MODELING OF NANOSCALE MOSFETS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF
ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES OF
STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Changhoon Choi

April 2002

© Copyright by Changhoon Choi 2002

All Rights Reserved

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Robert W. Dutton
(Principal Advisor)

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Krishna Saraswat
(Associate Advisor)

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Zhiping Yu

Approved for the University Committee on Graduate Studies:

Abstract

As CMOS technology dimensions are being aggressively scaled to reach a limit where device performance must be assessed against fundamental limits, nanoscale device modeling is needed to provide innovative new MOS devices as well as to understand the limits of the scaling process. This thesis describes advanced modeling of nanoscale MOSFETs from the viewpoint of device physics, which consists of three parts – gate, source/drain, and channel modeling. In the gate modeling part, MOS C–V characteristics of gate oxide in the sub 2.0 nm region are modeled with an empirical, hybrid formulation for QM corrections implemented in a 2D device simulator. The sharp decrease in capacitance for gate oxides below 2.0 nm is modeled by using a distributed RC network that includes the gate tunneling current which is calculated using a Green's function solver. Conversely, a reconstruction technique to extract the intrinsic gate capacitance, based on distorted $C - V$ curves in high leakage dielectric MOSFETs has been developed. An accurate direct tunneling model for circuit simulation is developed that incorporates an explicit surface potential model and quantum-mechanical corrections. In addition, CMOS circuit robustness in the presence of gate tunneling currents has been studied using circuit simulation, combined with a macro-circuit model of gate tunneling current and analytic estimation of the effects. Based on the simulation study, expected values of oxide thickness, needed to ensure the off-state gate leakage requirement of the ITRS roadmap, are outlined.

The source and drain modeling part describes parasitic resistance issues of the source

and drain regions in the presence of very shallow junction depths. It addresses accurate modeling of the extrinsic resistance and application in the modeling of ultra-shallow junction MOSFETs in order to optimize device performance. Analysis and optimization of device performance for an experimental process technology called Laser Thermal Process (LTP) are also discussed.

In the channel modeling part, a simple MOSFET on-current model, taking into account the off-equilibrium transport is proposed based on use of Lundstrom's carrier transport model formulation, augmented with analytical calculations that are calibrated using Hydro-Dynamic (HD) carrier transport simulations. The model shows good agreement with these simulation results for 50 nm MOSFETs.

Acknowledgments

First of all, I would like to acknowledge my adviser, Professor Robert W. Dutton. He has been a great mentor with his management skills and enthusiasm. He also has been an exceptional role model of life. He has done his utmost to help relieve concerns that his advisees have. It has been memorable that he encouraged and supported me while I was in deep depression.

I also would like to thank my associate adviser, Professor Krishna Saraswat, who gave me strong motivation in my research throughout his classwork. I am indebted to Dr. Zhiping Yu, who served on my oral exam and reading committees. His incisive comments were a great help whenever I met unsolved problems. I am also grateful to Professor Dan Boneh, who agreed to be the chairman of my oral committee in spite of his busy schedule.

I wish to thank the National Science Foundation (NSF) for support of this work through the Distributed Center of Advanced Electronics Simulations (DesCArtES). Special thanks go to Fely Barrera, Daniel Yergeau, Miho Nishi and Maria Perea for all their help. I also thank friends in the Center for Integrated Systems (CIS) and the Stanford TCAD group for their helpful discussions and friendship: Hyun-Jin Cho, Jung-Suk Goo, Kwang-Hoon Oh, Wonill Ha, Jaejune Jang, Ki-Young Nam, Moon-Jung Kim, Tae-Young Oh, SoYoung Kim, Xin-Yi Zhang, Xiaoning Qi, Michael Kwong, Yi-Chang Lu, Atsushi Kawamoto, Choshu Ito, Dr. U-In Chung, Dr. Chin-Shan Hou and prof. Ming-Jer Chen.

I would like to acknowledge many industrial people: Drs. Amr Bayoumi, Min Cao,

Paul Vande Voorde, Dieter Vook and Carlos H. Diaz of Hewlett-Packard ULSI Laboratories for providing the C–V data; Drs. Ming-Ren Lin and Bin Yu of the Advanced Micro Devices (AMD) for the data of 50–70 nm MOSFETs; Drs. P.R. Chidambaram, Charles F. Machala, Dennis Buss and Yoshio Nishi of the Texas Instruments for their help in obtaining the experimental data of polydepletion effects; Mr. Young-Kwan Park, Drs. Kyung-Ho Kim and Jeong-Taek Kong of Samsung Electronics for their firm support during my PhD program.

Specially, I am deeply indebted to my family – my mother, my brothers, and my wife, Ju-Hee, and my kids, Seung-Yeon and Seung-Woo.

Finally, I would like to devote this work to my late father's memory.

Contents

Abstract v

Acknowledgments vii

List of Tables xii

List of Figures xiii

1 Introduction **1**

1.1 Motivation 1

1.2 Challenges for CMOS Technology 5

1.3 Scope and Organization 10

2 Gate Modeling **13**

2.1 Introduction 13

2.2 Gate Capacitance Modeling 14

2.2.1 Tunneling through Silicon Dioxide 14

2.2.2 Polysilicon-Gate Depletion Effects 17

2.2.3 Quantum Effects on C–V 28

2.2.4 Anomalous C–V Behavior in Thin-Oxide MOS 30

2.2.5 QM Models of C–V and Gate Tunneling Current 34

2.2.6 Equivalent Circuit Modeling 40

2.2.7 Channel Length Dependence of Gate Current 44
2.2.8 Summary 48
2.3 C–V Reconstruction 49
2.3.1 Previous Reconstruction Method 49
2.3.2 Optimization Technique 51
2.3.3 Application Results 53
2.3.4 Summary 56
2.4 Performance-based CMOS Scaling 58
2.4.1 New CMOS delay model 58
2.4.2 Model Evaluation 60
2.4.3 Delay Optimization 65
2.4.4 Summary 67
2.5 Direct Tunneling Current Model for Circuit Simulation 69
2.5.1 Surface Potential Based Tunneling Model 69
2.5.2 Quantum-Mechanical Effects 72
2.5.3 Results and Discussions 73
2.5.4 Summary 83
2.6 Impact of Gate Tunneling Current on Circuit Performance 84
2.6.1 Introduction 84
2.6.2 Gate Current Modeling 85
2.6.3 Circuit Application 93
2.6.4 Review of ITRS Oxide Scaling Roadmap 110
2.6.5 Summary 111

3 Source and Drain Modeling **113**

3.1 Introduction 113

3.2 Shallow SDE Effects on External Resistance 114
3.2.1 Source/Drain Resistance 114
3.2.2 Calculation of External Resistance 119
3.2.3 Shallow SDE and Gate Overlap Effects 123
3.2.4 Summary 129
3.3 Analysis and Optimization of 70 nm Laser Thermal Processed MOSFETs . 130
3.3.1 Dopant Profiles of LTP and RTP 131
3.3.2 Calculation of External Resistance 136
3.3.3 Device Optimization in LTP Transistor 139
3.3.4 Summary 143

4 Channel Modeling **144**

4.1 Introduction 144
4.2 Non-local Effects in Small Devices 146
4.3 Fundamental Drain Current Model 147
4.4 Mobility and External Resistance Modeling 156
4.5 Potential and E-field Modeling 157
4.6 Summary 163

5 Conclusions **164**

5.1 Introduction 164
5.2 Gate Modeling 165
5.3 Source/Drain Modeling 166
5.4 Channel Modeling 167
5.5 Recommendations for Future Work 168

List of Tables

1.1 ITRS roadmap (2000 Edition) 8

2.1 Optimized parameters for target delays being 10, 16, and 22 ps when fanout is 1, 2, and 3, respectively (V_{dd} = 1 V, $t_{ox,phy}$ = 20 Å). 66

2.2 ITRS logic technology roadmap (2000 edition) [3] and oxide thicknesses to ensure edge direct tunneling leakage (dielectrics with $\kappa = 2\epsilon_{ox}$ are assumed for 2005–2011 node). 112

List of Figures

1.1 Short channel effect in a MOSFET. 3

1.2 Recent reported I_{dsat} values with respect to the channel length reduction by several industrial technologies, adopted from [12][13]. 5

1.3 A cross-section of a state-of-the-art CMOS (After [15]). 6

1.4 Device physics issues for future MOSFETs. 10

2.1 Electron tunneling through oxide layer in a MOS capacitor. (a) energy-band of an n^{+}-poly/oxide/p-Si MOS at flat-band condition. (b) Fowler-Nordheim tunneling. (c) direct tunneling (After [27]). 16

2.2 Band diagram of a poysilicon-gate depleted N-channel MOS capacitor (After [27]). 17

2.3 $C-V$ curves of N-channel MOS capacitors in the presence of polysilicon depletion effects. 18

2.4 Potential distribution of MOS from the top of the poly-gate to the silicon substrate considering poly-depletion effects. 20

2.5 Non-uniform, graded impurity distribution and corresponding energy-band diagram, showing the built-in electric field (E_x) and potential drop (ΔV_{p1}) in the depletion. 21

2.6 Polydepletion effect depending on the gate lengths; effective depletion width becomes wider as gate length is scaled down, leading to additional potential drop, ΔV_{p2}. 22

2.7 Impact of non-uniform dopant distribution on poly-depletion effects. (a) comparison between non-uniform ('A') and uniform ('B') dopant distribution, (b) potential drop (V_p) across the poly-gate for 'A' and 'B' cases. . . . 23

2.8 Channel length-dependent poly-depletion effects for non-uniformly doped gates; depletion width (x_d) of L_g = 35 nm is wider than that of L_g = 100 nm due to the additional depletion in sidewalls (ΔV_{p2}). 25

2.9 Simulated potential drops across poly-gates for various uniform impurity profiles. 25

2.10 Simulated non-uniform dopant profiles and corresponding potential drop (V_p), (a) non-uniform dopant profile generated from 2-D process simulation, (b) simulated potential drop in poly-gate. 26

2.11 Simulated gate capacitance of p-channel MOS capacitors for gate lengths, considering poly-depletion effects by non-uniform doping (t_{ox} = 2.0 nm). (a) simulated vertical dopant profile. (b) gate capacitance (C_g) for gate lengths (L_g = 250, 100, 50, and 25 nm). corrected for overlap capacitance (C_{ov}); $(C_g - C_{ov})/(C_{ox} - C_{ov})$. (c) C_{inv}/C_{ox} vs. gate lengths. (d) experimental C_{inv}/C_{ox} vs. gate lengths from Texas Instruments. 27

2.12 Quantum-mechanically calculated band banding and electron concentration distribution. (a) band banding and energy levels of inversion-layer electrons, E_0 is the ground state. (b) comparison between classical and QM calculations of electron profile. 29

2.13 Comparison between an ideal and real $C--V$ curves of thin oxide MOSFETs. 31

2.14 Measured gate capacitance curves for thin gate oxide (t_{ox} = 1.5, 1.8, and 2.1 nm) N-MOSFETs from the Hewlett-Packard (HP) Labs. 32

2.15 Small-signal equivalent MOS capacitor models. (a) single RC parallel circuit model of the measurement systems (LCR meter). (b) more accurate circuit model including series resistance (R_s). 33

2.16 Simulated and HP experimental $C-V$ curves for N-MOSFET with t_{ox} = 1.5 and 2.1 nm, circles denote measurements and lines indicate QM device simulation results. (a) t_{ox} = 2.1 nm. (b) t_{ox} = 1.5 nm. Note that QM capacitance modeling without considering gate current cannot predict the distorted $C-V$ curve for $t_{ox} < 2.0$ nm. 37

2.17 Electron wavefunction penetration through oxide layer. 38

2.18 Measured gate tunneling current from HP Labs., compared with the simulated one using 1D Green's function solver for t_{ox} = 1.3, 1.5, and 1.8 nm, respectively. 39

2.19 Equivalent circuit approach for ultra-thin gate oxide MOS with the source/drain grounded. (a) top view of the MOS structure with gate current flows. (b) distributed RC network used in this work. (c) RC circuit model used in capacitance measurements. 41

2.20 Simulated abnormal gate capacitance ($Im(Y_{in})$) obtained from the AC network analysis using the equivalent circuit in Figure 2.19(b) (lines), symbols represent measured HP data for t_{ox} = 1.3, 1.5, and 1.8 nm. 43

2.21 Quasi-Fermi potential distribution along the channel at V$_{gs}$ = V$_{ds}$ = 1.0 V for various channel lengths. 45

2.22 Modeling of gate tunneling current considering the drain bias effects, (a) equivalent circuit for a NMOS when the drain bias is given, (b) gate tunneling current ($I_g = I_{gs} + I_{gd}$). 46

2.23 Drain bias and gate channel length dependence of gate tunneling current of t_{ox} = 1.5 nm, (a) I_g versus V_{ds} for L_g = 0.5 and 1.0 μm when V_{gs} = 1.5 V, note that I_g of L_g = 1.0 μm is more than twice than that of L_g = 0.5 μm, (b) experimental ([43]) and modeled I_g with respect to L_g's at saturation mode (V_{gs} = V_{ds} = 1.5 V). 47

2.24 Reconstructed $C - V$ curves after the method in [24], circles represent the intrinsic $C - V$, and dotted and solid lines represent reconstructed $C - V$'s for the applied frequencies of 100 kHz–1 MHz and 100 kHz–5 MHz, respectively. 50

2.25 $C - V$ reconstruction results using the proposed technique. (a) measured (circles) from the North Carolina State University (NCSU) and reconstructed (solid line) $C-V$ curves by using the distributed RC network and optimization technique, dashed line indicates theoretical $C - V$ with QM device simulation. (b) comparison of extracted $C - V$ curves with respect to the number of segments in the RC ladder network, circles denote the measured anomalous $C - V$ and lines show the reconstructed intrinsic capacitance for the different number of segments–1, 20 and 40 segments. 54

2.26 Optimization results, measured C_m/G_m from NCSU, compared to final C/G ($Im(Y_{in})/Re(Y_{in})$) after the $C - V$ reconstruction. 55

2.27 Determination of effective channel length ($L_{eff} = L_{drawn} - \Delta L$) and source/drain resistance ($R_{sd} = R_{tot} - R_{ch}$) using the paired V_g method. . . 61

2.28 Comparisons of Idsat values between obtained from device simulation and Equation (2.16) as a function of physical gate oxide thickness. 62

2.29 Calculated CMOS propagation delays as functions of the physical gate oxide thickness and the fanouts. 63

2.30 Comparisons of experimental (HP) and simulated MOS C–V (classical Fermi-Dirac and QM correction models). 64
2.31 Normalized MOS gate capacitance versus physical oxide thickness. 65
2.32 Effective (electrical) gate oxide thickness (t_{ox}) as a function of physical oxide thickness. 66
2.33 Effective (electrical) gate oxide thickness (t_{ox}) as a function of physical oxide thickness. 67
2.34 CMOS speed optimization using the compact delay model. 68
2.35 Direct tunneling of electron in n^+-polysilicon / SiO_2 / p-Si MOS structure ($V_{ox} < \Phi_B$). 70
2.36 Voltage drop at oxide (V_{ox}) with respect to the applied gate bias for t_{ox} = 1.5 nm. Comparisons between $V_{ox} = V_g$ approximation, V_g - V_{poly}, and the surface potential based model. 74
2.37 Simulated gate current curves using V_g approximation, V_g - V_{poly}, and the surface potential based model, symbols are obtained from NEMO, an approximated Schrödinger equation solver (t_{ox} = 1.5 nm). 75
2.38 Results of the compact model. (a) gate currents vs. V_g by using the compact model for different t_{ox}'s (1.3, 1.5, and 1.8 nm), symbols denote gate current from the NEMO simulator [38]. (b) comparison between gate currents using the compact model and measured HP data for different t_{ox} values. . . 76
2.39 (a) Gate and series resistance (R_g and R_s) effects associated with gate current (I_g) in ultra-thin oxide, large area MOS transistor. (b) Circuit model of gate tunneling current for circuit simulation ($I_g = I_{gs} + I_{gd}$). 77
2.40 Simulated gate tunneling current versus V_{ds} for different V_{gs}, L_g = 10 μm and t_{ox} = 1.5 nm. 79

2.41 Gate length dependence of gate tunneling current at V_{ds} = V_{gs} = 1.5 V for t_{ox} = 1.5 nm, the simulated slope is comparable with ref. [41] due to consideration of DIBL. 79

2.42 Simulated drain current (I_d) versus V_d for t_{ox} = 1.5 nm, (a) L_g = 20 μm (b) L_g = 50 μm. 80

2.43 Transient circuit simulation results with the gate current model, (a) test circuit with NMOS (W/L = 100 μm/20 μm and t_{ox} = 1.5 nm) and 0.1 pF capacitor, (b) simulated voltages at node 'A' with and without gate tunneling current models. 82

2.44 Illustration of gate direct tunneling components of a very short-channel NMOSFET. I_{gso} and I_{gdo} are edge direct tunneling (EDT) currents. 85

2.45 Gate bias dependent band diagrams and electron tunneling in the channel (I_{gc}) and the gate edge (I_{gso} and I_{gdo}). (a) $V_g > 0$ V (inversion mode). (b) $V_{fb} < V_g < 0$ V (depletion mode). 87

2.46 Simulated gate currents using MEDICI [30] and comparison with the measured HP data for a long channel (L_g = 100 μm) NMOSFET. 89

2.47 Simulated I_{gc}, I_{gso} and I_{gg} (= I_{gc}+I_{gso}+I_{gdo}) for an NMOSFET with t_{ox} = 1.5 nm and L_g = 50 nm. 90

2.48 Simulated I_{gg} (= I_{gc}+I_{gso}+I_{gdo}) for different t_{ox}'s ranging 1.1 – 1.8 nm and L_g = 50 nm. 90

2.49 Transient simulation by using MEDICI for a single off-state NMOS transistor with t_{ox} = 1.3 nm, L_g = 70 nm and W = 10μm. (a) discharge through a single off-transistor, V_{out} (t = 0) = 2.5 V. (b) comparison of V_{out} values between with and without the direct tunneling model. 92

2.50 A macro-circuit model for direct tunneling current combined circuit simulation. 93

2.51 V_{OH} and V_{OL} modeling considering gate tunneling currents of a CMOS inverter. (a) V_{in} = logic-low = 0 V. (b) V_{in} = logic-high = V_{dd}. 95
2.52 CMOS inverter simulation results including gate direct tunneling for three gate oxide thicknesses, V_{dd} = 2.5 V (C_{out} = 0.08 pF (FO4) and $L_n = L_p$ = 50 nm). (a) input and output waveforms. (b) power consumptions. 96
2.53 Domino CMOS AND-2 gate. 99
2.54 Simulated waveforms of the domino AND-2 gate. (a) clock and input signals. (b) output waveforms for V_{dd} = 2.5 V. 100
2.55 Simulated waveforms of the domino NAND-2 gate for V_{dd} = 1.5 V. (a) clock and inputs (b) output waveforms. 101
2.56 Modeling of discharge and charge sharing behaviors during first evaluation period of domino AND-2 gate. (a) discharge through tunneling resistance before V_A switches to logic-high (t = 0.20–0.23 μs). (b) charge sharing of C_1 with C_3 after the V_A switches to logic-high ($t > 0.23$ μs). 102
2.57 CMOS sample and hold (S/H) circuit and simulated waveforms for different t_{ox}'s. (a) CMOS S/H circuit schematic and RC circuit model during hold period. (b) waveforms for V_{dd} = 2.5 V. (c) waveforms for V_{dd} = 1.5 V. . 105
2.58 Simulated gate tunneling current of L_g = 50 nm NMOSFET with an alternative gate dielectric of Si_3N_4, $\kappa = 2\epsilon_{ox}$ and thickness is 2.6 nm. 108
2.59 Voltage bootstrapping circuit and simulation results. (a) circuit schematic and its equivalent circuit when V_{in} switches to 0 V. (b) simulated waveforms with an alternative gate dielectric (Si_3N_4) and pure oxides. 109
3.1 External resistance components in source/drain extension (SDE) in an NMOSFET. (a) resistance components (R_{acc} : accumulation, R_{sp} : spreading, R_{co} : contact, R_{sh} : sheet resistances.) (b) current flow and resistivity components. 116

3.2 Comparisons of experimental (HP Labs.) and simulated total resistance (R_{tot} = R_{sd} + R_{ch}) of L_{eff} = 0.25 μm NMOS for V_{ds} = 0.1 V, the unified mobility model for inversion / accumulation layers is used in device simulation. 122

3.3 Simulated external resistance with respect to V_{gs} bias (V_{ds} = 0.05 V). L_{eff} = 0.08 μm and X_j = 40 nm. (a) Resistivity along the channel direction, t_{ox} =2.0 nm. (b) R_{ext} (= $R_s + R_d$) with respect to V_{gs} when t_{ox}'s are 2.0 and 4.0 nm, respectively. 125

3.4 The ratios of R_{ext}/R_{tot} and R_{acc}/R_{tot} for different drawn gate lengths (L_{gate}) ranging from 0.25 to 0.07 μm, V_{gs} = 1.5 V and V_{ds} = 0.05 V 126

3.5 Simulated external resistance with respect to the source-drain extension (SDE) depth (X_j) when V_{gs} = 1.5 V and V_{ds} = 0.05 V (L_{eff} = 0.08 μm and t_{ox} = 4.0 nm). (a) Resistivity along the channel direction for X_j = 40, 50, and 65 nm. (b) R_{ext} (= $R_s + R_d$) with respect to different X_j and L_{ov}. . . 127

3.6 External resistance with respect to X_j in the saturation region (L_{eff} = 0.08μm and t_{ox} = 4.0 nm). (a) R_{ext} with respect to X_j in the saturation region, V_{ds} = V_{gs} = 1.5 V. (b) Simulated on-current (I_{on}) for different SDE X_j's at V_{gs} = 1.0 V, each V_{dsat} for different X_j's is chosen at the V_{ds} when 1 nA/μm of off-leakage current is produced. 128

3.7 Comparison of lateral source/drain profiles (After [84]). 131

3.8 Process flow for an LTP NMOS transistor (After [85]). 132

3.9 Measured I_{ds}–V_{ds} and I_{ds}–V_{gs} of the AMD 70 nm LTP NMOSFETs. 133

3.10 Arsenic dopant profiles for LTP (SIMS and process simulation) and RTP (process simulation), Arsenic implant dose and energy are $6\times10^{14}/cm^2$ and 2 KeV, RTP temperature and time are 1010°C and 5 seconds, respectively. 134

3.11 Comparisons of simulated arsenic dopant profiles (contour) of 70 nm NMOS transistor, (a) LTP case, (b) RTP case. 135

3.12 Simulated net dopant profiles along the silicon surface for LTP and RTP. . . 136

3.13 Measured (AMD) and simulated total resistance (R_{tot}=R_{ch}+R_{sd}) vs. V$_{gs}$ using the unified mobility model, V_{ds} = 0.1 V and L_g = 70 nm NMOS. . . . 137

3.14 Computed sheet resistivity along the channel direction for LTP and RTP devices, (a) comparisons of resistivity distributions for LTP and RTP, V$_{gs}$ = 1.5 V and V$_{ds}$ = 0.1 V (b) variations of sheet resistivity vs. gate biases for LTP case, V$_{ds}$ = 0.1 V. 138

3.15 I_{on} and C_{ov} versus L_{ov} for LTP and RTP devices. (a) I_{on} of LTP and RTP for different gate overlaps (L_{ov}), I_{on} was chosen at the drain bias (V$_{ds,on}$), producing off-current (I_{off}) of 1 nA/μm (V_{gs} = 0.9 V). (b) Normalized I_{on} and gate overlap capacitance (C_{ov}) for various overlap lengths (L_{ov}). 140

3.16 R_{ext} and I_{on} current for various SDE junction depths (X_j) in LTP process, (a) R_{ext} vs. X_j (V$_{ds}$ = 0.1 V and V$_{gs}$ = 1.2 V), (b) I_{on} vs. X_j (V$_{ds}$'s at I$_{off}$ = 1 n A/μm and V$_{gs}$ = 0.9 V). 141

3.17 Computed sheet resistivity along the channel direction for different SDE depths (X_j). 142

3.18 Impact of spacer length on external resistance and on-current. 142

4.1 Regions of validity for various device simulation models (After [89]). . . . 145

4.2 Essential physical picture for carrier transport in MOSFETs, after [94]. (a) conduction band from the source and drain, (b) fluid flow analogy under high drain and gate bias. 149

4.3 MOSFET showing source and drain external resistance, effective (intrinsic) gate voltage is denoted by $V_{gs'}$. 151

4.4 (a) NMOSFET structure with L_g = 50 nm, V_T = 0.28 V, X_j = 30 nm, and t_{ox} = 2.5 nm. (b) Electron velocities along the channel obtained from HD and DD transport models for the L_g = 50 nm NMOSFET, HD has been calibrated according to the Monte-Carlo results in ref. [90]. 153

4.5 Simulated drain current and carrier velocity. (a) Simulated drain current between HD and DD models, V_{gs} = 1 V. (b) Electron velocity near the source extracted from hydrodynamic simulation result, V_{gs} = 1 V. (c) I_{ds} comparison between HD simulation and Equation (4.3), V_{gs} = 1 V. 155

4.6 Potential and electric field distribution for HD and DD device simulation (a) Quasi-Fermi potential distribution along the channel for HD and DD models, L_g = 50 nm and V_{gs} = V_{ds}= 1 V. (b) Lateral electric field distribution along the channel between HD and DD models, and relationship l and $E_X(0^+)$, V_{gs} = V_{ds}= 1 V. 159

4.7 Potential and electric field distributions calculated by using Equation (4.6) and dV/dx. V_{gs} = 1 V, symbols represent HD simulation results. (a) potential (b) e-field. 160

4.8 Critical distance (l) from the source calculated by using Equation (4.8). . . 161

4.9 Calculated $E_X(0^+)$ for drain bias based on Eqs. (4.6) and (4.8), V_{gs}=1 V. . . 162

4.10 Drain current by using the proposed $v(0)$ model, symbols represent HD simulation, V_{gs} = V_{ds} = 1 V. 162

Chapter 1

Introduction

1.1 Motivation

CMOS technology dimensions are being aggressively scaled to reach a plateau where device performance must be assessed against fundamental limits. At the same time, there has been concurrent scaling of supply voltage in concert with device dimensions to keep power manageable and in order to meet reliability requirements. For logic devices, increased power consumption is a big concern with increases in clock speed. The reduction of the supply voltage is the most effective method to reduce the power. However, reduced drive current for high speed circuit operation offsets gains due to power reduction. For this purpose, aggressive reduction of gate oxide thickness as well as gate length is desirable [1]. In practice, the major constraint, limiting the performance of sub-100 nm devices, is the thickness of the gate insulator.

