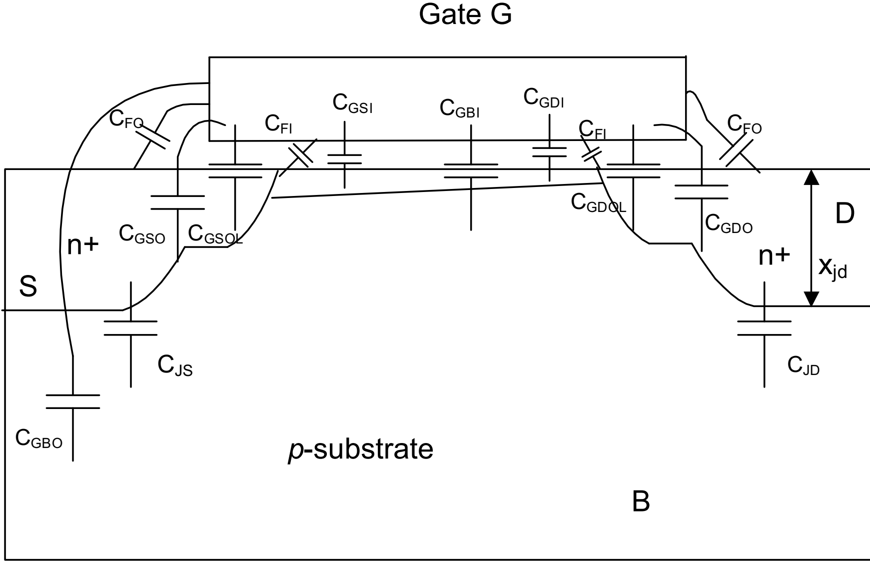


FIGURE 3.7
Various capacitors within an MOS transistor.

two components: intrinsic and extrinsic. The region between the metallurgical source and the drain junction where the gate to source/drain region is at flat band voltage, is referred to as the intrinsic region. The extrinsic capacitances are divided into five components: These are [192, 30]

1. outer fringing capacitances between the poly-silicon gate and the source/drain (S/D) region: C_{FO},
2. inner fringing capacitances between the poly-silicon gate and the S/D region: C_{FI},
3. the overlap capacitances between the gate and the heavily doped S/D regions (as well as the bulk region), $C_{GS0}, C_{GD0}(C_{GB0})$,
4. overlap capacitances between the gate and the lightly doped S/D regions C_{GSOL}, C_{GDOL}
5. source/drain junction capacitances C_{JS} and C_{JD}.

3.8.1 Characterization of Intrinsic Capacitances

In the Meyer's approach [130] for characterizing the intrinsic capacitances, the intrinsic capacitances are treated as lumped capacitances: gate-to-source capacitance C_{GS}, gate-to-drain capacitance C_{GD} and gate-to-bulk capacitance C_{GB}. The variations of the gate charges in the three terminals caused by variations of the corresponding terminal voltages leads to the various intrinsic

capacitances and are expressed in a compact fashion as follows.

$$C_{GZ} = \frac{\partial Q_G}{\partial V_{GZ}} \tag{3.39}$$

Here C_{GZ} is the intrinsic capacitance between the gate and the terminal Z(S/D) and V_{GZ} is the corresponding voltage difference. From charge neutrality condition, it follows that

$$Q_g(y) = -\left(Q_n(y) + Q_b(y)\right) \tag{3.40}$$

where Q_n and Q_b are the inversion charge density and bulk depletion charge density respectively and Q_g is the charge on the gate per unit area. In a strong inversion region, the inversion charge density is given by (3.14) and is repeated here for convenience

$$Q_n(y) = -C_{ox}\left(V_{GS} - V_T - V_{CS}\right) = -C_{ox}V_{GCT} \tag{3.41}$$

where V_T is the threshold voltage and V_{CS} is the channel potential. The drain-to-source current is given by

$$I_{DS}dy = -\mu_{ns} W Q_n . dV_{CS} \tag{3.42}$$

$$I_{DS} = \mu_{ns} W Q_n(y) \frac{dV_{GCT}(y)}{dy} \tag{3.43}$$

The integration is done within an appropriate limit after substituting Q_n and is written as

$$I_{DS}\int_0^L dy = \mu_{ns} W C_{ox} \int_{V_{GD}-V_T}^{V_{GS}-V_T} V_{GCT} dV_{GCT} \tag{3.44}$$

The drain current is given by

$$I_{DS} = \frac{\mu_{ns} W C_{ox}}{2L}\left[\left(V_{GS} - V_T\right)^2 - \left(V_{GD} - V_T\right)^2\right] \tag{3.45}$$

Considering the variation of the charges along the channel length, the total gate charge is given by

$$Q_G = -W\int_0^L Q_n(y)dy - W\int_0^L Q_b(y)dy = -W\int_0^L Q_n(y)dy - Q_B \tag{3.46}$$

where Q_B is the total bulk depletion charge. Substituting appropriate equations, this is given by

$$Q_G = \frac{2}{3} W L C_{ox}\left[\frac{\left(V_{GD} - V_T\right)^3 - \left(V_{GS} - V_T\right)^3}{\left(V_{GD} - V_T\right)^2 - \left(V_{GS} - V_T\right)^2}\right] - Q_B \tag{3.47}$$

The intrinsic capacitances in the linear region are determined from the following equations

$$C_{GS} = \left|\frac{\partial Q_G}{\partial V_{GS}}\right|_{V_{GD},V_{GB}} \tag{3.48}$$

$$C_{GD} = \left|\frac{\partial Q_G}{\partial V_{GD}}\right|_{V_{GS},V_{GB}} \tag{3.49}$$

$$C_{GB} = \left|\frac{\partial Q_G}{\partial V_{GB}}\right|_{V_{GS},V_{GD}} \tag{3.50}$$

Therefore, differentiating (3.47) following (3.48), (3.49) and (3.50), we get the following expressions for the three intrinsic capacitances in linear region

$$C_{GS} = \frac{2}{3}WLC_{ox}\left[1 - \frac{(V_{GD}-V_T)^2}{(V_{GS}-2V_T+V_{GD})}\right] \tag{3.51}$$

$$C_{GD} = \frac{2}{3}WLC_{ox}\left[1 - \frac{(V_{GS}-V_T)^2}{(V_{GS}-2V_T+V_{GD})}\right] \tag{3.52}$$

$$C_{GB} = 0 \tag{3.53}$$

The fact that C_{GB} is zero at the strong inversion region is explained by the fact that the inversion layer in the channel from the source to the drain screens the silicon bulk from the gate charge. Therefore, any change in substrate bias does not affect the gate charge.

In the saturation region, the gate-to-drain voltage is

$$V_{GD} = V_{GS} - V_{DSsat} = V_T \tag{3.54}$$

The total gate charge is therefore,

$$Q_G = \frac{2}{3}WLC_{ox}\left(V_{GS}-V_T\right) - Q_B \tag{3.55}$$

The various intrinsic capacitances in the saturation region are thus given by

$$C_{GS} = \frac{2}{3}WLC_{ox} \tag{3.56}$$

$$C_{GD} = 0 \tag{3.57}$$

$$C_{GB} = 0 \tag{3.58}$$

The physical explanation for (3.58) is same as that given for (3.53). The physical explanation for the fact $C_{GD} = 0$ is that in the saturation region the channel is pinched off at the drain end of the channel, so that the channel is electrically isolated from the drain. Therefore, the gate charge is not influenced by the change in drain voltage and thus $C_{GD} = 0$.

In the weak inversion region, the inversion charge is negligible compared

to the bulk depletion charge, so that the charge neutrality condition is given by

$$Q_g = -Q_b = C_{ox}\gamma\sqrt{\psi_a} \tag{3.59}$$

where ψ_a is the surface potential in the weak inversion region and is given by [194]

$$\psi_a = \left(\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}}\right)^2 \tag{3.60}$$

The total gate charge in the weak inversion region is given by

$$Q_G = -\frac{1}{2}WLC_{ox}\gamma^2\left[1 - \sqrt{1 + \frac{4}{\gamma^2}(V_{GB} - V_{FB})}\right] \tag{3.61}$$

The various intrinsic capacitances in the weak inversion region are determined as follows

$$C_{GS} = 0 \tag{3.62}$$

$$C_{GD} = 0 \tag{3.63}$$

$$C_{GB} = \frac{WLC_{ox}}{\sqrt{1 + \frac{4}{\gamma^2}(V_{GB} - V_{FB})}} \tag{3.64}$$

It is observed from (3.56) that $C_{GS} = 2/3WLC_{ox}$ when $V_{GS} = V_T$ in the saturation region. However, when $V_{GS} < V_T$, $C_{GS} = 0$ according to (3.62). In order to avoid this discontinuity it is proposed that C_{GS} decreases linearly from $2/3WLC_{ox}$ at $V_{GS} = V_T$ to zero at $V_{GS} = V_T - \Phi_F$. This is justified in the sense that the channel charge cannot become zero until the inversion layer vanishes totally.

In the accumulation region, $C_{GB} = C_{ox}WL$ and $C_{GS} = C_{GD} = 0$.

3.8.1.1 Charge Partitioning

The Meyer's model assumes that the node capacitance elements are all reciprocal (e.g., $C_{GD} = C_{DG}$) and computes the gate capacitance elements by considering only the gate charge. However, this model yields non-physical results when used for the simulation of circuits that have charge storage nodes, e.g., MOS charge pumps, switched capacitor circuits etc. This is referred to as the charge non-conservation problem. SPICE Level 2 models onward use the Ward-Dutton model [200] for capacitance calculations. This model considers the capacitive elements as non-reciprocal and assigns charges to each of the device terminals.

In the Ward-Dutton model, the inversion charge $Q_n.W.L$ is divided between the source and the drain terminals. This is achieved by introducing a parameter XQC, defined by

$$Q_D = XQC \times Q_n.W.L \tag{3.65}$$

$$Q_S = (1 - XQC) \times Q_n.W.L \tag{3.66}$$

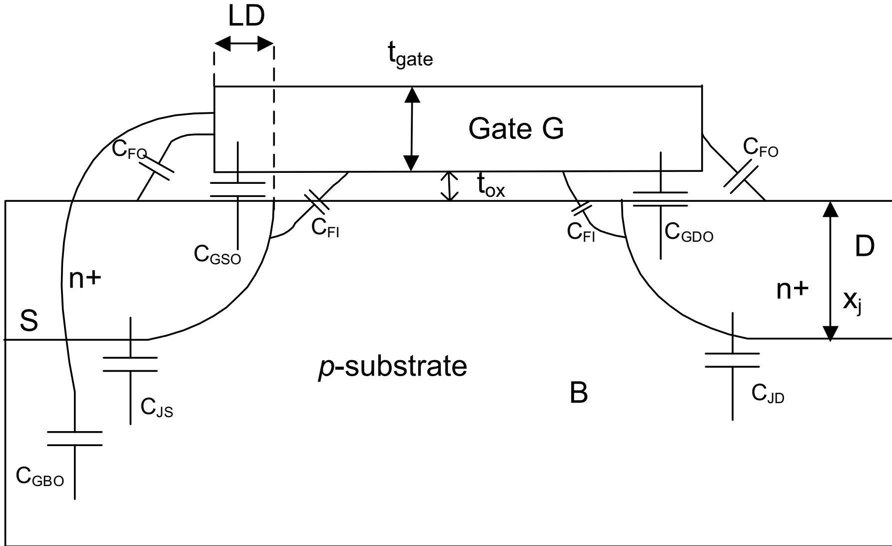


FIGURE 3.8
The overlap, fringing, and junction capacitances.

The default value of the factor XQC is 0.5 signifying equal charge partitioning. However, the value of this factor is often considered to be 0.4, which gives good matching with experimental results.

3.8.2 Characterization of Extrinsic Capacitances

The extrinsic components of MOS capacitances are of three broad types: (i) gate overlap capacitances in S/D and bulk region, (ii) inner and outer fringing capacitances and (iii) S/D junction capacitances.

3.8.2.1 Overlap and Fringing Capacitances

The various overlap and fringing capacitances are shown in Fig. 3.8. The overlap is due to the lateral diffusion of the source and the drain underneath the polysilicon. The overlap capacitance is approximated as

$$C_{GXov} = LD.W.C_{ox} = \frac{\epsilon_{ox} W.LD}{t_{ox}} = CGXO.W \tag{3.67}$$

where $CGXO$ is the overlap capacitance per unit length, W is the effective channel width, C_{ox} is the oxide capacitance per unit area and LD is the amount of overlap caused due to the lateral diffusion. Besides the gate-to-source/drain overlap, there is an additional parasitic capacitance between the gate and the bulk caused by the over-layer of the poly-silicon gate required

at one or both ends. The width of the over-layer is the channel length of the device. Thus the gate-to-bulk overlap capacitance is given by

$$C_{GBov} = C_{GBO}L \tag{3.68}$$

where C_{GBO} is the gate-to-bulk overlap capacitance per unit length. This capacitance exists only in the cut-off region.

In the linear and saturation regions, the overlap capacitances are given as

$$C_{GSov} = CGSO.W \tag{3.69}$$

$$C_{GDov} = CGDO.W \tag{3.70}$$

$$C_{GBov} = 0 \tag{3.71}$$

These capacitances are added to the intrinsic capacitance of the corresponding regions as determined from the Meyer's formula. It may be noted that the overlap length LD needs to be interpreted as the equivalent overlap length rather than the physical overlap length. Because of the lateral source-drain doping gradient at the surface, the overlap capacitance depends upon the drain bias. As a result of the enhanced drain depletion region with the application of drain bias, the equivalent overlap length LD reduces and the overlap capacitance value decreases slightly. This is especially the case with modern MOS transistors having source/drain extensions [192]. It has been reported that a minimum length of direction overlap region of the order of $LD \approx (2-3)t_{ox}$ is required to avoid reliability problems caused due to hot-carrier injection into the ungated region [192].

By solving Laplace's equation analytically with appropriate boundary conditions, the outer and inner fringing capacitances are written as [172]

$$C_{FO} = \frac{2\epsilon_{ox}W}{\pi}\ln\left(1+\frac{t_{gate}}{t_{ox}}\right) \tag{3.72}$$

$$C_{FI} = \frac{2\epsilon_{Si}W}{\pi}\ln\left(1+\frac{X_j}{2t_{ox}}\right) \tag{3.73}$$

where t_{gate} is the height of the poly-silicon gate and x_j is the depth of the source or drain junction. It may be noted that the contribution of inner fringe capacitance is comparatively larger than the outer fringe capacitance because of higher dielectric constant of silicon. However, it is present only in the weak inversion region. In the strong inversion region, the inversion layer shields any electrostatic coupling between the gate and the inner edges of the source or drain junctions. A similar situation happens in the accumulation region also, when the applied gate bias is negative for the n-channel MOS transistor. Therefore, except in the weak inversion region, the overlap capacitance consists of the direct overlap and the outer fringe components [192].

3.8.2.2 Junction Capacitances

The junction or diffusion capacitances arise from the depletion charge between the S/D and the substrate. With the variation of source or drain voltages, the

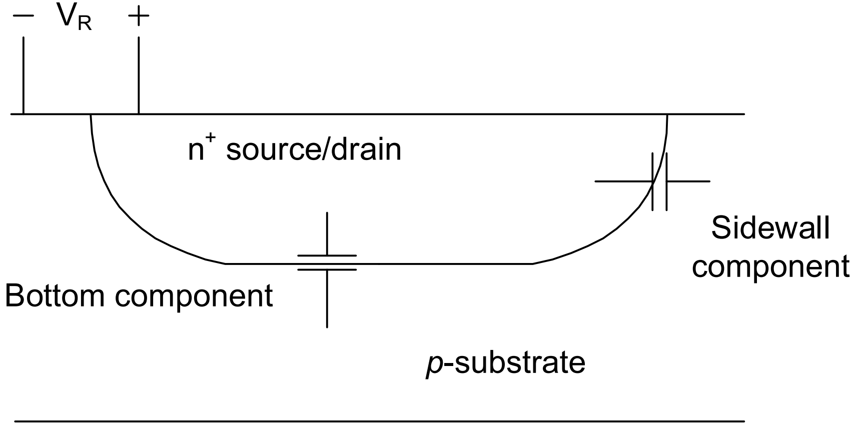


FIGURE 3.9
Bottom and sidewall component of a junction capacitance.

depletion charge increases or decreases accordingly. The junction capacitance per unit area of an abrupt p-n junction is [192]

$$C_J = \frac{\epsilon_{Si}}{W_{dj}} = \sqrt{\frac{\epsilon_{Si} q N_{sub}}{2\left(\psi_{bi} + V_R\right)}} = \left[\frac{\epsilon_{Si} q N_{sub}}{2\left(\psi_{bi} + V_R\right)}\right]^m \tag{3.74}$$

Here W_{dj} is the depletion layer width underneath the S/D junctions with respect to the substrate, ψ_{bi} is the built-in potential as defined below

$$\psi_{bi} = \frac{kT}{q} \ln\left(\frac{N_{sub} N_{SD}}{n_i^2}\right) \tag{3.75}$$

and N_{sub} is the substrate concentration, N_{SD} is the S/D concentration and V_R is the reverse bias voltage across the junction. In (3.74), m is the grading coefficient whose value is 1/2 for abrupt p-n junction. For zero-bias, the capacitance C_{J0} is defined as

$$C_{J0} = \left[\frac{\epsilon_{Si} q N_{sub}}{2\psi_{bi}}\right]^m \tag{3.76}$$

With this, (3.74) can be algebraically written as

$$C_J = C_{J0}\left(1 + \frac{V_R}{\psi_{bi}}\right)^{-m} \tag{3.77}$$

The junction capacitance has two components: a bottom component and a sidewall/perimeter component. This is illustrated in Fig. 3.9. The total junction capacitance is thus written as

$$C_J = C_{JB} A + C_{JSW} P \tag{3.78}$$

C_{JB} is the bottom component of the junction capacitance per unit area, A is the total junction area, C_{JSW} is the sidewall component of the junction capacitance per unit length and P is the total junction perimeter. Using (3.77), these components are written as

$$C_{JB} = C_{JB0}\left(1 + \frac{V_R}{\psi_{bi}}\right)^{-m_B} \tag{3.79}$$

$$C_{JSW} = C_{JSW0}\left(1 + \frac{V_R}{\psi_{bi}}\right)^{-m_{SW}} \tag{3.80}$$

Here m_B is the grading coefficient for the bottom component and m_{SW} is the sidewall component.

The variations of the capacitances C_{GS} and C_{GD} with the drain-to-source bias V_{DS} for three different gate biases as obtained from SPICE simulations are shown in Fig. 3.10(a) and Fig. 3.10(b) respectively. Some observations from the graphs are (1) In the cut-off region/sub-threshold region, $C_{GS} = 1fF, C_{GD} = 1fF$. Although the intrinsic capacitances are zero, the gate-to-source/drain overlap capacitances contribute in this region. These are calculated as $CGSO/CGDO \times W_{eff}$. (2) For $V_{GS=3V}$, the transistor remains in the linear region up to $V_{DS} = 2V$. On the other hand, for $V_{GS} = 1V$, the transistor remains in the linear region up to $V_{DS} = 0.5V$. Therefore, the two graphs merge at the saturation region. (3) The value of C_{GS} in the saturation region is greater than in the linear region. (4) The value of C_{GD} in the saturation region is due to the extrinsic components.

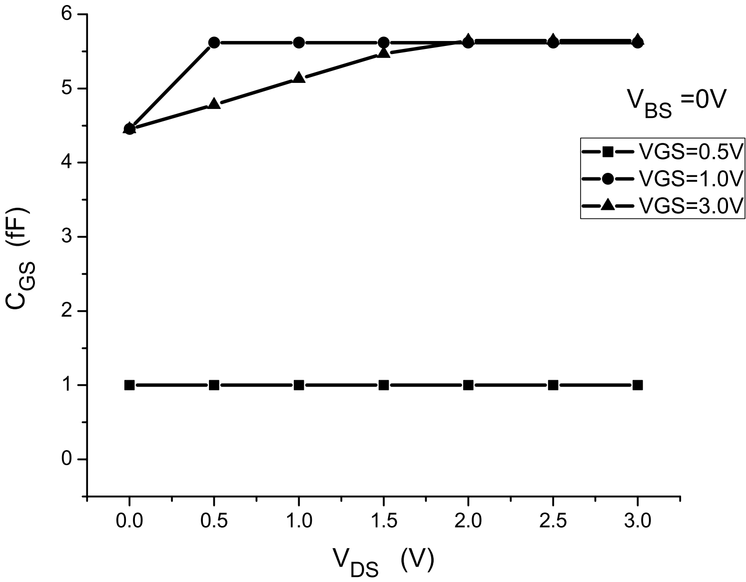
3.9 Short-Channel MOS Transistor

MOS transistors are classified into long or scaled (short) MOS transistors based on the relation of the channel length to the depletion widths under the gate for $V_{DS} \approx 0$. When the channel length of a MOS transistor becomes comparable to the sum of the source and drain depletion widths, the transistor is said to be a scaled or short-channel transistor. The minimum channel length for which a MOS transistor may be considered to be long-channel MOS transistor is given by the following criterion [189].

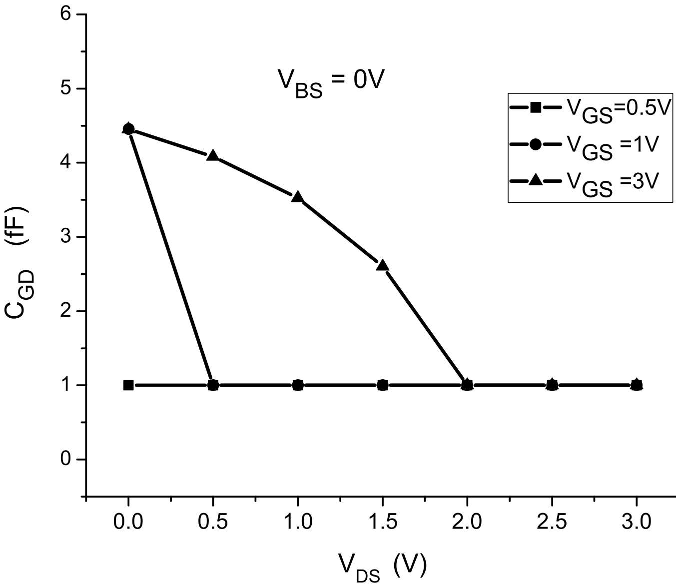
$$L \geq C_1\left[x_j t_{ox}\left(W_S + W_D\right)^2\right]^{1/3} \tag{3.81}$$

where x_j is the S/D junction depth, t_{ox} is the oxide thickness, $(W_S + W_D)$ is the sum of the source and drain depletion widths. For $V_{DS} = 0, W_D = W_S$.

The characteristics of a scaled MOS transistor differ from those of a long-channel MOS transistor primarily due to the following reasons. First, the electric field in the channel of a scaled MOS transistor is two dimensional in



(a) Variation of C_{GS} with drain bias



(b) Variation of C_{GD} with drain bias

FIGURE 3.10
Variation of intrinsic capacitances with drain bias as obtained from SPICE simulation.

contrast to that of a long-channel MOS transistor. For short-channel MOS transistors, the two-dimensional Poisson's equation is given by [189]

$$\frac{\partial^2 \psi_i}{\partial x^2} + \frac{\partial^2 \psi_i}{\partial y^2} = -\frac{\rho}{\epsilon_{Si}} \tag{3.82}$$

where ρ is the space charge density within the channel. Considering uniform substrate concentration and under depletion approximation, the Poisson's equation is written in terms of the electric field as

$$\frac{\partial \xi_x}{\partial x} + \frac{\partial \xi_y}{\partial y} = \frac{\rho}{\epsilon_{Si}} = -\frac{qN_{sub}}{\epsilon_{Si}} \tag{3.83}$$

Here ξ_x is the vertical component of the electric field and ξ_y is the lateral component of the electric field. The vertical component of the electric field is originated due to the gate voltage. On the other hand, the lateral component of the electric field is originated due to the S/D regions for short-channel length MOS transistors. The gradual channel approximation, which neglects the lateral component electric field, i.e., ξ_y is thus not applicable for a scaled MOS transistor. The existence of the 2-D channel field in short-channel MOS transistor is the fundamental difference between long-channel and short-channel MOS transistors. This lateral electric field ξ_y is responsible for various short-channel effects in scaled MOS transistors.

The electric field in the channel of a short-channel MOS transistor is high compared to that of a long-channel MOS transistor. It was shown in Chapter 1 that the scaling of supply voltage has not been achieved at the same rate as the scaling of feature sizes. Therefore, the electric field in the channel becomes high due to scaling of transistor dimensions. This leads to several phenomena related to the high electric field in scaled MOS transistors.

Miniaturization of feature size introduces several technology related phenomena such as unwanted parasitics, quantization of energy levels, non-uniform substrate etc. Therefore, scaled MOS transistors often demand use of new materials and non-conventional structures.

The important effects in scaled MOS transistors that have been reported in literature [30] are

1. Short-channel effects and narrow width effect
2. Carrier mobility degradation due to gate field
3. Carrier velocity saturation
4. Parasitic source and drain resistance effect
5. Poly-gate depletion layer effect
6. Hot electron effect and dielectric breakdown
7. Punchthrough effect
8. Gate induced drain leakage

9. Carrier energy quantization
10. Ballistic transport

These effects critically affect the various electrical characteristics of a scaled MOS transistor. The primary electrical quantities which are significantly affected are the threshold voltage, various terminal currents, small signal parameters and other performance parameters such as input referred noise etc. These are discussed in the subsequent sections.

3.10 Threshold Voltage for Short-Channel MOS Transistor

For a scaled MOS transistor the value of the threshold voltage reduces when the channel length decreases and the reduction is aggravated with increase of the drain bias. This phenomenon is known as the short-channel effect and is primarily caused by the combination of two effects: (i) source/drain charge sharing, (ii) drain induced barrier lowering. These are discussed below in the subsequent sub-sections.

3.10.1 Source/Drain Charge Sharing

The two-dimensional field pattern in scaled MOS transistors arises because of the closeness of source and drain regions. Similar to the depletion region underneath the gate, there exist depletion regions surrounding the source and the drain regions because of the p-n junction between the source/drain and substrate regions. For a long-channel MOS transistor, the source and drain regions are far apart and hence do not contribute to the field pattern in most part of the device. On the other hand, for short-channel MOS transistors, the field lines that terminate on the fixed depletion charges originate not only from the gate but also from the source and drain regions. This is schematically illustrated in Fig.3.11. This is referred to as the source/drain charge sharing. As is observed from Fig.3.11, for low drain bias, the field lines that terminate within the trapezoidal region PSTQ may be considered to originate from the gate. The rest of the field lines originate either from the source or from the drain. The total depletion charge within the trapezoidal region PSTQ is proportionally less than the total depletion charge within the rectangular region PRUQ for the long-channel case. Therefore, a smaller amount of gate voltage is able to induce inversion in short-channel MOS transistors compared to long-channel transistors. Consequently the threshold voltage magnitude is smaller in short-channel MOS transistors compared to long-channel transistors. As the channel length is reduced, the threshold voltage falls from its long-channel value. This is known as threshold voltage roll off.

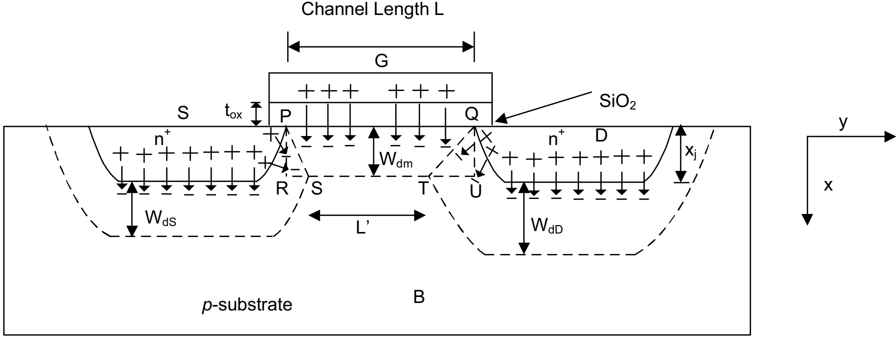


FIGURE 3.11
Source/drain charge sharing leading to threshold voltage reduction.

3.10.1.1 Level 2 Compact Model for V_T

First order estimation of the reduction of threshold voltage can be made by considering the charge partition. The total bulk depletion charge is

$$Q_B = WqN_{sub}W_{dm}\left(\frac{L+L'}{2}\right) \tag{3.84}$$

For small drain bias, a valid assumption is $W_{dS} = W_{dD} = W_{dm}$. From trigonometric analysis, it is easy to show that

$$L' = L - 2\left(\sqrt{x_j^2 + 2W_{dm}x_j} - x_j\right) \tag{3.85}$$

Consequently the threshold voltage shift from the long-channel behavior is given by [206].

$$\Delta V_T = -\frac{qN_{sub}W_{dm}x_j}{C_{ox}L}\left(\sqrt{1+\frac{2W_{dm}}{x_j}} - 1\right) \tag{3.86}$$

The negative sign indicates the fact that the threshold voltage is lowered. It is observed from (3.86), that the reduction of the maximum depletion depth W_{dm} and source/drain junction depth x_j are essential to reduce the threshold voltage roll off. To consider the drain bias and the substrate bias, (3.86) is modified as follows [1, 189]

$$\Delta V_T = -\frac{qN_{sub}W_{dm}x_j}{2C_{ox}L}\left[\left(\sqrt{1+\frac{2W_{dS}}{x_j}} - 1\right) + \left(\sqrt{1+\frac{2W_{dD}}{x_j}} - 1\right)\right] \tag{3.87}$$

where W_{dS} and W_{dD} are given as [189]

$$W_{dS} \approx \sqrt{\frac{2\epsilon_{Si}}{qN_{sub}}\left(\psi_{bi} - \psi_s - V_{BS}\right)} \tag{3.88}$$

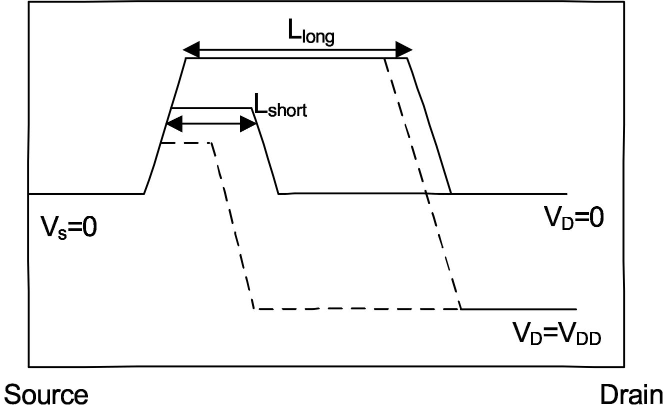


FIGURE 3.12
Illustration of barrier lowering for a short-channel MOS transistor with high drain bias.

$$W_{dD} \approx \sqrt{\frac{2\epsilon_{Si}}{qN_{sub}}\left(\psi_{bi} + V_{DS} - \psi_s - V_{BS}\right)} \tag{3.89}$$

It has been assumed here that the vertical depletion widths of the source/drain are equal to W_{dS} and W_{dD}. The threshold voltage expression with substrate bias as used in SPICE Level 2 compact model is written as [1]

$$V_T = V_{FB} + 2\Phi_F - \gamma f_s \sqrt{2\Phi_F - V_{BS}} \tag{3.90}$$

where f_s is known as the form factor and is defined as

$$f_s = 1 - \frac{x_j}{2L}\left[\left(\sqrt{1 + \frac{2W_{dS}}{x_j}} - 1\right) + \left(\sqrt{1 + \frac{2W_{dD}}{x_j}} - 1\right)\right] \tag{3.91}$$

For more aggressively scaled channel lengths, the Level 3 model includes a correction in the form factor based upon some empirical assumptions. This makes the expression of the form factor extremely complicated without offering any extra physical significance and hence is not given here.

3.10.2 Drain-Induced Barrier Lowering

The physics of the drain induced barrier lowering (DIBL) can be understood by considering the potential barrier (to electrons for an n-channel MOS transistor) at the surface between the source and the drain. This is illustrated in Fig. 3.12. When the transistor is in OFF state, the potential barrier (p-type region) prevents the carriers to flow from the source to the drain. With the increase of gate voltage the surface potential increases and consequently the energy barrier difference between the source and the channel decreases. This

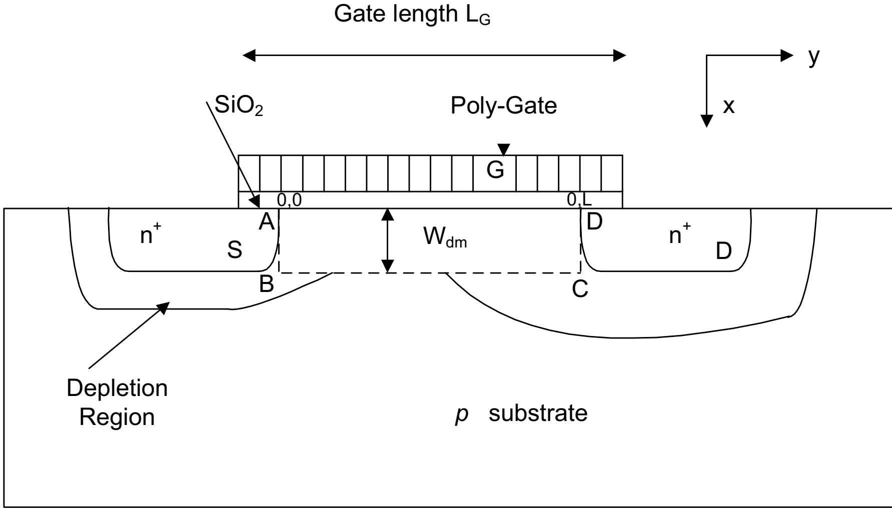


FIGURE 3.13
Gaussian box for the 2D analysis of V_T roll-off and DIBL.

leads to increase of carrier injection from the source to the channel over the lowered energy barrier. In a long-channel MOS transistor, the potential barrier is flat over most part of the device. The drain bias can change the effective channel length, but the barrier height between the source and the drain remains constant. In a scaled MOS transistor, on the other hand, the drain and source fields penetrate deeply into the middle of the channel, which lowers the potential barrier between the source and the drain. Therefore, at lower gate voltage, the carrier can overcome the barrier between the source and the channel. Therefore, the threshold voltage becomes lower than the long-channel value. With the application of high drain bias to a scaled MOS transistor, the barrier height is lowered even more, resulting in further decrease of the threshold voltage. The maximum barrier shifts toward the source. This phenomenon is called drain induced barrier lowering.

3.10.2.1 Level 3 Model of DIBL

The Level 3 model provides an empirical expression of the following form to account for the DIBL.

$$\Delta V_{T,\mathrm{DIBL}} = \sigma V_{DS} = \eta \frac{8.15 \times 10^{-22}}{C_{ox} L^3} V_{DS} \tag{3.92}$$

where η is often referred to as the coefficient of static feedback.

3.10.3 BSIM3/BSIM4 Compact Model for Threshold Voltage

A major limitation of the charge sharing model is that the assumption of constant surface potential for high drain and substrate bias is not valid for short-channel MOS transistors. In addition, the $1/L$ dependence of threshold voltage does not match with experimental data well. A fairly accurate model for the threshold voltage of a short-channel transistor is derived here using a quasi-two-dimensional approach [119]. Because of its simple functional form and computational efficiency, the model is suitable for device design and circuit simulation purposes.

Let us consider Fig. 3.13, which shows the cross-section of a MOS transistor at threshold voltage. The depletion region is formed due to the vertical field caused by the applied gate-to-source bias and the lateral field caused by the applied drain-to-source bias. Let us consider a rectangular Gaussian box ABCD, which is bounded between (i) source $y = 0$ and drain $y = L$ in the lateral dimension and (ii) the interface $x = 0$ and the edge of the neutral substrate region $x = W_{dm}$.

The 2-D Poisson equation in the channel is written as

$$\frac{\partial \xi_x}{\partial x} + \frac{\partial \xi_y}{\partial y} = -\frac{qN_{sub}}{\epsilon_{Si}} \tag{3.93}$$

where N_{sub} is the substrate concentration. The boundary conditions for the potential are written as

$$\psi(x,0) = \psi_s(0) = \psi_{bi} \quad \text{along AB} \tag{3.94}$$

$$\psi(x,L) = \psi_s(L) = V_{DS} + \psi_{bi} \quad \text{along DC} \tag{3.95}$$

$$\psi(W_{dm},y) = 0 \quad \text{along BC} \tag{3.96}$$

where ψ_{bi} is the built-in potential of the p-n junction formed between the source and the substrate regions. This is given as

$$\psi_{bi} = \frac{kT}{q} \ln\left(\frac{N_{sub}.N_{SD}}{n_i^2}\right) \tag{3.97}$$

where N_{SD} is the source-drain concentration. Now let us individually consider the two terms in the L.H.S of (3.93).

$$\frac{\partial \xi_x}{\partial x} = \frac{[\xi_x(0,y) - \xi(W_{dm},y)]}{W_{dm}} \tag{3.98}$$

Now we have

$$V_{GS} - V_{FB} - \psi_s(y) = \psi_{ox} = \xi_{ox}(0,y).t_{ox} \tag{3.99}$$

Therefore, we have

$$\xi_{ox}(0,y) = \frac{V_{GS} - V_{FB} - \psi_s(y)}{t_{ox}} \tag{3.100}$$

From the continuity of the displacement vector, $\epsilon_{ox}\xi_{ox}(0,y) = \epsilon_{Si}\xi_x(0,y)$ we have

$$\xi_x(0,y) = \frac{\epsilon_{ox}}{\epsilon_{Si}}\left(\frac{V_{GS} - V_{FB} - \psi_s(y)}{t_{ox}}\right) \tag{3.101}$$

Substituting (3.101) in (3.98) and considering the fact that at the depletion depth, the vertical electric field is zero, we have

$$\frac{\partial \xi_x}{\partial x} = \frac{\epsilon_{ox}}{\epsilon_{Si}}\frac{V_{GS} - V_{FB} - \psi_s(y)}{t_{ox}W_{dm}} \tag{3.102}$$

Now let us consider the second term in the L.H.S. of (3.93). For the considered Gaussian box, the term $\partial\xi_y/\partial y$ is a function of x along the depth. However, for simplicity the term is considered to be uniform along the depth, with an average value given as

$$\frac{\partial \xi_y}{\partial y} = \frac{1}{\eta}\frac{d\xi_{sy}}{dy} \tag{3.103}$$

where ξ_{sy} is the electric field at the surface ($x = 0$) and η is a fitting parameter. Therefore, substituting (3.102) and (3.103) in (3.93), we get

$$\frac{\epsilon_{ox}}{\epsilon_{Si}}\frac{V_{GS} - V_{FB} - \psi_s(y)}{t_{ox}W_{dm}} + \frac{1}{\eta}\frac{d\xi_{sy}}{dy} = -\frac{qN_{sub}}{\epsilon_{Si}} \tag{3.104}$$

This can be written as

$$\frac{d\xi_{sy}}{dy} + \left(\frac{V_{GS} - V_{FB} - \psi_s(y)}{l_t^2}\right) = -\frac{\eta q N_{sub}}{\epsilon_{Si}} \tag{3.105}$$

In (3.105), l_t as defined below is the characteristic length of the lateral electric field in the channel

$$l_t = \sqrt{\frac{\epsilon_{Si}t_{ox}W_{dm}}{\eta\epsilon_{ox}}} \tag{3.106}$$

Now considering the fact that $\xi_{sy} = -d\psi_s/dy$, we arrive at the following

$$\frac{d^2\psi_s(y)}{dy^2} - \frac{\psi_s(y)}{l_t^2} = \frac{qN_{sub}\eta}{\epsilon_{Si}} - \frac{(V_{GS} - V_{FB})}{l_t^2} \tag{3.107}$$

The solution to (3.107) is given by

$$\psi_s(y) = \psi_{sL} + (\psi_{bi} + V_{DS} - \psi_{sL})\frac{\sinh(y/L)}{\sinh(L/l_t)} + (\psi_{bi} - \psi_{sL})\frac{\sinh\left[(L-y)/l_t\right]}{\sinh(L/l_t)} \tag{3.108}$$

In (3.108), $\psi_{sL} = V_{GS} - V_{T0} + \psi_s$ represents the long-channel surface potential and V_{T0} is the long-channel threshold voltage. The channel potential expressed by (3.108) may be considered to be long-channel surface potential modified

by the source-drain field. It is to be noted that the depletion depth W_{dm} is assumed to be constant throughout the channel while deriving the equations. However, in reality, the depletion depth W_{dm} is a function of the drain voltage and channel length. This is incorporated by considering W_{dm}/η to be the average depletion depth in (3.106).

It is thus observed that with this approach, the surface potential in the channel does not remain constant in the channel, rather it varies along the channel. The location y_0 in the channel where the surface potential is minimum is derived from the condition [119]

$$\psi_{s\min} = \psi_s(y_0) \tag{3.109}$$

$$\left.\frac{d\psi_s}{dy}\right|_{y=y_0} = 0 \tag{3.110}$$

For the condition $V_{DS} << (\psi_{bi} - \psi_{sL}), \psi_{s\min}$ can be obtained from (3.110), assuming $y_0 = L/2$. This is given as

$$\psi_{s\min} = \psi_{sL} + [2(\psi_{bi} - \psi_{sL}) + V_{DS}] \frac{\sinh(L/2l_t)}{\sinh(L/l_t)} \tag{3.111}$$

At threshold voltage $\psi_{s\min} = 2\Phi_F$. From the above considerations, the threshold voltage for short-channel MOS transistor is given by

$$V_T = V_{T0} - \Delta V_T \tag{3.112}$$

where V_{T0} is the long-channel threshold voltage, given by (3.2) and ΔV_T is the amount of reduction of threshold voltage due to short-channel effect and ΔV_T is given by

$$\Delta V_T = \theta_T(L)[2(\psi_{bi} - \psi_s) + V_{DS}] \tag{3.113}$$

In (3.113) $\theta_T(L)$ is the short-channel effect coefficient depending on the channel length and is given by

$$\theta_T(L) = \frac{1}{2\cosh\left(\frac{L}{l_t}\right) - 2} \tag{3.114}$$

In order to make the model valid for different technologies, several empirical parameters have been included in the model. The short-channel roll off model is written as [55]

$$\theta_T(\mathrm{SCE}) = \frac{DVT0}{2\cosh\left(DVT1.\frac{L}{l_t} - 1\right)} \tag{3.115}$$

With this the threshold voltage reduction due to short-channel effect only is written as

$$\Delta V_T(\mathrm{SCE}) = \theta_T(\mathrm{SCE})(\psi_{bi} - \psi_s) \tag{3.116}$$

The characteristic length is modified to [55]

$$l_t = \sqrt{\frac{\epsilon_{Si} t_{ox} W_{dm}}{\epsilon_{ox}}} \left(1 + DVT2.V_{BS}\right) \tag{3.117}$$

Correspondingly the DIBL effect is expressed as [55]

$$\theta_T \left(\text{DIBL}\right) = \frac{1}{\cosh\left(DSUB.\frac{L}{l_{t0}} - 1\right)} \tag{3.118}$$

$$\Delta V_T(\text{DIBL}) = \theta_T(\text{DIBL})(ETA0 + ETAB.V_{BS})V_{DS} \tag{3.119}$$

with l_{t0} written as

$$l_{t0} = \sqrt{\frac{\epsilon_{Si} t_{ox} W_{dm0}}{\epsilon_{ox}}} \tag{3.120}$$

and

$$W_{dm0} = \sqrt{\frac{2\epsilon_{Si}\psi_s}{qN_{sub}}} \tag{3.121}$$

It may be noted that $DVT1$ may be considered to be equal to $1/\sqrt{\eta}$, $DVT2$ and $ETAB$ account for substrate bias effects on SCE and DIBL respectively. For non-uniformly doped substrate, N_{sub} is to be replaced by the channel doping concentration, e.g., N_{DEP}.

The SPICE simulation results (using PTM- 45nm, BSIM 4 model) illustrating the short channel effects on the threshold voltage (V_T roll-off and DIBL phenomenon) of an n-channel MOS transistor is shown in Fig. 3.14(a) and Fig. 3.14(b) respectively. The width of the transistor is taken to be $5\mu m$ for simulation purposes, in order to avoid any contribution due to narrow width effect. The channel length is taken to be $65nm$. The supply voltage is $1V$. The physical oxide thickness is $1.1nm$ and the electrical oxide thickness is $1.75nm$. The substrate is uniformly doped and the concentration is $3.24E18/cm^3$. The source/drain concentration is $2E20/cm^3$. A comparison between the threshold voltage roll off and DIBL characteristics as obtained from BSIM simulation results and calculated analytically using the models discussed above is shown in Fig. 3.15.

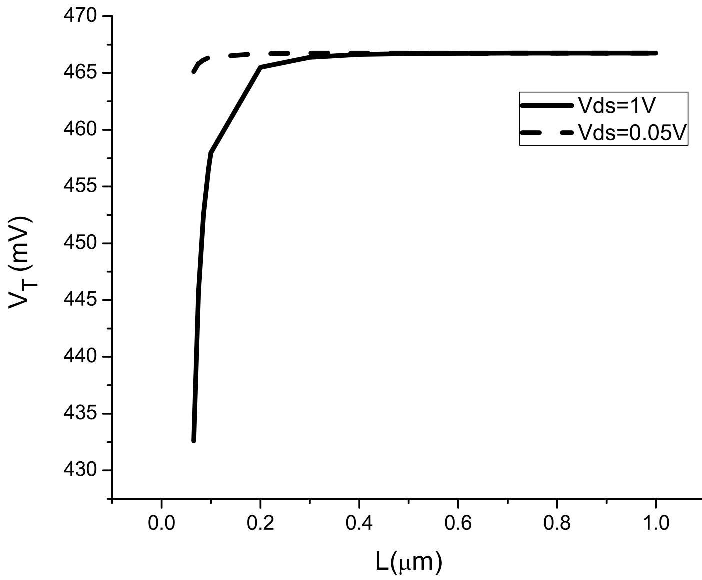
3.10.4 Short-Channel Effect Immunity

In order to minimize the short-channel effect, the aspect ratio of the MOS transistor should be sufficiently large. The aspect ratio is defined as

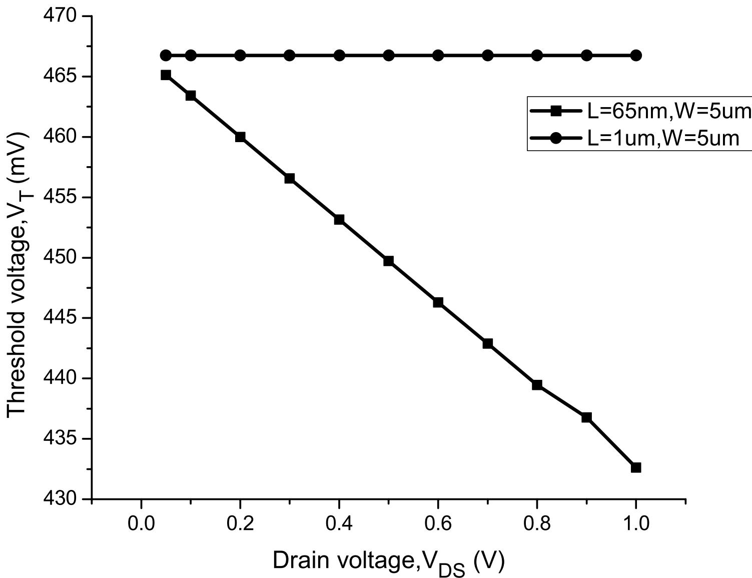
$$AR = \frac{\text{dimension}_{\text{lateral}}}{\text{dimension}_{\text{vertical}}} \tag{3.122}$$

For a MOS transistor, the aspect ratio can be expressed as [154]

$$AR = \frac{L}{\left[t_{ox}\left(\epsilon_{Si}/\epsilon_{ox}\right)\right]^{1/3} x_j^{1/3} W_{dm}^{1/3}} \tag{3.123}$$



(a) Threshold voltage roll off



(b) DIBL phenomenon

FIGURE 3.14
SPICE simulation results for short channel effects on threshold voltage, $V_{BS} = 0V$.

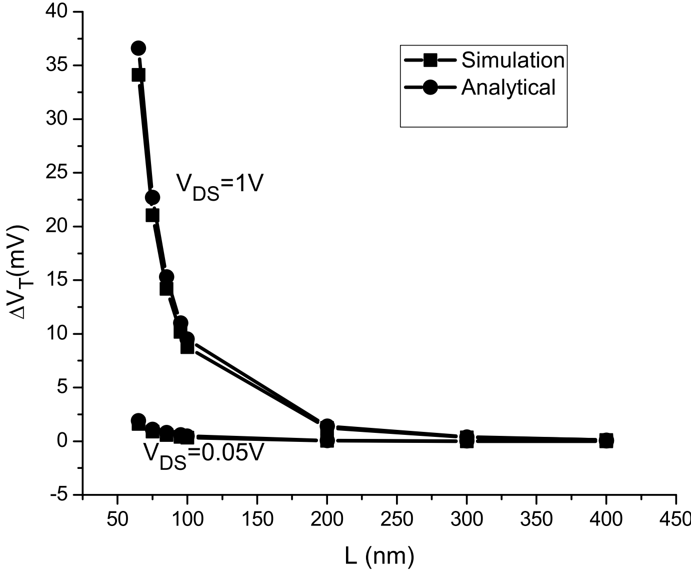


FIGURE 3.15
Comparison between BSIM simulation results and analytical results for threshold voltage roll off and DIBL characteristics.

where ϵ_{Si} and ϵ_{ox} are the permittivities of the silicon and silicon-dioxide respectively, x_j is the source/drain junction depth, W_{dm} is the maximum depletion depth and t_{ox} is the oxide thickness. It is therefore, observed that reduction of oxide thickness, junction depth and maximum depletion depth are the keys to minimize the short channel effects in a scaled MOS transistor. The values of the aspect ratio for a MOS transistor of channel length $65nm$ at three different technology nodes are summarized in Table 3.1. It is observed that with scaling down of the technology nodes, the various parameters are scaled in a manner such that the aspect ratio is improved.

TABLE 3.1
Aspect Ratios for Transistor of Channel Length $L = 65nm$ with Technology Nodes

Node	AR	t_{ox}	N_{sub}	x_j
32nm	24.45	$1.65nm$	$4.12E18/cm^3$	10nm
45nm	21.03	$1.75nm$	$5.24E18/cm^3$	14nm
65nm	18.07	$1.85nm$	$2.54E18/cm^3$	19.6nm

3.11 I-V Model for Short-Channel MOS Transistor

Drain characteristics of scaled MOS transistors depend upon two significant physical phenomena, namely the carrier mobility degradation and the carrier velocity saturation. Therefore, these are discussed first.

3.11.1 Carrier Mobility Degradation

The degradation of mobility of the carriers in the channel region due to applied gate voltage plays a significant role in the drain current characteristics of a MOS transistor. The drain current of a MOS transistor is determined by the mobility of the carriers in the channel region, which is significantly different from the bulk mobility. The reasons are discussed below.

3.11.1.1 Surface Mobility

The surface mobility of the carriers differ from the bulk mobility primarily because of several scattering mechanisms which take place in the channel region. These are the [192, 189].

1. The phonon scattering
2. The Coulombic scattering
3. The surface roughness scattering

The lattice vibrations in silicon at the interface emit and absorb phonons while exchanging energy with the carriers resulting in lattice scattering. The frequency of such scattering phenomenon increases as the temperature is increased, since the thermal agitation of the lattice increases. The phonon scattering leads to reduced mobility of the carriers at the surface.

The carriers in the channel region suffer collision with the ionized impurities, interface state charges and fixed oxide charges. The Coulombic attraction and repulsion of the carriers with the interface state charges and fixed oxide charges leads to scattering of the carriers from the normal drift motion. This is referred to as the Coulombic scattering, the rate of which is proportional to the channel doping concentration. The Coulombic scattering reduces as the temperature is reduced. This is because of the fact, that with increase in temperature, the thermal velocity of the carrier increases and thereby the carriers can overcome the Coulombic interaction. The Coulombic scattering plays a dominant role in the weak and the moderate inversion condition compared to the strong inversion condition, because under the latter condition due to the screening phenomenon by the mobile inversion carriers, the effect of ionized scattering gets weakened.

The surface irregularities in the Si-SiO_2 act as scattering centers to the carriers. The surface roughness has significant effect in determining the mo-

bility of the carriers in the channel. At low temperature, surface roughness scattering is the dominant mechanism at high electric fields.

These three scattering mechanisms determine the surface carrier mobility through Matthiessen's rule, as [192, 189]

$$\frac{1}{\mu_s} = \frac{1}{\mu_c} + \frac{1}{\mu_{ph}} + \frac{1}{\mu_{sr}} \tag{3.124}$$

where $1/\mu_c$, $1/\mu_{ph}$ and $1/\mu_{sr}$ are the mobility components related to the coulomb scattering, the phonon scattering and the surface roughness scattering respectively and μ_s is the surface mobility of the carriers.

3.11.1.2 Mobility Dependence on Gate Field

Increasing the vertical electric field, i.e., the gate field, forces the carriers in the channel to come closer to the surface of the silicon. There the surface roughness impedes the movement of the carriers, thus reducing mobility. The average vertical electric field depends on the inversion and bulk charge under the gate through Gauss's law and is given by [85]

$$\xi_x = -\frac{1}{\epsilon_{Si}} \left(Q_b + \eta Q_n\right) \tag{3.125}$$

where η is approximately 1/2 for electrons and 1/3 for holes, the exact values depend upon the process technology. Substituting the following expressions

$$Q_n = -C_{ox} \left(V_{GS} - V_T\right) \tag{3.126}$$
$$Q_b = -C_{ox} \left(V_T - V_{FB} - 2\Phi_F\right) \tag{3.127}$$

the average vertical electric field is given by

$$\xi_x = \frac{(V_{GS} + V_T) - 2V_{FB} - 4\Phi_F}{6t_{ox}} \approx \frac{(V_{GS} + V_T) + 0.2V}{6t_{ox}} \tag{3.128}$$

for n^+ poly-silicon gate n-channel MOS transistors. The dependence of the surface mobility of the carriers on this average electric field and hence on the gate bias is given by the following empirical relationship [1, 85]

$$\mu_s = \frac{\mu_0}{1 + \left(\frac{\xi_x}{\xi_0}\right)^{\nu}} \tag{3.129}$$

Here μ_0 is the surface mobility of the carriers in the absence of any gate field, ν is a constant whose value is nearly 1.85 for electrons at the surface and nearly 1.0 for holes at the surface. ξ_0 is the critical electric field (nearly 0.9 MV/cm for electrons at surface and nearly 0.45MV/cm for holes at the surface). The model (3.129) although fitting experimental data well, is difficult to compute because it involves a power function. Therefore, making a Taylor series expansion of (3.129) in terms of some fitting parameters whose values

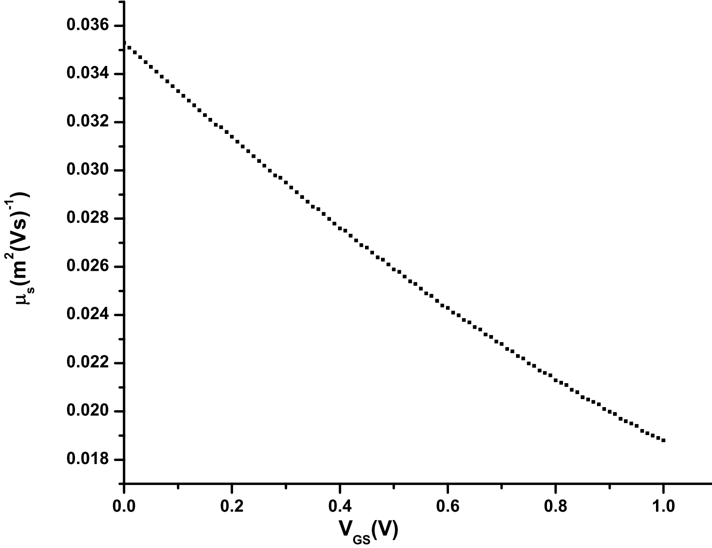


FIGURE 3.16
Variation of surface mobility of the carriers with gate bias.

are to be determined from experimental data, the following mobility models are proposed by BSIM [30].

$$\mu_s = \frac{\mu_0}{1 + U_A\left(\frac{V_{GS}+V_T}{t_{ox}}\right) + U_B\left(\frac{V_{GS}+V_T}{t_{ox}}\right)^2} \tag{3.130}$$

where U_A and U_B are the two fitting parameters. The substrate bias dependence of the mobility is incorporated by introducing another parameter, U_C in (3.130). With this, the model becomes

$$\mu_s = \frac{\mu_0}{1 + (U_A + U_C V_{BS})\left(\frac{V_{GS}+V_T}{t_{ox}}\right) + U_B\left(\frac{V_{GS}+V_T}{t_{ox}}\right)^2} \tag{3.131}$$

$$\mu_s = \frac{\mu_0}{1 + \left[U_A\left(\frac{V_{GS}+V_T}{t_{ox}}\right) + U_B\left(\frac{V_{GS}+V_T}{t_{ox}}\right)^2\right](1 + U_C V_{BS})} \tag{3.132}$$

These different models are incorporated in the SPICE simulator by using suitable mobility selector flags. It is observed that the mobility in a strong inversion region is a function of the gate bias. The variation of surface mobility with gate bias is shown in Fig. 3.16. However, in a weak inversion region, the variation of inversion charge with gate voltage cannot be modeled accurately.

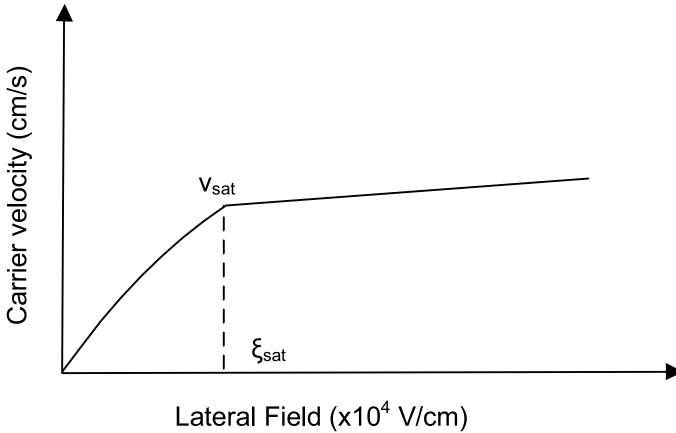


FIGURE 3.17
Illustration of carrier velocity saturation.

Hence the mobility of the carriers is considered to be constant in the weak inversion region [85, 30].

It may be noted that carrier mobility degradation affects long-channel transistors too. However, the effect is serious for scaled MOS transistors. This is because of the fact that ξ_x is higher for scaled MOS transistors because the power supply voltage is often not scaled as much as suggested by constant field scaling when the gate-oxide thickness is reduced.

3.11.2 Carrier Velocity Saturation

Carrier velocity saturation is another important physical phenomenon that critically affects the I-V characteristics of a scaled MOS transistor. If the lateral electric field is small, the drift velocity of the carriers is linearly proportional to the applied lateral field and is given by [1, 85]

$$v_d = \mu_s \xi_y \tag{3.133}$$

Here μ_s is the surface carrier mobility and is independent of the lateral field ξ_y. However, as the lateral field ξ_y becomes high, the carrier velocity no longer follows (3.133). With the increase of lateral field, the kinetic energy of the carriers increases. When the energy of a carrier exceeds the optical phonon energy, it generates an optical phonon and much of its velocity is lost in this process. Consequently, the kinetic energy and therefore the drift velocity cannot exceed a certain value. The limiting velocity is called saturation velocity [85]. This is illustrated in Fig. 3.17. Measured data indicate that the saturation velocity of an electron in a surface channel is between 6 and $10 \times 10^6 cm/s$ and that of a hole is between 4 and $8 \times 10^6 cm/s$. At lower fields, the velocity

is fitted by an equation of the form [85, 132]

$$v_d = \frac{\mu_s \xi_y}{\left(1 + \frac{\xi_y}{\xi_{sat}}\right)} \qquad \xi_y < \xi_{sat} \tag{3.134}$$

On the other hand, for fields greater than the saturation field ξ_{sat}, the velocity becomes constant and is given by

$$v_d = v_{sat} \qquad \xi_y > \xi_{sat} \tag{3.135}$$

The parameter ξ_{sat} is obtained by solving the field for which $v_d = v_{sat}$ as predicted by (3.134) so that the $v_d - \xi_y$ relation is continuous. Hence,

$$\xi_{sat} = \frac{2v_{sat}}{\mu_s} \tag{3.136}$$

Despite the various simplifications and assumptions made to derive this model, this is found to be useful for predicting drain current in scaled MOS transistors.

3.11.3 Drain Current in Scaled MOS Transistor

To derive the I-V model of a scaled MOS transistor, we start with (3.13), which is repeated here for convenience.

$$I_{DS} = -WQ_n(y)\mu_{ns}\frac{\partial V(y)}{\partial y} \tag{3.137}$$

where μ_{ns} is the surface mobility of electron. Substituting the inversion charge expression for strong inversion from (3.35), and applying the mobility expression from (3.134) we have

$$I_{DS} = WC_{ox}\left(V_{GS} - \alpha V_{CS} - V_T\right)\frac{\mu_{ns}\frac{dV_{CS}}{dy}}{1 + \left(\frac{dV_{CS}}{dy}\right)/\xi_{sat}} \tag{3.138}$$

As done earlier by integrating this expression within appropriate limits, we have

$$\int_0^L I_{DS}dy = \int_0^{V_{DS}} \left[WC_{ox}\mu_{ns}\left(V_{GS} - \alpha V_{CS} - V_T\right) - I_{DS}/\xi_{sat}\right]dV_{CS} \tag{3.139}$$

Therefore, the drain current in a scaled MOS transistor is given by [85, 132]

$$I_{DS} = \frac{\frac{W}{L}C_{ox}\mu_{ns}\left(V_{GS} - V_T - \frac{\alpha}{2}V_{DS}\right)V_{DS}}{1 + \frac{V_{DS}}{\xi_{sat}L}} \tag{3.140}$$

When the channel length L is large, the denominator becomes unity and (3.140) reduces to (3.36).

Because of the carrier velocity saturation phenomenon, it is expected that the drain current saturates when the carriers arrive at the drain with their limiting velocity v_{sat}. Let the drain voltage at which the carriers saturate be designated by V_{DSsat}. The inversion charge density at the drain region is thus

$$Q_{nD}(y) = -WC_{ox}\left(V_{GS} - V_T - \alpha V_{DSsat}\right) \tag{3.141}$$

Therefore, the saturation drain current is given by

$$I_{DSsat} = WC_{ox}\left(V_{GS} - V_T - \alpha V_{DSsat}\right) v_{sat} \tag{3.142}$$

By equating (3.140) and (3.142) at $\xi_y = \xi_{sat}$ and $V_{DS} = V_{DSsat}$, we arrive at the following expression for the drain-to-source saturation voltage [85, 132]

$$\frac{1}{V_{DSsat}} = \frac{\alpha}{V_{GS} - V_T} + \frac{1}{\xi_{sat} L} \tag{3.143}$$

Thus it is observed that the drain-to-source voltage for a scaled MOS transistor is an average of $\xi_{sat}L$ and the long-channel $V_{DSsat} = \left(V_{GS} - V_T\right)/\alpha$. The variation of the saturation voltage V_{DSsat} with $(V_{GS} - V_T)$ is shown in Fig. 3.18. It is observed that the velocity saturation effect reduces the drain saturation voltage with scaling of channel length. In addition, the gate oxide thickness also has a significant effect on drain saturation voltage.

Some salient features of the drain current model for scaled MOS transistor are discussed below.

By comparing (3.36) and (3.140), it is observed that scaling only minimally modifies the behavior of drain current in triode/linear region. The only significant change is caused by mobility degradation, which is higher for scaled MOS transistors.

Substituting (3.143) in (3.142), the drain current of a scaled MOS transistor in the saturation region is given by [85]

$$I_{Dssat} = \frac{W}{2\alpha L}\mu_{ns} C_{ox} \frac{\left(V_{GS} - V_T\right)^2}{1 + \dfrac{V_{Gs} - V_T}{\alpha \xi_{sat} L}} = \frac{\text{long-channel } I_{Dssat}}{1 + \dfrac{V_{Gs} - V_T}{\alpha \xi_{sat} L}} \tag{3.144}$$

Two special cases are distinguished.

1. For long-channel or low V_{GS} case, $\xi_{sat}L >> (V_{GS} - V_T)$

$$V_{DSsat} = \frac{\left(V_{GS} - V_T\right)}{\alpha} \tag{3.145}$$

$$I_{DSsat} = \frac{W}{2\alpha L}\mu_{ns} C_{ox}\left(V_{GS} - V_T\right)^2 \tag{3.146}$$

For low power analog circuit design, the overdrive voltage is small so that $\xi_{sat}L >> (V_{GS} - V_T)$ and scaled transistors exhibit long-channel characteristics.

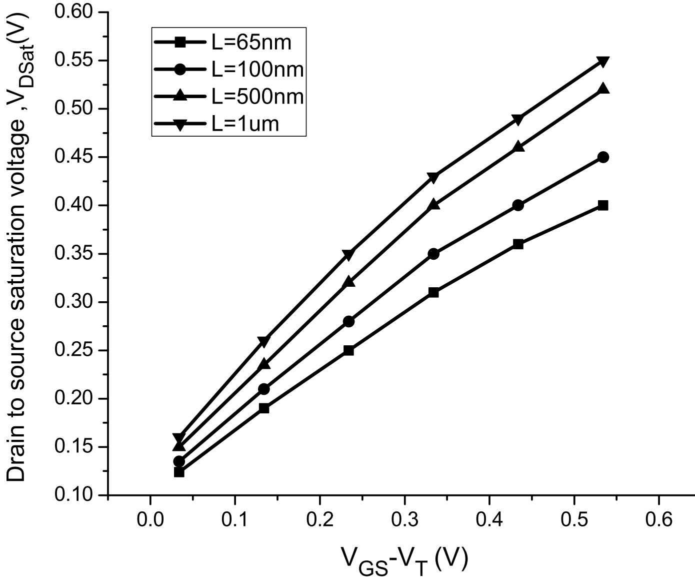


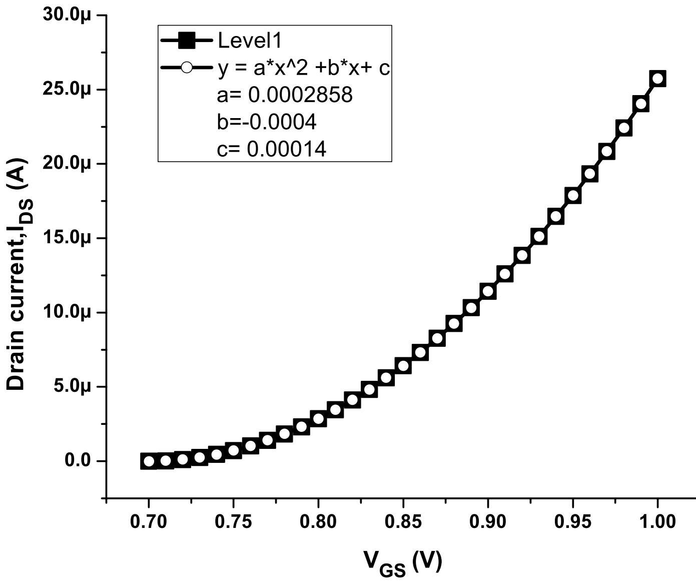
FIGURE 3.18
Variation of the drain saturation voltage with $(V_{GS} - V_T)$ in the presence of velocity saturation with channel length as a parameter.

2. For very short-channel transistors, $\xi_{sat}L << (V_{GS} - V_T)$, $V_{DSsat} \approx \xi_{sat}L < \dfrac{(V_{GS} - V_T)}{\alpha}$

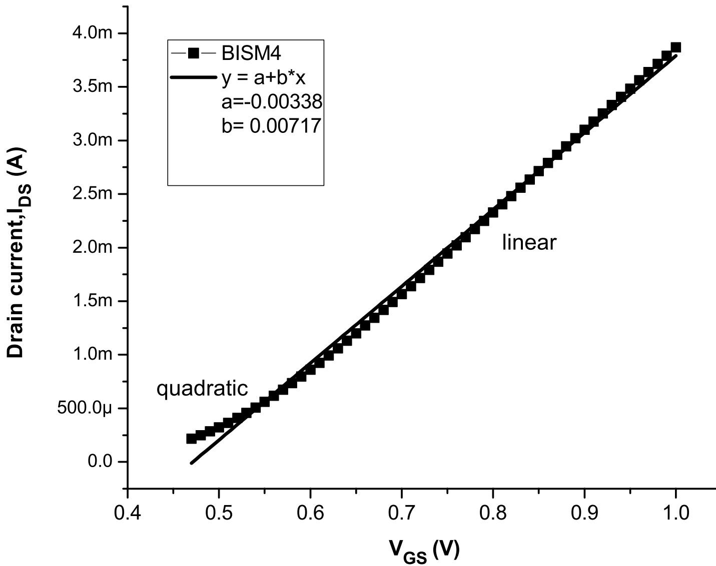
$$I_{DSsat} \approx Wv_{sat}C_{ox}\left(V_{GS} - V_t - \alpha\xi_{sat}L\right) \tag{3.147}$$

   It is interesting to note that I_{DSsat} is proportional to $(V_{GS} - V_T)$ rather than $(V_{GS} - V_T)^2$ and does not explicitly depend upon the channel length in comparison to a long-channel transistor. A changing L affects I_{DSsat} only through its influence on the term V_{DSsat}. In order to increase I_{DSsat} though device design, it is necessary to reduce the oxide thickness t_{ox} and minimize V_T. For circuit designers, the approach will be to increase the channel width W and use higher gate voltage V_{GS}. The scaling of oxide thickness is limited by tunneling leakage and oxide reliability, the threshold voltage V_T is limited by the transistor leakage in the off state, and the maximum V_{GS} is limited by concerns over circuit power consumption and device reliability.

The comparison between the gate characteristics for long channel and short channel MOS transistors, as obtained from SPICE simulation results is shown in Fig. 3.19(a) and Fig. 3.19(b) respectively. It is observed that the long-channel drain current shows a quadratic dependence on the gate voltage



(a) Long-channel MOS transistor



(b) Short-channel MOS transistor

FIGURE 3.19
Comparison of gate characteristics for long-channel ($1\mu m$) and short-channel ($65nm$) MOS transistors.

beyond the threshold voltage. On the other hand, the short-channel drain current shows a linear dependence. However, for low V_{GS}, the short channel characteristic shows a quadratic dependence, similar to that of the long-channel characteristic. This demonstrates the concept just given above.

The concept of pinch off is applicable for long-channel MOS transistors only where it suggests that drain current I_{DS} saturates when the inversion charge becomes zero at the drain end of the channel. For scaled MOS transistors, the saturation of the drain current is explained by the fact that the carrier velocity saturates to a limiting value v_{sat} at the drain. Thus instead of the pinch-off region, there is a velocity saturation region near the drain where the inversion charge density $Q_{nD}(y)$ as given by (3.141) is a constant.

3.12 Weak Inversion Characteristics of a Scaled MOS Transistor

The discussion made so far regarding the flow of drain current through a MOS transistor is based on the fact that the transistor is operating in the strong inversion region, where the applied gate-to-source voltage V_{GS} is assumed to cause only changes in the channel/inversion charge and not in the depletion-region charge. In addition to the normal region, i.e., the strong inversion region, there is another region of operation for a MOS transistor, which is referred to as the weak inversion region [192, 189]. In this region, the applied V_{GS} is less than the threshold voltage but high enough to create a depletion region at the surface of the silicon. The weak inversion region is defined by the condition $\Phi_F \leq \psi_s \leq 2\Phi_F$ [192]. In the weak inversion region, the inversion charge is much less than the depletion charge and the drain current conduction is dominated by the diffusion current due to gradient in minority carrier concentration. In the weak inversion region, the operation of an n-channel MOS transistor is like an npn bipoloar transistor, where the source acts as the emitter, substrate acts as the base and the drain acts as the collector [189]. The weak inversion drain current is often alternatively referred to as the subthreshold characteristics, especially in digital circuit design.

Let us assume that both the source and the substrate terminals are grounded and applied drain bias $V_{DS} > 0$. Then increasing the gate-to-source voltage V_{GS} increases the surface potential ψ_s which in turn forward biases the substrate-source (p-n) junction. From the basic theory of minority carrier injection it follows that the minority carrier concentration (note that electrons coming out from source are minority carriers in the substrate region) on each side of a p-n junction varies with the applied bias because of the variations in the diffusion of carriers across the junction. The minority carrier concentration

in the substrate at the source side is given by [189]

$$n_p(0) = n_{p0} \exp\left(\frac{\psi_s}{U_T}\right) \tag{3.148}$$

where n_{p0} is the equilibrium concentration of electrons (minority carriers) in the substrate and $U_T = kT/q$ is the thermal voltage. Similarly the concentration of electrons in the substrate at the drain side is given by [189]

$$n_p(L) = n_{p0} \exp\left(\frac{\psi_s - V_{DS}}{U_T}\right) \tag{3.149}$$

where V_{DS} is the applied drain bias. It is easy to understand the integration limits by considering Fig. 3.13 and the boundary conditions mentioned therein. Due to the diffusion of electrons within the depletion region due to concentration gradient, the weak inversion current flows. The electron density per unit area at the source end is

$$N'(0) = \int_0^{W_{dm}} n_p(0)dx = n_{p0} \int_{\psi_s}^{0} \exp\left(\frac{\psi_s}{U_T}\right) d\psi_s \tag{3.150}$$

Since the potential distribution inside the depletion region is known, this electron density is evaluated to be [189]

$$N'(0) = U_T \sqrt{\frac{\epsilon_{Si}}{2q\psi_s N_A}} n_{p0} \exp\left(\frac{\psi_s}{U_T}\right) \tag{3.151}$$

The electron density at the drain end is lowered exponentially by the applied drain bias and is given by

$$N'(L) = N'(0) \exp\left(\frac{-V_{DS}}{U_T}\right) \tag{3.152}$$

The drain diffusion current due to electron density gradient is written as

$$\begin{aligned} I_{DS} &= -WqD_n \frac{dN'(y)}{dy} \approx WqD_n \frac{N'(0) - N'(L)}{L} \\ &= \frac{W}{L} \mu_n U_T^2 \sqrt{\frac{q\epsilon_{Si} N_a}{2\psi_s}} \left(\frac{n_i}{N_A}\right)^2 \exp\left(\frac{\psi_s}{U_T}\right) \left[1 - \exp\left(-\frac{V_{DS}}{U_T}\right)\right] \end{aligned} \tag{3.153}$$

Here Einstein's relationship $\dfrac{D_n}{\mu_n} = U_T$ and $n_{p0} = \dfrac{n_i^2}{N_A}$ have been used, where n_i is the intrinsic carrier concentration.

In the weak inversion region, the depletion region exists in the substrate underneath the interface. Under this condition, the MOS capacitor consists of two capacitors in series: the oxide capacitor C_{ox} and the depletion layer

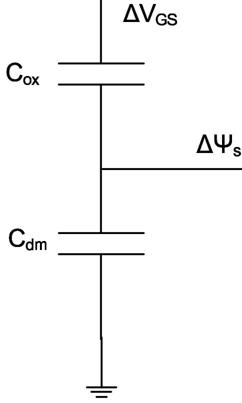


FIGURE 3.20
Equivalent capacitance network in the weak inversion region.

capacitor C_{dm} [192, 85]. This is as shown in Fig. 3.20. Therefore, a change in the applied gate voltage V_{GS} leads to a change in the surface potential through the voltage divider between C_{ox} and C_{dm}. Therefore,

$$\frac{d\psi_s}{dV_{GS}} = \frac{C_{ox}}{C_{dm} + C_{ox}} = \frac{1}{n} \tag{3.154}$$

where $n = 1 + \dfrac{C_{dm}}{C_{ox}}$ is the sub-threshold swing factor [85]. It may be noted the sub-threshold swing factor n should be equal to m, i.e., the body-effect coefficient according to (3.9). However, in practice, the value of n is somewhat larger than m [85]. The reason is that the value of the oxide capacitance is smaller in the weak inversion region compared to that in the strong inversion region. Nonetheless, n and m are very closely related and some authors use m in place of n while dealing with the weak inversion region model.

Assuming that the surface potential varies linearly with the gate bias in the weak inversion region, as shown in Fig. 3.21, just at the point where the surface potential $\psi_s = 2\Phi_F$, it can be written that [194]

$$\frac{d\psi_s}{dV_{GS}} = \frac{\psi_s - 2\Phi_F}{V_{GS} - V_T} = \frac{1}{n} \tag{3.155}$$

Substituting this in (3.153) and with proper simplifications we get [192]

$$I_{DS} = \mu_n \frac{W}{L} \sqrt{\frac{q\epsilon_{Si} N_A}{2\psi_s}} U_T^2 \exp\left(\frac{V_{GS} - V_T}{nU_T}\right) \left[1 - \exp\left(-\frac{V_{DS}}{U_T}\right)\right] \tag{3.156}$$

This is the final expression for the drain current that flows under the weak inversion condition. This is sometimes alternatively written as [192]

$$I_{DS} = \mu_n C_{ox} \frac{W}{L} (n-1) U_T^2 \exp\left(\frac{V_{GS} - V_T}{nU_T}\right) \left[1 - \exp\left(-\frac{V_{DS}}{U_T}\right)\right] \tag{3.157}$$

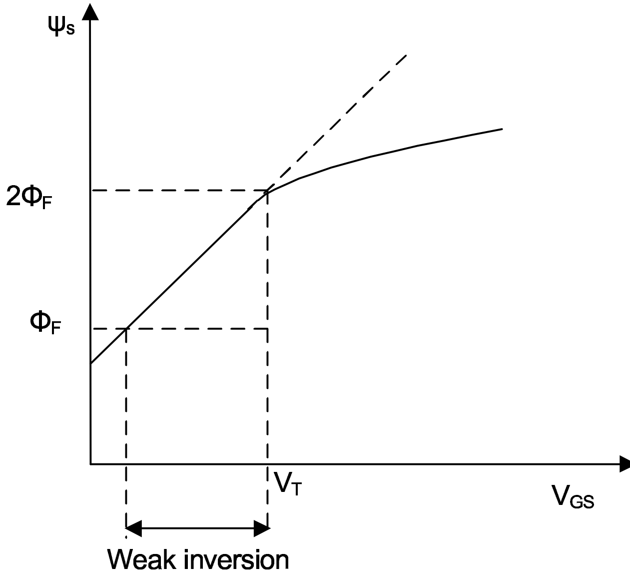


FIGURE 3.21
Variation of surface potential with gate bias in weak inversion region.

3.12.1 Subthreshold Swing

The variation of the weak inversion drain current with the applied gate bias is shown in Fig. 3.22. The parameter to quantify how sharply the transistor is turned off by the applied gate voltage is referred to as the subthreshold swing S (inverse of subthreshold slope). This is defined as the gate voltage change required to induce a drain current change of one order of magnitude and is expressed as [192, 189]

$$\begin{aligned} S &= \left(\frac{d\left(\log_{10} I_{DS}\right)}{dV_{GS}}\right)^{-1} = 2.3\frac{nkT}{q} \\ &= 2.3\frac{kT}{q}\left(1+\frac{C_{dm}}{C_{ox}}\right) \\ &\approx n \times 60mV \end{aligned} \tag{3.158}$$

The value comes out to be typically 70-100mV/decade. Therefore, I_{DS} drops by 10 times for every $n \times 60mV$. Therefore, if $n = 1.5$, I_{DS} drops by every $90mV$.