The insulator, currently using silicon dioxide, is 2–3 nm thick and further scaling requires reduction to 1.5 nm by 2004, which corresponds to about 5 layers of silicon atoms [2]. According to the CMOS technology roadmap proposed in November 2000 by

the International Technological Roadmap for Semiconductors (ITRS) [3], device designers would like to use 1.5 nm insulator layers as early as 2001, although it may not be practical in production until 2004 [4]. In this thin gate oxide regime, the direct tunneling will be exponentially increased, leading to significant power dissipation and device performance deterioration, which is a primary concern for MOS scaling [2]–[5]. Even though the tunneling process does not appear to damage the oxide, the resulting gate leakage can cause circuit failures because the designs assume that there is no appreciable gate current [5]. Another concern of the gate oxide scaling is the loss of inversion charge and transconductance due to the quantization of carriers in the inversion layer. Using Quantum-Mechanical (QM) carrier models, electron energy quantization occurs at the Si-SiO_2 interface. In the earliest MOS devices, its effects resulted in a reduction of low-field mobility compared with values in the bulk silicon. In modern devices, the quantization effects tend to become more relevant and easily detectable, as higher channel doping concentrations are widely employed to suppress short-channel effects. As the centroid of the charge distribution is displaced from the surface, gate capacitance is reduced in accordance with $C_{ox} = \epsilon_{ox}/t_{ox}$; such a reduction is more perceptible as t_{ox} is scaled down [6]. It is reported that carriers in the inversion layer peak at around 10 Å below the silicon surface and gate capacitance and inversion charge are both effectively reduced to values of an equivalent oxide which is a few angstroms thicker than the physical oxide in modern devices [7]. In addition, mobility degradation of the inversion charge due to the large vertical fields and IR drops in the gate electrode, in combination with tunneling currents, have become serious concerns. Nevertheless, scaling of the gate dielectric is essential to enhance capacitive coupling between the gate and the channel that reduces device on-resistance. If the gate insulator cannot be scaled, device performance will not continue to improve [5].

Another limiting factor on CMOS scaling is the effect of external source and drain

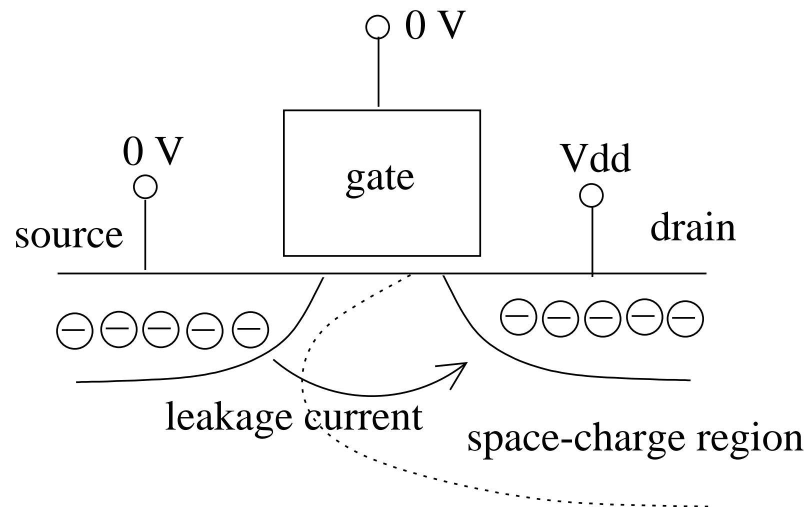


Figure 1.1: Short channel effect in a MOSFET.

resistance in shallow junction devices [8][9]. One of the fundamental challenges in modern device performance is the trade-off between short-channel effects and the impact of source-drain series resistance. When making the gate length small, the space-charge region near the drain touches the source in a location below the surface where the gate bias cannot control the potential, resulting in a leakage current between the source and drain via the space-charge region, as shown in Figure 1.1 – the so-called short-channel effect of MOSFETs. For a MOSFET to operate as a VLSI component, the capability of switching off this current path and suppressing short-channel effects is a major priority in MOSFET design [9]. In the on-state, reduction of gate length is desirable to minimize the channel resistance. However, when the channel resistance becomes as small as the source and drain resistance, further improvements in drain current cannot be expected because increases in these resistances tend to suppress the short-channel effects. Even though the ITRS roadmap predicts Source and Drain Extension (SDE) depths as shallow as 10 nm to achieve a 50 nm

transistor, too shallow a junction leads to high external resistance, which in turn results in degradation of driving current. Therefore, further down-scaling without improvement of MOSFET performance is meaningless, even if the short-channel effect is completely suppressed by introducing the shallow SDE.

Recently, several ultra-small transistors that employ shallow SDE junctions have been reported [10][11]. However, their drive currents have not been improved commensurate with channel length reduction, as reflected in Figure 1.2. In this figure, I_{dsat} values of the modern devices (i.e. L_g = 0.09 – 0.25 μm), adopted from recently published articles [12], are compared with those from previous generations of Intel logic technology (i.e. L_g = 0.25 – 1.5 μm) [13]. Note that only slight improvements in drive current are achieved by the scaling from 0.25 μm to 0.09 μm, which is in contrast to the scaling for long channel devices, from 1.5 μm to 0.25 μm. This poor I_{dsat} improvement can be attributed to the higher external source and drain resistance effects due to the shallow SDE. In design of sub-100 nm MOSFETs, modeling of the source and drain resistance is needed to accurately predict the drain current and optimize trade-offs between the short-channel effect and the driving current.

Recent work also shows that modern devices now operate surprisingly close to their ballistic limits [14], in which the steady-state current is controlled by how rapidly carriers are transported across a low field region near the beginning of the channel. In addition, transistors have reached dimensions approaching the mean distance between carrier collisions, only a few collisions occur as carriers move across the device. Under these circumstances, carrier velocity can exceed the saturation velocity, which results in velocity overshoot. Thus, the drift-diffusion (DD) theory which treats carrier transport in some average fashion in thermal equilibrium with the silicon lattice is no longer sufficient; understanding of non-equilibrium carrier transport effects is needed to develop nanoscale MOSFETs. For engineering-oriented device design of nanoscale MOSFETs, modeling that provides a

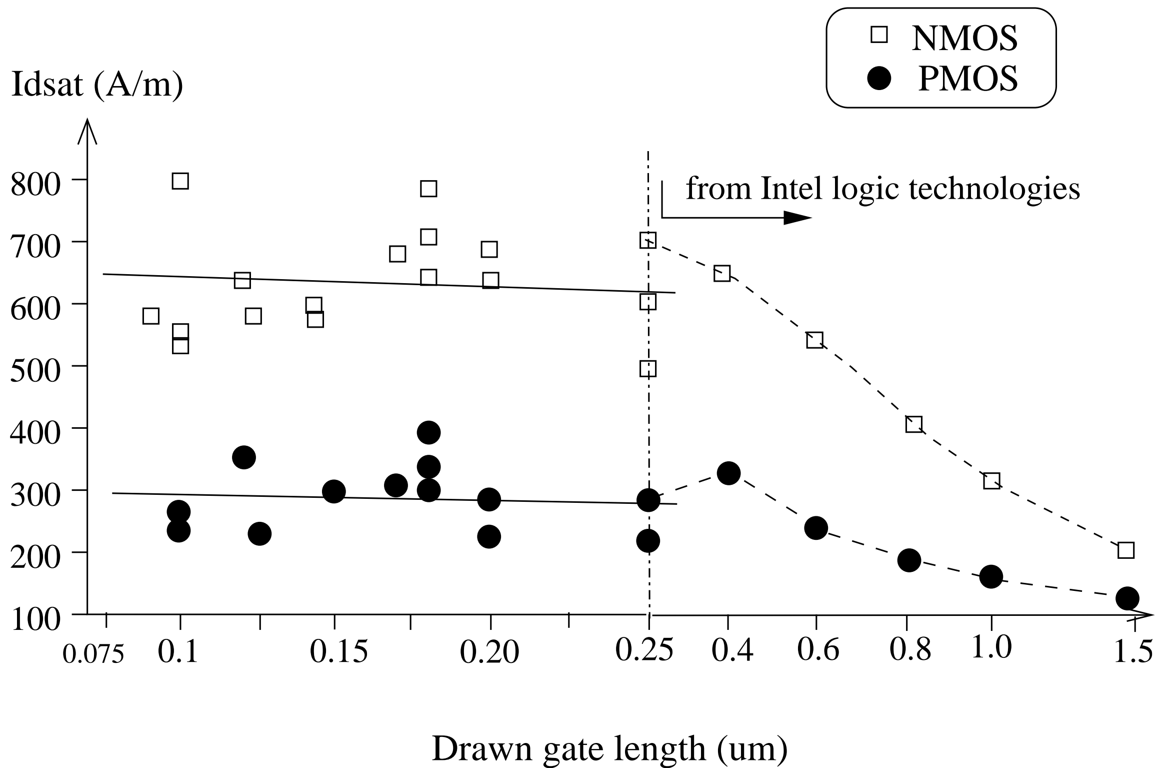


Figure 1.2: Recent reported I_{dsat} values with respect to the channel length reduction by several industrial technologies, adopted from [12][13].

physical view of carrier transport is necessary. One such non-equilibrium transport model uses the well-known hydro-dynamic (HD) approach from the computational fluid field to model electron transport.

1.2 Challenges for CMOS Technology

CMOS transistors with conventional processing technologies have been encountering performance limits as the device dimensions and operating voltages are scaled down. To

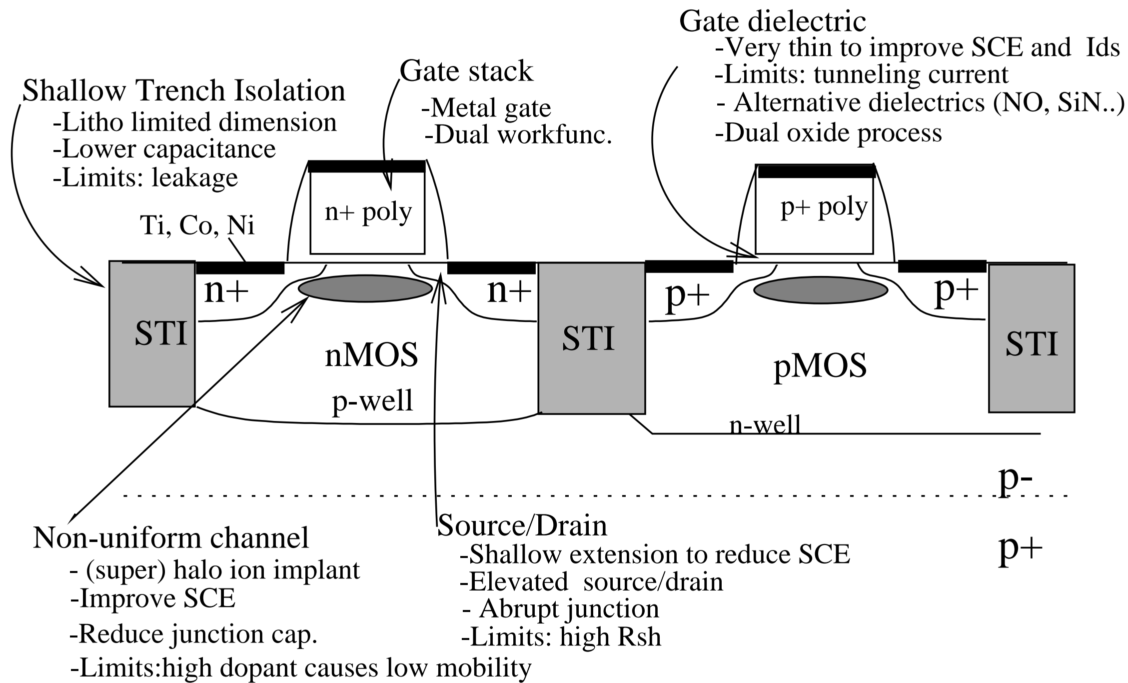


Figure 1.3: A cross-section of a state-of-the-art CMOS (After [15]).

overcome these limits while maintaining compatibility with conventional processes, new technologies continue to be developed. Figure 1.3 illustrates important features of a state-of-the-art CMOS technology [15]-[18]. The gates are fabricated with n or p-type polysilicon so that both NMOS and PMOS are surface channel devices. The gates are topped with a metal silicide to reduce gate series resistance. The gate dielectric must be very thin to provide sufficient current drive, it is typically around 3 nm for the 100 nm gate length CMOS generation. The source and drain require shallow, moderately doped extensions under the gate edges and sidewalls to reduce short channel effects.

Shallow trench isolation between devices involves etched trenches, filling with deposited dielectrics, and polishing of the surface, which allows devices to be placed much

closer together with higher circuit density than with the local oxidation of silicon (LOCOS). Self-aligned silicide has been introduced in the mid-1980s to reduce source/drain contact resistance as well as gate resistance. In achieving the shortest possible channel length, the engineering of the doping profiles in the channel region has been introduced. Halo implants at the source and drain edges are used to adjust threshold voltages and suppress punch-through. It also suppresses source and drain-induced threshold voltage shifts as well as drain-induced barrier lowering by shielding the effects of drain potential on the channel region. This increases junction capacitance, while carrier mobility is reduced due to the higher impurity scattering terms. Recently, a highly non-uniform 2D dopant profile called *super-halo* has been introduced. It demonstrated that a properly scaled *super-halo* can suppress the potential barrier lowering both in the inversion layer and the body depletion region, eliminating the need for overly aggressive scaling of source and drain extensions [19]. The substrate impurity concentration has already reached values above $1.0{\times}10^{18}$ cm^{-3}. If the concentration is further increased, the source-substrate and drain-substrate junctions become highly doped p-n junctions and act as tunnel diodes; isolation of the source and drain from the substrate cannot be maintained [1], which is a major concern for the super-halo technology.

There are many technological challenges facing CMOS in the near future; gate oxide thickness (t_{ox}), shallow source/drain junction (X_j) and gate length (L_g) are the most critical dimensions that need to be engineered for future generations of CMOS technology. These three parameters are now discussed in greater detail.

- Channel length : The gate length is the smallest feature of the MOSFET patterned by lithography and etching. From the processing technology point of view, employing a wavelength of the optical lithography light source that is shorter than 193 nm faces

Table 1.1: ITRS roadmap (2000 Edition)

year	1999	2002	2005	2008	2011	2014
DRAM min. feature (nm)	180	115	80	60	40	30
MPU min. feature (nm)	120	80	65	40	30	20
min. Logic V_{dd} (V)	1.8–1.5	1.5–1.2	1.2–0.9	0.9–0.6	0.6–0.5	0.6–0.3
t_{ox} equivalent (nm)	2–3	1.5–2	< 1.5	< 1.0	< 1.0	< 1.0
S/D junction depth (nm)	45–70	30–50	25–40	20–28	13–20	10–14

many obstacles. Today's most advanced production lithography equipment uses excimer laser sources with a wavelength of 248 nm (KrF), producing a resolution about 0.25 μm. Further improvements in resolution may be attained at the 193 nm wavelength (ArF) which has demonstrated adequate lithographic resolution down to 0.18 μm [2]. Resolution enhancement techniques, such as phase shifting, are capable of imaging features in 0.10 – 0.12 μm. However, there are limited expectations that these optical lithographic techniques will extend the resolution down to the sub-100 nm regime. For the fabrication of such small devices, optical lithography may be used for non-critical levels, where the features are defined by electron beam or X-ray lithography. However, the challenges in implementing X-ray lithography lie in difficulties of mask fabrication; challenges in electron beam technology come from the low throughput [15]. Even though sub-lithographic feature size may be obtained either by optical image-transfer techniques or new lithographical sources, these have yet to be demonstrated in a manufacturing environment. Development of reliable and manufacturable lithographic techniques is essential to achieve nanoscale CMOS technology. However, even if smallest features are successfully achieved due to the new lithographic sources for printing, electrical effects in the devices – surface scattering, dopant fluctuations [20] and non-equilibrium effects – are critical issues from

the viewpoint of achieving device performance in nanoscale MOSFETs.

- Gate oxide thickness : As indicated in Table 1.1, the thickness of the gate dielectric must be decreased along with the channel length to enhance the gate control in order to overcome short-channel effects and transconductance degradation. As devices continue to scaled down, tunneling through the thin-oxide becomes a limiting factor. Moreover, the tunneling occurs not only in the inversion region, but also in the accumulation region. Even though the gate voltage is sufficient for the off-state in the channel, tunneling current is appreciable when the drain is biased (i.e. off-state leakage current). While the gate leakage current may be at a negligible level compared with the on-current in the channel of a device, it will affect the chip standby current. Accordingly, thicker gate insulators with a higher dielectric constant than SiO_2 are being considered for the range of equivalent oxide thicknesses below 2 nm, as a way to reduce gate tunneling currents [21] [22]. However, the added process complexity and non-ideal properties at the interface between insulator/silicon are issues that have yet to be solved.

- Junction depth : Junction depths have been controlled by ion-implantation of the dopants into selected regions; they are limited by diffusion during subsequent thermal cycles. Future generations of technology require steeper doping profile, but present annealing technology is unable to produce sufficiently steep doping gradients at the edge of the source/drain regions. Moreever, higher dopant re-activation while maintaining low resistivity is critical in nanoscale CMOS technology.

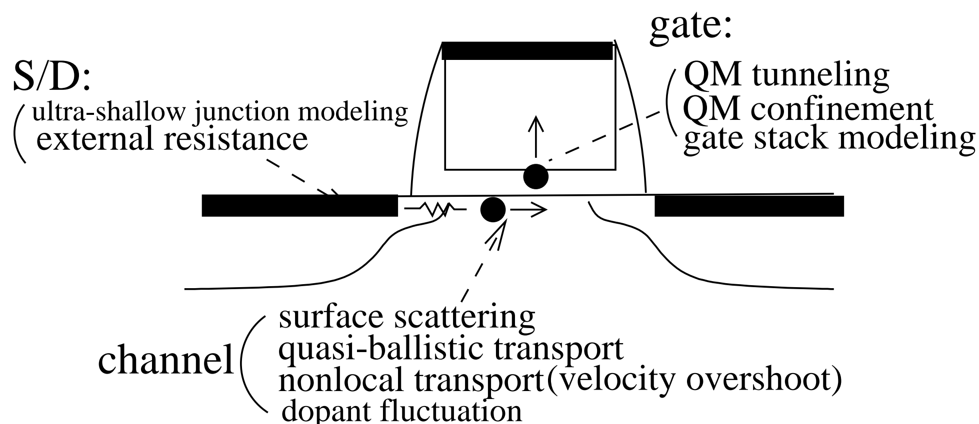


Figure 1.4: Device physics issues for future MOSFETs.

1.3 Scope and Organization

As traditional MOS scaling methods become less effective, nanoscale device modeling is needed to provide innovative new MOS devices as well as to understand the limits of the scaling process. Modeling provides insight into the operation of modern semiconductor devices and circuits, and simulation dramatically reduces the development costs and time-to-market. Modeling is also needed to explore new device structures that may operate on principles different than those currently used [23].

There are several device physics issues of importance for the future MOSFET, as shown in Figure 1.4. The challenge facing device model developers is the formulation and efficient numerical implementation of physically-based models that are predictive for practical device structures. This thesis focuses on issues of CMOS devices as they are scaled to nanoscale dimensions. The main objective of this work is to carefully consider TCAD models in order to quantify potential roadblocks for MOS scaling. This includes development of new modeling approaches that are needed, especially with respect to gate tunneling.

Following this chapter, the thesis describes advanced modeling of nanoscale MOSFETs from the viewpoint of device physics, which consists of three parts – gate, source/drain, and channel modeling. The first part of the thesis, Chapter 2, is devoted to a thorough examination of gate modeling in ultra-thin oxide MOSFETs. Modeling of gate capacitance in ultra-thin gate oxide MOSFETs is presented based on the experimental C–V data; it incorporates quantum-mechanical and direct tunneling effects. In addition, a capacitance reconstruction method is discussed, based on effects of measured anomalous C–V curves. A compact direct tunneling current model, useful for circuit simulation, is developed. In turn, the impact of the gate direct tunneling to circuit performance is finally considered.

Chapter 3 describes parasitic resistance issues of the source and drain regions. The modeling of ultra-shallow junctions becomes of great importance, as low resistance source and drain extensions are necessary for future high-performance devices. Thermal budgets in modern CMOS technology have been reduced to the point that profiles are dominated by damage, transient annealing and interface effects. In terms of process modeling, obtaining parameters for diffusion and reaction kinetics is a key challenge to allow greater understanding of fabrication processes using new dopants and materials. In this area, more detailed understanding of implant damage and subsequent re-crystallization during anneal is important, as these phenomena can critically affect dopant profiles. Traditionally, channel resistance has been the dominant factor limiting current transport in MOSFETs. In ultra-small devices with source/drain extensions, however, parasitic series resistance has become comparable or even higher to channel resistance, thus it has become imperative to accurately model the extrinsic region of the device. The objectives of this chapter address accurate modeling of the extrinsic resistance and application of the modeling to ultra-shallow junction MOSFETs to optimize device performance. Analysis and optimization of device performance for an experimental process technology called Laser Thermal Process (LTP) are also discussed.

Finally, traditional MOSFET channel current formulations have to date continued to use the drift-diffusion (DD) approximation which treats carrier transport in some average fashion, considering carriers to be in thermal equilibrium with the silicon lattice. The drift-diffusion model breaks down in ultra-small devices where high fields and rapid spatial variations of the electric field are presented. In such cases, the scattering events are no longer localized, and some fraction of the carriers may acquire thermal energy near the drain. These carriers are no longer in thermal equilibrium with the silicon lattice and are referred to as hot carriers. Under these circumstances, it is possible for the carrier velocity to exceed the saturation velocity, which results in velocity overshoot. To date a more rigorous treatment of carrier transport under spatially nonuniform high-field conditions has been carried out primarily using a Monte Carlo solution of the Boltzmann transport equation for electron distribution function. However, it is difficult to gain physical insight from the complex non-local transport used in Monte-Carlo simulation. Chapter 4 presents a simple MOSFET current model that takes into account the non-equilibrium channel effects based on the hydro-dynamic (HD) carrier transport model. The modeling results are discussed and compared with device simulations for 50 nm NMOSFETs.

Chapter 5 summarizes the findings of this research and suggests possible areas for further investigation.

Chapter 2

Gate Modeling

2.1 Introduction

It is not an overstatement to say that ongoing VLSI growth has been achieved primarily through scaling down the size of the MOSFET. Enhanced performance includes fast switching speed, lower power dissipation, and smaller areas for devices and circuits. However, such scaled MOSFETs have suffered from various undesirable phenomena. Especially the scaling of gate oxide thickness has reached both technology and circuit limits.

The ITRS roadmap predicts that CMOS with the gate length of 100 nm requires an oxide thickness of less than 1.5 nm in 2005 (Table 1.1), which corresponds to 3 – 4 atomic layers of oxide [5]. With such thin oxide layers, direct tunneling results in an exponential increase in gate leakage current such that series resistance in MOS capacitors becomes significant, owing to the low impedance of the capacitor [24]. As a result, measured $C-V$ curves for ultra-thin gate MOS devices show capacitance attenuation, both in the inversion and accumulation regions [25] and determination of an effective oxide thickness from the measured $C-V$ curves has become problematic. Although recent studies have reported

that the degradation of gate capacitance in MOSFETs can be suppressed by utilizing shorter channel lengths [26], this requires redesign of test patterns and more importantly, it may be difficult to extract intrinsic gate capacitances due to the various parasitic effects in short channel devices. Since $C-V$ measurement is a fundamental technique and workhorse for MOSFET characterization, $C-V$ curves are crucial in providing device information. A reconstruction technique of gate capacitance from measured $C-V$ curves is needed for ultra-thin gate dielectric MOSFETs. In addition, for determining directions of future gate oxide scaling, limitations due to gate tunneling current must be critically evaluated in terms of circuit operation in order to determine the circuit immunity against gate tunneling current.

This chapter includes modeling of gate capacitance and tunneling current along with validation of a distributed RC approach for representing ultra-thin gate oxide MOSFETs. A $C-V$ reconstruction technique based on distorted measured $C-V$ curves is discussed and application to very thin nitride/oxide gate dielectric MOSFETs is considered. In addition, a compact gate tunneling current model for circuit simulation and the impact of the gate current on circuit performance is introduced.

2.2 Gate Capacitance Modeling

2.2.1 Tunneling through Silicon Dioxide

Consider an N-channel MOS capacitor with a gate electrode consisting of heavily doped n-type polysilicon. When biased at the flat-band condition, the energy-band diagram is shown in Figure 2.1(a), where Φ_{ox} denotes the Si/SiO_2 interface energy barrier for electrons – approximately 3.1 eV [27]. When a large positive bias is applied to the gate electrode, electrons in the strongly inverted surface can tunnel through the oxide layer and hence

produce a gate current. Similarly, if a large negative voltage is applied to the gate electrode, electrons from the n^+ polysilicon can tunnel in the opposite direction.

Fowler-Nordheim (F-N) tunneling occurs when electrons tunnel into the conduction band of the oxide layer, as shown in Figure 2.1(b). At an oxide field strength of 8 MV/cm, the measured F-N tunneling current density is about 5×10^{-7} A/cm^2, which is quite small, and can be neglected for normal device operation.

If the oxide layer is very thin, say 4 nm or less, electrons from the inverted silicon surface can tunnel directly through the forbidden energy gap of the oxide layer, as illustrated in Figure 2.1(c). The theory of the direct tunneling is more complicated than that of F-N tunneling, because it should be explained based on the Quantum-Mechanical (QM) theory. The direct tunneling current can be very large for thin oxide layers, hence it is very important in MOSFETs of very small dimensions, where the gate oxide layers are less than 2 nm.

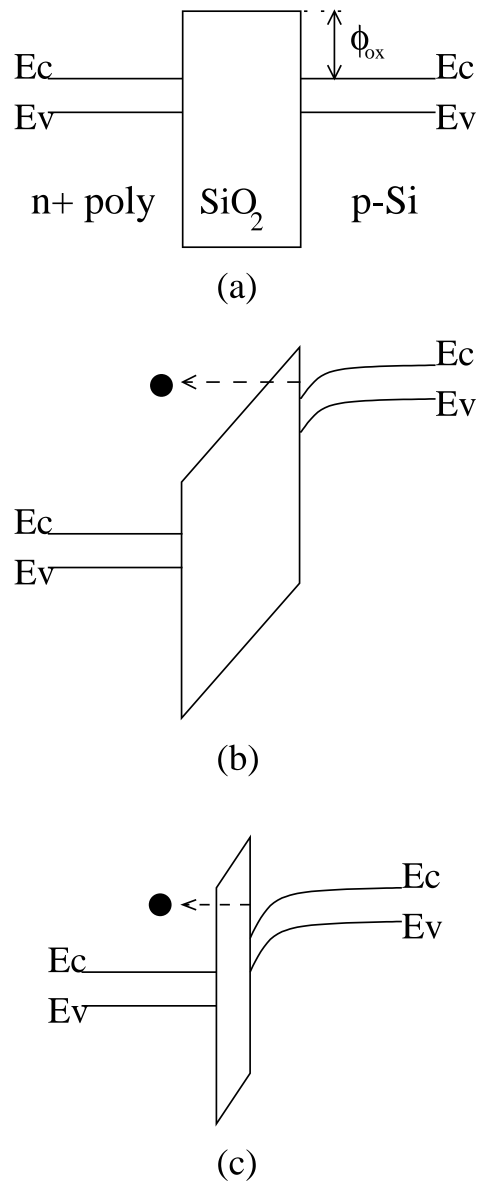


Figure 2.1: Electron tunneling through oxide layer in a MOS capacitor. (a) energy-band of an n^+-poly/oxide/p-Si MOS at flat-band condition. (b) Fowler-Nordheim tunneling. (c) direct tunneling (After [27]).

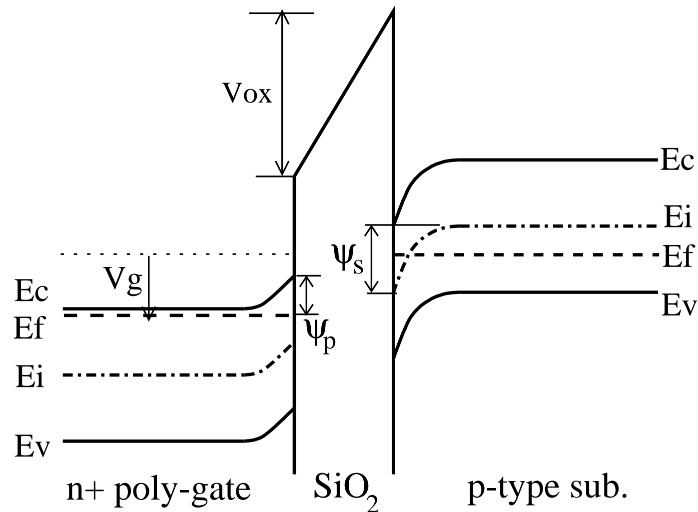


Figure 2.2: Band diagram of a poysilicon-gate depleted N-channel MOS capacitor (After [27]).

2.2.2 Polysilicon-Gate Depletion Effects

When the polysilicon gate is not doped heavily enough, a concern with the dual gate process in which the gates are doped by ion implantation, gate depletion results in an additional capacitance in series with the oxide capacitance. This effect leads to a reduced inversion charge density and transconductance. Consider the band diagram of an n^+-polysilicon gated n-channel MOS structure biased into inversion as shown in Figure 2.2. Since the oxide field is in the direction which accelerates negative charge toward the gate, the bands in the n^+-polysilicon bend slightly upward near the interface. The field depletes the surface of electrons and forms a thin space-charge region in the polysilicon layer, which lowers the total capacitance as given by:

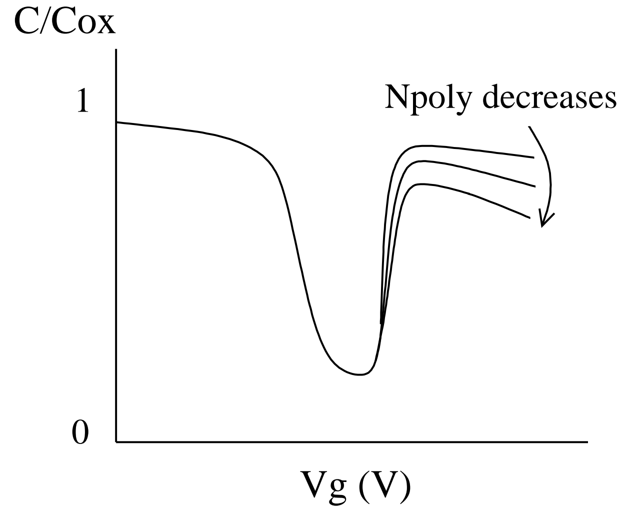


Figure 2.3: $C - V$ curves of N-channel MOS capacitors in the presence of polysilicon depletion effects.

$$\frac{1}{C} = \frac{1}{C_{ox}} + \frac{1}{C_{si}} + \frac{1}{C_{dep}} \tag{2.1}$$

Typical low-frequency $C - V$ curves in the presence of this gate depletion effect are shown in Figure 2.3. The maximum inversion capacitance is less than C_{ox}, and the higher the ploysilicon doping concentration the less pronounced is the gate depletion effect.

Impact of non-uniform dopant profile on poly-depletion effects

The potential drop (V_p) in the presence of the poly-depletion effect is represented in Figure 2.4. When the gate is doped by ion implantation, the dopant distribution as a function of position from the top of gate to the gate oxide direction can be represented as in Figure 2.5. Since the dopant density and carrier concentration vary with position, a built-in

electric field exists leading to built-in depletion. Under thermal equilibrium, since there is no current flow, the built-in field (E_x) in an n-type poly-gate can be expressed as [28]:

$$E_x = -\frac{d\phi}{dx} \approx -\frac{kT}{q}\frac{1}{N_d}\frac{dN_d}{dx} \qquad d\phi = \frac{kT}{q}\frac{dN_d}{N_d} \tag{2.2}$$

A potential drop (ΔV_{p1}) is established at the interface due to the graded dopant distribution between x_1 and x_2 which can be approximated as:

$$\Delta V_{p1} = \phi_1 - \phi_2 = \frac{kT}{q} ln \frac{N_{d1}}{N_{d2}} \tag{2.3}$$

When N_d changes from 10^{20} to 10^{18} cm^{-3}, the potential drop is about 0.12 V from Equation (2.3) even with no gate bias, which is on the order of the threshold voltage for sub-100 nm MOSFETs. This additional potential drop across the polysilicon gate (ΔV_{p1}) should be added to the voltage drop based on the uniform dopant concentration (V_p).

Polysilicon depletion effects depending on the gate length are shown in Figure 2.6; additional depletion effects at the gate sidewalls due to the fringing gate fields, result in additional potential drops. Similar to that of the narrow-width effect, the portion of the additional charge (ΔQ) in total charge (Q) becomes higher as the gate length is scaled down, resulting in wider depletion widths and additional potential drops for shorter gate lengths (L_2). Let A denote the triangular area of the additional charge ΔQ, then the additional potential drop (ΔV_{p2}) due to the sidewall depletion is approximated as [29]:

$$\frac{1}{2}\Delta Q \approx qN_d\frac{A}{L}\ (C/cm^2) \qquad \Delta V_{p2} \approx \frac{\Delta Q}{C_d} = \frac{2qN_dA}{LC_d} \tag{2.4}$$

This ΔV_{p2} should be also added to V_p to reflect further poly-depletion for very short gate lengths.

Comparison of the potential drop (V_p) has been made for uniform (A) and non-uniform

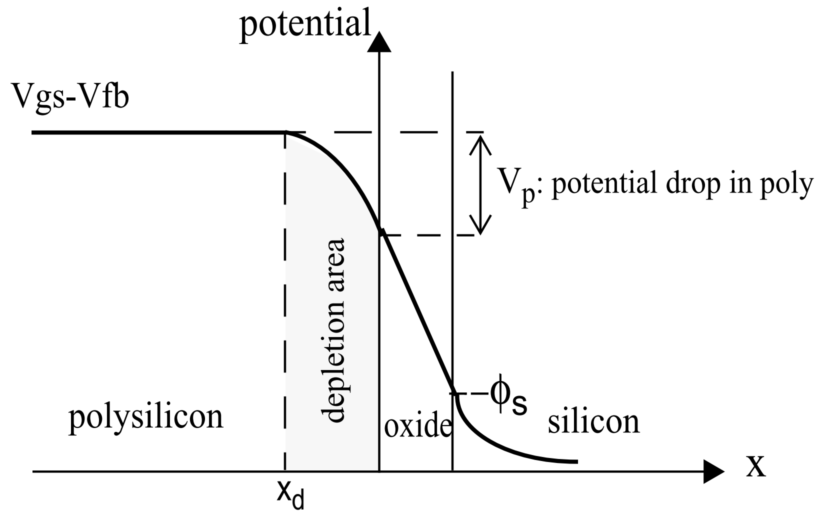


Figure 2.4: Potential distribution of MOS from the top of the poly-gate to the silicon substrate considering poly-depletion effects.

doping (B) cases, as represented in Figure 2.7(a). V_p values calculated by using a two-dimensional device simulator, MEDICI [30], are shown in Figure 2.7(b). It is instructive to note that V_p for case B is less than that of A, in spite of its lower average doping concentration, which can be attributed to the graded impurity distribution effects, reflected by ΔV_{p1}.

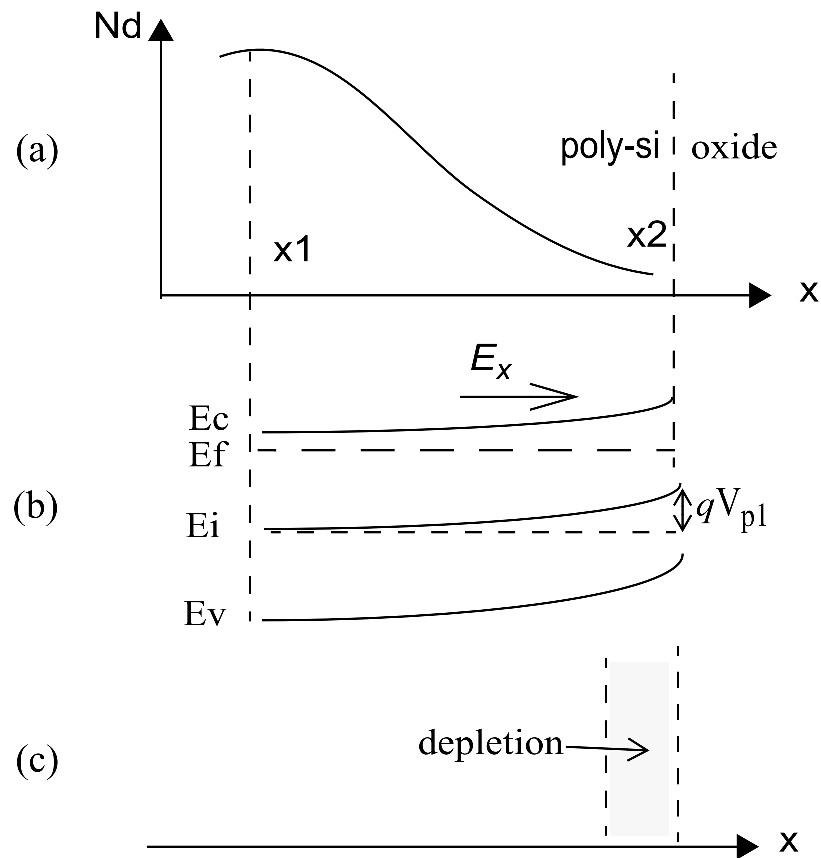


Figure 2.5: Non-uniform, graded impurity distribution and corresponding energy-band diagram, showing the built-in electric field (E_x) and potential drop (ΔV_{p1}) in the depletion.

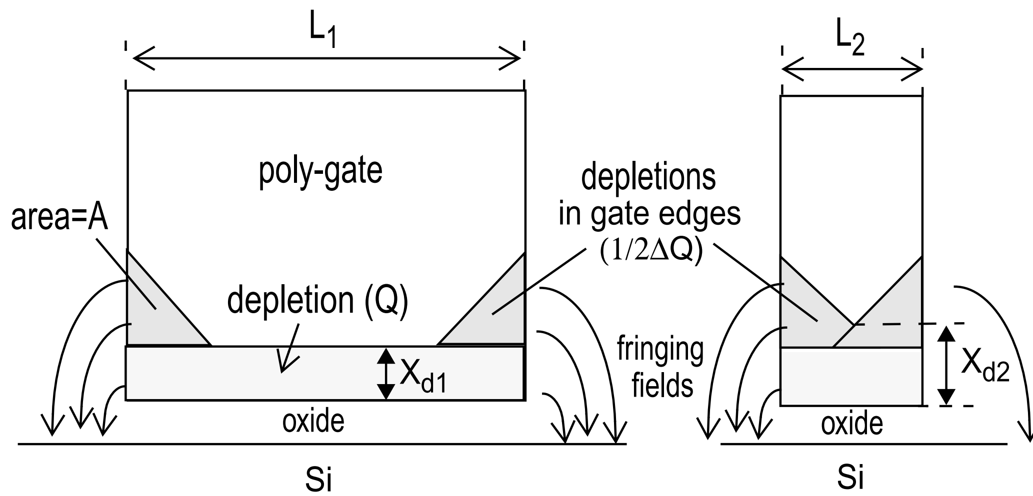


Figure 2.6: Polydepletion effect depending on the gate lengths; effective depletion width becomes wider as gate length is scaled down, leading to additional potential drop, ΔV_{p2}.

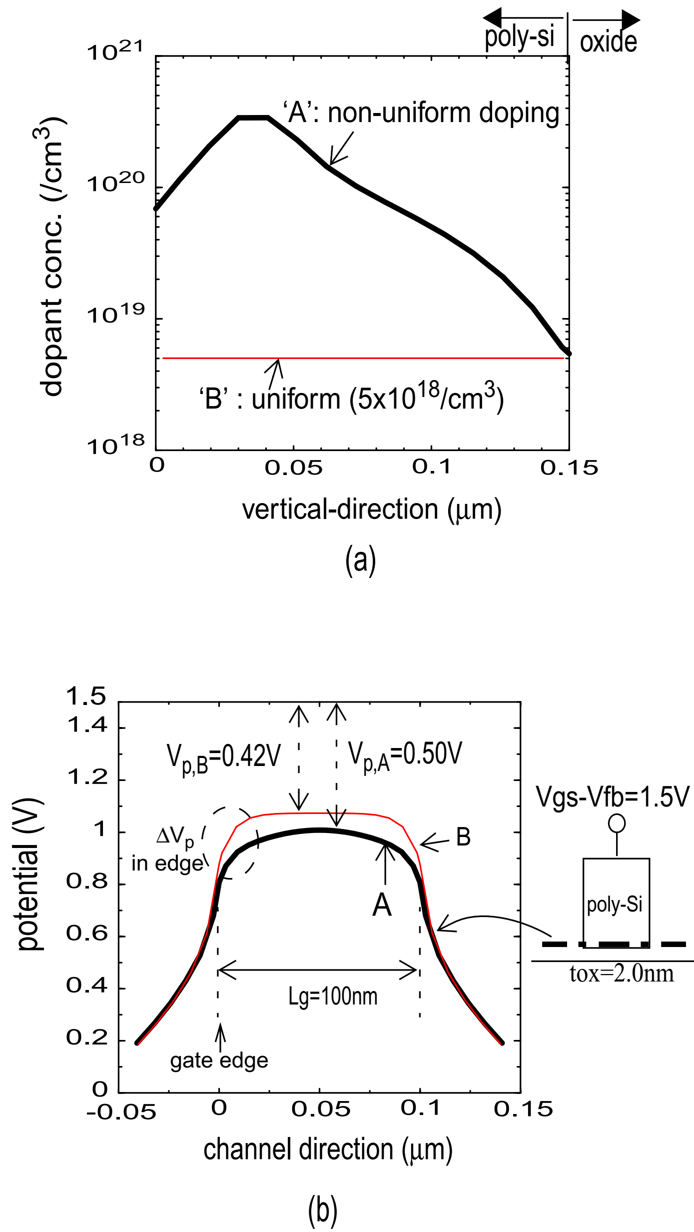


Figure 2.7: Impact of non-uniform dopant distribution on poly-depletion effects. (a) comparison between non-uniform ('A') and uniform ('B') dopant distribution, (b) potential drop (V_p) across the poly-gate for 'A' and 'B' cases.

Regarding the gate length dependence, the depletion width for a device with gate length L_g = 35 nm is wider than that for L_g = 100 nm, resulting in ∼0.1V larger V_p due to the ΔV_{p2} effect, as shown in Figure 2.8.

This gate length dependency on polysilicon depletion is minor for uniform dopant concentrations, as shown in Figure 2.9. However, it becomes more apparent for the non-uniform dopant distribution case as shown in Figure 2.10(a),(b).

Inversion capacitance (C_{inv}) is reduced due to the potential drops in gate region. Device simulation results for p-channel MOS capacitors show that a non-uniform poly-gate dopant profile represented in Figure 2.11(a) produces significant C_{inv} reductions, as shown in Figure 2.11(b). The portion of maximum C_{inv} relative to C_{ox} becomes lower as gate length is scaled down, as shown in Figure 2.11(c), due to the gate-length dependent poly-depletion effects. Figure 2.11(d) shows experimental data of C_{inv}/C_{ox} provided by Texas Instruments for devices ranging in channel length from 0.4 μm down to 0.12 μm; the results show trends consistent with simulations shown in Figure 2.11(c).

In conclusion, the impact of the non-uniform dopant distribution (ΔV_{p1}) and the short gate length (ΔV_{p2}) on the poly-depletion effects has been modeled and verified by using 2D device simulation. Non-uniform dopant gradients resulting from ion implant can worsen the poly-depletion effects, especially for sub-50 nm gate lengths. Achieving less steep dopant gradients as reflected in case of 'A' in Figure 2.10 can be a potential solution to overcome this problem.

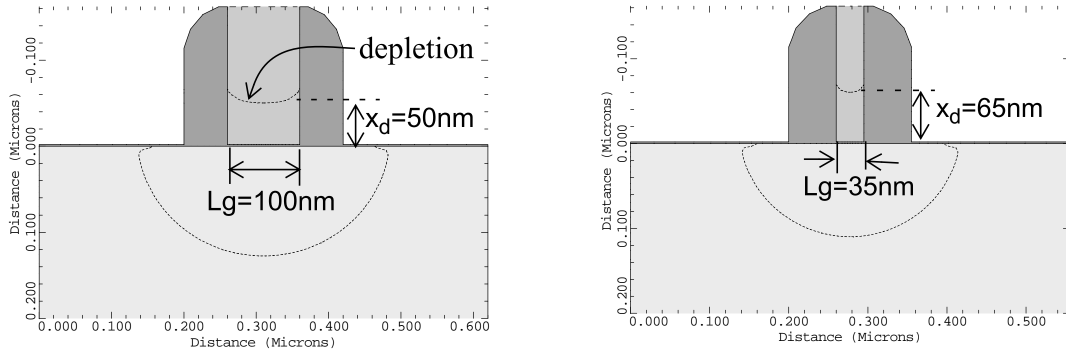


Figure 2.8: Channel length-dependent poly-depletion effects for non-uniformly doped gates; depletion width (x_d) of L_g = 35 nm is wider than that of L_g = 100 nm due to the additional depletion in sidewalls (ΔV_{p2}).

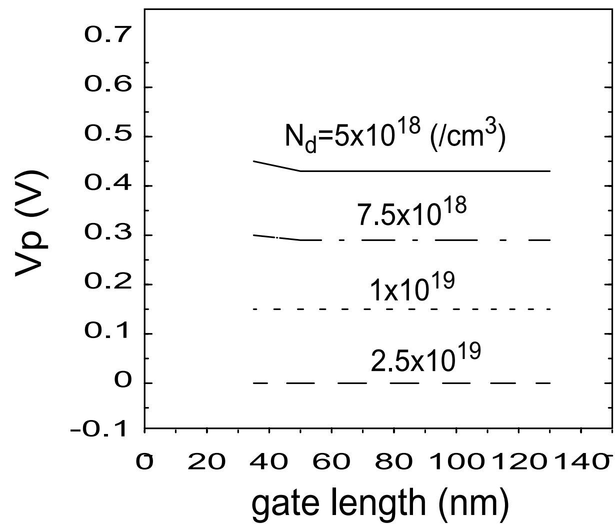
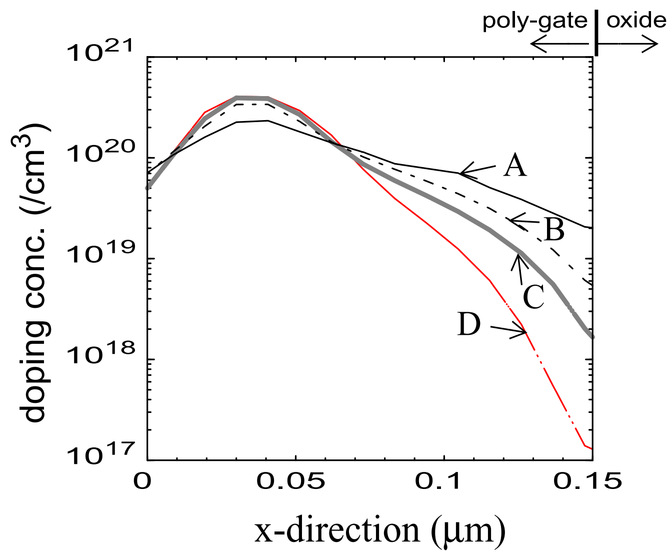
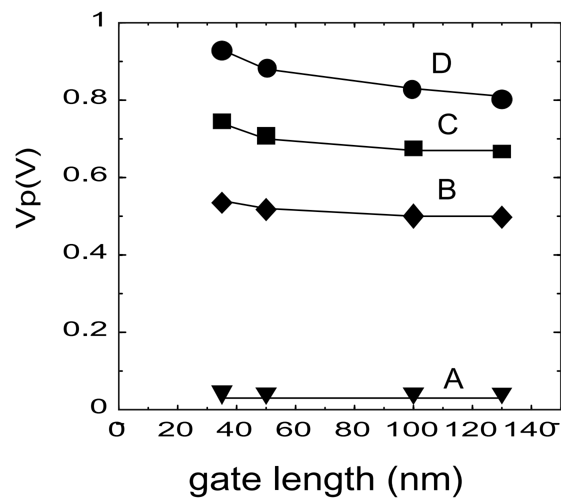


Figure 2.9: Simulated potential drops across poly-gates for various uniform impurity profiles.



(a)



(b)

Figure 2.10: Simulated non-uniform dopant profiles and corresponding potential drop (V_p), (a) non-uniform dopant profile generated from 2-D process simulation, (b) simulated potential drop in poly-gate.

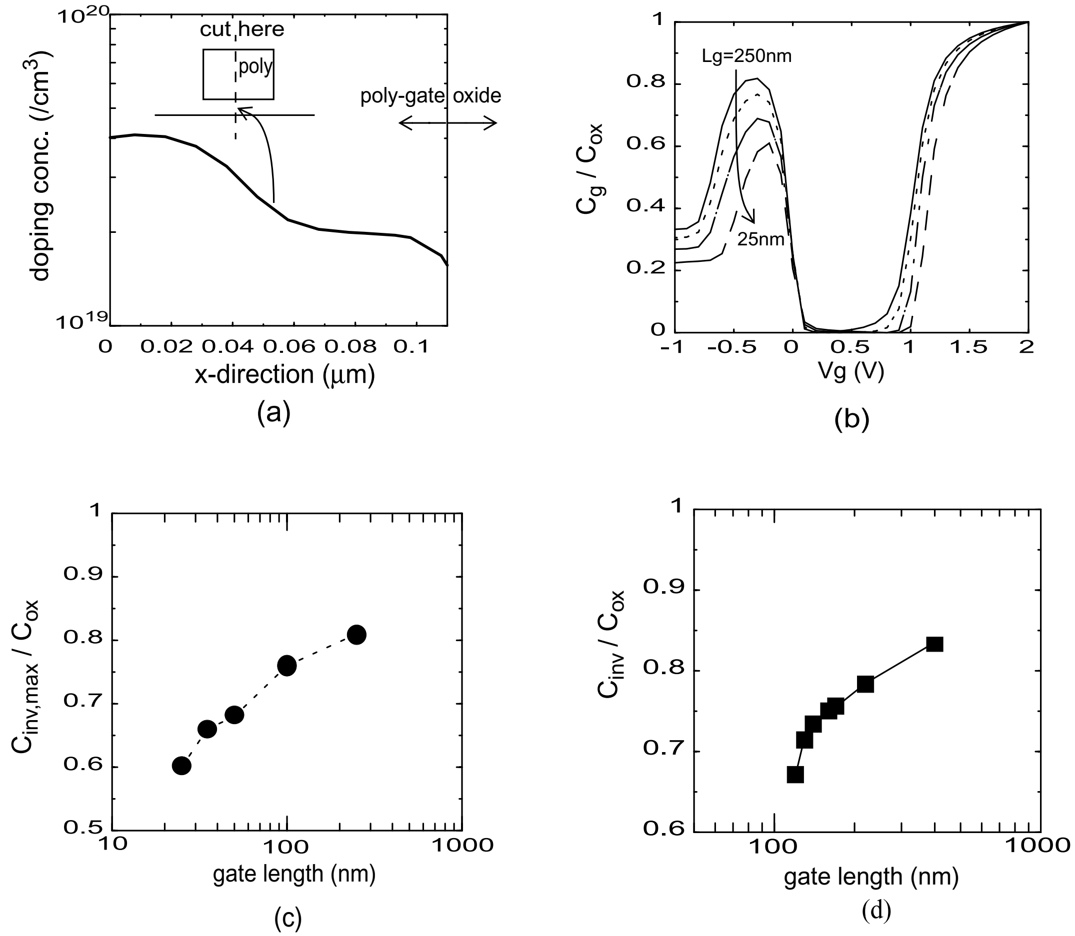
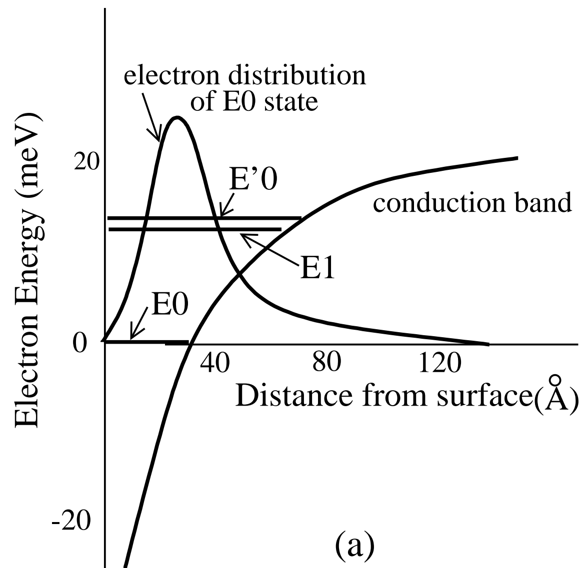
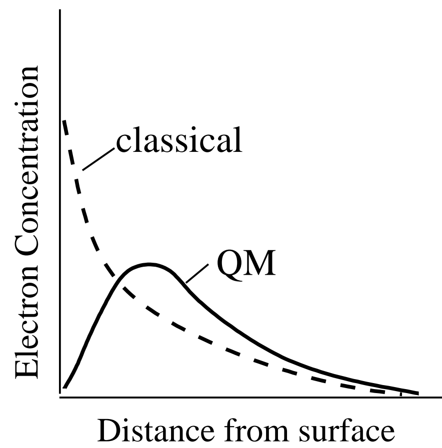


Figure 2.11: Simulated gate capacitance of p-channel MOS capacitors for gate lengths, considering poly-depletion effects by non-uniform doping (t_{ox} = 2.0 nm). (a) simulated vertical dopant profile. (b) gate capacitance (C_g) for gate lengths (L_g = 250, 100, 50, and 25 nm). corrected for overlap capacitance (C_{ov}); $(C_g - C_{ov})/(C_{ox} - C_{ov})$. (c) C_{inv}/C_{ox} vs. gate lengths. (d) experimental C_{inv}/C_{ox} vs. gate lengths from Texas Instruments.