An useful technique to measure the threshold voltage of a MOS transistor is the constant current technique [85]. This is measured by the amount of gate-to-source voltage corresponding to which the drain current flowing through the transistor is $100nA \times W/L$. However, some designers may choose some other values of the drain current, e.g., $70nA$ or $200nA$. The sub-threshold

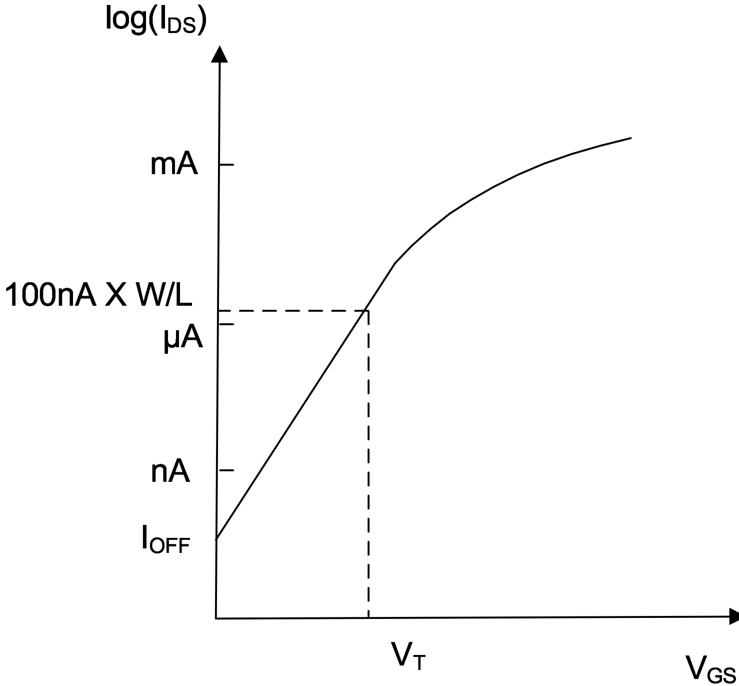


FIGURE 3.22
Variation of weak inversion drain current with gate bias at $V_{DS} = V_{DD}$.

current which flows at $V_{GS} = 0$ and $V_{DS} = V_{DD}$ is referred to as the OFF current. It is easy to derive that [85]

$$I_{OFF}(nA) \approx 100.\frac{W}{L}10^{-V_T/S} \tag{3.159}$$

There are two possible ways to minimize I_{OFF}. The first way is to increase the threshold voltage V_T. But this is not viable, because this degrades the ON current and hence the circuit speed. The second technique is to reduce the sub-threshold swing S. This may be achieved either by reducing the oxide thickness t_{ox} or by increasing the depletion depth W_{dm}. Another alternative approach to reduce the sub-threshold swing is to operate the transistor at a temperature much lower than the room temperature [192]. The variation of the weak inversion drain current as obtained from SPICE simulation results is shown in Fig. 3.23 and the corresponding parameters are summarized in Table 3.2.

In the presence of significant interface-trap density D_{it}, the associate capacitance $C_{it}(= q^2 D_{it})$ will act in parallel with the depletion layer capacitance C_{dm}. With these considerations, the sub-threshold swing factor n is written

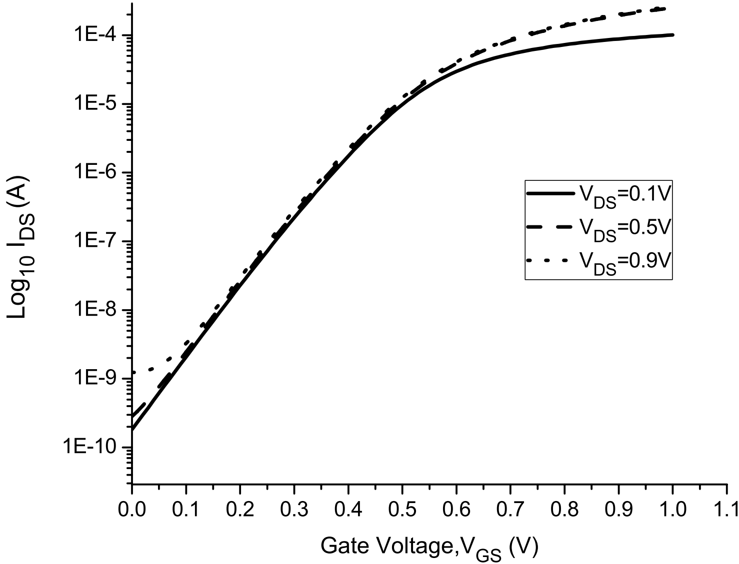


FIGURE 3.23
Variation of weak inversion drain current with gate bias for different drain biases as obtained from SPICE simulation results.

TABLE 3.2
Summary of OFF Current and Subthreshold Swing as Obtained from SPICE Simulation Results

V_{DS}	I_{OFF}	S
0.1V	0.184nA	94mV/dec
0.5V	0.320nA	96mV/dec
0.9V	1.270nA	101mV/dec

in BSIM compact model as [30]

$$n = 1 + \mathrm{NFACTOR}\frac{C_{dm}}{C_{ox}} + \frac{C_{it} + C_{DSC}}{C_{ox}} \tag{3.160}$$

The parameter NFACTOR is for compensating any error while calculating the depletion width capacitance. The value of this parameter is close to unity. The capacitance C_{DSC} is sensitive to the body bias as well as to the drain bias which is incorporated in the model as follows [30].

$$C_{DSC} = (C_{DSC} + C_{DSCD}V_{DS} + C_{DSCB}V_{BS})\,\frac{0.5}{\cosh\left(DVT1.\dfrac{L}{l_t}\right) - 1} \tag{3.161}$$

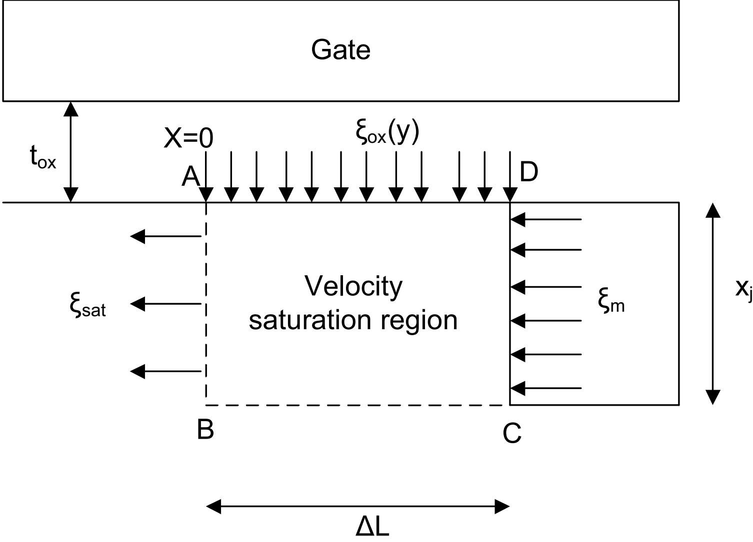


FIGURE 3.24
Schematic diagram of a MOS transistor in the velocity saturation region and the Gaussian box for computing the spatial distribution of electric field.

3.13 Hot Carrier Effect

The supply voltage scaling has been historically slow because of the system consideration and also the push for higher speed. Therefore, the electric field does not remain constant, rather it increases with scaling of dimensions. The electric field near the drain region causes impact ionization at a significant rate. Under sufficiently high electric field, an electron in the conduction band can gain sufficient energy to transfer an electron from the valence band to the conduction band, thus generating one free electron in the conduction band and one hole in the valence band. This is referred to as the impact ionization process [192]. Impact ionization is the physical mechanism for the generation of substrate current [192]. Therefore, it is essential to understand the electric field and its spatial distribution in the channel region of a scaled MOS transistor.

3.13.1 Spatial Distribution of Lateral Electric Field

When the applied drain bias exceeds the drain-to-source saturation voltage V_{DSsat}, the carriers travel at their limiting velocity, i.e., the saturation velocity v_{sat} through a portion of the channel region near the drain. This portion of the channel region is referred to as the velocity saturation region. The distance between the saturation point and the drain, ΔL (see Fig. 3.5.), is the amount

of channel length modulation by the drain voltage. With increasing drain voltage, ΔL increases so that the drain current continues to increase in the saturation region.

Let us consider the cross sectional geometry in the velocity saturation region near the drain end as shown in Fig. 3.24. The velocity saturation region is defined by the boundaries $y' = 0$ (saturation point) to $y' = \Delta L$ (drain end) and $x = 0$ (surface) to $x = x_j$ (drain junction depth). We consider a Gaussian box $ABCD$ that encloses a portion of the depletion region under the gate in the velocity saturation region near the drain. Applying Gauss's law, we have

$$-\xi_{sat} x_j W + \xi(y') x_j W + \frac{\epsilon_{ox}}{\epsilon_{Si}} \int_0^{y'} \xi_{ox}(y) dy W = \frac{Q}{\epsilon_{Si}} \tag{3.162}$$

where Q is the amount of charge enclosed in the Gaussian box and is given by the sum of depletion charge and mobile charge

$$Q = (N_A + n) q x_j y' W \tag{3.163}$$

Differentiating (3.162) with respect to y' and substituting (3.163) we get

$$x_j \frac{d\xi(y')}{dy'} + \frac{\epsilon_{ox}}{\epsilon_{Si}} \xi_{ox}(y') = \frac{q}{\epsilon_{Si}} x_j (N_A + n) \tag{3.164}$$

The oxide field ξ_{ox} at the interface is written as

$$\xi_{ox} = \frac{V_{GS} - V_{FB} - 2\Phi_F - V_{CS}(y')}{t_{ox}} \tag{3.165}$$

where $V_{CS}(y')$ is the quasi Fermi potential which increases from V_{DSsat} at $y' = 0$ to V_{DS} at $y = \Delta L$. Substituting (3.165) in (3.164), we obtain

$$x_j \frac{d\xi(y')}{dy'} + \frac{\epsilon_{ox}}{\epsilon_{Si}} \frac{[V_{GS} - V_{FB} - 2\Phi_F - V_{CS}(y')]}{t_{ox}} = \frac{q}{\epsilon_{Si}} x_j (N_A + n) \tag{3.166}$$

At $y' = 0$, all the silicon charges are still controlled by the gate, so that we have

$$\frac{V_{GS} - V_{FB} - 2\Phi_F - V_{DSsat}}{t_{ox}} = \xi_{ox}(y' = 0) = \frac{q}{\epsilon_{ox}} x_j (N_A + n) \tag{3.167}$$

From (3.166) and (3.167), we get

$$\epsilon_{Si} x_j \frac{d\xi(y')}{dy'} = C_{ox} [V_{CS}(y') - V_{DSsat}] \tag{3.168}$$

This is written as

$$\frac{d\xi(y')}{dy'} = \frac{[V_{CS}(y') - V_{DSsat}]}{l_t^2} \tag{3.169}$$

where the characteristic length l_t is given by

$$l_t = \sqrt{\frac{\epsilon_{Si}}{\epsilon_{ox}} t_{ox} x_j} \approx \sqrt{3 t_{ox} x_j} \tag{3.170}$$

The characteristic length is same as that derived earlier while considering the DIBL effect, with the minor difference that the depletion depth W_{dm} is assumed to be of same depth as the junction depth x_j.
(3.169) is a linear, second-order differential equation which can be solved with the boundary conditions $V_{CS}(0) = V_{DSsat}$ and $\xi(0) = \xi_{sat}$ to obtain

$$\xi(y') = \xi_{sat} \cosh\left(\frac{y'}{l_t}\right) \tag{3.171}$$

and

$$V_{CS}(y') = V_{DSsat} + l_t \xi_{sat} \sinh\left(\frac{y'}{l_t}\right) \tag{3.172}$$

It may be noted ξ_{sat} is as defined in (3.135) and is on the order of $5 \times 10^4 V/cm$ for electrons. The electric field is maximum at the drain end of the channel

$$\xi_m = \xi(y' = \Delta L) = \xi_{sat} \cosh\left(\frac{\Delta L}{l_t}\right) \tag{3.173}$$

and

$$V_{DS} = V_{DSsat} + l_t \xi_{sat} \sinh\left(\frac{\Delta L}{l_t}\right) \tag{3.174}$$

From these the length of the velocity saturation region is found to be

$$\Delta L = l_t \ln\left[\frac{\{(V_{DS} - V_{DSsat})/l_t\} + \xi_m}{\xi_{sat}}\right] \tag{3.175}$$

Therefore, the maximum electric field in the channel is written as

$$\xi_m = \left[\frac{(V_{DS} - V_{DSsat})^2}{l_t^2} + \xi_{sat}^2\right]^{1/2} \tag{3.176}$$

For higher drain voltage $(V_{DS} - V_{DSsat})/l_t >> \xi_{sat}$ so that the maximum electric field is approximated as

$$\xi_m \approx \frac{(V_{DS} - V_{DSsat})}{l_t} \tag{3.177}$$

It is important to have some physical insight of the electric field distribution.

1. At the origin of the Gaussian box, $y' = 0$, $V_{DS} = V_{DSsat}$. Up to this point, the gate has control over the mobile charges that are swept from the source to the drain. From this point onward the carriers travel with saturation velocity and start moving away from the

surface. In order to maintain the continuity of the drain current, the lateral electric field extending from the drain into the velocity saturation region increases by the same amount that the vertical gate field decreases. In terms of Poisson's equation, the charge released by the decreasing vertical field must be taken by the increasing lateral field. This is mathematically observed from (3.168). The right-hand side represents the amount of charge released by the vertical field due to the increase in quasi Fermi potential equal to $V_{CS}(y') - V_{DSsat}$. The left-hand side signifies the corresponding increase of the lateral field gradient that supports this charge.

2. From (3.171) it is observed that the lateral field increases almost exponentially toward the drain. This sharp increase is required to support the charge release by the vertical field.

3.13.2 Substrate Current Due to Hot-Carrier Effects

Several serious problems of scaled MOS transistor are caused due to hot-carrier effects. Even by considering the scaling of supply voltage, the electric field in the velocity saturation region is strong enough ($\approx 2{-}3\times10^4 V/cm$). The carriers are described as "hot" and because of their high energies are capable of physical effects that can degrade the MOS transistors. These highly energetic electrons are referred to as hot electrons because if their kinetic energy is expressed as kT_e, then T_e becomes as high as $1000K$, which is much higher than the lattice temperature.

The hot electrons collide with the bound electrons in the valence band to create impact ionization of silicon lattice atoms in scaled MOS transistors. Due to the impact ionization process, electron-hole pairs are generated. Among these pairs, the electrons are collected by the drain which increases the drain current. On the other hand, the holes are pushed toward the source, which in turn are directed toward the substrate due to the action of the vertical electric field. This creates a parasitic substrate current I_{sub}. It may be noted that this mode of carrier generation can lead to an avalanche breakdown. Fortunately, this does not happen because the generated holes move rapidly into a lower field region. The substrate current generation due to impact ionization process is schematically illustrated in Fig. 3.25. The critical parameter responsible for hot carrier effects is the maximum electric field ξ_m. From (3.176), it is observed that hot carrier effects are more significant at high drain voltage V_{DS}, short channel length (because of small V_{DSsat}), thin oxide thickness t_{ox} and shallow junction depth x_j. The latter two result in lower value of the characteristic length l_t.

The probability that an electron will generate an electron-hole pair per unit length by impact ionization process while traveling through the channel is determined by the ionization coefficient $\alpha_i(y)$. This is a strong function of the electron energy and consequently depends on the lateral electric field because the energy necessary for an ionizing collision is imparted to the electron by

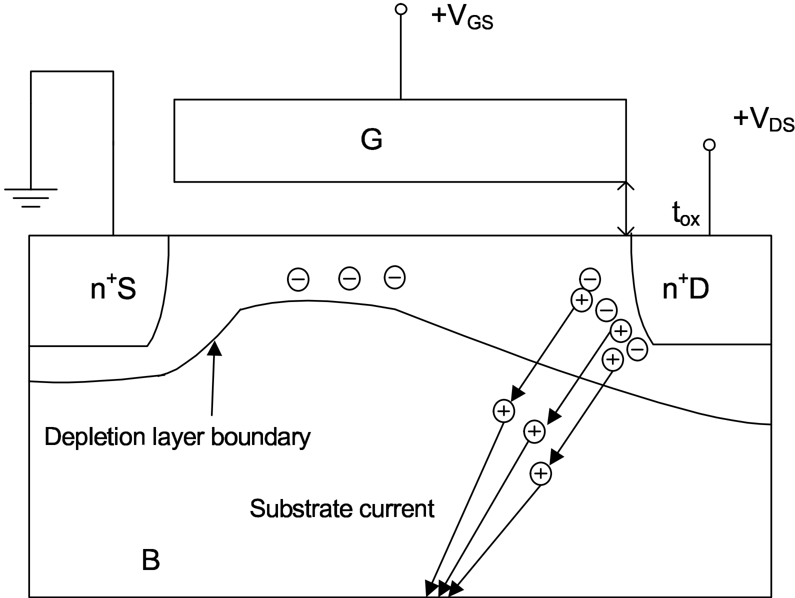


FIGURE 3.25
Illustration of substrate current due to hot carriers.

the field. The ionization coefficient is empirically given as

$$\alpha_i(y) = A_i \exp\left(-\frac{B_i}{\xi_y}\right) \tag{3.178}$$

where A_i and B_i are ionization parameters. The parameter B_i is greater for impact ionization for holes compared to that for electrons. This makes hole-induced electron-hole pair generation less significant. Thus the substrate current is given by

$$I_{sub} = I_{DS} \int_0^{\Delta L} \alpha_i(y)dy = A_i I_{DS} \int_0^{\Delta L} \exp\left(\frac{-B_i}{\xi_y}\right) dy \tag{3.179}$$

This can be written as

$$I_{sub} = -\int_{\xi_{sat}}^{\xi_m} I_{DS} A_i \exp\left[-\frac{B_i}{\xi(y)}\right] \xi^2(y) \frac{dy}{d\xi} d\left(\frac{1}{\xi}\right) \tag{3.180}$$

The exponential relationship between the lateral field ξ_y and the lateral channel distance is given by (3.173) and approximating $\cosh(y/l_t)$ by $e^{(y/l_t)}/2$ we have $\xi^2(y)\dfrac{dy}{d\xi} \approx l_t \xi$. Evaluating the integral (3.180) at ξ_m and taking it to be

constant we have

$$\begin{aligned}
I_{sub} &= A_i l_t \xi_m I_{DS} \int_{\xi_{sat}}^{\xi_m} \exp\left[-\frac{B_i}{\xi(y)}\right] d\left(\frac{1}{\xi}\right) \\
&= \frac{A_i}{B_i} l_t \xi_m I_{DS} \exp\left(-\frac{B_i}{\xi}\right)\Bigg|_{\xi_{sat}}^{\xi_m} \\
&\approx \frac{A_i}{B_i} l_t \xi_m I_{DS} \exp\left(-\frac{B_i}{\xi_m}\right) \qquad (3.181)
\end{aligned}$$

With proper substitutions, the substrate current due to hot carrier effects is given by

$$I_{sub} \approx \frac{A_i}{B_i}\left(V_{DS} - V_{Dssat}\right) I_{DS} \exp\left(-\frac{l_t B_i}{V_{DS} - V_{Dssat}}\right) \qquad (3.182)$$

This has been used widely to calculate the substrate current in scaled MOS transistors. The following model for substrate current is used in BSIM compact model

$$I_{sub} = \left(\alpha_1 + \frac{\alpha_0}{L}\right)\left(V_{DS} - V_{DSsat}\right) \exp\left(-\frac{\beta_0}{V_{DS} - V_{DSsat}}\right) I_{DS} \qquad (3.183)$$

where β_0 represents the product of B_i and l_t in (3.183), α_1 represents the A_i/B_i term in (3.183) and α_0/L is an empirical term to make the dependence of I_{sub} on effective channel length L more accurate. The term α_0 may be set to zero if the substrate current model gives correct scaleability without this term.

The substrate current causes an IR potential drop in the substrate. This leads to reverse body bias which causes the threshold voltage to drop. This triggers a positive feedback effect which further enhances the drain current. The substrate current induced body bias effect (SCBE) results in a current increase which is much larger than I_{sub} itself.

3.13.3 Gate Current Due to Hot-Carrier Effects: Lucky Electron Model

According to the lucky electron model for gate current, the electrons gain enough energy from the lateral field (without suffering any energy stripping collision within the channel) to surmount the potential barrier at the oxide-silicon interface and are then redirected toward the oxide-silicon interface by acoustic phonon scattering. Subsequently the electrons are swept toward the gate electrode by the aiding field in the oxide, provided that the gate is at a higher potential than the surface potential [132]. The trajectory of the electron in the energy diagram is shown in Fig. 3.26.

A "lucky electron" gains sufficient energy by traveling a long distance without suffering a collision. If the electron travels a distance d without suffering

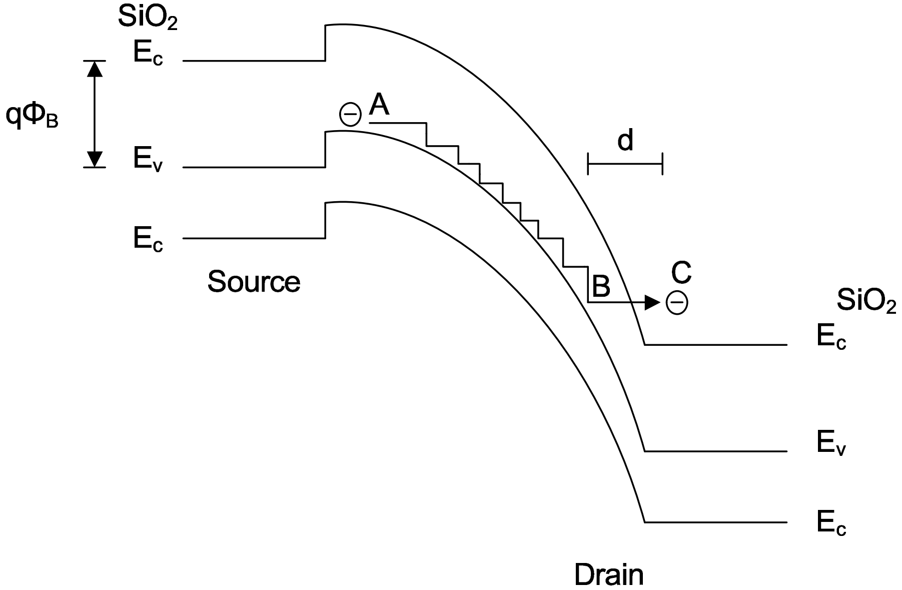


FIGURE 3.26
Trajectory of lucky electron.

any energy stripping collision, it would possess sufficient energy to enter the oxide region. The probability of this event is given by $\exp(-d/\lambda_m)$ where λ_m is the scattering mean free path of the hot electron, the dominant scattering mechanism being the optical phonon scattering. For a hot electron to surmount the silicon-oxide interface potential barrier of height $q\phi_B$, its kinetic energy must be greater than $q\phi_B$. For simplicity, d is taken to be equal to be (ϕ_B/ξ_y). Therefore, the probability that an electron acquires enough kinetic energy to surmount the potential barrier is $\exp\left[-\phi_B/(\xi_y\lambda_m)\right]$. The probability P of injection and collection of electrons is found to be a function of the oxide field ξ_{ox}. The gate current I_G is given by the product of three terms: (i) number of carriers (represented by the drain current), (ii) the probability of gaining energy, i.e., $\exp\left[-\phi_B/(\xi_y\lambda_m)\right]$ and (iii) probability P of injection and collection. Integrating this over the channel we have

$$I_G = I_{DS}\int_0^L \exp\left(-\frac{\phi_B}{\xi_y\lambda_m}\right) P(\xi_{ox})dy \tag{3.184}$$

This is approximated as

$$I_G = C I_{DS}\exp\left(-\frac{\phi_B}{\lambda_m\xi_m}\right) \tag{3.185}$$

where C is a constant $\approx 2\times 10^{-3}$ when $V_{GS} > V_{DS}$.

It may be noted that there is no phenomenon like hot hole injection for p-channel MOS transistors. This is because for p-channel MOS transistors the barrier height for hole injection into the oxide is very large and the mean free path for holes is also large compared to those for electrons. The gate current in a p-channel MOS transistor arises from electron injection into the oxide, with similar mechanism as that for an n-channel MOS transistor. For p-channel MOS transistors, the electrons are generated by the impact ionization process.

It has been found that the gate current is much smaller than the substrate current and is largest when the gate voltage and the drain voltage are approximately equal. The substrate current is larger and reaches maximum value at lower gate voltage than does the gate current.

3.13.4 Reduction of Drain Field through LDD Structure.

The most straightforward way to reduce the maximum electric field in the drain is to use low supply voltage. However, reduction of supply voltage means lowering of applied gate voltage and consequently lowering of saturation drain current I_{DSsat}, circuit speed and degradation of the overall performances of the system. Therefore, the alternative solution is to reduce the maximum electric field ξ_m by designing the transistor structure such that the excess drain voltage $V_{DS} - V_{DSsat}$ is not dropped across the velocity saturation region. This is achieved by fabricating a lightly doped (n^-) buffer region between the heavily doped drain and the channel. The resultant structure is referred to as the lightly doped drain (LDD) structure and is shown in Fig. 3.7. The fabrication of LDD structure consists of (i) patterning and etching of poly-silicon gate material, (ii) implantation of moderate amount of n-type dopant to form the source and drain regions, using the edge of the polysilicon to position one side of the implanted region, (iii) growing of an oxide sidewall space to cover the n^{-1} regions adjacent to the edges of the gate and finally (iv) fabrication of heavily doped source and drain regions through normal procedure. The lightly doped regions under the spacer are used to drop the excess the drain voltage. The doping of the LDD region has to be done in a controlled manner. If the doping is too low, the series resistance contributed by this structure will be excessive which limits the circuit performances. On the other hand, if the doping is too high, the purpose of adding the structure will not be served.

3.14 Source-Drain Resistance Model

It follows from (3.123) that for minimizing the short-channel effect the source/drain junction depth must be reduced. Therefore, additional process-

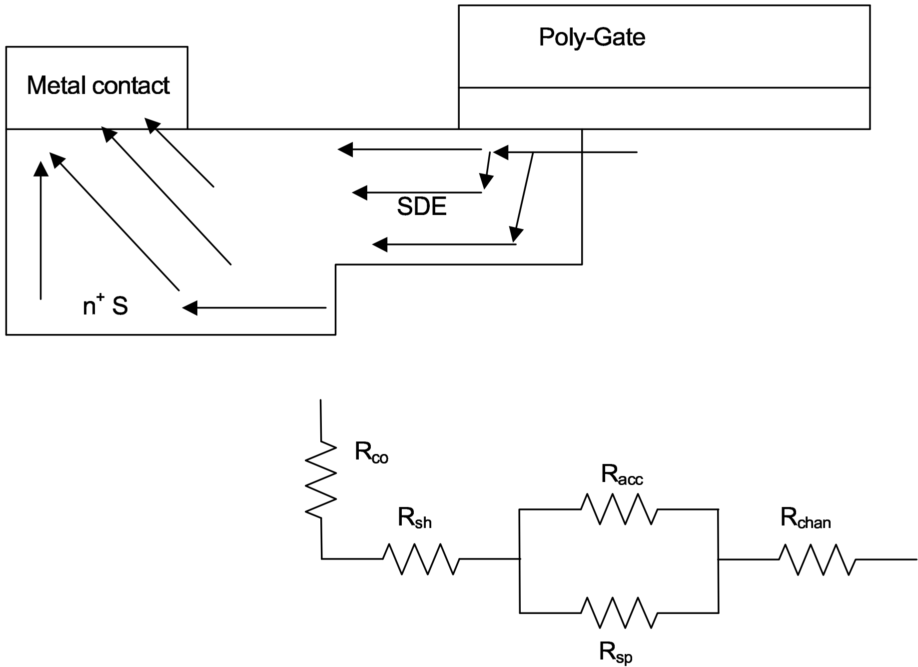


FIGURE 3.27
Schematic illustration of the current flow in the source-drain extension and source diffusion region leading to various parasitic components of the source resistance.

ing steps are performed to produce the shallow source-drain junction extension (SDE) between the deep junction and the channel. This also reduces the drain field as discussed in the previous section. The doping concentration in the shallow source-drain extensions are kept much lower compared to the doping density in the deep source/drain. The combined effect of shallow junction and light doping leads to an undesirable parasitic source-drain resistance. The various components of the external resistance of a MOS transistor are shown in Fig. 3.27. The physical origins of these components are as follows [1, 192].

1. The current flows through the inversion layer of the SDE overlap region, leading to the component R_{acc}.
2. There are two possible paths for the current, one through the surface of the SDE region and another through the bulk of the SDE region; the latter path leads to a spreading resistance R_{sp}
3. The resistance R_{sh} due to the sheet resistance of the SDE extension region where the current flows uniformly. The sheet resistance is determined by the SDE doping concentration and SDE junction depth.

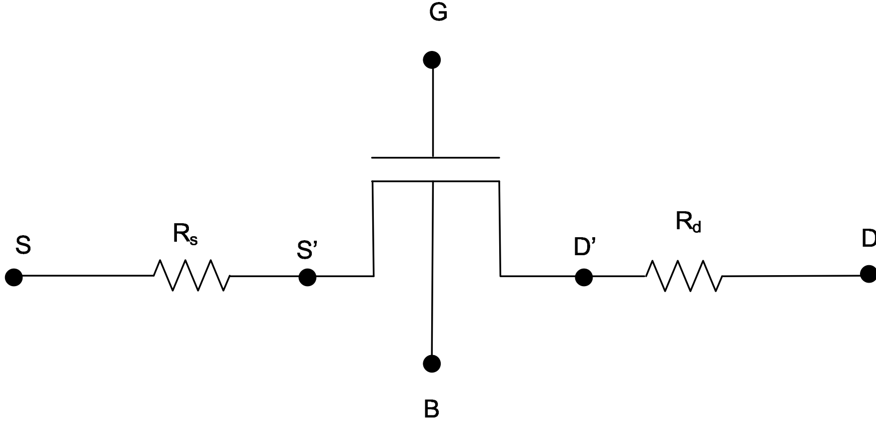


FIGURE 3.28
MOSFET representation including parasitic source and drain resistances.

4. The contact resistance R_{co} between metal/silicide and the SDE region.

The equivalent parasitic resistance thus can be written as [1]

$$R_s = (R_{co} + R_{sh}) + \left[\frac{R_{acc} \times R_{sp}}{R_{acc} + R_{sp}}\right] \tag{3.186}$$

The MOS transistor with parasitic resistances is represented in Fig. 3.28 where S and D are the extrinsic terminals and S' and D' are the intrinsic terminals, which are however, not accessible externally. This parasitic resistance critically affects the performances of a MOS transistor. With the scaling down of channel length, the parasitic resistance does not scale down proportionately.

3.14.1 Compact Modeling

The approach for the compact modeling of the parasitic source-drain resistance as used by BSIM is as follows. In the linear region, the drain current in the presence of the source-drain resistance is given by

$$I_{DS} = \frac{V_{DS}}{R_{chan} + R_{ds}} \tag{3.187}$$

where $R_{ds} = R_s + R_d$ is the total source-drain resistance and R_{chan} is the channel resistance. The channel resistance is given by $R_{chan} = V_{DS}/I_{DS0}$ where I_{DS0} is the drain current expression without considering the R_{ds} effect. Therefore, the drain current in presence of R_{ds} is given by

$$I_{DS} = \frac{I_{DS0}}{1 + \dfrac{R_{ds} I_{DS0}}{V_{DS}}} \tag{3.188}$$

The source-drain resistance R_{ds} is expressed in terms of the effective channel width W_{eff} as [30]

$$R_{ds} = R_{ds0} + \frac{R_{dsw}}{W_{eff}} \tag{3.189}$$

where R_{ds0} is a width independent component and R_{dsw} is the resistance per unit width. The effective channel width W_{eff} is given by [30]

$$W_{eff} = W_{drawn} - 2\Delta W = W_{drawn} - 2\left(\Delta W' + \Delta W_b\right) \tag{3.190}$$

Here W_{drawn} is the channel width specified by the circuit designer during the simulation procedure, ΔW_b is the change of the channel width caused by biases and $\Delta W'$ is the change of width caused due to process technology such as lithography, etc., and diffusion. A simple model of the bias dependent channel width variation is given by [30]

$$\Delta W_b = A\left(V_{GS} - V_T\right) = A\left[V_{GS} - V_{T0} - \gamma\left(\sqrt{\psi_s - V_{BS}} - \sqrt{\psi_s}\right)\right] \tag{3.191}$$

where A is a constant and γ is the body-effect parameter. Substituting (3.190) and (3.191) in (3.189), the drain-source resistance is given by

$$R_{ds} = R_{ds0} + \frac{R_{dsw}}{W_{drawn} - 2\Delta W' - 2A\left[V_{GS} - V_{T0} - \gamma\left(\sqrt{\psi_s - V_{BS}} - \sqrt{\psi_s}\right)\right]} \tag{3.192}$$

A more convenient form of (3.191) for circuit simulation purpose is given by

$$R_{ds} = R_{ds0} + \frac{R_{dsw}\left[1 + A\left(V_{GS} - V_T\right) - B\left(\sqrt{\psi_s - V_{BS}} - \sqrt{\psi_s}\right)\right]}{W'_{eff}} \tag{3.193}$$

where W'_{eff} is the effective channel width without the bias dependence. A and B are the two fitting parameters. This is the form used by BSIM3 compact model, which is written as

$$R_{ds} = \frac{Rdsw\left[1 + P_{RWG}\left(V_{GS} - V_T\right) + P_{RWB}\left(\sqrt{\psi_s - V_{BS}} - \sqrt{\psi_s}\right)\right]}{\left(10^6 W'_{eff}\right)^{Wr}} \tag{3.194}$$

Due to the parasitic drain-source resistance, the drain-to-source saturation voltage V_{DSsat} is enhanced. This is as follows [85]:

$$V_{DSsat} = V_{DSsat0} + I_{DSsat}(R_s + R_d) \tag{3.195}$$

where V_{DSsat0} is the V_{DSsat} in absence of R_s and R_d.

3.14.2 Salicide Technology

The sheet resistance component, i.e., R_{sh} and the contact resistance component R_{co} are significantly reduced in advanced CMOS technologies with

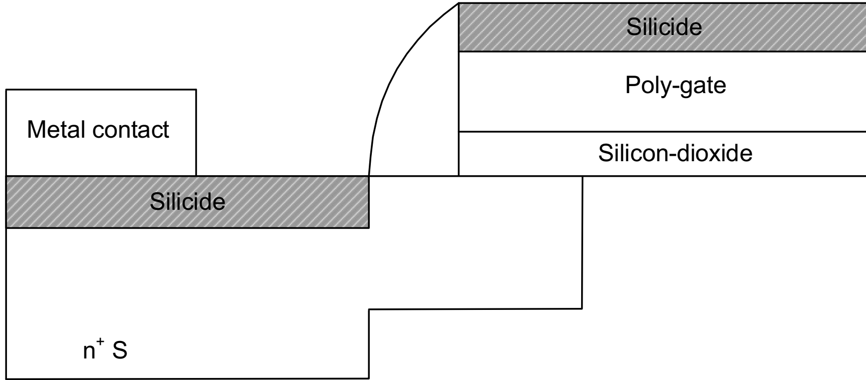


FIGURE 3.29
Self-aligned source/drain contact.

self-aligned silicide. The self-aligned silicide process is popularly coined as "salicide". The salicide process is described as follows [189]. After the gate definition, a dielectric spacer is formed on the sides of the gate. The selected metal layer for silicidation is deposited uniformly over the gate as well as over the source/drain regions. Upon heating at temperatures ($\approx 450^0 C$), the metal reacts with the exposed silicon, while no reaction occurs where the metal is over the dielectric. The unreacted metal is then removed by wet etching in a solution that removes the metal, but not its silicide. A further heat treatment is often done which converts the silicide to its low resistivity form while no conducting material remains over the spacer regions. As shown schematically in Fig. 3.29, a highly conductive ($\approx 2 - 10\Omega/\Box$) silicide film is formed over the gate and the source-drain surfaces separated by a dielectric spacer. Examples for silicides are $CoSi_2, NiSi_2, TiSi_2$ and $PtSi_2$.

For salicided transistors, the only significant contribution to R_{sh} is from the nonsilicided region under the dielectric spacer.

3.15 Physical Model for Output Resistance

For analog circuit applications, the output resistance of a MOS transistor plays a significant role in determining the intrinsic voltage gain of the transistor [70]. For long-channel MOS transistors, the output resistance is governed by the channel length modulation (CLM) phenomenon only. However, for scaled MOS transistors, the output resistance depends on several other phenomenon like DIBL and substrate current body effect (SCBE) [87]. For scaled MOS transistors, the value of the output resistance is observed to be much lower

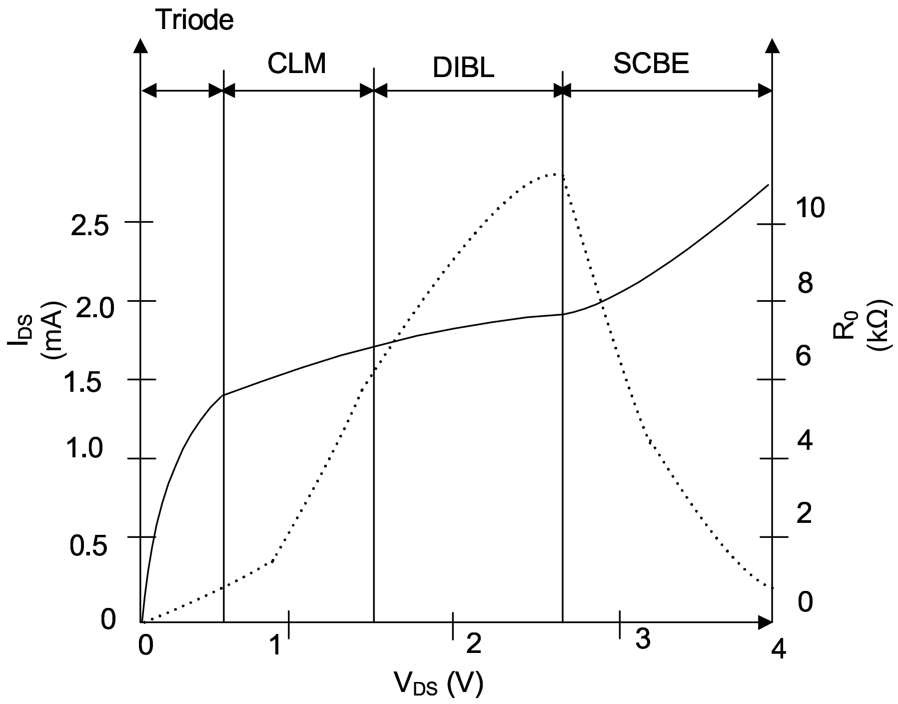


FIGURE 3.30
Schematic illustration of the typical drain current and output resistance

compared to that of long-channel MOS transistors. Therefore, the intrinsic voltage gain of a scaled MOS transistor is much lower compared to that of the long-channel MOS transistor. The schematic illustration of the drain current and output resistance variation of a MOS transistor is shown in Fig. 3.30 [30]. It is observed that the output resistance curve is divided into four regions, each region being governed by a physical mechanism: (i) triode region, (ii) channel length modulation region, (iii) drain induced barrier lowering region and (iv) substrate current body effect region.

3.15.1 Compact Modeling

In general, it can be assumed that the drain current in the saturation region is a weak function of the drain bias. Therefore, a first order Taylor series expansion for the saturation region yields

$$\begin{aligned} I_{DS}(V_{GS}, V_{DS}) &= I_{DS}(V_{GS}, V_{DSsat}) + \frac{\partial I_{DS}(V_{GS}, V_{DS})}{\partial V_{DS}}(V_{DS} - V_{DSsat}) \\ &= I_{DSsat} + \frac{\partial I_{DS}(V_{GS}, V_{DS})}{\partial V_{DS}}(V_{DS} - V_{DSsat}) \\ &= I_{DSsat}\left(1 + \frac{V_{DS} - V_{DSsat}}{V_A}\right) \end{aligned} \tag{3.196}$$

Here

$$I_{DSsat} = W v_{sat} C_{ox} \left(V_{GS} - V_T - \alpha V_{DSsat}\right) \tag{3.197}$$

$$V_A = I_{DSsat}\left(\frac{\partial I_{DS}}{\partial V_{DS}}\right)^{-1} \tag{3.198}$$

where V_A is called the Early voltage. V_A has three components, i.e., V_{ACLM}, V_{ADIBL} and V_{ASCBE}, corresponding to CLM, DIBL and SCBE respectively. Each component is to be computed separately such that

$$\frac{1}{V_A} = \frac{1}{V_{ACLM}} + \frac{1}{V_{ADIBL}} + \frac{1}{V_{ASCBE}} \tag{3.199}$$

If channel length modulation is the only physical mechanism to be taken into account, then the corresponding Early voltage is computed as

$$V_{ACLM} = I_{DSsat}\left(\frac{\partial I_{DS}}{\partial L}\frac{\partial L}{\partial V_{DS}}\right)^{-1} = \frac{\alpha \xi_{sat} L + (V_{GS} - V_T)}{\alpha \xi_{sat}}\left(\frac{\partial \Delta L}{\partial V_{DS}}\right)^{-1} \tag{3.200}$$

where ΔL is the length of the velocity saturation region, as discussed earlier. Substituting appropriate expressions and with little simplifications, the following model is derived

$$V_{ACLM} = \frac{\alpha \xi_{sat} L + (V_{GS} - V_T)}{\alpha \xi_{sat} l_t}\left(V_{DS} - V_{DSsat}\right) \tag{3.201}$$

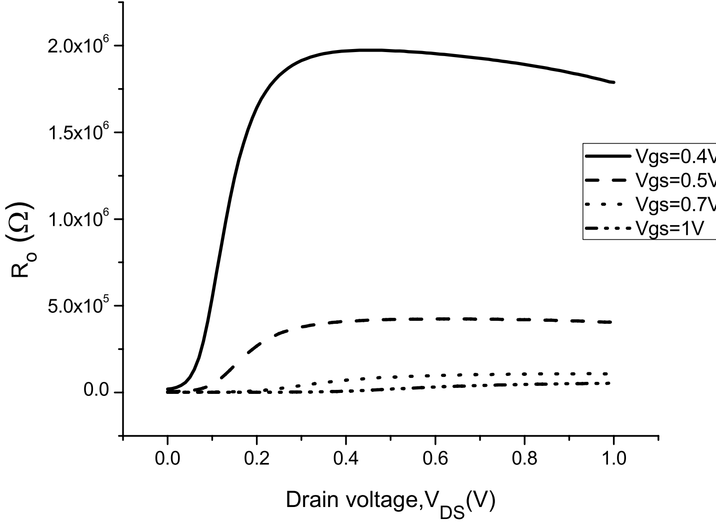


FIGURE 3.31
Simulation results showing the variation of output resistance.

BSIM introduces a fitting parameter P_{CLM} to compensate for any error occurring due to uncertainties in calculating l_t. Therefore,

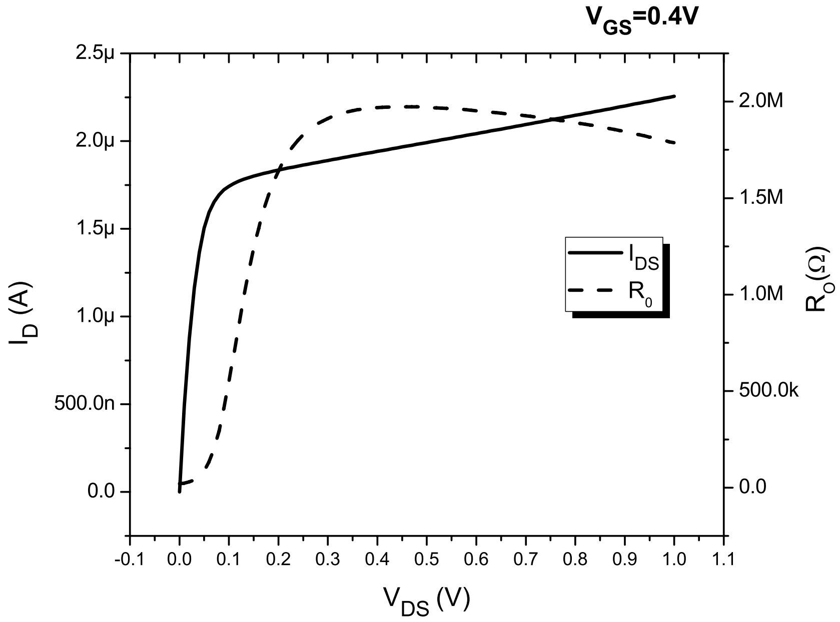
$$V_{ACLM} = \frac{1}{P_{CLM}} \frac{\alpha \xi_{sat} L + (V_{GS} - V_T)}{\alpha \xi_{sat} l_t} (V_{DS} - V_{DSsat}) \tag{3.202}$$

The Early voltage for the DIBL effect may be calculated as

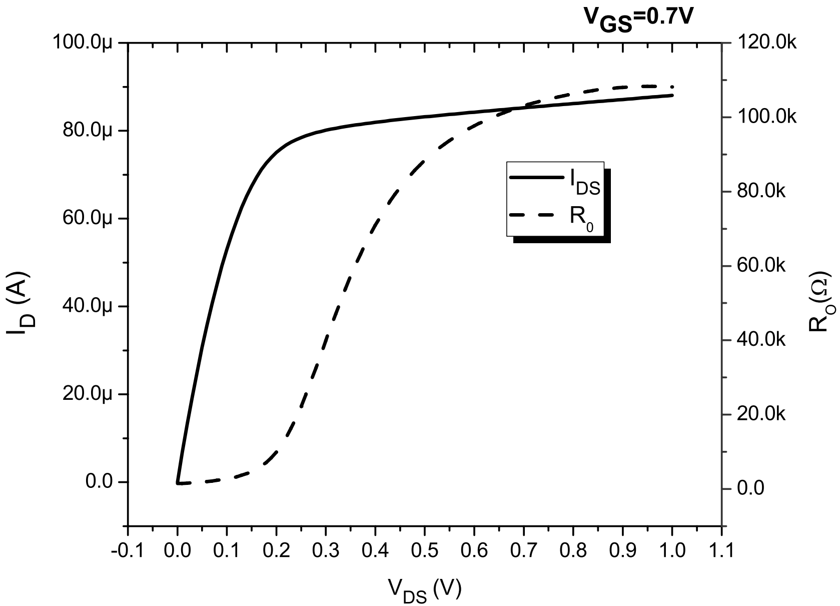
$$V_{ADIBL} = I_{DSsat} \left(\frac{\partial I_{DS}}{\partial V_T} \frac{\partial V_T}{\partial V_{DS}} \right)^{-1} \tag{3.203}$$

The analytical model becomes complicated and hence is not given here. Rather it is suggested to compute the various values of the Early voltage from the slope of the individual regions of the output resistance versus drain bias curve and then combine the individual V_A values through (3.199). It may be noted that with reduced supply voltage, the SCBE phenomenon is not that much dominant. The primary cause for the degradation of output resistance in a scaled MOS transistor is therefore the DIBL phenomenon apart from the standard CLM phenomenon.

The variations of the output resistance with drain bias for different gate biases are shown in Fig. 3.31. It is observed that the magnitude of the output resistance is higher for low gate bias compared to that for high gate bias. This is easy to justify because for low gate bias, the drain current is small compared to that at high gate bias. The degradation of the output resistance is much higher for low gate bias compared to that at high gate bias. This is illustrated



(a) $V_{GS} = 0.4V$



(b) $V_{GS} = 0.7V$

FIGURE 3.32
Variation of output resistance with drain bias for $V_{GS} = 0.4V$ and $V_{GS} = 0.7V$.

more clearly in Fig. 3.32(a) and Fig.3.32(b) respectively. This is because the increase of the drain current flow due to the barrier lowering by drain bias is higher for low gate bias, compared to that at the high gate bias.

3.16 Poly-Silicon Gate Depletion Effect

The conventional gate material that is used in the bulk CMOS technology is poly-silicon. The primary advantage of using poly-silicon gate material is compatibility with silicon processing and the capability to withstand the high temperature anneal procedure that is required after self-aligned source/drain implantation. Another significant factor is that the work function of the material can be varied by doping it into $n-$type and $p-$type. On the other hand, an important limitation of using poly-silicon as the gate material is the finite gate depletion width at the oxide interface. This reduces the effective gate capacitance and becomes more severe with thinner oxide. This is referred to as the poly-silicon gate depletion effect [1, 85]. This is becoming significant in nano-scale CMOS technology. In addition, the relatively high resistance of the poly-silicon gate critically affects the high-frequency performances of the transistor. The poly-depletion effect is illustrated in Fig. 3.33.

Suppose the doping concentration of the poly-silicon gate is N_p, the potential drop across depletion region is ψ_p and the thickness of the depletion region in the poly-silicon gate is W_p. The thickness W_p is related to the potential ψ_p

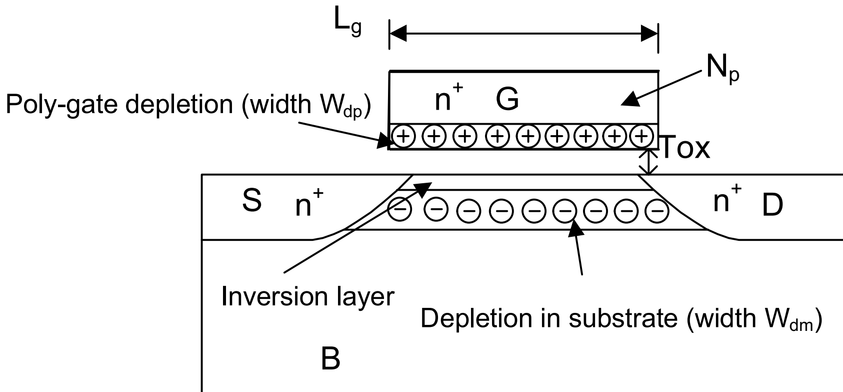


FIGURE 3.33
Illustration of poly-silicon gate depletion effect.

through the standard depletion region model

$$W_p = \sqrt{\frac{2\epsilon_{Si}\psi_p}{qN_p}} \tag{3.204}$$

The normal component of electric displacement is continuous across the interface. This is written as

$$\epsilon_{ox}\xi_{ox} = \epsilon_{Si}\xi_p = qN_pW_p \tag{3.205}$$

where ξ_{ox} is the electric field in the gate oxide and ξ_p is the electric field at the poly-gate/oxide interface. Therefore, the poly-depletion width W_p is written as

$$W_p = \frac{\epsilon_{ox}\xi_{ox}}{qN_p} = \frac{\epsilon_{ox}\psi_{ox}}{t_{ox}qN_p} \tag{3.206}$$

From (3.204) and (3.206), the potential drop in the oxide layer is written as

$$\psi_{ox} = \frac{t_{ox}qN_pW_p}{\epsilon_{ox}} = \frac{qN_pW_p}{C_{ox}} = \frac{\sqrt{2\epsilon_{Si}\psi_p qN_p}}{C_{ox}} = \gamma_p\sqrt{\psi_p} \tag{3.207}$$

where

$$\gamma_p = \frac{\sqrt{2q\epsilon_{Si}N_p}}{C_{ox}} \tag{3.208}$$

The potential balance equation is written as

$$V_{GS} = V_{FB} + \psi_p + \psi_{ox} + \psi_s \tag{3.209}$$

From (3.207) and (3.209), the following quadratic equation can be derived

$$(V_{GS} - V_{FB} - \psi_s - \psi_p)^2\,\frac{1}{\gamma_p^2} - \psi_p = 0 \tag{3.210}$$

Solving the quadratic equation and taking the positive root, the effective gate voltage is given by

$$V_{GS_{eff}} = V_{GS} - \psi_p = V_{FB} + \psi_s + \frac{\gamma_p^2}{2}\left(\sqrt{1 + \frac{4\,(V_{GS} - V_{FB} - \psi_s)}{\gamma_p^2}} - 1\right) \tag{3.211}$$

Substituting the value of γ_p, the effective gate voltage is written as

$$V_{GSeff} = V_{FB} + \psi_s + \frac{q\epsilon_{Si}t_{ox}^2N_p}{\epsilon_{ox}^2}\left(\sqrt{1 + \frac{2\epsilon_{ox}^2\,(V_{GS} - V_{FB} - \psi_s)}{q\epsilon_{Si}t_{ox}^2N_p}} - 1\right) \tag{3.212}$$

It may be noted that in the BSIM compact model, N_p is denoted by N_{GATE} and is considered as a model parameter [30]. The variations of effective gate voltage relative to the applied gate voltage with oxide thickness as a parameter

and the poly concentration as a parameter are shown in Fig. 3.34(a) and 3.34(b) respectively. It is observed from Fig.3.34(a) that when $t_{ox} = 1.75nm$, the reduction of effective gate voltage can be as low as by 30%. In addition, the effective gate voltage becomes much less than the applied gate voltage when the poly concentration is low.

3.16.1 Electrical Oxide Thickness

The oxide thickness value which is used for all sort of calculations and simulations is actually the electrical oxide thickness t_{oxe}, which is somewhat different from the physical oxide thickness t_{oxp}. The difference in the two may be attributed to two factors: (i) inversion-layer thickness and (ii) poly gate depletion width [85].

The majority of the analytical models assume that the inversion layer is a thin sheet of charge at the $Si-SiO_2$ interface. However, in reality the inversion charge profile is determined by the solution of the Schrödinger equation and Poisson equation. The average location or centroid of the inversion charge below the $Si-SiO_2$ interface is called the inversion layer thickness. The effect of the inversion layer thickness is that the oxide thickness is effectively increased by $t_{inv}/3$, where $\epsilon_{Si}/\epsilon_{ox} \approx 3$ [85].

Because of the poly-gate depletion effect and inversion layer thickness, the effective MOS capacitance in the strong inversion region thus becomes

$$C = \left(\frac{1}{C_{ox}} + \frac{1}{C_{poly}} + \frac{1}{C_{inv}}\right)^{-1} = \frac{\epsilon_{ox}}{t_{oxp} + W_p/3 + t_{inv}/3} = \frac{\epsilon_{ox}}{t_{oxe}} \quad (3.213)$$

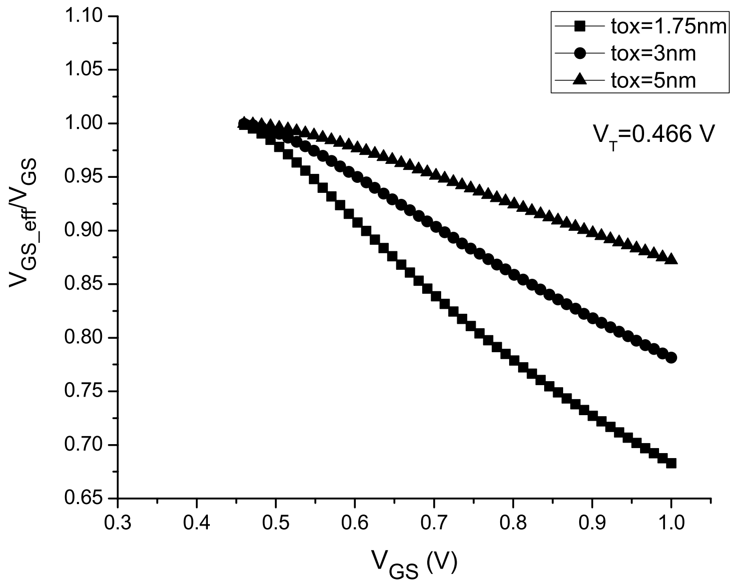
where t_{oxe} is the electrical oxide thickness. Typically t_{oxe} is larger than t_{oxp} by $6 - 10A^0$. In BSIM, this is taken care of by the terms $TOXE, TOXP$ and $DTOX$ [30].

3.16.2 Reduction of Poly-Gate Depletion

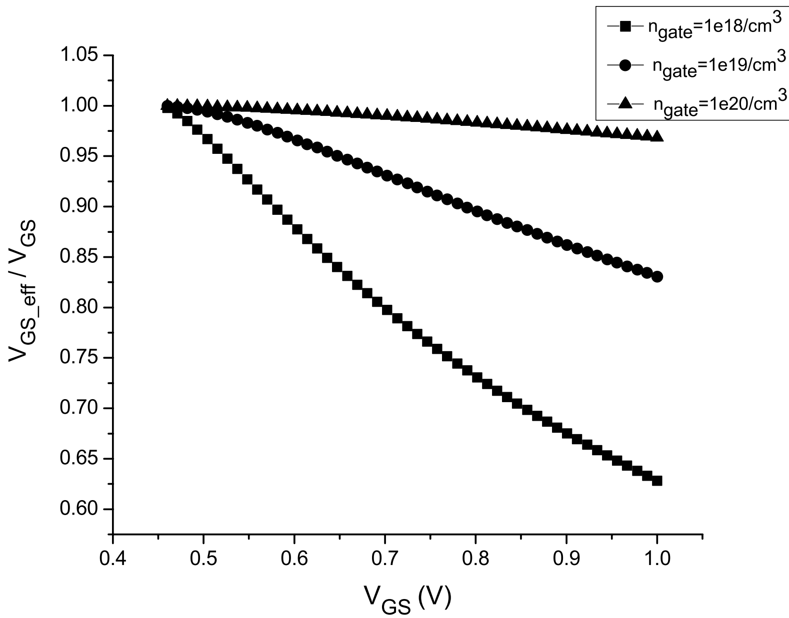
The poly-depletion effect can be reduced by doping the poly-Si gate heavily. This also reduces the gate resistance. However, too high doping sometimes cause dopant penetration from the gate through the oxide into the substrate. The poly-gate depletion effect is eliminated in the modern CMOS technology through the use of metal gates.

3.17 Effective Channel Length and Width

The concept of effective channel length is for the analog IC designers to understand and is discussed in details below.



(a) t_{ox} as parameter



(b) N_{GATE} as parameter

FIGURE 3.34
Variation of effective gate voltage with applied gate voltage.

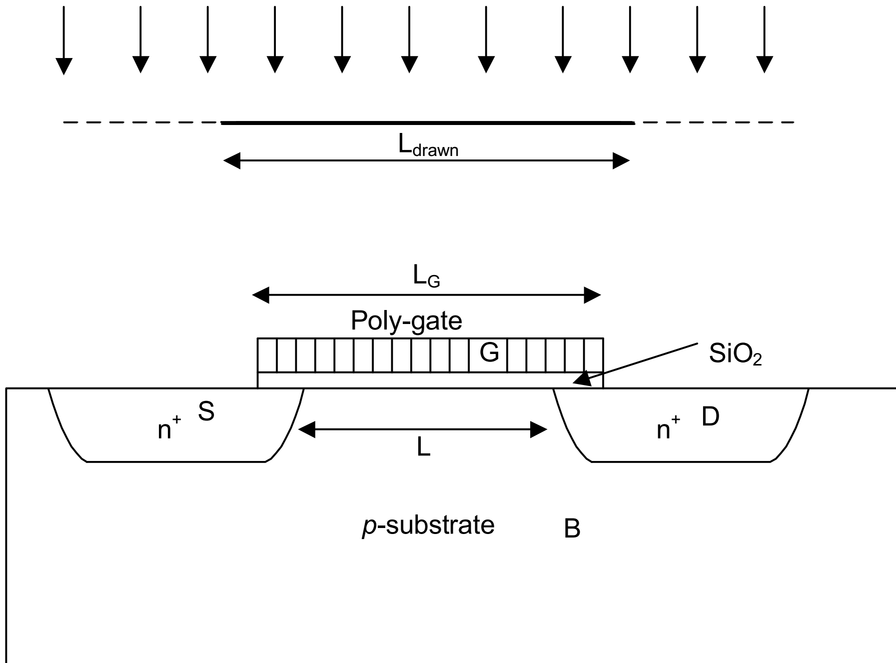


FIGURE 3.35
Schematic illustrating the definitions of drawn channel length, gate length, and effective channel length.

3.17.1 Effective Channel Length

There are altogether three types of channel length encountered in MOS transistor design and theory: (i) drawn channel length L_{drawn}, (ii) gate length L_G and (iii) effective channel length L. These are illustrated in Fig. 3.35. The channel length that is specified by the designers for simulation is the same as that considered while drawing the circuit layout. This is known as the drawn channel length L_{drawn}. This layout is transferred to a photomask. Therefore, L_{drawn} is sometime denoted as L_{mask}. After the photolithography and etching process, the physical gate length of the transistor is known as the gate length L_G. Depending on the lithography and etching biases, L_G can be either larger or shorter than L_{drawn}. For the same L_{drawn} design, L_G may vary from chip to chip, wafer to wafer and run to run. There is no simple way to measure L_G; usually it is done through scanning electron microscope. Due to the overlap between the SDE and the gate due to lateral diffusion, the effective channel length of a MOS transistor is reduced from the value of L_G. The reduced channel length is referred to as the effective channel length L. All device analysis and modeling is based upon the effective channel length L. For the designers,

it is useful to know the difference between L_{drawn} and L. This is written as

$$L = L_{drawn} - \Delta L \tag{3.214}$$

where ΔL is assumed to be constant, independent of L_{drawn}. All process related issues as well as the lateral diffusion of the SDE regions are lumped into ΔL. A very useful technique to determine the effective channel length L of a MOS transistor is the channel resistance method, which is described below. This technique is followed while extracting parameters for BSIM compact models [192, 85].

3.17.1.1 Extraction of the Effective Channel Length

In absence of the source-drain series resistance, the channel resistance is determined by the current flowing in the linear region and is given by

$$R_{chan} = \frac{V_{DS}}{I_{DS}} = \frac{L}{\mu_{ns} C_{ox} W \left(V_{GS} - V_T - \alpha V_{DS}/2\right)} \tag{3.215}$$

However, in the presence of the source-drain series resistance, the following may be written

$$\frac{V_{DS}}{I_{DS}} = R_{ds} + R_{chan} = R_{ds} + \frac{L_{drawn} - \Delta L}{\mu_{ns} C_{ox} W \left(V_{GS} - V_T - \alpha V_{DS}/2\right)} \tag{3.216}$$

Fig. 3.36 shows the variations of the measured V_{DS}/I_{DS} against L_{drawn} for three MOS transistors with different drawn lengths, otherwise identical. The drawn currents are measured at low V_{DS} and at least for two different overdrive voltages. The two straight lines intersect at a point where V_{DS}/I_{DS} is independent of $(V_{GS} - V_T)$. According to (3.216), this happens when $L_{drawn} = \Delta L$ and $V_{DS}/I_{DS} = R_{ds}$. Once ΔL is known, L can be calculated from (3.214). It may be noted that in this method, sometimes the V_{DS}/I_{DS}-vs-L_{drawn} curves may not intersect at a common point. Significant errors may result if only a limited number of $V_{GS} - V_T$ are considered.

The effective channel length as used in the drain current of BSIM model is given by [30]

$$L = L_{drawn} + XL - 2dL \tag{3.217}$$

Here the XL parameter accounts for the photolithography/etching effect and dL is given by

$$dL = LINT + \frac{LL}{L^{LLN}} + \frac{LW}{W^{LWN}} + \frac{LWL}{L^{LLN} W^{LWN}} \tag{3.218}$$

Here $LINT$ is extracted through channel resistance method discussed earlier. The rest of the parameters in (3.218) are in most cases considered to be zero.

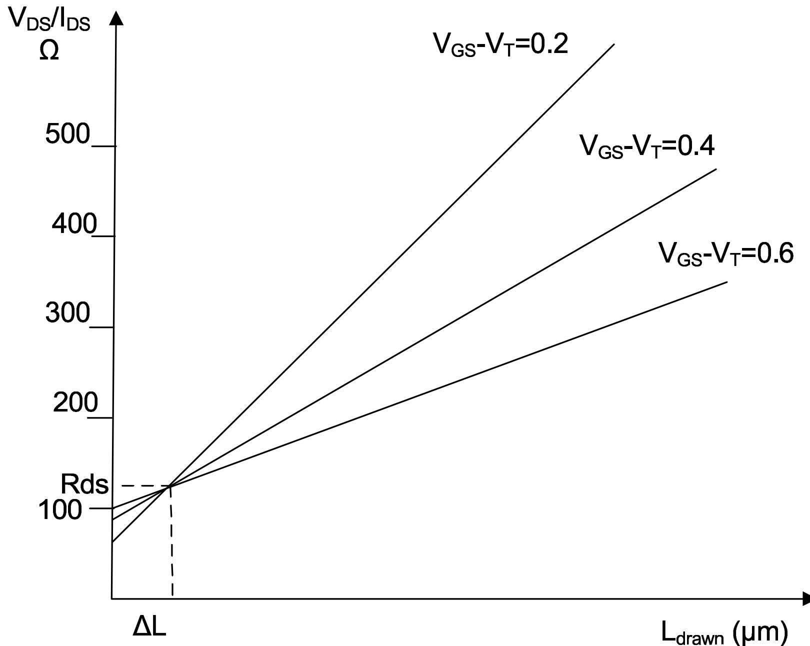


FIGURE 3.36
Channel resistance method for extraction of R_{ds} and ΔL.

3.18 Summary and Conclusion

This chapter provides a comprehensive overview of the physics-based modeling of various performances of the nano-scale MOS transistor. The importance of compact device models for VLSI circuit simulation has been emphasized. The commercial compact models that are used by SPICE simulation tools are discussed. The BSIM compact model has been considered to be the reference for discussion. The theory of the long-channel MOS transistor has been discussed in details since it provides the background for understanding the advanced theories for scaled MOS transistors. The various short channel effects like threshold voltage roll off, and DIBL phenomenon have been discussed in detail through physical concepts, mathematical modeling and exact SPICE simulation results. The drain current model of scaled MOS transistors which is very important for the VLSI designers to understand has been described in detail. The weak inversion current model has also been discussed. The effects of source-drain resistance on the drain current has been shown. The polygate depletion effect has been mentioned, explaining its importance through simulation results. This chapter therefore provides the essential background materials for VLSI designers for circuit design purposes.

4

Performance and Feasibility Model Generation Using Learning-Based Approach

4.1 Introduction

Machine learning is a branch of artificial intelligence that deals with the construction and study of systems that learn from data [77]. There are three cases of machine learning: (i) supervised, (ii) unsupervised and (iii) reinforcement learning. The problem of supervised learning involves learning a function from examples of its inputs and outputs. The problem of unsupervised learning involves learning patterns in the input where no specific output values are supplied. On the other hand, the reinforcement learning refers to the case when the learning agent learns from reinforcement.

The present chapter deals with the problem of supervised learning only. The two types of learning machines that have been considered in the present chapter are (i) artificial neural network (ANN) and (ii) least squares support vector machines (LS-SVM). In both the cases, the supervisor is the SPICE simulator. The learning problems considered are (i) the function estimation problem and (ii) the classification problem. The function estimation problem is formulated for the construction of performance models and the classification problem is formulated for the construction of feasibility models.

4.2 Requirement of Learning-Based Approaches

Before going into the details, it is necessary to understand why learning-based approaches are preferred for construction of the performance and feasibility models for nano-scale analog circuits. The conventional approaches for high-level model generation are either SPICE simulation based or through analytical equations. The former leads to accurate models but is too time consuming (CPU time) for large circuits. On the other hand, the analytical models are developed using a mixture of simplified component theory, heuristic interpretation and representation, and/or fitting of experimental data. These models are fast to evaluate but are limited in terms of accuracy, especially in the

nano-scale domain. In addition, sometimes it is also not possible to construct analytical models for certain performance parameters and design constraints. The learning-based approach is a new type of modeling approach where the model is developed by learning from accurate data of any integrated circuit or component. After training, the learning network becomes a fast and accurate model representing the original behaviors.

4.3 Regression Problem for Performance Model Generation

The regression based performance model generation problem is formally stated as follows: Given a set of m responses (or outputs) of interest, $\rho_1, \rho_2, ..., \rho_m$ and a set of n input variables, $\alpha_1, \alpha_2, ..., \alpha_n$, the objective is to determine a set of simplified empirical formulae:

$$\begin{aligned}
\hat{\rho}_1 &= \hat{\mathcal{P}}_1(\alpha_{11}, \alpha_{12}, ..., \alpha_{1n}) \\
\hat{\rho}_2 &= \hat{\mathcal{P}}_2(\alpha_{21}, \alpha_{22}, ..., \alpha_{2n}) \\
\hat{\rho}_3 &= \hat{\mathcal{P}}_3(\alpha_{31}, \alpha_{32}, ..., \alpha_{3n}) \\
\\
\hat{\rho}_m &= \hat{\mathcal{P}}_n(\alpha_{m1}, \alpha_{m2}, ..., \alpha_{mn}) \qquad (4.1)
\end{aligned}$$

In (4.1), $\mathcal{P}$ is the original function whose form is unknown and $\hat{\mathcal{P}}$ is the approximation of $\mathcal{P}$, or in other words, the regression model of $\mathcal{P}$. The more closely $\hat{\mathcal{P}}$ corresponds to $\mathcal{P}$, the more accurate is the constructed regression model. It may be noted that $\mathcal{P}$ is usually implemented through a numerical SPICE simulation technique.

4.4 Some Related Works

A fairly complete survey of the related works is given in [160]. An analog performance estimation (APE) tool for high-level synthesis of analog integrated circuits is described in [136, 48]. It takes the design parameters (e.g., transistor sizes, biasing) of an analog circuit as inputs and determines its performance parameters (e.g., power consumption, thermal noise) along with anticipated sizes of all the circuit elements. The estimator is fast to evaluate but the accuracy of the estimated results with respect to real circuit-level simulation results is not good. This is because the performance equations are based on

simplified MOS models (SPICE level 1 equations). A power estimation model for ADC using empirical formulae is described in [109]. Although this is fast, the accuracy with respect to real simulation results under all conditions is off by orders of magnitude. The technique for generation of posynomial equation-based performance estimation models for analog circuits like opamps, multi-stage amplifiers, switch capacitor filters, etc., is described in [78, 121]. An important advantage of such a modeling approach is that the circuit sizing and specification translation process can be formulated as a geometric program, which is easy to solve through very fast techniques. However, there are several limitations of this technique. The derivation of performance equations is often a manual process, based on simple MOS equations. In addition, although many analog circuit characteristics can be cast in posynomial format, this is not true for all characteristics. For such characteristics, often an approximate representation is used. An automatic procedure for generation of posynomial models using fitting technique is described in [37, 116]. This technique overcomes several limitations of the handcrafted posynomial modeling techniques. The models are built from a set of data obtained through SPICE simulations. Therefore, full accuracy of SPICE simulation is achieved through such performance models. A neural network-based tool for automated power and area estimation is described in [145]. Circuit simulation results are used to train a neural network model, which is subsequently used as an estimator. Fairly recently, a support vector machine (SVM) has been used for modeling of performance parameters for RF and analog circuits [150, 101, 47]. In [114], SVM-optimized by GA has been used to develop a soft fault diagnosis method for analog circuits. In [15], GA and SVM have been used in conjunction with developing a feasibility model which is then used within an evolutionary computation-based optimization framework for analog circuit optimization. An artificial neural network has been used in [12] for deciding the MOSFET channel length and width for analog integrated circuits. A technique for technology independent neural network modeling for fundamental blocks of analog circuits has been discussed in [97]. ANN has been used in [203] for modeling and design of a CMOS operational amplifier circuit.

Several approaches are available in literature for identification of the feasible performance region for analog circuits. In [76], a nonlinear regression technique is used to generate feasibility macromodels for analog circuits. [181] presents two approaches for identifying the feasible performance region for analog circuits. In one approach the normal boundary intersection method is applied for computing the Pareto optimal front between a set of competing performance requirements. In another approach, polytopal approximation of the entire feasibility region is obtained using a linear model of the circuit performances and the Fourier–Motzkin elimination principle. Identification of the entire range of feasible performance values for analog circuits using support vector machine principles is described in [16]. In [46], the feasible design space is characterized hierarchically through directed interval based search space profiling.

4.5 Preliminaries on Artificial Neural Network

Artificial neural networks (ANNs) are information-processing systems that emulate biological neural networks [77]. These networks are inspired by the ability of the human brain to learn from observations and generalize by abstraction. ANNs gather their knowledge by detecting the patterns and relationships in data and learn (or are trained) through experience, not from programming. The procedure used to perform the task of learning is called the learning algorithm. Generalization refers to the task of producing reasonable outputs by the neural networks for inputs not encountered during learning.

4.5.1 Basic Components

A typical ANN structure has two types of basic components [77]: (i) processing elements and (ii) interconnections between the processing elements. The processing elements are called neurons and the interconnections between the neurons are known as links or synapses. Each link has a weight associated with it. Each neuron receives stimulus from the other neurons connected to it, processes the information, and produces an output. The neurons are typically arranged in layers. The input patterns are presented to the network via the "input layer" which contains a set of neurons referred to as input neurons. The input layer communicates to one or more "hidden layers" where the actual processing is done via a system of weighted connections. Each hidden layer consists of a set of neurons referred to as hidden neurons. The hidden layers then link to an "output layer" which finally transfers the output. The output layer consists of a set of neurons referred to as output neurons. The various components are shown in Fig. 4.1.

4.5.2 Mathematical Model of Neuron

A simple mathematical model for a neuron is shown in Fig. 4.2. A link from the neuron j to the neuron i propagates the input stimulus x_j from j to i. Each link has an associated numeric weight $\omega_{j,i}$, which determines the strength and sign of the connection. Each neuron first computes the weighted sum of the various inputs and then adds a bias term b_i (also called threshold) as follows [77]

$$in_i = \sum_{j=0}^{n} \omega_{j,i} x_j + b_i \tag{4.2}$$

Then an activation function g is applied to this activation to derive the output as follows

$$y_i = g(in_i) = g\left(\sum_{j=0}^{n} \omega_{j,i} x_j + b_i\right) \tag{4.3}$$

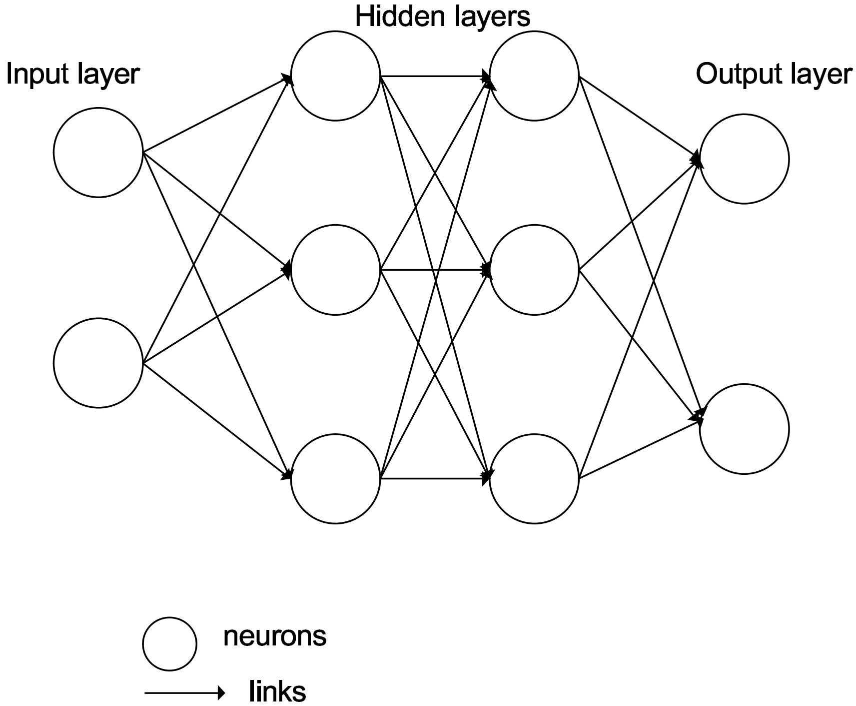


FIGURE 4.1
Basic components of an ANN structure.

The most commonly used neuron activation function is the sigmoid function given by [77]

$$g(in) = \frac{1}{(1 + e^{-in})} \tag{4.4}$$

Other functions that can also be used are the arc-tangent function, hyperbolic tangent functions, etc. These are all smooth switch functions that are bounded, continuous, monotonic and continuously differentiable. Biologically this corresponds to the firing of a neuron depending on whether the information collected from the incoming signals exceeds a certain threshold value. The input neurons use a relay activation function as there the main task is to relay the inputs to the hidden layer neurons. The activation functions for the output neurons are either logistic functions (e.g., sigmoid) or simple linear functions that compute the weighted sum of the stimuli.

4.5.3 MLP Feed-Forward NN Structure

There are two main categories of neural network structures: acyclic or feed-forward networks and cyclic or recurrent networks. In the feed-forward struc-

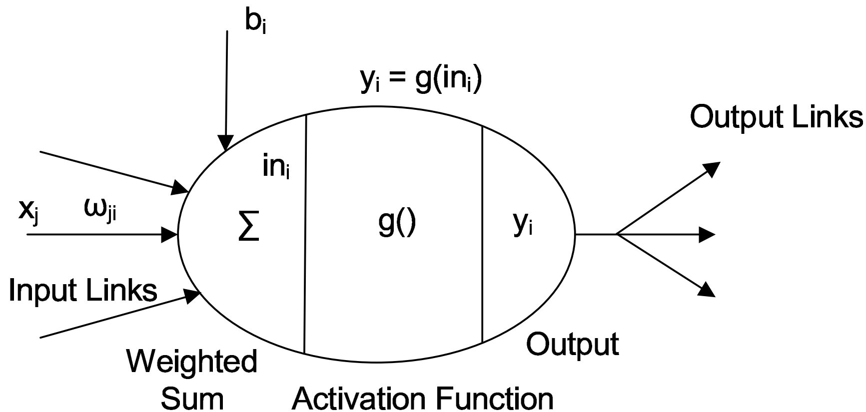


FIGURE 4.2
Mathematical model of a neuron.

ture, the structure does not have any loop. The outputs of the structure are functions of the current inputs. In the recurrent structure, the outputs are fed back to the inputs. Therefore, the dynamics of the recurrent structure may reach a stable state or exhibit oscillations or even chaotic behavior. This chapter considers the feed-forward network only.

The multilayer perceptron (MLP) is a popularly used feed-forward neural network structure. Typically the network consists of a set of source neurons that constitute the input layer, one or more hidden layers of computation neurons, and an output layer of output neurons. The input signal propagates through the network in a forward direction, on a layer-by-layer basis. For example, an MLP neural network with an input layer, one hidden layer, and an output layer, is referred to as a three-layer MLP (or MLP3). A very simple feed-forward NN structure with an input layer of two input neurons, two hidden layers of three neurons in each and one output layer with two neurons is shown in Fig. 4.1.

4.5.4 Feed-Forward Computation

In order to explain the principle of operation of a feed-forward MLP NN structure, let us consider a simple NN structure with two inputs, one hidden layer of two neurons and one output neuron as shown in Fig. 4.3. For simplicity, the bias input has been neglected. Let the two inputs be x_1 and x_2 and the output be y. Thus the activations of the input neurons are $a_1 = x_1$ and $a_2 = x_2$. The activations of the neuron $n3$ and $n4$ are a_3 and a_4 respectively. The activation of the output neuron $n5$ is $a_5 = y$. This is given as

$$\begin{aligned} y &= g\left(\omega_{3,5}a_3 + \omega_{4,5}a_4\right) \\ &= g\left(\omega_{3,5}g\left(\omega_{1,3}x_1 + \omega_{2,3}x_2\right) + \omega_{4,5}g\left(\omega_{1,4}x_1 + \omega_{2,4}x_2\right)\right) \end{aligned} \quad (4.5)$$

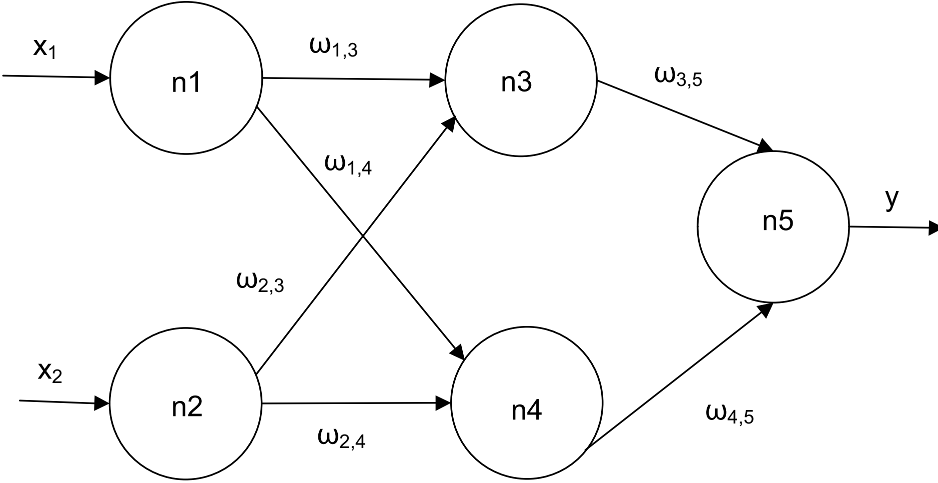


FIGURE 4.3
Simple MLP NN structure for illustration of the principle of feed forward computation.