2.2.3 Quantum Effects on C–V

A quantum mechanical analysis, in which the wave nature of electrons is emphasized, is necessary if the dimension of the confining potential is comparable to the deBroglie wavelength, $\lambda = h/\sqrt{3m^*k_BT}$, of electrons, which at room temperature is approximately 150 Å [27]. Modern MOSFETs with very thin gate oxides have steeper potential wells near the interface. The wells are formed by the oxide barrier and the silicon conduction band, which bends down steeply toward the surface due to the applied gate field, as illustrated in 2.12(a). Carriers in the inversion layer are confined in the potential well very close to the silicon surface and should be treated quantum-mechanically as a two-dimensional gas, especially at high normal fields. The energy levels of electrons are grouped in discrete subbands, each of which corresponds to a quantized level for motion in the normal direction. If the electrons are represented as wavefunctions, then the nature of the electron distribution in the inversion-layer differs significantly from the case in which the electrons are treated as classical particles [31], as shown in Figure 2.12(b).





(b)

Figure 2.12: Quantum-mechanically calculated band banding and electron concentration distribution. (a) band banding and energy levels of inversion-layer electrons, E_0 is the ground state. (b) comparison between classical and QM calculations of electron profile.

Assuming the potential barrier at the interface is infinitely large, the wavefunction vanishes at the interface since the probability of finding an electron there is nearly zero. Hence, the electron concentration peaks below the interface, which is in contrast to the classical model in which the electron concentration peaks at the surface. As a result, the threshold voltage becomes higher since more band bending is required to populate the lowest subband, which is a finite energy above the bottom of the conduction band. Once the inversion layer forms below the surface, it takes a higher gate voltage over-drive to produce a given level of inversion charge density [32]. Hence, the effective gate oxide thickness is slightly larger than the physical thickness, such that the maximum capacitance in the $C - V$ curves is less than C_{ox} for thin oxide MOSFETs, as depicted in Figure 2.13. The real gate capacitance appears to be less than the ideal capacitance in both the accumulation and the inversion modes, but there is an additional capacitance drop in the inversion mode due to the ploysilicon gate depletion effects.

2.2.4 Anomalous C–V Behavior in Thin-Oxide MOS

Recent measured $C - V$ for oxides thinner than 2.0 nm have shown a sharp decrease of capacitance both in the inversion and accumulation regions [25]. This capacitance attenuation was found to increase with increasing gate area and thinning gate oxide thickness.

Figure 2.14 shows experimental $C - V$ curves for gate oxide thicknesses of 1.5, 1.8, and 2.1 nm N-MOSFETs with an area of $100\ \mu$m $\times$ $100\ \mu$m, which were measured using an LCR capacitance meter in the RC parallel mode at a frequency of 100 kHz. Thickness extraction of the oxides was initially performed using ellipsometry using a 122 Å oxide standard; high resolution TEM was then used to correct the shift between ellipsometry and TEM. In Figure 2.14, sharp decreases of gate capacitance appear for the oxide thicknesses of 1.5 and 1.8 nm, which is related to gate current and series resistance associated with the

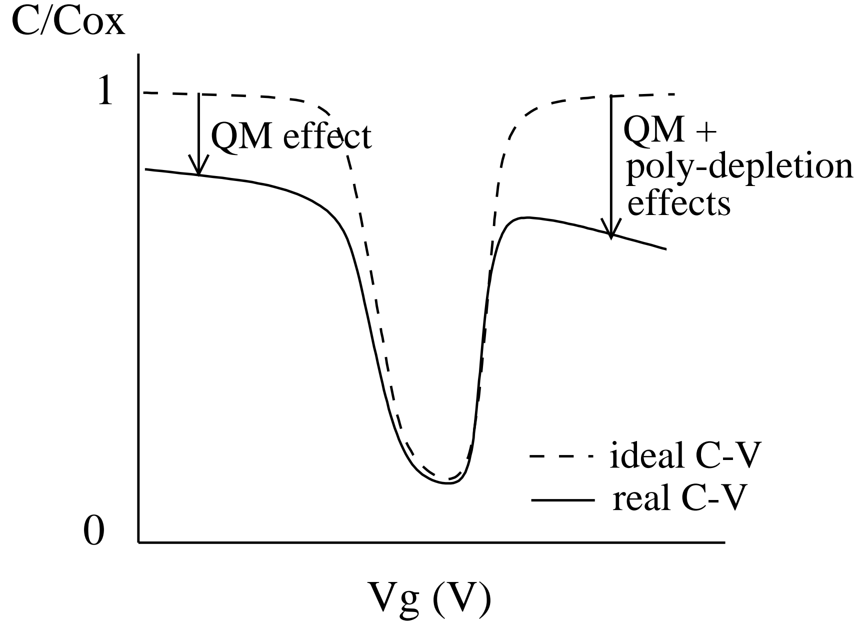


Figure 2.13: Comparison between an ideal and real $C--V$ curves of thin oxide MOSFETs.

equivalent circuit in $C-V$ measurement systems.

Figure 2.15(a) shows the RC parallel circuit model of the LCR capacitance meters. G_m and C_m are the measured conductance and capacitance, respectively, which are determined from a single measurement of magnitude and phase of impedance. Because this circuit model neglects the series resistance (R_s) associated with the bulk and contacts, it is most suitable for low series resistance devices. By contrast, a circuit model that more closely represents the real devices, including R_s, is shown in Figure 2.15(b). G_t and C_t represent the tunneling conductance and the true, intrinsic capacitance, respectively. The relationship between C_m and C_t is expressed as [33]:

$$C_m = \frac{C_t}{(G_t R_s + 1)^2 + \omega^2 C_t^2 R_s^2} \tag{2.5}$$

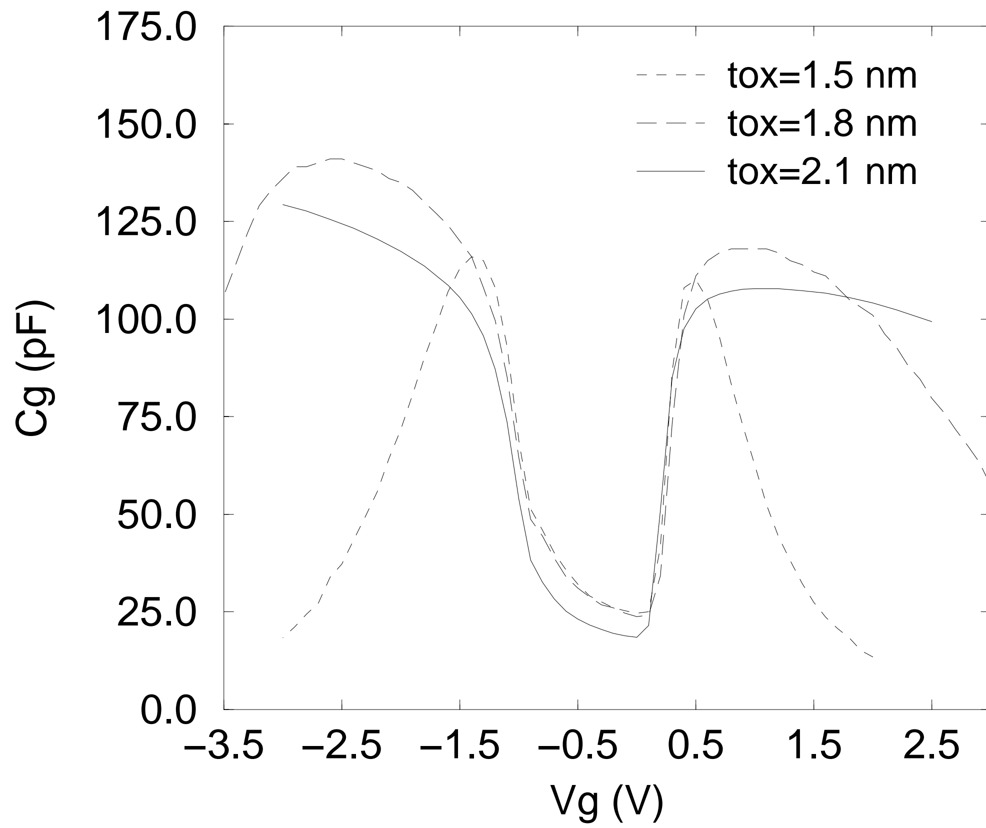


Figure 2.14: Measured gate capacitance curves for thin gate oxide (t_{ox} = 1.5, 1.8, and 2.1 nm) N-MOSFETs from the Hewlett-Packard (HP) Labs.

Here, C_m and C_t are identical when R_s is very small, but the magnitude of C_m decreases to a value less than C_t when R_s increases.

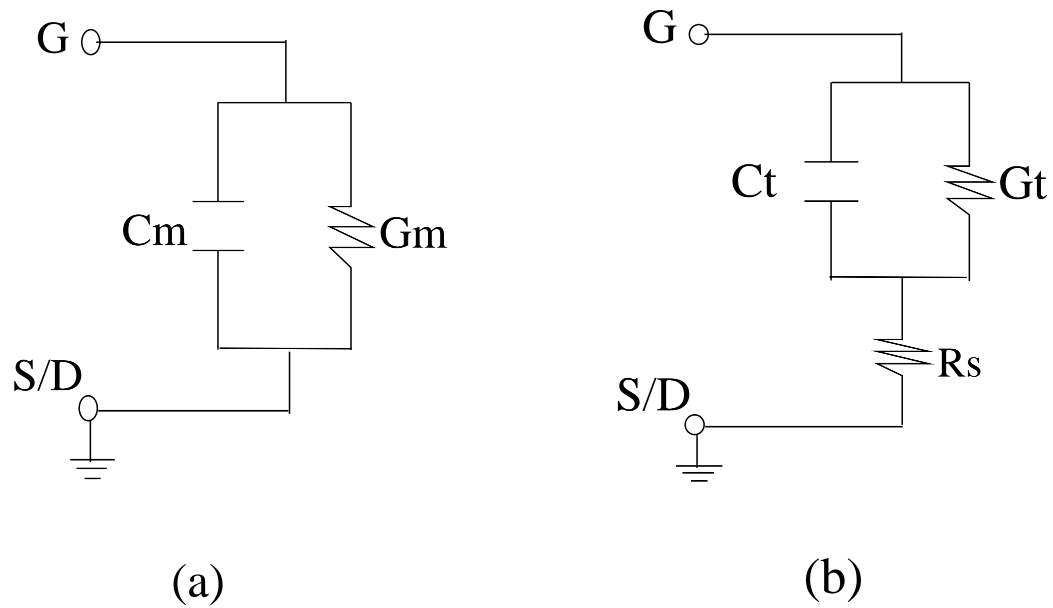


Figure 2.15: Small-signal equivalent MOS capacitor models. (a) single RC parallel circuit model of the measurement systems (LCR meter). (b) more accurate circuit model including series resistance (R_s).

In spite of these limitations, $C-V$ measurements of thin oxides with large leakage currents are widely performed, often using the parallel RC circuit model in Figure 2.15(a) [24]. The true capacitance (C_t) of the devices cannot be ascertained simply by the measurement of magnitude and impedance phase (i.e., G_m and C_m). Namely, when G_t becomes large, due to high leakage for thin dielectrics, R_s is amplified, leading to decreased capacitance (C_m) as the magnitude of gate bias increases [34]. As a result, determination of intrinsic gate capacitance and effective oxide thickness from $C-V$ measurements becomes difficult in the presence of leaky, very-thin oxide MOSFETs.

2.2.5 QM Models of C–V and Gate Tunneling Current

For gate oxide thicknesses less than 3.0 nm, fields in the oxide reach a maximum of 5 MV/cm, a regime in which tunneling and charge confinement effects become noticeable. Classical inversion charge modeling is no longer sufficient for these device characteristics. In the classical case, the electron density has its maximum value at the Si-SiO_2 interface, while in the quantum mechanical case the electron density is diminished at the interface, increases to its maximum value and decreases with the distance from the silicon surface [29]. Thus, a quantum-mechanical (QM) model of the inversion charge profile peaks below the silicon surface that inversion charge is effectively reduced to that of an equivalent oxide which is thicker than the physical oxide. For the modeling of gate characteristics an empirical, hybrid model for the QM correction in MOS structures is proposed and implemented in a 2D device simulator, which combines both the van Dort [35] and Hansch models [36].

In order to model the carrier confinement, the effective density of conduction band states in the vicinity of the Si/SiO_2 interface is expressed as:

$$N_C(z) = N_C[1 - e^{-(z+z_0)^2/\lambda_{th}^2}] \tag{2.6}$$

where N_C is the normal effective density of conduction band states, which is position-independent, and λ_{th} is a thermal wavelength determined by the carrier effective mass. The $z-$axis is perpendicular to the Si-SiO$_2$ interface and z_0 is introduced to represent the finite carrier concentration at the surface ($z = 0$).

The van Dort model employs bandgap broadening in the surface channel region to capture the dominant QM effects. The original form was proposed for the inversion region only. As a result, a singularity in the simulated $C-V$ at the flatband voltage region was not carefully considered. While the QM effect is not a concern in this region, the discontinuous behavior of the model greatly limits its applications and causes numerical instabilities in multi-dimension device simulations. To eliminate the flatband singularity of the van Dort model, an alternative bandgap broadening expression is used in this work. The bandgap correction term with a 2/3 power-law dependence of the transverse surface electric field, F_S, was implemented in the van Dort model formalism as follows:

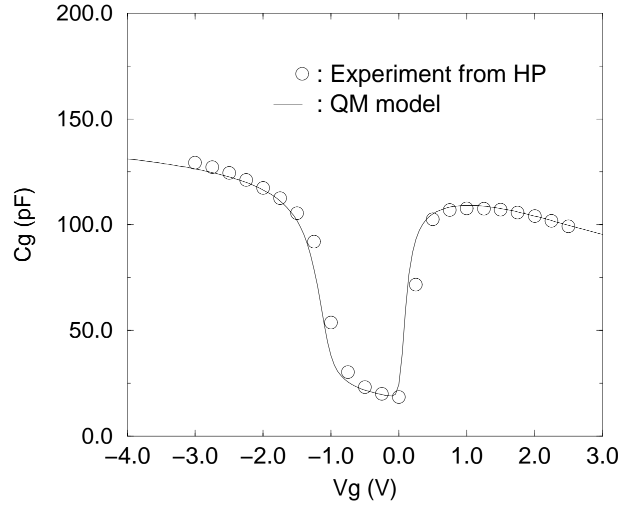
$$\Delta E_g(z) = \frac{\kappa\beta}{2k_B T} F_S^{2/3} \left[\frac{2e^{(z/L)^2}}{1 + e^{-2(z/L)^2}} \right] \tag{2.7}$$

where β is a physical constant (see [35]), κ is a fitting parameter with theoretical value of unity and L is a characteristic decay length. The singularity originates from the derivative of this quantity with respect to $F_S = 0$ in the gate capacitance computation. To eliminate the flatband singularity, the $F_S^{2/3}$ term in Equation (2.7) is replaced with

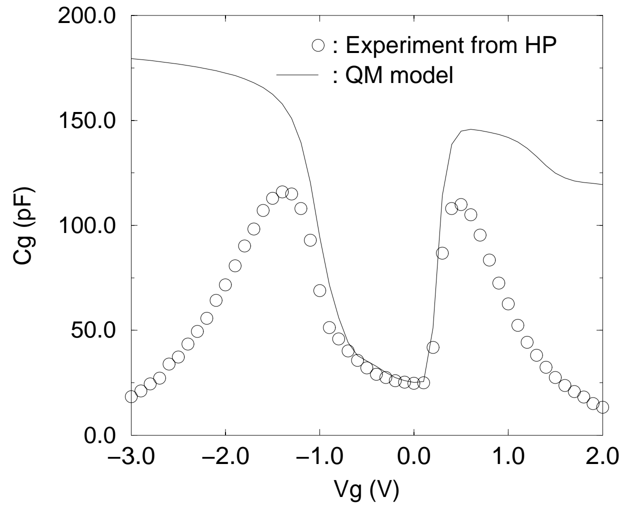
$$F_S^{2/3} \rightarrow \quad \frac{F_S^2}{ae^{-(F_S/\sigma)^2} + F_S^{4/3}} \tag{2.8}$$

where both a and σ are adjustable parameters. This expression preserves the asymptotic dependence of the bandgap correction on $F_S^{2/3}$ when $F_S \rightarrow \pm\infty$ while eliminating the flatband singularity.

MOS gate oxides were grown using rapid thermal oxidation (RTO) in the temperature range of 900–950°C. Figure 2.16 shows the comparison between experimentally measured and simulated $C-V$ curves for 1.5 and 2.1 nm oxide N-MOSFETs with an area of 100 μm$\times$ 100 μm. For this analysis, the substrate doping profile is obtained using process simulation that includes TED (transient enhanced diffusion) effects [37]. The $C-V$ curves obtained by utilizing the QM-corrected behavior, based on device simulation, show good agreement with experimental data for t_{ox} = 2.1 nm, as shown in Figure 2.16(a). The reduction of gate capacitance in the accumulation region is related to the QM effect; the reduction in the inversion region is related to both the QM and poly-depletion effects. However, as shown in Figure 2.16(b), the simulations cannot predict the characteristics of t_{ox} = 1.5 nm; discrepancies in the deep inversion and accumulation regions can be attributed to neglecting gate tunneling current effects. To model the anomalous $C-V$ behavior for gate oxides thinner than 2.0 nm, tunneling current effects should be incorporated along with the QM capacitance model.



(a)



(b)

Figure 2.16: Simulated and HP experimental $C - V$ curves for N-MOSFET with t_{ox} = 1.5 and 2.1 nm, circles denote measurements and lines indicate QM device simulation results. (a) t_{ox} = 2.1 nm. (b) t_{ox} = 1.5 nm. Note that QM capacitance modeling without considering gate current cannot predict the distorted $C - V$ curve for $t_{ox} < 2.0$ nm.

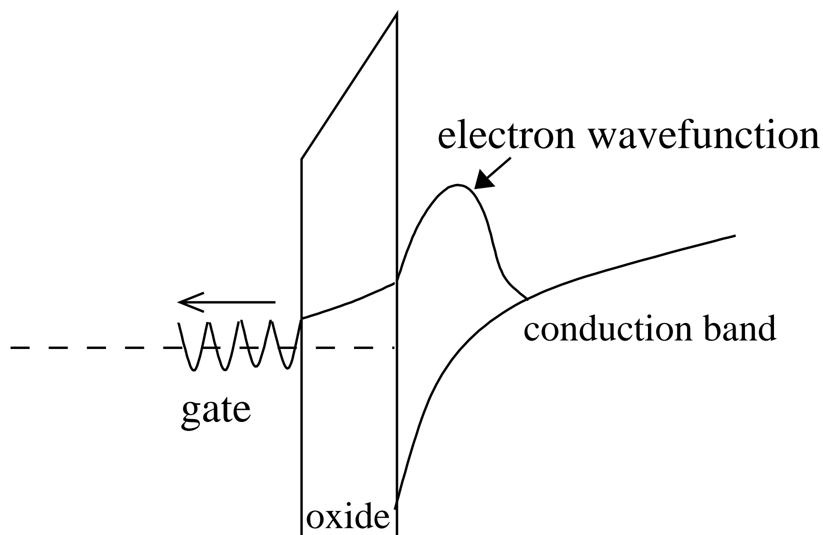


Figure 2.17: Electron wavefunction penetration through oxide layer.

As devices shrink to nanometer scales, their dimensions begin to approach the wavelength of the electron. Electrons must then be treated as quantum mechanical entities. Rather than picturing them as tiny billiard balls, they must be regarded as waves traveling across the device, reflecting off boundaries and contacts, and interacting with other waves. More precisely, the probability of finding an electron at the Si-SiO_2 interface is not zero. Hence, in very-thin oxide MOS system, the electron wavefunction in the potential well can tunnel through into the oxide layer thereby reaching the gate electrode, as shown in Figure 2.17.

In this work, the tunneling current is calculated using a one-dimensional Green's function simulator, NEMO [38], which is a Schrödinger equation solver. NEMO considers the injection from both quasi-bound states and continuum states; the carrier density is calculated quantum mechanically in the device. NEMO also considers multiple-scattering effects which are important for very thin SiO_2 layers [39]. Figure 2.18 shows measured

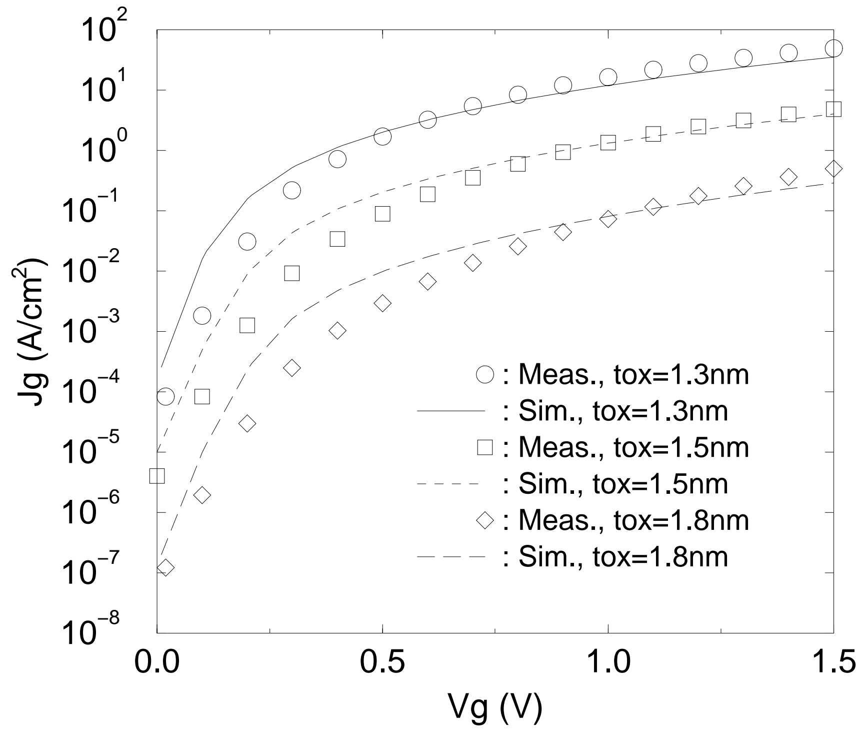


Figure 2.18: Measured gate tunneling current from HP Labs., compared with the simulated one using 1D Green's function solver for t_{ox} = 1.3, 1.5, and 1.8 nm, respectively.

and simulated gate tunneling currents for three different gate oxide thicknesses – 1.3, 1.5, and 1.8 nm. The substrate doping profile obtained from a calibrated process simulation was used; oxide thickness was the variable parameter (1.3–1.8 nm). However, these oxide thickness values were about 0.5 Å greater than that of the measured values due to effects of vanishing electron wavefunctions at the oxide interface and the displacement of the electron peak towards the bulk which is common in quantum simulations [40]. The discrepancies between the measured and simulated $I - V$ curves are probably caused by surface roughness and uncertainty in determining the effective oxide thickness.

2.2.6 Equivalent Circuit Modeling

In order to model the decrease of gate capacitance in ultra-thin gate MOS transistors, an equivalent RC network is used. To model gate tunneling currents considering parasitic resistance, the silicon surface region is divided into small rectangular segments perpendicular to the channel-current flow in order to consider the series resistance (R_s), as shown in Figure 2.19(a). The poly-Si area is vertically divided into small segments to reflect the polysilicon resistance effects (R_g). R_s represents the series resistance for each segment and R_g represents a distributed resistance in the polysilicon gate. Typical sheet resistance of the polysilicon (R_g) is 4–5 $\Omega/\Box$, which is much smaller than the series resistance (R_s $\sim$ 0.1–1K $\Omega/\Box$). Figure 2.19(b) shows the equivalent circuit for the single vertical segment represented in Figure 2.19(a). The equivalent circuit for the whole device structure is constructed by incorporating all the vertical segments. The tunneling current is modeled using a nonlinear, voltage-controlled current source ($i_g(V_g)$) derived from the gate current simulation discussed previously. The magnitude of current in each segment is then scaled according to the number of RC stages used. Also, the gate capacitance without the gate tunneling effects (or capacitance attenuation) is modeled using a nonlinear, voltage-controlled capacitance ($C_i(V_g)$).