It is observed that the neural network output is a function of the inputs of the network and the weights act as the parameters of the network function. This is the feed-forward computation of the MLP NN structure.

4.5.5 Success of MLP NN Structures

The success of the use of MLP -NN structures for construction of regression functions is attributed to the following reasons. The first is that these are universal approximators [80]. The universal approximation theorem states that there always exists a three-layer MLP NN structure that approximates any arbitrary nonlinear continuous multidimensional function to any desired accuracy. This forms the theoretical basis for using the MLP NN structures to approximate CMOS circuit behavior, which are multidimensional functions of physical/geometric and electrical bias parameters.

The second reason is that these structures are better able to cope with the *curse of dimensionality* problem. It has been shown that the NN structures can avoid the curse of dimensionality problem in the sense that the approximation error becomes independent of the dimension of the input space (under certain conditions) which is not the case for polynomial approximators.

4.5.6 Network Size and Layers

The accuracies of ANN structure depend upon the suitable number of hidden neurons. The number of hidden neurons depends upon the degree of nonlinearity of the chosen functional relationship and the dimensions of the input and the output design variable set. Approximating highly nonlinear relations requires more neurons compared to that of simpler relations. However, the universal approximation theorem does not specify anything about the size of the network. Therefore, the precise number of hidden neurons required for a given modeling task remains an open question. Most of the time this is fixed through a trial and error process. Sometimes some adaptive algorithms may also be used which dynamically add or delete neurons to reach a desirable accuracy limit for the NN structure. In general, the MLPs with one or two hidden layers (i.e., three- or four-layer MLPs) are commonly used in integrated circuit applications.

4.6 Neural Network Model Development

The various steps involved in the process of neural network model development are summarized in Fig. 4.4 [209]. The various steps are discussed in the following sub-sections.

4.6.1 Formulation of Inputs and Outputs

The first task in the development of an NN model is the identification of inputs ($\bar{\alpha}$) and outputs $\bar{\rho}$. The output design variables are determined based on the purpose of the NN model. For example, the voltage gain, bandwidth, and phase margin of an operational amplifier circuit could be the possible output design variables. The input design variables are the transistor sizes. The selection of input model variables is critical. Only the significant design variables should be considered as the model inputs. The input design variables form a hyperspace, referred to as the sample space. Selection of the sample points from a large sample space is very difficult. In addition, the complexity of creating a model with reasonable accuracy accelerates with increase in the number of input design variables. Design knowledge is to be used to reduce the dimensions of the sample space, such as transistor matching.

4.6.2 Data Range and Sample Distribution

The next task in the data generation procedure is to define the range of data and the distribution of $\bar{\alpha} - \bar{\rho}$ samples within that range. If the range of input design variables, in which the constructed NN model is to be used is $[\bar{\alpha}_l, \bar{\alpha}_u]$,

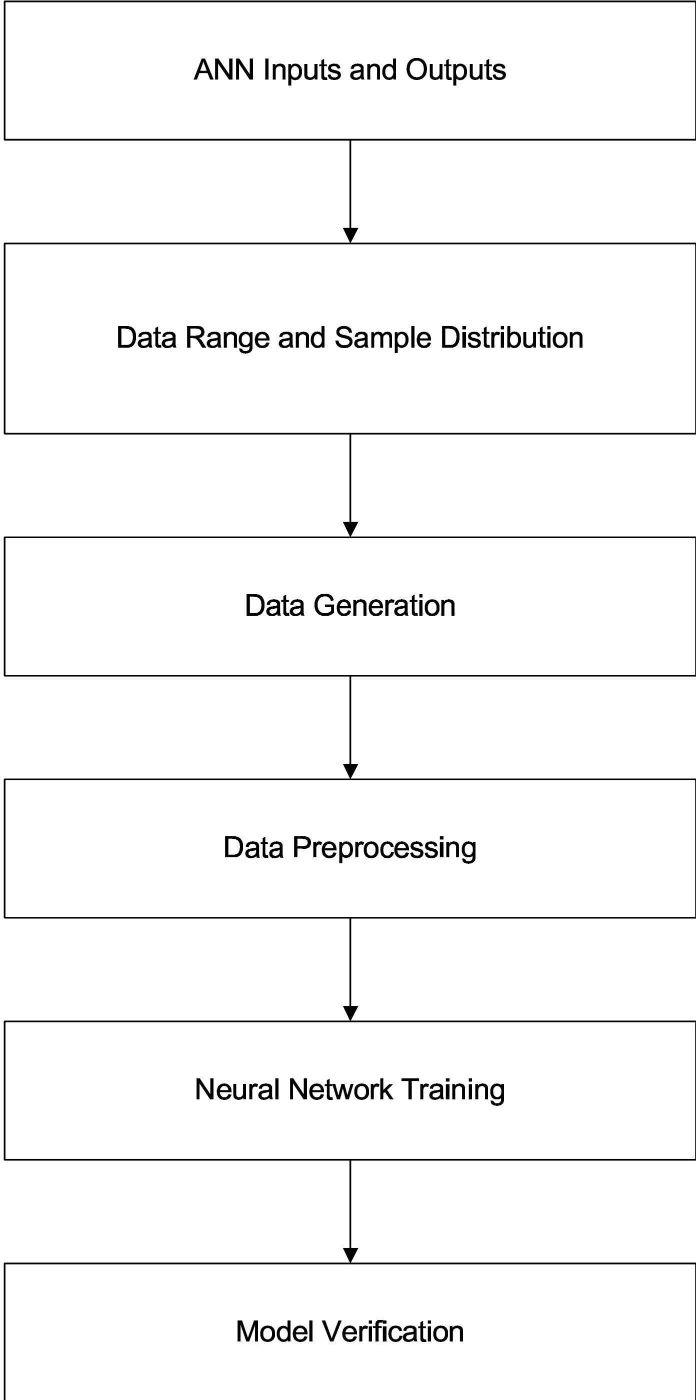


FIGURE 4.4
Steps for the development of NN model.

then the sample data should be collected slightly beyond the model utilization range, i.e., within $[\bar{\alpha}_l - \Delta, \bar{\alpha}_u + \Delta]$ in order to ensure the reliability of the NN model at the boundaries of the model utilization range. Samples, which are to be used for validation only are to be collected from the range $[\bar{\alpha}_l, \bar{\alpha}_u]$.

Once the data range is fixed, samples need to be extracted within this range. There are two approaches for the extraction of sample data: static sampling and dynamic sampling [95]. The static sampling process remains ignorant of how a sample will be used and applies some fixed criterion for the sampling process. Once the sampling is completed, the generated dataset corresponding to extracted samples is used to construct the ANN model. On the other hand, in the dynamic sampling process, the sampling process is controlled by a so-called learning machine, or learner. A static sampler does not take into account the shape of the function under scrutiny, whereas a dynamic sampler strives to reduce modeling error through strategic placement of sample points in the design space. The commonly used static sampling methods are uniform grid distribution, pseudo-random and quasi-random sampling technique, nonuniform grid distribution, design of experiments (DOE) technique [19]. In the uniform distribution, samples for each input design variable α_i are collected at equal intervals. Suppose the number of grids along the input dimension α_i is n_i. The total number of $\bar{\alpha} - \bar{\rho}$ samples is given by $\prod_{i=1}^{n} n_i$. Pseudo-random sampling is typically produced by pseudo-random number generators which are available in all programming languages. The pseudo-random number generator technique requires a seed, a number on which to base all future generated numbers. Once the seed is chosen, the procedure for the generation of numbers becomes deterministic. However, with change in the seed value, the sequence becomes altogether different. The distribution of numbers produced is usually uniform over the interval [0,1], although more specialized generators are available for producing Gaussian, or normal distributions. The sample points obtained through the pseudo-random technique are often found to be clustered in localized areas within the sample space and a small number of samples is distributed near the boundaries. The quasi-random sample generation technique, on the other hand, produces samples that are more uniformly distributed in the sample space; however, maintaining the random nature. However, the quasi-random generators lack the concept of a seed and will always produce the same sequence of numbers. An example of the quasi-random number generation technique is the Halton sequence generation technique [105].

In nonuniform grid distribution each input design variable is sampled at unequal intervals. This type of sample distribution strategy is useful when the circuit behavior is nonlinear in certain ranges of $\bar{\alpha}$ space and dense sampling is required in that specific region, e.g., the region near the cutin voltage in the dc characteristics of a p-n junction diode. For certain problems, the sample data generation procedure is an expensive one , e.g., sample data generation for modeling any semiconductor fabrication processes. For such problems, DOE-based sampling strategy has to be employed. A designed experiment is a series

of tests in which purposeful changes are made to the input variables of a process or system so that the causes of changes in the output response can be observed and identified. The commonly used DOE strategies are 2^n factorial experimental design, central composite experiment design, etc. [19].

4.6.3 Data Collection

For each input sample (transistor sizes) extracted through sampling strategy, the chosen circuit topology of a component block is simulated using SPICE. Appropriate BSIM model is to be used for simulation. For nano-scale technology, BSIM4 is to be used. This ensures that all important deep sub-micron effects of MOS transistors are considered while generating the dataset. Depending upon the selected input-output parameters of the NN model, it is necessary to construct a set of test benches that would provide sufficient data to facilitate automatic extraction of these parameters via postprocessing of SPICE simulation output files. The commonly used SPICE simulations are ac analysis, transient analysis, dc sweep etc. The voltages and currents at the various nodes of the circuit are also measured. In many cases, constraints are imposed upon the SPICE results to ensure that only feasible data are considered for model construction.

For simplicity of the following theoretical discussion and notations, let us assume that the MLP NN structure has only one output. Let the column vector $\bar{d}$ represent a set of outputs from simulation/measurement corresponding to the input vector $\bar{\alpha}$. With this the data collection process is defined as the use of SPICE simulation to collect the sample pairs $(\bar{\alpha}_k, d_k), k = 1, 2, ...S$ where S is the total number of samples. The general guideline in collecting the total number of samples is that for nonlinear high dimensional problems, large numbers of samples are required and for relatively smooth low dimensional problems, fewer samples are required.

4.6.4 Data Organization and Data Preprocessing

The total number of collected data samples $\{(\bar{\alpha}_k, d_k), k = 1, 2, ...S\}$ are divided into three sets: the training dataset, the validation dataset and the test dataset. The training data is used to train the NN structure. The NN weight parameter values are updated based on the training data. The validation dataset is used to check the quality of the trained model during the development procedure and to determine the stop criterion for the training process. On the other hand, the test dataset is used to independently assess the quality of the final constructed NN model in terms of accuracy and generalization capability. However, if the number of samples is small, then the total number of samples may be divided into two sets. One set may be used for the training and validation purpose and the other for the test purpose.

The orders of magnitude of various input and output values of integrated circuit problems are usually very different from one another. As such, a system-

atic preprocessing of construction data, referred to as data scaling, is required before constructing the model [146]. The commonly suggested scaling schemes are linear scaling, log scaling, and two-sided log scaling. The following formulae are used for linear and logarithmic scaling of the output data within an interval $[0, 1]$

$$\text{Linear} : d'_k = \frac{d_k - lb}{ub - lb} \tag{4.6}$$

$$\text{Logarithmic} : d'_k = \frac{\log\left(\frac{d_k}{lb}\right)}{\log\left(\frac{ub}{lb}\right)} \tag{4.7}$$

Here d_k is the unscaled k^{th} data of any parameter bounded within the interval $[lb, ub]$. Linear scaling of data balances the ranges of different inputs or outputs. Applying log scale to data with large variations balances large and small magnitudes of the same parameter in different regions of the model.

4.6.5 Neural Network Training

A flow chart summarizing the steps involved in the neural network training process is shown in Fig. 4.5. The neural network structure is selected first, which for the present purposes will be a MLP structure. The weight parameters are initialized. The widely used strategy is to initialize the weight parameters with small random values. During training the weights and the biases of the network are iteratively adjusted to minimize the network performance function. The training data consists of sample pairs $\{(\bar{\alpha}_k, d_k), k = 1, 2, ...T_r\}$, where T_r is the total number of training data. The network performance function is defined by the neural network training error as follows

$$E_{T_r}(\bar{\omega}) = \frac{1}{2}\sum_{k=1}^{T_r} |y_k(\bar{\alpha}_k, \bar{\omega}) - d_k|^2 \tag{4.8}$$

where d_k is the kth element of the column vector $\bar{d}$ and $y_k(\bar{\alpha}_k, \bar{\omega})$ is the kth element of the column vector $\bar{y}$ which is NN output for the input vector $\bar{\alpha}_k$. It may be noted that a single output NN structure has been considered for discussion. The various training algorithms use the gradient of the performance function for adjusting the weights in order to minimize the network performance. From the current values of the weight vector, the next weight vector is calculated as

$$\bar{\omega}_{i+1} = \bar{\omega}_i + \eta\bar{h} \tag{4.9}$$

where $\bar{h}$ denotes the direction in which the next weight vector is to be determined and η is a positive step size known as the learning rate. For example, if the next weight vector is calculated along the negative direction of the network performance, then $\bar{\omega}_{i+1} = \bar{\omega}_i - \eta\left(\partial E_{Tr}/\partial\bar{\omega}\right)$. The gradient is determined using a technique called backpropagation, which involves performing computations backwards through the network [77].

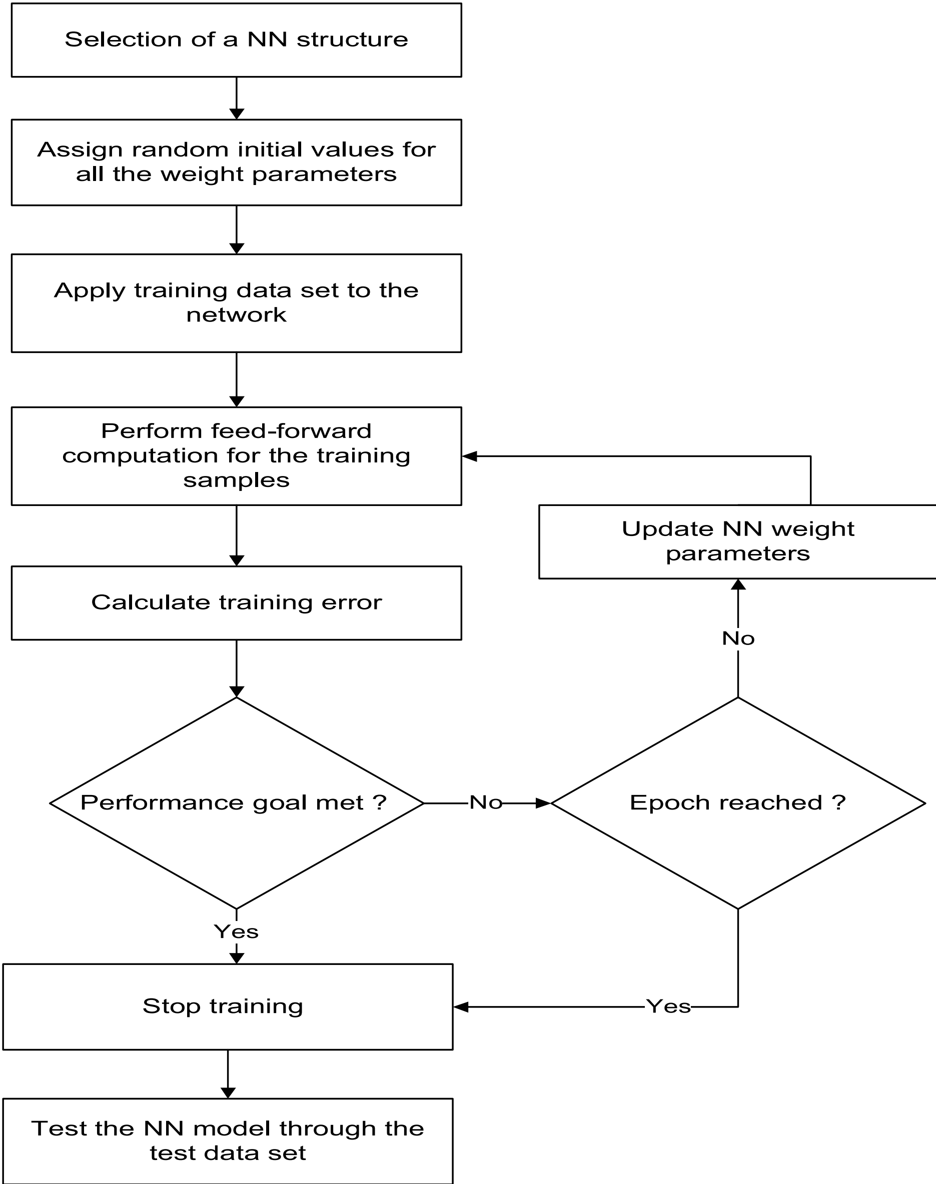


FIGURE 4.5
Flow chart illustrating the ANN training procedure.

There are two different ways in which this gradient descent algorithm can be implemented: incremental mode and batch mode. In the incremental mode, the gradient is computed and the weights are updated each time a training sample $(\bar{\alpha}_k, d_k)$ is presented to the network. In the batch mode all of the inputs are applied to the network before the weights are updated. For batch training the weights are only updated once in each epoch. The updating of the weight vector continues if the error goal is met, or if the maximum number of epochs is reached.

Apart from the backpropagation algorithm for computation of the gradients, there are several other high-performance techniques used for neural network training. These are classified into two broad categories: heuristic techniques and standard numerical optimization techniques discussed in Chapter 2. In the conjugate gradient algorithms, a search is performed along the conjugate directions, which produces generally faster convergence than the steepest descent directions. The Levenberg Marquardt algorithm is the most widely used algorithm to train the neural network and is used in the MATLAB® toolbox.

4.6.6 Quality Measures

Statistical functions are generally used to assess the quality of the generated NN model. The quality is evaluated with the independent test dataset. The average relative error (ARE) function defined as follows is one such measure.

$$ARE = \frac{1}{N_{Te}} \sum_{k=1}^{T_e} |y_k(\bar{\alpha}_k, \bar{\omega}) - d_k| \tag{4.10}$$

Here N_{Te} denotes the number of elements in the test dataset. Another commonly used measure is the correlation coefficient R. This is defined as follows:

$$R = \frac{N_{te} \sum yd - \sum y \sum d}{\sqrt{\left[N_{te} \sum y^2 - \left(\sum y\right)^2\right] \left[N_{te} \sum d^2 - \left(\sum d\right)^2\right]}} \tag{4.11}$$

The correlation coefficient is a measure of how closely the NN outputs fit with the target values. It is a number between 0 and 1. If there is no linear relationship between the estimated values and the actual targets, then the correlation coefficient is 0. If the number is equal to 1.0, then there is a perfect fit between the targets and the outputs. Thus, the higher the correlation coefficient, the better the model is.

4.6.7 Generalization Ability, Overlearning, and Underlearning

The ability of a neural network structure to estimate the output y_k accurately when the input set $\bar{\alpha}_k$ is presented to the network which has not been used

in the entire model development procedure is referred to as the generalization ability. Good learning of a NN structure is achieved when both the validation error and the test error are small ($e.g., 0.50\%$) and close to each other. However, if it happens that the training error E_{Tr} is small but the validation error is $E_V >> E_{Tr}$, the situation is referred to as the overlearning, i.e., the NN structure memorizes well but the generalization ability is not good. The possible remedies of overlearning are deletion of a certain number of hidden neurons or addition of more samples to the training data. The network is said to exhibit underlearning when in a certain epoch if the training error does not satisfy the performance goal. i.e., $E_{Tr} >> 0$. The possible remedies are addition of more hidden neurons and perturbation of the current network structure (i.e., for given weight and bias vectors) so as to escape from the trap of local minimum, and then continuation of the training procedure.

4.7 Case Study 1: Performance Modeling of CMOS Inverter

The basic circuit diagram of a CMOS inverter is shown in Fig 4.6. Let $\bar{\alpha}$ be the 3-dimensional input vector containing the circuit design variables, i.e., the channel width W_n of the NMOS transistor $M1$, the channel width W_p of the PMOS transistor $M2$ and the output load capacitor C_L. Let $\bar{\rho}$ be the 4 dimensional output vector containing the performance parameters of the design, i.e., output rise time (τ_R) and fall time (τ_F), inverter switching point (V_{SP}) and average power consumption P_{av}. Thus the inputs and outputs of the performance model are written as follows

$$\bar{\alpha} = [W_n, W_p, C_L] \tag{4.12}$$

$$\bar{\rho} = [\tau_R, \tau_F, V_{SP}, P_{av}] \tag{4.13}$$

The performance model is written as

$$\bar{\rho} = f(\bar{\alpha}) \tag{4.14}$$

This relationship between the circuit design variables and the performance parameter is generally strongly nonlinear and multi-dimensional. The function f is traditionally evaluated through SPICE simulation. The corresponding neural network model is written as

$$\hat{\bar{\rho}} = f_{ANN}(\bar{\alpha}, \bar{\omega}) \tag{4.15}$$

where f_{ANN} is the neural network, $\hat{\bar{\rho}}$ is the output vector of neural model responses, $\bar{\alpha}$ is the ANN input vector, $\bar{\omega}$ contains all the weight parameters

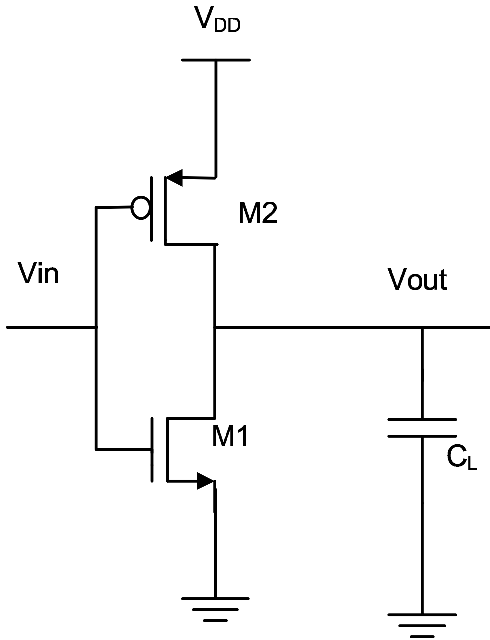


FIGURE 4.6
CMOS inverter.

required to construct the ANN structure. This case study describes the construction f_{ANN} such that it is a faithful approximation of the original function f [45].

In order to generate the training and test data, CMOS inverters are constructed corresponding to the circuit design variables listed in Table 4.1. The channel lengths of both transistors are fixed at minimum of the process technology, i.e., 45nm. The other process technology parameters are taken from the Berkeley Predictive Technology Model file [210]. Based on the Halton sequence generator, uniformly distributed samples are generated within the specified range. The training and test data corresponding to those sample points are generated through SPICE simulation using the BSIM4 model. Transient analysis and DC transfer sweep analysis are performed in order to extract the

TABLE 4.1
Range of Circuit Design Parameters

Parameters	Min	Max
$W_n(nm)$	90	1000
$W_p(nm)$	90	1000
$C_L(pF)$	1	5

TABLE 4.2
MLP NN Architecture for Case Study 1

Parameters	Optimized Values
Architecture	Feed-forward MLP
Training Algorithm	Levenberg–Marquardt
Hidden layers	2
neurons in the 1st hidden layer	18
neurons in the 2nd hidden layer	12
Hidden layer transfer function	Tan-sigmoid
Output layer transfer function	Linear
Maximum epoch	1000

performance parameters. It is observed from Table 4.1 that the input variables vary over a wide range. Similarly the output performance parameters vary over a wide range. Linear scaling of the data between 0 and 1 is used. A set of 1000 samples has been considered in the present work. Out of these, 800 samples have been taken for training purposes and the rest are taken for testing purposes.

The chosen ANN architecture is described in Table 4.2. The Levenberg–Marquardt (LM) back propagation method has been used as the training algorithm. The training goal is set to 10^{-7}. The training algorithm of the MATLAB toolbox has been used.

The neural network optimization technique is conducted for a maximum of 1000 iterations. However, after 725 epochs on an average the neural network model has reached the desired training goal. The neuron numbers in the first and second hidden layer are selected through the trial and error method. The architecture is shown in Fig.4.7. An important issue to be considered for training is to set the training goal. With too high a value of the training goal, the generalization error deteriorates, whereas for a low value, the model becomes under-learned [77]. This is illustrated with numerical results in Fig.4.8.

The training dataset has been generated through two steps. In the first step, using a Halton sequence generator, the input data are generated. The total CPU time consumption for this process is $\sim 0.5410s$. In the second step, using SPICE analysis, the output data are generated. The total CPU time consumption for this process is $\sim 364.34s$ Therefore, the time consumption for generating the training data is $\sim 364.881s$. On the other hand, the training time is found to be $\sim 412s$. The model construction time is the sum of data generation time and training time. In the present work, this comes out to be $\sim 777s$. This timing information is based on a PC with Core-2-duo processor and 2GB RAM.

Table 4.3 shows the average relative error (ARE) and the correlation coefficient R. It is observed that very good accuracy in terms of both ARE and R have been achieved. Figure 4.9(a), 4.9(b), 4.10(a) and 4.10(b) respectively

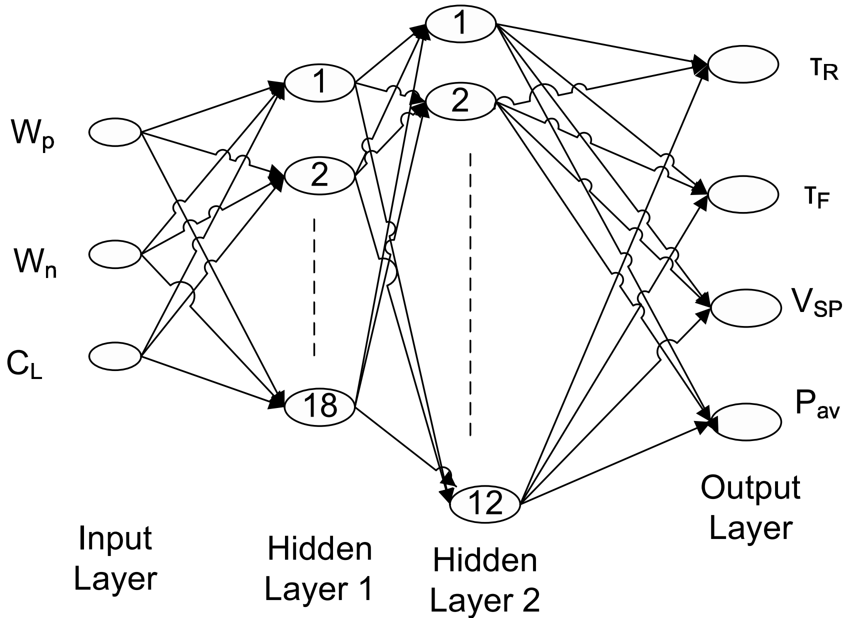


FIGURE 4.7
MLP-NN architecture for Case Study 1.

show τ_R, τ_F, V_{SP} and P_{av} for 100 designs, obtained through ANN and SPICE simulations. It is observed that all the ANN outputs show good matching with SPICE results. The scatter plots between the ANN predicted results and SPICE simulations are shown in Fig. 4.11(a), 4.11(b), 4.12(a) and 4.12(b). The scatter plots are nearly perfect diagrams with unity correlation coefficients. These demonstrate the accuracy of the constructed ANN model.

4.8 Case Study 2: Performance Modeling of Spiral Inductor

An on-chip spiral inductor is an important passive device component in high-frequency integrated circuit design. The most widely used type is the planar spiral inductor [135, 208]. The square inductor is the commonly used planar inductor because of its simple layout and fabrication process. The layout diagram of a square spiral inductor is shown in Fig. 4.13. The inductance and other electrical characteristics of a monolithic spiral inductor are determined by the following physical layout parameters (i) the outer diameter d_{out}, (ii) the number of turns n, (iii) the metal width W and (iv) the spacing between

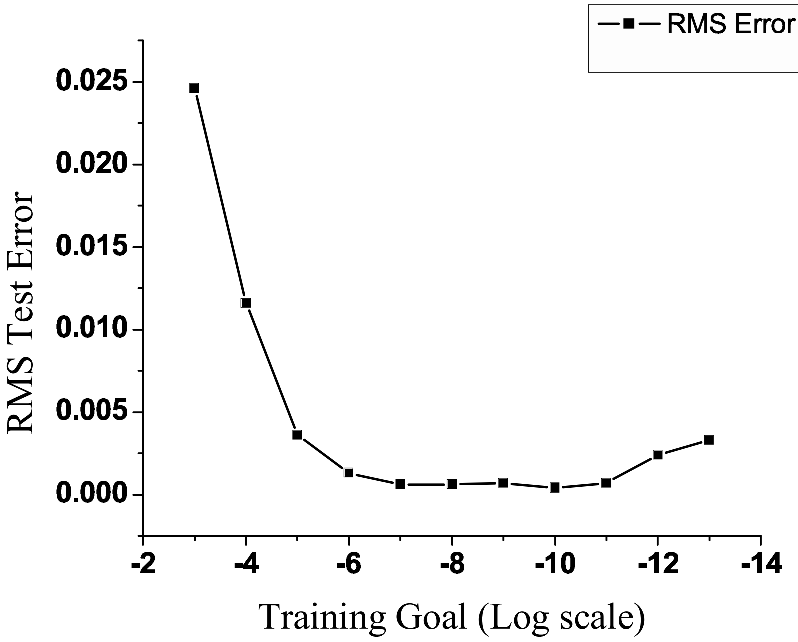
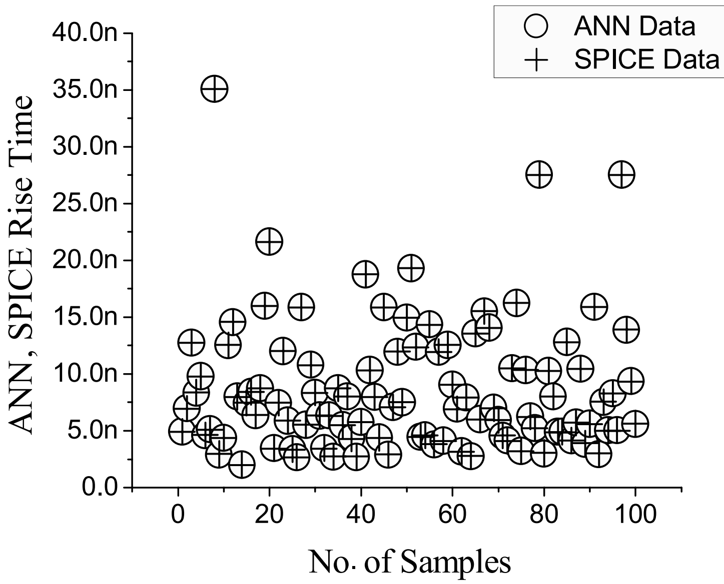
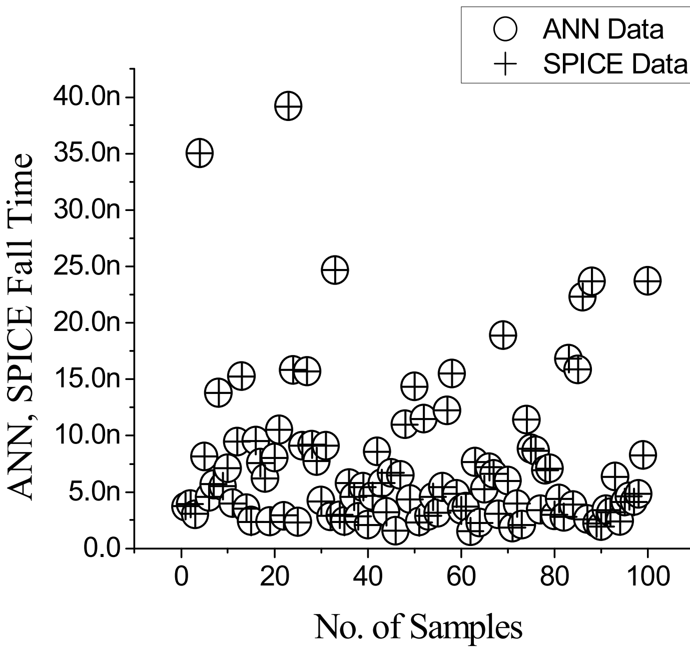


FIGURE 4.8
Effect of the value of training goal on test error.

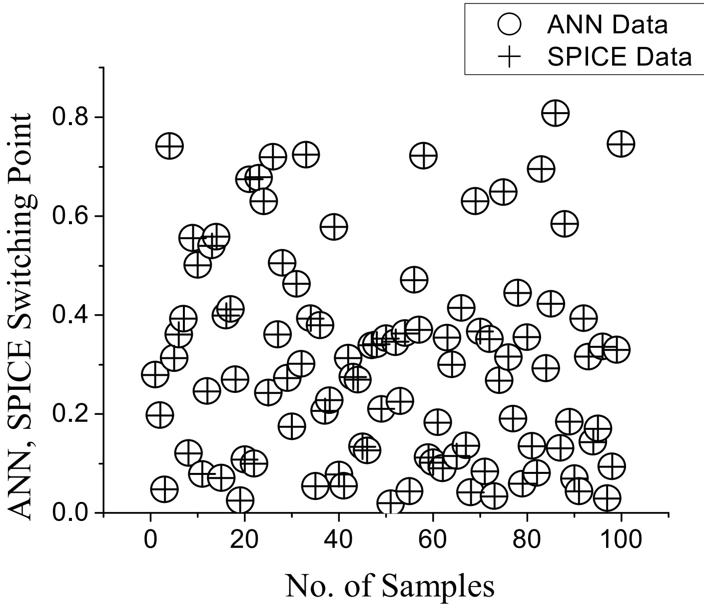


(a) Rise time

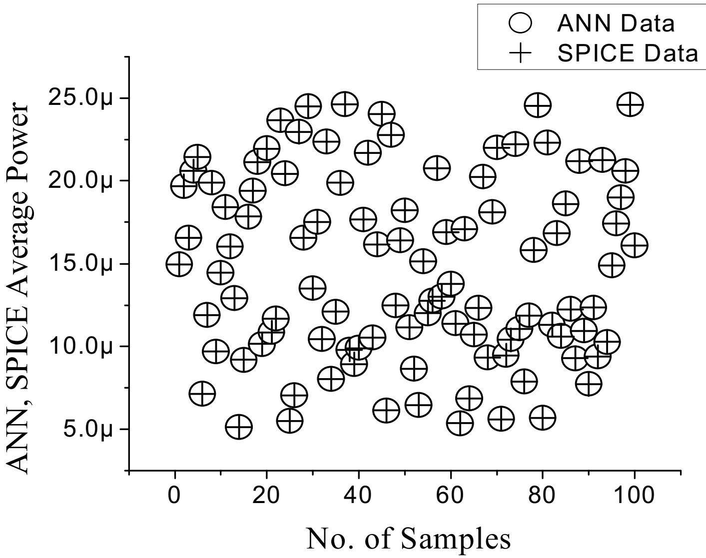


(b) Fall time

FIGURE 4.9
SPICE simulation vs. ANN prediction of rise time and fall time.

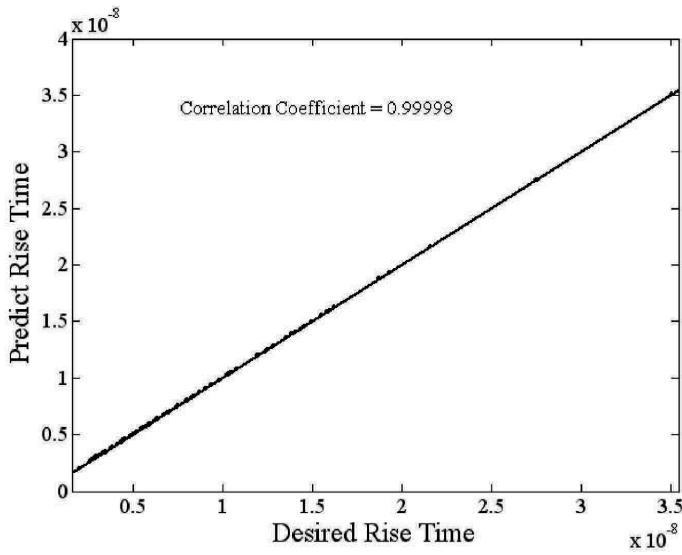


(a) Switching point

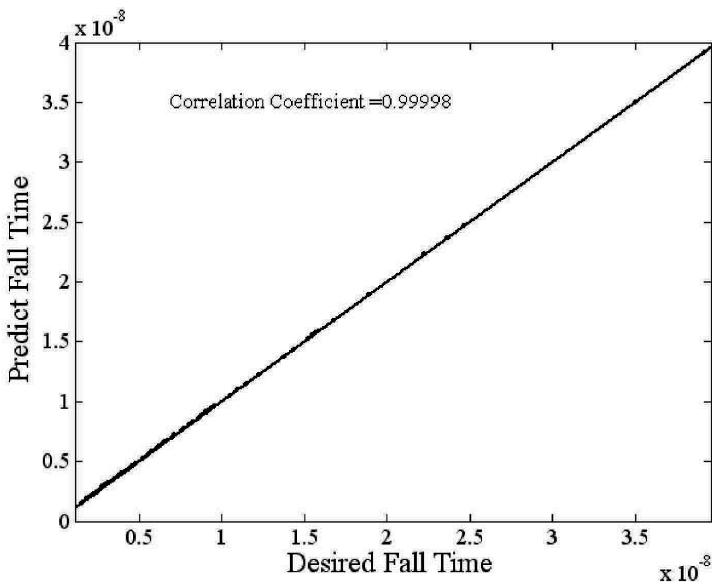


(b) Average power

FIGURE 4.10
SPICE simulation vs. ANN prediction switching point and average power dissipation.

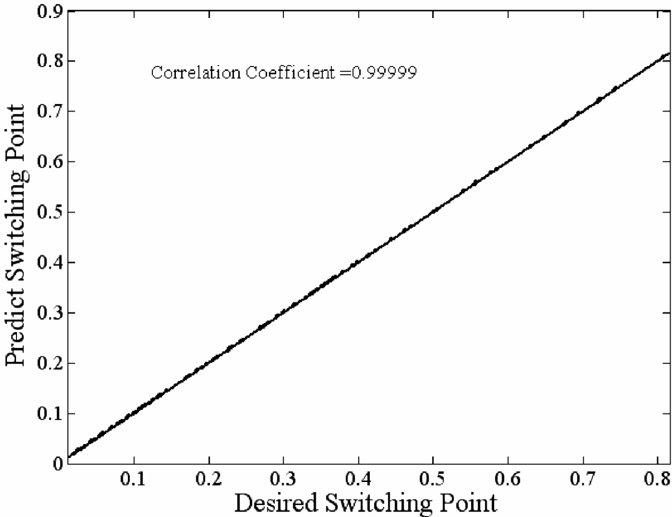


(a) Rise time

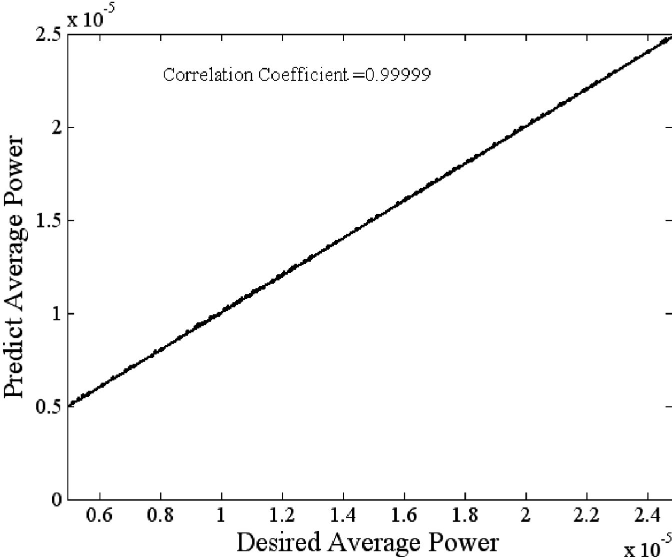


(b) Fall time

FIGURE 4.11
Scatter diagram plot for ANN predicted output for rise time and fall time.



(a) Switching point



(b) Average power

FIGURE 4.12
Scatter diagram plot for ANN predicted output for switching point and average power dissipation.

TABLE 4.3
ANN Model Accuracy for Case Study 1

Error	Output	Type of Dataset	
		Training	Test
E	τ_R	1.02	0.57
	τ_F	1.17	1.67
	V_{SP}	0.38	0.49
	P_{av}	0.42	0.27
R	τ_R	0.99999	0.9998
	τ_F	0.99998	0.99998
	V_{SP}	0.99999	0.99998
	P_{av}	0.99999	0.99999

the metal traces s. The performance parameters of a spiral inductor are (i) the inductance L, (ii) the quality factor Q and (iii) the self resonance frequency SRF. The performance model is thus written as

$$\bar{\alpha} = [d_{out} \quad W \quad n \quad s]^T \tag{4.16}$$

$$\bar{\rho} = [L \quad Q \quad SRF]^T \tag{4.17}$$

$$\bar{\rho} = \mathcal{P}(\bar{\alpha}) \tag{4.18}$$

It may be noted that some other process controlled parameters such as the thickness and resistivity of the metal layer, thickness of the oxide layer, substrate resistivity and the number of metal layers also affect the performances of a spiral inductor.

When the spiral inductor is used as a two-port device, the self inductance of a spiral inductor is obtained as follows [135, 208]

$$L = \frac{\mathrm{Im}\left(\frac{1}{Y_{11}}\right)}{2\pi f} \tag{4.19}$$

where f is the operating frequency. The quality factor Q of a device is defined as the ratio of the amount of energy stored per cycle in a device to the amount of energy dissipated per cycle. The electrical energy stored in the parasitic capacitance of a real inductor is an energy loss. Therefore, the total energy stored in an inductor is considered to be the difference between the average stored magnetic energy and the average stored electrical energy. In terms of impedance, the Q factor is calculated as follows [135, 208].

$$Q = \frac{\mathrm{Im}\left(\frac{1}{Y_{11}}\right)}{\mathrm{Re}\left(\frac{1}{Y_{11}}\right)} \tag{4.20}$$

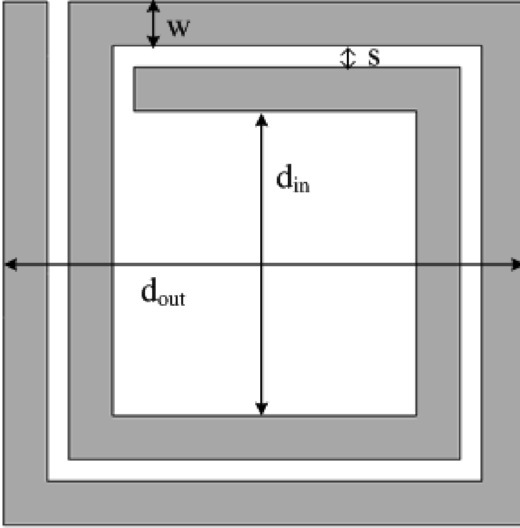


FIGURE 4.13
Layout diagram of an on-chip spiral square inductor.

where Y_{11} is the input impedance obtained from the measured 2-port S parameters of inductors. For on-chip spiral inductors, resonance occurs due to the parasitic effects of the substrate and the distributed characteristics of the metal layers. At the self resonance frequency, both the inductance and the Q values are zero. The self resonance frequency is measured from the Q plot at the frequency point where the Q value becomes zero.

This case study is based on the published literature [122]. The range of the design variables are (i) $100 - 340\mu m$ for the outer diameter d_{out}, (ii) $4 - 32\mu m$ for the width W, (iii) $2 - 8$ for the number of turns n and (iv) $1 - 5$ μm for the spacing s. The samples are extracted following the uniform grid distribution sampling strategy. A set of 500 realizable spiral inductors are selected and the data are generated by electromagnetic simulation. The input and the output data are normalized to the range $[-1, +1]$ through linear scaling. Out of 500 data, 80% are used for training and the rest are used for testing. A standard feed forward MLP architecture is selected to model the inductor performances for operating frequencies of $1GHz$, $2.5GHz$ and $3GHz$. The number of neurons in the hidden layers are taken to be 20 in both layers to maintain the training error within an acceptable range $(2-5\%)$. MATLAB toolbox is used for training purpose. The training goal is considered to be 0.001.

Table 4.5 shows percentage average error and correlation coefficient of each neural model output with respect to the EM simulated value. The average relative errors of L, Q and SRF are found to be less than 5%. This indicates reasonable accuracy of the trained neural network.

TABLE 4.4
ANN Structure for Inductor Modeling Problem

ANN structure/ Operating frequency	4-20-20-3 1 GHz	4-20-20-3 2.5 GHz	4-20-20-3 3 GHz
Training epochs	25	31	47
Time of each epoch (s)	1.67	1.72	1.83

TABLE 4.5
ANN Model Accuracy for the Inductor Modeling Problem

Error	Output	1GHz		2.5GHz		3GHz	
		Type of Dataset		Type of Dataset		Type of Dataset	
		Training	Test	Training	Test	Training	Test
E	L	3.81	3.73	4.44	3.81	3.50	2.23
	Q	1.88	2.16	0.87	0.73	1.05	0.98
	SRF	2.32	1.77	1.46	1.32	1.39	1.55

4.9 Dynamic Adaptive Sampling

Several data mining algorithms, including ANN have an important property that as the training set size increases, the accuracy increases until at some point it saturates, i.e., as the training set size increases beyond a certain value, the accuracy does not increase significantly [147]. A too-small training set will result in sub-optimal generalization performance. On the other hand, a too-large training set results in a lot of training time consumption without any significant advantage. In addition, the procedure of training sample generation is often very costly, especially for integrated circuit design applications. The task of determining an optimal training set size for acceptable accuracy is therefore an important challenge for developing an ANN-based predictive performance model.

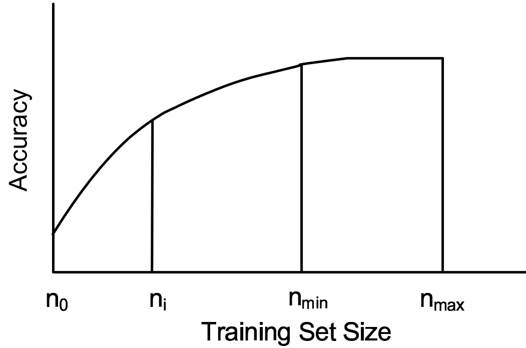


FIGURE 4.14
Hypothetical learning curve for ANN model.

4.9.1 Motivation of the Algorithm

The requirement of a dynamic sampling algorithm for construction of an ANN-based performance model is based upon three observations. First, the predictive performance/accuracy of the ANN increases initially with the increase of training dataset size, however, beyond a certain dataset size, the accuracy does not increase significantly. This is referred to as the learning characteristics of an ANN algorithm [147]. The curve describing the performance as a function of the sample size of the training data is often called the learning curve. A typical plot of the learning curve of an ANN predictive model is shown in Fig. 4.14. Learning curves typically have a steeply rising portion early in the curve, relatively gentle sloping in the middle portion of the curve and a plateau late in the curve [147]. It is observed from this curve, that a model built with a training set size lower than n_{min}, has lower accuracy compared to that of a model built with a training set size n_{min}. On the other hand, a model built with training set size greater than n_{min}, will not have any significant higher accuracy compared to that of the model built with training set size n_{min}. Second, the computational cost of training an ANN model increases as a function of the size of the training dataset. Third, the cost of training data generation for circuit performance modeling is quite high.

4.9.2 Simple Dynamic Sampling Algorithms

The essential idea of the dynamic sampling algorithm is to start with an initial sample size and then increase the sample size in a progressive manner until the desired accuracy of the constructed ANN model is met. In a simple dynamic sampling algorithm, the size of the sample set is increased through arithmetic progression, i.e., $\{n_0, n_0 + n_\delta, n_0 + 2n_\delta, ..., n_0 + k.n_\delta\}$. Arithmetic sampling has an obvious drawback. For a problem requiring a large number of samples, i.e., for large n_{min}, if the starting size is small, the algorithm

requires a large number of iterations. An alternative sampling scheme is the geometric sampling algorithm, where the size of the sample set is increased through geometric progression. While this algorithm may give good results for problems requiring large datasets, it often misses the n_{min} sample size by overshooting. This requires careful selection of the common ratio in the geometric progression series.

Simple progressive sampling algorithms do not generally address a dynamic, adaptive (that varies with the problem) approach to determine the number of instances required at each iteration. This requires the formulation of a dynamic adaptive sampling algorithm [166].

4.9.3 Dynamic Adaptive Sampling Algorithm

This sub-section presents a dynamic adaptive sampling algorithm using a heuristic technique described in [44]. The algorithm takes as inputs: (i) The maximum sample size n_{max}, corresponding to which the sample set may be able to generate, (ii) a very small value ϵ, which is used to formulate the stopping criteria and (iii) the desired accuracy Y of the model. It gives as output the minimum sample size n_{min} and the corresponding data sample D_{min}. The algorithm starts with an initial sample size $n_0 = 0.1 \times n_{max}$. Corresponding to this, the initial data set D_0 is generated through the Halton sequence generator and SPICE simulation. Subsequently the next sample size $n_{(i+1)}$ and the corresponding dataset $D_{(i+1)}$ are generated through a heuristic procedure. The algorithm terminates when a stopping criterion is satisfied. In addition, if the optimum value cannot be located within n_{max}, the algorithm breaks. The pseudo code of the algorithm is described in Fig. 4.15.

4.9.3.1 Initial Sample Size

From the preliminary knowledge about learning curve characteristics, a useful conjecture is to take a small initial sample size (determination of the starting sample size is an open problem). It is heuristically assumed to be $n_0 = 0.1 \times n_{max}$.

4.9.3.2 Sampling Schedule

A "myopic" strategy has been adopted, where it is assumed that the current performance measure of the ANN is the optimal one. The next sample size is believed to be distributed somewhat around the current sample size. This distribution is assumed to be Gaussian distribution. The mean of the Gaussian distribution is kept at the current point and the variance is assigned so as to have about 99.73% (equivalent to 3σ) of the points in the given domain ($n_0 \leq n_i \leq n_{max}$). The variance σ is found by solving the equation

$$3\sigma = \frac{(n_{max} - n_0)}{2} \tag{4.21}$$

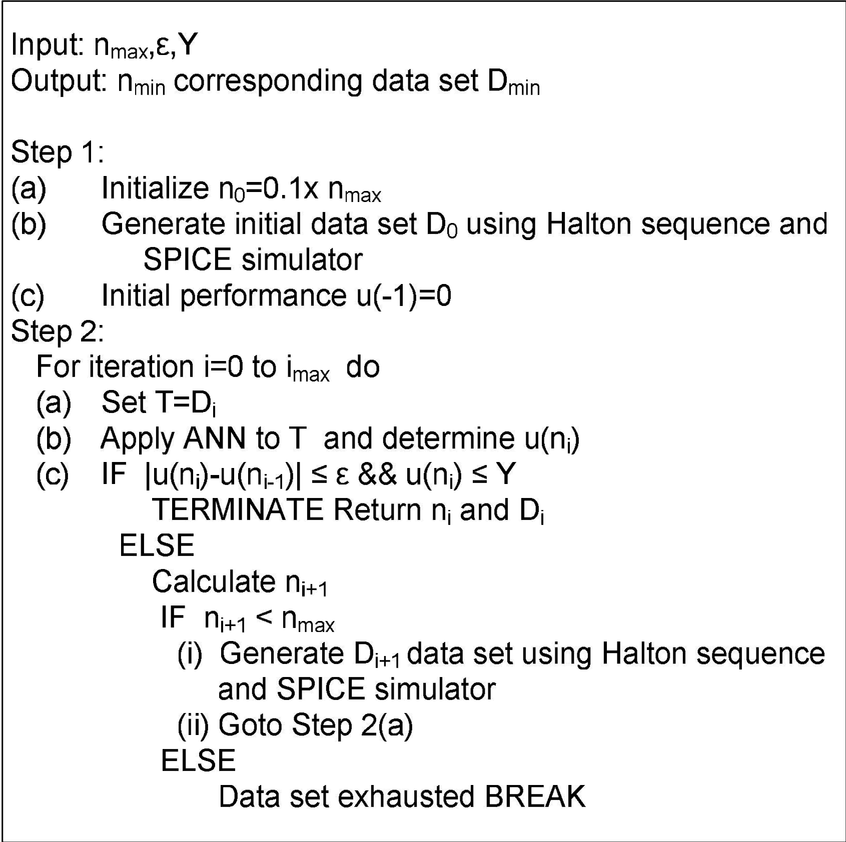


FIGURE 4.15
Dynamic adaptive sampling algorithm.

With this variance, the next sample size is calculated by the formula

$$n_{i+1} = \mu + \sigma N \tag{4.22}$$

where mean $\mu = n_i$ and N is a random number drawn from a Gaussian distribution with zero mean and unity standard deviation σ.

4.9.3.3 Stopping Criteria

An important component of the sampling algorithm is the stopping criteria. Let the current stage be i and the previous stage be $(i - 1)$ and the corresponding performance measures be $u(n_i)$ and $u(n_{i-1})$ respectively. The following inequality is considered as one of the stopping criteria,

$$|u(n_i) - u(n_{i-1})| \leq \epsilon \;\&\&u(n_i) \leq Y \tag{4.23}$$

TABLE 4.6
Comparison of the Total Number of Iterations and CPU Time Required for the Two Dynamic Sampling Methods to Reach Convergence for the Inverter Problem

Method	Iteration Count	CPU time
Arithmetic Sampling	16	$\sim 64min$
This Algorithm	7	$\sim 28min$

TABLE 4.7
ANN Model Accuracy

Error	Output	ARE (%) Test data
E	τ_R	0.42
	τ_F	1.41
	V_{SP}	0.74
	P_{SP}	0.39
R	τ_R	0.9999
	τ_F	0.9998
	V_{SP}	0.9999
	P_{SP}	0.9999

where ϵ is a very small value, depending upon the chosen application. Simultaneously the desired accuracy of the model has to be satisfied. It may be noted that the performance measure $u(n_i)$ is calculated based on the average relative error E, as discussed in (4.10). In addition, if the algorithm does not find the value of the optimal sample set within the given bound of the sample size, the algorithm will terminate.

4.9.4 Demonstration with CMOS Inverter Problem

The algorithm is demonstrated for the CMOS inverter problem. The data generation procedure has been carried out using the standard progressive sampling schemes as well as the present dynamic adaptive sampling algorithm. The optimum sample size n_{min} is found to be equal to 828 as obtained from the adaptive algorithm and 850 using the arithmetic progressive sampling algorithm. Using the geometric sampling algorithm, the convergence could not be achieved for the present problem. The arithmetic sampling technique reaches the optimum point with more iterations compared to that required for the present algorithm. A quantitative comparison between the algorithms is provided in Table 4.6. The CPU time excludes the data generation time, however, includes the training time. This timing information is based on a PC with Core-2-duo processor and 2GB RAM.

In order to verify the quality of the resultant ANN, the various quality metrics are measured. The percentage E and the correlation coefficients measured on test data for all the outputs are summarized in Table 4.7. It is observed that very good accuracy has been obtained in each case.

4.10 Introduction to Least Squares Support Vector Machines

Support Vector Machines (SVM) were first proposed in the year 1995 to solve machine learning problems [196]. Traditional neural network approaches have suffered difficulties with generalization, producing models that can overfit the data. These are consequences of the optimization algorithms used for parameter selection and the statistical measures used to select the "best" model. SVM's are based on the structural risk minimization (SRM) principle, which has been shown to be superior [72] to the traditional empirical risk minimization (ERM) principle employed by the conventional neural networks. SRM minimizes an upper bound on the expected risk, as opposed to ERM that minimizes the error on the training data. It is this difference which equips SVM with a greater ability to generalize, which is the goal in statistical learning.

SVMs were originally developed to solve the classification problem, but thereafter these have been extended to the domain of regression problems [197]. In the literature, the terminologies for SVMs are slightly confusing. The term SVM is typically used to describe classification with support vector methods and support vector regression is used to describe regression with support vector methods. In this text the term SVM has referred to both classification and regression methods, and the terms Support Vector Classification (SVC) and Support Vector Regression (SVR) have been used for support vector machines based classification and regression respectively. A modified version of SVM techniques, referred to as the least squares SVM (LS-SVM), had been proposed by Suykens et al. [187]. LS-SVM technique simplifies the traditional SVM technique to some extent. In this text, the LS-SVM technique has been used. In the following sections least squares support vector regression and classification are discussed in detail.

4.10.1 Least-Squares Support Vector Regression

Consider a given set of training samples $\{\bar{\alpha}_k, d_k\}_{k=1,2,\ldots,n}$ where $\bar{\alpha}_k$ is the input value, and d_k is the output value of the kth sample. Let y_k be the corresponding target value for the kth sample. With a SVR, the relationship between the input vector and the target vector is given as

$$y(\bar{\alpha}) = \bar{\omega}^T \phi(\bar{\alpha}) + b \tag{4.24}$$

where ϕ is the mapping of vector $\bar{\alpha}$ to some (probably high-dimensional) feature space, b is the bias and $\bar{\omega}$ is the weight vector of the same dimension as the feature space. The mapping $\phi(\bar{\alpha})$ is generally nonlinear which makes it possible to approximate nonlinear functions. The approximation error for the kth sample is defined as

$$e_k = d_k(\bar{\alpha}_k) - y(\bar{\alpha}_k) \tag{4.25}$$

For a given data, the weights which give the smallest summed quadratic error of the training samples are determined. Since this can easily lead to overfitting, ridge regression (a form of regression) technique is used to smooth the approximation. The minimization of the error together with the regression is given as

$$\min \;\; \mathcal{J}(\bar{\omega}, \bar{e}) = \frac{1}{2}\bar{\omega}^T\bar{\omega} + \gamma\frac{1}{2}\sum_{k=1}^{n} e_k^2 \tag{4.26}$$

with equality constraint

$$d_k = \bar{\omega}^T\phi(\bar{\alpha}_k) + b + e_k, \;\; k = 1, 2, ..., n \tag{4.27}$$

where γ is the regularization parameter. The first term of the cost function (4.26) is a so-called L_2 norm on the regression weights. The second term takes into account the regression error for all the samples.

The optimization problem (4.26) is considered to be a constrained optimization problem and a Lagrange function is used to solve it. Instead of minimizing the primary objective (4.26), a dual objective, the so-called Lagrangian, is formed, of which the saddle point is the optimum. The Lagrangian for this problem is given as

$$\mathcal{L}(\bar{\omega}, b, \bar{e}, \bar{\alpha}) = \mathcal{J}(\bar{\omega}, \bar{e}) - \sum_{k=1}^{n} \lambda_k \left(\bar{\omega}^T\phi(\bar{\alpha}_k) + b + e_k - d_k\right) \tag{4.28}$$

where λ_ks' are called the Lagrangian multipliers. The saddle point is found by setting the derivatives equal to zero:

$$\frac{\partial \mathcal{L}}{\partial \bar{\omega}} = 0 \rightarrow \bar{\omega} = \sum_{k=1}^{n} \lambda_k \phi(\bar{\alpha}_k) \tag{4.29}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \rightarrow \sum_{k=1}^{n} \lambda_k = 0 \tag{4.30}$$

$$\frac{\partial \mathcal{L}}{\partial e_k} = 0 \rightarrow \lambda_k = \gamma e_k \tag{4.31}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_k} = 0 \rightarrow \bar{\omega}^T\phi(\bar{\alpha}_k) + b + e_k - d_k = 0 \tag{4.32}$$

By eliminating e_k and $\bar{\omega}$ through substitution, the final model is expressed

as a weighted linear combination of the inner product between the training points and a new test object. The output is given as

$$y(\bar{\alpha}) = \langle \bar{\omega}, \phi(\bar{\alpha}) \rangle \quad (4.33)$$

$$= \left\langle \sum_{k=1}^{n} \lambda_k \phi(\bar{\alpha}_k), \phi(\bar{\alpha}) \right\rangle + b \quad (4.34)$$

$$= \sum_{k=1}^{n} \lambda_k \langle \phi(\bar{\alpha}_k), \phi(\bar{\alpha}) \rangle + b \quad (4.35)$$

$$= \sum_{k=1}^{n} \lambda_k K(\bar{\alpha}_k, \bar{\alpha}) + b \quad (4.36)$$

where $K(\bar{\alpha}_k, \bar{\alpha})$ is the kernel function. The elegance of using the kernel function lies in the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\phi(\bar{\alpha})$ explicitly. Any function that satisfies Mercer's condition [187] can be used as the kernel function. The Gaussian kernel function defined as

$$K(\bar{\alpha}_k, \bar{\alpha}) = \exp\left(- ||\bar{\alpha}_k - \bar{\alpha}||^2/\sigma^2\right) \quad (4.37)$$

is commonly used, where σ^2 denotes the kernel bandwidth.

The two hyperparameters, namely the regularization parameter γ and the kernel bandwidth σ^2, have to be tuned by the model developers. These can be optimized by the use of the Vapnik–Chervonenkis bound, k-fold cross validation technique or Bayesian learning. In this text, these have been determined through a grid-search-based technique and a GA-based technique.

The implementation of the entire LS-SVR technique is available in a MATLAB toolbox *lssvmlab* [186] developed by the authors of [187]. This has been extensively utilized in the case studies discussed in this text.

4.10.2 Least-Squares Support Vector Classification

The classification problem is restricted to the consideration of the two-class problem without any loss of generality. In this problem, the goal is to separate two classes of data by a function which is induced from available examples. The goal is to produce a classifier that will work well on unseen examples, i.e., it generalizes well. The general technique deals with a set of training data $\{\bar{\alpha}_k\} = \{\bar{\alpha}_1, \bar{\alpha}_2, ..., \bar{\alpha}_n\} \subseteq \Re^p$ and their corresponding levels $\{d_k\} \subseteq \{-1, 1\}$. The problem is to find a decision function $f : \Re^n \to \{-1, 1\}$ that predicts the label of new, previously unseen data points to minimize the probability of misclassification.

We consider each of n sample points $\bar{\alpha}_k \in \Re^p, k = 1, 2, ..., n$ to be associated with a label $d_k \in \{-1, +1\}$ which classifies the data into one of the two sets. In the simplest SVM formalism where the training data points are linearly separable, the problem is tackled by constructing a hyper-pane $\bar{\omega}^T \bar{\alpha}_k + b$.

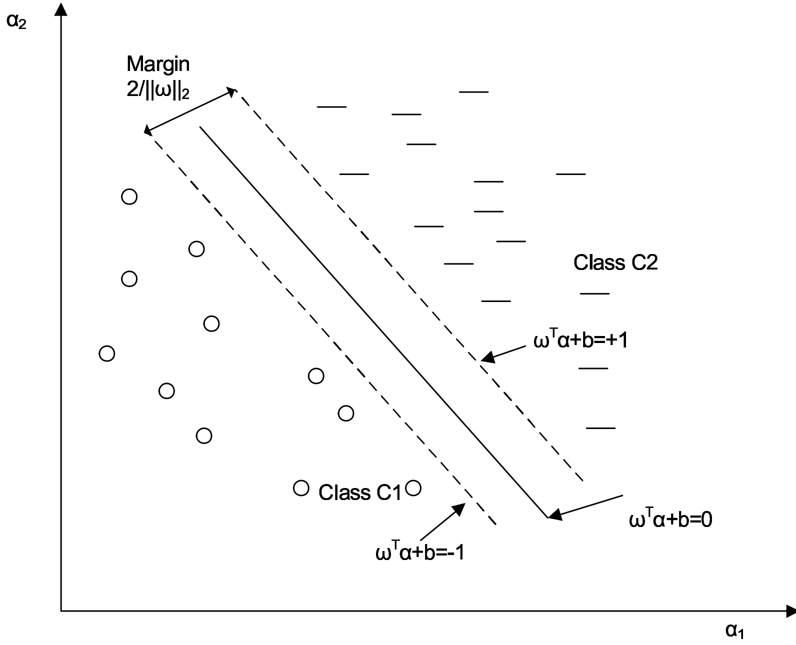


FIGURE 4.16
Classification problem for linearly separable dataset.

The points close to the hyperplane satisfy $\left|\bar{\omega}^T\bar{\alpha}_k + b\right| = 1$. Thus the margin which provides "maximal separation" between points $\bar{\alpha}_k$ belonging to the two classes is $2/\|\omega\|^2$. The situation is illustrated in Fig. 4.16. The linear classifier is

$$y(\bar{\alpha}) = \text{sign}\left[\bar{\omega}^T\bar{\alpha} + b\right] \tag{4.38}$$

When the data of the two classes are separable, it can be written that

$$\bar{\omega}^T\bar{\alpha}_k + b \quad \geq \quad +1, \quad \text{if} \;\; y_k = 1 \tag{4.39}$$
$$\bar{\omega}^T\bar{\alpha}_k + b \quad \leq \quad -1, \quad \text{if} \;\; y_k = -1 \tag{4.40}$$

These two sets of inequalities are combined into one single set as follows

$$y_k\left[\bar{\omega}^T\bar{\alpha}_k + b\right] \geq 1, \qquad k = 1, ..., n \tag{4.41}$$

In order to correctly classify the training data points, the margin of separation should be maximized.

The following modification to the original SVM is proposed by Suykens in the LS-SVC formulation. The optimization problem is formulated as

$$\begin{aligned}\mathcal{J}(\bar{\omega}, \bar{e}) \quad &= \quad \frac{1}{2}\bar{\omega}^T\bar{\omega} + \gamma\frac{1}{2}\sum_{k=1}^{n} e_k^2 \\ \text{such that} \;\; y_k\left[\omega^T\bar{\alpha}_k + b\right] \quad &= \quad 1 - e_k \quad k = 1, 2, ...n\end{aligned} \tag{4.42}$$

The constraints are formulated so that the nearest points $\bar{\alpha}_k$ with labels $+1, -1$ are at least $1/\|\omega\|^2$ distant from the separating hyper-plane. The problem is solved through the Lagrange multiplier technique.

In order to extend the linear method to nonlinear SVM classifiers, the input dataset is replaced by a nonlinear function $\phi(\bar{\alpha})$ operating on the input data. This can be thought of as mapping the input data to a higher (possibly infinite) dimensional space, to enable linear separation of data which is not possible in the original input space (see Fig. 4.17 for illustration). With this the classifier becomes

$$y(\bar{\alpha}) = \text{sign}\left[\bar{\omega}^T \phi(\bar{\alpha}) + b\right] \tag{4.43}$$

With this the Lagrangian of the problem becomes

$$\mathcal{L}(\bar{\omega}, b, \bar{e}; \lambda) = \mathcal{J}(\bar{\omega}, \bar{e}) - \sum_{k=1}^{n} \lambda_k \{y_k \left[\bar{\omega}^T \phi(\bar{\alpha}_k) + b\right] - 1 + e_k\} \tag{4.44}$$

where the $\bar{\lambda}_k$ values are the Lagrange multipliers, which can be positive or negative due to the equality constraints. The conditions for optimality are

$$\frac{\partial \mathcal{L}}{\partial \omega} = 0 \quad \frac{\partial \mathcal{L}}{\partial b} = 0 \quad \frac{\partial \mathcal{L}}{\partial e_k} = 0 \quad \frac{\partial \mathcal{L}}{\partial \lambda_k} = 0 \tag{4.45}$$

Following similar techniques as employed in LS-SVR construction, the final LS-SVC is given by

$$y(x) = \text{sign}\left[\sum_{k=1}^{n} \lambda_k y_k K(\bar{\alpha}, \bar{\alpha}_k) + b\right] \tag{4.46}$$

where $K(\bar{\alpha}, \bar{\alpha}_k)$ is the kernel function.

4.10.2.1 Classifier Accuracy

For evaluating the performances of the SVM classifiers, a set of test samples is identified and three quality metrics, *viz.,* sensitivity (Sen), specificity (Sp) and accuracy (Acc) are measured. These are defined as follows:

Let I denote the entire design space, $\mathcal{D}$ be the feasible design space, and $\mathcal{D}'$ be the approximated feasible design space. Thus I is divided by $\mathcal{D}$ and $\mathcal{D}'$ into four subspaces: TP of true positives, TN of true negatives, FP of false positives, and FN of false negatives. This is shown in Fig. 4.18. Sensitivity (Sen) is defined as the percentage of true positives relative to all the positive instances.

$$Sen = \frac{|TP|}{|TP| + |FN|} \tag{4.47}$$

Specificity (Sp) is defined as the percentage of true negatives relative to all the negative instances.

$$Sp = \frac{|TN|}{|TN| + |FP|} \tag{4.48}$$

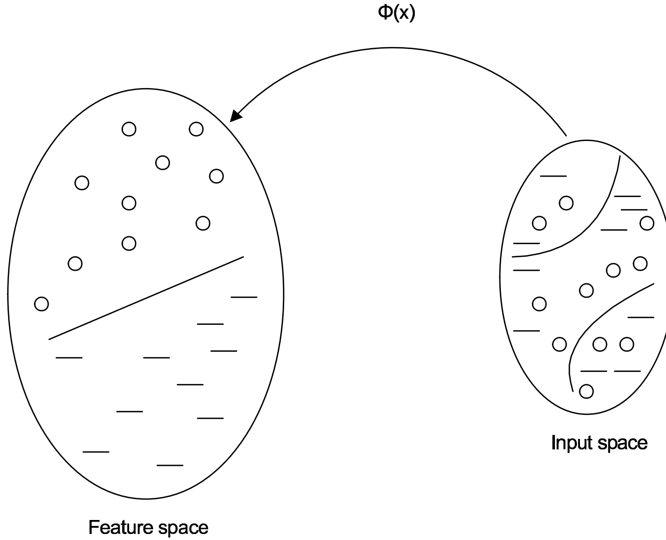


FIGURE 4.17
Mapping of the input space to a high dimensional feature space where linear separation of nonseparable data is possible.

Accuracy (Acc) is defined as the percentage of correctly classified instances in the dataset.

$$Acc = \frac{|TP| + |TN|}{|I|} \tag{4.49}$$

For a good classifier, these values ideally should be equal to unity.

4.10.3 Choice of Kernel Functions and Hyperparameter Tuning

The use of a kernel function allows the SVM representation to be independent of the dimensionality of the input space. The first step of construction of an LS-SVM model is the selection of an appropriate kernel function. For the choice of kernel function $K(\bar{\alpha}_k, \bar{\alpha})$, there are several alternatives. Some of the commonly used functions are listed in Table 4.8, where d, σ, κ and θ are constants, referred to as hyperparameters. In general, in any classification or regression problem, if the hyperparameters of the model are not well selected, the predicted results will not be good enough. Optimum values for these parameters therefore need to be determined through proper tuning methods. Note that the Mercer condition holds for all σ and d values in the radial basis function (RBF) and the polynomial case, but not for all possible choices of κ and θ in the multi-layer perceptron (MLP) case. Therefore, the MLP kernel will not be considered.

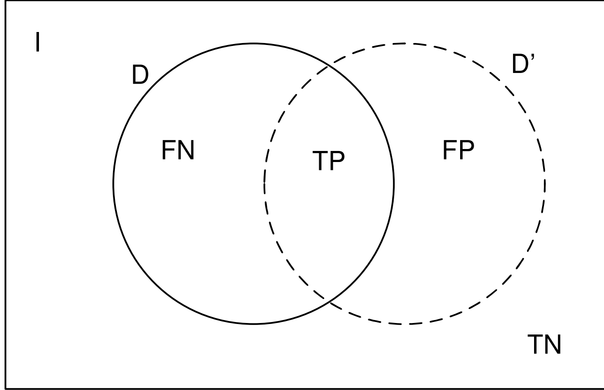


FIGURE 4.18
Feasible design space and its subspaces.

TABLE 4.8
List of Kernel Functions

Name	Function Expression
Linear Kernel	$K(\bar{\alpha}_k, \bar{\alpha}) = \bar{\alpha}_k^T \bar{\alpha}$
Polynomial Kernel	$K(\bar{\alpha}_k, \bar{\alpha}) = \left(1 + \bar{\alpha}_k^T \bar{\alpha}\right)^d$
RBF Kernel	$K(\bar{\alpha}_k, \bar{\alpha}) = \exp\left\{\frac{-\|\bar{\alpha}_k - \bar{\alpha}\|_2^2}{\sigma^2}\right\}$
MLP Kernel	$K(\bar{\alpha}_k, \bar{\alpha}) = \tanh\left(\kappa \bar{\alpha}_k^T \bar{\alpha} + \theta\right)$

As mentioned earlier, when designing an effective LS-SVM model, the hyperparameter values have to be chosen carefully. The regularization parameter γ, determines the trade-off cost between minimizing the training error and minimizing the model error. The kernel parameter σ or d defines the nonlinear mapping from the input space to some high dimensional feature space [53].

Optimal values of the hyperparameters are usually determined by minimizing the estimated generalization error. The generalization error is a function that measures the generalization ability of the constructed models, i.e., the ability to predict correctly the performance of an unknown sample. The two commonly used techniques for estimating the generalization error are [53]:

1. Hold-out method: This is a simple technique for estimating the generalization error. The dataset is separated into two sets, called the training set and the test set. The SVM is constructed using the training set only. Then it is tested using the test dataset. The test data are completely unknown to the estimator. The mean test error which is computed by considering the errors over all test samples is used to evaluate the model. This method is very fast. However, its

evaluation can have a high variance. The evaluation depends heavily on the data points that end up in the training set and on those which end up in the test set. Consequently the evaluation may be significantly different depending on how the division is made.

2. k-fold cross validation method: In this method, the training data is randomly split into k mutually exclusive subsets (the folds) of approximately equal size [53]. The SVM is constructed using $k-1$ of the subsets and then tested on the subset left out. This procedure is repeated k times. An estimation of the expected generalization error is obtained by averaging the test error over the k trials. The advantage of this method is that the accuracy of the constructed SVM does not depend upon the division of data. The variance of the resulting estimate is reduced as k is increased. The disadvantage of this method is that it is time consuming.

Primarily there are three different approaches for optimal determination of the SVM hyperparameters: the heuristic method, the local search method and the global search method. The σ value is related to the distance between training points and the smoothness of the interpolation of the model. A heuristic rule has been discussed in [158] for estimating the σ value as $[\sigma_{min}, \sigma_{max}]$ where σ_{min} is the minimum distance (non-zero) between two training points and σ_{max} is the maximum distance between two training points. The regularization parameter γ is determined based upon the trade-off between the smoothness of the model and its accuracy. The bigger its value the more importance is given to the error of the model in the minimization process. Choosing a low value is not suggested while using exponential RBF to model performances which are often approximately linear or weakly quadratic in most input variables. While constructing a LS-SVM-based analog performance model, the heuristic method has been applied for determining the hyperparameters in [101]. The hyperparameters generated through the heuristic method are often found to be sub-optimal as demonstrated in [141]. Therefore, determination of hyperparameters through formal optimization procedure is suggested [53].

The present text discusses two algorithmic techniques for selecting optimal values of the model hyperparameters. The first one is a grid search technique and the other one is a genetic algorithm-based technique. These are explained below considering the RBF as the kernel function. For other kernels, the techniques are accordingly used.

4.10.3.1 Grid Search Technique

In the grid search technique, pairs of (γ, σ^2) are tried and the one with the best accuracy is chosen. The basic steps of the grid search-based technique is outlined below [187, 140]:

1. Consider a grid space of (γ, σ^2), defined by $\log_2 \gamma \in \{lb_\gamma, ub_\gamma\}$ and

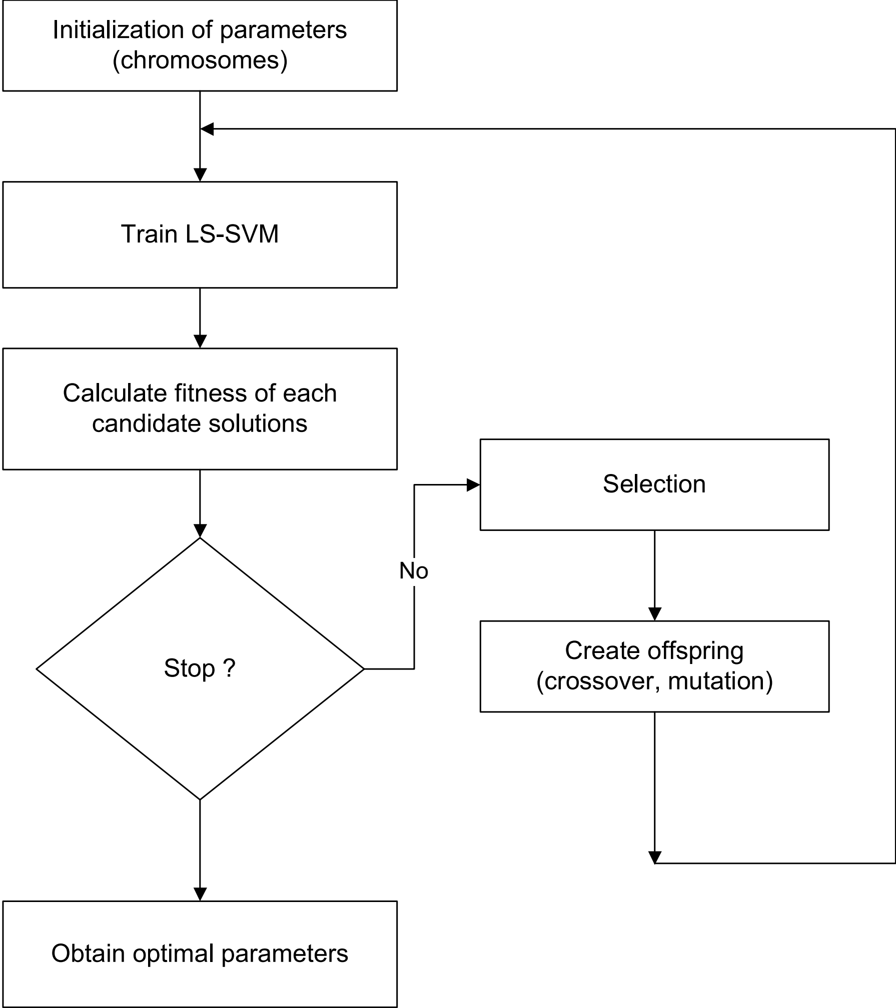


FIGURE 4.19
Outline of GA-based hyperparameter selection procedure.

$\log_2 \sigma^2 \in \{lb_{\sigma^2}, ub_{\sigma^2}\}$, where $[lb_\gamma, ub_\gamma]$ and $[lb_{\sigma^2}, ub_{\sigma^2}]$ define the boundary of the grid space.

2. For each pair within the grid space, estimate the generalization error through the hold-out/k-fold cross validation technique.
3. Choose the pair that leads to the lowest error.
4. Use the best parameter to create the SVM model as a predictor.

The grid search technique is simple. However, this is computationally expensive, since this is an exhaustive search technique. The accuracy and time cost of the grid method are trade-offs depending on the grid density. In general, with the increase in grid density, the computational process becomes expensive. On the other hand, sparse density lowers the accuracy. The grid search technique is therefore performed in two stages. In the first stage, a coarse grid search is performed. After identifying a better region on the grid, a finer grid search on that region is conducted in the second stage. In addition, the grid search process is a tricky task since a suitable sampling step varies from kernel to kernel and the grid interval may not be easy to locate without prior knowledge of the problem. These parameters are determined through a trial and error method.

4.10.3.2 Genetic Algorithm-Based Technique

In order to reduce the computational time required to determine the optimal hyperparameter values without sacrificing the accuracy, numerical gradient-based optimization technique can be used. However, it has been found that often the SVM model selection criteria have multiple local optima with respect to the hyperparameter values [197]. In such cases, the gradient-based method have chances to be trapped in bad local optima. Considering this fact, any heuristic global optimization technique is preferred for determining the hyperparameter values. The present text discusses a genetic algorithm based technique [140].

In the GA-based technique, the task of selection of the hyperparameters is the same as an optima searching task, and each point in the search space represents one feasible solution (specific hyperparameters). Each feasible solution is marked by its estimated generalization ability, and the determination of a solution is equal to the determination of some extreme point in the search space.

An outline of a simple GA-based process is shown in Fig. 4.19. The chromosomes consist of two parts, $\log_2 \gamma$ and $\log_2 \sigma^2$. The encoding of the hyperparameters into a chromosome is a key issue. A real-coded scheme is used as the representation of the parameters in this work. Therefore, the solution space coincides with the chromosome space. In order to produce the initial population, the initial values of the designed parameters are distributed in the solution space evenly. The selection of population size is one of the factors that affects the performance of GA. The GA evaluation duration is proportional to the population size. If the population size is too large, a prohibitive amount of time for optimization will be required. On the other hand, if the population size is too small, the GA can prematurely converge to a sub-optimal solution, thereby reducing the final solution quality. There is no generally accepted theory for determining optimal population size. Usually, it is determined by experimentation or experience.

During the evolutionary process of GA, a model is trained with the current hyperparameter values. The hold-out method as well as the k-fold cross

validation method are used for estimating the generalization error. The fitness function is an important factor for estimation and evolution of SVMs providing satisfactory and stable results. The fitness function expresses the users' objective and favors SVMs with satisfactory generalization ability. The fitness of the chromosomes in the present work is determined by the average relative error (ARE) calculated over the test samples. The fitness function is defined as

$$F = \frac{1}{ARE(\gamma, \sigma^2)} \tag{4.50}$$

Thus, maximizing the fitness value corresponds to minimizing the predicted error. The ARE function is as defined in (4.10). The fitness of each chromosome is taken to be the average of five repetitions. This reduces the stochastic variability of the model training process in GA-based LS-SVM.

The genetic operator includes the three basic operators such as selection, crossover, and mutation. The roulette wheel selection technique is used for the selection operation. The probability p_i of selecting the ith solution is given by

$$p_i = \frac{F_i}{\sum_{i=1}^{N_{pop}} F_i} \tag{4.51}$$

where N_{pop} is the size of the population. Besides, in order to keep the best chromosome in every generation, the idea of elitism is adopted. The use of a pair of real-parameter decision variable vectors to create a new pair of offspring vectors is done by the crossover operator. For two parent solutions $\bar{x}_1$ and $\bar{x}_2$, the offspring is determined through a blend crossover operator[41]. For two parent solutions $\bar{x}_1$ and $\bar{x}_2$, such that $\bar{x}_1 < \bar{x}_2$, the blend crossover operator (BLX-β) randomly picks a solution in the range $[\bar{x}_1 - \beta(\bar{x}_2 - \bar{x}_1), \bar{x}_2 + \beta(\bar{x}_2 - \bar{x}_1)$. Thus, if u be a random number in the range (0,1) and $\theta = (1 + 2\beta)u - \beta$, then the following is an offspring

$$\bar{x}^{new} = (1 - \theta\bar{x}_1) + \theta\bar{x}_2 \tag{4.52}$$

If β is zero, this crossover creates a random solution in the range $(\bar{x}_1, \bar{x}_2)$. It has been reported for a number of test cases that BLX-0.5 (with $\beta = 0.5$) performs better than BLX operators with any other β value. The mutation operator is used with a low probability to alter the solutions locally to hopefully create better solutions. The need for mutation is to maintain a good diversity of the population. The normally distributed mutation operator is used in this work. A zero mean Gaussian probability distribution with standard deviation η_i for the ith solution is used. The new solution is given as

$$\bar{x}^{new} = \bar{x_i} + N(0, \eta_i) \tag{4.53}$$

The parameter η_i is user-defined and dependent upon the problem. Also, it must be ensured that the new solution lies within the specified upper and lower limits. When the difference between the estimated error of the child population

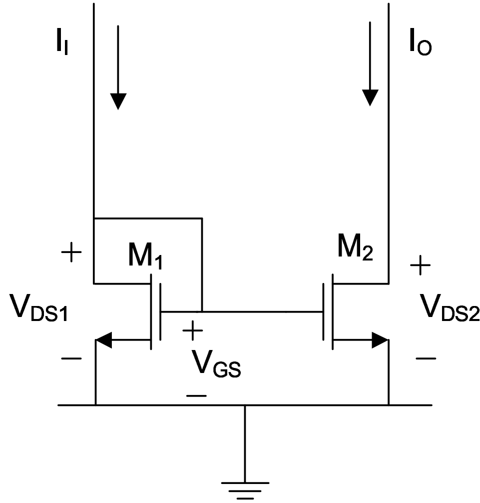


FIGURE 4.20
n-Channel simple current mirror circuit.

and that of the parent population is less than a pre-defined threshold over certain fixed generations, the whole process is terminated and the corresponding hyperparameter pair is taken as the output.

It may be mentioned here that there is no fixed method for defining the GA parameters, which are all empirical in nature. However, the optimality of the hyperparameter values is dependent upon the values of the GA parameters. The values of the GA parameters are selected primarily by the trial and error method over several runs.

4.11 Feasible Design Space and Feasibility Model

The task of identification of the feasible design space is an important task in the circuit sizing process. The circuit sizing process using numerical optimization techniques often generate pathological results. This means that the design on the one hand meets all specifications but on the other hand fails some basic design requirements, leading to a malfunctioning circuit. This situation arises due to lack of restricting the design space exploration process within a feasible design space. The experienced IC designers deal with the case of pathological circuit sizing by manually constraining the circuit to ensure proper biasing and good behavior of performance metrics. In addition, this also requires good understanding of the physics of the MOS transistors. The

methodology which is employed to automatically constrain a circuit in order to ensure proper operation is called the sizing rules [68].

A feasible design space is defined as a multidimensional space in which every design satisfies a set of feasible design constraints. These design constraints are classified into three categories [47]:

1. Geometry Constraints C_g: The geometry constraints are applied directly on the component sizes, such as the transistor dimensions, values of the resistors and capacitors. The constraints on the device sizes are usually given in the form of lower bounds and upper bounds. The lower bounds are generally determined by the feature size of a technology. The upper bounds, on the other hand, are selected by the designer such that the devices are not excessively large. The geometry constraints are specified in the form

$$C_g = \{f_i(\bar{\alpha}) \leq 0 \quad i = 1, 2, ...n_g\} \tag{4.54}$$

2. Functional Constraints C_f: These constraints are used to ensure the correct functionality of the circuits. These are applied on the current-voltage relationships of the circuit based on the physics of the MOS transistors and are given in analytic form. The functionality constraints are specified in the form

$$C_f = \{f_i(v, i) \leq 0 \quad i = 1, 2, ...n_f\} \tag{4.55}$$

This is explained with an example of a current mirror circuit as shown in Fig. 4.20. The working of the current mirror circuit is based on the principle that if the gate-source voltages of two identical MOS transistors are equal, the channel currents should be equal [5]. The current I_I is defined by some current source and I_O is the output or the mirrored current. Since the drain and the gate terminals of the transistor M_1 are shorted, this transistor remains in the saturation region. The transistor M_2 must also remain in saturation, the condition for which is $V_{DS2} > V_{GS2} - V_{T2}$. Assuming that the transistors are identical, it can be written that

$$\frac{I_O}{I_I} = \left(\frac{L_1 W_1}{W_1 L_2}\right)\left(\frac{1 + \lambda V_{DS2}}{1 + \lambda V_{DS1}}\right) \tag{4.56}$$

where W_1, L_1 are the widths and lengths of the transistor M_1, W_2, L_2 are the widths and lengths of the transistor M_2. It needs to be assumed that $V_{DS1} = V_{DS2}$ so that I_O/I_I becomes a function of the aspect ratio which is under the control of the designer [5]. Based on this concept, the functional constraints are listed as follows

- Strong inversion of M_1 and M_2: $V_{GS1} > V_T$ and $V_{GS2} > V_T$

- Saturation of M_2: $V_{DS2} > V_{GS2} - V_T$
- Equal aspect ratio for current mirror: $L_1 = L_2, W_1 = W_2$ (belongs to the category of geometry constraints)

The geometry and the functional constraints define the feasible design space. Apart from these two, there is another type of constraint which is applied on the performances of the circuit and is required for defining the feasible performance space.

3. Performance Constraints C_ρ: These are applied on the performance parameters depending upon the chosen application systems

$$C_\rho = \{f_i(\rho) \leq 0 \quad i = 1, 2, ...n_\rho\} \tag{4.57}$$

For example, the phase margin of an operational amplifier must be greater than 45°.

The total set of constraints for feasibility checking is thus

$$C = \{C_g \cup C_f \cup C_\rho\} \tag{4.58}$$

The feasible design space is thus defined by

$$\mathcal{D} = \{\bar{\alpha} | \bar{\alpha} \in C\} \tag{4.59}$$

This is somewhat the same as defined in Chapter 2. It is to be noted that through the process of feasibility checking, various simulation data are discarded. At a glance this may give an impression about wastage of costly simulation time. However, for an analog designer (who is a user of the model), this is an important advantage. This is because the infeasible data points will never appear as a solution whenever the model will be used for design characterization/optimization. Even from the model developer's perspective, this is not a serious matter considering the fact that the construction process is in general a one-time process [47]. The feasibility constraints remain invariant if the performance objectives are changed. Even if the design migrates by a small amount, these constraints usually do not change [69]. This however, demands an efficient determination of the feasibility constraints.

Since some of the constraints are not directly applied over the design variables $\bar{\alpha}$, it is very difficult to express $\mathcal{D}$ in analytic form. A feasibility model $\mathcal{F}(\bar{\alpha})$ is defined as one whose output only takes two values $\{+1, -1\}$, depending on whether $\bar{\alpha} \in \mathcal{D}$.

$$\mathcal{F}(\bar{\alpha}) = \begin{cases} +1 & \text{if } \bar{\alpha} \in \mathcal{D} \\ -1 & \text{if } \bar{\alpha} \notin \mathcal{D} \end{cases}$$

Feasibility modeling is treated as a classification problem, and existing classification techniques such as LS-SVC are applied to solve it. Instances from simulations are used to train a selected model with the objective of minimizing the classification error on the training set.

FIGURE 4.21
Two-stage CMOS operational amplifier circuit for Case Study 3.

4.12 Case Study 3: Combined Feasibility and Performance Modeling of Two-Stage Operational Amplifier

Identification of feasible design space is essential prior to the construction of performance models since it screens out infeasible designs of which performance parameters are essentially noise to a regression based model generation technique. This case study presents a combined feasibility and performance model generation for a two-stage operational amplifier circuit [47, 18]. The circuit schematic is shown in Fig. 4.21. In order to reduce the number of design variables, the lengths of all the transistors are assumed to be fixed at $L = 0.5\mu m$. The transistors M_1 and M_2 in the differential pair must have equal widths for matching purposes, i.e., $W_1 = W_2$. Similarly M_3 and M_4 must also have same widths so that $W_3 = W_4$. The width of the transistor M_6 is fixed to $W_6 = 10\mu m$. The width of the transistor M_8 is given by the relationship $W_8 = 2 \times \dfrac{W_7}{W_5} \times W_3$. The geometry constraints are summarized

TABLE 4.9
Transistor Sizes and Feasibility Constraints for Two-Stage OPAMP Circuit

	Parameters	Ranges
Transistor Sizes Geometry Constraints	$W_1 = W_2$	$[1\mu m, 100\mu m]$
	$W_3 = W_4$	$[1\mu m, 100\mu m]$
	W_5	$[1\mu m, 100\mu m]$
	W_7	$[1\mu m, 100\mu m]$
	W_8	$\frac{2 \times W_3 \times W_7}{W_5}$
	C_C	$[1pF, 20pF]$
Functional Constraints	Parameters	Range
	$V_{gs} - V_{th}$	$\geq 0.1V$
	V_{ds}	$\geq V_{gs} - V_{th} + 0.1V$
	V_{op}	$\approx 0.9V$
	V_{off}	$\leq 2mV$
Performance Constraints	Bandwidth	$\geq 2MHz$
	Phase margin	$\geq 45^0$

in Table 4.9. The functional and the performance constraints are also shown in Table 4.9. These constraints ensure that all the transistors are ON and operate in a saturation region with some margin.

4.12.1 Feasibility Model

The netlist of the circuit is simulated using the SPICE simulation tool. The model used for simulation is the BSIM3v3 model. Values of these design variables are randomly generated within upper and lower bounds to get a set of 10,000 tuples of design variables. SPICE simulation is run for this set of 10,000 tuples of design variables. Functional constraints and performance constraints are verified using SPICE simulation. The outputs corresponding to the tuples which satisfy the functional and performance constraints are taken as +1 otherwise the outputs are taken as −1. This results in 10,000 input and output data pair. Out of these data pairs, 6000 are used to train LS-SVM classifier and the rest are used for test purposes to check the accuracy of the classifier. The model is trained using an RBF kernel. For selecting the values of the hyperparameters, the heuristic rule suggested in [101] is followed and $\sigma^2 = 24, \gamma = 250$ have been used to get good results. The various classifier accuracy parameters, i.e., sensitivity (Sen), specificity (Sp) and accuracy (Acc) are provided in Table 4.10. The values of these parameters are close to unity which demonstrate the accuracy of the constructed feasibility model.

TABLE 4.10
Statistics of the Constructed Feasibility Models

No of Test data	σ^2	γ	Sen	Sp	Acc
4000	24	250	0.945	0.967	0.992

TABLE 4.11
Statistics of the Constructed Performance Models of Case Study 3

Model	σ^2	γ	RMS		$MAX\%$		T_{tr}
			Training	Test	Training	Test	(s)
$A_v(dB)$	18	178	0.047	0.050	0.220	0.225	128.54
$GB(\%)$	15	180	0.087	0.10	0.350	0.405	120.39
$PM(^0)$	23	220	0.032	0.035	0.720	0.985	139.09

4.12.2 Performance Model

Performance models of three performance parameters: open loop gain A_v, unity gain frequency GB, and phase margin PM are generated. These can be obtained by running one AC analysis. The data generated during feasibility model checking is reused here. The data that satisfies the feasibility constraints, and hence belong to the feasible category as predicted by the feasibility model, are used for constructing the performance model.

Let ρ' be the estimated performance parameter and ρ be the actual performance parameter. The error of models for gain A_v and phase margin PM are defined as

$$e = \rho' - \rho \tag{4.60}$$

and the error of GB is defined as

$$e = \frac{\rho' - \rho}{\rho} \tag{4.61}$$

that has a unit percentage. The maximum of the absolute values of the training error or test error is defined as

$$MAX = \max\{|e|\} \tag{4.62}$$

This denotes the worst case performance of the generated performance model. The root mean square error RMS and MAX on the training and test dataset for the three performance models are shown in Table.4.11. Reasonably good results for the model have been obtained.

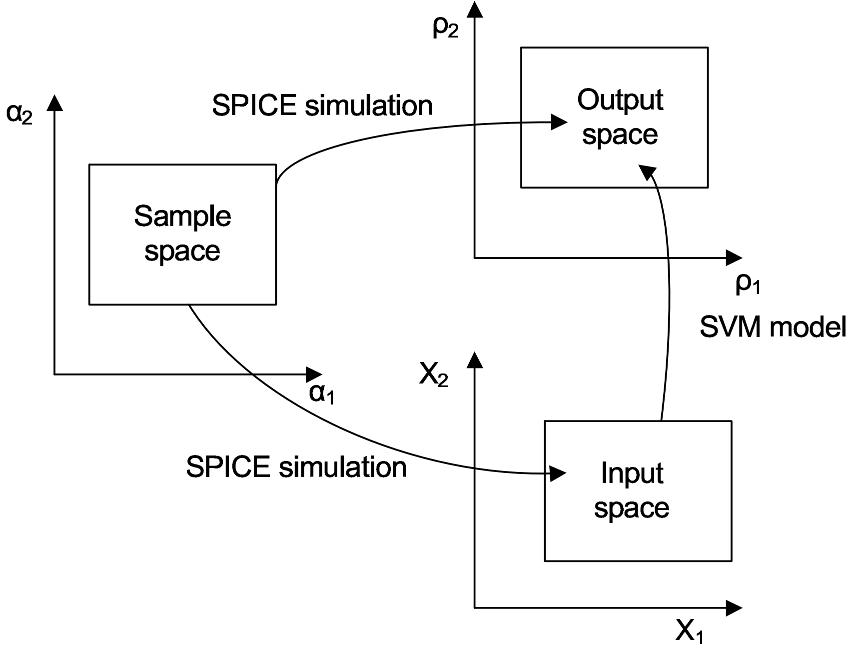


FIGURE 4.22
Nonlinear mapping from the sample space to the input and the output space for the construction of architecture level performance model.

4.13 Case Study 4: Architecture-Level Performance Modeling of Analog Systems

The performance models that are used in the architecture-level design abstraction are referred to as architecture-level performance models. An analog architecture-level performance model is a function that estimates the performance of an analog component block when some architecture-level design parameters of the block are given as inputs [141, 109]. This case study [141, 140] describes the construction of feasible architecture-level performance model of a two-stage CMOS operational transconductance amplifier (OTA), shown in Fig. 4.23. The technology is $0.18\mu m$ CMOS process, with a supply voltage of $1.8V$. The performance parameters are (1) input referred thermal noise (ρ_1), (2) power consumption (ρ_2), and (3) output impedance (ρ_3). These serve as the outputs of the performance model. The selected high-level design variables are functions of DC gain (X_1), bandwidth (X_2) and slew rate (X_3). These serve as the inputs of the performance model.

For the problem of construction of an architecture-level performance model, both the inputs and the output design variables of the model

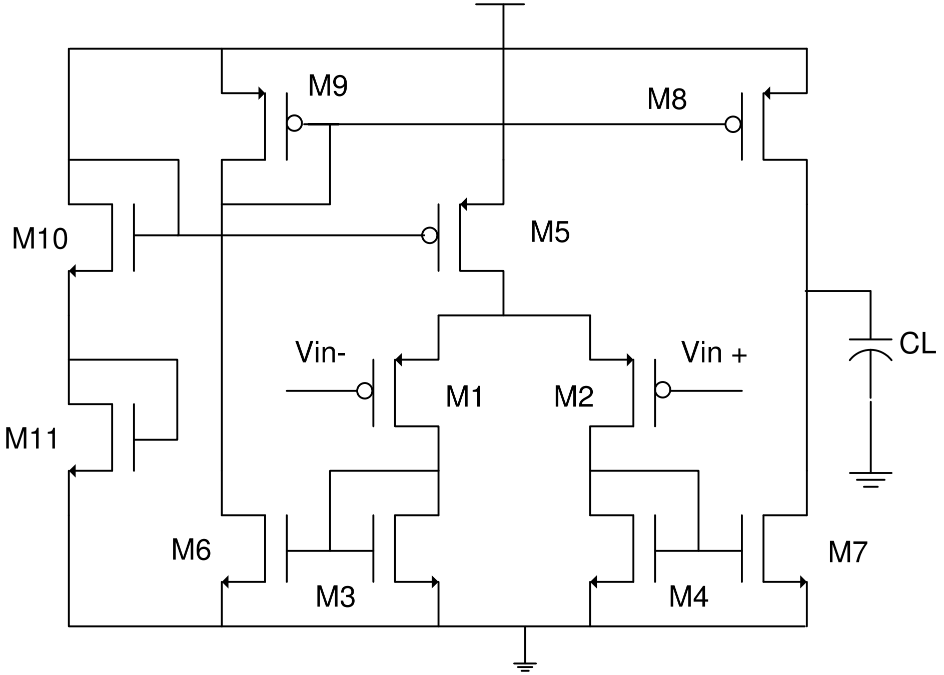


FIGURE 4.23
Two-stage CMOS OTA circuit for Case Study 4.

are functions of the transistor geometry parameters. This is expressed through

$$\bar{X} = \mathcal{R}(\bar{\alpha}) \tag{4.63}$$

$$\bar{\rho} = \mathcal{Q}(\bar{\alpha}) \tag{4.64}$$

The sample space is defined by a set of geometry constraints applied over the transistor dimensions. From this sample space, the SPICE simulation tool is used to generate the data samples corresponding to the input and the output model parameters. This is shown in Fig.4.22 [141]. The transistor level parameters along with the various feasibility constraints are listed in Table 4.12. From the sample space defined by the transistor sizes, a set of 5000 samples is generated using a Halton sequence generator. These are simulated through AC analysis, operating point analysis, noise analysis, and transient analysis using the SPICE program. The functional constraints ensure that all the transistors are ON and are in the saturation region with some user defined margin. Out of all the samples, only 1027 samples are found to satisfy the functional and performance constraints listed in Table 4.12.

The estimation functions are generated using the LS-SVM technique. The generalization errors are estimated through the hold-out method and the

TABLE 4.12
Transistor Sizes and Feasibility Constraints for OTA

	Parameters	Ranges
Transistor Sizes Geometry Constraints	$W_1 = W_2$	$[280nm, 400\mu m]$
	$W_3 = W_4 = W_6 = W_7$	$[1\mu m, 20\mu m]$
	$W_8 = W_9$	$[280nm, 10\mu m]$
	W_5	$[1\mu m, 50\mu m]$
	$W_{10} = W_{11}$	$[280nm, 400\mu m]$
	C_L	$[1pF, 10pF]$
Functional Constraints	Parameters	Range
	$V_{gs} - V_{th}$	$\geq 0.1V$
	V_{op}	$\approx 0.9V$
	V_{off}	$\leq 2mV$
Performance Constraints	Slew rate	$[0.1V/\mu s, 20V/\mu s]$
	Bandwidth	$\geq 2MHz$
	DC Gain	$\geq$ 70 dB
	Phase margin	$[45^0, 60^0]$

TABLE 4.13
Grid Search Technique Using Hold-Out Method

Model	σ^2	γ	$ARE(\%)$		R		T_{tr}
			Training	Test	Training	Test	(min)
ρ_1	3.43	173.26	1.82	2.48	0.999	0.998	118.19
ρ_2	2.10	112.04	2.32	4.18	0.918	0.905	117.83
ρ_3	5.43	387.55	2.02	3.14	0.999	0.937	118.13

5-fold cross validation method. The hyperparameters are computed through the grid search and the GA-based technique. In the grid search technique, the hyperparameters (σ^2, γ) are restricted within the range $[0.1, 6.1]$ and $[10, 510]$. The grid search algorithm is performed with a step size of 0.6 in σ^2 and 10 in γ. These parameters are fixed based on heuristic estimations and repeated trials. The determined hyperparameter values along with the quality measures and the training time are reported in Table 4.13 and Table 4.14 for the hold-out method and the cross validation method respectively. From the results, it is observed that the average relative errors for the test samples are low (i.e., the generalization ability of the models is high) when the errors are estimated using the cross validation method. However, the cross validation method is much slower compared to the hold-out method.

For GA, the population size is taken to be ten times the number of the optimization variables. The crossover probability and the mutation probability are taken as 0.8 and 0.05 respectively. These are determined through a trial and error process. The hyperparameter values and the quality measures are

TABLE 4.14
Grid Search Technique Using 5-Fold Cross Validation Method

Model	σ^2	γ	$ARE(\%)$		R		T_{tr}
			Training	Test	Training	Test	(min)
ρ_1	4.10	326.32	1.27	1.33	0.999	0.999	583.12
ρ_2	2.76	112.04	2.37	2.42	0.980	0.970	583.62
ρ_3	5.33	142.65	1.82	1.85	0.998	0.998	582.67

TABLE 4.15
GA Technique Using Hold-Out Method

Model	σ^2	γ	$ARE(\%)$		R		T_{tr}
			Training	Test	Training	Test	(min)
ρ_1	2.38	250.13	2.16	3.38	0.999	0.998	12.06
ρ_2	5.62	480.19	2.12	3.82	0.994	0.961	10.83
ρ_3	5.19	140.15	1.98	2.90	0.999	0.998	11.56

reported in Table 4.15 and 4.16. From the results the above observations are also noted.

A comparison between the grid-search technique and the GA-based technique with respect to accuracy (ARE), correlation coefficient (R) and required training time is made in Table 4.17. All the experiments are performed on a PC with PIV 3.00 GHz processor and 512 MB RAM. We observe from the comparison that the accuracy of SVM models constructed using the grid search technique and the GA-based technique are almost same. However, the GA-based technique is at least ten times faster than the grid search method. The construction cost of the GA-based method is much lower than the grid search-based method, since the data generation time is the same for both methods.

The scatter plots of SPICE-simulated and LS-SVM estimated values for normalized test data of the three models are shown in Fig. 4.24(a), Fig. 4.24(b) and Fig. 4.24(c) respectively. These scatter plots illustrate the correlation between the SPICE simulated and the LS-SVM estimated test data. The

TABLE 4.16
GA Technique Using 5-Fold Cross Validation

Model	σ^2	γ	$ARE(\%)$		R		T_{tr}
			Training	Test	Training	Test	(min)
ρ_1	3.98	350.13	1.35	1.36	0.999	0.999	46.66
ρ_2	3.02	150.19	2.12	3.02	0.994	0.980	44.83
ρ_3	5.32	540.15	1.81	1.90	0.999	0.990	46.61

TABLE 4.17
Comparison between GA and Grid Search (GS) Algorithm (Algo) for LS-SVM Construction

Model	Algo	σ^2	γ	$ARE(\%)$		R		T_{tr}
				Training	Test	Training	Test	(min)
ρ_1	GA	2.38	250.13	2.16	3.38	0.999	0.998	12.06
	GS	3.43	173.26	1.82	2.48	0.999	0.998	118.19
ρ_2	GA	5.62	480.19	2.12	3.82	0.994	0.961	10.83
	GS	2.10	112.04	2.32	4.18	0.980	0.905	117.83
ρ_3	GA	5.19	140.15	1.98	2.90	0.999	0.998	11.56
	GS	5.43	387.55	2.02	3.14	0.999	0.937	118.13

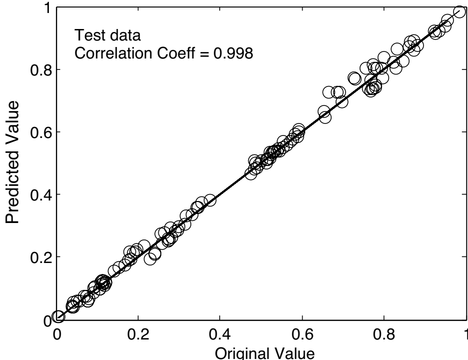
correlation coefficients are very close to unity. Perfect accuracy would result in the data points forming a straight line along the diagonal axis.

4.14 Meet-in-the-Middle Approach for Construction of Architecture-Level Feasible Design Space

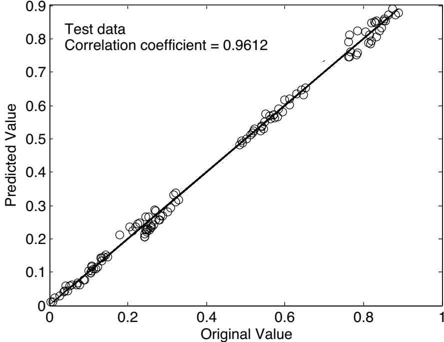
For architecture-level specification translation problems, the feasible design space is defined by the specification parameters of the component blocks of the chosen architecture of the system. The specification parameters need to be feasible for the chosen application and at the same time these should be realizable when the individual component blocks are implemented by transistors. This section presents a meet-in-the-middle approach for identification of the feasible design space at the architecture-level of design abstraction [139]. Two design spaces are identified: (i) application bounded space $\mathcal{D}_a(\bar{X})$ and (ii) circuit realizable space $\mathcal{D}_c(\bar{X})$, where $\bar{X}$ denotes the set of specification parameters of a component block at the architecture-level of design abstraction. The intersection of these two spaces defines the feasible design space. The construction of these two design spaces is described below.

4.14.1 Application Bounded Space $\mathcal{D}_a$

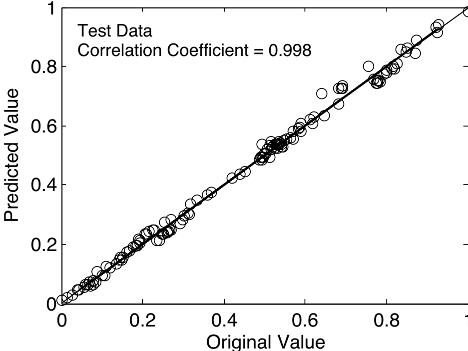
The application bounded space is defined by a set of constraints $f_a(\bar{X}) \leq 0$. The construction of this space is discussed with a simple example[139]. Consider a system with a voltage amplifier and a low pass filter connected in series. Let the desired specifications of the system be: total gain $A_T \in [A_{TL}, A_{TU}]$ and design constraints (1) maximum input signal = VmV, (2) bandwidth = $f_c KHz$ and (3) input signal frequency = $f_{in} MHz$. Suppose the design variable vectors are $\bar{X}_1 = (A_1, Lin_1, B_1)$ and $\bar{X}_2 = (A_2, Lin_2, B_2)$. These are the specification parameters of the component blocks, i.e., gain, input



(a) noise model



(b) power model



(c) impedance model

FIGURE 4.24
Scatter plot of the constructed models.

linearity (the range of the input signal in which the transfer characteristics of the circuit is linear) and bandwidth respectively. The following constraints are derived from circuit knowledge.

$$A_{TL} \leq A_1 \times A_2 \leq A_{TU} \tag{4.65}$$

$$A_1 - nA_2 = 0 \quad n = 1, 2.. \tag{4.66}$$

$$Lin_1 > V \tag{4.67}$$

$$B_1 > f_{in} \tag{4.68}$$

$$Lin_2 > A_1 \times V \tag{4.69}$$

$$B_2 = f_c \tag{4.70}$$

In (4.67), the value of n depends upon the designer's experience. The interaction between the gain of the two blocks is captured by these equations.

The problem of constructing the application bounded space $\mathcal{D}_a$ may be considered to be finding solutions of $f_a(\bar{X}) \leq 0$ over an interval of $\bar{X}$. An approach to solve this is through interval analysis technique. The interval analysis technique is based on the concepts of interval arithmetic [74]. In interval arithmetic, real numbers are replaced by intervals which are combinations of a lower bound and an upper bound on the allowable value range of a variable. All basic arithmetic operations like addition, multiplication, etc., are replaced by interval versions. Whenever there is more than one variable in the problem, the solution is enclosed within a multidimensional interval rectangle. The commonly used methods for solving equations/inequalities using interval analysis technique are the Krawczyk method, and the Hansen and Sengupta method [74]. The application bounded space $\mathcal{D}_a$ for a component block is constructed by combining the interval rectangles corresponding to all the specification parameters. The application bounded space is therefore constructed in a top-down fashion. This space is represented by a hyperbox as is shown in Fig. 4.25.

4.14.2 Circuit Realizable Space $\mathcal{D}_c$

A set of discrete tuples of circuit realizable specification parameters constitutes the circuit realizable space $\mathcal{D}_c$. This is constructed through the data generation technique discussed earlier. Each component block is implemented through transistors and is simulated through SPICE. A set of constraints is applied on the transistor sizes as well as on circuit performances to extract feasible tuples only. The applied circuit performance constraints are taken to be relatively weak compared to the box constraints derived in the top-down phase in order to ensure that several extracted parameter tuples lie within the hyperspace $\mathcal{D}_a$. The space $\mathcal{D}_c$ is thus constructed using a bottom-up approach.

4.14.3 Feasible Design Space Identification

The feasible design space $\mathcal{D}$ is a subset of the application bounded space, which is circuit realizable, as shown in Fig.4.25 [139]. The tuples of design parameters

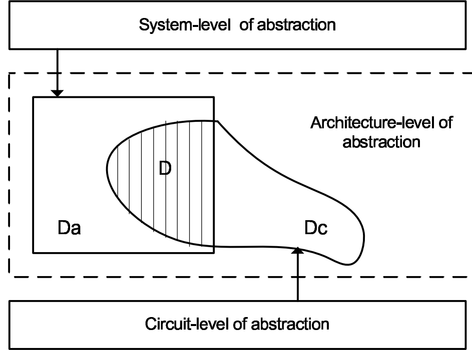


FIGURE 4.25
Meet-in-the-middle way of constructing the feasible design space $\mathcal{D}$[139].

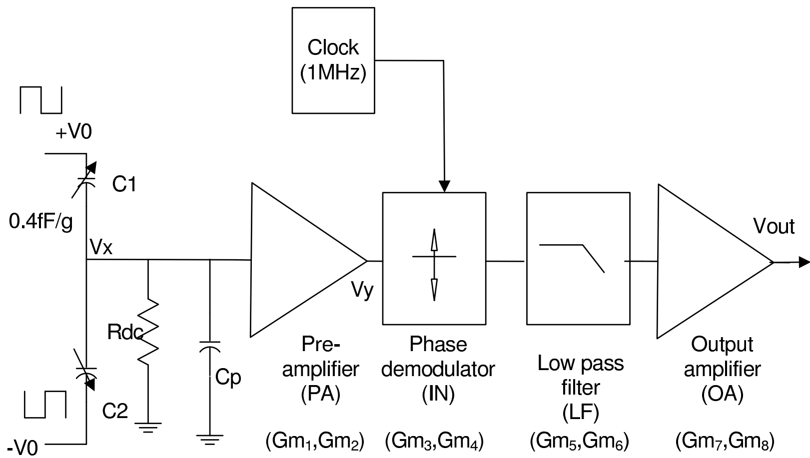
TABLE 4.18
Functional Specifications and Design Constraints of Case Study 5

Parameters	Desired Specs
Sensing Capacitance	100 fF
Capacitance Sensitivity	0.4 fF
Linear Range	$\pm$ 6 g
Modulation Frequency	1MHz
Modulation Voltage	500m V
Input Voltage Sensitivity	$1mV/g$
Output Voltage Sensitivity	$\geq$ 100 mV/g
Cut-Off Frequency	$\leq$ 45 KHz

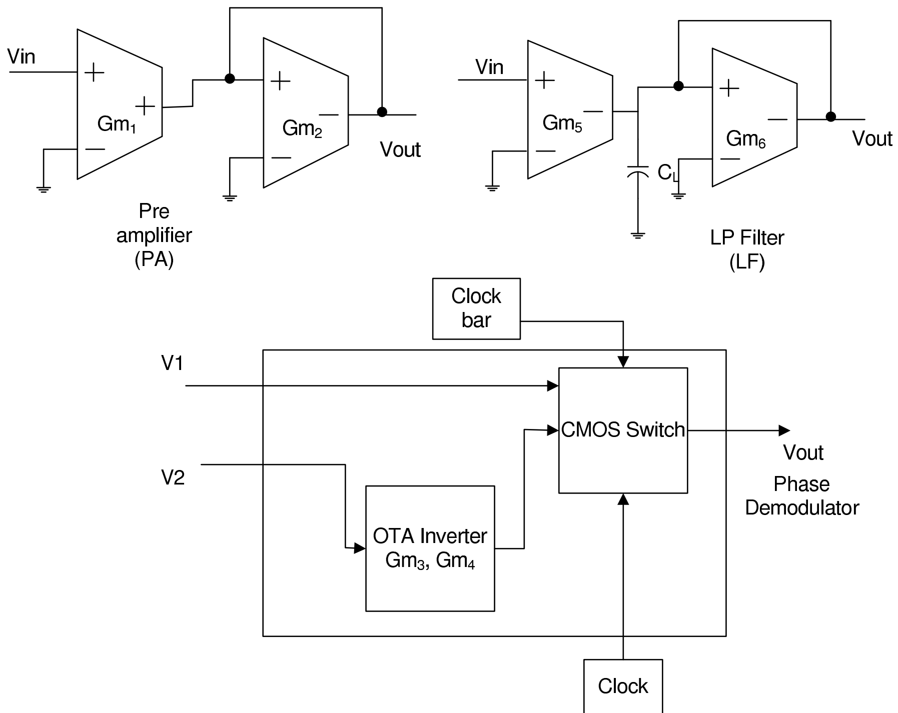
which lie within $\mathcal{D}$ are considered as feasible tuples and the rest as infeasible tuples. A two class LS-SVM classification technique is to be used to accurately identify the actual geometry of $\mathcal{D}$. The separating boundary between the two classes of tuples (feasible and infeasible) is implicitly described by a binary classification function $\mathcal{F}_i(\bar{X}_i) \rightarrow \{1, -1\}$. The value '1' signifies the feasible tuples whereas the value '-1' signifies the infeasible tuples.

4.15 Case Study 5: Construction of Feasibility Model at Architecture Level of an Interface Electronics for MEMS Capacitive Accelerometer System

The meet-in-the-middle approach presented above is illustrated with a case study [139]. Let us consider as an example of a complete system –the interface



(a) Voltage sensing configuration of the interface electronics for MEMS capacitive sensor



(b) OTA-C realizations of amplifier and filter

FIGURE 4.26
Considered system for Case Study 5.

electronics for a MEMS capacitive accelerometer sensor. The block diagram of the architecture of the system is shown in Fig. 4.26(a)[140]. In this configuration, a half-bridge consisting of the sense capacitors C_1, C_2 is formed and driven by two pulse signals with 180° phase difference. The amplitude of the bridge output V_x, is proportional to the capacitance change ΔC and is amplified by a voltage amplifier. The final output voltage V_{out}, is given by

$$V_{out} = V_0 \frac{2\Delta C}{2C_0 + C_p} A_v \tag{4.71}$$

where C_0 is the nominal capacitance value, C_p is the parasitic capacitance value at the sensor node, V_0 is the amplitude of the applied AC signal, and A_v is the gain of the system, depending upon the desired output voltage sensitivity. The topology employs a chopper modulation technique for low $1/f$ noise. The functional specifications and design constraints for the system are based on [204] and are listed in Table 4.18.

The synthesizable components are the pre-amplifier (PA), inverter (IN) of the phase demodulator, low pass filter (LF) and the output amplifier (OA). These are designed using OTAs and capacitors. The selected design variables are gain (A), input linearity (Lin), bandwidth (BW) and output swing (OS) of all the synthesizable blocks. The target output sensitivity A_D is considered as an interval $[105, 205]$. This is related to the gain of the individual components as

$$105 \leq A_D = A_{PA} \times A_{IN} \times A_{LF} \times A_{OA} \leq 205 \tag{4.72}$$

where A_{IN} and A_{LF} are the components whose gain is ideally unity and hence are represented in the interval $[0.9, 1.1]$. The gain of the pre-amplifier block PA is assumed to be $A_{PA} = 2.0 \times A_{OA}$. The intervals of the gain parameters of the individual component blocks are determined through the interval analysis method using MATLAB toolbox. The maximum input signal amplitude is determined from the input sensitivity and linear range and is given as $V = 6mV$. The linearity of the PA block should be such that it can accommodate the input signal within the linear range. Therefore, the lower bound of the input linearity parameter of the PA block is considered to be $Lin_{PA} = 2.5 \times V$. The same for the IN, LF and OA blocks are fixed at $A_{PAU} \times V, A_{PAU} \times A_{INU} \times V, A_{PAU} \times A_{INU} \times A_{LFU} \times V$ where $A_{PAU}, A_{INU}, A_{LFU}$ are the upper bounds of the corresponding gain parameters. The upper bound of the intervals of the linearity parameter of all the PA, IN, and OA component blocks are fixed at half of the supply voltage. The lower bound of the intervals of the bandwidth parameters are fixed at $2MHz$. The bandwidth of the LF block is equal to the cut-off frequency $= 45KHz$, fixed within an interval. For the swing OS variable, application constraints have not been imposed. The application bounded constraints for the chosen system are summarized in Table 4.19. These constraints define the space D_a. It is clear from the discussion herein that the application bounded constraints are determined through CAD techniques (interval analysis methods) as well as through designer's knowledge.

TABLE 4.19
Application Bounded Constraints: Case Study 5

Params	PA	IN	LF	OA
A	$[16.1, 18.40]$	$[0.9, 1.1]$	$[0.9, 1.1]$	$[8.05, 9.20]$
Lin (mV)	$[15, 900]$	$[276, 900]$	$[303.6, 900]$	$[334, 900]$
BW(MHz)	$[2, 10]$	$[2, 40]$	$[0.0447, 0.0453]$	$[2, 20]$

TABLE 4.20
Circuit Realizable Constraints: PA Block of Case Study 5

Sizes (L=1 um)	Ranges
$W_1 = W_2$	$[1\mu m$, 400 μm]
$W_3 = W_4 = W_6 = W_7$	$[1\mu m$, 100 μm]
$W_8 = W_9$	$[1\mu m$ 100 μm]

The geometry constraints on the OTA circuit used to realize the PA block are summarized in Table 4.20. A large set of circuit realizable specification data corresponding to the gain, linearity, bandwidth, and output swing with wide range of values is generated through SPICE simulation. Those data which satisfy the application bounded constraints are feasible data. In other words, those data which lie within the intersection space $\mathcal{D}$ in Fig. 4.25. Because all the component blocks have identical circuit topology, the dataset is reused. The statistical performances of the SVM feasibility models are reported in Table 4.21.

TABLE 4.21
SVM Performances: Case Study 5

Block	# Test data	σ^2	γ	Sen	Sp	Acc
PA	450	5.2	800	0.986	0.997	0.993
IN	560	3.8	80	0.972	0.936	0.992
LF	460	4.8	725	0.968	0.995	0.993
OA	540	5	800	0.988	0.963	0.992

4.16 Summary and Conclusion

This chapter described the methodology for construction of performance and feasibility models using a learning-based approach. The learning networks considered in this chapter are the artificial neural network and least squares support vector machine. Preliminary theoretical background on ANN and LS-SVM has been presented. This is followed by detailed discussion on the techniques for development of the ANN model. The same methodology is applicable for LS-SVM-based models. The ANN-based modeling technique has been demonstrated with two case studies highlighting practical results. The issue of dynamic sampling for training data generation has been discussed and a simple heuristic algorithm has been provided for determination of the optimal training dataset size. This chapter also provides detailed description about the construction of feasibility models. The methodology has been demonstrated with practical case studies of useful analog circuits and systems.

5

Circuit Sizing and Specification Translation

5.1 Introduction

Circuit sizing is defined as the process of finding out the parameter values of the dimensions of all the transistors (channel lengths and widths) and the values of the resistors and capacitors in a circuit, such that the desired performance objectives are optimized subject to a set of constraints that have to be satisfied [68]. The circuit sizing task is defined at the cell design level of abstraction in the IC design flow. The similar task performed at the architecture design level of abstraction is referred to as the specification translation. The specification translation is defined as the task of mapping the specifications of a block (e.g., a converter) under design to the individual specifications of the sub-blocks (e.g., a comparator) of the chosen block topology, such that the complete block meets its performance objectives optimally subject to a set of constraints that have to be satisfied [66, 139]. The task of circuit sizing and specification translation are thus conceptually the same thing. The concept is represented in Fig. 5.1 [76]. This chapter presents a comprehensive overview of the fundamental concepts behind the analog circuit sizing procedure.

5.2 Circuit Sizing as a Design Space Exploration Problem

The circuit sizing problem is formally cast as a design space exploration problem, which has been briefly introduced in Chapter 2 of this text. This is considered in detail below.

5.2.1 Problem Formulations

The design variables refer to those variables which are used by the circuit designers as decision variables. Design objectives include functional objectives and performance objectives. The functional objectives need to be met by the design in order to be functionally right. The performance objectives, on the

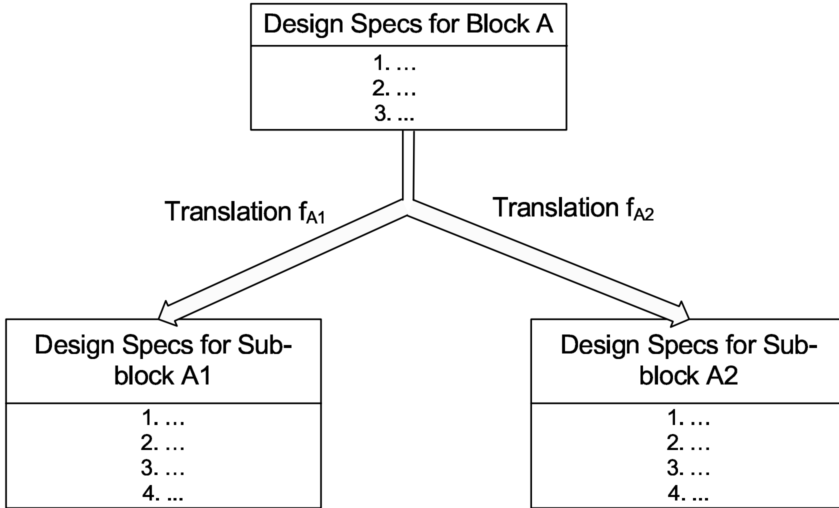


FIGURE 5.1
General design flow of a circuit sizing/specification translation task.

other hand, need to be minimized (or maximized). Let us consider the circuit of an output buffer, as shown in Fig.5.2 for illustration purpose [129]. The design variable set $\bar{\alpha}$ includes transistor dimensions and passive component values. The possible functional objective set is [129] $\bar{\rho_f}$ which includes (1) DC Gain ($A_0 >$ target), (2) input capacitance ($C_{in} <$ target), (3) 3-dB frequency ($f_{3dB} >$ target), and (4) output swing (target $< OS <$ target). The performance objective set $\bar{\rho_f}$ may be power consumption P. The design variable set $\bar{\alpha}^T = \{\alpha_1, \alpha_2, ..., \alpha_n\}^T$ defines a multi-dimensional design space. Each of the design variables α_i is bounded within an upper and a lower boundary. The functional objective and the performance objective parameters are expressed as functions of the design variables, i.e., $A_0(\bar{\alpha}), C_{in}(\bar{\alpha}), f_{3dB}(\bar{\alpha}), OS(\bar{\alpha}), P(\bar{\alpha})$. Then the problem of circuit sizing is formulated as a constrained optimization problem; in particular, for the case of the buffer of Fig. 5.2 as,

$$\begin{aligned}
\text{Minimize} \quad & P(\bar{\alpha}) \\
\text{subject to} \quad & A_0(\bar{\alpha}) > \text{target} \\
& C_{in}(\bar{\alpha}) < \text{target} \\
& f_{3dB}(\bar{\alpha}) > \text{target} \\
& \text{target} < OS(\bar{\alpha}) < \text{target} \\
\text{and} \quad & \alpha_{iL} < \alpha_i < \alpha_{iU} \quad i = 1, 2, ..., n
\end{aligned} \tag{5.1}$$

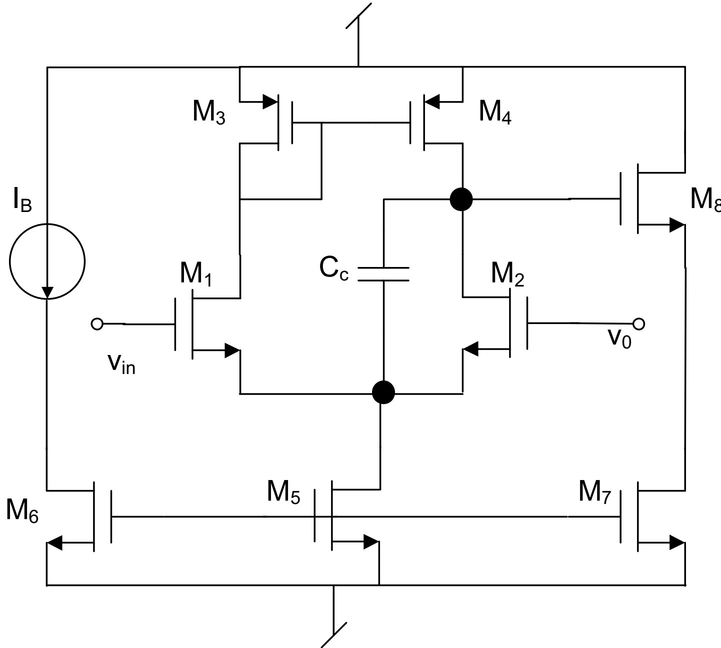


FIGURE 5.2
A CMOS output buffer circuit.

5.2.2 Solution Techniques

There are two broad categories of techniques for solving problems of (5.1). These are (1) analytical techniques and (2) iterative techniques. Unfortunately, even for elementary analog circuits like that shown in Fig. 5.2 exact analytical solution to the sizing problem is not possible. The primary reasons are

1. The design equations (i.e., the functional relationships between the various functional and performance objectives on one hand and the various design variables on the other hand, which are nothing but the performance and feasibility models) often cannot be expressed accurately in pure analytic form, especially in the nano-scale regime.
2. Even if analytical representations of the design equations are possible, these are often highly nonlinear, and consequently unsolvable analytically. The situation becomes more complicated when the dimensions of the design variable space and the design objective space become large.
3. Even if the design equations are simple, analytical solution of (5.1)

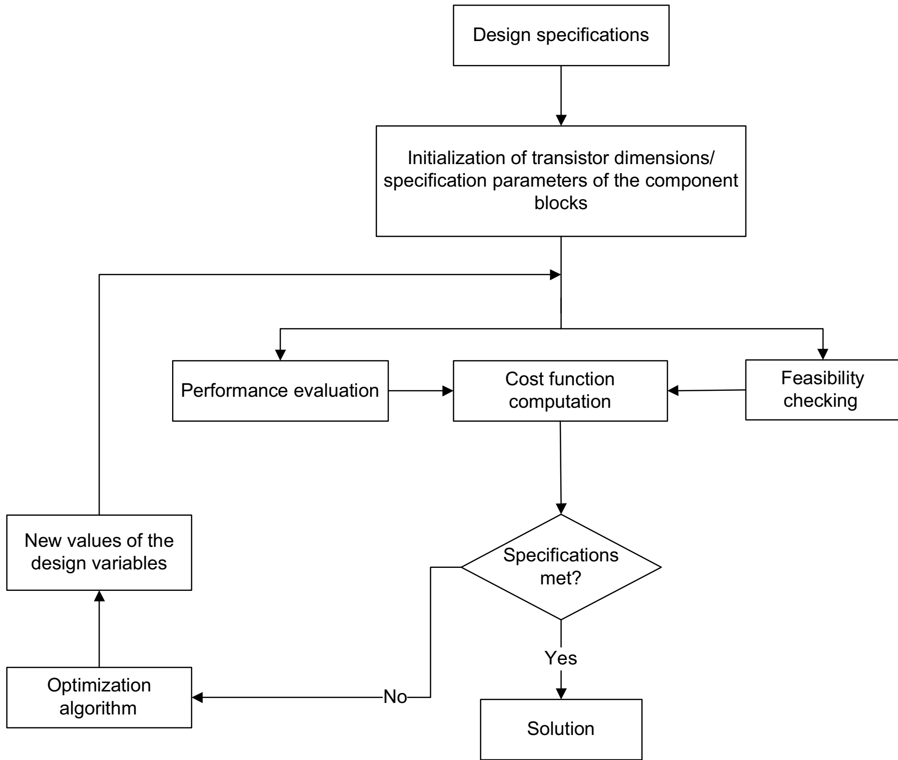


FIGURE 5.3
Design space exploration process for circuit sizing/specification translation task.

requires computation of the first and second derivatives of the design equations, which are in many cases very difficult to evaluate.

Therefore, analog circuits are most conveniently sized by using an iterative, dynamic process. This in turn can be done manually by the designers or through some automated design space exploration procedure.

5.2.3 Design Flow

The general flow of an automated design space exploration procedure for circuit sizing using optimization algorithm is shown in Fig. 5.3. The procedure starts with a set of design specifications. This consists of formal descriptions of the various functional and performance objectives and the boundaries of the design variables. The design variables are initialized to some random values within the boundaries. The various design equations (both performance models and feasibility models) are used to construct a cost function. The cost

function is evaluated with the initial values of the design variables and the evaluated results are checked against the design specifications. The subsequent values of the design variables are generated through some optimization algorithms. The process is done iteratively until a set of design variables is obtained for which the desired design specifications are satisfied.

The two important modules of said procedure are (1) evaluation of cost function through performance and feasibility models and (2) optimization algorithms. Chapter 4 of this book discusses in detail the construction of the performance and feasibility models. The implementation of the latter is discussed in the next section.

5.2.3.1 Evaluation of Cost Functions

The cost function is evaluated through some design equations. These design equations can be analytical equations or some learning network such as ANN and LS-SVM network, depending upon the complexity of the problem and the degree of accuracy required.

5.3 Particle Swarm Optimization Algorithm (PSO)

Particle Swarm Optimization is a population-based search algorithm inspired by the behavior of biological communities that exhibit both individual and social behavior; examples of these communities are flocks of birds, schools of fish, and swarms of bees. Kennedy and Eberhart introduced the concept of function-optimization by means of a particle swarm [99]. A swarm is a collection of individuals or particles. Particles are conceptual entities, which fly through the multi-dimensional design space. At any particular instant, each particle has a position and a velocity. The position vector of a particle with respect to the origin of the design space represents a trial solution of the design problem. Particles move randomly in the entire design space with velocity which is dynamically adjusted according to its own flying experience and the flying experience of the swarm. The movements of the particles are controlled by updating the position and velocity vectors of each individual particle in an effort to find the optimum solution. The position of each particle is represented by a set of coordinates in the n-dimensional (n being the number of design variables) design space of the exploration problem. The position vector of the ith particle in an n-dimensional design space is given as

$$\bar{x}_i = [x_{i1}, x_{i2}, ..., x_{in}]^T \tag{5.2}$$

and the corresponding velocity vector is given by

$$\bar{v}_i = [v_{i1}, v_{i2}, ..., v_{in}]^T \tag{5.3}$$

The position corresponds to the design variables that need to be optimized. At the beginning, a population of particles is initialized with random positions and random velocities. The population of such particles is called swarm S. At each time step, all particles adjust their positions and velocities, i.e., directions in which particles need to move in order to improve their current position, thus their trajectories. The changes in the positions of the particles in the design space are based on the social-psychological tendency of individuals to emulate the success of other individuals. Thus each particle tends to be influenced by the success of any other particle it is connected to.

5.3.1 Dynamics of a Particle in PSO

There are two main versions of the PSO algorithm, local and global. In the local version, each particle moves toward its best previous position and toward the best particle within a restricted neighborhood. In the global version of PSO, each particle moves toward its best previous position and toward the best particle of the whole swarm. The global version is actually a special case of the local version where the neighborhood size is the size of the swarm. The position and velocity of each particle is updated according to the following two equations [100]

$$\bar{v}_i^{(t+1)} = \omega.\bar{v}_i^{(t)} + c_1 r_1.\left(\bar{p}_{besti}^{(t)} - \bar{x}_i^{(t)}\right) + c_2 r_2.\left(\bar{g}_{besti}^{(t)} - \bar{x}_i^{(t)}\right) \quad (5.4)$$

$$\bar{x}_i^{(t+1)} = \bar{x}_i^{(t)} + \bar{v}_i^{(t+1)} \quad (5.5)$$

where $\bar{p}_{besti}$ represents personal best experience and $\bar{g}_{besti}$ represents the best position found so far in the neighborhood of the particle. When the neighborhood size is equal to the swarm size, $\bar{g}_{besti}$ is referred to as the globally best particle in the entire swarm.

The first term in the velocity updating formula represents the inertial velocity of the particle. ω is referred to as the "inertia factor". Since it is the tendency to maintain the previous direction, it is called inertia. The second term represents the competition between the personally best position $\bar{p}_{besti}$ that each individual particle has experienced and its current position. c_1 is termed as "self-confidence" [171]. The third term represents the particle's social cognition or cooperation between the globally best position that one particle of the swarm has found and the current position of the particle. c_2 is termed as "swarm confidence" [171]. r_1 and r_2 stand for a uniformly distributed random number in the interval $[0, 1]$. These are used to give diversity to the particles. The particle updates itself constantly sharing the information both from itself and the entire swarm in such a way that enables the particles to move toward the optimum solution. The dynamics of the particle in a PSO algorithm is illustrated in Fig. 5.4.

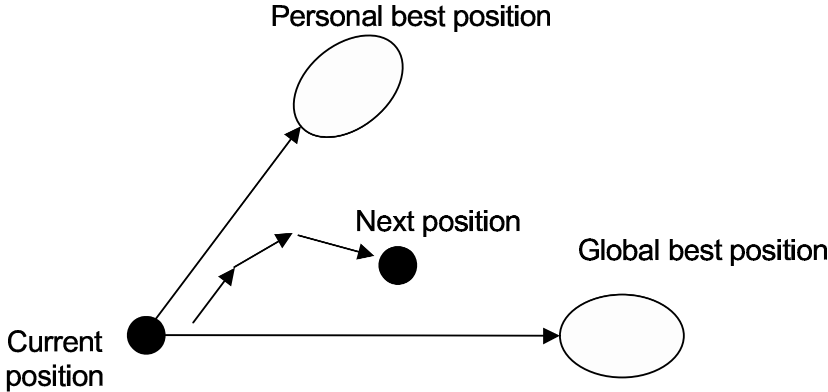


FIGURE 5.4
Illustration of the dynamics of a particle.

5.3.2 Flow of the Algorithm

The algorithm starts with random initialization of the position and velocity of all the particles in a swarm. At the start of the simulation run, this initial position of each particle is taken as the p_{best} of the respective particle and the first g_{best} is obtained from among these initial positions. A fitness function is used to search for the best position. As discussed in Chapter 2, this fitness function is computed based upon the cost function of the problem. The fitness function is computed based upon the current position of each particle. If the current fitness value of a particle is better than the corresponding p_{best}, then the p_{best} location is updated with the current position value. If the current best fitness value is better than the g_{best}, then g_{best} position is replaced by the current best position value of the entire swarm. The next position of each particle in the swarm is calculated based upon the dynamic equations (5.4) and (5.5). This iterative process continues until some termination criteria are satisfied. This process is iterated, in general, for a certain number of time steps, or until some acceptable solution has been found by the algorithms or until an upper limit of CPU usage has been reached. The flow chart of the PSO algorithm is shown in Fig. 5.5.

5.3.3 Selection of Parameters for PSO

The main parameters of a simple PSO algorithm are $\omega, c_1, c_2, V_{max}$ and the swarm size S. The settings of these parameters determine how it optimizes the design space exploration. However, it may be noted that the same parameter settings do not guarantee success in different problems. Therefore, it is essential for the designers to understand the effects of the different settings, so that it is possible to pick a suitable setting from problem to problem [39].

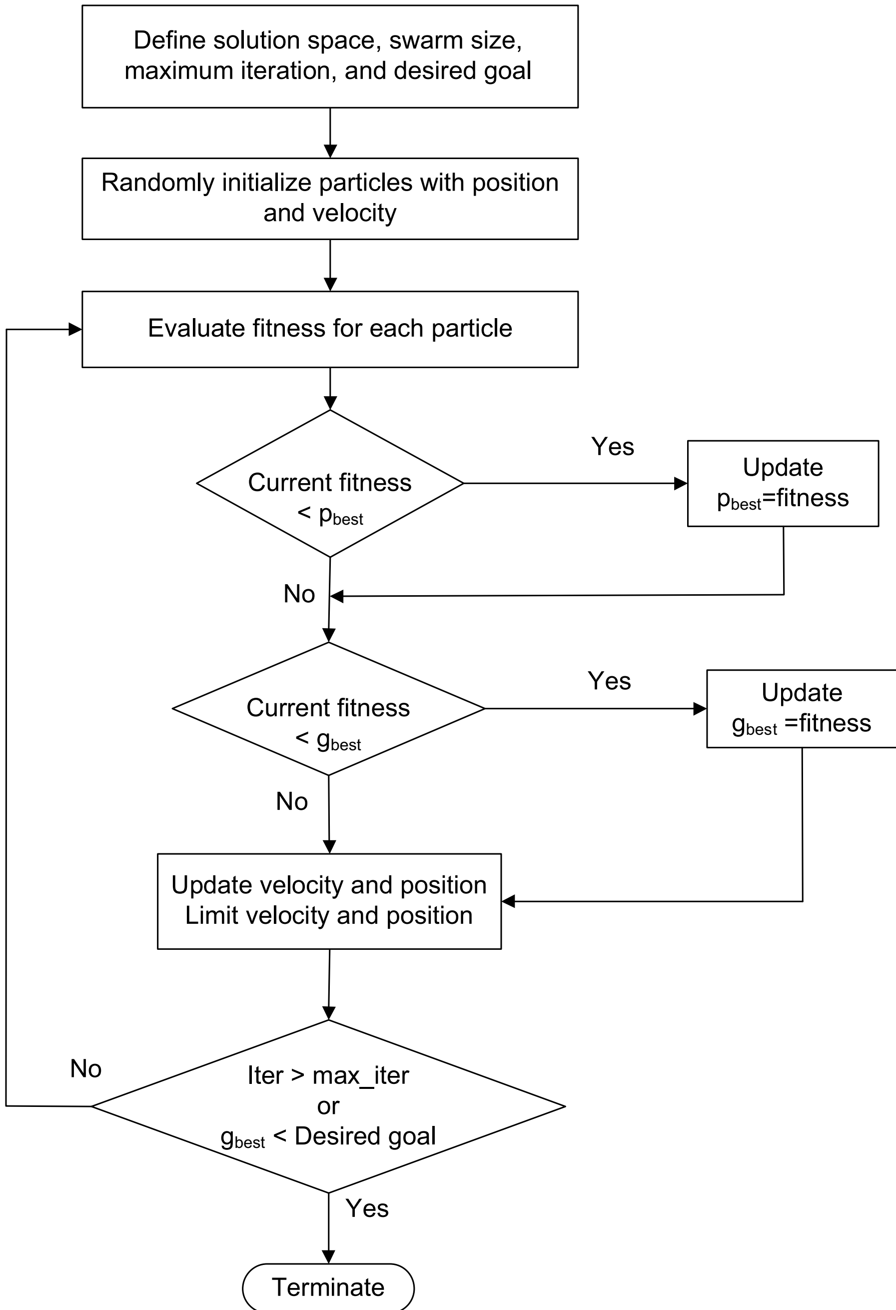


FIGURE 5.5
Flow chart of the PSO algorithm.

5.3.3.1 Inertia Weight ω

The momentum of the particle is controlled by the inertia weight. Therefore, if $\omega << 1$, quick changes of direction of the movement of a particle is possible. If $\omega = 0$, the concept of velocity is lost and the particle moves in each step without the knowledge of the previous velocity. On the other hand, setting ω high (> 1) produces the same effect as setting c_1 and c_2 low. The particles can hardly change their directions which implies a larger area of exploration as well as a reluctance against convergence toward optimum. Therefore, in short high settings near 1 facilitate global search, and lower settings in the range $[0.2, 0.5]$ facilitate rapid local search [39]. It has been reported that when V_{max} is not small (≥ 3), an inertia-weight of 0.8 is a good choice [170].

5.3.3.2 Maximum Velocity V_{max}

The maximum change that one particle can undergo in its positional coordinates during an iteration is determined by maximum velocity V_{max}. The commonly used approach is to set the entire range of the design space as the maximum velocity V_{max}. The original idea of using this parameter is to avoid explosion and divergence. However, with the use of ω in the velocity update formula, the maximum velocity parameter becomes unnecessary to some extent. Therefore, sometimes this parameter is not used. In spite of this fact, the maximum velocity limitation can still improve the search for optima in many cases [39].

5.3.3.3 Swarm Size S

A common practice in selecting the swarm size is to limit the number of particles to the range 2060 [39]. It has been shown that though there is a slight improvement of the optimal value with increasing swarm size, a larger swarm increases the number of function evaluations to converge to an error limit.

5.3.3.4 Acceleration Coefficient c_1 and c_2

An usual choice for the acceleration coefficients c_1 and c_2 is to take $c_1 = c_2 = 1.494$ [98]. An extensive study of the acceleration factor of PSO can be found in [170]. Some researchers prefer to change these parameters in an adaptive manner as follows [148]:

$$c_1 = (c_{1f} - c_{1i}).\frac{iter}{MAXITER} + c_{1i} \tag{5.6}$$

$$c_2 = (c_{2f} - c_{2i}).\frac{iter}{MAXITER} + c_{2i} \tag{5.7}$$

where c_{1i}, c_{1f}, c_{2i} and c_{2f} are constants, $iter$ is the current iteration number and $MAXITER$ is the number of maximum allowable iteration. The basic idea behind the adaptive change of the acceleration coefficients is to boost

the global search over the entire search space in the initial part of the search procedure and to encourage the particles to converge to global optima at the end of the search.

5.4 Case Study 1: Design of a Two-Stage Miller OTA

The two-stage OTA circuit that has been considered in the present case study is shown in Fig.5.6. The design variables are the transistor dimensions, bias current and the compensation capacitor C_c. The various design specifications are (1) open loop gain A_v, (2) gain-bandwidth GBW, (3) slew rate SR, (4) input common mode range $ICMR$, (5) output voltage swing and (6) power dissipation P_{diss}.

The various design equations related to the manual sizing of the two-stage CMOS OTA are summarized below based on [5]

1. From the desired phase margin, i.e., for a 60° phase margin, it is

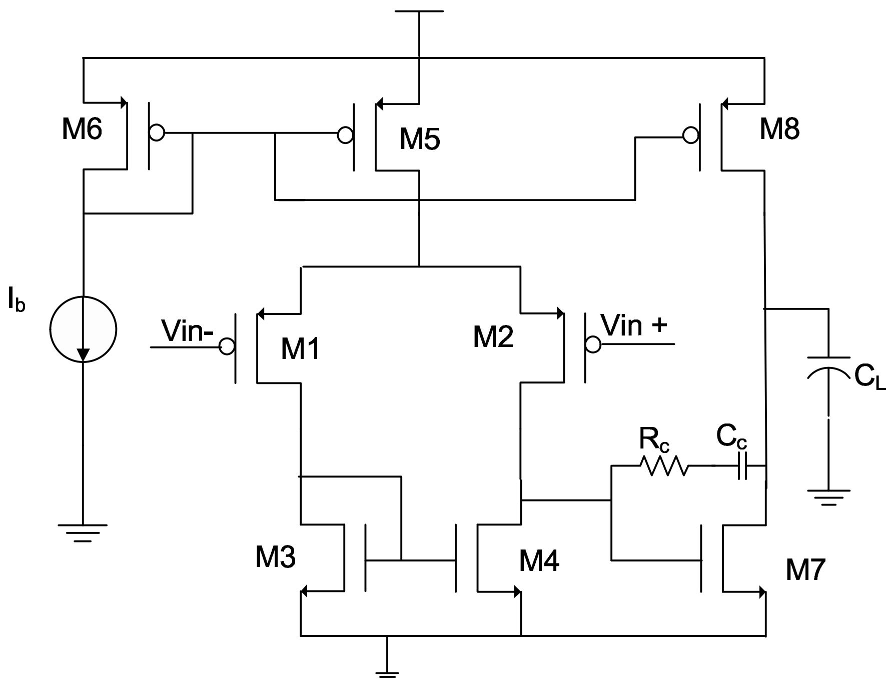


FIGURE 5.6
Schematic diagram of the Miller OTA.

chosen that

$$C_c > 0.22C_L \tag{5.8}$$

where C_c is the compensation capacitor and C_L is the load capacitor.

2. The tail current is selected from the slew rate requirement

$$I_5 = SR.C_c \tag{5.9}$$

3. Find g_{m1} from GBW and C_C using the formula

$$g_{m1} = GBW.C_c \tag{5.10}$$

Thereafter $(W/L)_1$ is calculated as

$$\left(\frac{W}{L}\right)_1 = \left(\frac{W}{L}\right)_2 = \frac{g_{m1}^2}{2K_pI_1} = \frac{g_{m1}^2}{K_pI_5} \tag{5.11}$$

where $I_1 = I_5/2$ and $K_p = \mu_0 C_{ox}$

4. From the positive ICMR requirement

$$V_{IC}(max) = V_{DD} - V_{SD5}(sat) - V_{SG1} \tag{5.12}$$

The current flowing through M1 is $I_5/2$. Therefore, it can be written that

$$V_{IC}(max) = V_{DD} - \sqrt{\frac{I_5}{\beta_1}} - |V_{T1}| - V_{SD5}(sat) \tag{5.13}$$

from which $V_{SD5}(sat)$ is calculated. Thereafter, $(W/L)_5$ is calculated as

$$\left(\frac{W}{L}\right)_5 = \frac{2I_5}{K_pV_{SD5(sat)}^2} \tag{5.14}$$

5. From the negative ICMR requirement, it follows that

$$V_{IC}(min) = V_{S1} - V_{SG1} = V_{SD1} + V_{GS3} - V_{SG1} = V_{GS3} - |V_{T1}| \tag{5.15}$$

since the transistor M1 is to remain in saturation. From this V_{GS3} is calculated, thereafter β_3 and hence $(W/L)_3$ is calculated using the following

$$V_{GS3} = \sqrt{\frac{I_5}{\beta_3}} + V_{T3} \tag{5.16}$$

Since $V_{GS3} = V_{GS4}$ and M3 and M4 form a current mirror, we have

$$\left(\frac{W}{L}\right)_3 = \left(\frac{W}{L}\right)_4 \tag{5.17}$$

6. For $60°$ phase margin, it is required that

$$g_{m7} \geq 10 \times g_{m1} \tag{5.18}$$

To achieve proper mirroring of the first stage current mirror load, it is required to ensure that $V_{GS3} = V_{GS7}$. It is easy to show that this requires

$$\left(\frac{W}{L}\right)_7 = \left(\frac{W}{L}\right)_3 . \frac{g_{m7}}{g_{m3}} \tag{5.19}$$

where g_{m3} is found out from

$$g_{m3} = \frac{2I_3}{V_{GS3} - V_{T3}} = \frac{I_5}{V_{GS3} - V_{T3}} \tag{5.20}$$

Hence calculate $\left(\frac{W}{L}\right)_7$

7. The current through M7 is calculated as

$$I_7 = \frac{g_{m7}^2}{2K_n \left(\frac{W}{L}\right)_7} \tag{5.21}$$

Since $I_7 = I_8$, and $V_{SG8} = V_{SG5}$, $\left(\frac{W}{L}\right)_8$ is calculated as

$$\left(\frac{W}{L}\right)_8 = \left(\frac{W}{L}\right)_5 . \frac{I_7}{I_5} \tag{5.22}$$

8. The gain is calculated as

$$A_v = \frac{2g_{m2}g_{m7}}{I_5 \left(\lambda_2 + \lambda_4\right) I_7 \left(\lambda_7 + \lambda_8\right)} \tag{5.23}$$

The PSO is utilized for design specifications of gain $A_v \geq 45dB$, phase margin $> 45^0$, $GBW \geq 2.5MHz$ $0.05V \leq ICMR \leq 0.6V$, $SR = 5V\mu s$. The target objective is to reduce the total MOS transistor area to smaller than $3\mu m^2$.

Since the target gain is not high and the target area is very small, it is preferred to take the channel length to be $0.1\mu m$. The design variables are the channel widths W and the compensation capacitor C_c. The inputs of PSO are set as $V_{DD} = 1V$, $V_{Tn} = 0.466V, |V_{Tp}| = 0.412V, K_n = 239\mu A/V^2, K_p = 36\mu A/V^2$. The swarm size is 35, $\omega = 0.99, C_1 = C_2 = 2$. The constraint functions are based on the design equations outlined above. The equality constraints are used to reduce the number of design variables. The target values of the specifications are selected so as to have sufficient guard band. The design process is iterated over 1000 epochs with a total execution time of about

TABLE 5.1
Aspect Ratios of Each Transistor of Case Study 1

Transistors	W/L
M1,M2	6.7
M3,M4	27.3
M5,M6	10.2
M7	52.69
M8	9.826
C_L	2.2pF

TABLE 5.2
Comparison between PSO and Simulation Results of Case Study 1

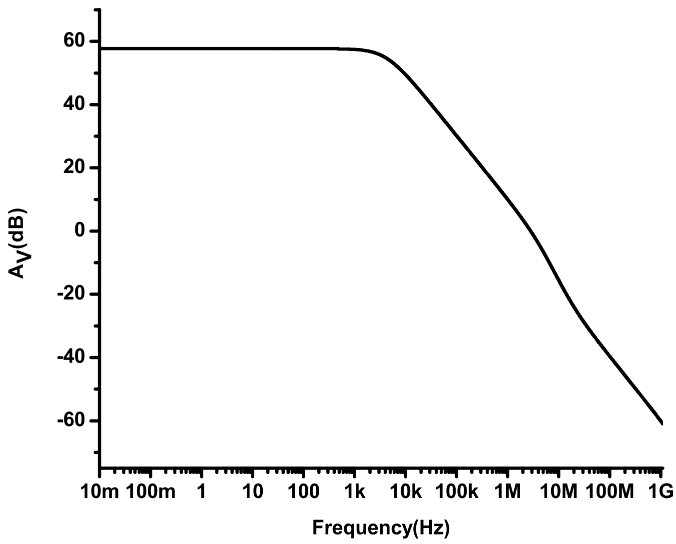
Parameters	PSO	SPICE
A_v	48 dB	57.7 dB
PM	60^0	50^0
GBW	3MHz	2.82MHz
CMRR		60.78 dB
ICMR	0.1 to 6V	0.05 to 0.60V
Slew rate	$5V/\mu s$	+ve $1.045V/\mu s$ and -ve $4.3V/\mu s$

5s with an Intel Core 2 duo processor. The (W/L) values of the various MOS transistors are tabulated in Table 5.1.

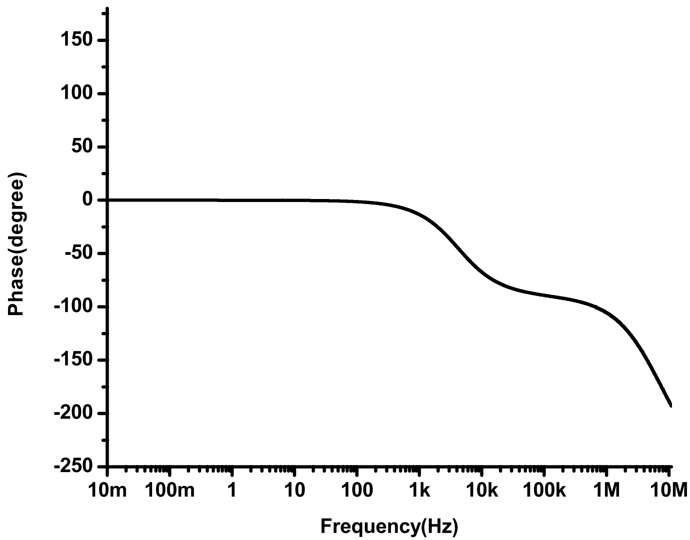
In order to validate the synthesized results, the design is implemented in a SPICE environment and simulated with 45nm CMOS technology with 1V supply. The simulation results are shown in Fig. 5.7(a), 5.7(b), 5.8(a), 5.8(b) and tabulated in Table 5.2

5.5 Case Study 2: Synthesis of on-Chip Spiral Inductors

This case study is in continuation of that described in Chapter 4 of this text. The task of on-chip spiral-inductor synthesis refers to the process of determining the layout geometric parameters from electrical specifications. The layout geometry parameters are (i) the outer diameter d, (ii) the number of turns N, (iii) the metal width W and (iv) the spacing between the metal traces s. The spiral-inductor-synthesis procedure helps the designer to make a trade-off analysis between the competing objectives, namely, Q, SRF, and outer diameter d, for a given L. The synthesis flow is shown in Fig. 5.9 [122]. The objective of the synthesis methodology is to find a set of layout parameters which will give the desired inductance value within acceptable error.

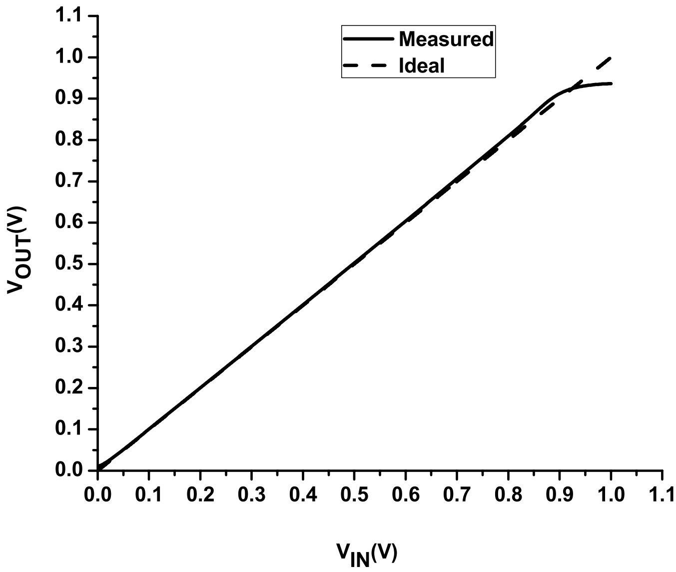


(a) Gain vs. frequency plot of the synthesized OTA in Case Study 1

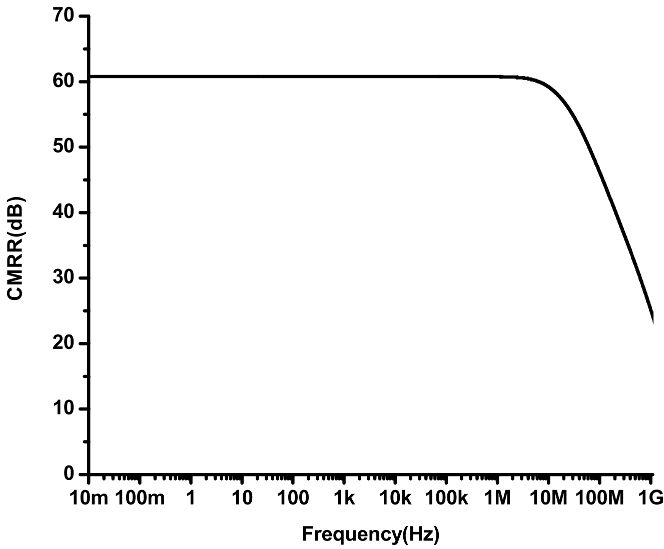


(b) Phase vs. frequency plot of the synthesized OTA in Case Study 1

FIGURE 5.7
AC simulation results of the synthesized OTA in Case Study 1.



(a) ICMR plot



(b) CMRR plot

FIGURE 5.8
ICMR and CMRR results of the synthesized OTA in Case Study 1.

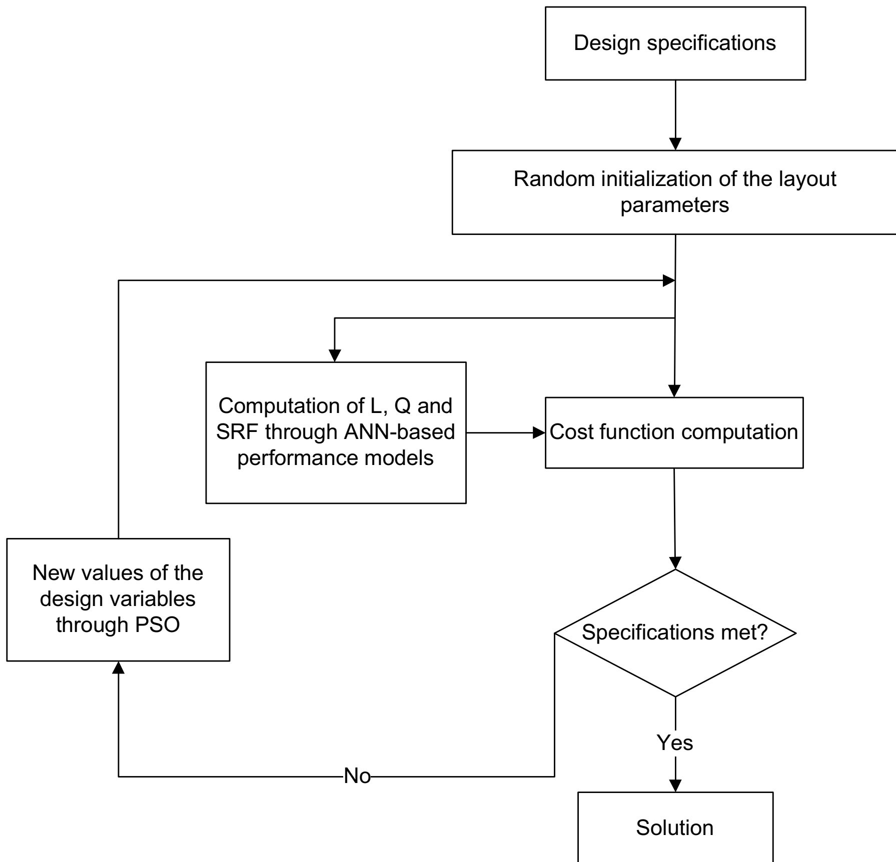


FIGURE 5.9
Flow chart of the on-chip spiral inductor synthesis problem.

TABLE 5.3
Synthesized Values of Inductor Layout Geometry Parameters

$L(nH)$	$d(\mu m)$	$W(\mu m)$	N	$s(\mu m)$	Q	$SRF(GHz)$
3.9999	275	15.1	4.0	2.4	3.6122	10.257
3.9968	284	13.1	3.0	1.4	3.639	11.587
4.0041	252	11.7	3.9	2.5	3.1102	11.207
4.0032	300	19.0	3.5	1.3	4.2953	9.1302

The design variable vector is $\bar{\alpha} = [d, N, W, s]^T$. The cost function is formulated as [122]

$$\begin{aligned} \text{Minimize} \quad & L_{target} - L_{ANN} \\ \text{subject to} \quad & N_{min} \leq N \leq N_{max} \\ & d_{min} \leq d \leq d_{max} \\ & W_{min} \leq W \leq W_{max} \\ & s_{min} \leq s \leq s_{max} \\ & d \geq 2N(W + s) - 2s \\ & SRF \geq SRF_{min} \end{aligned} \tag{5.24}$$

The variables bounds are $d = 100 - 300\mu m, W = 8 - 24\mu m, N = 2 - 6, s = 1 - 4\mu m$ and $SRF_{min} = 6GHz$.