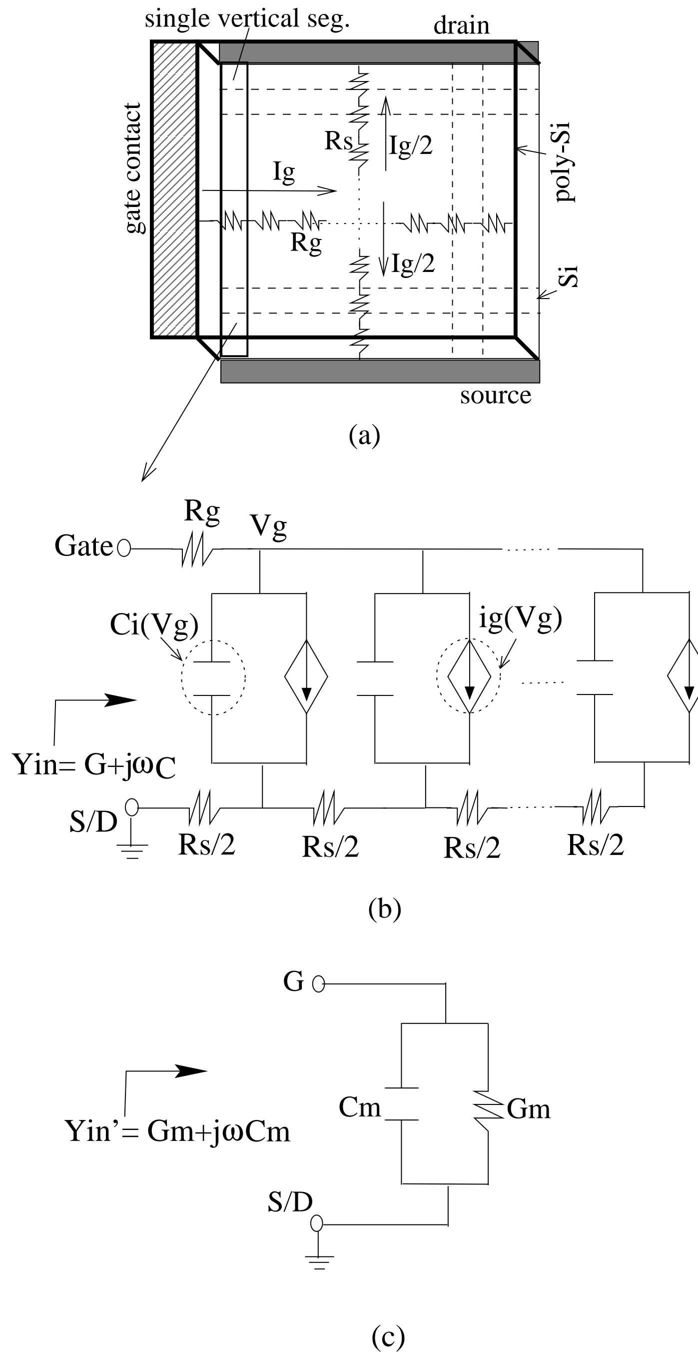


Figure 2.19: Equivalent circuit approach for ultra-thin gate oxide MOS with the source/drain grounded. (a) top view of the MOS structure with gate current flows. (b) distributed RC network used in this work. (c) RC circuit model used in capacitance measurements.

SPICE ac small-signal analysis is then performed to find the input admittance ($Y_{in} = G + j\omega C$) of the circuit at the given frequency (i.e. 100 kHz). The $C - V$ data including gate tunneling effects are finally obtained by taking the imaginary component of Y_{in} (i.e., $C = Im(Y_{in})/\omega$).

Figure 2.20 shows modeling results compared to the measurements for oxide thicknesses of 1.3, 1.5 and 1.8 nm, when the number of RC segments is 20. The sharp transition point in capacitance becomes lower as gate oxide thickness is reduced, which results from the increase in gate tunneling current with an exponential - like dependence between tunneling current and oxide thickness.

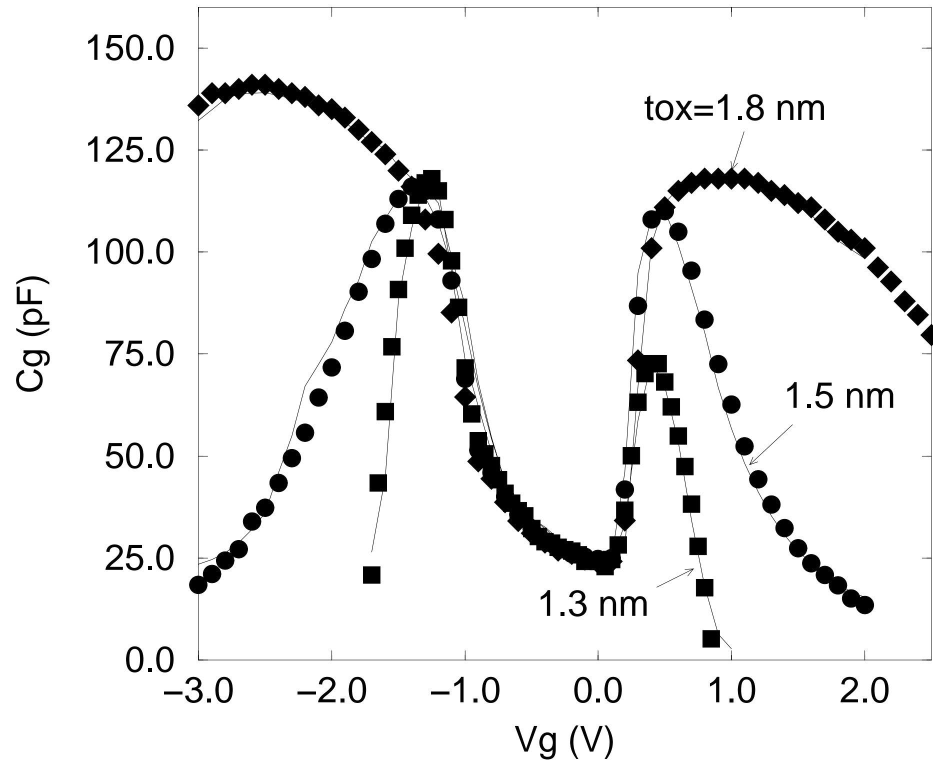


Figure 2.20: Simulated abnormal gate capacitance ($Im(Y_{in})$) obtained from the AC network analysis using the equivalent circuit in Figure 2.19(b) (lines), symbols represent measured HP data for t_{ox} = 1.3, 1.5, and 1.8 nm.

2.2.7 Channel Length Dependence of Gate Current

Contrary to general opinion, it was shown for the first time by Momese $et\ al.$ [41] that the direct tunneling gate current does not affect the MOSFET operation for devices down to the 1.5 nm gate oxide thickness with gate lengths less than 1 μm. The behavior of gate tunneling current strongly depends on the drain bias effects in on-state operation, especially in short channel MOSFETs, because the drain field easily penetrates to the channel region close to the source areas in the short-channel devices, as in Figure 2.21. In this figure, the potential near the source is higher than 0 V, and becomes especially severe for short-channel devices, which is the so-called drain-induced barrier lowering (DIBL) effect.

Gate tunneling current with respect to the drain bias (i.e., $V_d > 0$ V) is modeled using an equivalent circuit approach, as shown in Figure 2.22. Here, the channel length dependence of the gate tunneling current (I_g), considering the short channel effects, can be estimated due to a distributed channel resistance [42]. As a result, the simulated gate tunneling current is shown to decrease as the drain bias increases since the potential difference between the gate and the Si surface becomes smaller, as shown in Figure 2.23(a). In addition, effects of gate tunneling current on the drain current become less serious as the channel length is reduced, owing to the exponentially decreasing gate current. Figure 2.23(b) shows the simulated gate-channel dependence of I_g; I_g decreases in inverse proportion to $L_g^{1.5}$, which is comparable to the experiments by Momose $et\ al.$ [41][43], where the slopes were 1.8 and 1.5, respectively. This implies that effects of gate tunneling current on the drain current become less serious for the short channel MOSFET devices as the gate current decreases exponentially with reduced channel length. However, the total gate current for the entire chip may cause a serious problem for battery applications due to high off-state current; selective use of the thin gate oxide devices on a chip may be one solution [9].

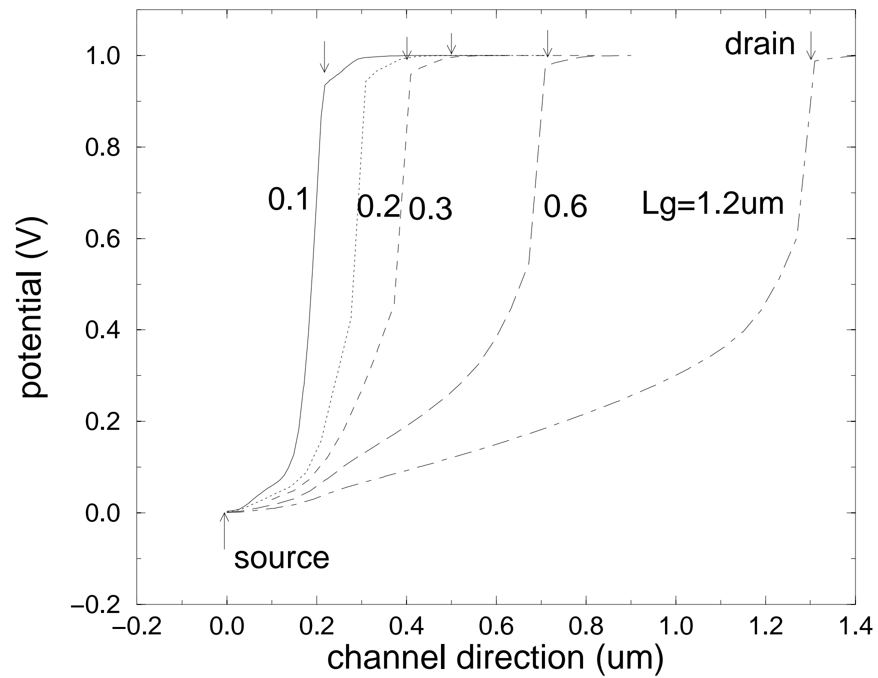


Figure 2.21: Quasi-Fermi potential distribution along the channel at $V_{gs} = V_{ds} = 1.0$ V for various channel lengths.

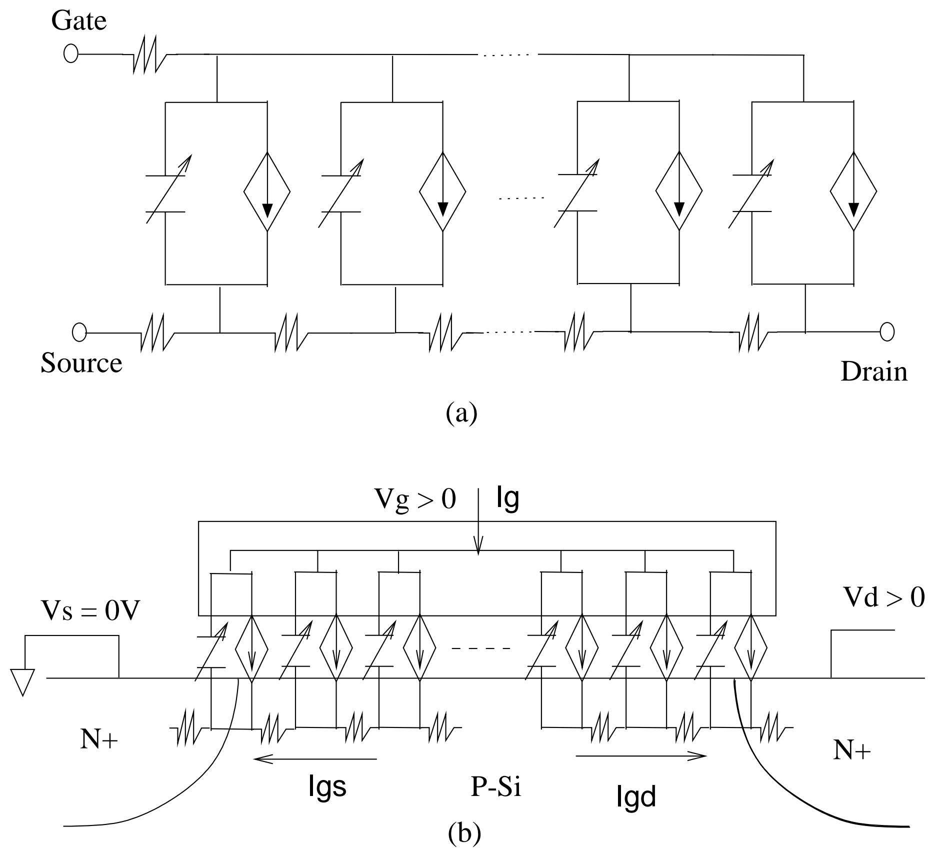
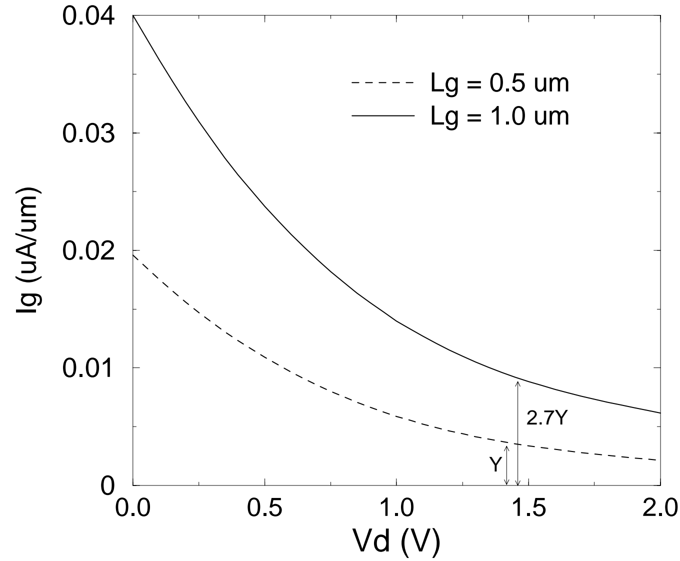
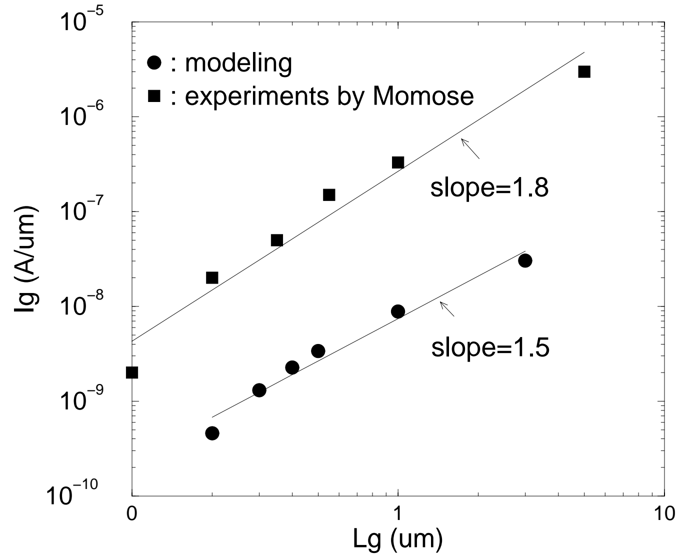


Figure 2.22: Modeling of gate tunneling current considering the drain bias effects, (a) equivalent circuit for a NMOS when the drain bias is given, (b) gate tunneling current ($I_g = I_{gs} + I_{gd}$).



(a)



(b)

Figure 2.23: Drain bias and gate channel length dependence of gate tunneling current of t_{ox} = 1.5 nm, (a) I_g versus V_{ds} for L_g = 0.5 and 1.0 μm when V_{gs} = 1.5 V, note that I_g of L_g = 1.0 μm is more than twice than that of L_g = 0.5 μm, (b) experimental ([43]) and modeled I_g with respect to L_g's at saturation mode ($V_{gs} = V_{ds}$ = 1.5 V).

2.2.8 Summary

MOS C–V characteristics of gate oxide in the sub 2.0 nm region are modeled with an empirical, hybrid formulation for QM corrections implemented in a 2D device simulator. Discrepancies in C–V modeling exist as the gate oxide becomes thinner than 2.0 nm due to the effects of gate tunneling current. The sharp decrease in capacitance for gate oxides below 2.0 nm in both the accumulation and inversion regions is modeled by using a distributed RC network that includes the gate tunneling current, which is calculated using a Green's function solver, NEMO.

2.3 C–V Reconstruction

Recently, alternate insulating materials with a dielectric constant much larger than that of SiO_2 are under evaluation to replace the conventional SiO_2 gate dielectrics and thereby overcome gate tunneling problems [44]. In particular, nitride/oxide (N/O) composite layer gate dielectrics have emerged as a promising alternative to SiO_2 in CMOS devices [45]. The composite N/O gate prevents boron diffusion from polysilicon gates during thermal dopant activation anneals. In addition, carrier mobility is enhanced because the ultra-thin buffer oxide of the composite layer improves the interface quality at the Si substrate. More importantly, lower tunneling currents can be obtained compared to pure oxide gates, due to the increased physical film thickness associated with the higher dielectric constant of the nitride layer ($\epsilon_{nitride} \sim$ 7.8). It was reported that 1.6 nm oxide equivalent N/O dielectric MOSFETs showed more than a 100 fold reduction in tunneling current compared with 1.6 nm single layer oxide devices [45]. However, even though tunneling current is much reduced in N/O composite MOSFETs compared to SiO_2 dielectrics, the $C-V$ curves of very-thin N/O MOSFETs still show capacitance attenuation. Therefore, a reconstruction technique of gate capacitance from the distorted $C-V$ is necessary to predict device performance [46].

2.3.1 Previous Reconstruction Method

The $C-V$ reconstruction technique reported previously in the literature [24] is efficient and useful for small-area MOS devices with low leakage current, where the $C-V$ curve is reconstructed from a derived equation, consisting of measured capacitance and conductance values obtained at two different ac frequencies (e.g., f_1=50 kHz, f_2=100 kHz or f_1=100 kHz, f_2=1MHz). However, since the technique treats a large-area MOS device as a single RC circuit (see Figure 2.15(b)), it has potential problems for large area devices with higher

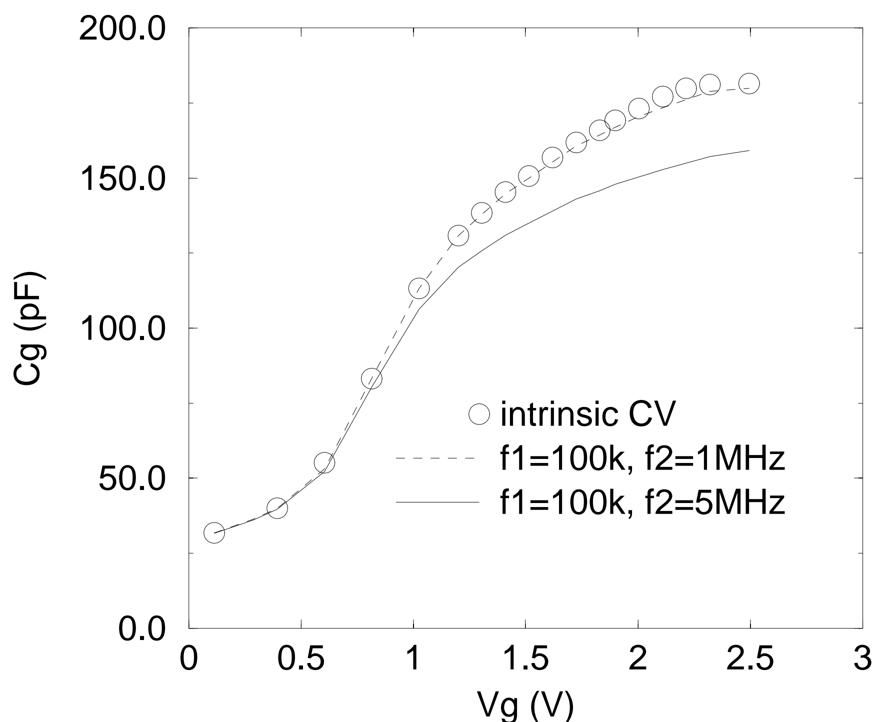


Figure 2.24: Reconstructed $C-V$ curves after the method in [24], circles represent the intrinsic $C-V$, and dotted and solid lines represent reconstructed $C-V$'s for the applied frequencies of 100 kHz–1 MHz and 100 kHz–5 MHz, respectively.

leakage currents. In practice the gate current and capacitance in long-channel devices is not uniform along the channel direction due to the distributed gate and series resistance effects; a single-lumped RC model is not an adequate circuit representation for large MOS devices. Also, choosing qualitatively dissimilar (measured) $C-V$ curves may cause substantial errors in the reconstruction process.

Figure 2.24 shows simulated $C-V$ curves for $100\ \mu$m $\times$ $100\ \mu$m P-MOSFETs after applying this method for reconstruction. Fairly large discrepancies are observed when

reconstructing the $C-V$ curve using (f_1=100 kHz, f_2=5 MHz). These $C-V$ curves are qualitatively dissimilar because the frequency of 5 MHz is above the cutoff frequency (f_T) of the device, estimated to be [47]:

$$f_T = \frac{\mu_n(V_G - V_T)}{2\pi {L_g}^2} \simeq 1.5\ MHz \tag{2.9}$$

where L_g = 100 μm, V_G = 2 V, μ_n = 600 $cm^2/V \cdot s$, and V_T = 0.4 V.

Thus, the single-lump RC small-signal model as in Figure 2.15(b) is no longer accurate for the case of 5 MHz; the series resistance becomes significant due to the low impedance of capacitance at higher frequencies. On the other hand, a low frequency $C-V$ measurement is not suitable for high leakage dielectric MOS capacitors since it causes considerable measurement error. The approximated instrumentation error for LCR meters is given as [34]:

$$error(\%) = 0.1\sqrt{1 + \left(\frac{G_m}{2\pi f C_m}\right)^2} \tag{2.10}$$

where C_m and G_m are measured capacitance and conductance, respectively. For example, the instrumentation errors calculated using Equation (2.10) are 12.7 % and 1.2 % for the applied frequency of 100 kHz and 1MHz, respectively; the measured G_m and C_m are 1.0 mS and 13.4 pF for a 1.5 nm oxide thickness (N-MOSFET) at V_g = 2.0 V. Hence, high frequency $C-V$ measurements and the incorporation of a distributed RC model are desirable in modeling of large-area, high-leakage dielectric MOSFETs.

2.3.2 Optimization Technique

For capacitance reconstruction, the distributed RC network for high leakage dielectric MOSFETs (Figure 2.19(b)) is used. The intrinsic gate capacitance, which is to be extracted, is modeled with a voltage-controlled capacitance ($C_i(V_g)$). The conductance is

modeled using a voltage-controlled current source ($i_g(V_g)$), which has been determined based on measured gate current data. Capacitance reconstruction consists of extraction of $C_i(V_g)$ from the distributed RC network such that the imaginary part and real part (C and G) of the simulated input admittance (Y_{in}) can be matched with the measured capacitance and measured conductance (C_m and G_m), shown in Figure 2.19(b),(c).

For the R_s extraction, sheet resistance (R_{sh}) of the device is calculated using conventional drift-diffusion theory of carrier transport [27]:

$$R_{sh}(x) = \frac{\frac{d}{dx}\phi_n(x)}{I_{ds}} \tag{2.11}$$

where x is the direction parallel to the Si surface and ϕ_n is the electron quasi-Fermi level obtained from 2D device simulation. I_{ds} is independent of x due to current continuity; the sheet resistivity is simply proportional to the lateral gradient of $\phi_n(x)$. This expression is valid as long as the device is sufficiently wide and the current flow is dominantly parallel to the x-direction.

Using device simulation and Equation (2.11), R_s was determined to be in the range between 60 – 70 Ω for a 100 μm $\times$ 100 μm P-MOSFET structure with a peak dopant concentration of 1.3×10^{18} cm^{-3}. R_g used in the extraction process was fixed at 5 $\Omega/\Box$, which is a typical sheet resistance of the polysilicon gate.

Given this set of parameters, the next step is to perform circuit optimization for the distributed RC network in order to extract the intrinsic capacitance ($C_i(V_g)$). A Levenberg-Marquadt algorithm implemented in HSPICE [48] is used which numerically combines the steepest-descent method with a Gauss-Newton approach to give the greatest stability for points far from the minimum while achieving rapid convergence for points near the minimum [49]. The quantity to be minimized is the norm of the error vector, $\boldsymbol{f}(\boldsymbol{p})$, given

by:

$$\|\boldsymbol{f}(\boldsymbol{p})\|^2 = \sum_k f_k(\boldsymbol{p})^2 = \sum_k \left[\frac{x_k(\boldsymbol{p}) - x_k^*}{\max(x_k^*, x_{min})}\right]^2 \tag{2.12}$$

where $\boldsymbol{p}$ represents the vector of optimized parameters and the vector $\boldsymbol{f}(\boldsymbol{p})$ consists of the errors evaluated at each gate bias point. x_k and x_k^* are the calculated capacitance/conductance and measured capacitance/conductance, respectively, at the k-th data point. For each iteration of the algorithm, the model is evaluated at each data point and the error vector $\boldsymbol{f}(\boldsymbol{p})$ is calculated; the parameter vector $\boldsymbol{p}$ is then adjusted to reduce $\|\boldsymbol{f}(\boldsymbol{p})\|^2$, the sum of the squares of the errors. For the $C_i(V_g)$ extraction, HSPICE ac small-signal analysis is performed, coupled with optimization to find the input admittance ($Y_{in} = G + j\omega C$) for the circuit.