The PSO algorithm generates a swarm of particles, each representing a combination of layout parameters in the given design space. For each combination of the design variables, the performance parameters are computed from a pre-constructed ANN-based performance model. Cost function is computed using these electrical parameter values. The design variables are then updated according to the minimum cost following the PSO algorithm. This process continues until a desired cost function objective is achieved or the maximum number of iterations is executed. The error value is set to $0.0001nH$ and the maximum number of iterations is taken to be 1000.

Table 5.3 shows the layout geometries of the inductors as synthesized by the proposed approach for a desired inductance value of $4nH$ at $1GHz$ operating frequency [122]. A set of sample 4 layout geometries are reported here. This helps the designer to make a trade-off between Q, area (outer diameter), and SRF. It is to be noted that it may not be feasible to fabricate all the inductor geometries synthesized by this approach due to the design rules of a particular process. For such cases the design values need to be rounded off to the nearest grid point while doing the layout. To validate the accuracy of the synthesis approach, the synthesized inductors are simulated with the IE3D EM simulator. The synthesized inductors satisfy the desired design specifications. This is demonstrated in Table 5.4. The L, Q, and SRF of these inductors were extracted from simulated S-parameters. The synthesized inductors show reasonable matching with the EM simulated results.

TABLE 5.4
Verification of the Synthesized Inductor Geometry through EM Simulation

	$L(nH)$	Q	$SRF(GHz)$	$d(\mu m)$	$W(\mu m)$	N	$s(\mu m)$
PSO	4.0032	4.2953	9.1302	300	19.0	3.5	1.3
EM	3.9389	4.150	9.5200	300	19.2	3.5	1.1
Error (%)	1.60	3.38	4.26				

5.6 Case Study 3: Design of a Nano-Scale CMOS Inverter for Symmetric Switching Characteristics

This case study, based on the published literature [131] presents a technique for the modeling and design of a nano-scale CMOS inverter circuit using an ANN and PSO algorithm such that the switching characteristics of the circuit is symmetric. This means that (i) the difference between the output rise time (τ_R) and fall time (τ_F) and (ii) the difference between the output propagation delay times, high-to-low (τ_{PHL}) and low-to-high (τ_{PLH}) should be minimum. The transistor channel widths W_n, W_p and the load capacitor C_L are the design parameters. The value of the rise/fall time of the input signal will be taken from the user. The problem is therefore written as [131]

$$\begin{aligned}
\text{Minimize} \quad & \frac{|\tau_F - \tau_R|}{\tau_F} + \frac{|\tau_{PHL} - \tau_{PLH}|}{\tau_{PHL}} \\
\text{subject to} \quad & (\tau_F)_{min} \leq \tau_F \leq (\tau_F)_{max} \\
& (\tau_R)_{min} \leq \tau_R \leq (\tau_R)_{max} \\
& (\tau_{PHL})_{min} \leq \tau_{PHL} \leq (\tau_{PHL})_{max} \\
& (\tau_{PLH})_{min} \leq \tau_{PLH} \leq (\tau_{PLH})_{max} \\
& 0.45 \times V_{SP} \leq V_{SP} \leq 0.55 \times V_{SP} \\
\text{and} \quad & (W_n)_{min} \leq W_n \leq (W_n)_{max} \\
& (W_p)_{min} \leq W_p \leq (W_p)_{max} \\
& (C_L)_{min} \leq C_L \leq (C_L)_{max}
\end{aligned} \tag{5.25}$$

The design flow of this circuit sizing problem is shown in Fig. 5.10. The various performance parameters are evaluated through an ANN-based performance model. Thus the PSO algorithm would result in the exact values of the design parameters which minimize the cost function value and satisfy the specified constraints. The supply voltage V_{DD} is taken to be 1.0V. The swarm size is taken to be 30. The acceleration parameters are taken as $c_1 = c_2 = 1.49618$ and the inertia weight factor is $\omega = 0.7298$. This ensures

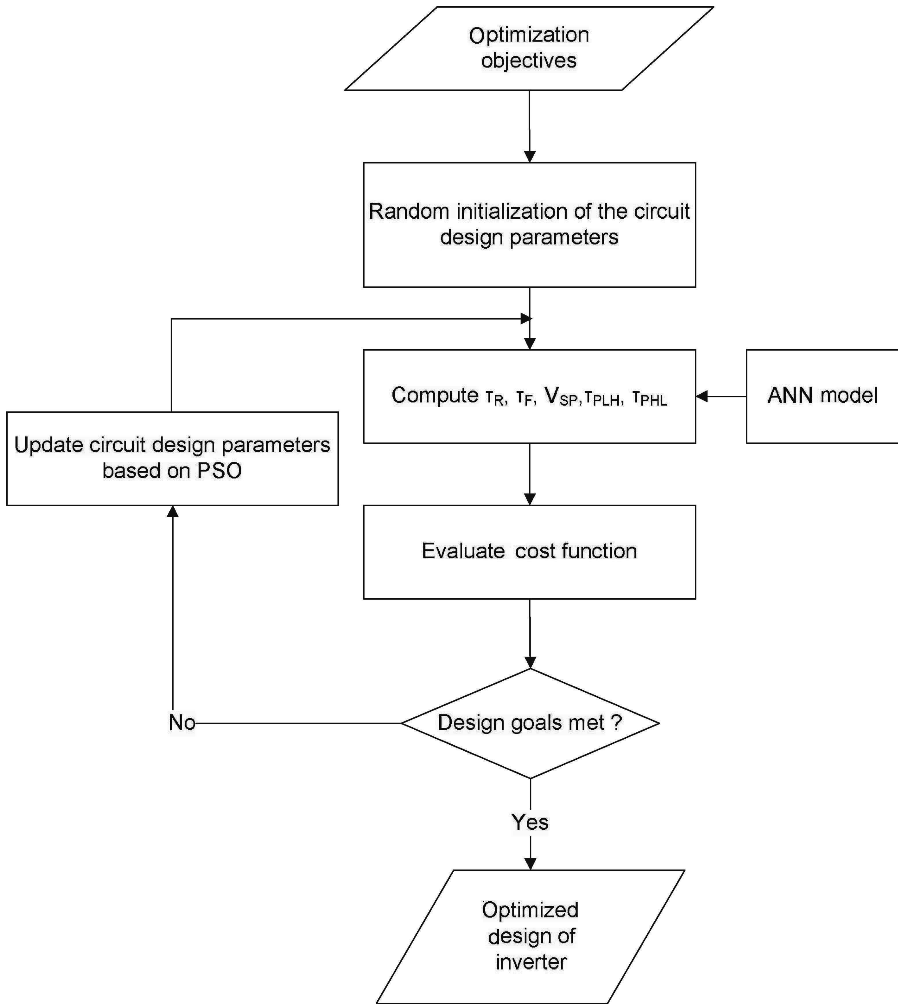


FIGURE 5.10
Flow chart of the inverter design problem.

TABLE 5.5
Delay Constraints and Design Parameter Bounds

Sample	$C_L(pF)$	$W_n(nm)$	$W_p(nm)$	$\tau_F(ns)$	$\tau_R(ns)$	τ_{PHL} (ns)	τ_{PLH} (ns)
1.	0.5-2.5	45-135	90-940	0.1-15	0.1-15	0.05-8.0	0.05-8.0
2.	0.5-2.5	45-110	90-620	0.1-15	0.1-15	0.05-8.0	0.05-8.0
3.	0.5-1.5	45-135	90-940	0.1-15	0.1-15	0.05-8.0	0.05-8.0
4.	1.0-3.0	60-160	160-945	0.1-15	0.1-15	0.05-8.0	0.05-8.0
5.	1.5-3.5	60-135	135-840	0.1-15	0.1-15	0.05-8.0	0.05-8.0
6.	0.3-2.0	45-90	90-540	0.1-8.0	0.1-8.0	0.05-6.0	0.05-6.0
7.	0.6-1.9	60-160	135-910	0.1-7.5	0.1-7.5	0.05-5.5	0.05-5.5

TABLE 5.6
Synthesis Results: $\tau_{in} = 1ns$

Sample	C_L (pF)	W_n (nm)	W_p (nm)	$\tau_F(ns)$	$\tau_R(ns)$	τ_{PHL} (ns)	τ_{PLH} (ns)	$V_{sp}(V)$
1.	0.83	128.37	221.93	5.1546	5.1363	2.5277	2.4648	0.4890
2.	0.76	69.44	100.08	10.5131	10.5820	4.8127	4.8025	0.4861
3.	0.81	125.82	217.53	5.0244	5.1146	2.6854	2.6731	0.4879
4.	1.02	157.91	273.01	5.2138	5.2015	2.5812	2.5637	0.4851
5.	1.66	134.70	232.88	8.8279	8.8588	4.6977	4.6785	0.4800
6.	0.42	65.42	113.10	5.4471	5.4233	2.5805	2.5332	0.4887
7.	0.61	95.52	165.14	5.3867	5.3678	2.7219	2.7581	0.4876

good convergence of the PSO algorithm. The maximum number of iterations that has been considered is 1000.

A set of seven samples has been chosen. For each sample, the desired rise time, fall time, low-to-high and high-to-low output propagation delay times are kept within a certain constraint, defined by an upper limit and a lower limit. Similarly, the design parameters are also kept within a specified bound. These are tabulated in Table 5.5 [131]. The value of the input rise time/fall time τ_{in} is assumed to be $1ns$. The synthesized values of the design parameters corresponding to which the cost function is minimized and the constraints are satisfied for all the case studies, are shown in Table 5.6 [131]. It also contains the corresponding values of the performance parameters. It is observed from Table 5.5 and 5.6, that the synthesized parameters satisfy the design constraints.

In order to validate the results obtained through PSO optimization, the design samples are selected and implemented at the transistor level. The PSO synthesized transistor widths and output load capacitor values have been considered. The channel length is taken as 45nm with 1.0V supply. Transient simulation is then performed using SPICE simulation. A comparison between the PSO generated results and SPICE results is provided in Table. 5.7-5.8

TABLE 5.7
Comparison between PSO Results and SPICE Results: τ_R and τ_F

Sample	PSO Results			SPICE Results		
	$\tau_R(ns)$	$\tau_F(ns)$	Diff(ns)	$\tau_R(ns)$	$\tau_F(ns)$	Diff(ns)
1.	5.1363	5.1546	0.0183	5.0674	5.2532	0.1858
2.	10.5820	10.5131	0.0689	10.6200	9.8351	0.7849
3.	5.1146	5.0244	0.09012	5.0535	5.3356	0.2821
4.	5.2015	5.2138	0.0123	5.0398	5.4025	0.3627
5.	8.8588	8.8279	0.0309	8.6485	8.5699	0.0786
6.	5.4233	5.4471	0.0238	5.2171	5.6551	0.438
7.	5.3678	5.3867	0.0189	5.2649	5.6741	0.4092

TABLE 5.8
Comparison between PSO Results and SPICE Results: τ_{PHL} and τ_{PLH}

Sample	PSO Results			SPICE Results		
	$\tau_{PHL}(ns)$	$\tau_{PLH}(ns)$	Diff(ns)	$\tau_R(PHL)$	$\tau_F(LH)$	Diff(ns)
1.	2.5227	2.4648	0.0579	2.6455	2.4367	0.2088
2.	4.8217	4.8025	0.0192	4.9417	4.9256	0.0161
3.	2.6854	2.6731	0.0123	2.8366	2.4292	0.4074
4.	2.5812	2.5637	0.0175	2.5157	2.4210	0.0947
5.	4.6977	4.6785	0.0192	4.7861	4.5264	0.2597
6.	2.5805	2.5332	0.0473	2.5771	2.4957	0.0814
7.	2.7219	2.7581	0.0362	2.8629	2.6367	0.2262

[131]. It is observed that the PSO generated designs yield very good results even when simulated at the SPICE level as far the symmetry of the switching characteristics is considered.

5.7 The g_m/I_D Methodology for Low Power Design

By now it is clear to the readers that sizing of an analog circuit while meeting simultaneously a large number of objectives like a prescribed gain-bandwidth product, minimal power consumption, minimal area, low-voltage design, dynamic range, non-linear distortion, etc., is a very difficult task. This becomes more complicated when the transistors are designed with nano-scale technology. Optimization algorithms are attractive without any doubt, but they require translating not always well-defined concepts into mathematical expressions. The interactions amid semiconductor physics and analog circuits are not always easy to implement [93].

This section presents a methodology for sizing of CMOS analog circuits so

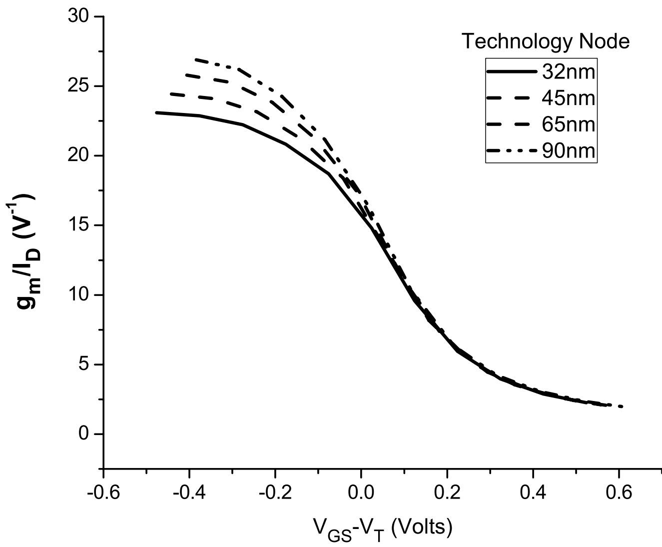
as to meet specifications such as gain-bandwidth while optimizing attributes like low power and small area. The sizing method takes advantage of the g_m/I_D ratio of a MOS transistor and makes use of a set of look-up tables. These tables are derived from physical measurements carried out on real transistors or advanced compact models such as BSIM4.

5.7.1 Study of the g_m/I_D and f_T Parameters for Analog Design

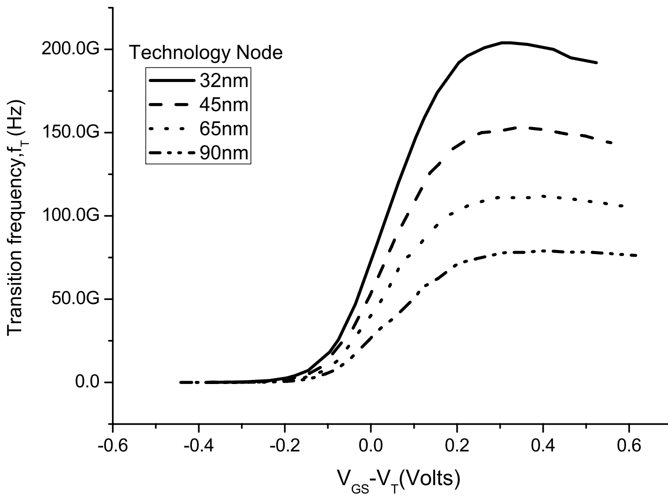
An important challenge in analog design is to achieve a good balance between the bandwidth and power efficiency of a circuit [133]. The two parameters which appear to be very much significant to the analog designers are (1) g_m/I_D and (2) $f_T = g_m/2(\pi C_{gg})$. The former signifies the amount of current to be used per transconductance and the later parameter signifies how much total gate capacitance C_{gg} must be driven at the controlling node per desired transconductance. The values of these quantities are found to be dependent on the region of operations of a MOS transistor. This is demonstrated in Fig. 5.11(a) and 5.11(b). It is observed that the g_m/I_D parameter has the maximum value in the weak inversion region and the value decreases as the operating point moves toward the strong inversion region. On the other hand, the f_T parameters has minimum value in the weak inversion region and the value increases as the operating point moves toward the strong inversion region. In addition, it is observed that the g_m/I_D parameter is not very sensitive to technology scaling. On the other hand, the values of f_T increase significantly with technology scaling.

It is interesting for the analog designers to study the variations of the product of g_m/I_D and f_T with the region of operation. This is shown in Fig. 5.12. This helps the designers to determine the overdrive voltage, i.e., $(V_{GS} - V_T)$ such that the bandwidth objective is met while operating at the corresponding maximum possible g_m/I_D (lowest power). It is observed that for a given technology node, the product quantity exhibits a “sweet spot” around a gate overdrive of 100 mV, which is a commonly found bias condition in many of today’s moderate-to-high speed designs [133]. On the other hand, working with high g_m/I_D greatly helps in reducing power consumption for applications that do not demand an extremely high bandwidth. For such applications, the MOS transistors can be biased in the weak inversion region. However, operating in the weak inversion region with high g_m/I_D comes at the cost of degraded linearity performance of the transistor. This is illustrated in Fig. 5.13 , which shows the linearity performance of various technologies versus g_m/I_D. The linearity is characterized through the parameter VIP3 which represents the extrapolated gate-voltage amplitudes, at which the third-order harmonics become equal to the fundamental tone in the drain current I_D. Mathematically, this is expressed as [202]

$$\mathrm{VIP3} = \sqrt{24\frac{g_m}{g_{m3}}} \tag{5.26}$$



(a) variation of g_m/I_D



(b) variation of f_T

FIGURE 5.11
Simulation results for the variations of g_m/I_D and f_T with the region of operation and technology nodes.

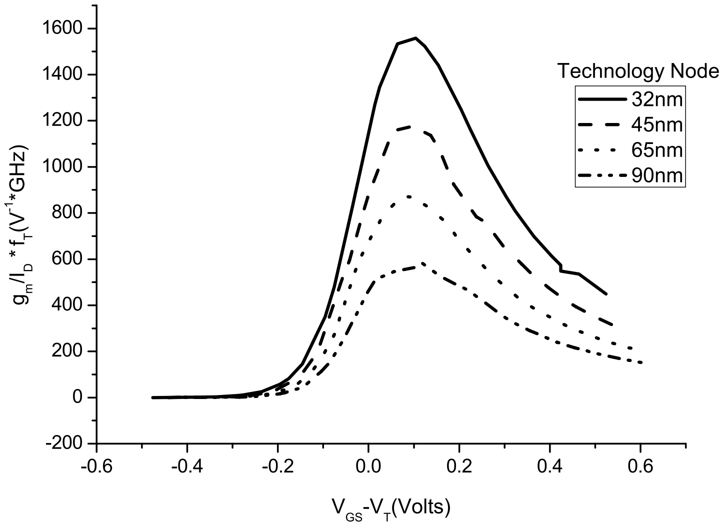


FIGURE 5.12
Simulation results showing the variations of the product of g_m/I_D and f_T with the region of operation and technology node.

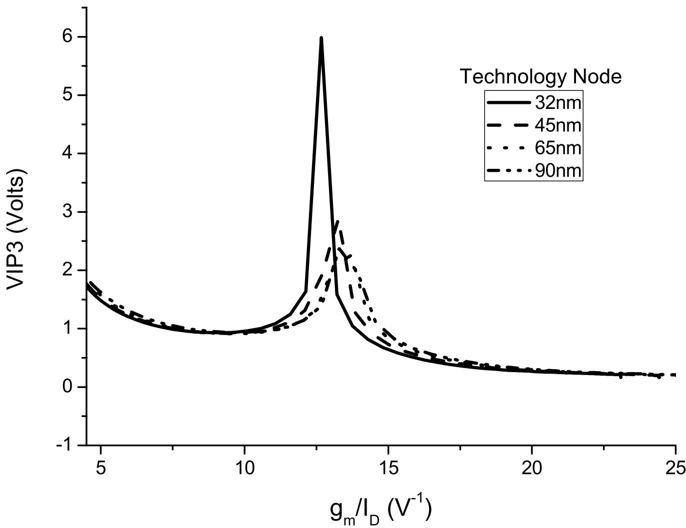


FIGURE 5.13
Simulation results showing the variations of the transconductance linearity with the region of operation and technology node.

where $g_{m3} = \dfrac{\partial^3 I_D}{\partial V_{GS}^3}$. The VIP3 peak, which is shown in Fig. 5.13, is because of the second-order-interaction effect and can be explained as a cancellation of the third-order nonlinearity coefficient by device internal feedback around a second-order nonlinearity. In practice, it is very hard to utilize this extremely linear point. It is observed that the linearity is degraded as the g_m/I_D ratio is increased.

It is also important to study the variations of the intrinsic capacitances of MOS transistors as functions of the g_m/I_D parameter. These are shown in Fig. 5.14(a) and 5.14(b). It is observed that the values of the both the capacitors are low, when g_m/I_D is high.

5.7.2 g_m/I_D Based Sizing Methodology

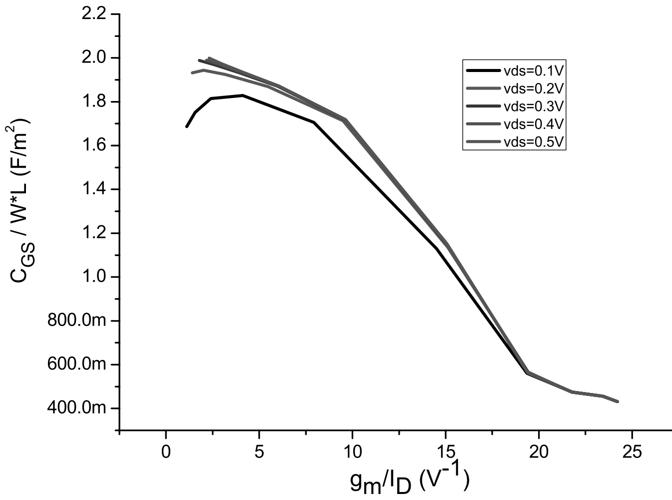
The g_m/I_D based circuit sizing procedure is based on the relation between the ratio of the transconductance over DC current g_m/I_D and the normalized current $I_N = I_D/(W/L)$[173, 34]. The selection of the g_m/I_D as the key parameter is due to three reasons. First, this parameter is strongly related to the analog performances. Second, it gives an indication of the operating region of a MOS transistor. Third, it provides a tool for calculating the transistor dimensions. This parameter is considered to be a universal characteristic of the transistors in the same process technology. The relation between the g_m/I_D parameter with the operating region of the transistor may be written as follows

$$\frac{g_m}{I_D} = \frac{1}{I_D}\frac{\partial I_D}{\partial V_{GS}} = \frac{\partial\left(\ln I_D\right)}{\partial V_{GS}} = \frac{\partial\left\{\ln\left[\dfrac{I_{DS}}{\left(\dfrac{W}{L}\right)}\right]\right\}}{\partial V_{GS}} \tag{5.27}$$

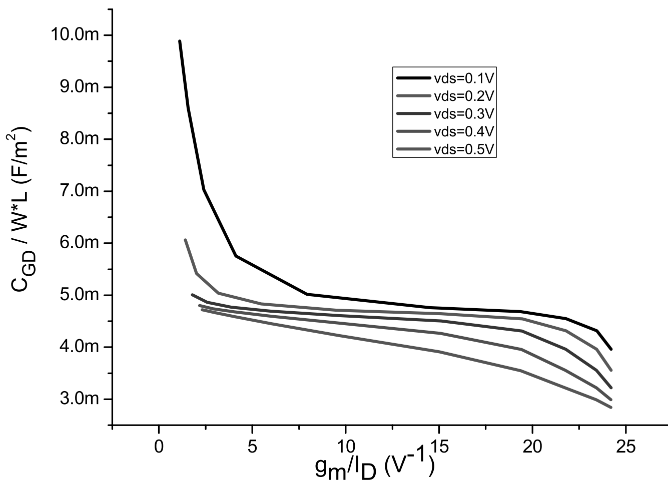
As it has been demonstrated in Fig. 5.11(a) the maximum value of the g_m/I_D ratio lies in the weak inversion region and the value decreases as the operating point moves toward strong inversion when I_D or V_{GS} are increased. It may be noted that the g_m/I_D ratio is independent of the transistor sizes. The normalized current I_N is also independent of the transistor sizes. Therefore, the relationship between the g_m/I_D and the normalized current is a unique characteristic for all transistors of the same type (n-channel MOS or p-channel MOS) in a given batch. This relationship is shown in Fig. 5.15(a) and Fig. 5.15(b). This universal characteristic of the g_m/I_D versus I_N curve is used to determine the aspect ratio of a transistor, which is then subsequently used to determine the channel width, assuming a fixed value of the channel length. The corresponding simulation graph is shown in Fig. 5.16.

For a MOS transistor, the magnitude of the intrinsic voltage gain is given by

$$A_v = g_m r_0 = \left(\frac{g_m}{I_D}\right)\left(I_D r_0\right) = \left(\frac{g_m}{I_D}\right)V_A \tag{5.28}$$

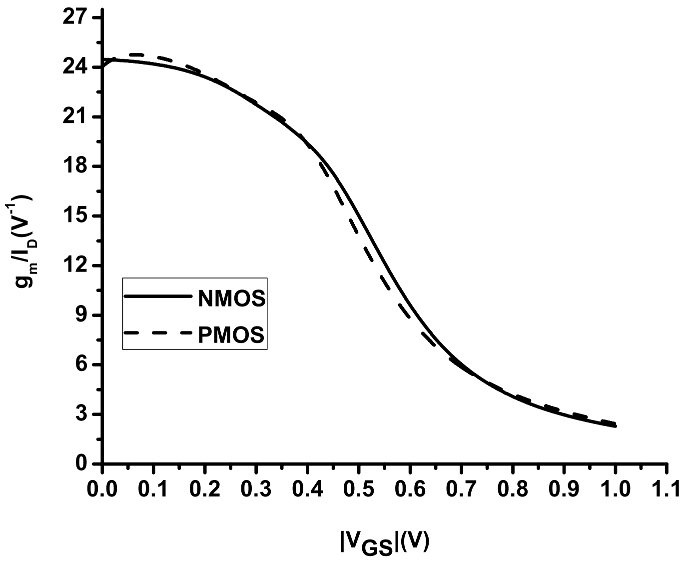


(a) variation of C_{GS}

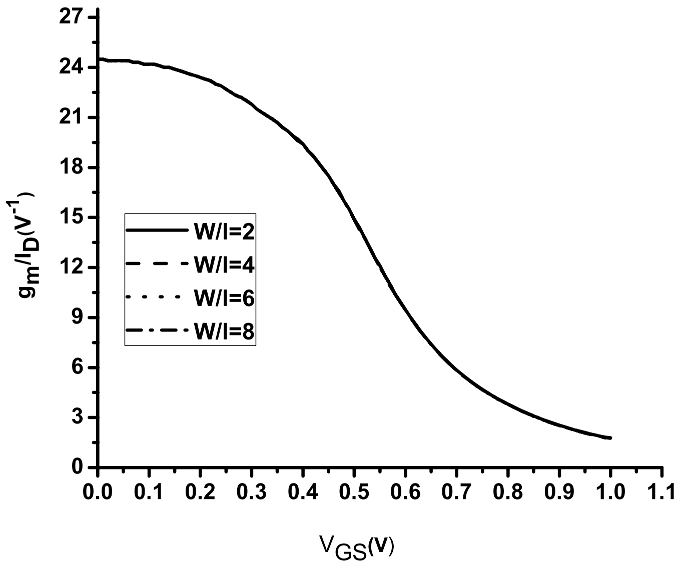


(b) variation of C_{GD}

FIGURE 5.14
Simulation results for the variations of C_{GS} and C_{GD} with the g_m/I_D.



(a) g_m/I_D for n-channel and p-channel MOS transistor



(b) g_m/I_D graph for different aspect ratios

FIGURE 5.15
Simulation results for the g_m/I_D variations.

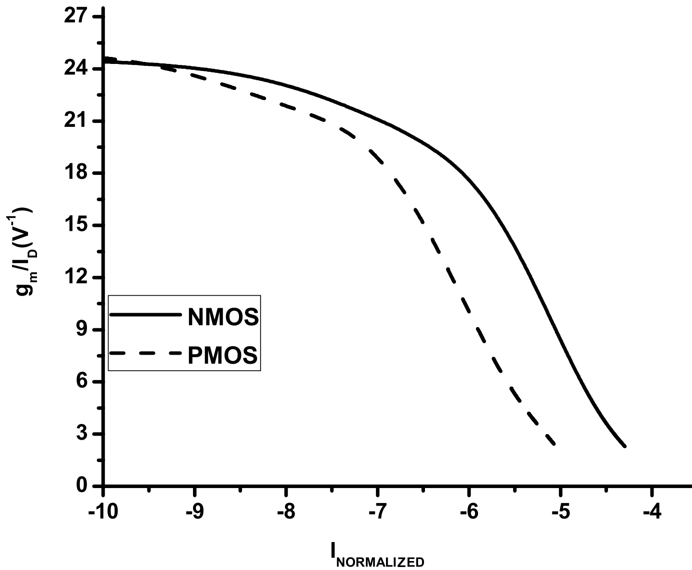


FIGURE 5.16
Simulation results showing the variations of the g_m/I_D with the normalized current I_N.

where V_A is referred to as the early voltage of the transistor. Assuming V_A to be constant for a particular channel length of a transistor, the intrinsic gain is determined by the g_m/I_D ratio. Therefore, the intrinsic gain of a MOS transistor is maximum in the weak inversion region and reduces as the operating point moves toward the strong inversion region. This is shown in Fig. 5.17. Therefore, an important guideline to get high gain for a MOS transistor, is to bias the transistor in the weak inversion region with as low V_{GS} as possible. An interesting thing observed from the curve is that under weak inversion regions very small amounts of drain current flows, which implies a small amount of power dissipation. Therefore, by biasing the MOS transistor in the weak inversion region, it is possible to obtain high gain with very small power dissipation.

The procedure for determining the aspect ratio through the g_m/I_D methodology is explained by a simple example. Let the current flow through the transistor be $I_D = 100nA$. The transistor is biased in the weak inversion region with $g_m/I_D = 21.8V^{-1}$ at $V_{GS} = 0.3V$. From the normalized current plot, it is observed that the corresponding $I_N = 41.64nA$. Therefore, the aspect ratio is found to be $W/L = 2.4$. Therefore, by assuming $L = 100nm$, the channel width is found to be $W = 0.24\mu m$. The power dissipation corresponding to a supply voltage of 0.5V will be 50nW.

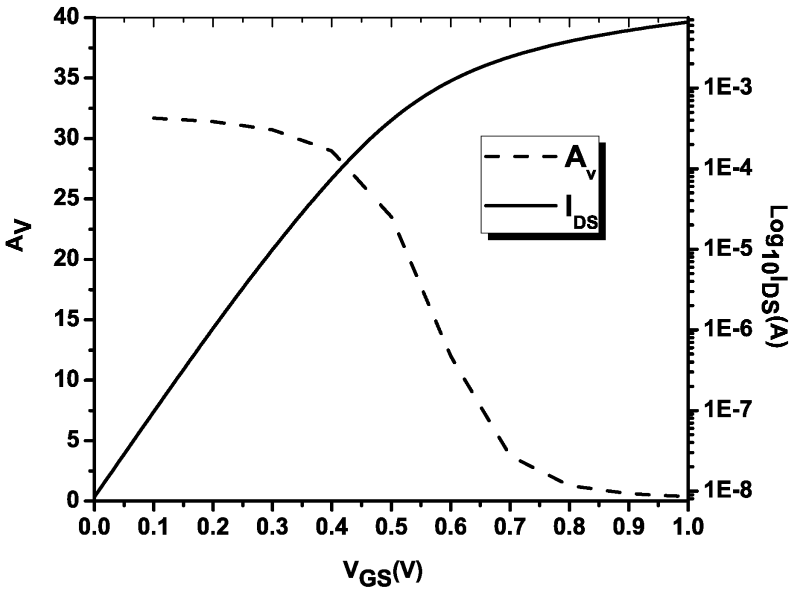


FIGURE 5.17
Simulation results showing the variations of the intrinsic gain and drain current with the region of operation.

The primary advantage of the g_m/I_D methodology over other methodologies is that this methodology uses a set of look-up tables as the main design tool, which is itself constructed from SPICE simulations. Therefore, the predicted results, as obtained after the circuit sizing process, appear to be quite close to the actual SPICE results. The same can be achieved by using ANN/LS-SVM based models. However, construction of look-up tables is much easier and less time consuming. Therefore, the g_m/I_D methodology is gaining importance day by day for nano-scale analog circuit design. In addition, the g_m/I_D methodology gives the designers the flexibility to operate the transistors in any region of operation. It may be noted that by using compact models, it is possible to compute the g_m/I_D graph analytically.

5.7.3 Case Study 4: Sizing of Low-Power Nano-Scale Miller OTA Using the g_m/I_D Methodology

The sizing methodology of a Miller OTA circuit, shown in Fig.5.6 , is based on what is presented in [173, 34]and is outlined below.

1. From the power consumption requirement, the total current I_T flowing through the circuit is calculated.
2. The compensation capacitor C_c is calculated from the 60° phase margin requirement, $C_c > 0.22C_L$.
3. The bias current is determined from the slew rate requirement. $I_b = I_5 = SR.C_c$
4. The second-stage branch current is calculated as $I_8 = I_7 = I_T - 2I_b$
5. From the gain-bandwidth requirement, the transconductance of M1 transistor is calculated as $GBW = \dfrac{g_{m1}}{2\pi C_c}$
6. Fix $\left(\dfrac{g_m}{I_D}\right)_1$ to operate the M1 transistor in weak inversion.
7. $\left(\dfrac{g_m}{I_D}\right)_1 = \left(\dfrac{g_m}{I_D}\right)_2$
8. The transistors M3 and M4 are operated in the weak inversion region since the current flowing through these transistors is small. Select the $\left(\dfrac{g_m}{I_D}\right)_3$ and $\left(\dfrac{g_m}{I_D}\right)_4$ sufficiently high.
9. The transistors M5, M6 and M8 are similarly made to operate in the weak inversion region and the $\left(\dfrac{g_m}{I_D}\right)$ ratios are determined accordingly.
10. From the relation $g_{m7} >> 10.g_{m1}$, find out g_{m7} and hence $\left(\dfrac{g_m}{I_D}\right)_7$

TABLE 5.9
Aspect Ratios and g_m/I_D Ratio of Each Transistor of Case Study 4

Transistor	W/L	g_m/I_D
M1	230	24.7
M2	230	24.7
M3	4	23.43
M4	4	23.43
M5	459	24.7
M6	459	24.7
M7	14	22.73
M8	2526	24.7

TABLE 5.10
Comparison between Analytical and Simulation Results for Case Study 4

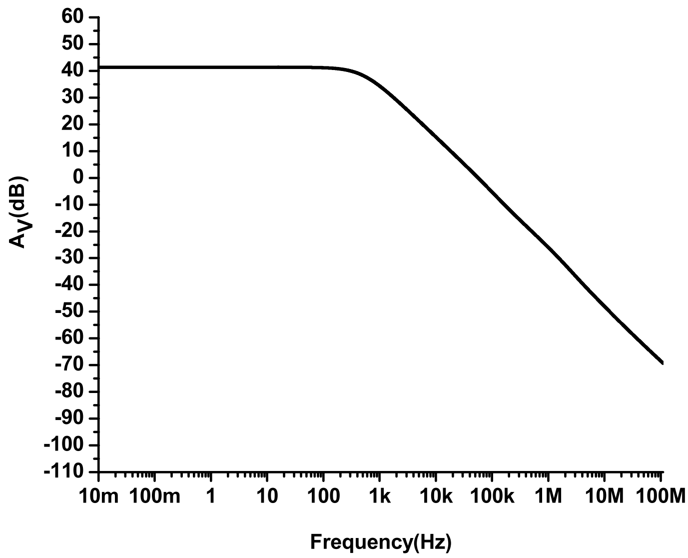
Performances	Analytical	Simulation
Gain (db)	60	41.4
GBW (KHz)	50	56
SR (V/μs)	0.025	0.02
CMRR(db)		71.9
ICMR(V)		(0.065 to 0.9)
Total current (nA)	300	299

11. Once the $\left(\frac{g_m}{I_D}\right)$ of all transistors and the corresponding drain currents are known, the normalized current are determined from the g_m/I_D vs I_N graph and hence the $\left(\frac{W}{L}\right)$ ratios for each transistor are determined from the corresponding normalized currents and drain currents.

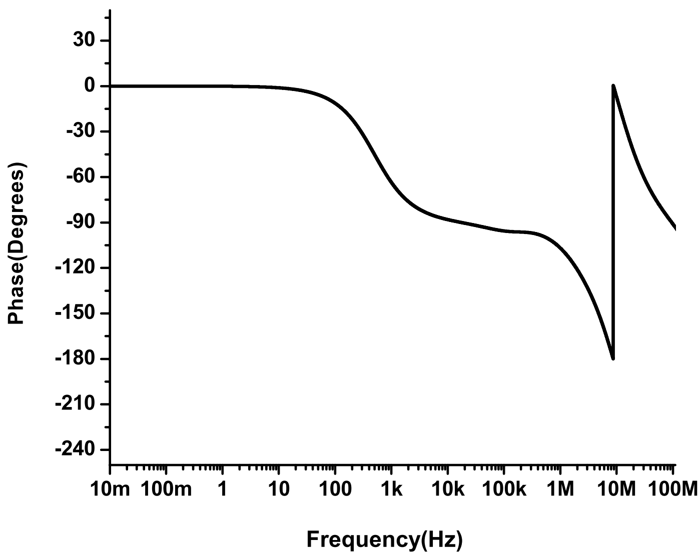
It may be noted that an important issue related to the operation of MOS transistors in a weak inversion region is that under this condition, the drain current mismatch due to threshold voltage mismatch is maximum. Therefore, often the transistors involved in the current mirror circuit are not operated in the weak inversion region. Therefore, the g_m/I_D ratios of such transistors are to be selected accordingly.

With the present methodology it is attempted to design a two-stage Miller OTA with gain $A_v > 40dB$, gain bandwidth product $GBW \geq 40KHz$, phase margin $PM > 60^0$, slew rate $SR = 25V/ms$ and power dissipation ¡350nW.

The transistor length is considered to be $0.1\mu m$. The length can be increased, if higher gain is required. The aspect ratios as well as the (g_m/I_D) ratio of each transistor are tabulated in Table. 5.9. The circuit is simulated using 45nm,1V CMOS technology using HSPICE simulation tool. Table 5.10

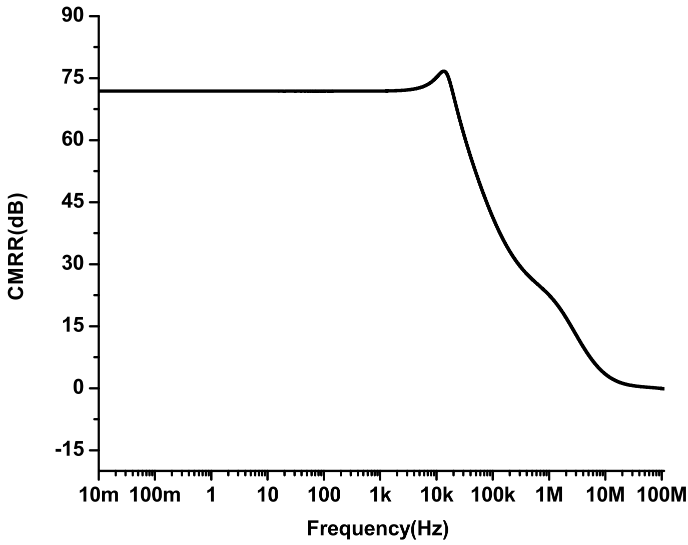


(a) Gain vs. frequency plot of the synthesized OTA

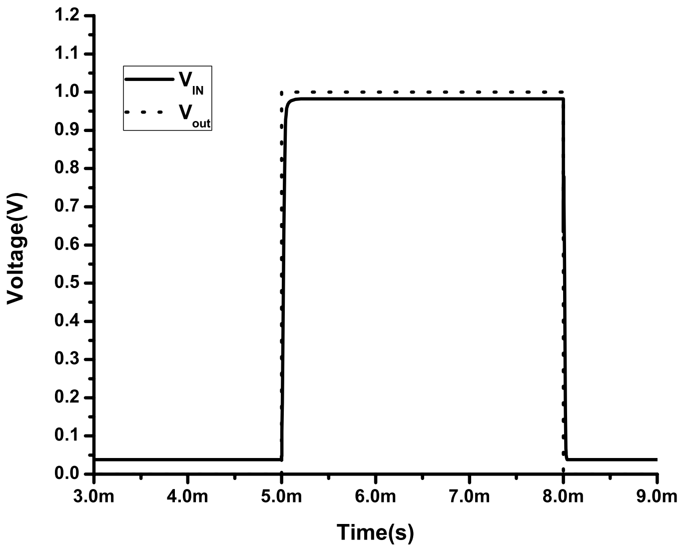


(b) Phase vs. frequency plot of the synthesized OTA

FIGURE 5.18
AC simulation results of the synthesized OTA in Case Study 4.



(a) CMRR



(b) Slew rate

FIGURE 5.19
CMRR and slew rate of the synthesized OTA in Case Study 4.

shows the comparison between the performances as calculated analytically and as obtained from SPICE simulation results. The AC simulation results illustrating the variations of the gain and phase with frequency for the synthesized OTA are shown in Fig. 5.18(a) and Fig. 5.18(b). The CMRR and the slew rate is obtained from Fig. 5.19(a) and Fig. 5.19(b) respectively. The difference between the analytical and simulation results occurs primarily because of the fact that the variations of the early voltage with drain bias are not taken into consideration in this method. This is an important issue in nano-scale design and needs to be taken care of judiciously.

5.8 High-Level Specification Translation

At the architectural level of abstraction of a hierarchical analog design methodology, the overall architecture of the system is first decomposed into several component blocks. The specifications of these component blocks are then derived from the specifications of the complete system so that they can be designed separately. This is referred to as the process of high-level specification translation [139]. For example, if the system to be considered is a $\Sigma - \Delta$ ADC, then at the architecture level, the component blocks are the integrator, comparator, etc. The specifications of these component blocks need to be derived such that the system specifications are optimally satisfied. On the other hand, at the same time, it has to be ensured that the specifications of the component blocks can actually be realized in practice, finally when the various component blocks are to be implemented using transistor-level circuits. Therefore, the task of high-level specification translation is a challenge to the analog designers.

The task of construction of feasible design space as an intersection of an application bounded space (constructed by top-down procedure through interval analysis technique) and circuit realizable space (constructed by bottom procedure through actual circuit simulation) and subsequent identification through LS-SVM classifier method has been discussed in [139]. This has been described in detail in Chapter 4 of this text. The identified feasible design space needs to be explored through any design space exploration algorithm such as the particle swarm optimization algorithm or genetic algorithm etc., in order to find out the various values of the design parameters. Recently a geometric programming-based methodology has been used for high-level specification translation [38].

5.9 Summary and Conclusion

This chapter discusses in detail the task of automated sizing of analog circuits. The particle swarm optimization algorithm has been described as a popular design space exploration algorithm. This is also demonstrated with the case study of synthesizing an OTA circuit. The cost functions which are to be computed by the design space exploration algorithm during the sizing procedure are often made simple, based on the square law model of MOS transistors. As a result, often the synthesized results are found to deviate greatly when the designs are actually simulated through SPICE. This is sometimes judiciously avoided by considering a large guard band for the specifications. The alternative is to embed accurate models such as ANN/SVM-based learning models. These are also discussed in the present text and demonstrated through case studies. Finally, this chapter presents a look-up table based approach, based on the g_m/I_D methodology. This approach is found to be simple and suited for nano-scale analog circuit sizing, however with a scope of improvement.

6

Advanced Effects of Scaled MOS Transistors

6.1 Introduction

It has been emphasized in the earlier chapters, that the physics of scaled MOS transistors plays a significant role in determining the performances of CMOS analog circuits and systems. Therefore, comprehensive knowledge of device physics is essential in understanding the behavior and characteristics of nano-scale analog circuits. Without such understanding, the development of CAD tools will simply be a futile exercise. Chapter 3 of the present text discusses some fundamental issues related to the physics of scaled MOS transistors that have profound effects on circuit performances. This chapter presents some advanced effects of scaled MOS transistors which are gaining importance day by day as the circuits are designed with sub-90nm CMOS technology.

6.2 Narrow Width Effect on Threshold Voltage

MOS transistors are considered to be narrow when the channel width of the transistor is of the same order of magnitude as the thickness of the depletion region under the gate oxide [3]. For CMOS ICs with embedded SRAM (static random access memory), there is a logic portion with relatively wide transistors and a memory portion that has much narrower MOS transistors. The study of the effect of channel width reduction thus becomes essential for memory cell size reduction. The narrow MOS transistors are associated with interesting characteristics that are the result of several effects occuring at the edges of the gate. The primary reason behind these effects is isolation that exists at the two sides of the channel width. The edge effects strongly increase with reduction of the feature size. Two major isolation technologies have developed, the semi-recessed or the local oxidation of silicon (LOCOS) technology and the fully-recessed or the shallow trench isolation (STI) technology [194]. The following subsections discuss the two technologies and their associated effects.

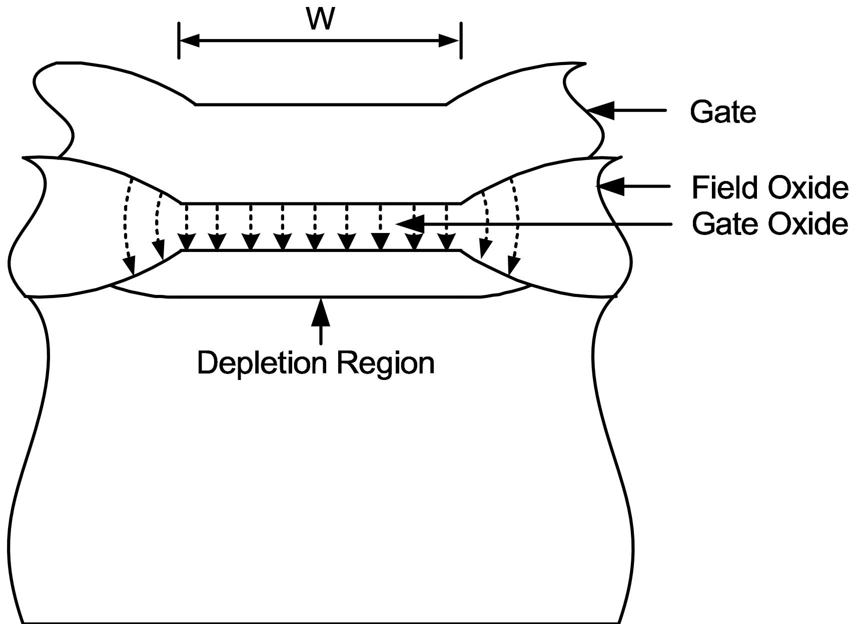


FIGURE 6.1
Cross-section along the width of a MOS transistor with LOCOS isolation.

6.2.1 LOCOS Isolated MOS Transistors

Figure 6.1 shows the cross-section along the width of a MOS transistor with LOCOS isolation. As seen in the figure, there is a gradual transition from the thick (field) oxide to the thin (gate) oxide. This region is the “bird’s beak” region. The channel width is equal to the extent of the thin oxide region. It is to be noted that the depletion region is not limited to just the area below the thin gate oxide. This is because some of the field lines emanating from the gate charges terminate on the ionized acceptor atoms on the sides [194]. These lines are known as the fringing field lines and are also shown as dotted lines in Fig. 6.1. For wide devices (W large), the portion of the depletion region on the two sides is a small percentage of the total depletion volume and thus may be neglected. However, for narrow devices, this constitutes a considerable portion of the entire depletion volume and becomes non-negligible. The gate is thus responsible for depleting a larger region as compared to one’s usual assumption. Thus, it takes a higher V_{GS} value to deplete that amount before an inversion layer is formed. Effectively, the depletion charge increases and thus the threshold voltage of the transistor increases. This increase in the threshold voltage of the device with a reduction in the width of the transistor is known as the narrow width effect (NWE) [113].

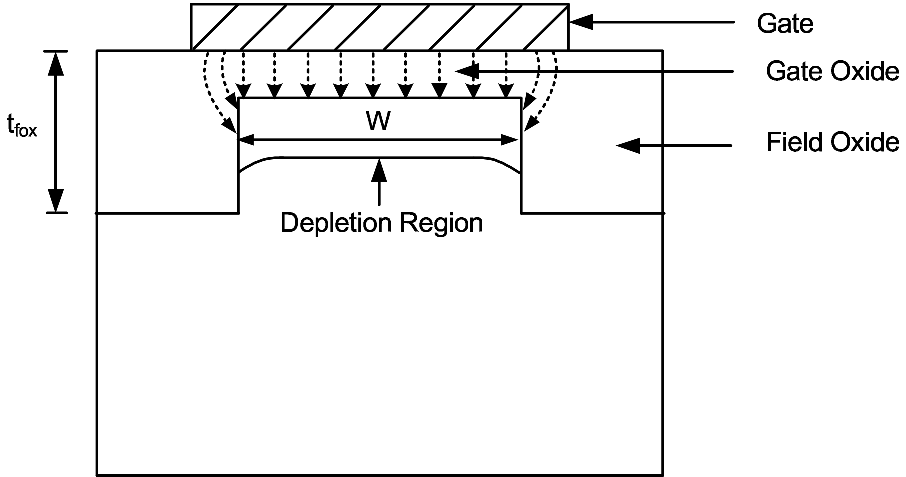


FIGURE 6.2
Cross-section along the width of a MOS transistor with STI.

6.2.2 Shallow Trench Isolated (STI) MOSFETs

The continued increase in circuit density requires devices not only with short channel lengths and narrow widths, but with a small active area pitch. This requires non-encroaching isolation oxides. The trench or fully-recessed isolation oxides meet this requirement. Hence, STI technology is extensively used in the sub-250 nm CMOS generations. Figure 6.2 shows the cross-section along the width of a MOS transistor with STI technology. The threshold voltage of an STI MOSFET decreases with a reduction of the width of the device. This is in contrast to the LOCOS isolated MOS transistors where the threshold voltage increases as the device width is reduced. The phenomenon of threshold voltage reduction with a reduction in the width of an STI MOS transistor is conventionally known as the inverse narrow width effect (INWE) [82].

As seen in Fig. 6.2 the fringing field lines from the gate terminate on the charges at the sidewalls of the channel. When the width of the channel is large, the fringing effect may be neglected. However, when the width of the channel is small, fringing contributes significantly to the total gate flux and hence cannot be neglected. Thus, the fringing capacitances are significant. The total gate capacitance is then considered to be an ideal thin oxide capacitor in parallel with two sidewall capacitors and may be written as [3]

$$C_G = C_{ox}WL + 2C_{STI} \tag{6.1}$$

where C_{ox} is the ideal thin oxide capacitance per unit area and C_{STI} is the sidewall fringe capacitance. This is found to be [1, 3]

$$C_{STI} = \left(\frac{2\epsilon_{ox}L}{\pi}\right)\ln\left(\frac{2t_{fox}}{t_{ox}}\right) \tag{6.2}$$

where t_{fox} is the depth of the recessed oxide. Thus

$$C_G = \left(1 + \frac{F^*}{W}\right) WLC_{ox} \tag{6.3}$$

where

$$F^* = \left(\frac{4t_{ox}}{\pi}\right) \ln\left(\frac{2t_{fox}}{t_{ox}}\right) \tag{6.4}$$

is the fringing field factor.

The fringing effect is exhibited by an increase in the surface potential and an increase in the depletion depth near the trench oxide sidewalls. Further, the electrostatic potential at the trench oxide sidewalls varies quadratically along the depth within the depletion layer of the transistor. The enhanced electrostatic potential and the depletion depth at the sidewalls imply that the depletion charge is effectively reduced in the narrow devices so that the threshold voltage gets lowered in the narrow transistors. The threshold voltage of the narrow channel transistors is thus written as

$$V_T = V_{FB} + 2\Phi_F + \frac{Q_b^{'}}{WLC_{ox}} \tag{6.5}$$

where $Q_b^{'}$ is the reduced depletion charge density. Again, if the depletion charge is assumed to be the same as in a wide device the threshold voltage of the narrow device may be written as

$$V_T = V_{FB} + 2\Phi_F + \frac{Q_b}{C_G} \tag{6.6}$$

The BSIM compact model employs an empirical approach to model the narrow width effect as given below

$$\Delta V_T(NWE) = (K3 + K3BV_{BS}) \left(\frac{t_{ox}}{W + W0}\right) 2\Phi_F \tag{6.7}$$

where $K3, K3B$ and $W0$ are the empirical fitting parameters and W is the effective channel width W_{eff}. W_0 is an offset parameter. $K3$ is positive for NWE and negative for INWE.

Figure 6.3 shows the typical width dependence of threshold voltage for LOCOS and STI MOS transistors. It has been recently reported in [142] that the INWE is caused due to the combined effect of both the gate fringing field effect and the phenomenon of dopant redistribution.

It is to be noted that change in the threshold voltage with the reduction in the channel width is also affected in the opposite way by STI-induced mechanical stress effects in the width direction [138]. With the scaling of the minimum feature sizes, inverse narrow width effect and STI-induced stress are becoming increasingly important. However, for device widths below $1\mu m$ the INWE dominates and the effect of STI mechanical stress in the channel width direction becomes significant at larger widths [138].

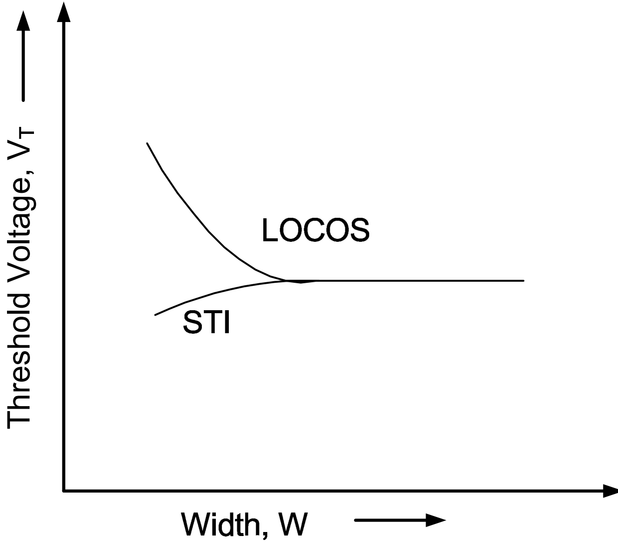


FIGURE 6.3
Width dependence of threshold voltage for LOCOS and STI MOS transistors.

6.3 Channel Engineering of MOS Transistor

In order to combat the degradation of circuit performances of scaled MOS transistors, various techniques are employed to suppress the short channel effects. Among the various approaches, vertical channel engineering with retrograde doping, lateral channel engineering with pocket implantations like single-halo (SH), also known as lateral asymmetric channel (LAC), or double-halo (DH) etc., have been reported. These are collectively referred to as the channel engineering of MOS transistors [191] and is discussed in this section.

6.3.1 Non-Uniform Vertical Doping

Localized manipulation of doping is a useful technique often used by the device designers for engineered characteristics of the MOS transistor. The threshold voltage and the depletion depth are coupled with each other through the parameter N_A, as observed from the following expressions:

$$V_T = V_{FB} + 2\Phi_F + \frac{\sqrt{4\epsilon_{ox} q N_A \Phi_F}}{C_{ox}} \tag{6.8}$$

$$W_{dm} = \sqrt{\frac{4\epsilon_{Si}\Phi_F}{qN_A}} \tag{6.9}$$

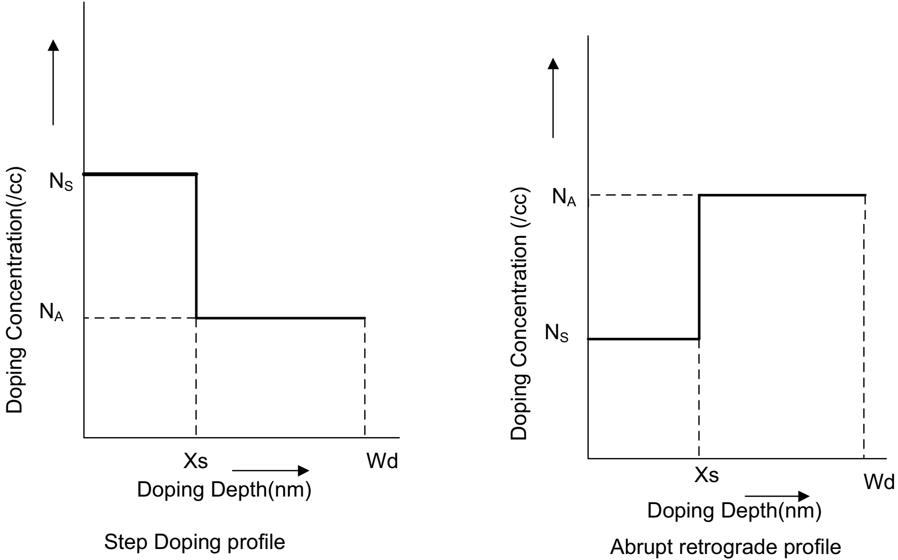


FIGURE 6.4
Step doping and abrupt retrograde channel doping profile.

If it is required to scale down the threshold voltage, it is required to reduce N_A, which in turn increases the depletion depth, which again is not desirable for better short channel effect immunity. Non-uniform channel doping provides the device designer an additional degree of freedom to tailor the profile for improved device performances [192]. For example, lighter doping at a greater depth from the interface reduces the drain-substrate capacitance and also the substrate-bias effect [189]. A lighter doping near the interface lowers the threshold voltage, reduces the field, and improves mobility. On the other hand, higher doping at a deeper region reduces punch-through between source and drain [189].

The inhomogeneity in the substrate, due to the channel doping profile is approximated to belong to the following two broad categories from a modeling viewpoint: (i) Step profile: abrupt high-to-low profile and (ii) Abrupt retrograde: low-to-high profile. These are illustrated in Fig. 6.4. It may be noted that in reality, these are observed to follow a Gaussian distribution.

6.3.1.1 High-to-Low Profile

The general equation for the long-channel threshold voltage is given by

$$V_T = V_{FB} + \psi_s - \frac{Q_b}{C_{ox}} \tag{6.10}$$

$$= V_{FB} + 2\Phi_F - \frac{q}{C_{ox}} \int_0^{W_{dm}} N(x)dx \tag{6.11}$$

The long-channel maximum depletion depth is determined from the Poisson equation at the onset of strong inversion and is given by [192, 189]

$$\psi_s = 2\Phi_F = \frac{q}{\epsilon_{Si}} \int_0^{W_{dm}} xN(x)dx \tag{6.12}$$

This integration is easy to carry out for the high-to-low profile and thus the surface potential is given as

$$\psi_s = \frac{qN_S}{2\epsilon_{Si}} x_s^2 + \frac{qN_A}{2\epsilon_{Si}} \left(W_{dm}^2 - x_s^2 \right) \tag{6.13}$$

From this the depletion depth is calculated to be

$$W_{dm} = \sqrt{\frac{2\epsilon_{Si}}{qN_A} \left(\psi_s - \frac{q\left(N_S - N_A\right) x_s^2}{2\epsilon_{Si}} \right)} \tag{6.14}$$

The threshold voltage is then given by

$$\begin{aligned} V_T &= V_{FB} + 2\Phi_F + \frac{1}{C_{ox}} \sqrt{2q\epsilon_{Si} N_A \left(2\Phi_F - \frac{q\left(N_S - N_A\right) x_s^2}{2\epsilon_{Si}} \right)} \\ &\quad + \frac{q\left(N_S - N_A\right) x_s}{C_{ox}} \\ &= V_{FB} + 2\Phi_F + \frac{1}{C_{ox}} \sqrt{2q\epsilon_{Si} N_A \left(2\Phi_F - \frac{q\Delta N x_s^2}{2\epsilon_{Si}} \right)} + \frac{q\Delta N x_s}{C_{ox}} \end{aligned} \tag{6.15}$$

where $\Delta N = N_S - N_A$.

The effect of the high-to-low doping profile is therefore reduction of the depletion layer width and increase of the depletion charge within $0 \leq x \leq x_s$ by $(N_S - N_A)x_s$.

The non-uniform profile shown in Fig. 6.4 is basically an approximation of the Gaussian profile, which is generally the real profile. This is because of the ion implantation and the subsequent thermal annealing procedure. The implant dose is defined as

$$D_I = \left(N_S - N_A\right) x_s \tag{6.16}$$

The Gaussian profile may be written as

$$N(x) = \frac{D_I}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(x - x_c\right)^2}{2\sigma^2} \right) \tag{6.17}$$

which is centered at $x_c = x_s/2$ and σ is the implant straggle. The threshold voltage is therefore, written as

$$V_T = V_{FB} + 2\Phi_F + \frac{1}{C_{ox}} \sqrt{2\epsilon_{Si} q N_A \left(2\Phi_F - \frac{qD_I x_c}{\epsilon_{Si}} \right)} + \frac{qD_I}{C_{ox}} \tag{6.18}$$

The maximum depletion depth is given by

$$W_{dm} = \sqrt{\frac{2\epsilon_{Si}}{qN_A}\left(2\Phi_F - \frac{qD_I x_c}{\epsilon_{Si}}\right)} \tag{6.19}$$

For the limiting case of a delta function of dose localized at the interface ($x_c = 0$), the depletion width does not change and the threshold voltage shift is given by

$$\Delta V_T \approx \frac{qD_I}{C_{ox}} \tag{6.20}$$

It may be noted that all expressions derived for the depletion capacitance, subthreshold slope and body-effect coefficient for the uniform doped case remains valid, by replacing the depletion width of the uniform case with that derived for the non-uniform case. For the high-to-low profile, the depletion depth decreases so that the depletion capacitance increases. This results in larger (less steep) subthreshold swing.

6.3.1.2 Low-to-High Retrograde Profile

To reduce the threshold voltage without significantly increasing the gate depletion width, a retrograde channel profile is used. The derivations of the maximum depletion width and threshold voltage remain identical, with appropriate change of signs, and are given as follows [192, 189]

$$W_{dm} = \sqrt{\frac{2\epsilon_{Si}}{qN_A}\left(2\Phi_F + \frac{q\Delta N x_s^2}{2\epsilon_{SI}}\right)} \tag{6.21}$$

$$V_T = V_{FB} + 2\Phi_F + \frac{1}{C_{ox}}\sqrt{2\epsilon_{Si} qN_A\left(2\Phi_F + \frac{q\Delta N x_s^2}{2\epsilon_{Si}}\right)} - \frac{q\Delta N x_s}{C_{ox}} \tag{6.22}$$

The threshold voltage is thus decreased and the depletion depth is increased.

6.3.1.3 Compact Modeling of Vertical Non-Uniform Doping Effect

This non-uniformity makes the body-effect parameter γ in (3.5) a function of both the depth from the interface and the substrate bias. If the depletion depth is less than X_s, as shown in Fig. 6.4, N_A in (3.6) is equal to N_{DEP}, otherwise it is equal to N_{SUB}. Then the threshold voltage for non-uniform vertical doping is proposed to be [30]

$$V_T = V_{T0} + K1\left(\sqrt{\psi_s - V_{BS}} - \sqrt{\psi_s}\right) - K2V_{BS} \tag{6.23}$$

$K1$ and $K2$ are usually determined by fitting (6.23) to the measured threshold voltage data.

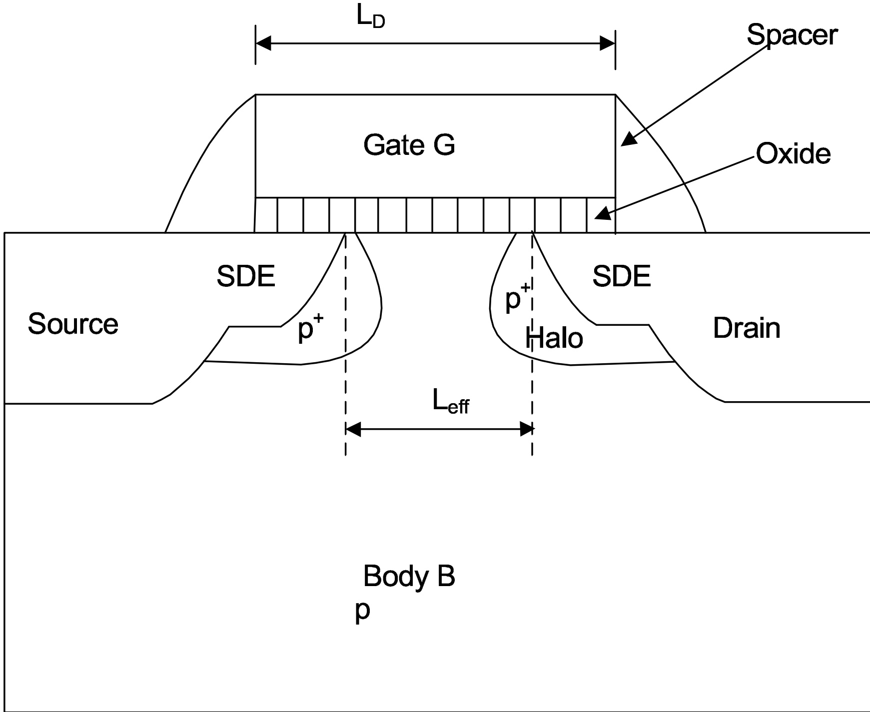


FIGURE 6.5
Cross-sectional view of a MOS transistor with double halo channel engineering.

6.3.2 Pocket (Halo) Implantation

In this case, the doping concentration near the source/drain junctions is higher than that in the middle of the channel. With this type of channel engineering, the doping concentration in the channel along the channel length becomes non-uniform. The halo/pocket implantation can be made either near the source or near both the source and drain regions [191]. Accordingly, the transistors are referred to as single halo (SH) or double halo (DH) transistors. The schematic diagram of a VLSI MOS transistor using double halo pocket implantation is shown in Fig. 6.5. The lateral non-uniform doping with higher doping concentration near the source/drain extension regions results in an increase in the average doping concentration in the channel. This in turn leads to an increase in the threshold voltage.

The non-uniform doping concentration along the channel length may be approximated by a step doping profile as shown in Fig. 6.6. The average

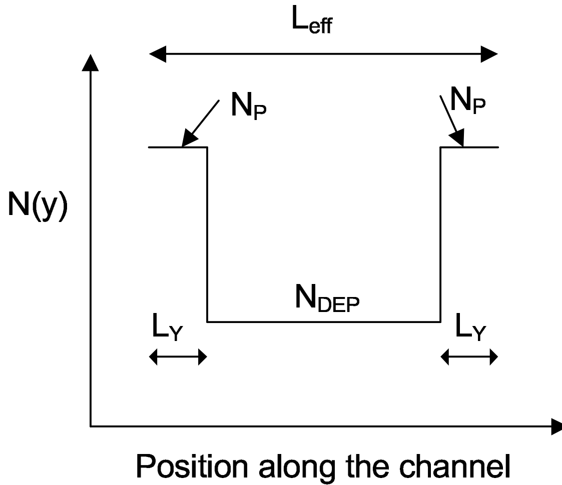


FIGURE 6.6
Lateral variation of channel doping.

channel doping is written as [30]

$$\begin{aligned} N_{eff} &= \frac{N_{DEP}\,(L-2L_Y)+N_p 2L_Y}{L} = N_{DEP}\left(1+2\frac{L_Y}{L}\frac{N_p-N_{DEP}}{N_{DEP}}\right) \\ &\approx N_{DEP}\left(1+\frac{L_{PE0}}{L}\right) \end{aligned} \tag{6.24}$$

In (6.24), N_p is the pocket concentration and L_{PE0} is a fitting parameter whose value is to be extracted from measurement results [55] and L is the effective channel length.

With the introduction of lateral and vertical channel engineering, the threshold voltage for a long channel MOS transistor is finally expressed as [55]

$$\begin{aligned} V_T &= V_{T0} + K1\left(\sqrt{2\Phi_F - V_{BS}} - \sqrt{2\Phi_F}\right)\sqrt{1+\frac{L_{PEB}}{L}} \\ &- K2V_{BS} + K1\left(\sqrt{1+\frac{L_{PE0}}{L}}-1\right)\sqrt{2\Phi_F} \end{aligned} \tag{6.25}$$

L_{PEB} is a fitting parameter which signifies the lateral non-uniform doping effect on $K1$. The value of this parameter also needs to be extracted from measurement results.

For n-channel MOS transistors, halo regions near the two ends of the channel are beneficial for the suppression of short-channel effects by compensating the charge-sharing effects from the source-drain fields. This is the significance of halo implantation.

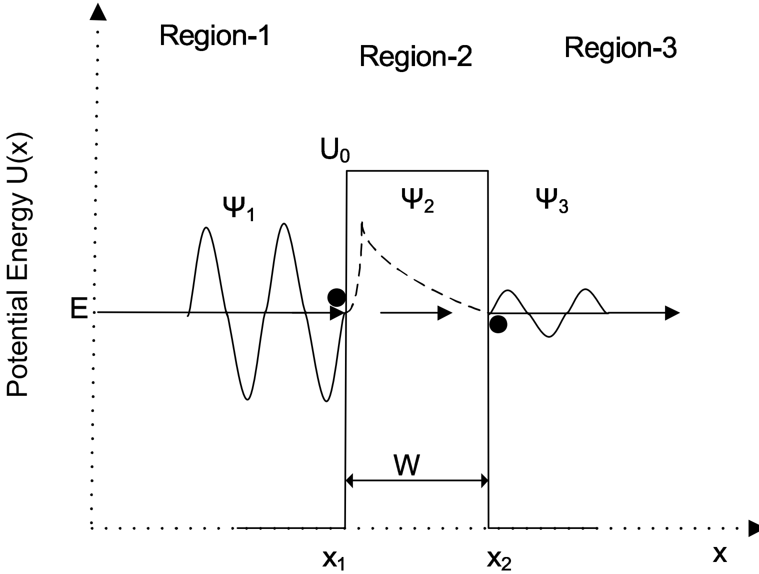


FIGURE 6.7
Quantum-mechanical tunneling.

6.4 Gate Leakage Current

It has been discussed in Chapter 1, that the thickness of the gate dielectric scales down with scaling of the technology. The most serious limitation of the gate oxide scaling is that the gate terminal which is conventionally considered to be non-conducting shows finite conductivity and current flows through the gate terminal under a biased condition [1]. This has been briefly mentioned in Chapter 1. The gate leakage current occurs due to quantum mechanical tunneling. This is introduced in the next subsection and the other advanced issues related to gate leakage current are dealt with in this section. A remedial approach is to use materials other than SiO_2 as the gate dielectric. This is also introduced here.

6.4.1 Basic Ideas about Quantum Mechanical Tunneling

According to quantum mechanics, each particle in a physical system is described by a wave function $\Psi(x, y, z, t)$. This function and its space derivative $(\partial\Psi/\partial x + \partial\Psi/\partial y + \partial\Psi/\partial z)$ are continuous, finite, and single-valued [184]. The probability of finding a particle in an arbitrary volume $dxdydz$ is

$\Psi^*\Psi dxdydz$ which is normalized so that

$$\int_{-\infty}^{\infty} \Psi^*\Psi dxdydz = 1 \tag{6.26}$$

Let us consider Fig. 6.7, which illustrates a situation where an electron while traveling finds a potential barrier with potential energy U_0 (this represents a potential barrier of finite height) that is higher than the electron energy E. According to classical mechanics, it is not possible for the electron to cross the barrier. However, according to quantum mechanics, the electron wave function has a finite non-zero value within the barrier and also on the other side. This implies that there is some probability of finding the electron beyond the barrier. The mechanism by which the electron penetrates the barrier is called tunneling [184, 192, 189]. The electron wave function will emerge from the barrier with reduced amplitude.

To calculate the tunneling probability, the wave function Ψ has to be determined from the Schrödinger equation, in the three regions that are given as

1. Region-1 ($x < x_1$), potential $U(x) = 0$. The wave function Ψ_1, an oscillatory function which has both forward and backward components is determined from

$$\frac{d^2\Psi_1}{dx^2} + \frac{8\pi^2 m^* E}{h^2}\Psi_1 = 0 \tag{6.27}$$

where h is Planck's constant and m^* is the effective mass of the electron.

2. Region-2 ($x_1 < x < x_2$), $E < U_0$

$$\frac{d^2\Psi_2}{dx^2} + \frac{8\pi^2 m^*}{h^2}\left[E - U(x)\right]\Psi_2 = 0 \tag{6.28}$$

3. Region-3 ($x > x_2$), $U(x) = 0$. The wave function Ψ_3 is an oscillatory function, which has a forward component only and no backward component and is determined from

$$\frac{d^2\Psi_3}{dx^2} + \frac{8\pi^2 m^* E}{h^2}\Psi_3 = 0 \tag{6.29}$$

The solutions of (6.27), (6.28) and (6.29) are given as follows

$$\Psi_1 = A_1 \exp[j(k_1 x)] + B_1 \exp[-j(k_1 x)] \tag{6.30}$$

$$\Psi_2 = A_2 \exp[j(k_2 x)] + B_2 \exp[-j(k_2 x)] \tag{6.31}$$

$$\Psi_3 = A_3 \exp[j(k_3 x)] \tag{6.32}$$

where

$$k_1 = k_3 = \frac{2\pi}{h}\sqrt{2m^* E} \tag{6.33}$$

$$k_2 = \frac{2\pi}{h}\sqrt{2m^*(E - U_0)} \tag{6.34}$$

A_1 and B_1 represent the amplitudes of the incident and reflected waves respectively in Region-1, A_2 and B_2 represent the same in Region-2 and A_3 is the amplitude of the transmitted wave in Region-3. The boundary conditions that are to be satisfied by the electron wave function are that the wave function and its derivatives are continuous at the boundaries.

The tunneling probability is calculated to be [189]

$$\begin{aligned} T_t &= \frac{|A_3|^2}{|A_1|^2} = \left[1 + \frac{U_0^2 \sinh^2(|k_2|W)}{4E(U_0 - E)}\right]^{-1} \\ &\approx \frac{16E(U_0 - E)}{U_0^2} \exp\left(-2\sqrt{\frac{8\pi^2 m^*(U_0 - E)}{h^2}}W\right) \end{aligned} \tag{6.35}$$

The two important features regarding tunneling phenomenon are

- It is observed from (6.35) that the tunneling probability has a negative exponential dependence on the barrier thickness W. This means that a thin barrier increases the tunneling probability.
- For highly energetic particles, the tunneling probability through the barrier is high. This is responsible for the gate leakage current due to hot electrons.

For complicated barrier shapes such as triangular, trapezoidal, etc., the Schrödinger equation is simplified through the WKB (Wentzel–Krammer–Brillouin) approximation provided the potential $U(x)$ does not vary rapidly. The tunneling probability is given by [189]

$$T_t \approx \exp\left\{-2\int_{x_1}^{x_2} \sqrt{\frac{8\pi^2 m^*}{h^2}[U(x) - E]}dx\right\} \tag{6.36}$$

The tunneling current J_t is calculated as the product of the number of electrons in Region-1 and the number of empty states in Region-2, and is given by [189]

$$J_t = \frac{4\pi q m^*}{h^3}\int f_1 N_1 T_t (1 - f_2) N_2 dE \tag{6.37}$$

where f_1 and N_1 represent the Fermi–Dirac distribution and density of states of electrons in Region 1 and f_2, N_2 correspond to that in Region 2 respectively.

6.4.2 Gate Oxide Tunneling Current

The scaling of gate oxide thickness results in an increase in the field across the oxide. The high electric field coupled with low oxide thickness leads to the tunneling of electrons from substrate to gate and also from gate to substrate through the gate oxide, resulting in the gate oxide tunneling current [154]. The physical mechanism of the gate oxide tunneling current, elaborated through the energy band theory is discussed below.

6.4.2.1 Energy Band Theory Model

Let us consider a MOS capacitor with p-type substrate and n^+ doped polysilicon gate. The energy band diagram under flat-band condition[1] is shown in Fig. 6.8(a), where $\phi_{ox}(\approx 3.1eV)$ is the Si-SiO_2 interface energy barrier for electrons. The energy barrier ϕ_{ox} is the difference in energy between the conduction band of SiO_2 and the conduction band of Si. Upon application of a large positive gate bias, the energy band diagram changes as shown in Fig. 6.8(b). The electrons from the strongly inverted surface can tunnel into or through the oxide layer leading to a gate current. On the other hand, with the application of a large negative bias to the gate electrode, the energy band diagram becomes as shown in Fig. 6.8(c). The electrons from the n^+ polysilicon can tunnel into or through the oxide layer leading to a gate current.

The mechanism of tunneling between substrate and gate polysilicon is primarily divided into two parts, namely: (1) Fowler–Nordheim (F-N) tunneling; and (2) direct tunneling [192]. These are discussed below.

6.4.2.2 Fowler–Nordheim Tunneling

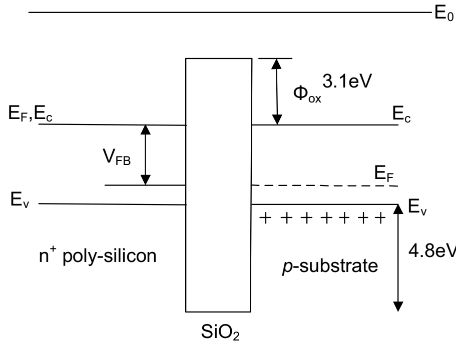
The Fowler–Nordheim tunneling involves tunneling of electrons from the conduction band of the inverted silicon surface to the conduction band of the SiO_2 layer and then hopping of electrons along in the oxide to the gate electrode [184, 192]. The F–N tunneling involves the triangular barrier and the mechanism is illustrated through a band diagram in Fig. 6.9(a). The F–N current is given by [192]

$$J_{FN} = \frac{q^3 \xi_{ox}^2}{8\pi h \phi_{ox}} \exp\left[\frac{-8\pi\sqrt{2m^*}\phi_{ox}^{3/2}}{3hq\xi_{ox}}\right] \tag{6.38}$$

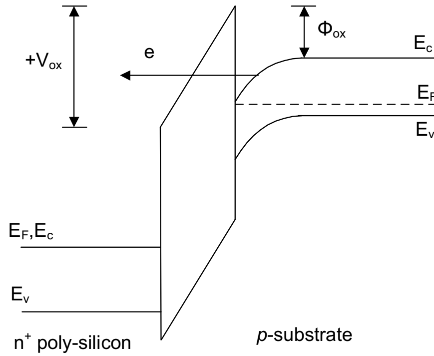
where ξ_{ox} is the electric field in the oxide and ϕ_{ox} is the interface energy barrier for electrons. The F–N current equation represents the tunneling through the triangular potential barrier and is valid for $V_{ox} > \phi_{ox}$, where V_{ox} is the voltage drop across the oxide. At an oxide field of $8MV/cm$, the measured F–N tunneling current density is about $5 \times 10^{-7} A/cm^2$, which is very small.

It may be noted that as electrons are tunneled from Si to SiO_2, the actual

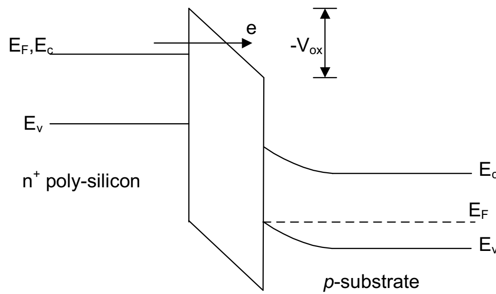
[1] Under the flat-band condition the energy band (E_C, E_V) of the substrate is flat at the Si-SiO_2 interface. The surface electric field in the substrate is zero, so that the electric field in the oxide is also zero.