2.3.3 Application Results

The N/O composite dielectric P-MOSFETs (N/O $\sim$ 1.4 nm/0.7 nm) with p^+-poly gates implanted with BF_2 (30 keV, 5×10^{15} cm^{-2}) were fabricated on 1.3×10^{18} cm^{-3} n-type Si <100> substrates. The bottom oxide layer was formed using N_2O remote-plasma oxide and the RPECVD nitride was then deposited using SiH_4 and N_2 as source gases to form the N/O structures. The nitride thickness is determined from the deposition rate and the oxide thickness was extracted using Auger electron spectroscopy [45]. $C - V$ measurements for the nitride/oxide P-MOSFET (N/O $\sim$ 1.4 nm/0.7 nm) were performed using the RC parallel mode, as shown in Figure 2.15(a). The device area is $100\ \mu$m $\times$ $100\ \mu$m and the applied small-signal frequency is 1 MHz. Measured $C - V$ data are represented by circles in Figure 2.25(a) and show a sharp decrease of the capacitance in the accumulation region for gate biases greater than 1.7 V. The solid line in the figure represents the reconstructed gate capacitance from the method described above. The reconstructed gate capacitance

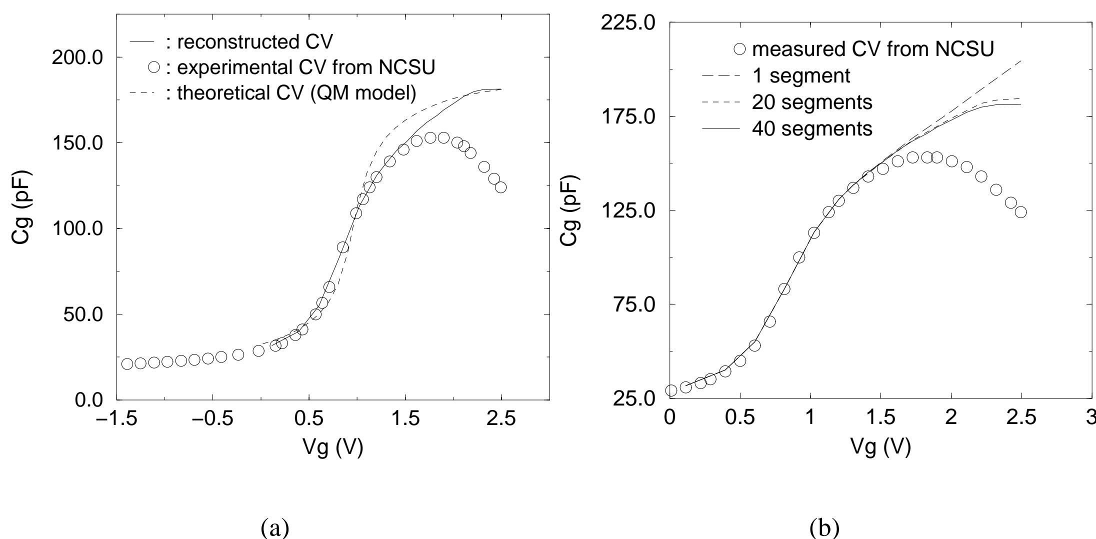


Figure 2.25: $C - V$ reconstruction results using the proposed technique. (a) measured (circles) from the North Carolina State University (NCSU) and reconstructed (solid line) $C - V$ curves by using the distributed RC network and optimization technique, dashed line indicates theoretical $C - V$ with QM device simulation. (b) comparison of extracted $C - V$ curves with respect to the number of segments in the RC ladder network, circles denote the measured anomalous $C - V$ and lines show the reconstructed intrinsic capacitance for the different number of segments–1, 20 and 40 segments.

of the N/O composite MOS device in the deep accumulation region is 180.4 pF, which corresponds to the gate capacitance of an equivalent oxide thickness of 1.9 nm ($t_{ox,eq}$ = 1.9 nm), assuming a classical charge model. Also, the capacitance corresponds to a gate capacitance with an equivalent oxide thickness of 1.4 nm ($t_{ox,eq-qm}$ = 1.4 nm), assuming the QM model.

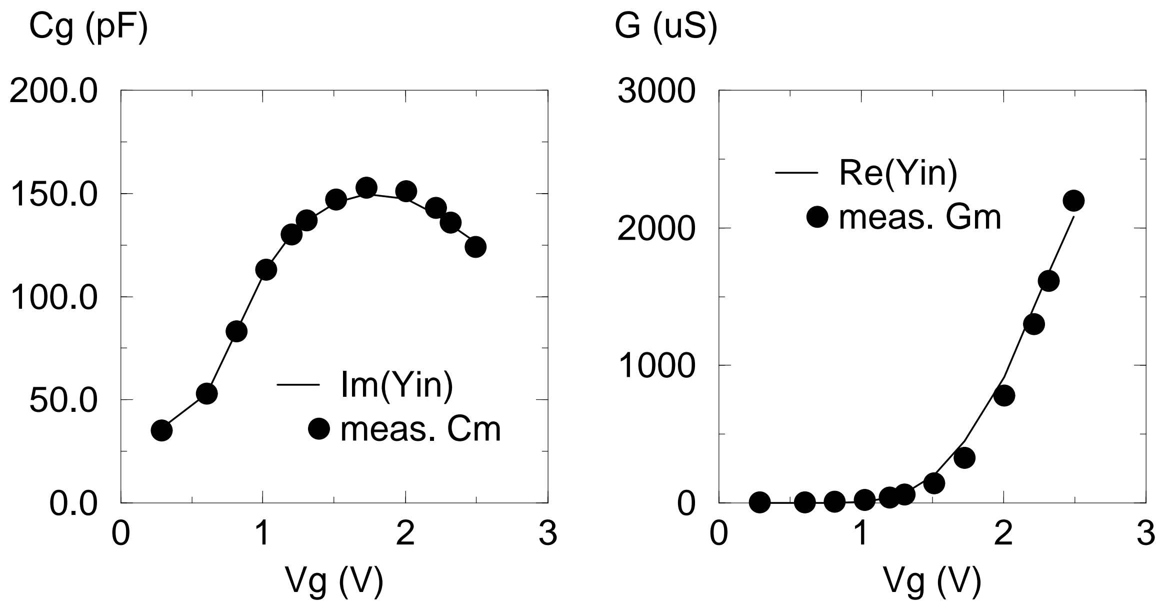


Figure 2.26: Optimization results, measured C_m/G_m from NCSU, compared to final C/G ($Im(Y_{in})/Re(Y_{in})$) after the $C-V$ reconstruction.

Regarding validation of the reconstructed $C-V$ curves, direct comparison with the experimentally reconstructed curves is the most desirable. For such a comparison, however, test structures with channel lengths of 1–10 μm are necessary to avoid capacitance attenuation, as reported in [26]. Instead, theoretical $C-V$ calculations are used, where the QM model has been considered for N/O ($\sim$ 1.4 nm/0.7 nm) dielectric P-MOSFET. The objective of the theoretical calculation is to find a reasonable capacitance range for the layer thickness, not to produce a perfect fit. As a result, the reconstructed gate capacitance shows good correspondence to the theoretical curve in deep accumulation, as shown in Figure 2.25(a); discrepancies between the theoretical and reconstructed curves may be attributed to the uncertainties in the series resistance value and not having modeled interface state properties specifically [33].

Figure 2.25(b) shows the reconstructed $C-V$ curves with respect to the number of segments in the RC ladder network. The shape of the reconstructed $C-V$ for a single segment case is unrealistic, while the reconstructed capacitance values in deep accumulation are almost identical for 20 and 40-segment cases compared to the theoretical calculations.

Figure 2.26 shows comparison between the measured C_m/G_m and the simulated C/G (= $Im(Y_{in})/Re(Y_{in})$) when the final intrinsic capacitances have been extracted, the optimization error between the measured and the simulated capacitance/conductance is 2 %.

2.3.4 Summary

A reconstruction technique for $C-V$ curves, based on distorted $C-V$ curves in high leakage dielectric MOSFETs has been developed. Anomalous $C-V$ behavior has been modeled using a distributed RC network to account for the QM effects as well as the distributed RC nature of MOSFETs; the intrinsic, pure gate capacitance is conversely extracted from the anomalous measured $C-V$ curves. The reconstructed $C-V$ data is comparable

to the theoretical QM calculations for very-thin nitride/oxide gate dielectric ($t_{ox,eq-qm} \sim$ 1.4 nm) MOSFETs.

2.4 Performance-based CMOS Scaling

Remarkable improvements in CMOS VLSI have been achieved using the scaling rules proposed by R. Dennard $et\ al$ [50]. However, traditional scaling theories based on constant voltage or electric field have difficulty to be extended into future nanometer-regime scaling due to physical limitations of the parameters. Gate oxide suffers from the quantum mechanical and direct tunneling effects for thicknesses less than 3.0 nm and parasitic capacitance and source/drain series resistance becomes important. Polysilicon gate depletion effects also reduce gate capacitance and inversion charge density because the voltage drop across the gate oxide and substrate is reduced in the presence of the depletion region in the polysilicon, which cannot be neglected for oxides thinner than 4.0 nm. It is reported that 2.0 nm oxide results in a loss of about 20 % for inversion charge at 1.5 V gate voltage due to the QM and polysilicon depletion effects [2]. On the other hand, as the miniaturization of MOSFET proceeds, parasitic components of transistors such as junction capacitance and source/drain resistance act to degrade the switching speed [51]. Thus, it is desirable to develop new scaling guidelines that take into account the above effects in choosing design parameters for future CMOS technology.

2.4.1 New CMOS delay model

The CMOS propagation delay (t_{pd}) can be defined as [52],

$$t_{pd} = \frac{C_L V_{dd}}{a} \left(\frac{1}{I_{dsat,n}} + \frac{1}{I_{dsat,p}} \right) \tag{2.13}$$

where 'a' is a constant with a typical value of 4.0 and C_L is the loading capacitance. Let $k = \frac{W_p}{W_n}$ and assume that the junction capacitances (C_j) of NMOS and PMOS are identical so that:

$$C_L = W_n\,(k+1)(C_j + C_{ov} + mFL_gC_{ox}) + C_{int} \tag{2.14}$$

where C_j and C_{ov} are the junction and the overlap capacitance per unit gate width, respectively, and C_{int} is the lumped interconnection capacitance. The 'm' factor reflects the Miller effect, channel charge partitioning, fringing capacitance, and other effects of devices when operated in an inverter. Besides, $C_{ox} = \epsilon_{ox}/t_{ox}$, and F is fanout.

Lundstrom has proposed a drain current equation including carrier velocity effects at the source end of the channel [53], so that an approximate I_{dsat} including source resistance, r_s, can be obtained when $V_{gs} = V_{dd}$ as [54],

$$I_{dsat} \approx W\ v_{ch}Q_i = W\ v_{ch}C_{ox}\,(V_{gs} - I_{dsat}\,r_s - V_{th}) \tag{2.15}$$

$$I_{dsat} = W\epsilon_{ox}\ v_{ch}\ \frac{V_{dd} - V_{th}}{t_{ox} + W\epsilon_{ox}v_{ch}r_s} \tag{2.16}$$

where v_{ch} is carrier velocity at the source end of the channel, $v_{ch} = v_T/(1 + \frac{v_T}{\mu_{eff}E_s})$ with v_T being the thermal velocity ($\sim 6.5 \times 10^6 cm/s$ for NMOS and $\sim 5.8 \times 10^6 cm/s$ for PMOS at room temperature [56]). E_s is related to v_{ch}, the electric field along the channel, and can be approximated as $E_s = \frac{V_{dd}}{L_{eff}}$. The effective mobility, μ_{eff}, has a relationship with V_{th} and t_{ox} as [52][55],

$$\mu_{eff,n} = \frac{540}{1 + \left(\frac{v_{dd}+V_{th}}{5.4t_{ox}}\right)^{1.85}} \tag{2.17}$$

$$\mu_{eff,p} = \frac{185}{1 + \left(\frac{v_{dd}+1.5V_{th}}{3.38t_{ox}}\right)} \tag{2.18}$$

Finally, combining the above set of empirical and first-order approximations a concise inverter delay model, parameterized in terms of t_{ox} is proposed as follows:

$$t_{pd} = \frac{k+1}{a} \left(\frac{m\ F\ L_g}{t_{ox}} + \frac{c_j + c_{ov}}{\epsilon_{ox}} + \frac{C_{int}}{W_n(k+1)}\right) \cdot \left[\frac{W_n\epsilon_{ox}\ v_{ch,n}\ r_{s,n} + t_{ox}}{v_{ch,n}\ (1 - \frac{V_{th,n}}{V_{dd}})} + \frac{W_p\epsilon_{ox}\ v_{ch,p}\ r_{s,p} + t_{ox}}{k\ v_{ch,p}\ (1 - \frac{V_{th,p}}{V_{dd}})}\right] \tag{2.19}$$

2.4.2 Model Evaluation

In order to evaluate the new delay formula, a computational model for a hypothetical 0.1μm NMOS device is built by using two-dimensional process simulation with gate oxide thickness of 20 Å. Shallow n^+ source/drain extensions are used in conjunction with deeper n^+ region, implemented after spacer etch. Trench isolation is included to extract accurate junction capacitance. In this case, QM and polysilicon depletion effects are not considered. Moreover, V_{dd} = 1 V and the threshold voltages of the N and PMOS transistors are $V_{th,n}$ = 0.22 V, $V_{th,p}$ = 0.19 V, respectively.

For extraction of the effective channel length (L_{eff}) and source resistance (r_s) the paired V_g algorithm [57] is used. The drain-source resistance of the MOSFET linear region expression is generalized as:

$$R_{tot}(V_{gs}) = R_{ch}(V_{gs}) + R_{sd}(V_{gs}) \tag{2.20}$$

where R_{ch} is channel resistance, and R_{sd} represents all resistance outside the channel that includes source/drain contact and series resistances. ΔL defined as the difference between the drawn gate length (L_{drawn}) and the effective channel length (L_{eff}) is given by:

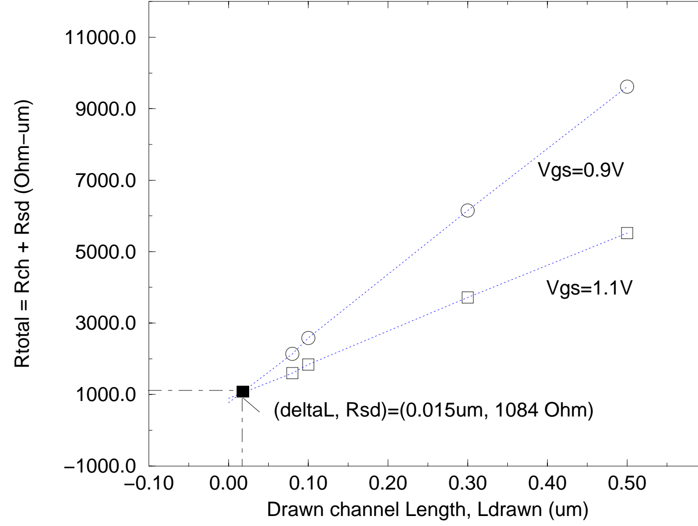


Figure 2.27: Determination of effective channel length ($L_{eff} = L_{drawn} - \Delta L$) and source/drain resistance ($R_{sd} = R_{tot} - R_{ch}$) using the paired V_g method.

$$L_{eff}(V_{gs}) = L_{drawn} - \Delta L(V_{gs}) \tag{2.21}$$

In the algorithm, plotting R_{tot} versus various L_{drawn} is required in order to extract ΔL and R_{ext} at the intersect of two lines corresponding to closely spaced curves separated by V_{g1} and V_{g2}, as shown in Figure 2.27. As a result, for 0.1 μm NMOS, extracted ΔL is 0.015 μm and $r_{s,n}$ and $r_{s,p}$ are 542 $\Omega \cdot \mu$m and 460 $\Omega \cdot \mu$m, respectively, when V_{g1} = 0.9 and V_{g2} = 1.1 V. The extracted junction capacitance (C_j) and overlap capacitance (C_{ov}) from device simulation are 0.60 and 0.15 fF/μm, respectively.

Figure 2.28 shows a comparison of saturation currents between device simulation and

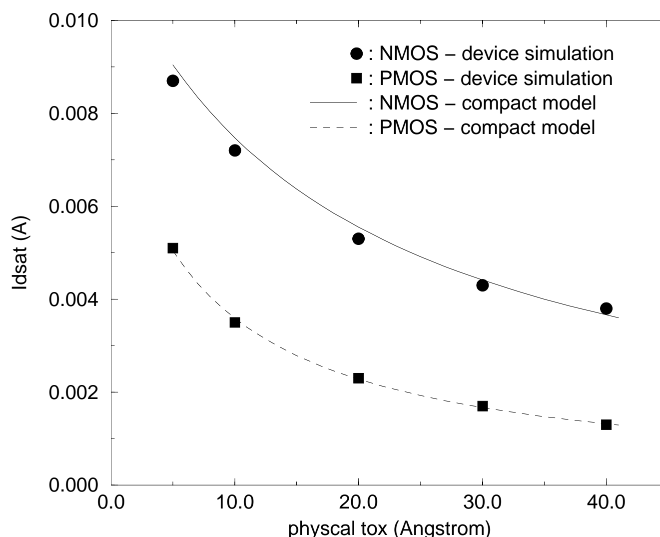


Figure 2.28: Comparisons of Idsat values between obtained from device simulation and Equation (2.16) as a function of physical gate oxide thickness.

the analytic Equation (2.16) for various physical gate oxide thicknesses ranging from 5 – 40 Å at the bias condition of $V_{ds} = (V_{gs} - V_{th})$ = 1 V.

Figure 2.29 shows propagation delays obtained from the proposed model as functions of physical gate oxide thickness and fanouts of 1, 2, and 3, when C_{int} = 0.1 fF. The values of the empirical factors such as ‘a’ and ’m’ are 4 and 1.56, respectively. Symbols represented as circles, squares, and diamonds indicate the results extracted from the device simulation with a sequence as follows: i) 2D device simulation for I-V and C-V data, ii) DC & AC parameter extraction, iii) SPICE simulation for propagation delays. Drive current degradation owing to QM and polysilicon depletion effects has not been taken into account.

Again, Figure 2.30 shows the comparison between the experimentally measured and

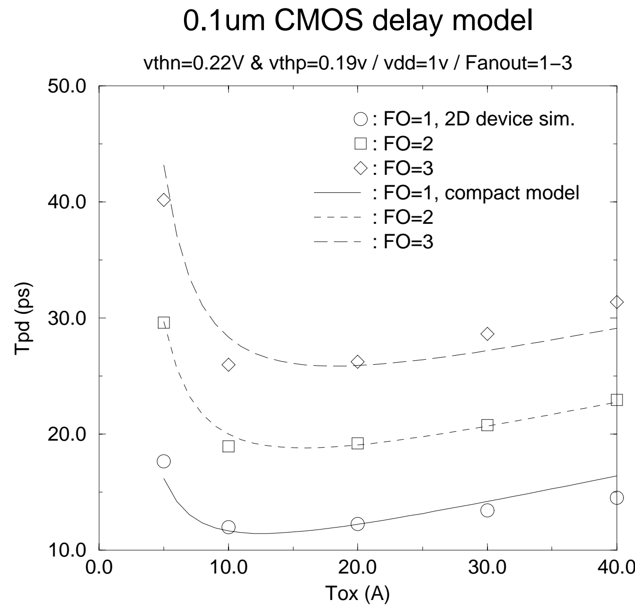


Figure 2.29: Calculated CMOS propagation delays as functions of the physical gate oxide thickness and the fanouts.

simulation MOS capacitor C-V curves when oxide thickness is 20 Å for an area of 100 $\mu m\times$ $100\ \mu m$. The results of device simulation via the QM correction model agree well with the measurements both in the accumulation and inversion regions, while discrepancies exist for the classical drift-diffusion model. The reduction of gate capacitance in the accumulation region is related to the QM effect, and the reduction in the inversion region is related to both the QM and the poly depletion effects. It is shown that 20 Å of gate oxide loses about 30% of the inversion charge due to these effects at $\sim V_g = 1.0$ V.

The normalized gate capacitance as a function of oxide thickness in the inversion and accumulation regions (C_{inv}/C_{ox}, C_{acc}/C_{ox}) are plotted in Figure 2.31. Here, it is obvious that the portion of gate capacitance attenuation becomes more significant as oxide thickness is scaled down.

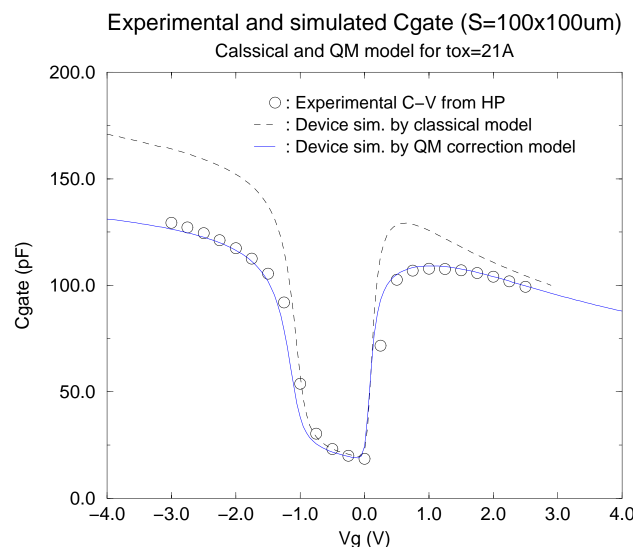


Figure 2.30: Comparisons of experimental (HP) and simulated MOS C–V (classical Fermi-Dirac and QM correction models).

The effective oxide thickness is determined from a regression model as a function of the physical oxide thickness. Also, a linear regression model is derived for the relationship between the effective oxide thickness and physical oxide thickness and can be expressed as '$t_{ox,eff} = 1.00\ t_{ox,phy} + 10.59$', as shown in Figure 2.32. The accuracy of this extraction method is confirmed by Lo and $et\ al.$, with measurements using ellipsometry [58]. Thus, the effective oxide thickness under 20 Å is also projected, which is not easily achievable with current technology. As a result, for instance, the estimated effective oxide thickness of t_{ox} = 5 Å corresponds to 15.59 Å.

Figure 2.33 shows simulation results obtained from the new CMOS delay model, considering the effective oxide thickness obtained from the linear regression expression. Unlike Figure 2.29, an abrupt increase of t_{pd} disappears for gate oxide thicknesses less than 10 Å.

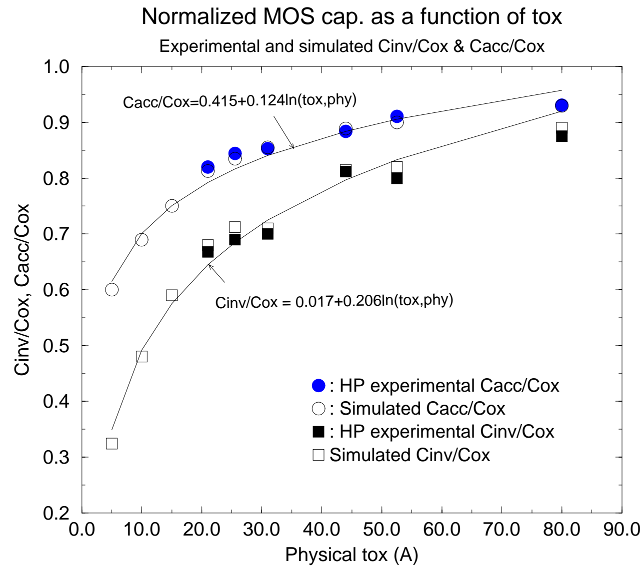


Figure 2.31: Normalized MOS gate capacitance versus physical oxide thickness.

2.4.3 Delay Optimization

There are several parameters affecting the CMOS delay time, and in order to reduce the overall delay, parameter optimization is applied using the model to target delay results. The Levenberg-Marquadt algorithm in Equation (2.12) can also be applied to delay optimization with respect to the calculated and target CMOS delay times.

Figure 2.34 shows the predicted propagation delay as a function of effective oxide thickness. In order to reduce delay times, parameter optimization is applied using the delay model to the targets denoted as filled symbols.

Table 2.1 shows the optimization results for the target delays being 10, 16, and 22 pS when the number of fanout is 1, 2, and 3, respectively (V_{dd} = 1 V, $t_{ox,phy}$ = 20 Å). As a result, gate length (L_g) and threshold voltages (V_{thn}, V_{thp}) are shown to be the dominant parameters in determining a target delay; junction capacitance (C_j) follows as the third

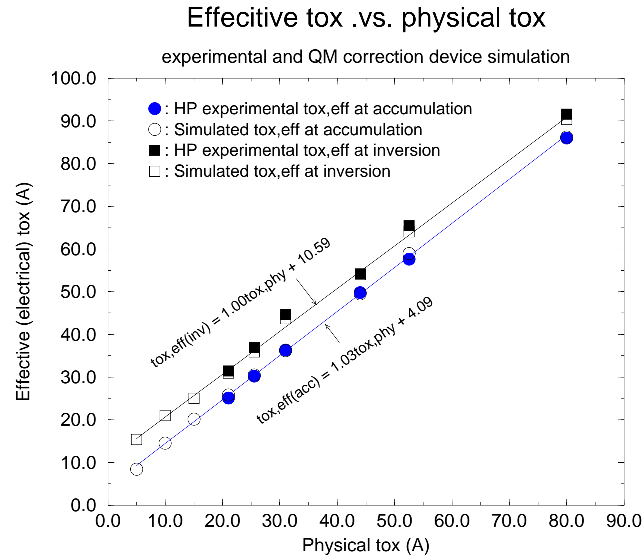


Figure 2.32: Effective (electrical) gate oxide thickness (t_{ox}) as a function of physical oxide thickness.

Table 2.1: Optimized parameters for target delays being 10, 16, and 22 ps when fanout is 1, 2, and 3, respectively (V_{dd} = 1 V, $t_{ox,phy}$ = 20 Å).

parameters	**sensitivity (%)**	**optimized value**
V_{thp}	24.71	0.15 V
L_g	24.18	0.1 μm
V_{thn}	21.34	0.15 V
C_j	10.46	0.2 fF/μm
r_{sn}	9.69	280 $\Omega \cdot \mu m$
r_{sp}	5.94	343 $\Omega \cdot \mu m$
C_{ov}	3.66	0.07 $fF/\mu m$
C_{int}	0.0	0.1 fF

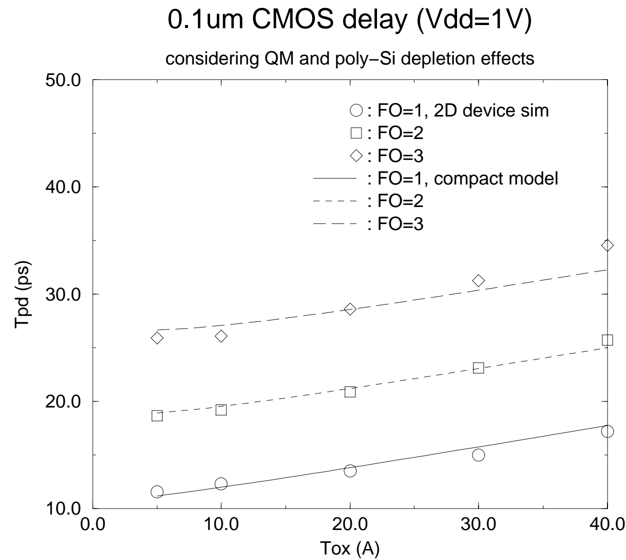


Figure 2.33: Effective (electrical) gate oxide thickness (t_{ox}) as a function of physical oxide thickness.

most important parameter. The RMS error between the calculated and targeted delay values is within 4 %.