(a) Energy band diagram at flat-band condition

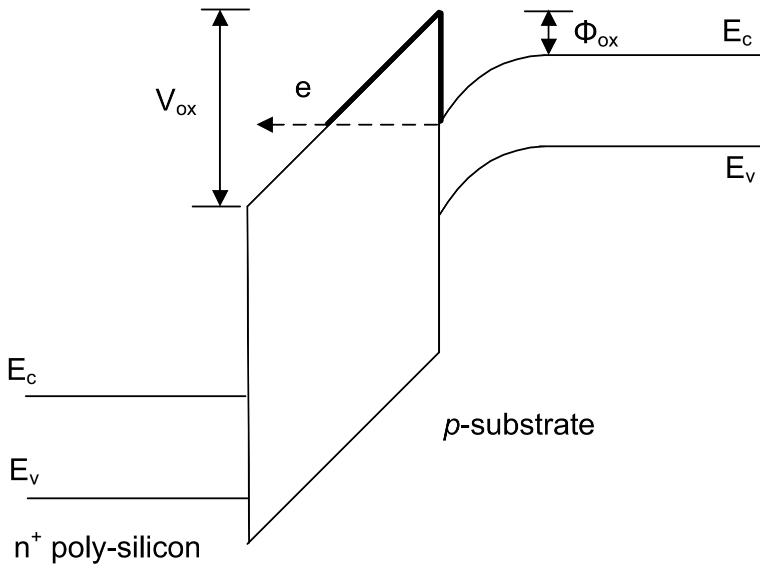


(b) Energy band diagram with positive gate bias showing tunneling of electrons from substrate to gate

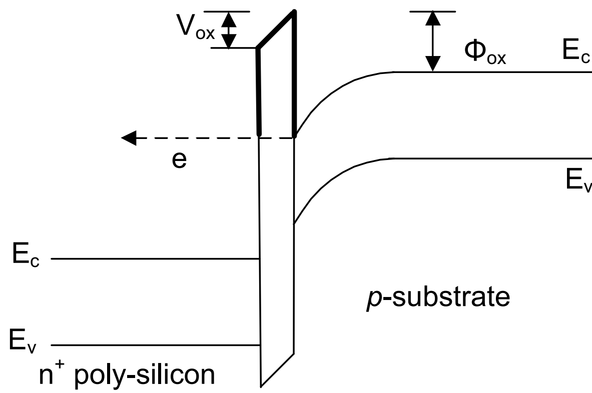


(c) Energy band diagram at negative gate bias showing tunneling of electrons from gate to substrate

FIGURE 6.8
Tunneling of electrons through a MOS capacitor.



(a) Fowler–Nordheim tunneling



(b) Direct tunneling

FIGURE 6.9
Energy band theory for gate oxide leakage tunneling.

TABLE 6.1
Leakage Current Mechanisms under Different Operating Conditions

Mechanisms	I_{GC}	I_{GB}	
Region	Inversion	$V_G > 0$	$V_G < 0$
p-channel	HVB	ECB	EVB
n-channel	ECB	ECB	EVB

energy barrier to tunneling is less than ϕ_{ox} by an amount equal to [192]

$$\Delta\phi = \sqrt{\frac{q^3\xi_{ox}}{4\pi\epsilon_{ox}}} \tag{6.39}$$

This is referred to as the image-force-induced barrier-lowering effect [189].

6.4.2.3 Direct Tunneling

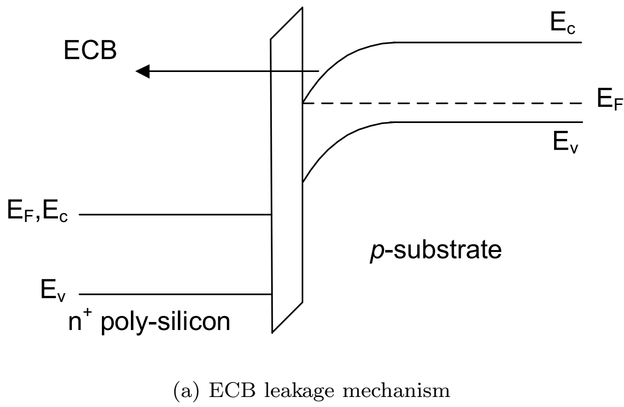
If the oxide layer is very thin, then instead of tunneling into the conduction band of the SiO_2 layer, electrons from the inverted silicon surface tunnel directly through the forbidden energy gap of the SiO_2 layer. The direct tunneling involves a trapezoidal barrier and the mechanism is illustrated through the band diagram in Fig. 6.9(b). The electrons tunnel through a trapezoidal potential barrier instead of a triangular potential barrier. Hence, the direct tunneling occurs at $V_{ox} < \phi_{ox}$. The direct tunneling current density is given by [169]

$$J_{DT} = \frac{q^3\xi_{ox}^2}{8\pi h\phi_{ox}} \exp\left[\frac{-8\pi\sqrt{2m^*}\phi_{ox}^{3/2}\left\{1-\left(1-\frac{V_{ox}}{\phi_{ox}}\right)^{3/2}\right\}}{3hq\xi_{ox}}\right] \tag{6.40}$$

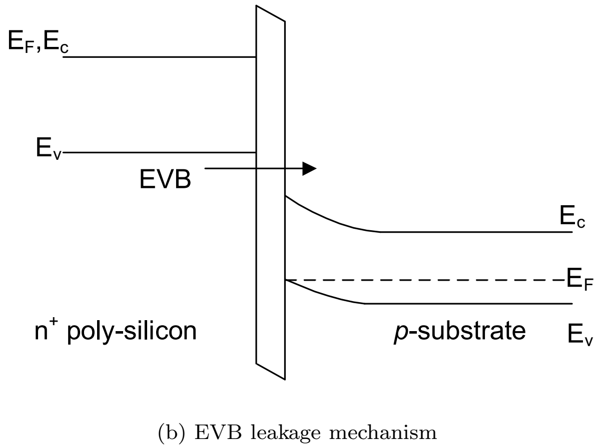
The observation of direct tunneling is limited to oxide thickness less than $50°A$ because the tunneling probability for thicker oxide is small.

6.4.3 Gate Leakage Mechanisms and Leakage Components for MOS Transistors

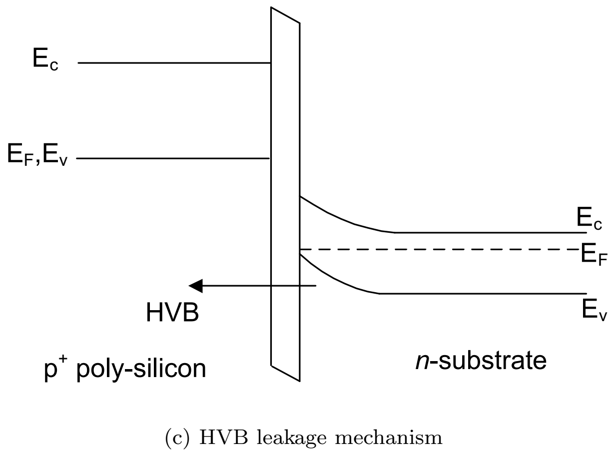
There are three mechanisms of gate dielectric direct tunneling leakage. These are (1) electron tunneling from the conduction band (ECB), (2) electron tunneling from the valence band (EVB), and (3) hole tunneling from the valence-band (HVB)[110, 22]. The last component is also described as valence-band electron tunneling into the valence band. These are schematically shown in Fig. 6.10(a)–6.10(c). Each mechanism is dominant or important in different regions of operation for n-channel and p-channel transistors as listed in Table



(a) ECB leakage mechanism



(b) EVB leakage mechanism



(c) HVB leakage mechanism

FIGURE 6.10
Different mechanisms of gate leakage current in MOS transistor.

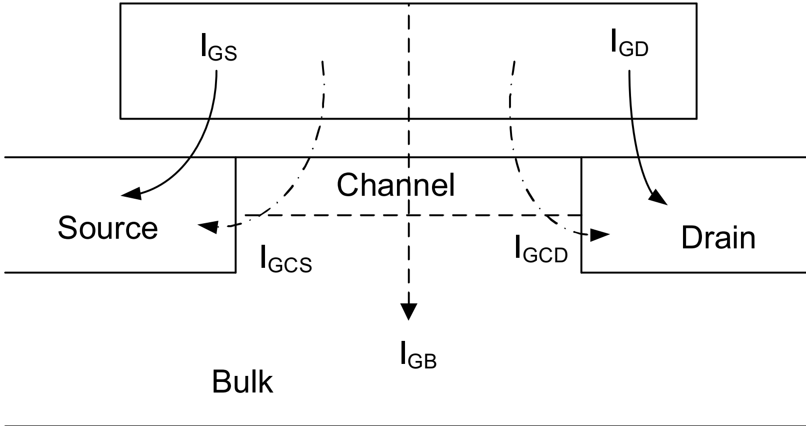


FIGURE 6.11
Schematic diagram showing gate leakage components in n-channel MOS transistor.

6.1. When a transistor operates in an inversion region, tunneling current flows between the gate and the channel, which is represented by (I_{GC}). The gate-to-channel current is partitioned into two parts: one part goes to the source (I_{GCS}) and the other part goes to the drain (I_{GCD}). Tunneling current flows between the gate and the body(I_{GB}) under both accumulation and inversion conditions.

The various components of gate oxide tunneling current are shown in Fig. 6.11. For n-channel MOS transistors, the ECB mechanism controls the gate-to-channel tunneling current in inversion condition, whereas gate-to-body tunneling is controlled by ECB in depletion-inversion and EVB in accumulation. On the other hand, for p-channel MOS transistors, HVB controls the gate to channel leakage in inversion, whereas gate-to-body leakage is controlled by EVB in depletion-inversion and ECB in accumulation. The barrier height for HVB (4.5 eV) is found to be considerably higher than barrier height for ECB (3.1 eV). Consequently, the tunneling current associated with HVB is much less than the current associated with ECB. This leads to a lower gate leakage current in p-channel MOS transistors than in n-channel MOS transistors.

Under all operating conditions, there is a tunneling in the region where the gate overlaps the source and the drain (I_{GS}, I_{GD}). The overlap tunneling current is also known as edge direct tunneling (EDT) and is significant for nano-scale transistors for which the ratio of the source-drain extensions to the channel length is high. It has been observed that the EDT current is more significant compared to other leakage mechanisms such as gate-induced-drain leakage and band-to-band tunneling current for ultra thin gate oxides.

6.4.4 Compact Modeling

The expression for direct tunneling current density given by (6.40) includes a number of approximations which lead to inaccuracies. The use of WKB approximation for ultra-thin gate oxide is questionable. Further, the assumption of a constant effective mass for all energies (all locations at any oxide thickness and gate bias) is also not accurate. Therefore, a quasi-empirical model is suggested in [110] as follows

$$J_{DT} = \frac{q^3}{8\pi h \phi_{ox} \epsilon_{ox}} C(V_G, V_{ox}, t_{ox}, \phi_{ox}) \times \exp\left\{ \frac{-8\pi\sqrt{2m^*}\phi_{ox}^{3/2}\left[1-\left(1-\frac{|V_{ox}|}{\phi_{ox}}\right)^{3/2}\right]}{3hq\,|\xi_{ox}|} \right\} \tag{6.41}$$

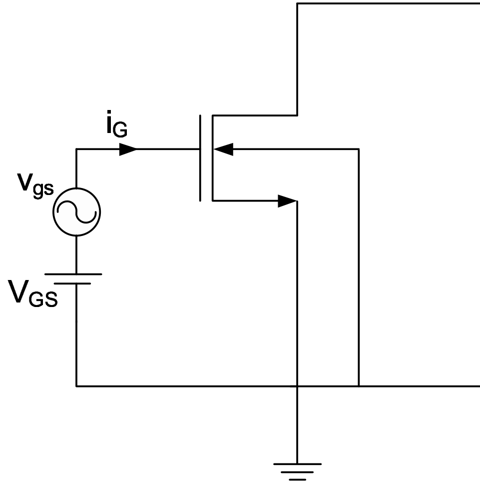
BSIM4 gate current model is based on this semi-empirical approach and uses a common expression for all the tunneling current components and mechanisms. The BSIM4 model equation for the tunneling current is [22, 1]

$$J_{DT} = A\left(\frac{T_{oxref}}{T_{ox}P}\right)^{ntox} \frac{V_{aux}V_{appl}}{(T_{ox}P)^2}.\exp\left[-B\left(\alpha-\beta\,|V_{ox}|\right)\left(1+\gamma\,|V_{ox}|\right)T_{ox}P\right] \tag{6.42}$$

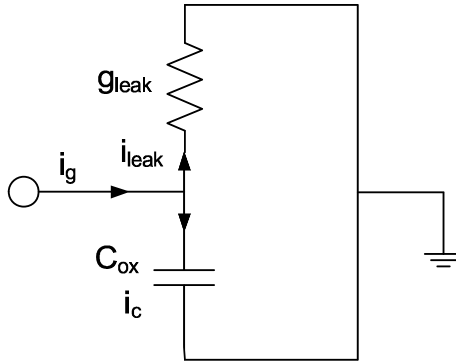
where $A = q^2/(8\pi h\phi_{ox}), B = 8\pi\sqrt{2qm^*}(q\phi_{ox})^{3/2}/(3h), T_{oxref}$ is the reference oxide thickness for parameter extraction, $ntox$ is a fitting parameter, and V_{aux} is an auxiliary function modeling the density of tunneling carriers and available states in various regions of operations. V_{appl} is the applied voltage which has different meaning depending on the current components and the parameters α, β, γ, P depend on the tunneling mechanism (ECB, EVB or HVB), region of operation, and the current components. For detailed discussions, the readers may consult the BSIM4 manual [55].

6.4.5 Effects of Gate Leakage

One obvious implication of a non-zero gate leakage current is that the gate terminal now includes a tunnel conductance in parallel with the traditional capacitance [6, 112]. In order to study the implications of the gate leakage current, let us consider the simple circuit shown in Fig. 6.12(a). The small signal equivalent circuit behavior of the gate terminal can be represented as shown in Fig. 6.12(b) [6, 112]. The variation of the gate current for an n-channel MOS transistor with the applied gate bias and the signal frequency is shown in Fig. 6.13(a)–6.13(b). It is observed from Fig. 6.13(a) that the gate current increases sharply with V_{GS} in the strong inversion region. Further, the leakage current increases with the scaling down of process technology. At the 32 nm technology node, the gate leakage current has increased by nearly

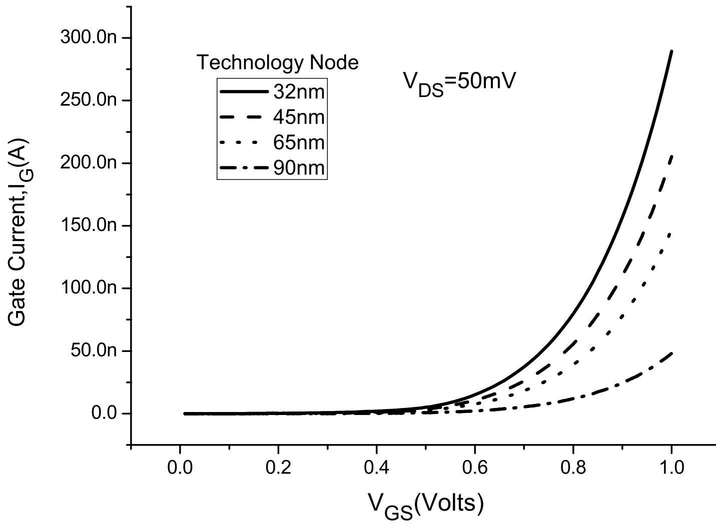


(a) Schematic circuit diagram

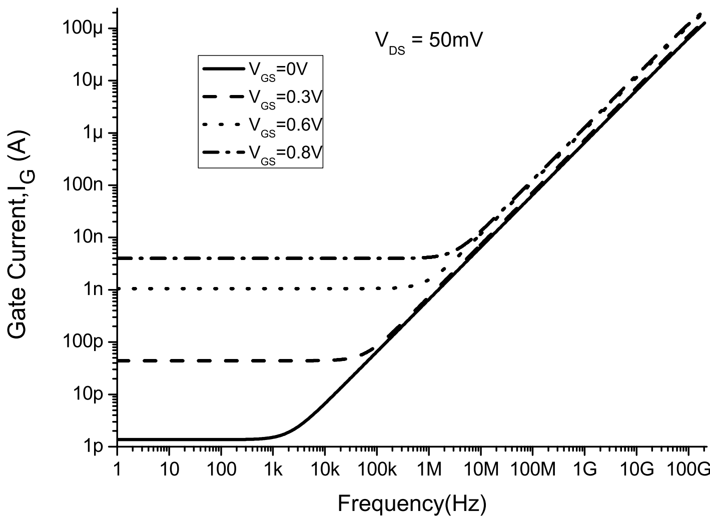


(b) Small signal equivalent model of the gate terminal considering low to moderate frequencies and strong inversion

FIGURE 6.12
Investigation of the gate leakage current of an n-channel MOS transistor.



(a) variation with gate bias



(b) variation with the signal frequency

FIGURE 6.13
Variation of gate current for n-channel MOS transistor.

six orders of magnitude as compared to the level for the 90 nm technology node. It is observed from Fig. 6.13(b) a non-zero amount of gate current flows even in the low frequency region of the applied gate signal. When the value of the applied signal frequency is increased after a certain threshold value, the magnitude of the gate current increases considerably.

The component of the gate current flowing through the capacitive component, i.e., i_c depends upon the signal frequency, while the component i_{leak} does not depend upon the frequency. Therefore, there exists a characteristic frequency f_{gate} where the two current components have equal magnitude. This frequency is given by [112]

$$f_{gate} = \frac{g_{leak}}{2\pi C_{ox}} \tag{6.43}$$

For frequencies much larger than f_{gate}, i_c becomes much larger than i_{leak}, the gate current becomes mostly capacitive and the gate behaves like a conventional MOS gate. Otherwise, below f_{gate}, it is mainly resistive and the gate leakage is dominant. The value of this frequency is extracted from SPICE simulation results by plotting the phase of the AC small signal gate current versus frequency and observing the frequency at which the phase shift is 45°. The variation of f_{gate} with the applied gate bias for the various technology nodes are shown in Fig. 6.14(a) and Fig. 6.14(b). It has been found that f_{gate} is independent of the gate area and the drain bias. An empirical expression of f_{gate} has been derived in [6] as

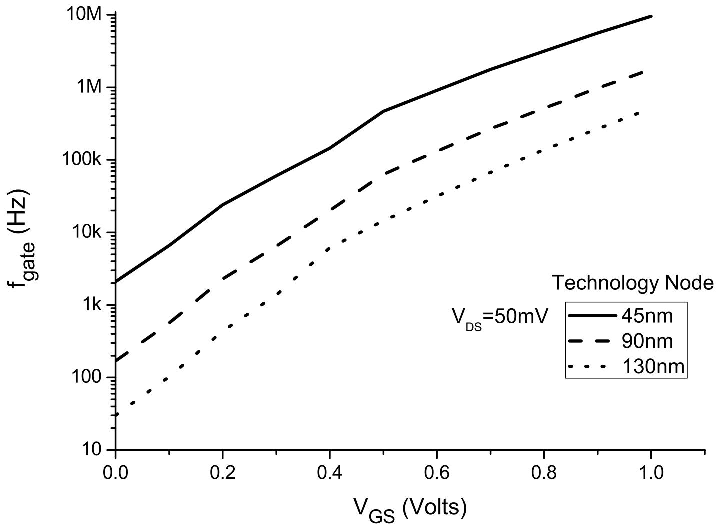
$$\begin{aligned} f_{gate} &\approx 0.5 \times 10^{16} \times v_{GS}^2 . \exp\left[t_{ox}\left(v_{GS} - 13.6\right)\right] \text{ for PMOSFET} \quad (6.44) \\ &\approx 1.5 \times 10^{16} \times v_{GS}^2 . \exp\left[t_{ox}\left(v_{GS} - 13.6\right)\right] \text{ for NMOSFET} \quad (6.45) \end{aligned}$$

A significant effect of the enhanced gate leakage current is the limited current gain. The variation of the low-frequency current gain of MOS transistors in advanced CMOS technologies as a function of gate length is shown in Fig. 6.15.

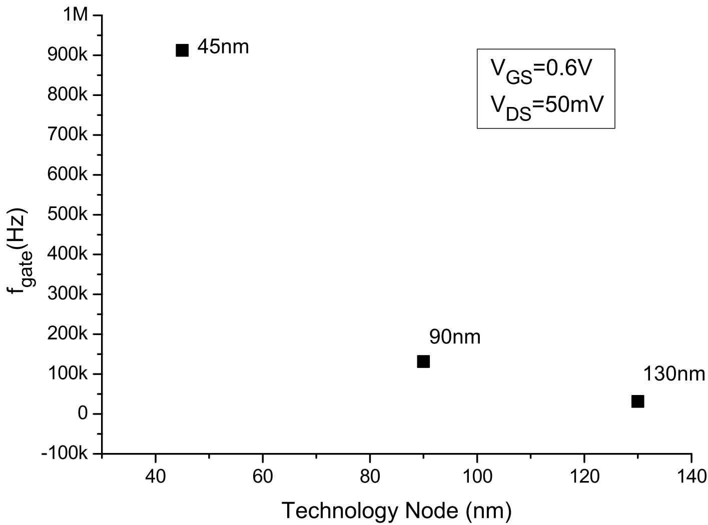
6.5 High-κ Dielectrics and Metal-Gate/High-κ CMOS Technology

6.5.1 High-κ Dielectric Materials

It is clear from the previous discussion that the primary cause of gate oxide tunneling current is the reduced thickness of the gate oxide region. In order to avoid the tunneling problem, there has been an intense search for alternative dielectrics with high-κ (permittivity) which have properties very close to SiO_2 but offer the opportunity to use a higher thickness for the gate dielectric for the same value of the gate oxide capacitance. The gate capacitance of a MOS



(a) As function of gate voltage



(b) Typical values for various technology nodes

FIGURE 6.14
Investigation of f_{gate} of an n-channel MOS transistor.

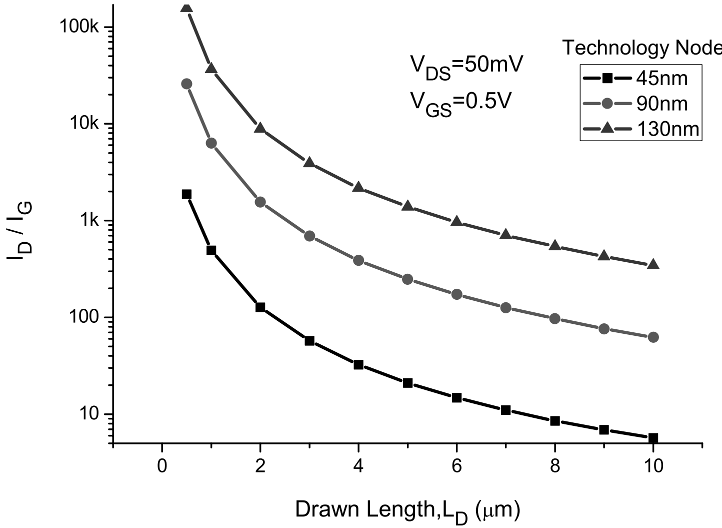


FIGURE 6.15
Limited current gain with scaling of technology.

transistor using any dielectric material with thickness T_d is given by

$$C_G = \frac{\epsilon_0 \kappa_d A}{T_d} \tag{6.46}$$

where ϵ_0 is the permittivity of free space, κ_d is the relative permittivity of the dielectric material, A is the area of the conducting plates, and T_d is the gate dielectric thickness. The thickness of the high-κ dielectric insulator is obtained from the following relation [1].

$$T_{ox} = \text{Effective oxide thickness (EOT)} = \frac{\kappa_{\text{SiO}_2}}{\kappa_d} T_d \tag{6.47}$$

The relative permittivity of hafnium dioxide HfO_2 is six times more than that of SiO_2. Therefore, a film of HfO_2 of thickness $6nm$ has an EOT of $1nm$, in the sense that both the films produce the same oxide capacitance. However, the HfO_2 film is physically much thicker compared to that of the SiO_2 film. Consequently, the electric field across the HfO_2 gate material will be much less compared to that across the SiO_2 gate material. Some other popular high-κ dielectric insulator materials are ZrO_2 and Al_2O_3.

It is a challenging task in itself to integrate these materials into the conventional CMOS processes. The several issues include chemical reactions between these materials and the silicon substrate, lower surface mobility and more oxide trapped charges. These problems are reduced to some extent by inserting a thin SiO_2 interfacial layer between the silicon substrate and the high-κ dielectric insulator.

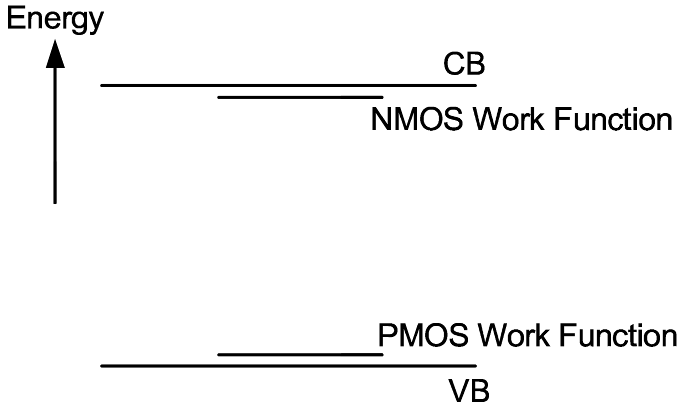


FIGURE 6.16
Requirement of work function of metal gates for n-channel and p-channel MOS transistors.

It is essential to include high-κ dielectric material-based MOS transistors in the CAD framework. These are possible through two ways [103]: (i) varying the model parameter in the SPICE model file that denotes relative permittivity (EPSROX) and/or (ii) finding the EOT for a dielectric under consideration. The first approach is not sufficient for modeling the behavior of non-classical nano-CMOS with non-SiO_2 dielectrics as it does not correctly account for the barrier height of non-SiO_2 dielectrics. Therefore, the second approach is the preferred one [103].

6.5.2 Metal Gate

Degenerately doped polycrystalline silicon (poly-Si) is used as the gate electrode due to its compatibility with SiO_2. Limited carrier densities used in poly-Si give a depletion depth of a few angstroms. This leads to the poly-gate depletion effect as discussed in Chapter 3. However, good metals with higher carrier densities have a depletion depth of only 0.5 angstrom. Hence, the depletion effect in poly-Si may be greatly reduced by replacing poly-Si with a metal, typically like TiN. Again, some high-κ oxides react with poly-Si. Thus, replacing poly-Si by a metal not only is desirable but is essential too.

The gate metal is chosen predominantly for its work function and thermal robustness [153]. The purpose of the gate electrode in CMOS is to shift the Fermi level of the Si channel to the appropriate band edge, to create inversion. This is illustrated in Fig. 6.16 that shows the work function requirement of metal gates for n-channel MOS and p-channel MOS transistors. A simple choice for both n-channel MOS and p-channel MOS in CMOS circuits is to use the same metal for both of them. The work function would then correspond to the mid-gap energy of Si, about 4.6eV. However, this easy choice is the worst

in terms of device properties so far because the threshold voltage for the metal gate MOSFET is larger than that for poly-Si MOSFET. The other option is to use different metals for n-channel MOS and p-channel MOS transistors. An nMOS, consisting of a p-Si substrate requires a metal with work function close to the Si conduction band energy (4 eV) below the vacuum level. These metals are quite reactive. The p-channel MOS, on the contrary, requires a metal with work function close to the Si valence band energy (5.1 eV). These metals are very noble like, Pt, and are difficult to etch. Elemental metals cause problems as they tend to react with SiO_2 or the high-κ oxide. Instead, high stability "diffusion barrier" materials like nitrides, carbides, and silicides of transition metals (e.g., TiN) might be used. These materials however, do not have a wide range of work functions. CMOS fabrication with molybdenum (Mo) as the gate material has been reported in literature. Mo exhibits a high work function that is suitable for bulk p-MOSFETs and this value may be lowered for n-MOSFETs with nitrogen (N) implantation. The choice of the gate metal is thus a critical issue in the CMOS circuits and needs to be made carefully.

6.6 Advanced Device Structures of MOS Transistors

This section presents three non-conventional device architectures of MOS transistors that offer significant improvements in device performances and power consumption. These are the silicon-on-insulator (SOI) MOS transistor, double gate (DG)-MOS transistor and FinFET. These are briefly discussed below.

6.6.1 SOI MOS Transistor

The silicon-on-Insulator or SOI MOS transistor involves development of the conventional MOS transistor on very thin layers of crystalline silicon [191, 192, 189]. The thin layer of silicon is isolated from the substrate by a thick layer of buried oxide (typically $1000A^{\circ}$ or more). This layer electrically isolates the transistors from the underlying silicon substrate and from each other. The schematic structure of an SOI MOS transistor is shown in Fig. 6.17. SOI structures are of two types [192]. These are partially-depleted SOI (PDSOI) and fully-depleted SOI (FDSOI). In PDSOI, the sum of the gate depletion widths from the front and back ends is smaller than the silicon film thickness, t_{Si}. These devices exhibit a floating body effect (kink effect). In FDSOI, the silicon film is thin enough that the entire film is depleted before the threshold condition is attained. The sub-threshold slope of these devices is steeper than that of the bulk devices. The FDSOI devices perform better than the PDSOI. The SOI structures exhibit lower leakage, low junction capacitance, low latch-up, and better sub-threshold swing. The performance of

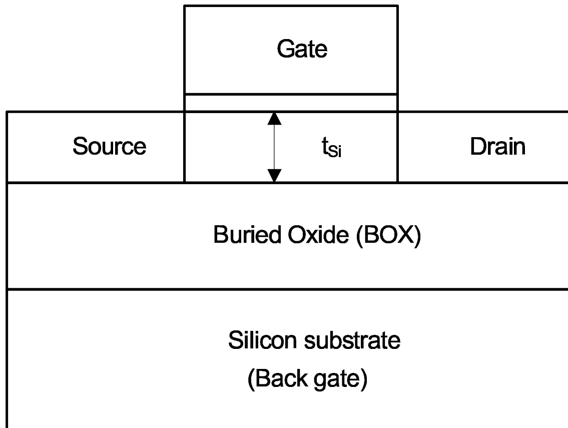


FIGURE 6.17
Thin film SOI MOS transistor.

these devices depends on the silicon film thickness, BOX layer thickness and the doping concentrations.

6.6.2 Double Gate (DG)-MOS Transistors

The double gate (DG)-MOS transistors improve the performance of CMOS devices and overcome some of the difficulties faced in the downscaling of MOS transistors [32]. The DG-MOSFET was originally proposed in 1984 as "XMOS." Threshold voltage roll-off, DIBL, off-state leakage, etc., are significantly reduced with a nearly ideal sub-threshold swing of 60mV/decade and hence these devices are preferred in nano-scale circuits [96]. Figure 6.18 shows the schematic diagram of a typical DG-MOS transistor. It consists of a silicon slab sandwiched between two oxide layers. The gate is a metal or a polysilicon film. The front and back gate electrodes create inversion layers near the two Si-SiO_2 interfaces upon the application of a suitable bias. Thus, there are two MOS transistors that share the same source, drain, and substrate. The short channel effects of DG MOS transistors are controlled by device geometry and not by doping (channel doping or halo doping). This is in contrast to the conventional bulk MOS transistors that suffer from a degraded swing due to a high channel doping concentration. The two gate electrodes simultaneously control the carriers. Consequently, the effect of the drain field on the channel is screened out. In addition, the thin silicon channel leads to a stronger coupling of the gate potential with the channel potential. The reduced SCEs lead to greater scalability than bulk MOS transistors. This allows the use of thicker oxides in DG-MOS transistors compared to the bulk MOS transistors, thereby reducing the gate leakage current. The undoped channel reduces mobility degradation by eliminating impurity scattering, thereby improving

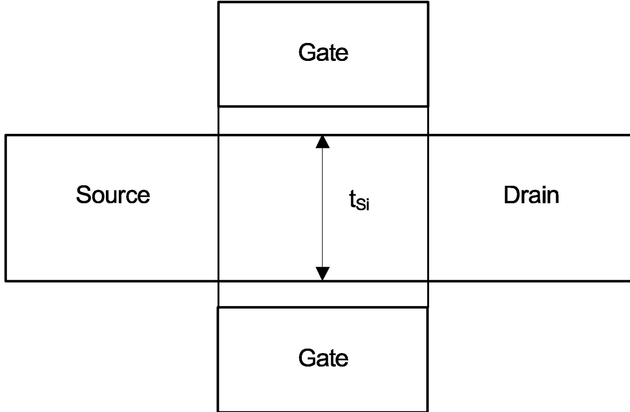


FIGURE 6.18
Schematic diagram of a double gate MOS transistor.

the carrier transport. The random dopant fluctuations are also avoided. The current drive (or gate capacitance) per unit area is increased. The threshold voltage in DG-MOS transistors with intrinsic silicon channel may be adjusted by gate electrodes with mid-gap work function, e.g., tungsten.

The DG-MOS transistors are broadly classified as symmetric DG-MOS and asymmetric DG-MOS [32]. In the former, the two oxide thicknesses are the same and the two gates have the same flat-band voltage i.e., same work function (near mid-gap metals) and the gates are connected together. In the latter, the two oxide thicknesses are different and the work functions of the gate materials may differ.

Considering the gate materials and body doping, the following three types of DG-MOSFETs have been proposed. (i) a doped body symmetric device with poly-gates (n+ poly for NMOS, SymDG), (ii) an intrinsic body symmetric device with near mid-gap metal gates (MGDG) and (iii) an intrinsic body asymmetric device with different front and back-gate work functions (e.g., n+ poly/ p+ poly, AsymDG). In the SymDG-MOS transistor, the threshold voltage is adjusted by using body doping, in MGDG MOS transistor this is done using the metal work function and in the AsymDG-MOS transistor the threshold voltage is controlled by the work function difference between the front and back gates.

It is to be noted that in DG-MOS transistors with undoped channel (nearly intrinsic), there is almost no depletion charge and the average vertical electric field is dependent only on the inversion charge density. Thus, compared to the bulk devices, the reduction in the vertical field in DG-MOS transistors improves the carrier mobility. Thus, the gate-to-channel tunneling current is less as compared to the bulk MOS transistors.

The various types of DG-MOS transistor structures are

1. Planar DG-MOS Structure: This structure as shown in Fig. 6.19(a)., is similar to a planar MOSFET except that it has a bottom gate. Though this structure offers good control of the silicon channel thickness, the fabrication of the self-aligned bottom gate is very challenging.
2. Vertical DG-MOS Structure: In these device structures as shown in Fig. 6.19(b) and 6.19(c), the current flow is perpendicular to the wafer. The vertical structure is very attractive for DRAM applications since the gate length is decoupled from the packing density.

The DG-MOS transistors are associated with several phenomena like "self-heating", quantum mechanical effects, "volume inversion" and misalignment of the top and the bottom gates. In an ultra-thin DG MOS transistor the carriers are distributed throughout the entire silicon volume. This increases the mobility of the carriers by reducing their scattering at the oxide and interface traps.

6.6.3 FinFETs

The FinFET is believed to be one of the promising structures that improves the gate controllability and thus minimizes the short channel effect (SCE) through adoption of multiple gates [32]. A typical FinFET structure is shown in Fig. 6.19(c).

The thickness or width of a single fin is equal to the silicon film thickness, t_{Si}. The channel width is basically twice the fin-height plus the fin-width. The device width is quantized into units of the fins. Greater widths are obtained by using multiple fins. The conducting channel is wrapped around the surface of the fin and resembles the fin of a fish. Hence, the name "FinFET". Since the source/drain and gate are much thicker (taller) than the fin, the device structure is quasi-planar. The three-dimensional (3D) structure of the device requires 3D analyses to obtain a reasonable prediction of device performance and structural optimization. In fact, the short channel effects of the FinFETs are essentially three-dimensional phenomena which are sensitive to the geometry of the device. Therefore, 3D-process and device simulations are indispensable to the design of FinFETs.

It is interesting to note that when the height of the fin is much larger than the thickness of the silicon film or the top gate oxide is much thicker than the front and back gate oxides, the FinFET is approximately treated as a DG-MOSFET. The horizontal cross-section then appears very similar to the conventional DG-MOSFET structure.

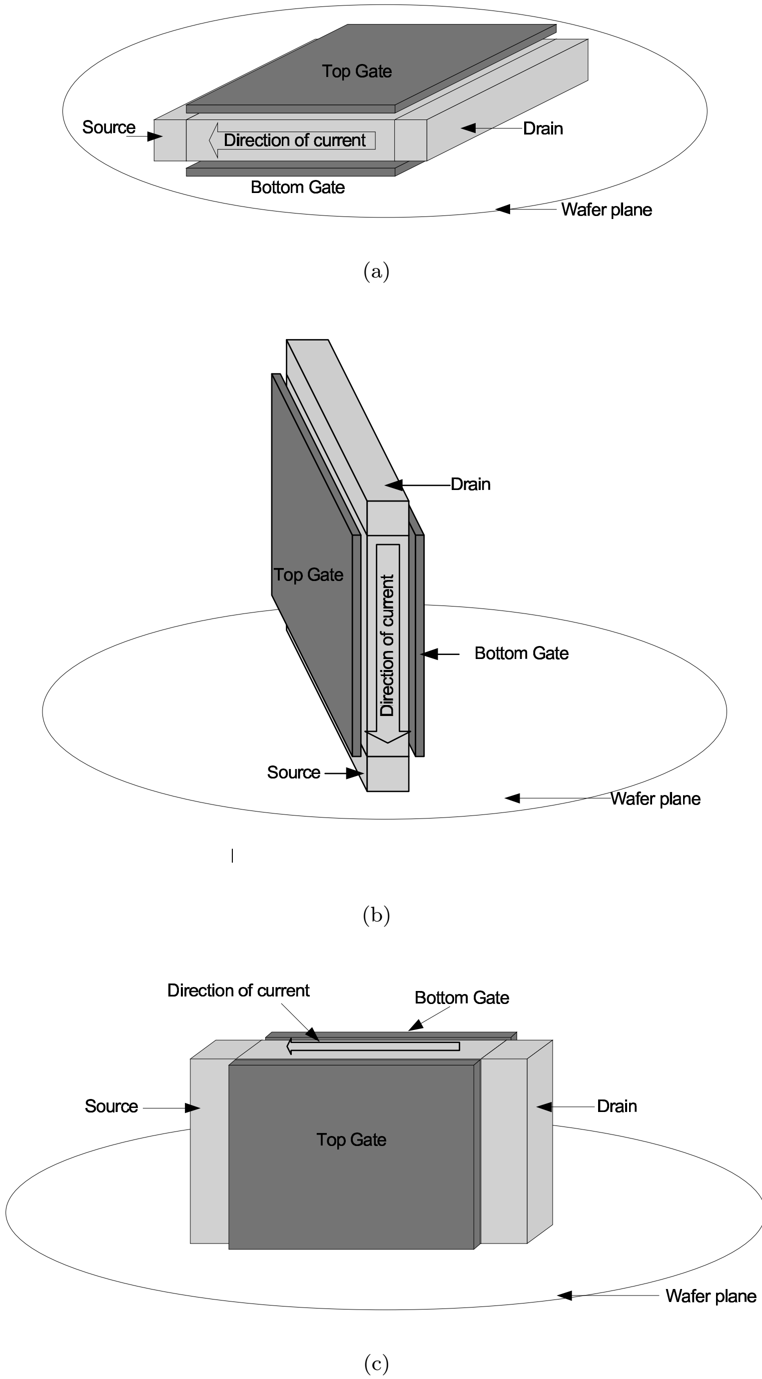


FIGURE 6.19
Different architectures of DG-MOSFET: (a) planar DG-MOSFET (b), (c) vertical DG-MOSFETs.

6.7 Noise Characterization of MOS Transistors

Noise in electronic devices originates from internal random fluctuations of a deterministic signal inherent to the physics of the device [73, 194]. This type of noise is known as true noise and arises from physical processes governing the electron transport in the medium. True noise cannot be eliminated, but it is possible to reduce it by proper design of the devices and circuits. An excellent text dealing with the low frequency noise characterization of MOS transistors is that by Hartman and Ostling [73], which may be referred to by the readers for details of the following discussions.

The current through a device may be written as [73]

$$I(t) = \bar{I} + i_n(t) \tag{6.48}$$

where $\bar{I}$ is the average bias current and $i_n(t)$ is a randomly fluctuating current. Since $i_n(t)$ cannot be predicted, noise is described with averages measured over a long time. In practice, all fluctuating currents and voltages in electronic devices follow the Gaussian (normal) distribution due to the central limit theorem that states the sum of a large number of independent random variables has a normal distribution. An important exception is the switching of the signal between two levels, random-telegraph signal (RTS) noise, which is a Poisson process. A useful approach to describe noise is to convert the problem from the time domain to the frequency domain by Fourier transformation.

Noise power spectral density (PSD) gives information about the distribution of noise power in frequency. The PSD of noise current and noise voltage have dimensions of A^2/Hz and V^2/Hz, respectively. Noise with a constant PSD for all frequencies is said to be "white." It is usually observed that noise PSD is dependent on frequency at low frequencies, and becomes white thereafter.

6.7.1 Fundamental Sources of Noise

Random fluctuations in the current (or voltage) in a device are generated by some fundamental processes in the device. The average current in a slab of length L may be written as [73]

$$\bar{I} = q\overline{Nv_d}/L \tag{6.49}$$

where q is the electron charge, N is the number of free carriers in the slab, and v_d is the drift velocity of the electrons. Since both N and v_d can fluctuate, it can be written that

$$I(t) = \sum_{j=1}^{N(t)} q\frac{v_{dj}(t)}{L} \tag{6.50}$$

where j is for an individual carrier and

$$N(t) = \overline{N} + \Delta N(t) \tag{6.51}$$

$$v_{dj}(t) = \overline{v_{dj}} + \Delta v_{dj}(t) \tag{6.52}$$

Since the average drift velocity is the same for each carrier, the fluctuating current is written as

$$\Delta I(t) = \frac{q}{L}\overline{v_d}\Delta N(t) + \frac{q}{L}\sum_{j=1}^{\overline{N}} \Delta v_{dj}(t) \tag{6.53}$$

Here, the first term is attributed to the fluctuations in the carrier number and the second term to the fluctuations in the carrier velocity. These fluctuations give rise to the fluctuations in the current and voltage. However, the fluctuations in the carrier number and velocity are caused by separate physical mechanisms. Again, since the drift velocity is proportional to the electric field, the velocity fluctuations may alternatively imply carrier mobility fluctuations.

The various types of noise present in an electronic device are as follows [73]:

1. Thermal (Johnson/Nyquist) noise: This arises from the random thermal motion of electrons in a material. When an electron gets scattered, its velocity is randomized. Thus, at a particular instant, the number of electrons moving in a certain direction may be more than that in another direction and a small net current flows. This current fluctuates in magnitude and direction, but the average over a long time is always zero. The PSD of thermal noise current in a material of resistance R and temperature T is

$$S_i = \frac{4kT}{R} \tag{6.54}$$

   This is sometimes alternatively written as

$$S_v = 4kTR \tag{6.55}$$

   Thermal noise is not white up to infinitely high frequencies. It exists in every resistive medium and is unavoidable. However, it may be minimized by proper circuit designing. For instance, input matching techniques using reactive elements can be used to lower the noise in amplifiers since reactive elements do not generate thermal noise. Also, system bandwidth should be kept as small as possible to pass the desired signal since unused portions of the bandwidth cause unnecessary noise.

2. Shot noise: The current flowing through a potential barrier (as in a p-n junction) does not remain continuous due to the discrete nature of the electronic charge. When the electrons cross the barrier at random, a shot noise current is generated. Shot noise is a Poisson process.

3. Generation-recombination noise (GR noise): This noise in semiconductors is caused by traps that randomly capture and release carriers, thereby leading to fluctuations in the carrier number. Electronic states within the forbidden gap of a semiconductor are known as traps that exist due to the presence of impurities or defects within the semiconductor and/or at its surface. Traps may be neutral or charged in its empty state. A trapped charge may further induce fluctuations in the carrier mobility, diffusivity, electric field, transition region width, etc. The PSD of GR noise is proportional to the number of traps and inversely proportional to the carrier number and is of the Lorentzian type.
4. Random telegraph signal (RTS/burst/popcorn) noise: This is a special case of GR noise. When only a few traps are involved, the current switches between two or more states resembling a RTS waveform in the time domain. The PSD is of Lorentzian type. GR noise may be considered to be a sum of many RTS processes. RTS noise is observed in MOS devices with small gate area. It is sensitive to current crowding in the device or a poor contact.
5. 1/f or flicker noise: This noise fluctuation has a PSD proportional to $1/f$. The PSD is of the form [73, 194]

$$S_{I_D} = \frac{K_F . I_{DS}^{AF}}{f C_{ox} W L} \tag{6.56}$$

where K_F is the flicker noise coefficient and AF is the flicker noise exponent. The value of the parameter AF lies in the range of 0.5 to 2. The flicker noise coefficient is proportional to the interface trap density, which is technology specific. It is observed that the flicker noise spectrum at frequencies above $100 MHz$ becomes negligible compared to that of the thermal noise. The flicker noise spectrum reduces as the gate area is increased. It has been found that the value of the flicker noise coefficient is less for p-channel MOS transistors compared to that of the n-channel MOS transistors. Therefore, p-channel MOS transistors are used in designing low noise circuits, at least in the first stage.

6.7.2 Characterization of Thermal Noise in MOS Transistors

The intrinsic thermal noise of a MOS transistor originates from the channel resistance due to the random thermal motion of the carriers. The channel of a MOS transistor may be considered to be divided into several resistive segments and each of these segments contributes to thermal noise [194, 1].

The thermal noise PSD is given as

$$S_{I_D} = \frac{8kT}{3} g_m \tag{6.57}$$

In order to match the model with experimental results, especially in the linear region where the transconductance is zero, (6.57) is modified as follows

$$S_{I_D} = \frac{8kT}{3}\left(g_m + g_{ds} + g_{mb}\right) \tag{6.58}$$

where g_{ds} and g_{mb} are the output conductance and body transconductance respectively. A more rigorous approach for characterizing the thermal noise is given below.

Let us consider an infinitesimally small section of the noiseless channel of a MOS transistor of length dy, the resistance of which is dR. If the channel voltage across this section be dV_{CS}, then the drain current (assumed noiseless), in the absence of velocity saturation is given by

$$dV_{CS} = I_{DS}.dR = -\mu_s W Q_n \frac{dV_{CS}}{dy}.dR \tag{6.59}$$

Then dR is given as

$$dR = -\frac{dy}{W\mu_s Q_n} \tag{6.60}$$

The PSD for the elemental noise voltage is [194]

$$dS_{Vc} = 4kTdR = -4kT\frac{dy}{W\mu_s Q_n} \tag{6.61}$$

Corresponding to this, the elemental noise current PSD is given by

$$dS_{I_D} = g_c^2 dS_{V_c} \tag{6.62}$$

where g_c is the conductance of the elemental channel segment and is given by

$$g_c = \frac{dI_{DS}}{dV_{CS}} = -\frac{d}{dV_{CS}}\left[\frac{W}{L}\mu_s \int Q_n dV_{cs}\right] = -\mu_s \frac{W}{L} Q_n \tag{6.63}$$

Therefore, by substitution the elemental noise current PSD is given by

$$dS_{I_D} = -4kT\frac{\mu_s}{L^2} W Q_n dy \tag{6.64}$$

Integrating over the entire channel length, the total noise current power spectral density of the thermal noise is given by

$$S_{I_D} = -4kT\frac{\mu_s}{L^2}\int_0^L Q_n W dy \tag{6.65}$$

The integral term multiplied by W represents the total inversion charge under the gate. Therefore, the final expression for the total noise current PSD of the thermal noise is given by

$$S_{I_D} = 4kT\frac{\mu_s}{L^2} Q_{INV} \tag{6.66}$$

where Q_{INV} represents the total inversion charge under the gate. This is the channel thermal noise model used in BSIM3v3 with proper substitution of the inversion charge [30]. The variation of the drain current noise with respect to frequency for n-channel and p-channel MOS transistors as obtained from SPICE simulation are shown in Fig. 6.20(a)−6.20(b). It is observed that the channel thermal noise PSD is independent of frequency, at least in the range of frequencies where the assumption of quasi-static behavior is valid [194]. The variation of the thermal noise for an n-channel MOS transistor with applied gate and drain bias is shown in Fig. 6.21(a)−6.21(b). The thermal noise increases with the increasing gate voltage V_{GS}. However, it depends weakly on the drain voltage V_{DS}. This indicates that the noise contribution from the velocity saturation region of the channel is negligible. This is theoretically justified from the thermal noise model.

6.7.3 Characterization of Flicker Noise in MOS Transistors

Low $1/f$ noise in MOS transistors is an important requirement for low-noise and RF/analog applications. Accurate noise models are therefore essential for the the VLSI designers in order to reduce the $1/f$ noise in the MOS transistors. The physical mechanism of flicker noise is important for the designers to understand [73].

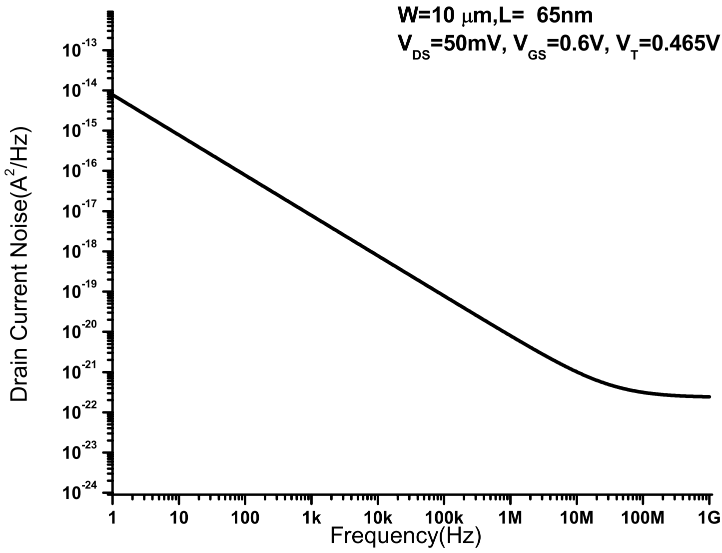
6.7.3.1 Physical Mechanism of Flicker Noise

There are several different theories for explaining the physical cause of flicker noise. These are broadly classified into three different categories [73, 194]: (1) carrier density fluctuation model, (2) mobility fluctuation model, and (3) correlated carrier density and mobility fluctuation model.

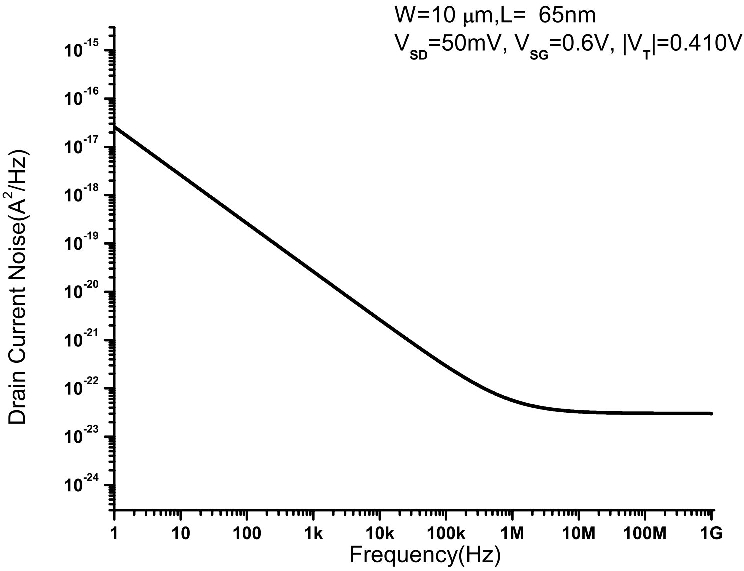
The carrier density fluctuation model attributes the origin of flicker noise to the random fluctuation of the number of carriers in the channel, due to fluctuations in the surface potential, which in turn are caused by the trapping and releasing of carriers by the traps located near the semiconductor-oxide interface [194]. The carrier density fluctuation model is observed to successfully explain the flicker noise spectrum in n-channel MOS transistors.

The mobility fluctuation model attributes the origin of flicker noise to the fluctuations of the mobility of the carriers, which are caused by the interactions of the carriers with lattice fluctuations. The mobility fluctuation model successfully explains the flicker noise spectrum in p-channel MOS transistors.

According to the correlated carrier density and mobility fluctuation model, also referred to as the unified flicker noise model [89, 88], when an interface trap captures an electron from the inversion layer, it becomes charged and reduces the carrier mobility due to Coulombic scattering. Thus according to this model, both the carrier number and the carrier mobility fluctuates due to trapping and de-trapping of the carriers by the interface traps.

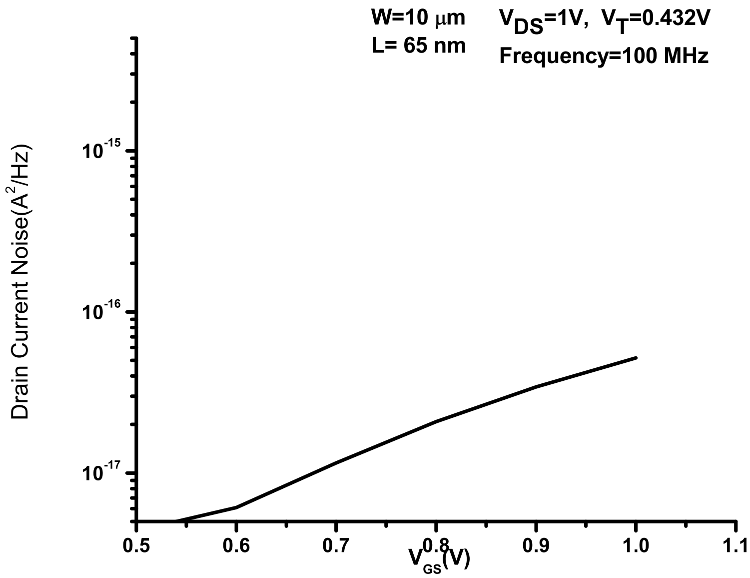


(a) NMOS

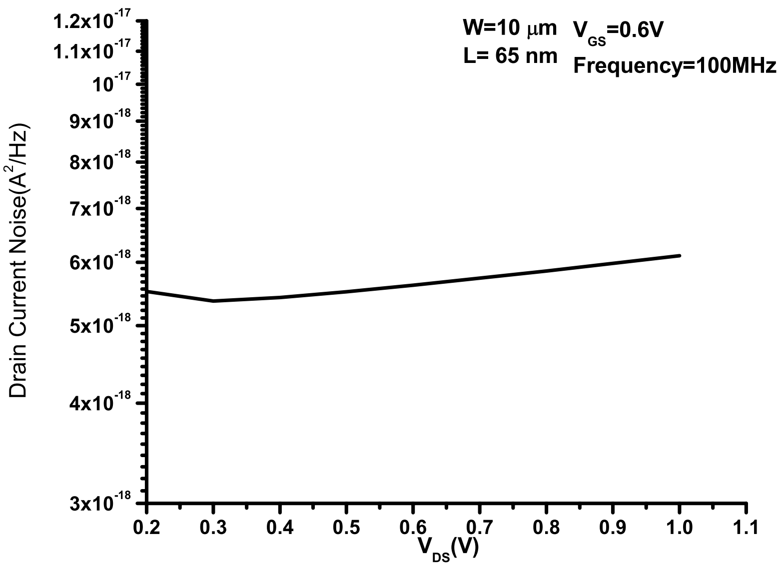


(b) PMOS

FIGURE 6.20
Noise PSD of n-channel and p-channel MOS transistors.



(a) variation with gate bias



(b) variation with drain bias

FIGURE 6.21
Simulated thermal noise PSD for n-channel MOS transistor.

6.7.3.2 Physics-Based Modeling of Flicker Noise

Let us consider a section of the channel of width W and length Δy. The drain current is given by

$$I_{DS} = W\mu_s qN\xi_y \tag{6.67}$$

where μ_s is the carrier mobility at the surface, q is the electron charge, N is the number of channel carriers per unit area, and ξ_y is the lateral channel field. The fluctuation in the local drain current caused due to the combined effects of the carrier number fluctuation and mobility fluctuation is given by [89, 88]

$$\frac{\delta I_{DS}}{I_{DS}} = -\left(\frac{1}{\Delta N}\frac{\delta \Delta N}{\delta \Delta N_t} \pm \frac{1}{\mu_s}\frac{\delta \mu_s}{\delta \Delta N_t}\right)\delta \Delta N_t \tag{6.68}$$

Here $\Delta N = NW\Delta y, \Delta N_t = N_t W\Delta y$, where N_t is the number of occupied traps per unit area and N is the inversion carrier density. The $\pm$ sign denotes whether the trap is neutral or charged when filled.

The ratio of the fluctuations in the carrier number to fluctuations in occupied trap number is given by a general expression [89, 88]

$$R = \frac{\delta \Delta N}{\delta \Delta N_t} = -\frac{N}{N + N^*} \tag{6.69}$$

where $N^* = \left(kT/q^2\right)\left(C_{ox} + C_{dm} + C_{it}\right)$. Typical values of N^* is $1 - 5 \times 10^{10}/cm^2$

The carrier mobility is related to the oxide trap density as follows

$$\frac{1}{\mu_s} = \frac{1}{\mu_n} + \frac{1}{\mu_{Cit}} = \frac{1}{\mu_n} + \alpha_{sc} N_t \tag{6.70}$$

where μ_n is the effective surface mobility limited by ionized impurity scattering, surface roughness scattering, and phonon scattering. The quantity $\mu_{Cit} = 1/\alpha_{sc} N_t$ is the mobility limited by Coulombic scattering of the mobile carriers at trapped charges near the Si-SiO_2 interface. The scattering coefficient α_{sc} is a function of the local carrier density due to the screening effect as well as the distance of the trap from the interface. The relationship is given as follows [195].

$$\alpha_{sc} = \frac{1}{\mu_{Co}\sqrt{N}} \tag{6.71}$$

The reduction of α_{sc} with increase of N may be physically understood as follows. As the inversion carrier density increases, the screening length and the scattering cross section due to the screening by minority carriers reduces and hence the scattering parameters decrease. In a weak inversion region, screening due to minority carriers becomes less significant compared to that by majority carriers. Because the majority carrier concentration does not change much in the weak inversion region, the scattering cross section remains almost constant with inversion carrier density. Therefore, in the weak inversion region, α_{sc} saturates to a value[195].

By differentiating (6.70) and substituting (6.69), the local drain current fluctuation is given by

$$\frac{\delta I_{DS}}{I_{DS}} = -\left[\frac{R}{N} \pm \alpha_{sc}\mu_s\right]\frac{\delta\Delta N_t}{W\Delta y} \tag{6.72}$$

The corresponding PSD is written as follows

$$S_{\Delta I_{DS}}(y,f) = \left(\frac{I_{DS}}{W\Delta y}\right)^2\left(\frac{R}{N} \pm \alpha_{sc}\mu_s\right)^2 S_{\Delta N_t}(y,f) \tag{6.73}$$

Here $S_{\Delta N_t}(y,f)$ is the PSD of the mean square fluctuations in the number of occupied traps over the area $W\Delta y$ and is given by

$$S_{\Delta N_t}(y,f) = N_t(E_{fn})\frac{kTW\Delta y}{\gamma f} \tag{6.74}$$

In (6.74), E_{fn} is the electron quasi-Fermi level, and γ is the attenuation coefficient of the electron wave function in the oxide. For the Si-SiO$_2$ system, $\gamma = 10^8/cm$. From (6.73) and (6.74), we get

$$S_{\Delta I_{DS}}(y,f) = \left(\frac{I_{DS}}{W\Delta y}\right)^2\left(\frac{R}{N} \pm \alpha_{sc}\mu_s\right)^2 N_t(E_{fn})\frac{kTW\Delta y}{\gamma f} \tag{6.75}$$

The total drain current noise PSD is given as

$$S_{\Delta I_{DS}}(f) = \frac{1}{L^2}\int_0^L S_{\Delta I_{DS}}(y,f)\Delta y dy \tag{6.76}$$

Substituting (6.75) in (6.76) and changing the variables of integration, we write

$$S_{\Delta I_{DS}}(f) = \frac{qkTI_{DS}\mu_s}{\gamma f L^2}\int_0^{V_{DS}} N_t(E_{fn})\left(1 \pm \alpha_{sc}\mu_s\frac{N}{R}\right)^2\frac{R^2}{N}dV \tag{6.77}$$

This is written in a compact way as follows

$$S_{\Delta I_{DS}}(f) = \frac{qkTI_{DS}\mu_s}{\gamma f L^2}\int_0^{V_{DS}} N_t^*(E_{fn})\frac{R^2}{N}dV \tag{6.78}$$

Here $N_t^*(E_{fn})$ is the equivalent oxide trap density that produces the same noise power in the absence of mobility fluctuations and is given as

$$N_t^*(E_{fn}) = N_t(E_{fn})\left(1 \pm \alpha_{sc}\mu_s\frac{N}{R}\right)^2 \tag{6.79}$$

In order to make the unified noise model suitable for circuit simulation purposes, the BSIM compact model approximated $N_t^*(E_{fn})$ as a three parameter function of the channel carrier density as

$$N_t^*(E_{fn}) = A + BN + CN^2 \tag{6.80}$$

where A, B and C are technology dependent model parameters. The integration variable in (6.78) is changed as follows

$$S_{\Delta I_{DS}}(f) = \frac{q^2 kTI_{DS}\mu_s}{\gamma f L^2 C_{ox}} \int_{N_S}^{N_D} N_t^*(E_{fn}) \frac{R^2}{N} dN \tag{6.81}$$

where N_S and N_D are the inversion charge density at the source end and the drain end of the channel respectively. The drain current noise PSD at the three regions of operations is written as follows [30]:

Linear Region in strong inversion
In the strong inversion region, the charge density of the carrier is written as

$$qN(y) = C_{ox}\left[V_{GS} - V_T - \alpha V(y)\right] \tag{6.82}$$

From this, we have

$$qN_S = qN(0) = C_{ox}\left[V_{GS} - V_T\right] \tag{6.83}$$
$$qN_D = C_{ox}\left[V_{GS} - V_T - \alpha V_{DS}\right] \tag{6.84}$$

where α is the bulk-charge factor. Therefore, the drain current noise PSD is written as

$$S_{\Delta I_{DS}}(f) = \Gamma\left[A \ln\left(\frac{N_S + N^*}{N_D + N^*}\right) + B\left(N_S - N_D\right)\frac{1}{2}C\left(N_s^2 - N_D^2\right)\right] \tag{6.85}$$

where

$$\Gamma = \frac{q^2 kTI_{DS}\mu_s}{\alpha\gamma f L^2 C_{ox}} \tag{6.86}$$

Saturation Region in strong inversion
The channel is divided into two parts: one part is from source $L = 0$ to $L = L_s$ (velocity saturation region) and the other part L_d is from the velocity saturation point to the drain. Accordingly the flicker noise includes two parts. The compact expression is given as follows

$$\begin{aligned} S_{\Delta I_{DS}}(f) &= \Gamma\left[A \ln\left(\frac{N_S + N^*}{N_D + N^*}\right) + B\left(N_S - N_D\right) + \frac{1}{2}C\left(N_s^2 - N_D^2\right)\right] \\ &+ \Delta L \frac{kTI_{DS}^2}{\gamma f W L^2} \frac{A + BN_D + CN_D^2}{\left(N_D + N^*\right)^2} \end{aligned} \tag{6.87}$$

where ΔL refers to the channel length reduction due to the channel length modulation phenomenon.

Weak Inversion Region
In the weak inversion region, it is reasonable to assume that $N << N^*$ and

$N_t^*(E_{fn}) = A + BN + CN^2 \approx A$. The flicker noise in the weak inversion region is written in a simplified manner as follows:

$$S_{\Delta I_{DS}}(f) = \frac{AkTI_{DS}^2}{WL\gamma f N^{*2}} \tag{6.88}$$

The variation of the flicker noise normalized PSD for n-channel MOS transistor under all operating regions as obtained from SPICE simulation results is shown in Fig. 6.22(a). In the strong inversion region, due to large numbers of inversion carriers, the screening effect is significant. As a result, the carrier scattering is less and hence the noise is less. However, as the gate bias is reduced, the number of inversion carriers reduce and the screening effect is reduced. Consequently the carrier scattering increases, and hence the flicker noise increases. In the weak inversion region, the inversion carrier concentration does not change much so that the flicker noise remains almost constant. From Fig. 6.22(b), the weak dependence of flicker noise on drain bias is observed.

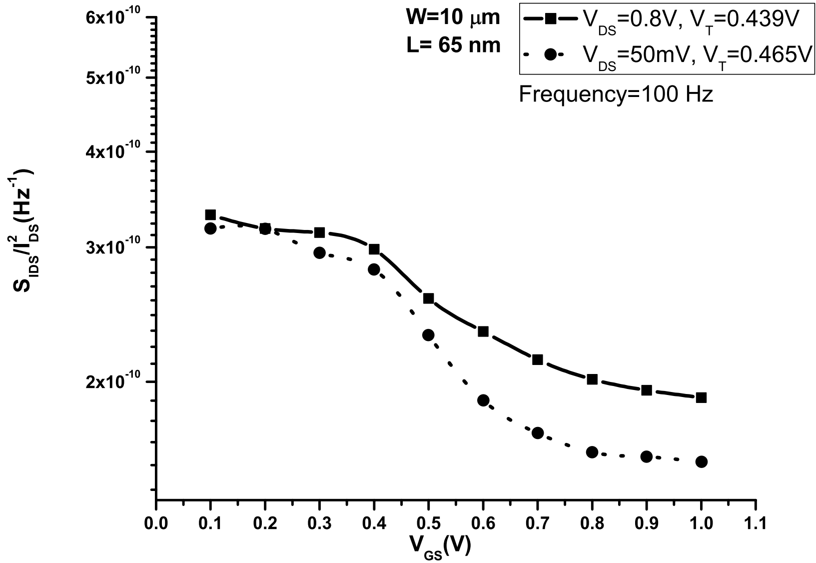
The variation of the normalized flicker noise PSD for p-channel MOS transistor under all operating regions is shown in Fig. 6.23. It may be noted that for p-channel MOS transistors, the mobility fluctuations play the dominant role in comparison to the number fluctuations in determining the noise contributions. However, the BSIM model does not properly explain the flicker noise of p-channel MOS transistor especially in the weak inversion region.

Another interesting observation is that the flicker noise in the p-channel MOS transistor is less than that of the n-channel MOS transistor by at least an order of magnitude.

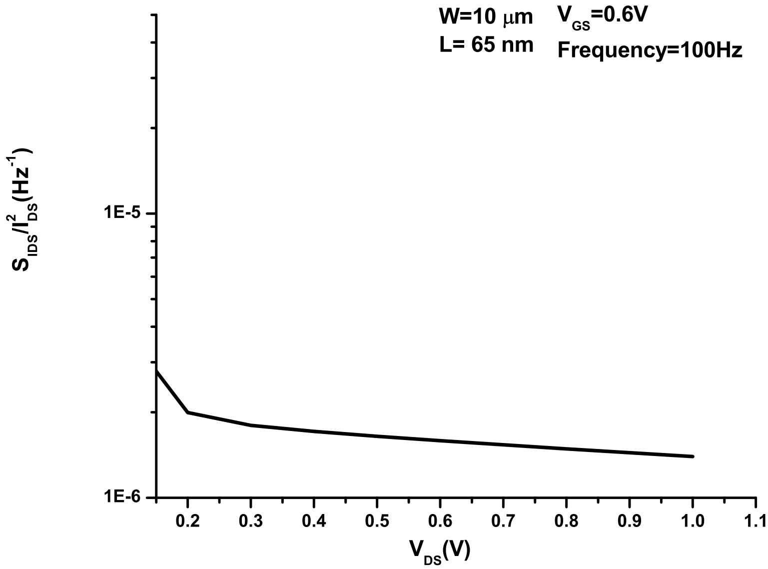
In Fig. 6.24, the $1/f$ noise is shown for different gate lengths. The channel width is fixed at $10\mu m$ and the gate lengths are varied. It is observed that the $1/f$ noise power in the transistor is decreased as the gate length increases.

6.8 Gate Resistance and Substrate Network Model of MOS Transistor for RF Applications

With the continuous scaling of CMOS technology, MOS transistors have become attractive candidates for radio frequency (RF) applications, because of low cost, high integration, and easy access to the technology [202]. It may be noted that any frequency lying between 30KHz-300GHz is considered to belong to the RF range. The compact models for MOS transistors were originally developed for digital and low frequency analog circuits. These models focus on DC drain current, conductance, and intrinsic charge/ capacitance behavior up to the megahertz range. However, with the increase of the operating frequency to GHz range, the extrinsic components of a MOS transistor become equally significant compared to the intrinsic core components of the



(a) variation with gate bias



(b) variation with drain bias

FIGURE 6.22
Variation of normalized flicker noise PSD for n-channel MOS transistor with gate bias and drain bias.

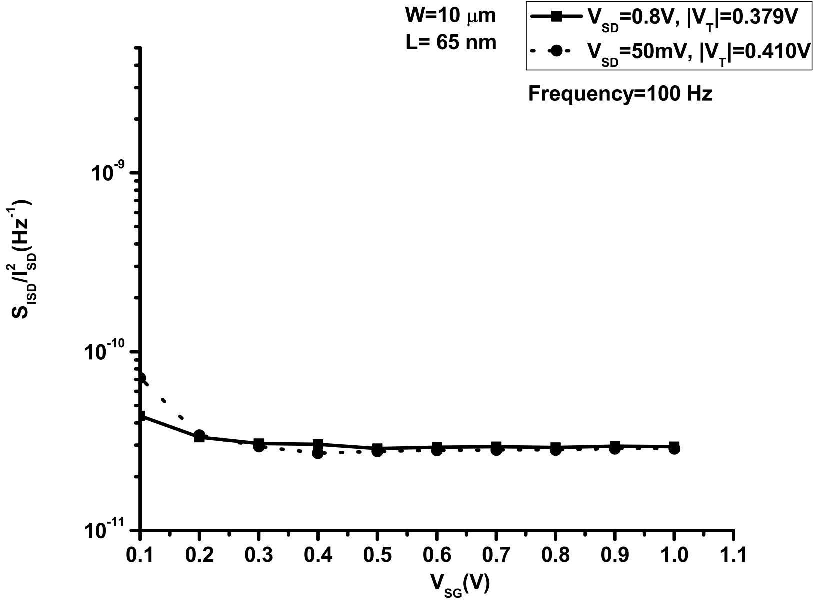


FIGURE 6.23
Variation of normalized flicker noise PSD for p-channel MOS transistor with gate bias.

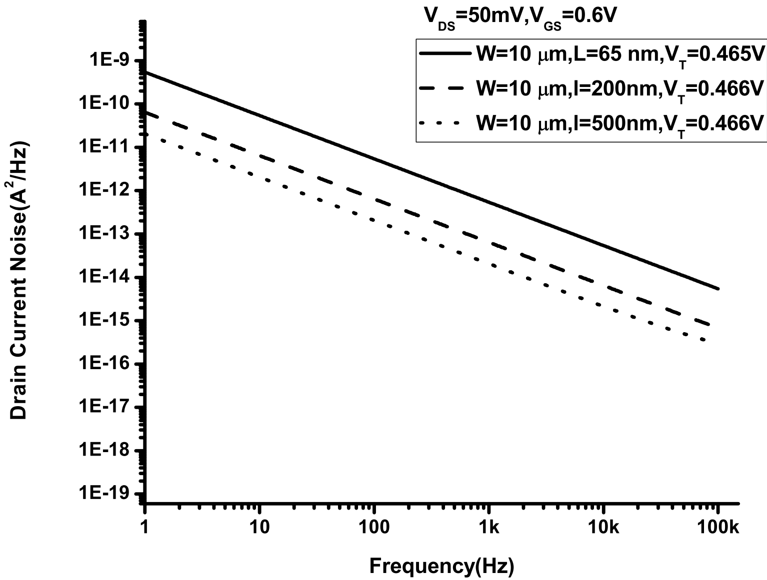


FIGURE 6.24
Flicker noise spectrum of n-channel MOS transistor with gate length as parameter.

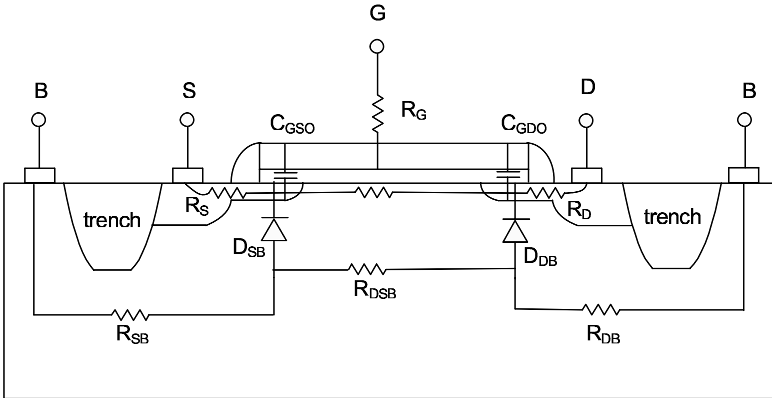


FIGURE 6.25
Schematic diagram of a MOS transistor illustrating the parasitic components.

MOS transistor. Therefore, there are several additional requirements for MOS models for RF circuit design. Some major additional requirements are [29] (1) the model should include the non-quasistatic (NQS) effect, (2) efficient modeling of gate resistance, (3) extrinsic source/drain resistances should be modeled as real resistors, and (4) the substrate coupling in a MOS transistor needs to be modeled physically.

6.8.1 Parasitic Components of MOS Transistors

The various parasitic components of a MOS transistor are identified in Fig. 6.25 [29]. The parasitic components of a MOS transistor are (1) gate resistance R_G, (2) gate-to-source/drain overlap capacitors C_{GSO} and C_{GDO}, (3) source series resistance R_S and drain series resistance R_D, (4) source-to-bulk junction diode D_{SB} and drain-to-bulk junction diode D_{DB}, and (5) substrate resistances R_{SB}, R_{DB} and R_{DSB}. The source and the drain series resistors are added outside the intrinsic MOS model. This is because the internal series resistors, discussed in Chapter 3 of the text, are used to calculate the drain current considering the DC voltage drop across the resistors, but they do not contribute anything in AC simulation in the sense that they do not add any poles to the system transfer function. The gate resistance, although not part of the compact model, plays a significant role in RF circuits. The substrate resistors are added to account for the signal coupling through the substrate [57].

The effects of these parasitic components are usually not profound in low frequency analog circuit design, but play significant roles in determining the high frequency circuit performances.

6.8.2 Gate Resistance Modeling

The gate resistance influences the input matching to achieve maximum power transfer and noise performances due to the thermal noise introduced by the gate resistance. Therefore, it is essential to characterize the gate resistance of a MOS transistor accurately for RF IC design [29, 94].

At DC and low frequency, the gate resistance mainly consists of the polysilicon sheet resistance. The resistance is thus given as

$$R_G = \frac{W}{L} R_{Gsh} \tag{6.89}$$

where R_{Gsh} is the gate sheet resistance per square of the material, W is the effective channel width, and L is the effective channel length. The typical value of the sheet resistance of a polysilicon gate material is $20 - 40\Omega/\Box$. This can be reduced by a factor of 10 with a silicide process. At high frequency, two additional physical effects come into action: (1) distributed transmission line effect on the gate and (2) distributed or NQS effect in the channel.

The distributed nature of the gate is shown in Fig. 6.26(a). Because of the distributed nature, the actual gate-to-source voltage V_{GS} decreases as z increases because of the finite voltage drop across the gate material. Fortunately, as demonstrated in [117], it is possible to model an equivalent lump resistance of magnitude

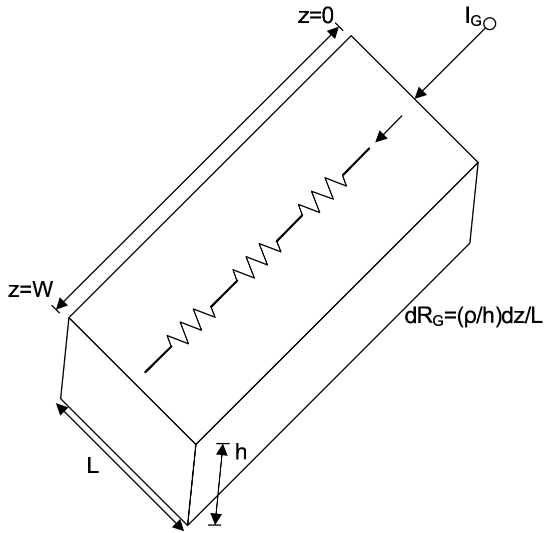
$$R_{Gpoly} = \frac{1}{3} \frac{W}{L} R_{Gsh} \tag{6.90}$$

The factor 3 accounts for the distributed RC effects when the gate electrode is contacted at only one end. This factor is found to be 12 when the electrode is contacted at both ends. The gate resistance model considering the distributed transmission line effect is written as [29, 94]

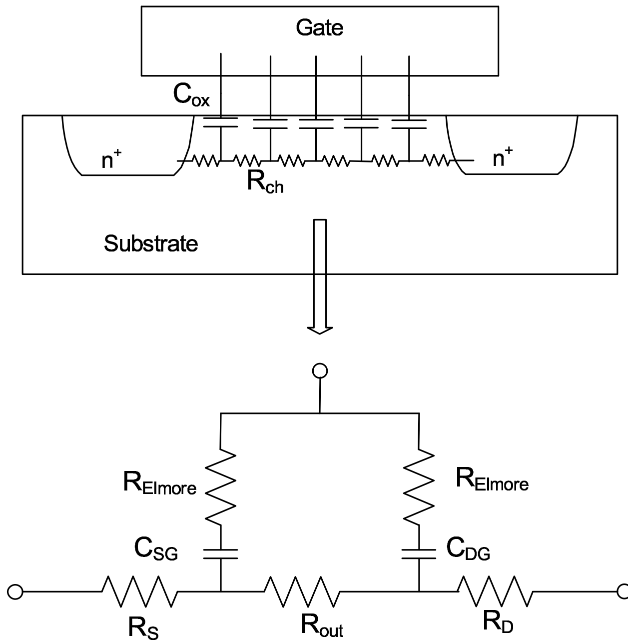
$$R_{Gpoly} = R_{Gsh} \left(\alpha W/L + W_{ext} \right) \tag{6.91}$$

where α is either 1/3 or 1/12 and W_{ext} is the extension of the polysilicon gate over the active region.

In addition to the above, the distributed RC effect of the channel, i.e., the non-quasi-static effect of the channel, needs to be considered for characterizing the RF behavior of a MOS transistor. Under normal quasi-static operation it is assumed that the potential and charge density at any given point in the channel follows the applied bias voltage without any delay [1]. This is true when the rise/fall time of the voltage change is greater than the transit time of the carriers from source to drain. The channel charge is assumed to achieve equilibrium once biases are applied, thus the finite charging time of the carriers in the inversion layer is ignored. However, if the frequency of the applied signal is very high and the rise/fall time becomes less than the transit time of the carriers through the channel, the assumption that the channel charge is only a function of terminal voltages does not hold good [1]. The channel of a MOSFET can be viewed as a bias dependent RC distributed transmission



(a) Distributed gate resistance



(b) Distributed channel resistance

FIGURE 6.26
Schematic diagram illustrating the distributed nature of gate and channel resistance.

line as shown in Fig. 6.26(b). Utilizing Elmore's approach, the RC distributed channel can be approximated by a simple RC equivalent, which retains the lowest frequency pole of the original RC network. The Elmore resistance in strong inversion region is given by [23]

$$R_{\text{Elmore}} = \frac{L}{E\mu_s W Q_n} \tag{6.92}$$

where $Q_n = C_{ox}(V_{GS} - V_T)$ is the inversion charge density and E is the Elmore constant to match the lowest frequency pole. The value of this parameter is found to be nearly 3 and is invariant with respect to W and L. However, in BSIM3 model, this value if chosen to be 5 for better accuracy. The overall channel resistance is given as

$$R_{Gchan} = \gamma \left(\frac{1}{R_{st}} + \frac{1}{R_{\text{Elmore}}} \right) \tag{6.93}$$

where R_{st} is the static channel resistance and is given by

$$\begin{aligned} R_{st} &= \int dV/I_{DS} && (6.94) \\ &= V_{DS}/I_{DS} \quad \text{linear region} && (6.95) \\ &= V_{DSsat}/I_{DS} \;\; \text{saturation region} && (6.96) \end{aligned}$$

and γ is a fitting parameter.