2.4.4 Summary

A design approach is demonstrated using a new CMOS delay model. Effective oxide thickness is determined using MOS capacitor C-V data both from experimental and QM-corrected device simulation; results are applied to circuit level delay modeling. Optimization of circuit speed as a function of device parameters is achievable using the approach based on consideration of physical parameters as well as parasitic components.

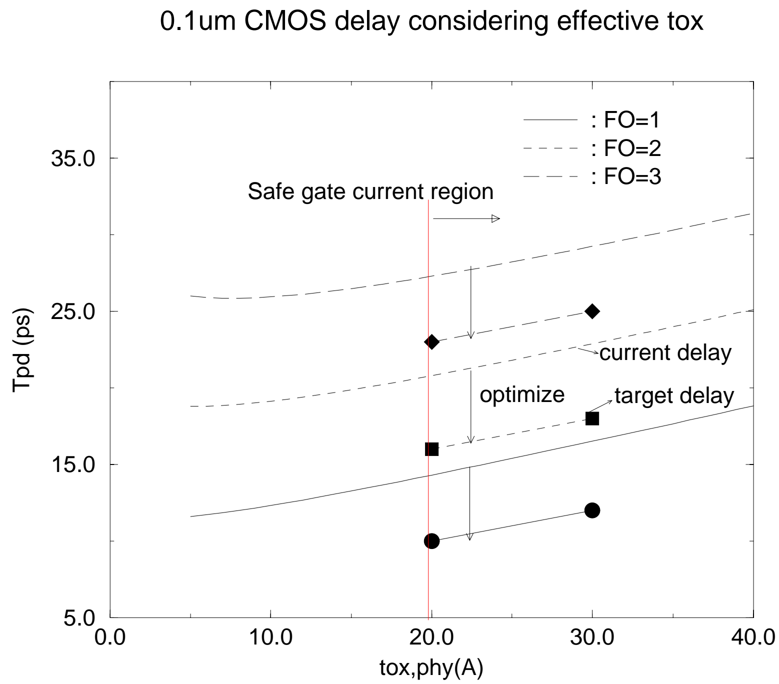


Figure 2.34: CMOS speed optimization using the compact delay model.

2.5 Direct Tunneling Current Model for Circuit Simulation

Direct tunneling current increases exponentially with decreasing oxide thickness and gate bias, which introduces a potential limitation for aggressive scaling of the gate oxide in advanced CMOS technology. Even though there are many reports concerning the effects of the gate leakage current on MOS transistor operation, few studies have been made regarding the impact of the gate current on circuit operation, due to lack of a circuit simulation model for gate tunneling current. For long channel devices with very thin gate oxides, device characteristics have been shown [41], where the gate current is of the same order of magnitude; this current significantly alters I-V characteristics. Meanwhile, the gate current effects on I-V characteristics become minimized as the channel length becomes shorter due to the electric field effects near the drain. Since gate current becomes an integral part of intrinsic device operation, circuit simulation which includes accurate modeling of tunneling characteristics is required. Hence, a compact gate tunneling current model for circuit simulation is needed to analyze circuit immunity against the gate leakage current including dependencies on architecture and operating conditions. In this section, a compact model for direct tunneling in very-thin gate oxide MOS transistors suitable for circuit simulation is discussed [59].

2.5.1 Surface Potential Based Tunneling Model

Figure 2.35 shows direct tunneling of electrons across the gate oxide from the p-type Si substrate to n^+-poly Si gate. When a positive gate bias is applied to the device with a grounded substrate, electrons in the inversion layer pass through the oxide quantum-mechanically.

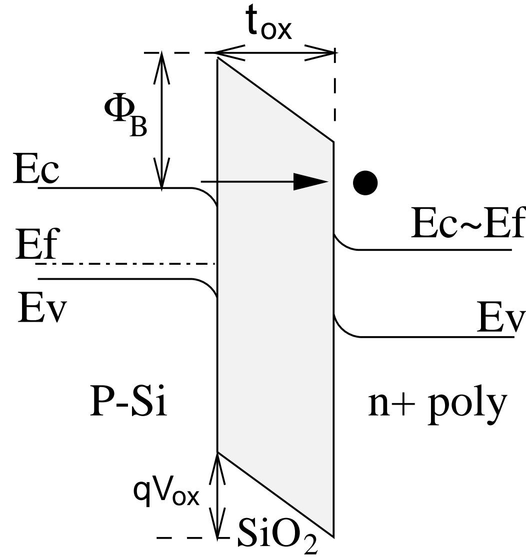


Figure 2.35: Direct tunneling of electron in n$^+$-polysilicon / SiO_2 / p-Si MOS structure ($V_{ox} < \Phi_B$).

For a negative gate voltage the holes in the accumulation layer can be the source of tunneling current.

This direct tunneling current model is expressed as [60],

$$J_{DT} = C(V_{ox}/t_{ox})^2 e^{\frac{-B(1-(1-V_{ox}/\Phi_B)^{3/2})}{V_{ox}/t_{ox}}} \tag{2.22}$$

where the pre-exponent C and slope B are given as:

$$C = \frac{q^3 m}{8\pi h m_{ox} \phi_b}, \quad B = \frac{8\pi(2m_{ox})^{1/2}}{3qh}\phi_b^{3/2} \tag{2.23}$$

where ϕ_b is barrier height between the emitting electrode and oxide expressed in electron volts, h is Plank's constant. Coefficients m and m_{ox} are electron effective mass in free space and in the oxide, respectively.

In order to apply Equation (2.22) to the calculation of gate tunneling current, it is necessary to relate the oxide voltage (V_{ox}) to the applied voltage (V_{gs}), since V_{ox} depends on V_{gs} as well as the surface potential (ψ_s) as follows:

$$V_{ox} = V_{gs} - V_{fb} - \psi_s \tag{2.24}$$

However, ψ_s can only be solved accurately using an iterative numerical approach, requiring expensive computations, which is not desirable for circuit simulation. In order to reduce computation time while retaining an accurate relationship between the surface potential and the terminal voltages, an explicit formulation for surface potential (ψ_s) [61] is used in this work.

Under the assumption of having a gradual channel and using the charge sheet approximation for an ideal n-type MOS transistor, the surface potential in the weak inversion region ($0 < \psi_s < 2\phi_f$) can be approximated as [29]:

$$\psi_{s,weak} = V_{gs} - V_{fb} + \frac{\gamma^2}{2} - \gamma\sqrt{V_{gs} - V_{fb} + \frac{\gamma^2}{4}} \tag{2.25}$$

where γ is the body factor defined by $\sqrt{2q\epsilon_{Si}N_a}/C_{ox}$.

In the strong inversion region ($\psi_s > 2\phi_f$) ψ_s becomes

$$\psi_{s,strong} = 2\phi_f + V \tag{2.26}$$

where V denotes the electron quasi-Fermi potential, ranging from V_{sb} at the source to V_{db} at the drain side. The surface potential at the drain node ($\psi_{s,drain}$) is modeled to consider

the drain bias (V_{db}) effect as:

$$\psi_{s,drain} = \psi_s(V_{db} + V_{sb}), \quad \psi_{s,source} = \psi_s(V_{sb}) \tag{2.27}$$

This enables reduction of gate tunneling current at the drain as the drain bias increases due to the decrease of the potential difference between the gate and drain (i.e. decrease of V_{ox} at the drain). In other words, as the drain bias is increased with the source at zero, the channel potential varies from the source to the drain , thus the oxide field is reduced towards the drain end of the device and the gate current decreases. In addition, ΔV_{gs} is introduced to model the channel length dependence of gate current caused by the drain-induced barrier lowering (DIBL) effect as:

$$\Delta V_{gs} = \delta_{DIBL}\sqrt{2\phi_f + V_{sb}} \cdot V_{db} \tag{2.28}$$

where $\delta_{DIBL} = \delta_O(\frac{L}{L_R})^{n_{dibl}}$, L_R is the reference gate length and n_{dibl} is an exponent of the length dependence. This ΔV_{gs} is added to V_{gs} in Equation (2.24). As a result, the effect of gate length dependence on gate tunneling current is taken into account because ψ_s is abruptly increased as the channel length is scaled down.

2.5.2 Quantum-Mechanical Effects

For gate oxide thicknesses less than 2.0 nm, quantum-mechanical effects become dominant. In the classical case, the electron density has its maximum value at the Si-SiO$_2$ interface, while in the quantum mechanical case the electron density is diminished at the interface, increases to its maximum value and decreases with the distance from the surface [29]. Hence, in the quantum-mechanical model the inversion charge profile peaks at around 10 Å below the silicon surface such that inversion charge is effectively reduced to that of an

equivalent oxide which is a few angstroms to one nanometer thicker than the physical oxide.

For calculation of the surface potential, considering the QM effects, the intrinsic carrier concentration (n_i), is evaluated by introducing a bandgap broadening mechanism which is approximated using van Dort's bandgap broadening approach which is expressed as [42]:

$$\Delta E_g = \frac{\beta}{2kT} E_n^{2/3} \tag{2.29}$$

where E_n is the normal electric field at the Si-SiO$_2$ interface which is gate bias dependent (determined by V_{ox} and t_{ox}) and β is a parameter that is determined experimentally. The value of ΔE_g is used to calculate a new intrinsic carrier concentration (n_i) and Fermi potential (ϕ_f).

2.5.3 Results and Discussions

Figure 2.36 shows the simulated oxide voltage (V_{ox}) with respect to gate bias, comparing the V_{ox} = V_g approximation, V_{ox} = V_g - V_{poly} [60], and the surface potential based model. Figure 2.37 shows simulated gate current using the compact model, the conventional methods [60][62] and results from the 1D Green's function solver, NEMO [38], that approximates the solution to the Schrödinger equation. Gate tunneling current in the low gate bias regime obtained from the compact model agrees well with that from the numerical solver as a result of the surface potential model accuracy in the weak inversion region. Comparing NEMO and measurements, simulated gate currents for different oxide thicknesses are shown in Figure 2.38. The simulated gate currents using the compact model agree well with those from NEMO for oxide thicknesses of 1.3, 1.5, and 1.8 nm due to consideration of the surface potential as well as quantum mechanical effects. Discrepancies between the

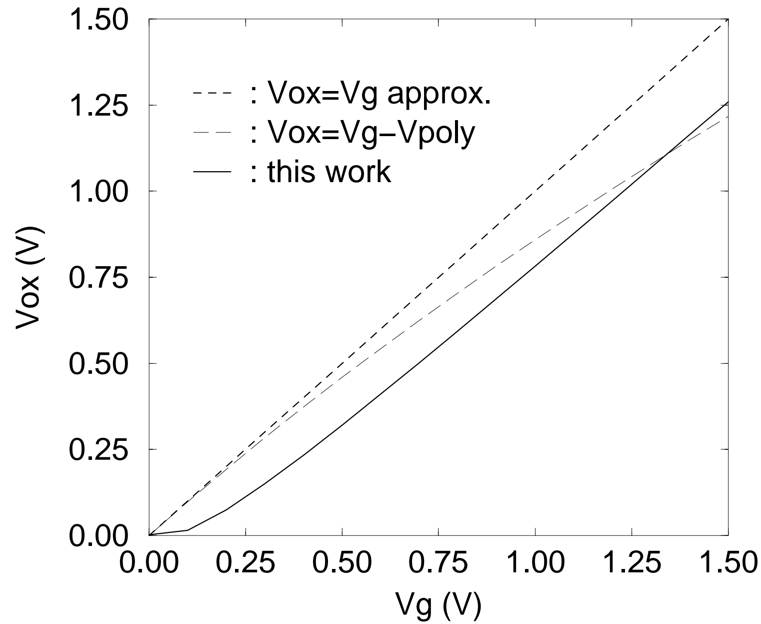


Figure 2.36: Voltage drop at oxide (V_{ox}) with respect to the applied gate bias for t_{ox} = 1.5 nm. Comparisons between $V_{ox} = V_g$ approximation, V_g - V_{poly}, and the surface potential based model.

measured and simulated data in Figure 2.38(b) are probably caused by surface roughness, uncertainty in determining the effective oxide thickness, and the IR drop at the poly gate and the channel due to the gate leakage current. In reality, the gate tunneling current behavior is affected by the three-dimensional geometric effects associated with gate (R_g) and series resistance (R_s), as shown in Figure 2.39 [42].

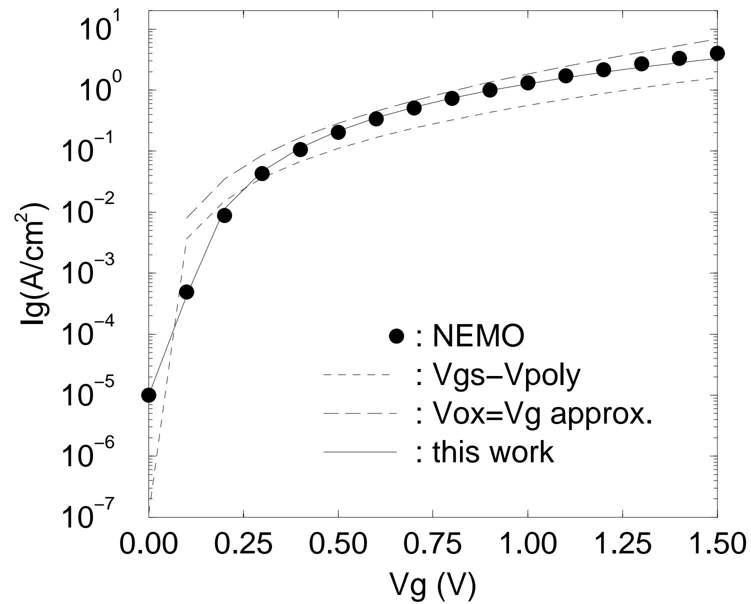
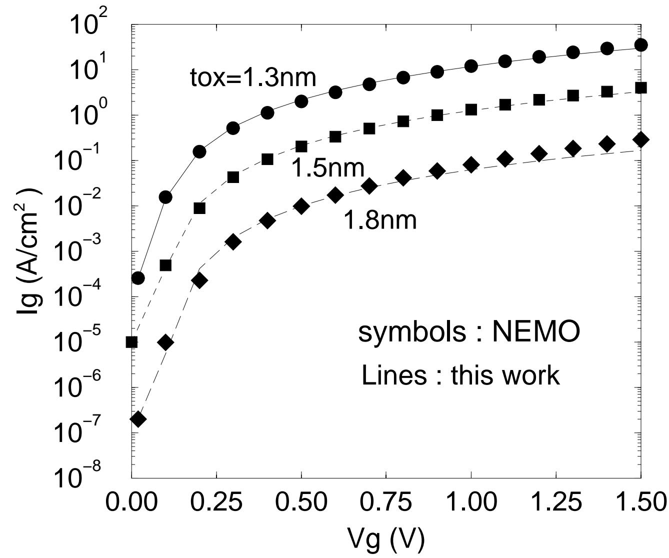
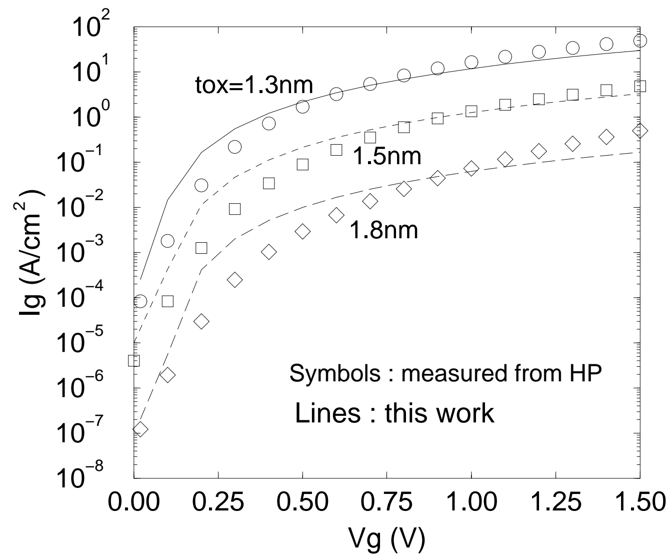


Figure 2.37: Simulated gate current curves using V_g approximation, V_g - V_{poly}, and the surface potential based model, symbols are obtained from NEMO, an approximated Schrödinger equation solver (t_{ox} = 1.5 nm).

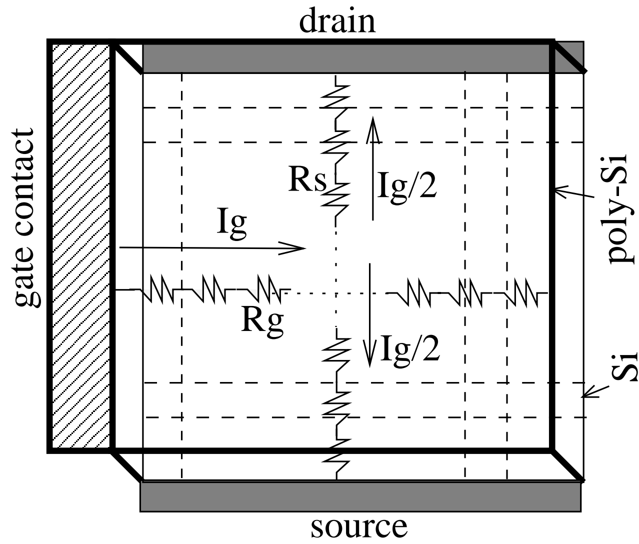


(a)

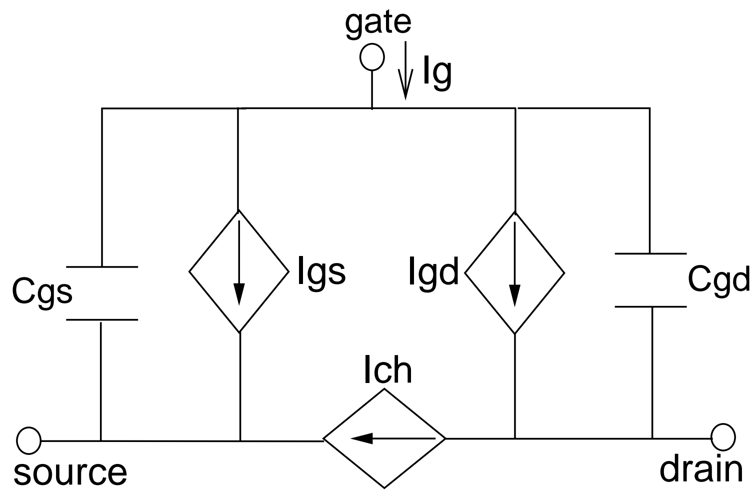


(b)

Figure 2.38: Results of the compact model. (a) gate currents vs. V_g by using the compact model for different t_{ox}'s (1.3, 1.5, and 1.8 nm), symbols denote gate current from the NEMO simulator [38]. (b) comparison between gate currents using the compact model and measured HP data for different t_{ox} values.



(a)



(b)

Figure 2.39: (a) Gate and series resistance (R_g and R_s) effects associated with gate current (I_g) in ultra-thin oxide, large area MOS transistor. (b) Circuit model of gate tunneling current for circuit simulation ($I_g = I_{gs} + I_{gd}$).

Overall, comparison of results for the compact model and numerical solver show good agreement due to consideration of the surface potential as well as quantum mechanical effects. In order to consider drain bias effects ($V_d > 0$ V) an equivalent circuit is used as in Figure 2.39(b), where gate current (I_g) is composed of gate-to-source current (I_{gs}) and gate-to-drain current (I_{gd}). Gate current in the drain (I_{gd}) is determined by surface potential at the drain ($\psi_{s,drain}$) for a given drain bias. Namely, I_{gs} and I_{gd} are computed independently by considering the surface potential of gate-to-source ($\psi_{s,source}$) and gate-to-drain ($\psi_{s,drain}$) in Equation (2.27). In this work, these voltage-controlled current sources, I_{gs} and I_{gd}, are described by utilizing the behavior-level modeling in SPICE; the BSIM3 model is used as a channel current (I_{ch}) model. As a result, as shown in Figure 2.40 the simulated gate current decreases as the drain bias increases due to the increase of surface potential at the drain, which leads to the V_{ox} reduction. The direction of gate current near the drain region can even become reversed when the potential of the drain is higher than that of the gate. Hence, total gate tunneling current becomes lower as the drain bias moves from the linear to saturation region operation; the gate current effect is dominant at high V_{gs} and low V_{ds} bias conditions.

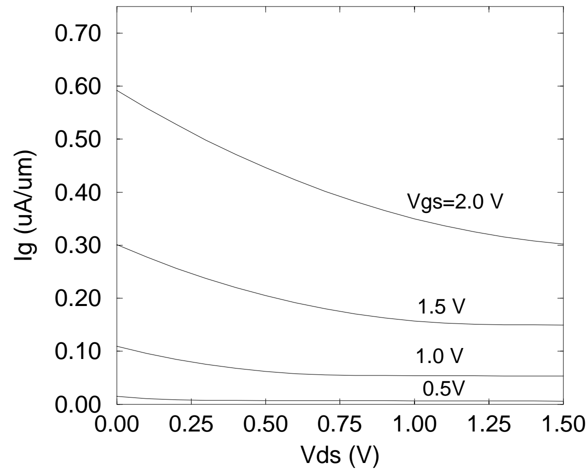


Figure 2.40: Simulated gate tunneling current versus V_{ds} for different V_{gs}, L_g = 10 μm and t_{ox} = 1.5 nm.

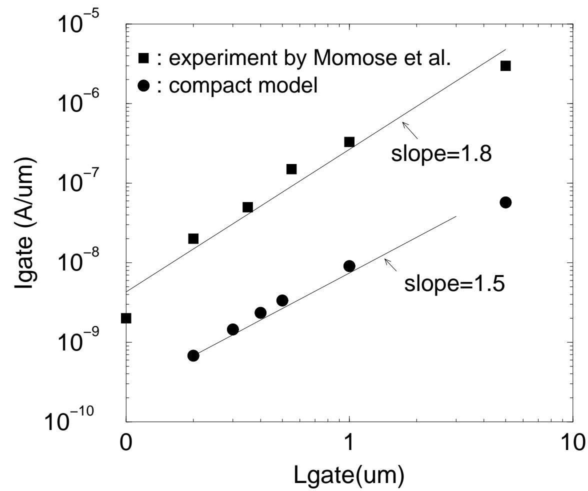
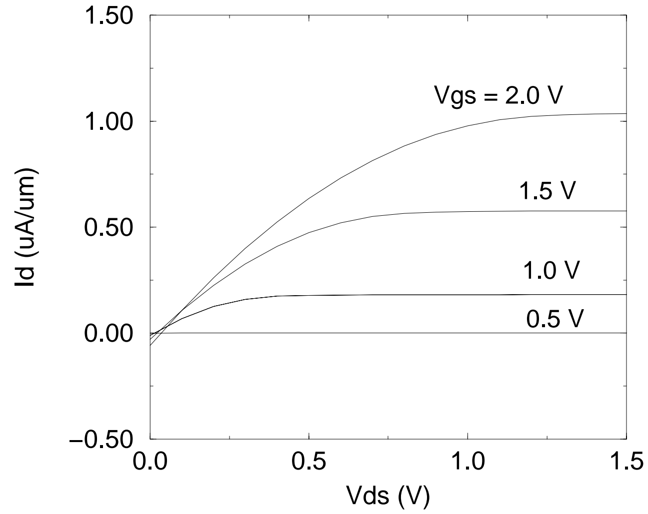
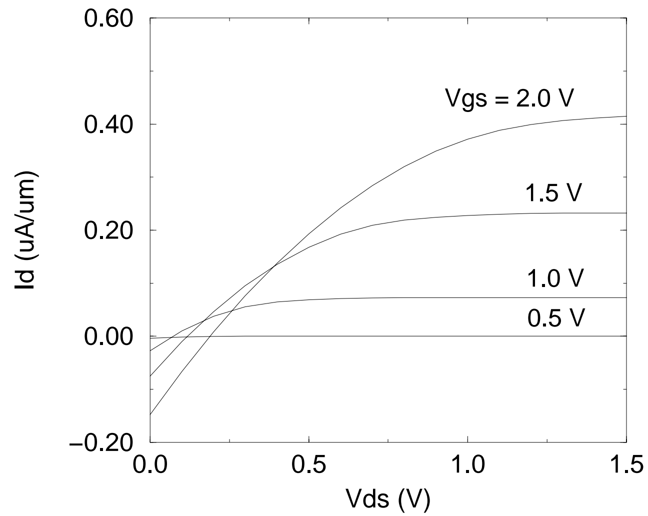


Figure 2.41: Gate length dependence of gate tunneling current at V_{ds} = V_{gs} = 1.5 V for t_{ox} = 1.5 nm, the simulated slope is comparable with ref. [41] due to consideration of DIBL.



(a)



(b)

Figure 2.42: Simulated drain current (I_d) versus V_d for t_{ox} = 1.5 nm, (a) L_g = 20 μm (b) L_g = 50 μm.

Figure 2.41 illustrates the simulated channel-length dependence of gate current at V_{ds} = V_{gs} = 1.5 V by consideration the DIBL effect; I_g decreases in proportion to $L_g^{1.5}$, which is comparable to the slope of the experiment slope (= 1.8) by Momose $et\ al$ [41]. It is obvious that effects of gate tunneling current on the drain current becomes less problematic for the short channel MOSFETs because the channel current is much higher than gate current and gate current decreases exponentially as the channel length is reduced. However, even though these gate currents for an individual transistor may not be significant, the sum of all gate currents for the entire chip will become a serious problem for applications with battery operation [9].

Figure 2.42(a) and (b) show the simulated drain currents for L_g = 20 μm and L_g = 50 μm when gate tunneling effects are considered. In the long channel regime of L_g = 50 μm, anomalous electric characteristics are observed at very low V_{ds}, because the magnitude of gate current is comparable to the drain current in this bias range.