The effective gate resistance is thus given by [29, 94]

$$R_G = R_{Gpoly} + R_{Gchan} \tag{6.97}$$

In order to extract R_G, two port S-parameters are converted to Y-parameters and the input resistance is given by

$$R_{in} = \text{real}\,(1/Y_{11}) \tag{6.98}$$

where the gate is connected to Port 1 and the drain is connected to Port 2. The extracted resistance value includes the values of R_G as well as the source/drain resistance R_S/R_D. Knowing the values of the latter from DC measurements, the values of R_G can be extracted from R_{in}.

6.8.2.1 Minimization of Gate Resistance

The distributive channel resistance component of the gate resistance is a fundamental component. On the other hand, it is possible to reduce the distributed gate component in an effective way through multi-finger layout. The concept is illustrated in Fig. 6.27[85]. A MOS transistor of large width, say $10\mu m$ can be designed by connecting 10 MOS transistors in parallel, each having a width of $1\mu m$. Therefore, the gate electrode resistance is reduced by a factor of 100, because the resistance of each small transistor is 10 times

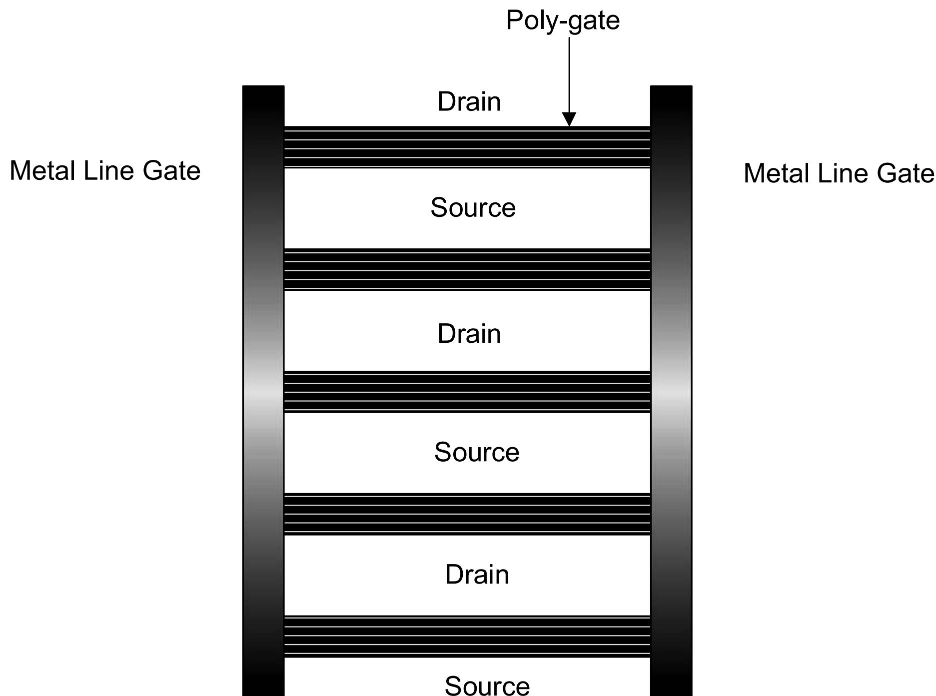


FIGURE 6.27
Illustration of the multi-finger layout concept.

smaller. With the concept of multi-finger the distributed gate resistance component is written as

$$R_{Gpoly} = \frac{R_{Gsh}}{N_F} \left(\alpha W/L + W_{ext}\right) \tag{6.99}$$

where N_F is the number of fingers. Using the multi-finger layout and metal as the gate material, the distributed gate component can be minimized to a large extent, and it can be neglected.

6.8.3 Substrate Network Modeling

The influence of the substrate resistance is non-negligible for RF IC design. This is because the signal at the drain couples to the source and bulk terminals through the source/drain junction capacitance and the substrate resistance [193, 29]. The substrate resistance primarily influences the output characteristics. It is desirable to consider the distributed nature of the substrate resistances. However, it makes the models too complex to be considered for circuit simulation purposes. Therefore, a lumped RC network model, accurate up to the frequency of operation, is the preferred approach. A simple

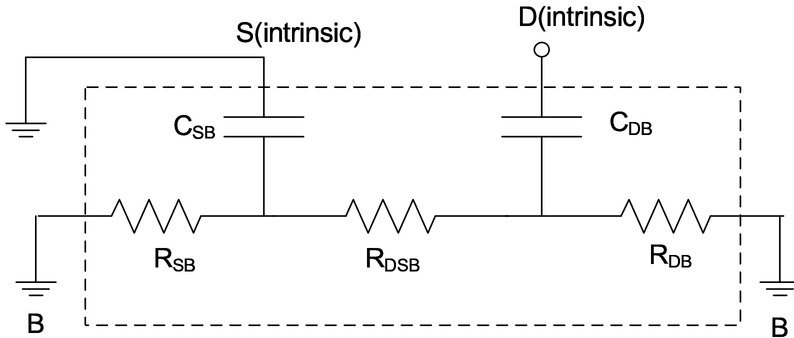


FIGURE 6.28
Equivalent circuit of the substrate network.

equivalent circuit for the substrate network is shown in Fig. 6.28. The various resistor components in the equivalent circuit may be computed as follows

$$R_{DSB} = \frac{R_{DSBsh}L}{N_F W} \tag{6.100}$$

$$R_{DB} \approx \frac{r_{dbw}}{W} \tag{6.101}$$

$$R_{SB} \approx \frac{r_{sbw}}{W} \tag{6.102}$$

where R_{DSBsh} is the sheet resistance in the substrate between the source and drain, and r_{dbw}, r_{sbw} are the substrate resistances per unit channel width.

The gate and the substrate resistance model can be added with the intrinsic MOS transistor and implemented in any circuit simulation framework. The sub-circuit approach presented in this section is found to be quite useful and accurate for RF IC design at least up to 10GHz.

6.9 Summary and Conclusion

This chapter presents some advanced issues related to nano-scale MOS transistors which have significant effects in the circuit performances. With the scaling down of the aspect ratio, the narrow channel effect is gaining importance day by day. The effect on the threshold voltage can no longer be neglected. The degradation of the circuit performances due to scaling of the MOS transistors can be mitigated to a large extent at the device-level abstraction through channel and gate engineering in an elegant way. The vertical and lateral channel engineered devices are now gaining importance for circuit design. This chapter presents a comprehensive treatment on this issue. With the scaling of

oxide thickness, the gate leakage current no longer becomes negligible. Leakage current now plays a critical role in determining the circuit performances, by limiting the intrinsic current gain. The conventional representation of the gate of the transistor as a capacitor is no longer valid up to a certain input signal frequency. The basic mechanisms of the gate leakage current are discussed in this chapter, along with the effects of the leakage current in a comprehensive manner. The use of high-κ dielectric material as the gate material now becomes almost compulsory. The use of metals instead of the doped polysilicon is also becoming important. The noise of MOS transistors is another important issue which plays a critical role in determining the noise performance of the total circuit. This chapter presents a detailed discussion about the modeling and characterization of the two important noise sources, i.e., the thermal noise and the flicker noise. Finally, the effects of various parasitic components such as the gate resistance and the substrate resistance of MOS transistors are also discussed in the context of RF IC design.

7

Process Variability and Reliability of Nano-Scale CMOS Analog Circuits

7.1 Introduction

With the scaling of CMOS technology to the sub-90nm domain, yield and reliability of integrated circuits become increasing challenges to the designers as far as the design productivity and the design creativity gap is considered. Smaller devices combined with new materials are the cause of the increasing yield and reliability problems. The yield of a circuit is defined by the ratio of the number of fabricated circuits which meet the design specifications out of the total number of fabricated circuits, and is expressed in percentage. Reliability is defined as the ability of a circuit to conform to its true characteristics over a specified period of time under specified conditions. The two primary causes of nonideal behavior of analog circuits are (1) imperfections in the fabrication process which includes both random and systematic errors and (2) complicated physical behavior of MOS transistors in nano-scale regime under various operating conditions. The degradation of circuit performances is found to be both time independent and time dependent. The time independent variations of circuit performances are due to systematic and random variations of manufacturing process related parameters. On the other hand, the time dependent variations of circuit performance depend on the stress applied to the device, i.e., the voltages and currents applied to the transistor.

This chapter presents an introductory overview of the various physical causes of process variations and reliability issues on nano-scale analog circuits. The effects on the circuit performances and the approaches for computer-aided simulation of these are also discussed. This chapter therefore, attempts to make the designers aware of the two most critical challenges of nano-scale analog circuit design.

7.2 Basic Concepts on Yield and Reliability

7.2.1 Yield

The drift of process parameters causes transistors to have characteristics different from those desired. These parameter drifts are caused by inherent practical and physical limitations associated with manufacturing steps such as photolithography, etching, diffusion, etc. Failures due to parameter variations are termed as soft faults. An estimate of the yield of good dice can be found from probability theory in which n defects are randomly placed in a wafer containing N die sites. The probability P_k that a given die site contains exactly k defects is given by the binomial distribution

$$P_k = \frac{n!}{k!\,(n-k)!} N^{-n} \left(N-1\right)^{n-k} \tag{7.1}$$

For large n and N, this is approximated by the Poisson distribution

$$P_k = \frac{\lambda^k}{k!} \exp\left(-\lambda\right) \tag{7.2}$$

where $\lambda = n/N$ is the average number of defects per die. The yield is given by the probability that a die is found with no defects

$$Y = P_0 = \exp(-\lambda) \tag{7.3}$$

If A is the area of one die, then the area of the wafer is NA and D_0 is the defect density, then $D_0 = n/NA$. Therefore,

$$\lambda = \frac{n}{N} = D_0 A \tag{7.4}$$

The yield based on the Poisson distribution is thus [134]

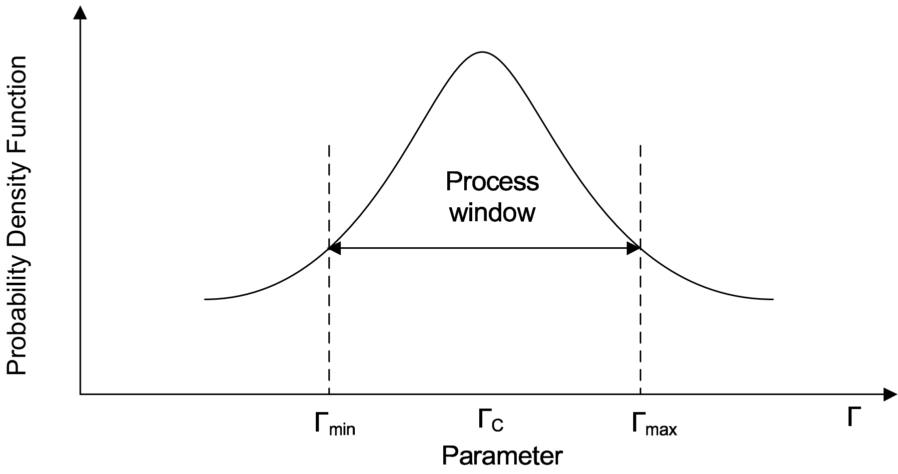
$$Y = \exp\left(-D_0 A\right) \tag{7.5}$$

The defect density is typically in the range of 1 to $2/cm^2$. For large dice with $D_0 A > 1$, the yield predicted by the above model is somewhat pessimistic. The two other commonly used models are the Seed's model and the Murphy model which are as follows [134]

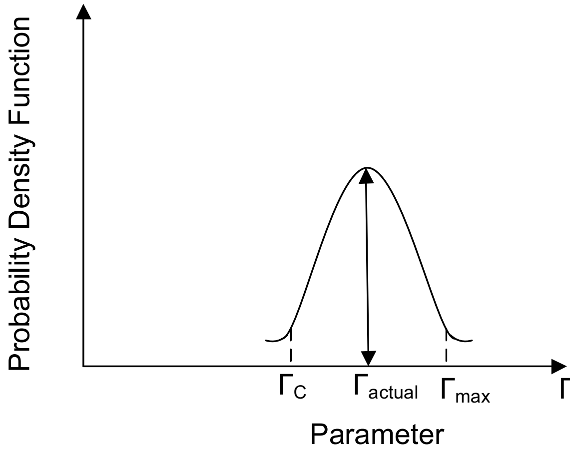
$$Y = \exp\left(-\sqrt{AD_0}\right) \tag{7.6}$$

$$Y = \left(\frac{1-\exp(-AD_0)}{AD_0}\right)^2 \tag{7.7}$$

The soft faults effect yield in a statistical sense and play a major role in determining the yield of analog circuits. The statistical distribution of a process



(a) Process window



(b) Wafer level variation

FIGURE 7.1
Statistical spreading of process parameter Γ.

parameter Γ measured over repeated processing runs is shown in Fig. 7.1(a) and Fig. 7.1(b). In Fig. 7.1(a), Γ_{min} and Γ_{max} define a process window within which the average value of the parameter lies. It is obvious that the cost of fabricating an IC increases with a decreasing process window. The statistical

variation of the process parameter across the wafer is shown in Fig. 7.1(b). Γ_{actual} is the average value of the parameter Γ on a given wafer.

It is observed that the statistical variance at the wafer level is generally much smaller than the variance at the processing level. The distribution function is usually found to be, or can be approximated to be a normal distribution and is mathematically modeled as [71]

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-m)^2}{2\sigma^2}} \tag{7.8}$$

where m and σ are the mean and standard deviation, respectively. The probability of a random variable x lying between two points x_1 and x_2 is given by the area under the normal curve from x_1 and x_2, i.e., $P(x_1 \leq x \leq x_2)$ and is given by

$$\begin{aligned} P &= \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{(x-m)^2}{2\sigma^2}} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-z^2/2} dz \end{aligned} \tag{7.9}$$

where $z = (x-m)/\sigma$ is called a normal variate. This can be written as

$$P = \frac{1}{\sqrt{2\pi}} \left[\int_0^{z_2} e^{-z^2/2} dz - \int_0^{z_1} e^{-z^2/2} dz \right] = P_2(z) - P_1(z) \tag{7.10}$$

The values of the integral are readily obtainable from tables of the normal probability distribution, which appear in most engineering mathematics textbooks [71]. Using this table it can be shown that (see Fig. 7.2.)

1. The area under the normal curve between $x = m - \sigma$ to $x = m + \sigma$ is nearly 0.6826 $\sim$ 68%. Thus approximately 2/3 of the samples lie within these limits.
2. The area under the normal curve between $x = m - 2\sigma$ to $x = m + 2\sigma$ is nearly 0.9544$\sim$ 95.5%.
3. The area under the normal curve between $x = m - 3\sigma$ to $x = m + 3\sigma$ is nearly 0.9973$\sim$ 99.73%.
4. The area under the normal curve between $x = m - 6\sigma$ to $x = m + 6\sigma$ is nearly 0.999999998$\sim$ 99.9999998%.

7.2.2 Design Tolerance and Capability Index

It is now obvious that it is impossible to get a design which has zero performance variations. Therefore, the designers need to establish specifications that define not only the target value of something but also acceptable limits about

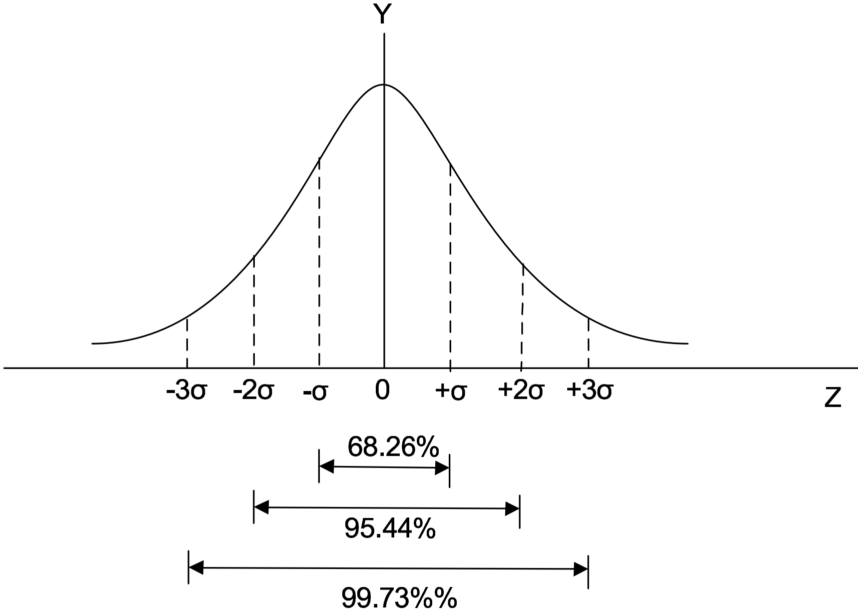


FIGURE 7.2
Area under normal distribution.

the target. These design limits are often referred to as the upper and lower specification limits or the upper and lower tolerance limits. Due to the various sources of process variations, to be discussed later, the performance parameters of a design are distributed around the statistical mean, the distribution in most cases is normal (or can be approximated to be normal). It is important for the designers to be aware of a relationship among the distributed values of the various performance parameters and the design tolerance range.

The potential capability index is defined as [2]

$$C_p = \frac{y_{UTL} - y_{LTL}}{6\sigma} \tag{7.11}$$

where y_{LTL} is the lower tolerance limit and y_{UTL} is the upper tolerance limit and the design variation width is defined by 6σ limit of the distributed performance values. This is shown in Fig. 7.3. In an idealized normal distribution case, a $C_p = 1$ indicates that the widths of the specification limits is the same as the $\pm 6\sigma$ width of the performance parameter variations. The value of the parameter $C_p < 1$ indicates that the design is not at all good as robustness of the design is concerned, $C_p = 1$ means OK. On the other hand, $C_p = 2$ represents that the design is very good. Therefore, a design for which $C_p = 2$ should be the desired goal for a designer.

However, the parameter C_p alone is not sufficient to estimate the robustness of a design. This is illustrated in Fig. 7.4. Another metric referred to as

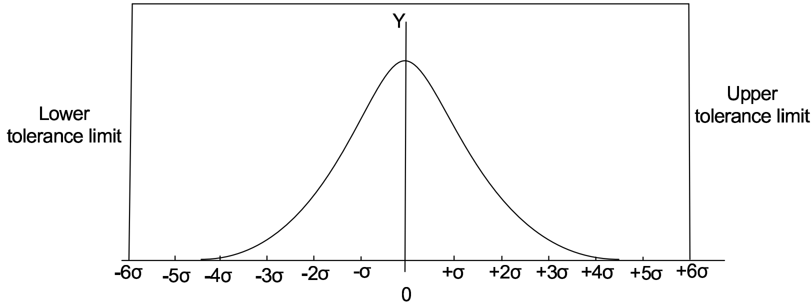


FIGURE 7.3
Schematic illustration of potential capability index.

the process capability index which signifies the position of the mean and tails of the design distribution relative to design specification is therefore, required. This is mathematically defined as [2]

$$C_{pk} = \min\left[\frac{\bar{y} - y_{LTL}}{3\sigma}, \frac{y_{UTL} - \bar{y}}{3\sigma}\right] \quad (7.12)$$

where $\bar{y}$ is the statistical mean of the distributed values of the performance parameter y, y_{LTL} is the lower tolerance limit, and y_{UTL} is the upper tolerance limit and the design variation width is defined by 3σ limit of the distributed performance values. The C_p parameter signifies the smartness of the design and C_{pk} signifies the positioning of the design within the specification limits. This leads to the concept of design centering, which is illustrated in Fig. 7.5 for a hypothetical design space. The basic idea of the design centering is to place the synthesized design at the center of the acceptable specification space, and that with the variation of the design parameters due to process error will not be able to push the design out of the acceptable specification space.

The probability that the design distribution lies inside the design specification window is given by

$$P = \frac{1}{\sqrt{2\pi}} \int_{-3C_p}^{3C_P} e^{\left(-\frac{z^2}{2}\right)} dz \quad (7.13)$$

In order to design a robust circuit, the designer may need to optimize a cost function consisting of both C_p and C_{pk} in some suitable format [2].

7.2.3 Reliability Bathtub Curve

Reliability is measured by the probability that a device will perform its required function under stated conditions for a specific period of time. The lifetime of an integrated circuit is divided into three distinct phases, which are shown in the famous reliability bathtub curve, shown in Fig. 7.6.

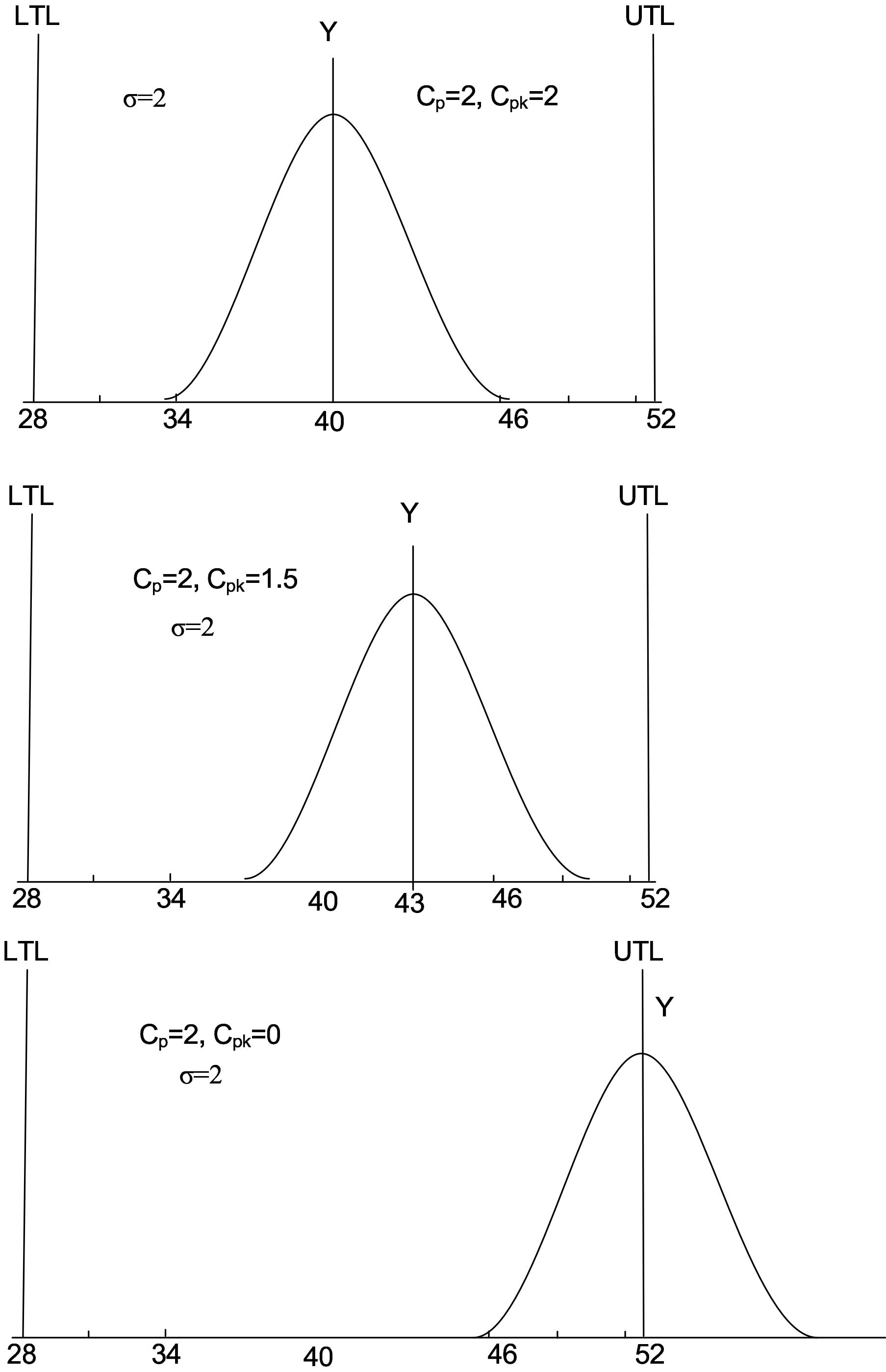


FIGURE 7.4
Schematic illustration of process capability index.

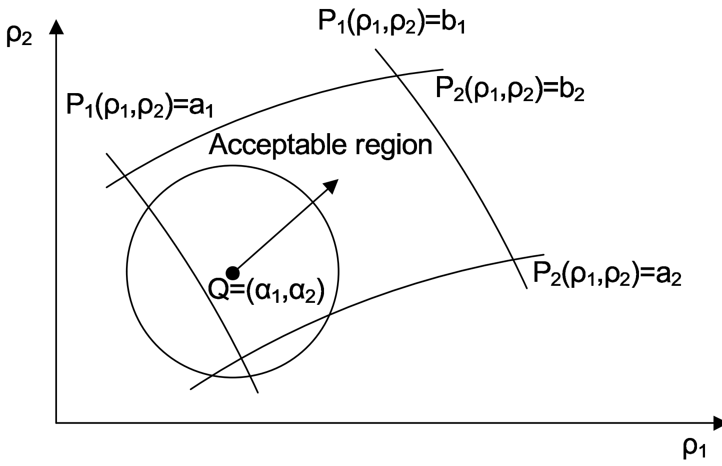


FIGURE 7.5
Schematic illustration of the concept of design centering.

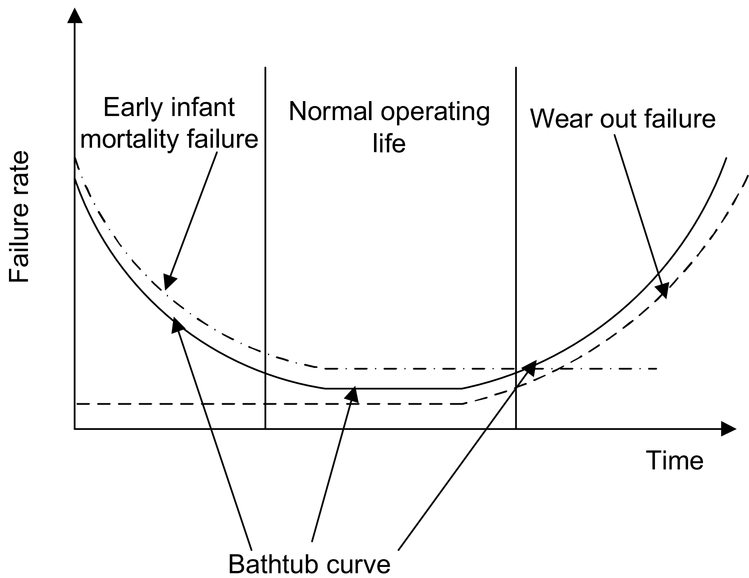


FIGURE 7.6
Reliability bathtub curve.

The first phase is referred to as the infant mortality rate. This is the phase during which the circuits which are manufactured with the extreme values in the tolerance region tend to fail very early in their operation. The designs which are not centered in the acceptable specification space therefore have a high probability of failure at this stage. This type of failure can be reduced either by improving the manufacturing process technology or by proper design centering. The first option is not in the hands of the designers, so that the second option is the preferred one. Another common practice done by the manufacturers is to identify the set of mal-manufactured devices and throw them away. All the manufactured circuits are subjected to elevated operating conditions for a short time to induce accelerated stress. With this the circuits which are close to the periphery of tolerance region eventually fail and are thrown away. Although this reduces the yield, it saves the manufacturer from the embarrassment of failure in the field during the promised warranty period.

The second phase is the normal operating life. This is the period in which the circuit operates per the specifications. The failure rate is low and remains fairly constant. The failure primarily occurs due to corner case operations that are not taken care of during the design, soft errors due to radiation, and exceeding the allowed operating conditions. However, the latter is of no concern to the designers and manufacturer since the operating conditions are in most cases detailed in the specification data sheet.

The third phase is referred to as the wear out phase. As the circuits begin to fatigue or wear out, failures occur at increasing rates. This is a natural phenomenon in all the semiconductor materials under stress. The various sources of wear out failures in integrated circuits are discussed in this chapter. The designer does not have any control over them except to employ safety margins in the design which may extend the normal operating life of the circuits.

It is interesting to note that the reliability curve of any human being is identical to that explained above. The mortality rate is higher for children, becomes constant at young ages and again increases when a person becomes old. The design specifications are the various disciplines that need to be followed in life for a normal and long healthy life.

7.3 Sources of Variations in Nanometer Scale Technology

The fundamental difference between nanometer-scale circuits and those built in their predecessor technologies is that the nano-scale circuits are subject to a wide range of new effects that induce on-chip variations. Such variations cause significant deviations from the prescribed specifications for a chip. The various sources of variations that critically affect the performances of a chip are broadly classified into three types [165]. These are (1) process variations,

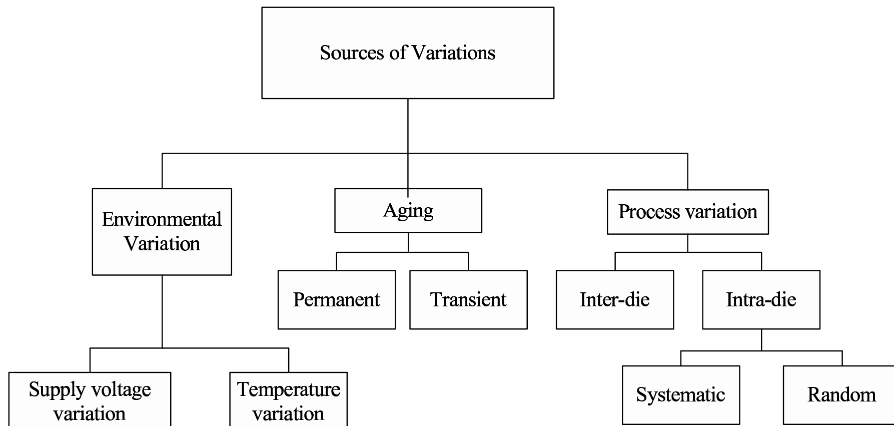


FIGURE 7.7
Classification of the various sources of variations.

(2) environmental variations, and (3) aging. These are in turn sub-divided into some classes as shown in Fig. 7.7.

7.3.1 Process Variations

The modern IC manufacturing process typically requires many steps. Despite the advances in fabrication techniques, there still exist systematic or random errors during each fabrication step. Process variations thus occur when a circuit is manufactured, and cause process parameters to drift from their designed values. From a circuit design perspective the parametric process variations are typically divided into two major groups which are [104]: (1) inter-die variations or global variations and (2) intra-die variations or local variations. These are shown in Fig. 7.8. The inter-die variations are characterized by lot-to-lot, wafer-to-wafer, or die-to-die fluctuations in the process. Such variations affect all transistors in a given circuit equally. On the other hand, the intra-die variations correspond to variations within a single die. Such variations may affect different transistors differently on the same circuit. The inter-die variations are systematic in nature and hence cause movement of the statistical mean. On the other hand, the intra-die variations are generally random in nature and cause variations around the statistical mean, as shown in Fig. 7.9. The inter-die variation is generally much larger than the intra-die variation. Systematic variations can be handled through layout design and more controlled resolution enhancement techniques (RETs). However, handling the effects of random variations requires innovative process and circuit design techniques and device modeling. Therefore, in nanometer technologies, random intra-die variations have become significant and can no longer be ignored [163].

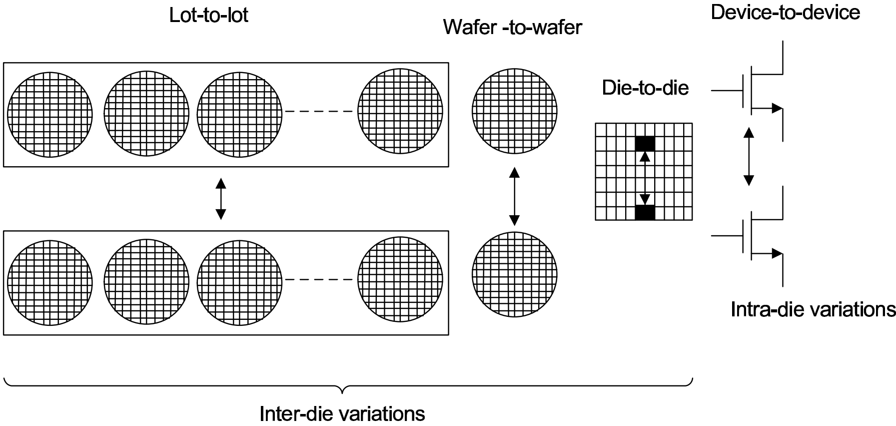


FIGURE 7.8
Inter-die and intra-die process variations.

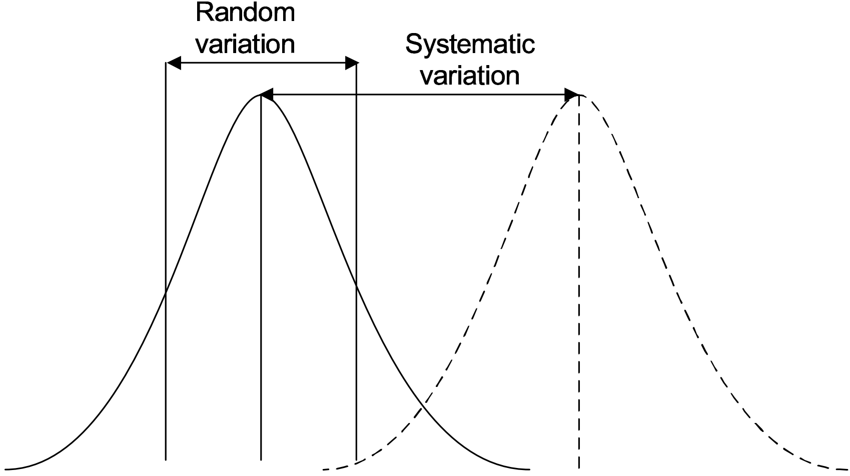


FIGURE 7.9
Systematic and random variations.

7.3.2 Environmental Variations

Environmental variations correspond to the changes during the operation of a circuit. These mainly include variation in the supply voltage (V_{DD}) and the temperature variation.

With the scaling of the technology node, the power supply level has reduced which in turn has decreased the tolerance to voltage changes within the power distribution networks in CMOS integrated circuits. Run-time fluctuations in the supply voltage levels in a chip cause significant variations in parameters such as gate delay. A small fluctuation in the gate-delay value may even result in the logic failure of the whole circuit.

The impact of temperature on the functioning of a chip is an important factor in inducing variation. It causes transistor threshold voltages to go down, and carrier mobilities to decrease [192]. Threshold voltage decrease tends to speed up a circuit, while mobility reduction slows it down. Depending on which effect wins, a circuit may show either negative or positive or mixed temperature dependence if the trend is nonuniform. Leakage power also increases with temperature [154]. If this increase is substantial, the increased power can raise the temperature further, causing a feedback cycle. This can even cause thermal runaway, where the increase in the power goes to a point that cannot be supported by the heat sink, and the chip burns out.

7.3.3 Aging Variations or Reliability

Reliability is an important issue in VLSI circuits. The reliability of a design is often degraded by various causes ranging from soft errors, electromigration, hot carrier injection, negative bias temperature instability, crosstalk, power supply noise, and variations to the physical design. With the continual scaling down of circuit designs, the issues pertaining to reliability have a greater impact within the design. Given this problem along with the demand for high-performance designs, chip designers are faced with the objective to design reliable circuits with high yield, high performance and energy-efficiency.

7.4 Systematic Process Variations

As the semiconductor process technology scales into nanometer dimension, the printability and process window of the finer lithographic patterns are significantly reduced due to the fundamental limit of the microlithography systems. As for now, leading IC fabs still use the 193nm lithography systems to print sub-wavelength feature size (e.g., 65nm or even 45nm), with the aid of various sophisticated resolution enhancement techniques (RET), such as optical proximity correction (OPC), phase shift mask (PSM), etc. [31]. However, the

complex steps of these lithographic techniques increase the risk of process variations even further. These process variations are found to be systematic in nature. Apart from the RETs, layout induced strain, well-proximity effect, etc., are some other sources of systematic process variations. The systematic sources of process variation can be mitigated either by more controlled RETs or can be modeled and subsequently incorporated in the design. This section introduces some major sources of systematic process variations.

7.4.1 Optical Proximity Correction

While the layout of an integrated circuit is transferred from the mask to the semiconductor by different imaging mechanisms, it may get distorted. This happens primarily due to the limitations of the light used to maintain the edge placement integrity of the original design, after processing, into the etched image on the silicon wafer. The projected images therefore, appear with irregularities such as line widths that are narrower or wider than designed; the sharp corners appear to be rounded, etc. The sharp features are lost because of the fact that the higher spatial frequencies are lost due to diffraction phenomenon. Optical proximity correction (OPC) is the technique of pre-distorting the mask patterns such that the printed patterns are as close to the desired shapes as possible [31]. The pre-distortion is achieved by moving edges or adding extra polygons to the pattern written on the photomask. The method of adding extra polygons is illustrated in Fig. 7.10. The amount of pre-distortion, i.e., the correction required is determined by different numerical techniques. There are mainly two approaches: rule-based and model-based correction, among which, in practical cases it is found that the model-based OPC is preferable [31]. However, the conventional model-based OPC assumes nominal process parameters. The OPC technique, being a photomask technique, inevitably suffers from process variations due to variations in focus and dosage. Therefore the correction technique has to be even more sophisticated.

7.4.2 Phase Shift Mask

Application of photo shift mask (PSM) in photolithography is another strong resolution enhancement technique [111, 31]. The photo-mask which is used in the photolithography technique consists of some opaque and transparent spaces of the mask. The imaging degrades because light from clear areas on the mask is diffracted into regions that ideally would be completely dark. The nominally dark region gets light diffracted into it from space on both the left and right. The idea behind PSM is to modify the mask so that alternating clear regions also cause the light to be phase shifted by 180°. The phase shifting mask consists of a normal transmission mask that has been coated with a transparent layer patterned to ensure that the optical phases of the nearest aperture are opposite. As a consequence, the diffracted light in the nominally dark area from the clear area to the left will interfere destructively with the

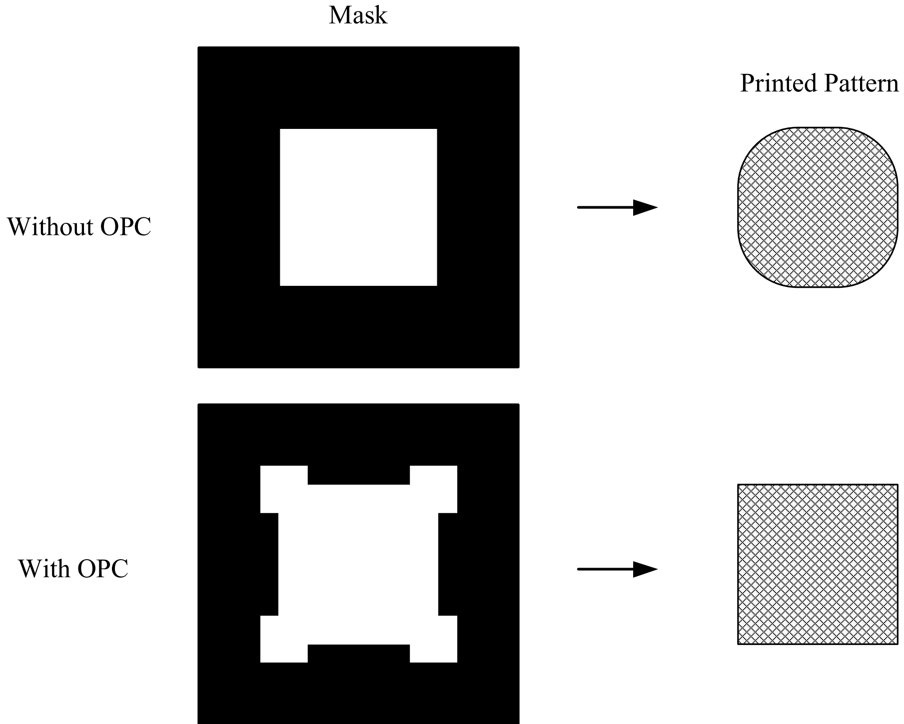


FIGURE 7.10
Comparison of the patterns obtained with and without OPC.

light diffracted from the right clear area, improving the image contrast (see Fig. 7.11). For best contrast the phase shift is needed to be accurately achieved by proper thickness of the phase shifter. However, this RET also brings in process variation through its involved lithographic steps because the steps are highly process sensitive. Due to rigorous diffraction effects the intensities of the lines deviate from each other causing $0/\pi$ CD (critical dimension) variation in the image. In addition to that, deviation from precise control over the phase determination introduces some variation in the desired design. Better controlled PSM can only reduce the process induced variation of the designed parameters.

7.4.3 Layout-Induced Strain

The performance of devices with identical geometry varies not only because of the lithography related geometry variations but also because of the layout determined strain variations due to different spacing between the devices, different distances to the shallow trench isolation, and different numbers and positions of contacts. The STI is the preferred isolation technique over the

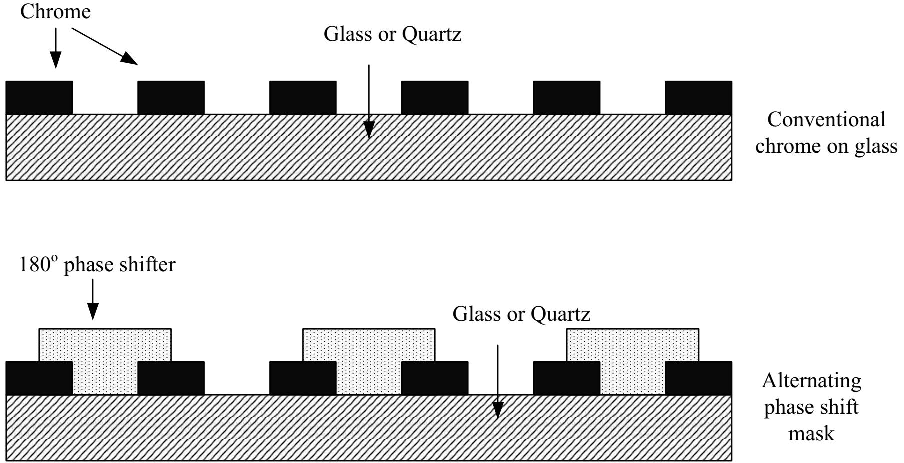


FIGURE 7.11
Use of alternating phase shift mask.

traditional LOCOS technique for the sub-0.5 m technology. But STI induces mechanical stress in the device which affects the device performance [51]. The stress gets transferred to the device channel and in fact, enhances carrier mobility for p-channel transistors but unfortunately reduces mobility for n-channel transistors. As a result the device performances suffer from variations from its expected nature. The effect of stress being well predictable can be simulated and modeled and hence can be incorporated in the design.

7.4.4 Well Proximity Effect

Highly scaled bulk CMOS technologies make use of high energy implants to form the deep retrograde well profiles needed for latch-up protection and suppression of lateral punch-through. During the implant process, some atoms can scatter laterally from the edge of the photoresist mask and become embedded in the silicon surface in the vicinity of the well edge [79], as illustrated in Fig. 7.12. The gate-to-well-edge distance SC determines the number of ions scattered into the channel region and controls the performance of the device. The result is a well surface concentration that changes with lateral distance from the mask edge, over the range of $1\mu m$ or more. This lateral non-uniformity in well doping causes the MOSFET threshold voltages and other electrical properties to vary with the distance of the transistor to the well-edge SC. This phenomenon is commonly known as the well proximity effect (WPE). However, the effect of WPE can be mitigated by increasing the distance from the gate edge to the well edge. When the distance from the gate to well edge is greater than $2\mu m$, the WPE effect is very small. But if the corresponding

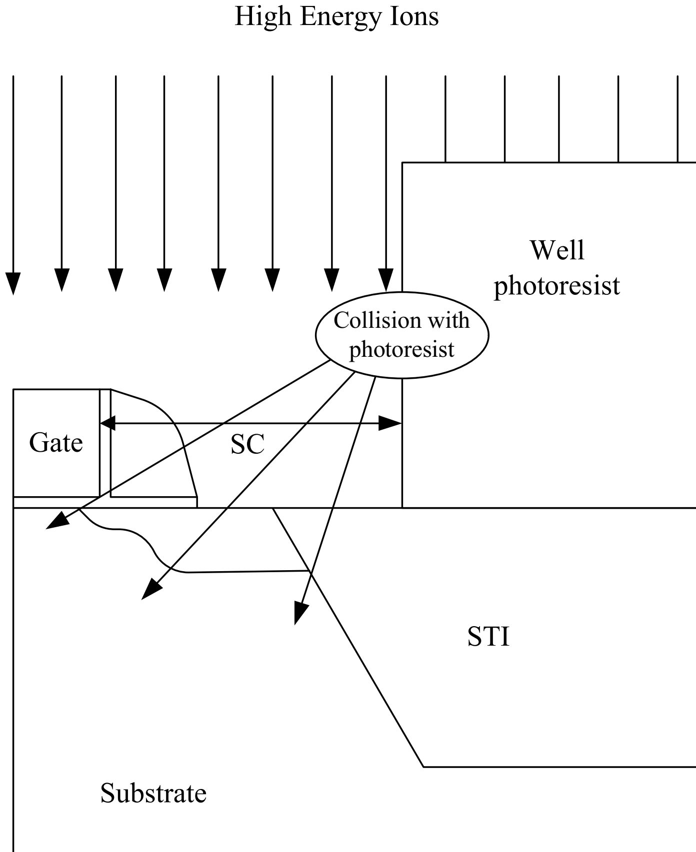


FIGURE 7.12
Illustration of WPE.

distance is less than $1\mu m$, significant increase of threshold voltage is possible [112].

7.5 Random Process Variations

Addressing the impact of random process variation on device/circuit performance is even more important in present day CMOS circuit design as innovative design techniques are sought after. The random process variations, which occur within a die (intra-die), between one device to another are completely statistical in nature. These sets of variations are the most critical concern in sub 90 nm VLSI design. Different sources of process variation which

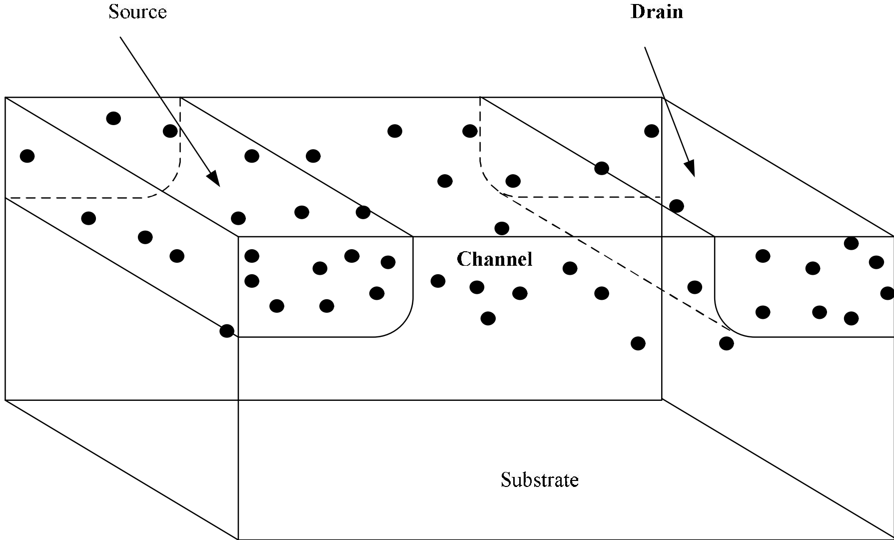


FIGURE 7.13
Illustration of random distribution of channel dopants.

fall under this category are random discrete dopant (RDD), line edge roughness/line width roughness (LER/LWR), oxide thickness variation (OTV), and poly-silicon/metal gate granularity.

7.5.1 Random Discrete Dopants

Random discrete dopants (RDD) which are currently the major sources of random process variations result from the discreteness of the dopant atoms in the channel region of a MOS transistor [163]. A common practice to control the threshold voltage of a MOS transistor is to dope the channel region with dopant atoms. For a MOS transistor with effective channel length L, effective channel width W and source-drain junction depth x_j, if the channel is doped with concentration N_{ch}, the total number of dopant atoms within the channel volume is given by

$$N_T \approx N_{ch}.W.L.x_j \tag{7.14}$$

It is therefore clear that with the continual down scaling of MOS technology, the total number of channel dopant atoms reduce even with increase in the channel doping concentration. It has been observed that a transistor in $1\mu m$ technology has about 5000 dopant atoms whereas that in 45-nm technology has only about 100 [104]. The number of dopant atoms in a transistor channel is a discrete statistical quantity, as shown in Fig. 7.13. Therefore, in an integrated circuit, the electrical characteristics of two transistors placed side by side will be different because of the randomness in a few dopant atoms. It

may be noted that such randomness in the number of dopant atoms occurs even in SDE doping and halo doping. The major effects of RDD are significant variations in the threshold voltage, variations in the overlap capacitance caused by uncertainty in the position of SDE dopants under the gate, and variations in the effective source-drain series resistance.

Considering uniform channel doping, the effect of RDD on threshold voltage fluctuation for a large geometry MOS transistor is given by [182, 192]

$$\sigma V_{T,RDD} = \frac{q}{C_{ox}}\sqrt{\frac{N_A W_{dm}}{3LW}} \tag{7.15}$$

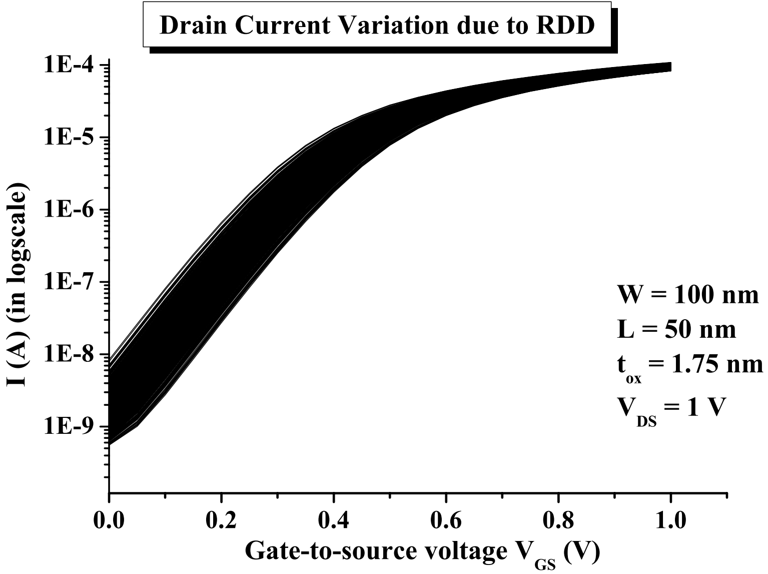
where W_{dm} is the depletion depth. For non-uniform doped transistors, N_A is to be replaced with N_{eff}, which is written as [190]

$$N_{eff} = 3\int_0^{W_{dm}} N(x)\left(1-\frac{x}{W_{dm}}\right)^2 \frac{dx}{W_{dm}} \tag{7.16}$$

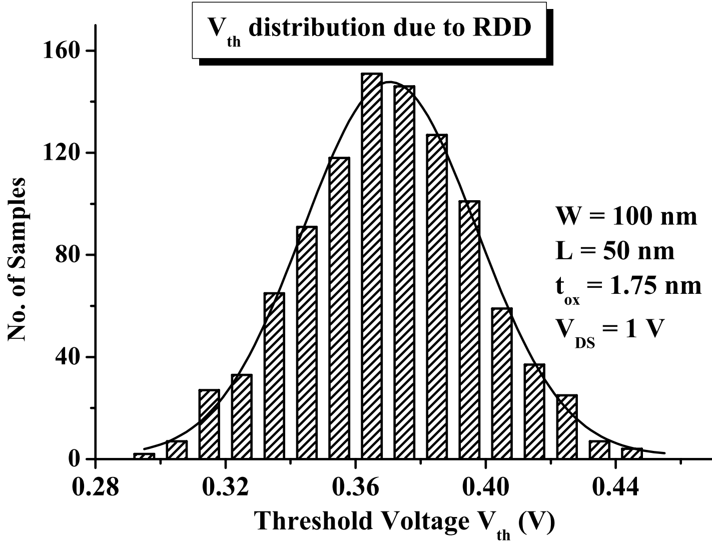
where $N(x)$ is the charge density along depth. With the scaling of CMOS technology, the device area LW decreases, so that the threshold voltage variability caused by RDD increases. However, RDD decreases with scaling of oxide thickness. It is found that RDD is a major contributor, over 60% to the threshold voltage mismatch. As a result, the process variations due to RDD cause significant variations in the drain current I_{DS} mismatch in analog circuits and also off-state leakage current in digital circuits and SRAM noise margins. The statistical fluctuations of the gate characteristics of an n-channel MOS transistor due to the RDD effect, as obtained from SPICE simulation results is shown in Fig. 7.14(a). The statistical distribution of the threshold voltage extracted from the curves is shown in Fig. 7.14(b). The SPICE simulation is performed through Monte Carlo analysis with the SPICE parameter $VTHO$. The doping concentration is taken to $3.24 \times 10^{18}/cm^3$. The $\sigma V_{T,RDD}$ comes out to be 26.52mV. From the distribution curve, the statistical mean of the threshold voltage is found to be 0.370V. The threshold voltage is extracted through the constant current method.

7.5.2 Line Edge Roughness

The critical physical dimensions of a semiconductor structure are defined by the process of lithography by exposing a light sensitive material (photo-resist). Until the 180nm node, the wavelength of light used to pattern these critical dimensions is scaled with the smallest of the dimensions to be patterned. In this regime lithography-induced variations were a result of lens imperfections, mask errors, illumination non-uniformity, and contributions arising from resist non-uniformities. At the 180nm node, scaling of the wavelength of light used for patterning ceased at 193nm due to increased cost of lithography technology, materials, and equipment development and deployment. The resulting lithographic defocus causes both systematic and random line edge and



(a) Statistical fluctuation of the gate characteristics of an n-channel MOS transistor



(b) Statistical distribution of the threshold voltage of a sample of n-channel MOS transistors

FIGURE 7.14
Effects of random discrete dopants as obtained from SPICE simulation results.

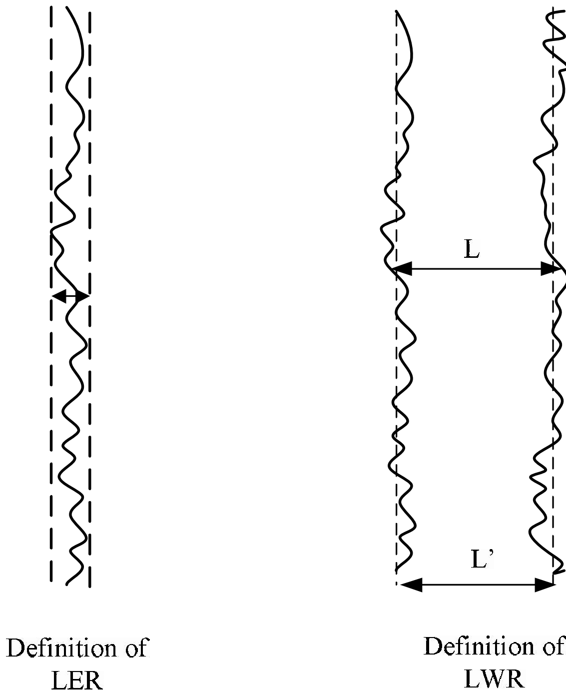


FIGURE 7.15
Relation between LER and LWR.

line-width variations in poly-gate patterning. During resist development, polymer aggregates along the edge of the mask are non-uniformly dislodged from their surrounding polymer matrix due to their different dissolution rates. This causes the formation of line-edge roughness (LER)[104]. Line edge roughness (LER) from two coupled edges leads to line width roughness (LWR). Therefore, LER is the phenomenon which causes LWR and it has been found that LWR is $\sqrt{2}$ times the LER [102]. This is shown in Fig. 7.15. It is observed that LER is the fluctuation of a line about its mean value for a given edge and LWR is the fluctuation of line width about its mean value averaged over the width W.

In order to characterize the distortion of the gate edge, a simplified model of a rough line as shown in Fig. 7.16 is considered [36]. The roughness in the gate edge is characterized by high frequency roughness and low frequency roughness. The gate is divided into segments with characteristic width W_c, which characterizes the change at which the low frequency part changes the gate length. Within this portion, only high frequency roughness is present. If there is no correlation of the variation along the two edges, standard deviation

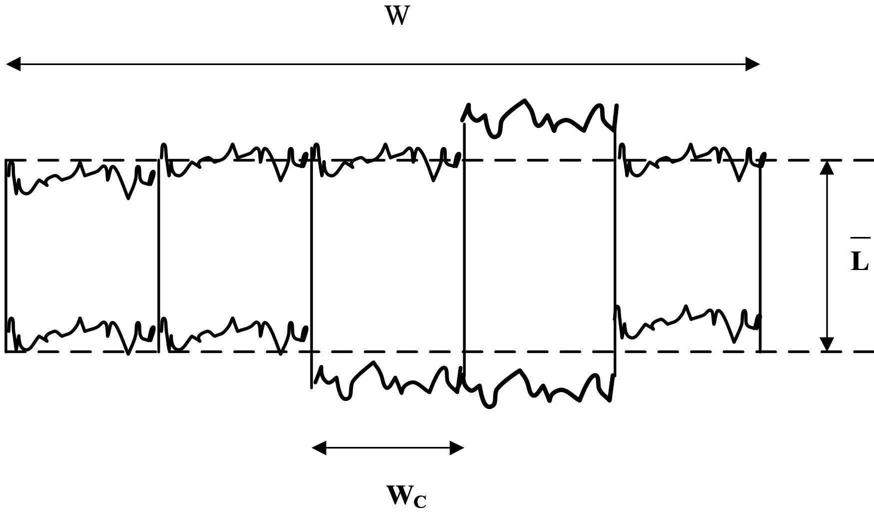


FIGURE 7.16
Simplified model for estimating the LER of a gate.

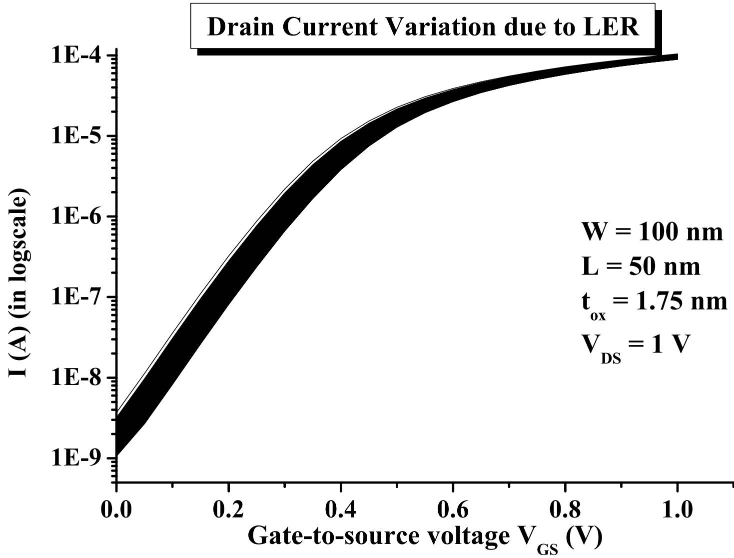
of the channel length variation due to LER/LWR is calculated to be [36]

$$\sigma L = \sqrt{\frac{2}{1+\frac{W}{W_c}}}\sigma_{LER} \tag{7.17}$$

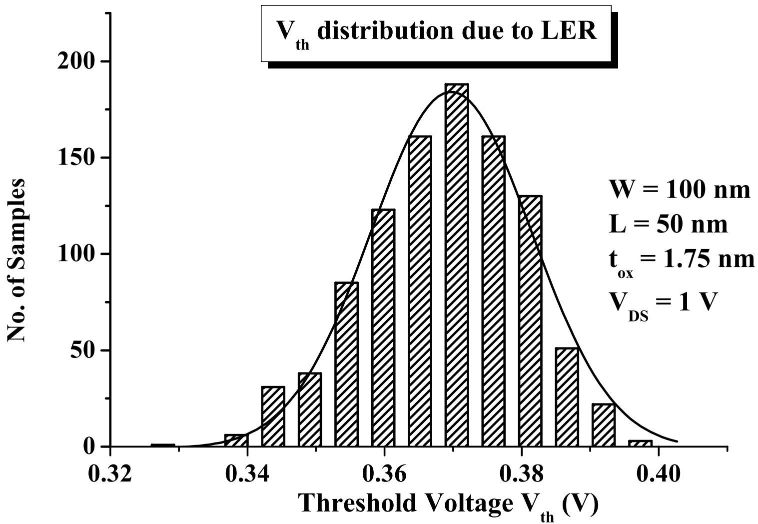
The edge locations of two different segments are uncorrelated and have a standard deviation σ_{LER}. The typical value of the amount of LER is $3\sigma_{LER} = 4nm$ [10]. As the edge fluctuation does not scale with technology scaling down, LER/LWR are expected to have more serious impact on sub-45nm devices. It may be noted that LER and RDD are statistically independent, such that [207]

$$\sigma_{V_T,total} = \sqrt{\left(\sigma_{V_T,RDD}\right)^2 + \left(\sigma_{V_T,LER}\right)^2} \tag{7.18}$$

The statistical fluctuations of the gate characteristics of an n-channel MOS transistor due to the LER effect, as obtained from SPICE simulation results is shown in Fig. 7.17(a). The statistical distribution of the threshold voltage extracted from the curves is shown in Fig. 7.17(b). The SPICE simulation is performed through Monte Carlo analysis with the SPICE parameter XL. The parameter W_c is taken to be 30nm. The standard deviation and statistical mean of the threshold voltage as extracted from the statistical distributions are found to be 11.30mV and 0.368V.



(a) Statistical fluctuation of the gate characteristics of an n-channel MOS transistor



(b) Statistical distribution of the threshold voltage of a sample of n-channel MOS transistors

FIGURE 7.17
Effects of line edge roughness as obtained from SPICE simulation results.

7.5.3 Oxide Thickness Variations

The gate oxide formation, though a well-controlled process, becomes critical when the oxide layer becomes only a few atomic layers. With SiO_2 gate oxide, variations of a single monolayer (approximately 0.2nm) are typical and result in 20% shifts in oxide thickness. In this condition physical limitations, such as interface roughness and oxide-layer non-uniformity lead to increased variability of the effective oxide thickness. Threshold voltage V_T which is strongly correlated with the oxide thickness t_{ox} fluctuates significantly with fluctuation of t_{ox} when MOSFET dimension goes below 30 nm, and becomes comparable to threshold voltage fluctuation due to RDD [11]. Variation in the oxide thickness can affect carrier mobility as well. The gate leakage current (due to tunneling) also depends on the oxide thickness. Therefore variation in oxide thickness will affect these device parameters as well, causing them to vary from their mean position.

The statistical fluctuations of the gate characteristics of an n-channel MOS transistor due to the OTV effect, as obtained from SPICE simulation results is shown in Fig. 7.18(a). The statistical distribution of the threshold voltage extracted from the curves is shown in Fig. 7.18(b). The SPICE simulation is performed through Monte Carlo analysis with the SPICE parameter $TOXE$. The threshold voltage variation due to oxide thickness is also statistically independent of RDD and LER so that [163]

$$\sigma_{V_T,total} = \sqrt{(\sigma_{V_T,RDD})^2 + (\sigma_{V_T,LER})^2 + (\sigma_{V_T,OTV})^2} \qquad (7.19)$$

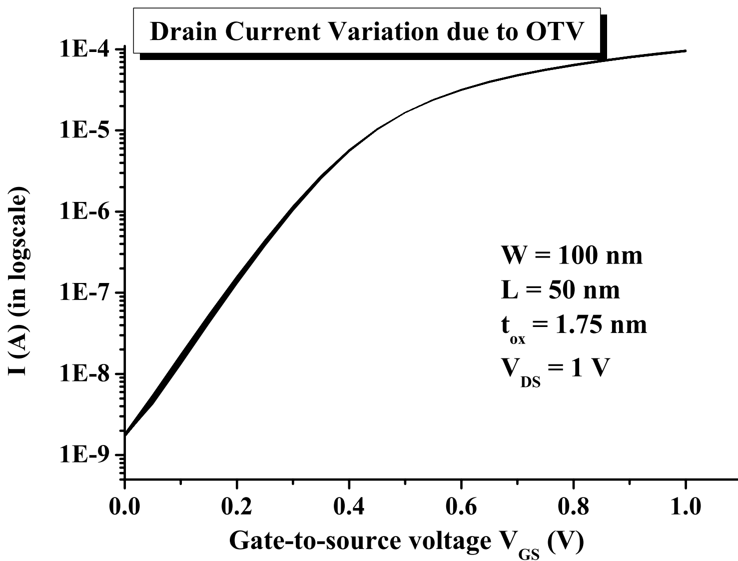
The contributions of the three effects on the two chosen performance parameters is shown in Fig. 7.19.

7.5.4 High-κ Dielectric Morphology and Metal Gate Granularity

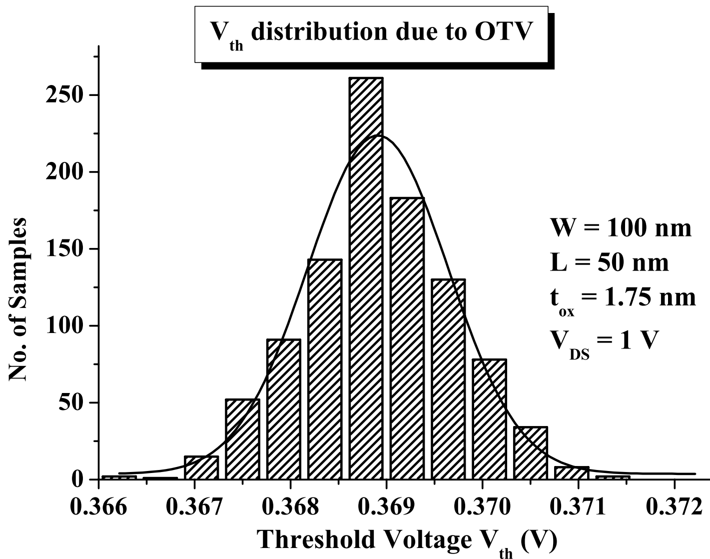
The high-κ-metal gate technology introduces a significant amount of process variability because of interface roughness between Si and the high-κ dielectric, and between the high-κ dielectric and the metal gate [163, 199]. The variation in the metal grain crystal orientation results in variation in the work function due to the different surface densities of polarization charges. Such work function variations lead to corresponding local threshold variations in the gate region.

7.6 Statistical Modeling

Because of variability constraints, a circuit optimized using conventional deterministic design methodology is more susceptible to random process fluctu-



(a) Statistical fluctuation of the gate characteristics of an n-channel MOS transistor



(b) Statistical distribution of the threshold voltage of a sample of n-channel MOS transistors

FIGURE 7.18
Effects of oxide thickness variations as obtained from SPICE simulation results.

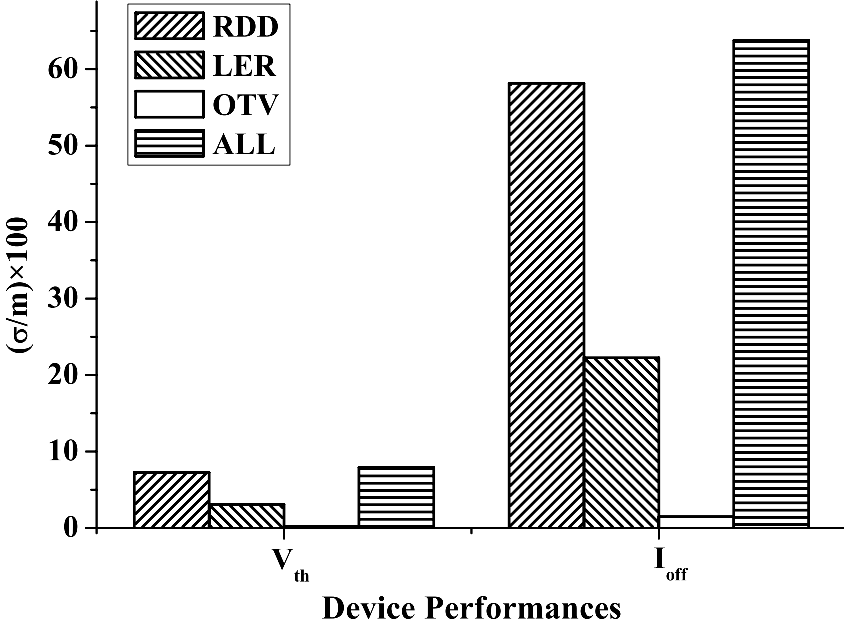


FIGURE 7.19
Contributions of each effect on the statistical variations of threshold voltage and off-current.

ations. Therefore, statistical design methodologies have become indispensable for modern VLSI circuit design [163]. Hence, accurate characterization and modeling of random intra-die process variations for circuit simulations are very important for accurate prediction of design yield and to achieve the benefits of the statistical design methodologies. Therefore it is essential to have accurate compact device models, an idea of the statistical variations of the compact model parameters, knowledge of the sensitivity of these models to process variations, and ultimately an awareness of the sensitivity of a given design to process variations. The various approaches to model the effect of process variation on CMOS circuits are discussed in the following sub-sections.

7.6.1 Worst Case Corner Analysis

Worst case analysis basically consists of considering the results of the worst combinations of the extreme fluctuations in an IC process in order to evaluate the range of circuit performance variations. The worst case corner models are generated by setting each process sensitive compact model parameters at a value deviated from their corresponding nominal values by some fraction of their respective standard deviations. In the case of MOSFETs, nominal values are captured in what is known as a ‘typical’ or TT library, while extreme

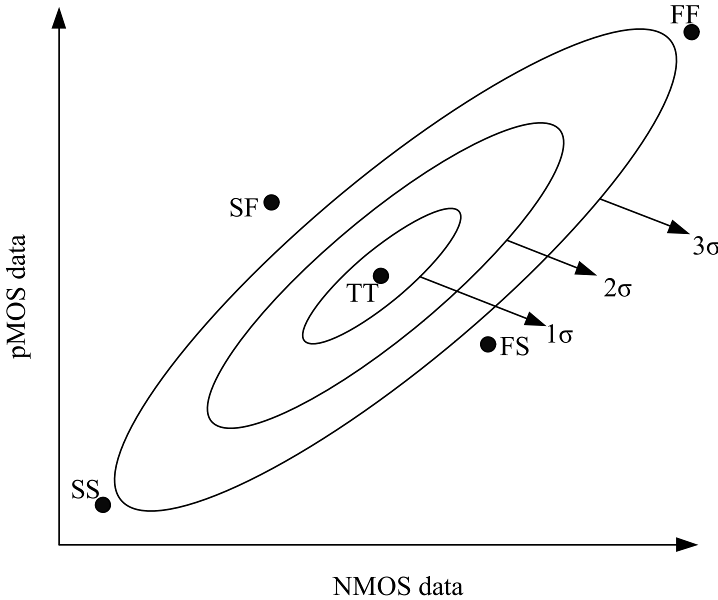


FIGURE 7.20
Data spread and design corners.

process values are captured in 4 corner libraries called FF, SS, FS, and SF (referring to fast NMOS and PMOS, slow N and P, fast N/slow P, and slow P/fast N, respectively). The TT model is generated from the measured data on a single golden wafer corresponding to the central-line process. Figure 7.20 describes the data spread from its typical value and position of the corners. Here σ represents the standard deviation calculated from the measured data. Conventionally, the process variability is modeled by the worst case four corners, two each for analog and digital applications. While the SS and FF corners are used for analog circuit modeling, the FS and SF corners utilized for digital circuit generation. The advantage of this design corner approach is that the corner models are supplied to the designers so that the circuits can be simulated at each of the four process corners. But, there are some problems with this approach. The fixed corners tend to be too pessimistic. As illustrated in Fig. 7.20, there are some extreme combinations of process parameters that are too unrealistic. Thus, a corner-based methodology often leads to over design. Moreover, while generating the corner parameters, the correlation between the core model parameters are ignored. Therefore, the design may still work, but it will take a larger die area and more design effort to achieve the same function. This approach therefore, does not provide adequate information about the robustness of the design. It may however, be noted that because of the

simplicity of the approach, until recently this is the most common technique used by the designer for estimating the robustness of the design.

7.6.2 Monte Carlo Simulation Technique

In the Monte Carlo methodology, the statistical model can be implemented based on measured/extracted SPICE parameters. The variation sources are usually represented by a set of independent random numbers with Gaussian distributions, which are directly or indirectly linked to the parameters in a compact model so that Monte Carlo simulation can be performed. This forms a large database of process related SPICE parameters and for each sample of the database, the performance parameters of the circuit are simulated through SPICE simulation. The statistical distribution of the performance parameters corresponding to each sample of the process database are estimated by determining the mean and the standard deviation. The yield is obtained as the fraction of samples that are accepted. The design flow of the Monte Carlo simulation-based statistical modeling is shown in Fig. 7.21. The problem of Monte Carlo simulations, however, is that hundreds of simulation runs have to be performed, and depending on the complexity of analog/mixed-signal circuits this may take several simulation hours to be completed. However, some strategies are available to reduce the sample size such as variance reduction techniques, stratified sampling, etc. [92].

7.6.3 Statistical Corner Technique

The basic idea behind the development of the statistical corner technique is to make the design corner technique more realistic by adding a realistic value of the standard deviation of the corresponding model parameter to its nominal value. For statistical corner analysis, statistical compact models are required, which are generated by the data obtained from different dies, wafers, and wafer lots collected over a period of time [13], [37]. The design flow of the statistical corner technique is shown in Fig. 7.22. The measured data may be the current-voltage (I-V) or the capacitance-voltage (C-V) measurement data for a MOS transistor device. In absence of actual measurement data, which are fairly common for new process technology, the production data may consist of calibrated TCAD simulation results. But for the realization of the statistical nature of the data, a large number of simulations are required to be performed. Now by applying suitable extraction procedure the SPICE model parameters are found from the measurement/simulated data. This type of model being derived from measured data, is much more realistic than WC corner models. These models give reasonably accurate results but are computationally intensive as parameters are required to be extracted from huge numbers of measurement/simulation data. Development of statistical compact models based on BSIM4 and PSP is reported in [28].

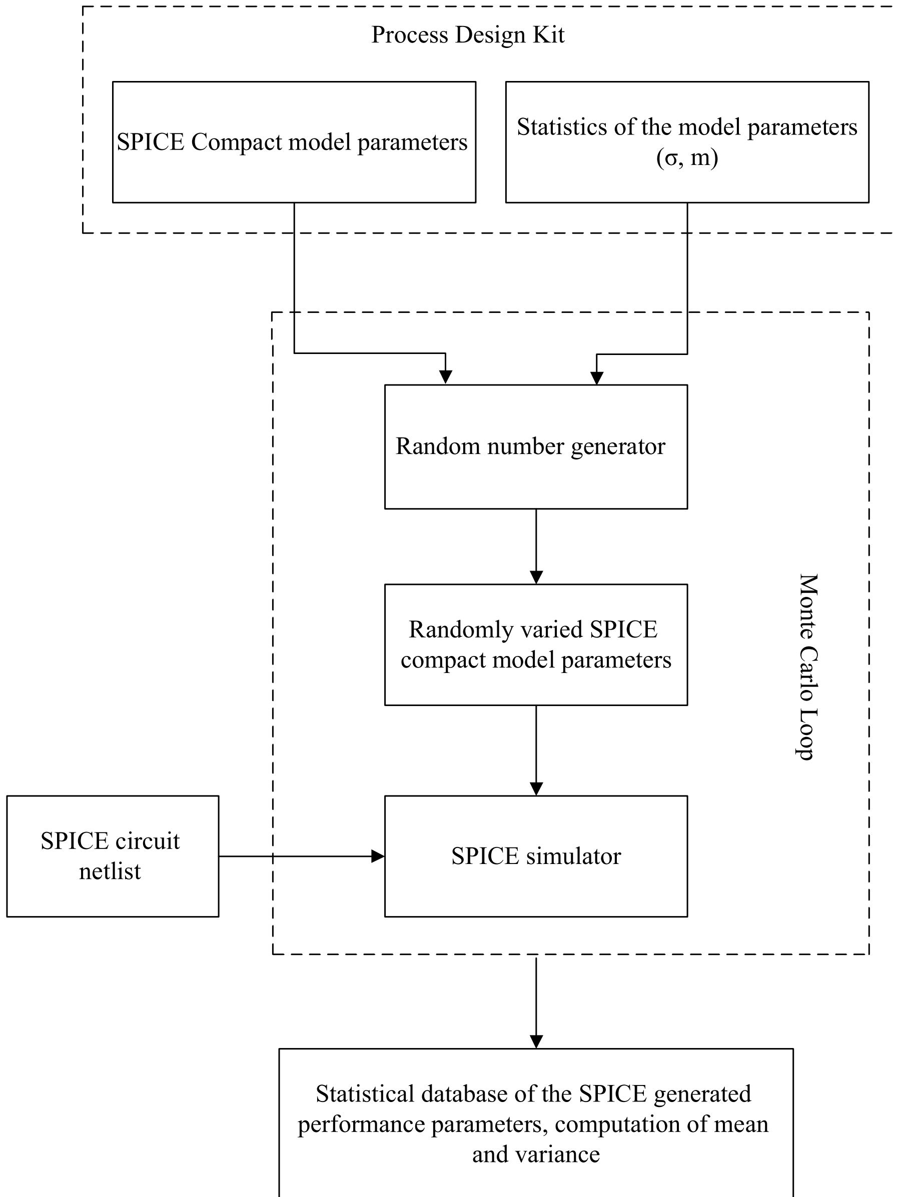


FIGURE 7.21
Statistical modeling using Monte Carlo simulation.

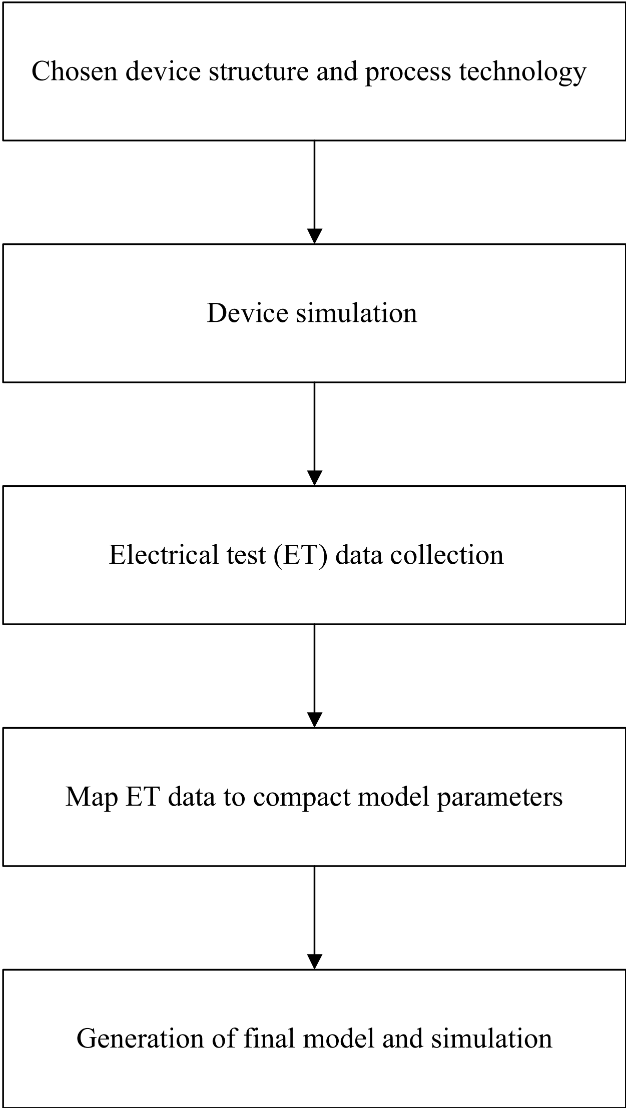


FIGURE 7.22
Statistical corner modeling using TCAD.

7.6.4 Mismatch in Analog Circuits

MOS transistor matching in analog CMOS applications deals with statistical device differences between pairs of identically designed and identically used transistors [143]. Mismatch is considered to be the key to precision analog IC design. Transistor mismatch limits the accuracy of circuits and has important implications on the performance of several circuits such as offsets of operational amplifiers or comparators, on the bit accuracy of ADC and DAC, and on the power- supply-noise and common-mode rejection ratios in differential structures.

The drain current mismatch of two closely spaced, identical transistors with zero source-bulk bias, is modeled by the random variation of the difference in their threshold voltage $\Delta V_T = \Delta P_1(say)$ and the current factor $\Delta\beta = \Delta P_2(say)$ where $\beta = \mu_s C_{ox}\,(W/L)$[52].