SPICE transient circuit simulation using the compact model is performed for the simple circuit as shown in Figure 2.43, composed of a single NMOS transistor with a 0.1 pF capacitor. To observe the distinct effects of gate tunneling current during circuit operation, a large transistor (W/L = 100 μm/20 μm) is used. With the compact model, gate tunneling current effects are quite noticeable; the node 'A' remains at a steady voltage due to the current injection from the gate, while it gradually decreases as a function of time when the gate current is not considered.

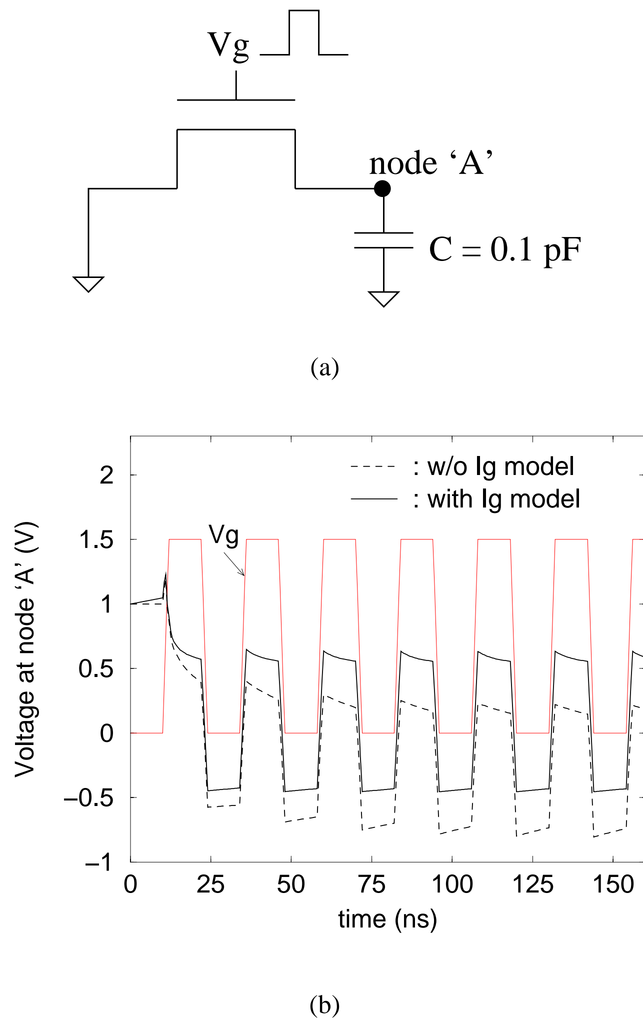


Figure 2.43: Transient circuit simulation results with the gate current model, (a) test circuit with NMOS (W/L = 100 μm/20 μm and t_{ox} = 1.5 nm) and 0.1 pF capacitor, (b) simulated voltages at node 'A' with and without gate tunneling current models.

2.5.4 Summary

An accurate direct tunneling model for circuit simulation is developed that incorporates an explicit expression for surface potential with quantum-mechanical effect. Simulated gate currents from this model demonstrate good agreement with the results from a numerical solver and measured data for gate oxides ranging 1.3 – 1.8 nm. Long channel devices are used in the simulations which emphasize the tunneling effects.

2.6 Impact of Gate Tunneling Current on Circuit Performance

2.6.1 Introduction

To date, many studies have considered extensively gate oxide scaling issues in terms of direct tunneling current [41],[59]-[67]. Some reports have demonstrated that the gate current will deteriorate device performance and increase power consumption [5][65]. Others have shown that the gate current effect on the drain current is less serious owing to the exponentially decreases in gate current as the channel length is reduced, which in turn supports more aggressive oxide scaling [41].

For conventional CMOS devices, the dominant leakage mechanism is mainly due to short channel effects owing to drain induced barrier lowering (i.e. DIBL). In the ultra-thin gate oxide regime, however, the gate leakage current can significantly contribute to off-state leakage, which may result in faulty circuit operation since designers may assume that there is no appreciable gate current. A recent study has shown that direct tunneling current appearing between the Source-Drain Extension (SDE) and the gate overlap, the so-called Edge Direct Tunneling (EDT) effect, dominates off-state drive current, especially in very short channel devices [68][69]. This results from the factor that the ratio of the gate overlap to the total channel length becomes large in the short channel device compared to that of the long channel device. Thus, the gate current effect is expected to become appreciable in ultra-thin oxide, sub-100 nm channel length MOS circuits. Even though many researchers have discussed the effects of gate leakage current, scaling limitations due to gate tunneling current from the viewpoint of circuit operation have not been critically addressed. Assessment of circuit immunity against gate tunneling currents, depending on various device structures and bias conditions, is of great importance in determining directions for future

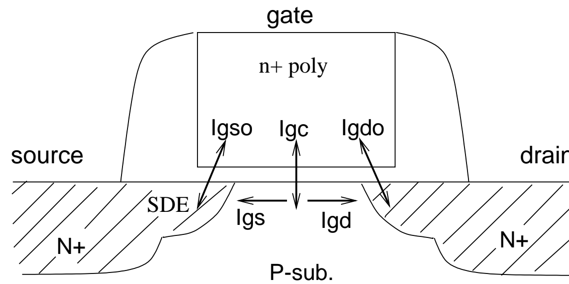


Figure 2.44: Illustration of gate direct tunneling components of a very short-channel NMOSFET. I_{gso} and I_{gdo} are edge direct tunneling (EDT) currents.

gate oxide scaling [66].

This section considers circuit stability and oxide scaling limitations for several typical logic and non-logic CMOS circuits using both device- and circuit-level simulation models.

2.6.2 Gate Current Modeling

Edge Direct Tunneling (EDT)

Gate direct tunneling current is produced by the penetration of quantum-mechanical wave-functions of the carriers through the gate oxide potential barrier into the gate, which depends not only on the device structure but also the bias conditions. Figure 2.44 illustrates various gate tunneling components in a scaled NMOSFET; the gate-to-channel current (I_{gc}) and the edge-direct-tunneling (EDT) currents – both I_{gso} and I_{gdo}, respectively.

In long channel devices, I_{gso} and I_{gdo} are less important than I_{gc} because the gate overlap length is small compared to the channel length. In very short channel devices, the portion of the gate overlap compared to the total gate length becomes a large fraction. For example, the typical gate overlap of a physical gate length for a 50 nm NMOSFET, estimated by two-dimensional process simulation, is around 20 nm which corresponds to 40 %

of the total. To reduce the direct tunneling current and Miller capacitance, a smaller overlap length is desirable. However, the overlap cannot be scaled easily in advanced MOS devices due to difficulties in controlling the doping profiles. Moreover, even if the devices can be scaled successfully, too short an overlap region may cause unacceptably high external resistance in the shallow source-drain junctions.

Figure 2.45 illustrates the band diagrams and electron tunneling directions along the gate-to-channel and gate-to-SDE directions in a highly doped drain (HDD) NMOSFET. For $V_g > 0$ V, the gate-to-channel tunneling current (I_{gc}) is the dominant current component, since higher gate oxide voltage (V_{ox}) appears between the gate and the channel, as shown in Figure 2.45(a). Namely, the V_{fb} of an NMOSFET with an n-type polysilicon gate (i.e. n^+-poly/SiO_2/p-substrate) is approximately -1 V, while the V_{fb} along the gate-to-SDE (i.e. n^+-poly/SiO_2/n^+ SDE) is nearly 0 V. In contrast, assuming that the overlap length is comparable to the channel length, the EDT currents (I_{gso} and I_{gdo}) can become dominant for bias conditions of $V_{fb} < V_g < 0$ V. For the gate-to-SDE case, accumulated electrons in the n^+-poly gate tunnel to the SDE region, which can lead to an appreciable off-state current. Meanwhile, operating in the depletion mode along the n^+-poly/SiO_2/p-substrate surface, few electrons are present in the channel that could in turn tunnel into the gate, as shown in Figure 2.45(b).

Direct Tunneling Device Simulation

The edge direct tunneling in the gate-to-SDE region must be treated as a two-dimensional problem in very short channel devices [69], owing to the laterally finite doping gradient in the SDE region and the drain electric field effects. The direct tunneling current compact model for circuit simulation has shown good agreement with experimental data for long channel devices [59]; however, it cannot accurately represent the direct tunneling current for highly non-uniform SDE and channel regions in sub-100 nm MOSFETs. The exact gate

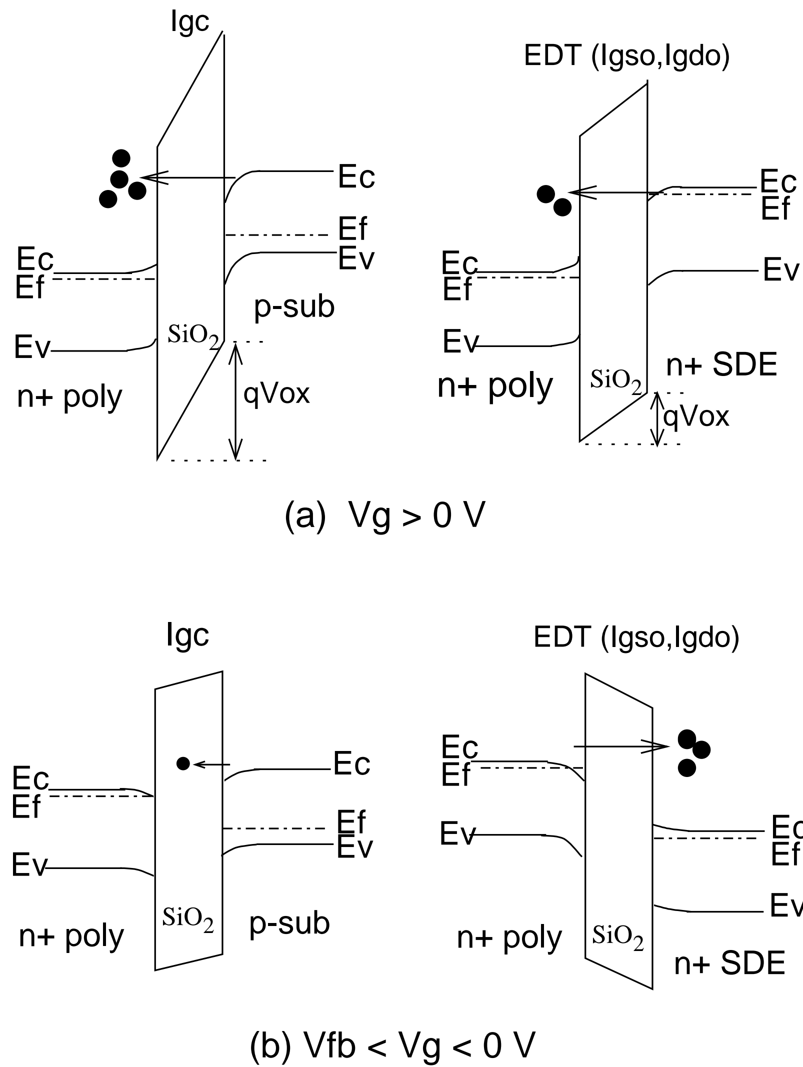


Figure 2.45: Gate bias dependent band diagrams and electron tunneling in the channel (I_{gc}) and the gate edge (I_{gso} and I_{gdo}). (a) $V_g > 0$ V (inversion mode). (b) $V_{fb} < V_g < 0$ V (depletion mode).

direct tunneling should be modeled based on solution of the two-dimensional Schrödinger

equation, coupled with the semiconductor transport equations. For practical reasons, models which maintain a macroscopic formulation while still incorporating the approximated quantum-mechanical wavefunctions are advantageous from the perspective of circuit and device designers [70].

In order to model the edge direct tunneling behavior, MEDICI [30] was used. The electron direct tunneling is calculated for two sources: conduction band electron tunneling (CBET) and valence band electron tunneling (VBET). CBET is the tunneling of electrons from the conduction band of the silicon substrate to the polysilicon, while VBET is the electron tunneling from the valence band of the silicon substrate due to generation of free holes. For CBET the net direct tunneling current is calculated for the conduction band electrons, using the independent electron approximation given by [30]:

$$J_{DT} = \frac{4\pi q m_1 k_B T}{h^3} \int_0^{E_b} TC(E)\, ln \left[\frac{e^{(E_{Fn1} - E_{c1} - E)/k_B T} + 1}{e^{(E_{Fn3} - E_{c3} - E)/k_B T} + 1} \right] dE \qquad (2.30)$$

where E_{Fn1}, E_{c1}, and m_1 are the electron quasi-fermi level, the conduction band edge, and the electron effective tunneling mass in the silicon region. E_{Fn3} and E_{c3} are the electron quasi-fermi level and the conduction band in the polysilicon region, respectively. The endpoint integration is determined by the barrier height in the silicon region, E_b. $TC(E)$ is the tunneling coefficient of an electron with kinetic energy of the incident electron (E) that describes the probability that an electron with a certain energy can tunnel through the oxide.

The direct tunneling current is calculated self-consistently along with the electron and hole continuity equations ; the Gundlach formula [71] is used to extract $TC(E)$ for a trapezoidal shaped barrier.

To validate the direct tunneling model, gate currents were simulated and compared to

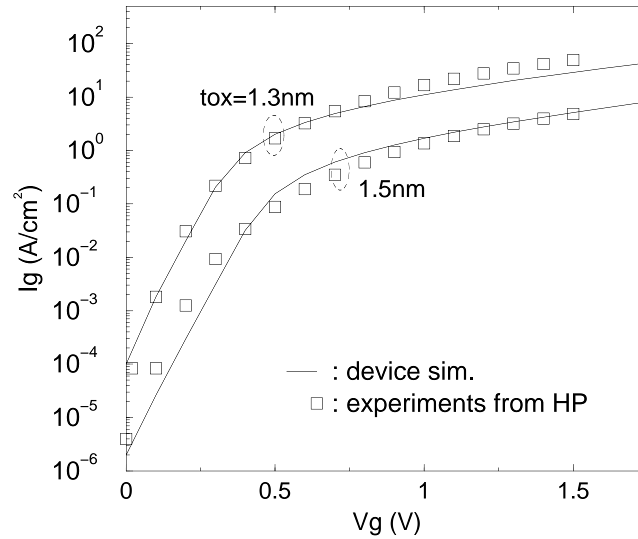


Figure 2.46: Simulated gate currents using MEDICI [30] and comparison with the measured HP data for a long channel (L_g = 100 μm) NMOSFET.

the experimental data of long channel NMOSFETs (i.e. W/L = 100 μm/100 μm and t_{ox} = 1.3 and 1.5 nm). Though agreement is not perfect, simulated gate currents from MEDICI show reasonable correspondence to the measurements, as reflected in Figure 2.46.

Device simulations were also performed for a very-short channel NMOSFET with 50 nm gate length; Figure 2.47 illustrates resulting gate currents for an NMOSFET with t_{ox} = 1.5 nm. The source and drain are tied to ground and the gate bias is swept from negative to positive values. Note that the EDT current (I_{gso}, I_{gdo}) is higher than the gate-to-channel current (I_{gc}) for gate biases of -1.5 V $< V_g <$ 0 V, implying that the EDT is the dominant leakage source for the off-state current in the low voltage range of operation for MOS circuits. Figure 2.48 shows the total simulated gate current ($I_{gg} = I_{gc} + I_{gso} + I_{gdo}$) for different gate oxide thicknesses, ranging from 1.1 – 1.8 nm; exponentially increasing gate current is produced as the gate oxide thicknesses are scaled down.

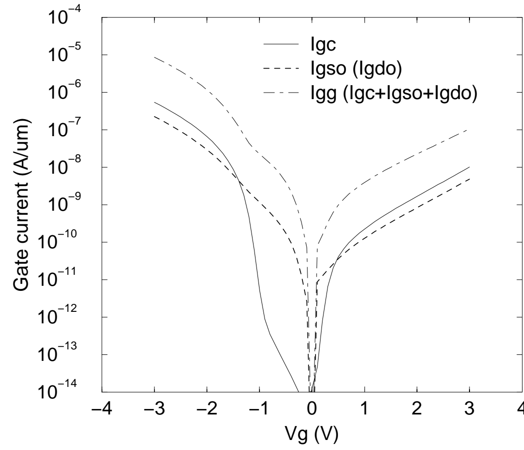


Figure 2.47: Simulated I_{gc}, I_{gso} and I_{gg} (= I_{gc}+I_{gso}+I_{gdo}) for an NMOSFET with t_{ox} = 1.5 nm and L_g = 50 nm.

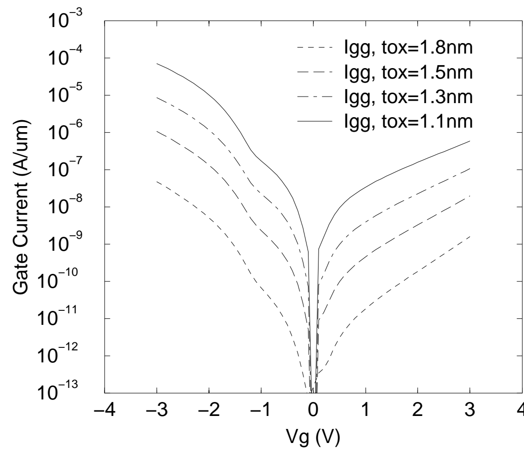
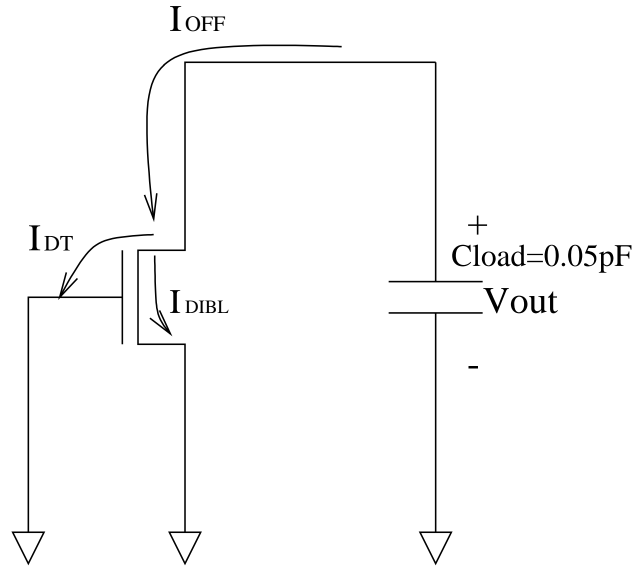
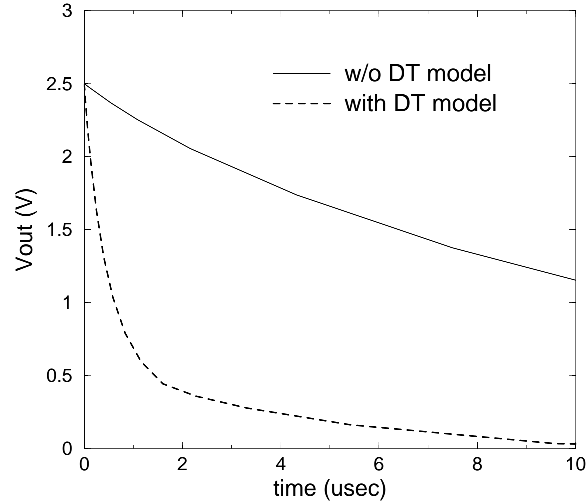


Figure 2.48: Simulated I_{gg} (= I_{gc}+I_{gso}+I_{gdo}) for different t_{ox}'s ranging 1.1 – 1.8 nm and L_g = 50 nm.

In order to observe the transient behavior of thin oxide MOSFETs with significant gate tunneling current, mixed mode circuit-device simulation is performed for a single off-state transistor with a loading capacitor initially charged to 2.5 V. The discharge of the output node connected to the NMOS drain is determined by the off-state current (I_{OFF}), which is the sum of the direct tunneling leakage (I_{DT}) and the DIBL leakage (I_{DIBL}) currents, as illustrated in Figure 2.49(a). The gate length and the oxide thickness are 70 nm and 1.3 nm, respectively. Even though the NMOS is in the off-state, it acts more like a resistor due to the off-state current of the transistor, therefore V_{out} decays with time, as shown in Figure 2.49(b). With the direct tunneling model, V_{out} drops more sharply than the case without considering the direct tunneling model. Defining an equivalent time constant (τ) as, $\tau = V_{dd}C_{load}/I_{OFF} = V_{dd}C_{load}/(I_{DT} + I_{DIBL})$ provides a useful estimate. Overestimating τ will be predictable without considering the I_{DT} term.



(a)



(b)

Figure 2.49: Transient simulation by using MEDICI for a single off-state NMOS transistor with t_{ox} = 1.3 nm, L_g = 70 nm and W = 10μm. (a) discharge through a single off-transistor, V_{out} (t = 0) = 2.5 V. (b) comparison of V_{out} values between with and without the direct tunneling model.

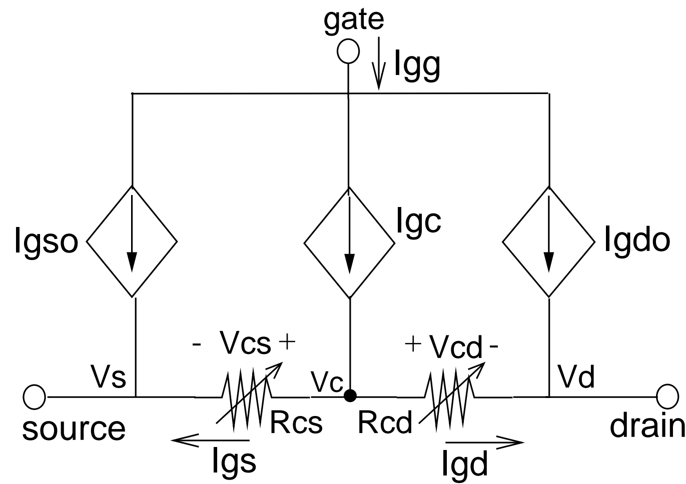


Figure 2.50: A macro-circuit model for direct tunneling current combined circuit simulation.

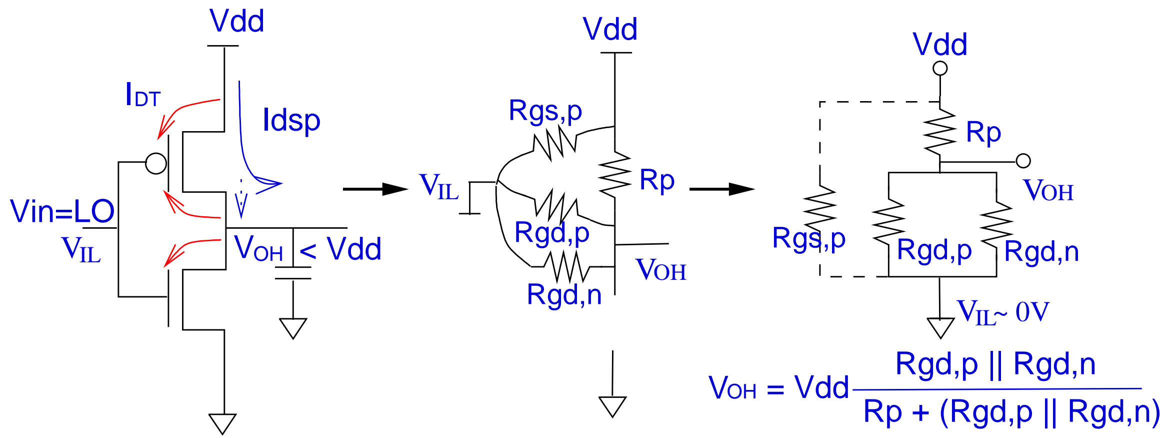
2.6.3 Circuit Application

In order to evaluate circuit performance by considering gate direct tunneling effects, a two-lump (macro) circuit model has been constructed in the circuit simulator, HSPICE [48]. Gate direct tunneling currents, obtained from the device simulation for the gate oxide thicknesses of 1.1, 1.3 and 1.5 nm, are described using voltage-dependent current sources as a function of the terminal voltage, as shown in Figure 2.50. The partitioning of I_{gc} into I_{gs} and I_{gd} is modeled by using variable resistances, R_{cs} and R_{cd}, respectively, in each part of the channel. R_{cs} and R_{cd} are the channel resistance corresponding to $0.5L_{ch}$, where $R_{cs} = V_{cs}/I_{cs}$ and $R_{cd} = V_{cd}/I_{cd}$. I_{cs} and I_{cd} are channel currents of each region; they have been obtained by adjusting the BSIM3-model parameters to fit the I-V curves generated from device simulation.

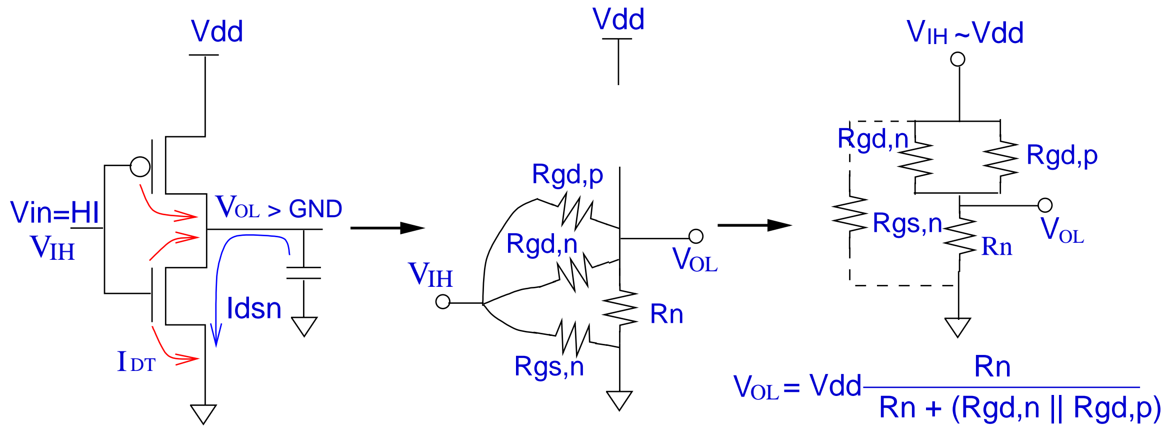
The macro-circuit model has been applied to several MOS circuits – CMOS inverter, dynamic AND gate, sample and hold (S/H) and bootstrapping