Assuming mismatch in the input parameters to be smaller than the parameter itself, we can expand the relative mismatch in the drain current as first-order Taylor series as

$$\frac{\Delta I_D}{I_D} = \frac{1}{I_D}\frac{\partial I_D}{\partial P_1}\Delta P_1 + \frac{1}{I_D}\frac{\partial I_D}{\partial P_2}\Delta P_2 \tag{7.20}$$

A first-order Taylor series assumes a linear mapping between the input process parameters and the electrical parameters. The statistical distribution of the mismatch of drain current between a pair of transistors may be characterized by its mean $m_{\Delta I_D/I_D}$ and standard deviation $\sigma_{\Delta I_D/I_D}$. These are written as

$$m_{\Delta I_D/I_D} = \frac{1}{I_D}\frac{\partial I_D}{\partial P_1}m_{\Delta P_1} + \frac{1}{I_D}\frac{\partial I_D}{\partial P_2}m_{\Delta P_2} \tag{7.21}$$

$$\sigma_{\Delta I_D/I_D} = \left[\left(\frac{1}{I_D}\frac{\partial I_D}{\partial P_1}\right)^2\sigma^2_{\Delta P_1} + \left(\frac{1}{I_D}\frac{\partial I_D}{\partial P_2}\right)^2\sigma^2_{\Delta P_2} + \frac{2}{I_D^2}\frac{\partial I_D}{\partial P_1}\frac{\partial I_D}{\partial P_2}cov\left(\Delta P_1, \Delta P_2\right)\right]^{1/2} \tag{7.22}$$

Here $cov(\Delta P_1, \Delta P_2)$ represents the covariance between the mismatches in P_1 and P_2.

The drain current mismatch due to threshold voltage mismatch can be written as

$$\frac{\Delta I_D}{I_D} = \frac{1}{I_D}\frac{\partial I_D}{\partial V_T}\Delta V_T \cong -\frac{g_m}{I_D}\Delta V_T \tag{7.23}$$

Therefore,

$$\sigma^2_{\Delta I_D/I_D} = \left(\frac{g_m}{I_D}\right)^2\sigma^2_{\Delta V_T} \tag{7.24}$$

The drain current mismatch due to the current factor mismatch may be written as

$$\Delta I_D = \frac{\partial I_D}{\partial \beta}\Delta\beta \tag{7.25}$$

Both in the saturation and linear region, it is easy to show that

$$\frac{\Delta I_D}{I_D} = \frac{\Delta\beta}{\beta} \tag{7.26}$$

Neglecting the mobility fluctuation and assuming that μ_s is constant, the current factor mismatch arises only from OTV and channel length mismatch. Therefore,

$$\Delta\beta = \frac{\partial\beta}{\partial L}\Delta L + \frac{\partial\beta}{\partial t_{ox}}\Delta t_{ox} \tag{7.27}$$

$$= -\frac{\beta}{L}\Delta L - \frac{\beta}{t_{ox}}\Delta t_{ox} \tag{7.28}$$

Hence,

$$\frac{\Delta\beta}{\beta} = -\frac{\Delta L}{L} - \frac{\Delta t_{ox}}{t_{ox}} \tag{7.29}$$

The variance in drain current mismatch due to current factor mismatch becomes

$$\sigma^2_{\Delta I_D/I_D} = \sigma^2_{\Delta\beta/\beta} = \sigma^2_{\Delta L/L} + \sigma^2_{\Delta t_{ox}/t_{ox}} \tag{7.30}$$

Therefore, the complete drain current mismatch model becomes

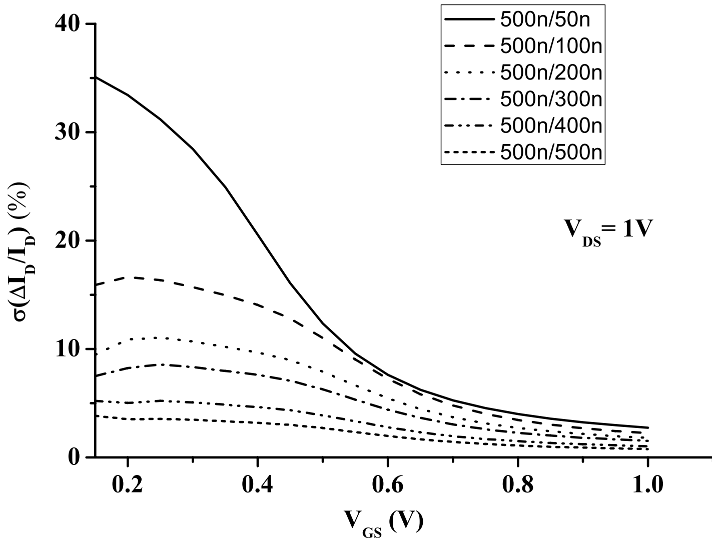
$$\frac{\Delta I_D}{I_D} = \left.\frac{\Delta I_D}{I_D}\right|_{\Delta V_T} + \left.\frac{\Delta I_D}{I_D}\right|_{\Delta\beta/\beta} \tag{7.31}$$

$$\sigma^2_{\Delta I_D/I_D} = \left(\frac{g_m}{I_D}\right)^2 \sigma^2_{\Delta V_T} + \sigma^2_{\Delta\beta/\beta} - 2\frac{g_m}{I_D}.\frac{1}{\beta}cov\left(\Delta\beta, \Delta V_T\right) \tag{7.32}$$

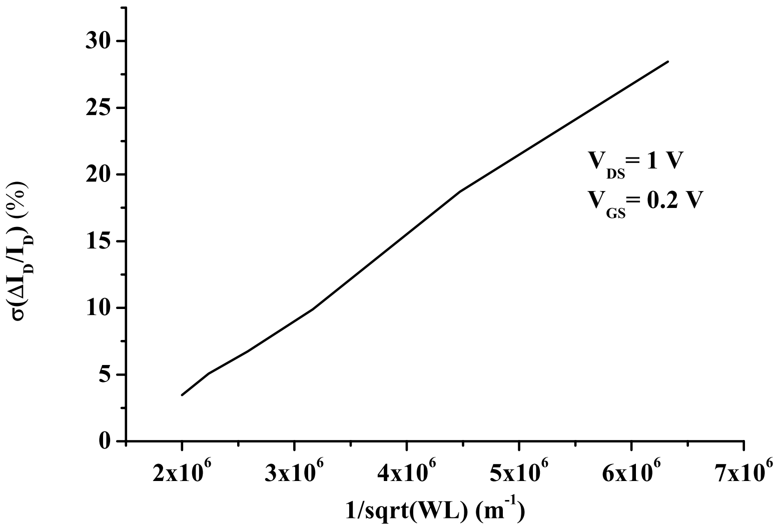
The variation of the drain current mismatch of a MOS transistor with respect to the gate-bias and the corresponding dependence on the area of the transistor is shown in Fig. 7.23(a) and 7.23(b). It is observed that the drain current mismatch is maximum in the weak inversion region and reduces as the operating point moves toward the strong inversion region. Also, the mismatch reduces with increase in the area of the transistor. This is referred to as the Pelgrom's law [143].

7.7 Physical Phenomena Affecting the Reliability of Scaled MOS Transistor

The major physical phenomena affecting the reliability of analog circuits are (1) time-dependent dielectric breakdown, (2) hot carrier injection, and (3) negative bias temperature instability. These phenomena are described below at an introductory level in the following sub-sections.



(a) Variation of drain current mismatch between a pair of identical n-channel MOS transistors



(b) Area dependence of the mismatch

FIGURE 7.23
Drain current mismatch.

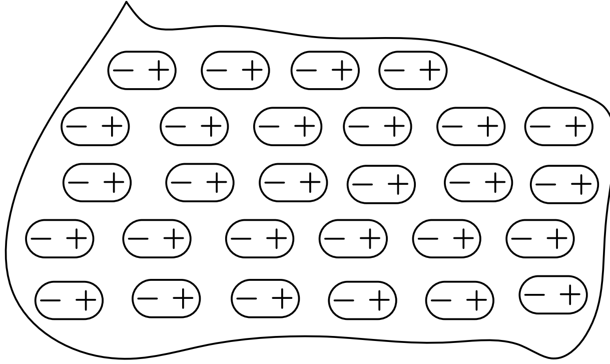


FIGURE 7.24
Polarized dielectric medium.

7.7.1 Time Dependent Dielectric Breakdown (TDDB)

The time dependent dielectric breakdown phenomenon in gate oxides is an irreversible reliability phenomenon that results in a sudden discontinuous increase in the conductance of the gate oxide at the point of breakdown, as a result of which the current through the gate insulator increases significantly [192]. The breakdown of a gate dielectric insulator occurs due to the application of a high electric field (such as 5MV/cm) over a period of time. There are several theories explaining the TDDB phenomenon. Among them the two most important theories are the anode hole injection theory and the percolation theory. However, before going into the details about the TDDB phenomenon in scaled MOS transistors it is important to understand the electrostatics of any dielectric medium.

7.7.1.1 Electrostatics in Dielectrics

An ideal dielectric material is one which does not contain any free charge [27]. The molecular charges present in a dielectric material are not free to move very far; they are called bound charges to distinguish them from the free charges of a conductor. Due to an applied external electric field, the entire positive charges inside the dielectric are displaced with respect to the negative charges. The dielectric under such conditions is said to be polarized (see Fig. 7.24). The electric displacement vector $\bar{D}$ is defined by

$$\bar{D} = \epsilon_0 \bar{\xi} + \bar{P} \tag{7.33}$$

where ϵ_0 is the free space permittivity and $\bar{P}$ is the electrical polarization vector and $\bar{\xi}$ is the electrical field due to both bound charges as well as free charges within the dielectric. The functional relationship between the polarization $\bar{P}$ and the electric field $\bar{\xi}$ is given by the following constitutive relation

$$\bar{P} = \epsilon_0 \chi \bar{\xi} \tag{7.34}$$

where χ is a dimensionless parameter, referred to as the electrical susceptibility. The displacement vector is therefore, related to the electric field as follows [27]

$$\bar{D} = \epsilon \bar{E} = \epsilon_0 \kappa \bar{\xi} \tag{7.35}$$

where ϵ is the permittivity of the dielectric material. The quantity κ given by

$$\kappa = \frac{\epsilon}{\epsilon_0} = 1 + \chi \tag{7.36}$$

is referred to the as the relative permittivity or the dielectric constant of the material. The dielectric material for which the dielectric constant κ is a constant is referred to as the linear dielectric. If the susceptibility χ of a dielectric material does not depend on the direction of the electric field, it is referred to as an isotropic dielectric. With the application of a sufficiently high electric field, the linear relation between $\bar{P}$ and $\bar{\xi}$ no longer remains valid; the dielectric breaks down, which implies that the electrons are pulled out of the molecules. The material ceases to act as an insulator since the electrons torn away from the molecules become conducting. The critical electric field at which the dielectric breakdown occurs is referred to as the dielectric strength of the material. For air, the typical value of the dielectric strength is about $3 \times 10^6 V/m$ and for most of the solid dielectrics, it is a few times $10^8 - 10^9 V/m$.

7.7.1.2 Energy Band Theory of Dielectric Breakdown

The phenomenon of dielectric breakdown can be also understood from the energy band theory [27]. According to the energy band theory, for dielectric materials, the band gap E_g between the bottom of the conduction band and the top of the valence band is very high. At ordinary temperature, the electrons do not possess sufficient thermal energy to cross the band gap and reach the conduction band. Even with small amount of electric field, the electrons in the valence band cannot carry a current because the valence band is completely filled up. However, if the applied field is sufficiently large, the electrons in the valence band can move into the conduction band by accepting energy from the field. These electrons in the conduction band become free electrons and hence take part in current conduction. Correspondingly holes are created in the valence band which also take part in current conduction. The dielectric breakdown has now occurred, and both the free electrons and holes carry currents.

A rough estimate of the dielectric strength can be made from the energy band theory. If ξ_c is the dielectric strength, the force on an electron of charge q is $q\xi_c$. The energy gained by an electron in moving between successive atoms is given by $q\xi_c.a$ where a is the lattice constant of the material. As a rough approximation, the dielectric breaks down when this energy equals the band gap energy E_g. Therefore, the dielectric strength can be estimated as follows,

$$\xi_c = \frac{E_g}{qa} \tag{7.37}$$

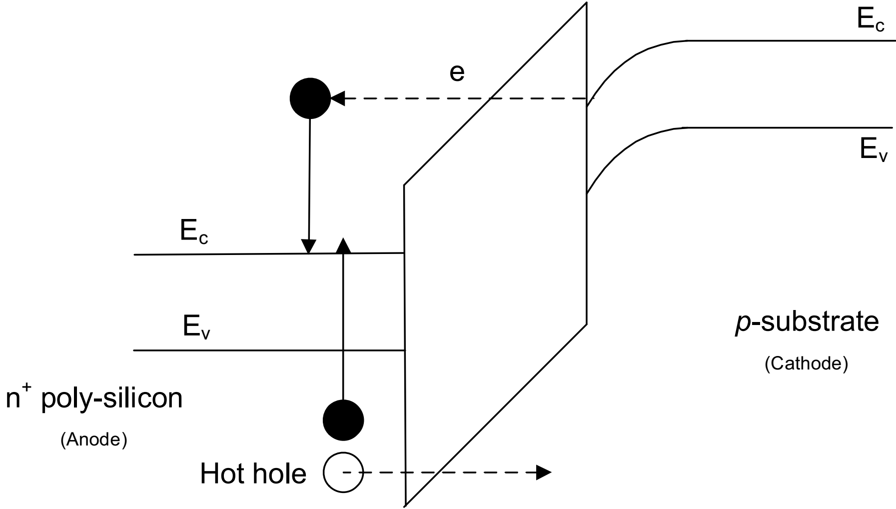


FIGURE 7.25
Schematic illustration of the anode hole injection mechanism.

7.7.1.3 Anode Hole Injection Theory

With the application of a large positive bias in the gate electrode, electrons in the strongly inverted surface tunnel into the conduction band of the oxide layer through the F–N tunneling mechanism [169, 185] discussed in Chapter 6 of this text. The electrons, by gaining energy from the electric field, become hot electrons in the gate oxide. These electrons arrive at the anode (electron sink, i.e., the gate electrode) and cause impact ionization in the anode near the oxide-anode interface. Electron-hole pairs are generated. Some of the tunneled electrons elastically transfer their entire energy to deep valence band electrons, so that the holes generated become hot holes. These hot holes have high probability of being injected into the oxide layer. The anode hole injection process is schematically illustrated in Fig. 7.25 [192]. The injected holes can be trapped in the oxide layer while traveling toward the cathode (electron source, i.e., the inverted surface). These cause an increase in the oxide field ξ_{ox} near the cathode and a decrease in the oxide field near the anode. This in turn increases the F–N tunneling current. The hole trapping in the oxide near the cathode therefore, provides a positive feedback to the electron tunneling process. Eventually this positive feedback leads to a runaway of the electron tunneling current at some local weak spots of the oxide.

A suitable parameter to characterize breakdown due to charge injection is the "charge-to-breakdown" Q_{BD}, which is defined as

$$Q_{BD} = \int_0^{t_{BD}} Jdt = Jt_{BD} \quad \mathrm{C/cm^2} \tag{7.38}$$

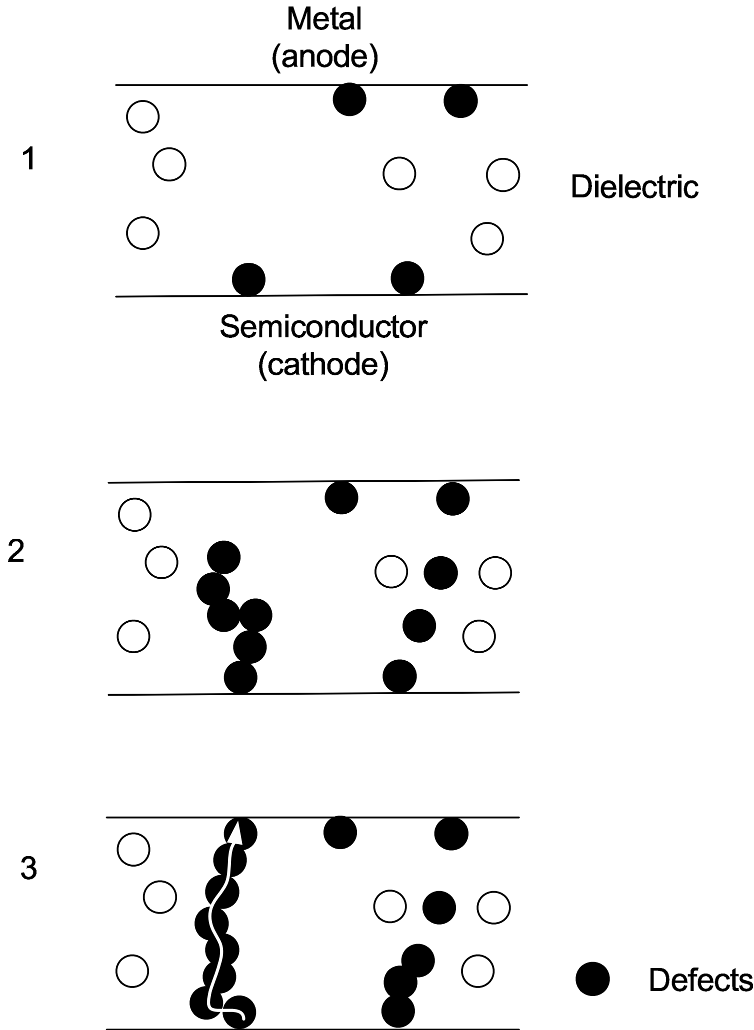


FIGURE 7.26
Percolation of defects and breakdown of ultra-thin gate oxide.

where t_{BD} is the time necessary to breakdown. Referring to the discussion in Chapter 6, regarding the F–N current, the time-to-breakdown t_{BD} is found to be inversely proportional to the applied electric field ξ_{ox}, i.e., $t_{BD} \propto \exp(\beta/\xi_{ox})$, where β is the electric field acceleration factor. The reciprocal field dependence is a consequence of the Fowler–Nordheim current, which is the driving force for the breakdown. Because of the $1/\xi_{ox}$ dependence, the anode hole injection theory is sometimes referred to as $1/E$ model where E signifies the energy.

7.7.1.4 Percolation Theory

The anode hole injection theory is found to give good explanations of experimental observations under high stress field. However, for the scaled MOS transistors with scaled oxide thickness and supply voltage, the percolation theory [185, 176] depending upon the generation of traps and conduction of current through these traps is found to be the better physical model for prediction purposes. This theory is explained as follows.

According to the percolation theory when the defects are dense enough to form a continuous chain connecting the gate to the semiconductor, a conduction path is created and catastrophic breakdown occurs. The concept does not depend in any way on the physics of defect generation. A series of schematics illustrating the percolation of defects and ultimate breakdown of the gate dielectric is illustrated in Fig. 7.26 [185]. Oxide traps/defects are generated throughout the volume of the dielectric. As the generation process continues, the defects can connect electrically to the anode, cathode, or to nearest neighbors. If two neighboring defects overlap or are so placed that they come in contact with one of the electrodes, conduction is possible. Breakdown occurs when a continuous path from one electrode to the other is created through the defects. Two important obvious conclusions are (1) if the oxide is made thinner a percolation path can result with a lower critical defect density N_{BD}, (2) a device with a larger area would also have a higher probability of having overlapping defects for the same oxide thickness and defect density.

7.7.1.5 Statistics of Gate Oxide Breakdown

The statistics of gate oxide breakdown are described using the Weibull distribution

$$F(x) = 1 - \exp\left[-\left(\frac{x}{\alpha}\right)^{\beta}\right] \tag{7.39}$$

where F is the cumulative failure probability, x can be either charge or time, i.e, Q_{BD} or t_{BD}, β is called the slope parameter or Weibull shape, the characteristic life α is percentile 63.2. The Weibull shape parameter β is experimentally observed to decrease as the oxide thickness is decreased. The failure distribution becomes wider (smaller β) as the oxide thickness is reduced. This is because of the fact that the critical defect density N_{BD} reduces as the oxide thickness is reduced.

7.7.1.6 Soft and Hard Breakdown

The gate oxide breakdown phenomenon is typically classified into soft or hard depending on the magnitude of the post-breakdown conduction. The hard breakdown occurs abruptly and is characterized by an abrupt increase in gate leakage or as sudden collapse in the gate voltage. Hard breakdown provokes the complete loss of the oxide dielectric properties with gate currents in the mA range at standard operating voltages. The soft breakdown, on the other

hand, occurs rather gradually or softly and is observed only as a slight change in voltage or current, usually accompanied by noise or signal fluctuations.

7.7.2 Hot Carrier Injection (HCI)

The hot carrier effect has already been introduced in Chapter 3 of this text. This effect also affects the reliability of MOS transistors which is one of the subject matters of discussion in this chapter.

The hot carrier injection phenomenon refers to the injection of carriers into the channel or gate insulator produced by impact ionization near the drain end of the channel creating interface and oxide trap damage [112, 84]. This hot carrier injection phenomenon leads to a long-term reliability problem, or "aging problem," where a circuit might degrade or fail after being in use for some time. The degradation may be attributed to increase of threshold voltage, reduction of carrier mobility, which in turn leads to lowering of drain current, transconductance, and switching speed [125]. As holes are much "cooler" than electrons, hot carrier effects in n-channel MOS transistors are found to be more significant than in p-channel MOS transistors. The degradation occurs by two mechanisms: (i) charge trapping in the oxide and (ii) generation of interface traps. Experimental results have suggested that of the two mechanisms, interface trap generation is the most important.

7.7.3 Negative Bias Temperature Instability (NBTI)

NBTI refers to the threshold voltage instability in p-channel transistors that is dependent on temperature and transistor geometry, particularly the channel width [112]. NBTI causes an increase in the absolute threshold voltage, degradation of the mobility, drain current, and transconductance of p-channel MOS transistors [126]. This process does not require the presence of a high lateral field and the resulting hot carriers. The mechanism is attributed to the breaking of SiH bonds at the Si-SiO_2 interface by a combination of electric field, temperature, and holes, resulting in dangling bonds or interface traps at that interface and positive oxide charge.

7.8 Physical Model for MOSFET Degradation Due to HCI

This section discusses a physical model [84, 125, 54] for computing the number of generated interface traps due to HCI and the application of the model for analog circuit design.

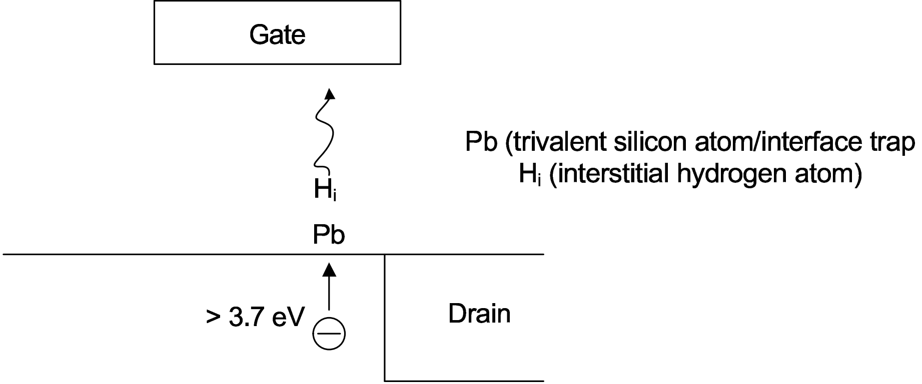


FIGURE 7.27
Schematic illustration of the generation of interface traps due to channel hot electrons.

7.8.1 Physical Mechanism for Interface Trap Generation

The silicon atom possesses four valence electrons and therefore requires four bonds to fully complete the valence shell. In the crystalline structure each silicon atom is attached to its four neighboring atoms, leaving no incomplete bond behind. However, at the surface of the silicon crystal, atoms are missing and traps are formed. The incomplete bonds form traps. After oxidation, the incomplete bonds are attached with oxygen atoms, thereby reducing the number of traps. The quality of the interface is further improved, i.e., the number of dangling valence bonds is further reduced by annealing the Si-SiO_2 interface in forming gas (N_2/H_2) mixture or other nitrogen containing gases such as ammonia. Atomic hydrogen passivates the silicon dangling bonds at the $Si(111) - SiO_2$ interface at the room temperature via the reaction $Pb + H^0 \rightarrow PbH$ where PbH represents the hydrogen-passivated dangling bond. With this treatment the amount of electrically active interface states can be reduced. However, the Si-H bonds can break at elevated temperatures and high electric fields due to their lower binding energy and re-activate the interface states. A hot electron breaks the PbH bond to produce Pb centers (trivalent silicon atom, i.e., the interface traps) and an interstitial hydrogen atom. It may be noted that if the resultant trivalent silicon atom recombines with hydrogen, an interface trap will not be generated. The interface trap will only be generated if the hydrogen atoms diffuses away from the interface. A schematic representation of the interface trap generation is shown in Fig. 7.27 [84].

7.8.1.1 Physical Model

From the lucky electron model concept, the rate of bond breakage is expressed as

$$R_B = K\left(\frac{I_{DS}}{W}\right)\exp\left(-\frac{\phi_{it}}{q\lambda\xi_m}\right) \quad (7.40)$$

where I_{DS} is the drain current, i.e., the rate of "cold" electron flow, W is the channel width, so that I_{DS}/W is proportional to the electron density, K is proportional to the density of PbH bonds, ϕ_{it} is the critical energy that a hot electron must have in order to create an interface trap, λ is interpreted as the hot electron mean free path, and ξ_m is the maximum electric field at the drain end.

The rate of recombination of the trivalent silicon atom with the hydrogen atom is expressed as

$$R_R = BN_{it}n_H(0) \quad (7.41)$$

where $n_H(0)$ is the concentration of the interstitial hydrogen atom at the interface and N_{it} is the interface trap density.

The net rate of interface traps generation is therefore,

$$\frac{dN_{it}}{dt} = K\left(\frac{I_{DS}}{W}\right)\exp\left(-\frac{\phi_{it}}{q\lambda\xi_m}\right) - BN_{it}n_H(0) \quad (7.42)$$

This is equal to the rate at which the hydrogen atom diffuses away from the interface. Therefore,

$$\frac{dN_{it}}{dt} = \frac{D_H n_H(0)}{X_H} \quad (7.43)$$

where D_H is the effective diffusion constant and X_H is the effective length of diffusion of hydrogen. Eliminating $n_H(0)$ between (7.42) and (7.43), we get

$$\frac{dN_{it}}{dt}\left(1 + BN_{it}X_H/D_H\right) = K\left(\frac{I_{DS}}{W}\right)\exp\left(-\frac{\phi_{it}}{q\lambda\xi_m}\right) \quad (7.44)$$

By integration, the following expression for the interface trap density is obtained.

$$\frac{BX_H}{2D_H}N_{it}^2 + N_{it} = Kt\left(\frac{I_{DS}}{W}\right)\exp\left(-\frac{\phi_{it}}{q\lambda\xi_m}\right) \quad (7.45)$$

The dynamics of N_{it} growth is very similar to the oxidation kinetics. If the interface trap density is small, then the rate is reaction limited and $N_{it} \propto t$. On the other hand, if the interface trap density is large, then the rate is diffusion limited and $N_{it} \propto t^{1/2}$. In general as in the case of thermal oxidation, we have

$$N_{it} = C\left[t\frac{I_{DS}}{W}\exp\left(-\frac{\phi_{it}}{q\lambda\xi_m}\right)\right]^n \quad (7.46)$$

with n in the range of 0.5-1. The exact values of ϕ_{it} and n are to be determined

experimentally. The device lifetime is determined by the time at which the interface trap density reaches a certain value and is given as [84]

$$\tau = C_1 \frac{W}{I_{DS}} \exp\left(\frac{\phi_{it}}{q\lambda\xi_m}\right) \tag{7.47}$$

where the constant C_1 contains C, n and the chosen N_{it} value.

7.8.1.2 Application to Analog Circuit Design

It is observed from (7.46) that the interface trap density is approximated by a t^n power law. However, for circuit simulation, it is difficult to evaluate. This is avoided by expressing $N_{it}(t)$ to be [54]

$$N_{it}(t) \simeq N_{it0} + \Delta N_{it}(t) \tag{7.48}$$

where N_{it0} is the initial trap density. Now, from the derived physical model (7.44) it can be written that

$$\Delta N_{it}(t) = K \frac{\exp\left(\frac{-V_C}{V_{DS0} - V_{DSsat0}}\right)}{1 + AN_{it0}} \frac{I_{DS0}}{W}.t \tag{7.49}$$

where $A = BX_H/D_H$ and K are technology dependent parameters. V_{DS} is the applied drain voltage and V_{DSsat} is the drain-to-source voltage and V_C is the voltage proportional to the critical energy ϕ_{it} required to create interface traps.

It has been found that the threshold voltage shift ΔV_T is proportional to ΔN_{it} such that [84, 54]

$$\Delta V_T \propto \Delta N_{it}(t) = \Gamma.t \tag{7.50}$$

where

$$\Gamma = C_k \frac{I_{DS0}}{W} \exp\left(\frac{-V_C}{V_{DS0} - V_{DSsat0}}\right) \tag{7.51}$$

The two unknown parameters in this model are C_k and V_C. The values of these parameters are to be extracted under different stress conditions, i.e., for different V_{DS} and I_{DS}. A routine for extraction of these parameters is detailed in [54].

The threshold voltage degradation can be directly related to the interface trap density as

$$\Delta V_T(t) = \frac{q\left(N_{it}(t) + N_{ot}\right)}{C_{ox}} \tag{7.52}$$

where C_{ox} is the oxide capacitance per unit area and N_{ot} is the number of oxide traps. The change in channel carrier mobility due to the generation of interface traps may be represented by [125]

$$\mu_s = \frac{\mu_0}{(1 + \theta(V_{GS} - V_T)(1 + \alpha\gamma\Delta V_T)} \tag{7.53}$$

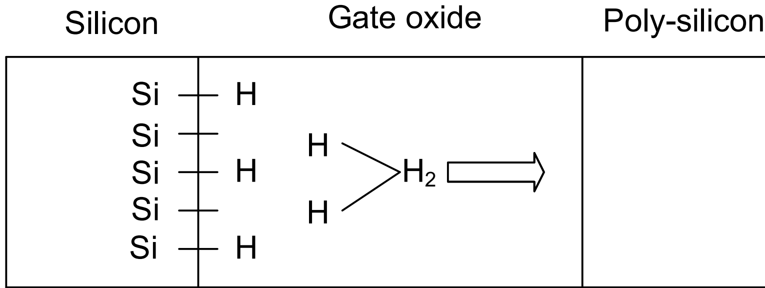


FIGURE 7.28
Schematic illustration of the reaction-diffusion model.

where θ and α are process dependent parameters. The value of γ depends on V_{DS}. The value of γ is close to zero when the transistor remains in the saturation region and is equal to unity when the transistor remains in the linear region. The degradation of the output conductance can be expressed as a change in the early voltage parameter due to threshold voltage shift.

The bias conditions that are being applied for reliability testing of a device may be identical to that in normal operating conditions or can be an "overstress" state, for example, by using a higher value of the supply voltage. The device is operated with the applied stresses for an extended period of time with the chosen performance parameters being continuously monitored. The stress test continues until the failure criterion is reached. The time τ to reach failure is called the hot-carrier-device lifetime or simply the MOS lifetime. The devices are designed so that they operate reliably for 10 years under normal operating conditions. However, for testing purposes the procedure cannot continue for 10 years. Therefore, the stress test is often accelerated so that the lifetime τ can be determined in a reasonable time. One possible approach may be to carry out the test procedure under overstress conditions so that the monitor quantity (e.g., I_{sub}) is high and therefore, τ is short. The lifetime τ and I_{sub} are related through a power-law function of the general form

$$\tau \approx K_1 \left(\frac{I_{sub}}{I_{DS}} \right)^{-m} \tag{7.54}$$

where K_1 and m are determined empirically. Some typical values are $m = 3; K_1 = 3$. The MOS lifetime τ are first measured at several high biases and the measured values are then extrapolated using (7.54) to predict τ under normal operating conditions.

7.9 Reaction-Diffusion Model for NBTI

Various NBTI models have been proposed. Some excellent review papers on this topic are [4, 177, 168]. The reaction-diffusion (RD) model is the most prevalent. The model assumes that when a gate voltage is applied, it initiates a field dependent reaction at the $\mathrm{Si}-\mathrm{SiO_2}$ interface that generates interface traps by breaking the passivated $\mathrm{Si}-\mathrm{H}$ bonds. Prior to this dissociation inversion layer holes are captured by the $\mathrm{Si}-\mathrm{H}$ covalent bonds which make the bonds weak. Therefore, the existing $\mathrm{Si}-\mathrm{H}$ covalent bonds are easily broken at high temperatures. The hydrogen atoms are thus released in the reaction phase. In the diffusion phase, the released hydrogen atoms diffuse away from the interface, leaving behind positively charged interface states. These states are responsible for higher threshold voltage and lower transconductance. The process is schematically explained in Fig. 7.28 [4].

The R-D model is described by the following equations

$$\frac{dN_{it}}{dt} = k_F\left(N_0 - N_{it}\right) - k_R N_H N_{it} \quad (x = 0) \tag{7.55}$$

$$\frac{dN_{it}}{dt} = D_H\left(\frac{dN_H}{dx}\right) + \delta/2\frac{dN_H}{dt}(0 < x < \delta) \tag{7.56}$$

$$D_H\left(\frac{d^2N_H}{dx^2}\right) = \frac{dN_H}{dt} \quad (\delta < x < t_{ox}) \tag{7.57}$$

$$D_H\left(\frac{dN_H}{dx}\right) = k_p N_H \quad (x.t_{ox}) \tag{7.58}$$

where $x = 0$ denotes the Si-SiO$_2$ interface, t_{ox} is the oxide thickness, N_{it} is the number of interface traps at any given instant, N_0 is the initial number of unbroken Si-H bonds, N_H is the hydrogen concentration, k_F is the oxide field dependent forward dissociation rate constant, k_R is the annealing rate constant, D_H is the hydrogen diffusion coefficient, and δ is the interface thickness.

7.10 Reliability Simulation for Analog Circuits

In order to reduce the design productivity and design creativity gap, reliability analysis is becoming mandatory within the design loop. This leads to the concept of design for reliability (DFR) which refers to design and verification methodology employed to assure that the IC performance does not substantially degrade over the anticipated lifetime of the product [112]. In order to predict the effects of various reliability phenomena on the performance of analog circuits, an equivalent transistor SPICE model can be used. Such a model

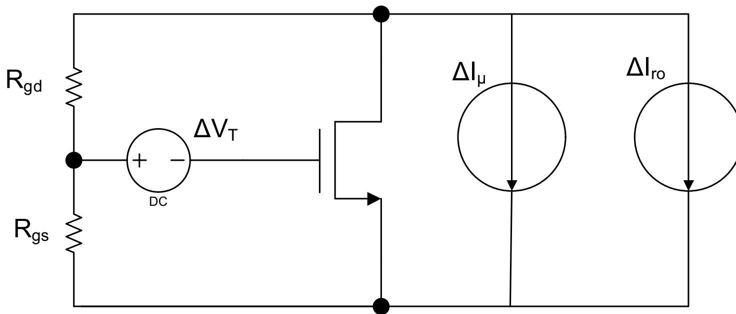


FIGURE 7.29
Equivalent model of MOS transistor considering all aging mechanisms.

Circuit netlist
Compact model parameters
SPICE simulator
Transient analysis and operating point analysis (Fresh Circuit)
Lifetime and age calculation
Aged compact model
SPICE simulator
Transient analysis and operating point analysis (Aged Circuit)
Comparison

FIGURE 7.30
Simplified CAD flow of the simulation of the reliability of an analog circuit.

is shown in Fig. 7.29 [124]. Here R_{gd} and R_{gs} are the time-varying equivalent gate-to-drain and gate-to-source resistances respectively, which model the TDDB effects. The effects of bias temperature instability and hot carrier effects on the threshold voltage and drain currents are modeled through additional current sources.

A simplified design flow for simulating the reliability of an analog circuit is shown in Fig.7.30 [83, 115, 118]. A fresh circuit simulation is done with the fresh and original device models using a circuit simulator such as SPICE. Saturation current I_{DSsat}, the threshold voltage V_T or the transconductance g_m are generally used as degradation monitoring parameters. Normally the stress time resulting in 10% decrease of one of these degradation monitoring parameters is arbitrarily set as the device lifetime. Using the currents obtained from the simulation, other reliability parameters and user defined reliability models, the AGE of the device is calculated. The AGE quantity is directly related to the amount of degradation that takes place. For HCI estimation, the AGE parameter is defined as [115]

$$AGE = \frac{I_{DS}}{WH}\left(\frac{I_{sub}}{I_{DS}}\right)^m .t \tag{7.59}$$

where H is a constant which depends upon the dielectric technology, m is also technology dependent parameters. The values of these two parameters are determined from experiments. W is the channel width. With this the I_{DSsat} can be written as

$$I_{DSsat}(t) = I_{DSsat}(t=0) \times [1 - AGE^n] \tag{7.60}$$

Some of the key compact model parameters such as $VTH0, U0$ are expressed as functions of the HCI and NBTI aging. This forms the aged compact model. An aged SPICE model for each transistor based on its particular degradation from circuit operation is generated. A second pass of the circuit simulation then has to be performed in order to obtain the aged characteristics of the circuit. It is therefore observed that generating the degraded models based on reliability models and circuit activities are the keys to accurate reliability simulation.

The design flow shown in Fig. 7.30 is somewhat similar to that used by the commercial tool of Cadence. The other commercial tool commonly used for reliability simulation purpose is that of Mentor Graphics.

7.11 Summary and Conclusion

This chapter presents an introductory overview of the two very critical challenges to nano-scale analog circuit design – the effects of statistical process

variability and reliability on the performances of nano-scale analog circuits. The major sources of variations in nanometer scale technology are identified. The statistical process variations are broadly classified into types − inter-die and intra-die process variations. The intra-die process variations are becoming very significant in the nano-scale domain. Among the various sources of variations, the systematic variations are manageable through advanced resolution enhancement techniques. The three major sources of random process variations are random discrete dopants, line edge roughness, and oxide thickness variations. The importance of process variations on circuit performances is so significant that a complete paradigm shift in the design methodology is required. The traditional worst case corner analysis is no longer sufficient to estimate the various sources of variations. The Monte Carlo simulation technique, though lengthy and tedious is useful for accurate statistical estimation. The mismatch of drain currents is discussed in detail and a simple model for estimating the drain current mismatch is derived.

The three important physical phenomena which critically affect the reliability of nano-scale analog circuits are the time dependent dielectric breakdown, the hot carrier injection effect, and the negative bias temperature instability. These effects have been discussed qualitatively. A simple process for estimating the reliability of circuits within the design flow is also introduced.

Bibliography

[1] A.B. Bhattacharyya. *Compact MOSFET Models for VLSI Design.* John Wiley and Sons (Asia) Pte Ltd, 2009.

[2] S.A. Aftab and M.A. Styblinski. IC Variability Minimization Using a New C_p and C_{pk} Based Variability/Performance Measure. In *Proceedings of the IEEE ISCAS*, pages 149–152, 1994.

[3] L.A. Akers. The Inverse Narrrow Width Effect. *IEEE Electron Device Letters*, EDL-7: 419–421, 1986.

[4] M.A. Alam and S. Mahapatra. A Comprehensive Model of PMOS NBTI Degradation. *Microelectronics Reliability*, Vol. 45: 71–81, 2005.

[5] P.E Allen and D.R Holberg. *CMOS Analog Circuit Design.* Oxford University Press, 2004.

[6] J.R. Amaya, J.M. de la Rosa, F.V. Fernandez, F. Medeiro, R. del Rio, B. Perez-Verdu, and A. Rodriguez-Vazquez. High-Level Synthesis of Switched-Capacitor, Switched-Current and Continuous-Time Sigma Delta Modulator Using SIMULINK-Based Time-Domain Behavioral Models. *IEEE Trans. Circuits and Systems-I*, Vol. 52: 1795–1810, September 2005.

[7] A.J. Annema, B. Nauta, R.J. Langevelde, and H. Tuinhout. Analog Circuit in Ultra-Deep Sub-Micron CMOS. *IEEE Journal of Solid-State Circuits*, Vol. 40(1): 132–143, January 2005.

[8] B.A.A. Antao. Trends in CAD of Analog ICs. *Proceedings of the IEEE Circuits and Devices*, pages 31–41, September 1996.

[9] B.A.A. Antao and A.J. Broderson. ARCHSIM: Behavioral Simulation for Analog System Design Verification. *IEEE Trans. VLSI Systems*, Vol. 3: 417–429, September 1995.

[10] A. Asenov, S. Kaya, and A.R. Brown. Intrinsic Parameter Fluctuations in Decananometer MOSFETs Introduced by Gate Line Edge Roughness. *IEEE Trans. Electron Devices*, Vol. 50: 1254–1260, 2003.

[11] A. Asenov, S. Kaya, and J.H. Davis. Intrinsic Threshold Voltage Fluctuations in Decanano MOSFETs Due to Local Oxide Thickness Variations. *IEEE Trans. Electron Devices*, Vol. 49: 112–119, 2002.

[12] M. Avci, M.Y. Babac, and T. Yildirim. Neural Network-Based MOSFET Channel Length and Width Decision Method for Analogue Integrated Circuits. *International Journal of Electronics*, Vol. 92: 281–293, May 2005.

[13] G. Baccarani, M.R. Wordeman, and R.H. Dennard. Generalized Scaling Theory and Its Application to a 1/4 Micrometer MOSFET Design. *IEEE Trans. Electron Devices*, 31: 452–462, 1984.

[14] P.K. Bandopadhyay. Moore's Law Governs the Silicon Revolution. *Proceedings - IEEE*, 86: 78–81, 1998.

[15] M. Barros, J. Guilherme, and N. Horta. Analog Circuits Optimization Based on Evolutionary Computation Techniques. *Integration the VLSI Journal*, 43: 136–155, January 2010.

[16] F.De. Bernardinis, M.I. Jordan, and A. Sangiovanni Vincentelli. Support Vector Machines for Analog Circuit Performance Representation. In *Proceedings of DAC*, pages 964–969, June 2003.

[17] M. Bohr. The New Era of Scaling in an SoC World. In *Proceedings of ISSCC*, pages 23–28, 2009.

[18] D. Boolchandani and V. Sahula. Exploring Efficient Kernel Functions for Support Vector Machine-Based Feasibility Models for Analog Circuits. *International Journal for Design, Analysis and Tools for Circuits and Systems*, Vol. 1: 1–8, June 2001.

[19] G.P. Box, W.G. Hunter, and J.S. Hunter. *Statistics for Experimenters: An Introduction to Design, Analysis and Model Building.* Wiley, New York, 1978.

[20] R.K. Brayton, G.D. Hachtel, and A.S. Vincentelli. A Survey of Optimization Techniques for lntegrated Circuit Design. *Proceedings of the IEEE*, Vol. 69: 1334–1362, October 1981.

[21] B.H. Calhoun et al. Digital Circuit Design Challenges and Opportunities in the Era of Nanoscale CMOS. *Proceedings of the IEEE*, 96: 343–365, February 2008.

[22] K.M. Cao et al. BSIM4 Gate Leakage Model Including Source-Drain Partition. In *Proceedings of IEEE Electron Device Meeting*, 2000.

[23] M. Chan, K.Y. Hui, C. Hu, and P.K. Ko. A Robust and Physical BSIM3 Non-Quasi-Static Transient and AC Small-Signal Model for Circuit Simulation. *IEEE Trans. Electron Devices*, Vol. 45: 834–841, 1998.

[24] H. Chang, L. Cooke, M. Hunt, A. Martin, G. McNelly, and L. Todd. *Surviving the SOC Revolution- A Guide to Platform-Based Design.* Norwell, MA: Kluwer, 1999.

[25] H. Chang et al. A Top-Down Constraint-Driven Design Methodology for Analog Integrated Circuits. In *Proceedings of CICC*, September 1992.

[26] P.K. Chatterjee, W.R. Hunter, T.C. Holloway, and Y.T. Lin. The Impact of Scaling Laws on the Choice of n-Channel or p-Channel for MOS VLSI. *IEEE Electron Device Letters*, 1: 220–223, 1984.

[27] D. Chattopadhyay and P.C. Rakshit. *Electricity and Magnetism*. Books and Allied (P) Ltd., India, 2003.

[28] B. Cheng et al. Statistical-Variability Compact-Modeling Strategies for BSIM4 and PSP. *IEEE Design & Test of Computers*, March/April 2010: 26–35, 2010.

[29] Y. Cheng, M.J. Deen, and C-H Chen. MOSFET Modeling for RF IC Design. *IEEE Trans. Electron Devices*, Vol. 52: 1286–1303, 2005.

[30] Y. Cheng and C. Hu. *MOSFET Modeling and BSIM3 User's Guide*. Kluwer Academic Publishers, 2002.

[31] C.C. Chiang and J. Kawa. *Design for Manufacturability and Yield for Nano-Scale CMOS*. Springer, 2007.

[32] J.P. Collinge. *FinFETs and Other Mult-Gate Transistors*. Springer, 2008.

[33] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to Algorithms*. Prentice Hall of India Private Limited, 2002.

[34] F.P. Cortes, E. Fabris, and S. Bampi. Analysis and Design of Comparators in CMOS 0.35 μm Technology. *Microelectronics Reliability*, Vol. 44: 657–664, 2004.

[35] J. Crols, S. Donnay, M. Steyaert, and G. Gielen. A High-Level Design and Optimization Tool for Analog and RF Receiver Front-Ends. In *Proceedings of ICCAD*, pages 550–553, 1995.

[36] J. A. Croon et al. Line Edge Roughness: Characterization, Modeling and Impact on Device Behavior. In *Proceedings of IEEE Electron Device Meeting*, pages 307–311, 2002.

[37] W. Daems, G. Gielen, and W. Sansen. Simulation-Based Generation of Posynomial Performance Models for the Sizing of Analog Integrated Circuits. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 22: 517–534, May 2003.

[38] S. Dam and P. Mandal. Modeling and Design of CMOS Analog Circuits through Hierarchical Abstraction. *Integration the VLSI Journal, http://dx.doi.org/10.1016/j.vlsi.2013.02.001*, 2013.

[39] S. Das, A. Abraham, and A. Konar. *Studies in Computational Intelligence*, chapter Particle Swarm Optimization and Differential Evolution Algorithms: Technical Analysis, Applications and Hybridization Perspectives, pages 1–38. Springer-Verlag Berlin Heidelberg, 2008.

[40] T.R. Dastidar, P.P. Chakrabarti, and P. Ray. A Synthesis System for Analog Circuits Based on Evolutionary Search and Topological Reuse. *IEEE Trans. Evolutionary Computation*, Vol. 9: 211–224, 2005.

[41] K. Deb. *MultiObjective Optimization Using Evolutionary Algorithms.* John Wiley and Sons Ltd., 2001.

[42] K. Deb. *Optimization for Engineering Design.* Prentice Hall of India, 2003.

[43] R.H. Dennard et al. Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions. *Proceedings - IEEE Journal Solid State Circuit*, 9: 256–268, 1974.

[44] D. Dhabak and S. Pandit. Adaptive Sampling Algorithm for ANN-based Performance Modeling of Nano-Scale CMOS Inverter. *Int. Journal Electrical and Electronics Engineering, World Academy of Science, Engineering and Technology*, Vol. 5: 214, 2011.

[45] D. Dhabak and S. Pandit. Performance Modeling of Nano-Scale CMOS Inverter Using Artificial Neural Network. In *Proceedings of IESPC*, pages 33–36, 2011.

[46] N. Dhanwada, A. Doboli, A. Nunez-Aldana, and R. Vemuri. Hierarchical Constraint Transformation Based on Genetic Optimization for Analog System Synthesis. *Integration the VLSI Journal*, Vol. 39: 267–290, June 2006.

[47] M. Ding and R. Vemuri. A Combined Feasibility and Performance Macromodel for Analog Circuits. In *Proceedings of DAC*, pages 63–68, June 2005.

[48] A. Doboli, N. Dhanwada, A. Nunez-Aldana, and R. Vemuri. A Two-Layer Library-Based Approach to Synthesis of Analog Systems from VHDL-AMS Specifications. *ACM Trans. Design Automation of Electronic Systems*, Vol. 9: 238–271, April 2004.

[49] A. Doboli and R. Vemuri. Exploration-Based High-Level Synthesis of Linear Analog Systems Operating at Low/Medium Frequencies. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 22: 1556–1567, November 2003.

[50] N. Dong and J. Roychowdhury. Piecewise Polynomial Model Order Reduction. In *IEEE DAC*, 2003.

[51] P.G. Drennan, M.L. Kniffin, and D.R. Locascio. Implications of Proximity Effects for Analog Design. In *Proceedings of IEEE CICC*, 2006.

[52] P.G. Drennan and C.C. McAndrew. Understanding MOSFET Mismatch for Analog Design. *IEEE Journal Solid State Circuits*, Vol. 38: 450–456, March 2003.

[53] K. Duan, S.S. Keerthi, and A.N. Poo. Evaluation of Simple Performance Measures for Tuning SVM Hyperparameters. *Neurocomputing*, Vol. 51: 41–59, 2003.

[54] B. Dubois, J-B. Kammerer, L. Hebrard, and F. Braun. Modelling of Hot-Carrier Degradation and Its Application for Analog Design for Reliability. *Microelectronics Journal*, pages 1274–1280, 2009.

[55] M.V. Dunga et al. *BSIM 4.6.0 MOSFET Model-User's Manual.* Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, 2006.

[56] R.W. Dutton and A.J. Strojwas. Perspective on Technology and Technology-Driven CAD. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 19: 1544–1560, December 200.

[57] C.C. Enz and Y. Cheng. MOS Transistor Modeling for RF IC Design. *IEEE Journal Solid State Circuits*, Vol. 35: 186–201, 2000.

[58] P. Feldmann and R.W. Freund. Efficient Linear Circuit Analysis by Pade Approximation via the Lanczos Process. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 14: 639–649, May 1995.

[59] K. Francken and Georges. G.E. Gielen. A High-Level Simulation and Synthesis Environment for Sigma-Delta Modulators. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 22: 1049–1061, August 2003.

[60] G. Gielen et al. Emerging Yield and Reliability Challenges in Nanometer CMOS Technologies. In *Proceedings of DATE*, 2008.

[61] G. Gielen and W. Sansen. *Symbolic Analysis for Automated Design of Analog Integrated Circuits.* Kluwer Academic Publishers, 1991.

[62] G. Gielen, H. Walscharts, and W. Sansen. OPTIMAN: Analog Circuit Design Optimization Based on Symbolic Simulation and Simulated Annealing. *IEEE Journal of Solid State Circuits*, Vol. 25: 707–713, June 1990.

[63] G. Gielen, P. Wambacq, and W. Sansen. Symbolic Analysis Methods and Applications for Analog Circuits: A Tutorial Overview. *Proceedings of the IEEE*, Vol. 82: 287–304, February 1994.

[64] G. Gielen. Modeling and Analysis Techniques for System-Level Architectural Design of Telecom Front-Ends. *IEEE Trans. MTT*, Vol. 50: 360–368, January 2002.

[65] G. Gielen. CAD Tools for Embedded Analogue Circuits in Mixed Integrated Systems on Chip. *IEE Proc.-Comput.Digit.Tech.*, Vol. 152: 317–332, May 2005.

[66] G. Gielen and R.A. Rutenbar. Computer-Aided Design of Analog and Mixed-Signal Integrated Circuits. *Proceedings of the IEEE*, Vol. 88: 1825–1852, December 2000.

[67] D.E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, Mass.: Addison-Wesley, 1989.

[68] H. Graeb, S. Zizala, J. Eckmueller, and K. Antreich. The Sizing Rules Method for Analog Integrated Circuit Design. In *Proceedings of IEEE/ACM ICCAD*, pages 343–349, 2001.

[69] H. E. Graeb. *Analog Design Centering and Sizing*. Springer, 2007.

[70] P.R. Gray, P.J. Hurst, S.H. Lewis, and R.G. Meyer. *Analysis and Design of Analog Integrated Circuit*. John Wiley & Sons Inc, fourth edition, 2001.

[71] B.S. Grewal and J.S. Grewal. *Higher Engineering Mathematics*. Khanna Publishers, 2001.

[72] S.R. Gunn, M. Brown, and K.M. Bossley. Network Performance Assessment for Neuro Fuzzy Data Modeling. In *Intelligent Data Analysis, LNCS*, pages 313–323, 1997.

[73] M. Haartman and M. Ostling. *Low Frequency Noise in Advanced MOS Devices*. Springer, 2007.

[74] E. Hansen and G.W. Walster. *Global Optimization Using Interval Analysis*. Marcel Dekker, 2004.

[75] R. Harjani. Designing Mixed-Signal ICs. In *IEEE Spectrum*, pages 49–51, November 1992.

[76] R. Harjani and J. Shao. Feasibility and Performance Region Modeling of Analog and Digital Circuits. *Analog Integrated Circuits and Signal Processing*, Vol. 10: 23–43, August 1996.

[77] S. Haykin. *Neural Networks and Learning Machines*. Pearson Education, 2009.

[78] M.M Hershenson, S.P. Boyd, and T.H. Lee. Optimal Design of a CMOS Op-Amp via Geometric Programming. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 20: 1–21, January 2001.

[79] T.B. Hook et al. Lateral Ion Implant Straggle and Mask Proximity Effect. *IEEE Trans. Electron Devices*, Vol. 50: 1946–1951, 2003.

[80] K. Hornik, M. Stinchcombe, and H. White. Multilayer Feedforward Networks Are Universal Approximators. *Neural Networks*, Vol. 2: 359–366, 1989.

[81] E. Horowitz, S. Sahni, and S. Rajasekaran. *Computer Algorithms.* Galgotia Publications Pvt. Ltd, 2003.

[82] K.K.L. Hseuh, J.J. Sanchez, T.A. Demassa, and L.A. Akers. Inverse-Narrow-Width Effects and Small-Geometry MOSFET Threshold Voltage Model. *IEEE Trans. Electron Devices*, Vol. 35: 614–622, 1988.

[83] C. Hu. IC Reliability Simulation. *IEEE Journal Solid State Circuits*, Vol. 27: 241–246, March 1992.

[84] C. Hu et al. Hot-Electron-Induced MOSFET Degradation-Model, Monitor, and Improvement. *IEEE Trans. Electron Devices*, Vol. ED-32: 375–384, February 1985.

[85] C.C. Hu. *Modern Semiconductor Devices for Integrated Circuits.* Pearson Education, Dorling Kindersley India Pvt. Ltd., First edition, 2010.

[86] C.Y. Huang, C.Y. Lai, and K.T. Cheng. *Electronic Design Automation.* Elsevier, Morgan Kaufman, 2009.

[87] J.H. Huang et al. A Physical Model for MOSFET Output Resistance. In *IEDM Tech Digest*, pages 21.5.1–21.5.4, 1992.

[88] K.K. Hung, P.K. Ko, C. Hu, and Y.C. Cheng. A Physics-Based MOSFET Noise Model for Circuit Simulators. *IEEE Trans. Electron Devices*, Vol. 37: 1323–1333, 1990.

[89] K.K. Hung, P.K. Ko, C. Hu, and Y.C. Cheng. A Unified Model for the Flicker Noise in Metal-Oxide Semiconductor Field-Effect Transistors. *IEEE Trans. Electron Devices*, Vol. 37: 654–665, 1990.

[90] ITRS. itrs.net.

[91] H. Iwai. CMOS Technology 2010 and Beyond. *Proceedings - IEEE Journal Solid State Circuit*, Vol. 34: 357–366, 1999.

[92] J. Jaffari and M. Anis. On Efficient LHS-Based Yield Analysis of Analog Circuits. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 30: 159–163, January 2011.

[93] P.G.A. Jespers. *The gm/ID Methodology, A Sizing Tool for Low-Voltage Analog CMOS Circuits.* Springer, 2010.

[94] X. Jin et al. An Effective Gate Resistance Model for CMOS RF and Noise Modeling. In *Proceedings of IEEE Electron Device Meeting*, 1998.

[95] G.H. John and P. Langley. Static versus Dynamic Sampling for Data Mining. In *Proceedings of Knowledge Discovery and Data Mining*, 1996.

[96] H.K. Jung and S. Dimitrijev. Analysis of Subthreshold Carrier Transport for Ultimate DG MOSFET. *IEEE Trans. Electron Devices*, Vol. 53: 685–691, 2006.

[97] N. Kahraman and T. Yildirim. Technology Independent Circuit Sizing for Fundamental Analog Circuits Using Artificial Neural Networks. In *Proceedings of PRIME*, pages 1–4, 2008.

[98] J. Kennedy. Small Worlds and Mega-Minds: Effects of Neighborhood Topology on Particle Swarm Performance. In *Proceedings of the 1999 Congress of Evolutionary Computation, Vol. 3, IEEE Press, New York*, 1999.

[99] J. Kennedy, R. Eberhart, and Y. Shi. *Swarm Intelligence.* Morgan Kaufmann, 2001.

[100] J. Kennedy and R.C. Eberhart. Particle Swarm Optimization. In *Proceedings of IEEE International Conference on Neural Networks*, 1995.

[101] T. Kiely and G. Gielen. Performance Modeling of Analog Integrated Circuits Using Least-Squares Support Vector Machines. In *Proceedings of DATE*, pages 448–453, Feb 2004.

[102] H-W. Kim, J-Y. Lee, J. Shin, S-G. Woo, H-K. Cho, and J-T. Moon. Experimental Investigation of the Impact of LWR on Sub-100-nm Device Performance. *IEEE Trans. Electron Devices*, Vol. 51: 1984–1988, 2004.

[103] E. Kougianos and S.P. Mohanty. A Comparative Study on Gate Leakage and Performance of High-κ Nano-CMOS Logic Gates. *International Journal of Electronics*, Vol. 27: 985–1005, 2010.

[104] K. Kuhn and others. Managing Process Variation in Intel's 45nm CMOS Technology. *Intel Technology Journal*, Vol. 12: 92–110, 2008.

[105] L. Kuipers and H. Niederreiter. *Uniform Distribution of Sequences.* Dover Publications, 1974.

[106] K. Kundert, H. Chang, D. Jefferies, G. Lamant, E. Malavasi, and F. Sendig. Design of Mixed-Signal Systems on Chip. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 19: 1561 – 1571, December 2000.

[107] K. Kundert. Principles of Top-Down Mixed-Signal Design. www.designers-guide.org, March 2005.

[108] R. Langevelde et al. Gate Current: Modeling, ΔL Extraction and Impact on RF Performance. In *Proceedings of IEDM Tech Digest*, pages 289–292, 2001.

[109] E. Lauwers and G. Gielen. Power Estimation Methods for Analog Circuits for Architectural Exploration of Integrated Systems. *IEEE Trans. VLSI Systems*, Vol. 10: 155–162, April 2002.

[110] W-C. Lee and C. Hu. Modeling CMOS Tunneling Currents through Ultrathin Gate Oxide Due to Conduction-and Valence-Band Electron and Hole Tunneling. *IEEE Trans. Electron Devices*, Vol. 48: 1366–1373, 2001.

[111] N.S. Levenson, M.D. and Viswanathan and R.A. Simpson. Improving Resolution in Photolithography with a Phase-Shifting Mask. *IEEE Trans. Electron Devices*, ED-29: 1828–1836, 1982.

[112] L.L. Lewyn, T. Ytterdal, C. Wulff, and K. Martin. Analog Circuit Design in Nanoscale CMOS Technologies. *Proceedings of the IEEE*, Vol. 97: 1687–1714, October 2009.

[113] E.H. Li, K.M. Hong, Y.C. Cheng, and K.Y. Chan. The Narrow Channel Effect in MOSFETs with Semi-Recessed Oxide Structures. *IEEE Trans. Electron Devices*, Vol. 37: 692–701, 1990.

[114] H. Li and Y.X. Zhang. An Algorithm of Soft Fault Diagnosis for Analog Circuit Based on the Optimized SVM by GA. In *Proceedings of ICEMI*, pages 4–1023 – 4–1027, 2009.

[115] X. Li et al. Deep Submicron CMOS Integrated Circuit Reliability Simulation with SPICE. In *Proceedings of IEEE ISQED'05*, 2005.

[116] X. Li, P. Gopalkrishnan, Y. Xu, and L.T. Pileggi. Robust Analog/RF Circuit Design with Projection-Based Performance Modeling. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 26: 2–15, January 2007.

[117] W. Liu and M-C. Chang. Transistors Transient Studies Including Transcapacitive Current and Distributive Gate Resistance for Inverter Circuits. *IEEE Trans. Circuits and Systems-I*, Vol. 45: 416–422, 1998.

[118] Z. Liu, B.W. McGaughy, and J. Z. Ma. Design Tools for Reliability Analysis. In *Proceedings of ACM/IEEE Design Automation Conference*, 2006.

[119] Z.H. Liu et al. Threshold Voltage Model for Deep Sub-Micrometer MOSFETs. *IEEE Trans. Electron Devices*, Vol. 40: 86–95, January 1993.

[120] P. Malcovati, S. Brigati, F. Francesconi, F. Maloberti, P. Cusinato, and A. Baschirotto. Behavioral Modeling of Switched-Capacitor Sigma Delta Modulators. *IEEE Trans. Circuits and Systems-I*, Vol. 50: 352–364, March 2003.

[121] P. Mandal and V. Visvanathan. CMOS Op-Amp Sizing Using a Geometric Programming Formulation. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 20: 22–38, January 2001.

[122] S.K. Mandal, S. Sural, and A. Patra. ANN-and PSO-Based Synthesis of on-Chip Spiral Inductors for RF ICs. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 27: 188–192, January 2008.

[123] G. Manganaro, S.U. Kwak, S. Cho, and A. Pulinchery. A Behavioral Modeling Approach to the Design of a Low Jitter Clock Source. *IEEE Trans. Circuits and Systems-II*, Vol. 50: 804–814, November 2003.

[124] E. Maricau and G. Gielen. Computer-Aided Analog Circuit Design for Reliability in Nanometer CMOS. *IEEE Trans. Emerging and Selected Topics in Circuits and Systems*, Vol. 1: 1576–1580, March 2011.

[125] E. Maricau, P.D. Wit, and G. Gielen. An Analytical Model for Hot Carrier Degradation in Nanoscale CMOS Suitable for the Simulation of Degradation in Analog IC Applications. *Microelectronics Reliability*, pages 1576–1580, 2008.

[126] G. Maricau, and G. Gielen. NBTI Model for Analogue IC Reliability Simulation. *Electronics Letters*, Vol. 46: 1–2, 2010.

[127] E. Martens and G. Gielen. Top-Down Heterogeneous Synthesis of Analog and Mixed-Signal Systems. In *Proceedings of DATE*, pages 1–6, March 2006.

[128] F. Medeiro, B. Perez-Verdu, A. Rodriguez-Vazquez, and J.L. Huertas. A Vertically Integrated Tool for Automated Design of Sigma-Delta Modulators. *IEEE J. of Solid State Circuits.*, Vol. 30: 762–772, July 1995.

[129] F. Medeiro, R. Rodriguez-Macias, F. Fernandez, R-C Dominguez, J. Huertas, and A. Rodriguez-Vazquez. Global Design of Analog Cells Using Statistical Optimization Techniques. *Analog Integr. Circ. Sig Process*, Vol. 6: 179–195, 1994.

[130] J. Meyer. MOS Models and Circuit Simulation. *RCA Review*, Vol. 32: 42–63, 1971.

[131] J. Mukhopadhyay and S. Pandit. Modeling and Design of a Nano Scale CMOS Inverter for Symmetric Switching Characteristics. *VLSI Design*, Article ID 505983:13 pages, 2012.

[132] R.S. Muller, T.I. Kamins, and M. Chan. *Device Electronics for Integrated Circuits.* John Wiley and Sons, Third edition, 2003.

[133] B. Murmann, P. Nikaeen, D.J. Connelly, and Dutton. R.J. Impact of Scaling on Analog Performance and Associated Modeling Needs. *IEEE Trans. Electron Devices*, Vol. 53: 2160–2167, 2006.

[134] B.T. Murphy. Cost-Size Optima of Monolithic Integrated Circuits. *Proceedings of the IEEE*, Vol. 52: 1537–1545, 1964.

[135] A.M. Niknejad and R.G. Meyer. Analysis, Design and Optimization of Spiral Inductors and Transformers for Si RF ICs. *IEEE Journal Solid State Circuits*, Vol. 34: 1470–1481, October 1998.

[136] A. Nunez and R. Vemuri. An Analog Performance Estimator for Improving the Effectiveness of CMOS Analog Systems Circuit Synthesis. In *Proceedings of DATE*, pages 406–411, 1999.

[137] E. Ochotta, R. Rutenbar, and R. Carley. Synthesis of High Performance Analog Circuits in ASTRX/OBLX. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 15: 273–294, March 1996.

[138] C. Pacha et al. Impact of STI-Induced Stress, Inverse Narrow Width Effect and Statistical V_T Variations on Leakage Currents in 120nm CMOS. In *Proceedings of IEEE ESSDERC*, pages 397–400, 2004.

[139] S. Pandit, S.K. Bhattacharya, C.R. Mandal, and A. Patra. A Fast Exploration Procedure for Analog High-Level Specification Translation. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 27: 1493–1497, August 2008.

[140] S. Pandit, C. Mandal, and A. Patra. A Methodology for Generation of Performance Models for the Sizing of Analog High-Level Topologies. *VLSI Design*, Vol. 2011 Article ID 475952: 1–17, 2011.

[141] S. Pandit, C.R. Mandal, and A. Patra. Systematic Methodology for High-Level Performance Modeling of Analog Systems. In *Proceedings of Int. Conf. on VLSI Design*, pages 361–366, 2009.

[142] S. Pandit and C.K. Sarkar. Analytical Modeling of Inverse Narrow Width Effect for Narrow Channel STI MOSFETs. *International Journal of Electronics*, Vol. 99: 361–372, 2012.

[143] M.J.M Pelgrom, A.C.J. Duinmaijer, and A.P.G. Welbers. Matching Properties of MOS Transistors. *IEEE Journal Solid State Circuits*, Vol. 24: 1433–1440, 1989.

[144] L.T. Pillage and R.A. Rohrer. Asymptotic Waveform Evaluation for Timing Analysis. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 9: 352–366,, April 1990.

[145] G.V. Plas, J. Vandenbussche, G. Gielen, and W. Sansen. EsteMate: A Tool for Automated Power and Area Estimation in Analog Top-Down Design and Synthesis. In *Proceedings of CICC*, pages 139–142, May 1997.

[146] P.W.C Prasad and A. Beg. Investigating Data Pre-Processing Methods for Circuit Complexity Models. *Expert Systems with Applications*, Vol. 36: 519–526, 2009.

[147] F. Provost, D. Jensen, and T. Oates. Efficient Progressive Sampling. In *Proceedings of Knowledge Discovery and Data Mining*, pages 23–32, 1999.

[148] A. Ratnaweera, S.K. Halgamuge, and H.C. Watson. Self-Organizing Hierarchical Particle Swarm Optimizer with Time-Varying Acceleration Coefficients. *IEEE Trans. Evolutionary Computation*, Vol. 8: 240–255., 2004.

[149] G. Reklaitis, A. Ravindran, and K.M. Ragsdell. *Engineering Optimization- Methods and Applications*. Wiley, New York, 1983.

[150] X. Ren and T. Kazmierski. Performance Modeling and Optimization of RF Circuits Using Support Vector Machines. In *Proceedings of MIXDES*, pages 317–321, 2007.

[151] C.-J. Richard Shi and X.D. Tan. Canonical Symbolic Analysis of Large Analog Circuits with Determinant Decision Diagrams. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 19: 1–18, January 2000.

[152] C.-J. Richard Shi and X.D. Tan. Efficient Approximation of Symbolic Expressions for Analog Behavioral Modeling and Analysis. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 23: 907–918, 2004.

[153] J. Robertson. High Dielectric Constant Gate Oxides for Metal Oxide Si Transistors. *Reports on Progress in Physics, IOP*, Rep. Prog. Phys. 69: 327–396, 2006.

[154] K. Roy, S. Mukhopadhyay, and H.M. Mahmoodi. Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits. *Proceedings of the IEEE*, Vol. 91: 305–327, February 2003.

[155] J. Roychowdhury. Reduced-Order Modelling of Time-Varying Systems. *IEEE Trans. Circuits and Systems-II*, Vol. 46, November 1999.

[156] J. Roychowdhury. Algorithmic Macromodeling Methods for Mixed-Signal Systems. In *Proceedings of VLSIID*, pages 63–68, 2004.

[157] J. Roychowdhury. An Overview of Automated Macromodelling Techniques for Mixed-Signal Systems. In *Custom Integrated Circuits Conference*, 2004.

[158] G. Rubio, H. Pomares, I. Rojas, and L.J. Herrera. A Heuristic Method for Parameter Selection in LS-SVM: Application to Time Series Prediction. *International Journal of Forecasting*, Vol. 27: 725–739, 2011.

[159] J. Ruiz-Amaya, J.M. de la Rosa, M. Delgado-Restituto, and A. Rodriguez-Vazquez. Behavioral Modeling, Simulation and High-Level Synthesis of Pipeline A/D Converters. In *Proceedings of ISCAS*, pages 5609– 5612, May 2005.

[160] R.A. Rutenbar, G. Gielen, and J.Roychowdhury. Hierarchical Modeling, Optimization, and Synthesis for System-Level Analog and RF Designs. *Proceedings of the IEEE*, Vol. 95: 640–669, March 2007.

[161] Y. Saad. *Iterative Methods for Sparse Linear Systems.* PWS Boston, 1996.

[162] S.K. Saha. Managing Technology CAD for Competitive Advantage: An Efficient Approach for Integrated Circuit Fabrication Technology Development. *IEEE Trans. Engineering Management*, 46: 221–229, May 1999.

[163] S.K. Saha. Modeling Process Variability in Scaled CMOS Technology. *IEEE Design and Test of Computers*, Vol. 27: 8–16, March/April 2010.

[164] W.M.C Sansen. *Analog Design Essentials.* Springer, 2006.

[165] S.S. Sapatnekar. Overcoming Variations in Nanometer-Scale Technologies. *IEEE Trans. Emerging and Selected Topics in Circuits and Systems*, Vol. 1: 5–18, March 2011.

[166] A. Satyanarayana and I. Davidson. A Dynamic Adaptive Sampling Algorithm for Real World Applications: Finger Print Recognition and Face Recognition. In *Proceedings of ISMIS*, pages 631–640. Springer Verlag, 2005.

[167] M. Schetzen. *The Volterra and Wiener Theories of Nonlinear Systems.* John Wiley, 1980.

[168] D.K. Schroder. Negative Bias Temperature Instability: What Do We Understand? *Microelectronics Reliability*, Vol. 47: 841–852, 2007.

[169] K.F. Schuegraf and C. Hu. Hole Injection SiO_2 Breakdown Model for Very Low Voltage Lifetime Extrapolation. *IEEE Trans. Electron Devices*, Vol. 41: 761–767, May 1994.

[170] Y. Shi and R.C. Eberhart. A Modified Particle Swarm Optimiser. In *IEEE International Conference on Evolutionary Computation*. IEEE Press, 1998.

[171] Y. Shi and R.C. Eberhart. Empirical Study of Particle Swarm Optimization. In *Proceedings of the 1999 Congress of Evolutionary Computation*. IEEE Press, 1999.

[172] R. Shrivastava and K. Fitzpatrick. A Simple Model for the Overlap Capacitance of a VLSI MOS Device. *IEEE Trans. Electron Devices*, ED-29: 1870, 1980.

[173] F. Silveira, D. Flandre, and P.G.A. Jesper. A g_m/I_D-Based Methodology for the Design of CMOS Analog Circuits and Its Application to the Synthesis of a Silicon-on-Insulator. *IEEE Journal Solid State Circuits*, Vol. 31: 1314–1319, 1996.

[174] B.De. Smedt and G. Gielen. WATSON: Design Space Boundary Exploration and Model Generation for Analog and RF IC Design. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 22: 213–224, February 2003.

[175] A. Somani, P.P. Chakrabarti, and A. Patra. An Evolutionary Algorithm-Based Approach to Automated Design of Analog and RF Circuits Using Adaptive Normalized Cost Functions. *IEEE Trans. Evolutionary Computation*, Vol. 11: 336–353, 2007.

[176] J.H. Stathis. Percolation Models for Gate Oxide Breakdown. *Journal of Applied Physics*, Vol. 86: 5757–5766, 1999.

[177] J.H. Stathis and S. Zafar. The Negative Bias Temperature Instability in MOS Devices: A Review. *Microelectronics Reliability*, Vol. 46: 270–286, 2006.

[178] G. Stehr, H. Graeb, and K. Antreich. Feasibility Regions and Their Significance to the Hierarchical Optimization of Analog and Mixed-Signal Systems. *Int. Ser. Numer. Math*, Vol. 146: 167–184, 2003.

[179] G. Stehr, H. Graeb, and K. Antreich. Performance Trade-Off Analysis of Analog Circuits by Normal-Boundary Intersection. In *Proceedings of DAC*, pages 958–963, 2003.

[180] G. Stehr, H. Graeb, and K. Antreich. Analog Performance Space Exploration by Fourier–Motzkin Elimination with Application to Hierarchical Sizing. In *Proceedings of ICCAD*, pages 847–854, November 2004.

[181] G. Stehr, H. Graeb, and K. Antreich. Analog Performance Space Exploration by Normal-Boundary Intersection and by Fourier–Motzkin Elimination. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 26: 1733–1748, Oct. 2007.

[182] P.A. Stolk, F.P. Widdershoven, and D.B.M Klaasse. Modeling Statistical Dopant Fluctuations in MOS Transistors. *IEEE Trans. Electron Devices*, Vol. 45: 1960–1971, 1998.

[183] G. Strang. *Linear Algebra and Its Applications.* Harcourt Brace Jovanovich, Publishers, San Diego, 1988.

[184] B.G. Streetman and S.K. Banerjee. *Solid State Electronic Devices.* Pearson Education, sixth edition, 2008.

[185] J.H. Suehle. Ultrathin Gate Oxide Reliability: Physical Models, Statistics, and Characterization. *IEEE Trans. Electron Devices*, Vol. 49: 958–971, June 2002.

[186] J.A.K. Suykens et al. LS-SVM Toolbox. http://www.esat.kuleuven.ac.be/sista/lssvmlab, February 2003.

[187] J.A.K Suykens, T.V Gestel, J.D Brabanter, B.D Moor, and V. Joos Vandewalle. *Least Squares Support Vector Machines.* World Scientific, 2002.

[188] C. Svensson and J.J. Wikner. Power Consumption of Analog Circuits: A Tutorial. *Analog Integrated Circuits and Signal Processing*, Vol. 65: 171–184, 2010.

[189] S.M. Sze and K.K. Ng. *Physics of Semiconductor Devices.* Wiley, India, Third edition, 2007.

[190] K. Takeuchi, T. Tatsumi, and A. Furukawa. Channel Engineering for the Reduction of Random-Dopant-Placement-Induced Threshold Voltage Fluctuation. In *Proceedings of IEEE Electron Device Meeting*, pages 841–844, 1997.

[191] Y. Taur et al. CMO Scaling into the Nanometer Regime. *Proceedings of the IEEE*, Vol. 85: 486–504, 1997.

[192] Y. Taur and T.H. Ning. *Fundamentals of Modern VLSI Devices.* Cambridge Univ. Press, 1998.

[193] S.F. Tin, A.A. Osman, K. Mayaram, and C. Hu. A Simple Subcircuit Extension of the BSIM3v3 Model for CMOS RF Design. *IEEE Journal Solid State Circuits*, Vol. 35: 612–624, 2000.

[194] Y. Tsividis and C.C. McAndrew. *Operation and Modeling of the MOS Transistor.* Oxford University Press, Second edition, 2010.

[195] E.P. Vandamme and L.K.J. Vandamme. Critical Discussion on Unified 1/f Noise Models for MOSFETs. *IEEE Trans. Electron Devices*, Vol. 47: 2146–2152, 2000.

[196] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.

[197] V. Vapnik. *Statistical Learning Theory.* Springer, New York, 1998.

[198] P. Venkataraman. *Applied Optimization with MATLAB Programming.* John Wiley & Sons Inc, 2002.

[199] X. Wang et al. Statistical Threshold-Voltage Variability in Scaled Decananometer Bulk HKMG MOSFETs: A Full-Scale 3-D Simulation Scaling Study. *IEEE Trans. Electron Devices*, Vol. 58: 2293–2301, 2011.

[200] D. Ward and R. Dutton. A Charge-Oriented Model for MOS Transistor Capacitances. *IEEE Journal Solid State Circuits*, Vol. SC-13: 703–708, 1978.

[201] Y. Wei, A. Doboli, and H. Tang. Systematic Methodology for Designing Reconfigurable Sigma-Delta Modulator Topologies for Multimode Communication Systems. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 26: 480–496, March 2007.

[202] P.H. Woerlee et al. RF-CMOS Performance Trends. *IEEE Trans. Electron Devices*, Vol. 48: 1776–1782, 2001.

[203] G. Wolfe and R. Vemuri. Extraction and Use of Neural Network Models in Automated Synthesis of Operational Amplifiers. *IEEE Trans. Computer Aided Design of Integrated Circuits and Systems*, Vol. 22: 198–212, February 2003.

[204] J. Wu, G.K. Fedder, and L.R. Carley. A Low-Noise Low-Offset Capacitive Sensing Amplifier for a 50-μg/$\sqrt{Hz}$ Monolithic CMOS MEMS Accelerometer. *IEEE J. of Solid-State Circuits*, Vol. 39: 722–730, May 2004.

[205] H. Yang et al. Current Mismatch Due to Local Dopant Fluctuations in MOSFET Channel. *IEEE Trans. Electron Devices*, Vol. 50: 2248–2254, November 2003.

[206] L.D. Yau. A Simple Theory to Predict the Threshold Voltage of Short Channel IGFETs. *Solid State Electronics*, 17: 1059–1063, 1974.

[207] Y. Ye, F. Liu, M. Chen, S. Nassif, and Y. Cao. Statistical Modeling and Simulation of Threshold Variation under Random Dopant Fluctuations and Line-Edge Roughness. *IEEE Trans. VLSI Systems*, Vol. 19: 987–996, June 2011.

[208] C.P. Yue and S.S. Wong. Physical Modeling of Spiral Inductors on Silicon. *IEEE Trans. Electron Devices*, Vol. 47: 560–568, March 2000.

[209] Q.J. Zhang, K.C. Gupta, and V.K. Devabhaktuni. Artificial Neural Networks for RF and Microwave Design: From Theory to Practice. *IEEE Trans. MTT*, Vol. 51: 1339–1350, April 2003.

[210] W. Zhao and Y. Cao. New Generation of Predictive Technology Model for Sub-45 nm Early Design Exploration. *IEEE Transactions Electron Devices*, Vol. 53: 2816–2823, November 2006.

ELECTRICAL ENGINEERING

Reliability concerns and the limitations of process technology can sometimes restrict the innovation process involved in designing nano-scale analog circuits. The success of nano-scale analog circuit design requires repeat experimentation, correct analysis of the device physics, process technology, and adequate use of the knowledge database.

Starting with the basics, **Nano-Scale CMOS Analog Circuits: Models and CAD Techniques for High-Level Design** introduces the essential fundamental concepts for designing analog circuits with optimal performances. This book explains the links between the physics and technology of scaled MOS transistors and the design and simulation of nano-scale analog circuits. It also explores the development of structured computer-aided design (CAD) techniques for architecture-level and circuit-level design of analog circuits.

The book outlines the general trends of technology scaling with respect to device geometry, process parameters, and supply voltage. It describes models and optimization techniques, as well as the compact modeling of scaled MOS transistors for VLSI circuit simulation.

- Includes two learning-based methods: the artificial neural network (ANN) and the least-squares support vector machine (LS-SVM) method
- Provides case studies demonstrating the practical use of these two methods
- Explores circuit sizing and specification translation tasks
- Introduces the particle swarm optimization technique and provides examples of sizing analog circuits
- Discusses the advanced effects of scaled MOS transistors such as narrow width effects, and vertical and lateral channel engineering

Nano-Scale CMOS Analog Circuits: Models and CAD Techniques for High-Level Design describes the models and CAD techniques, explores the physics of MOS transistors, and considers the design challenges involving statistical variations of process technology parameters and reliability constraints related to circuit design.

CRC Press
Taylor & Francis Group
an informa business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK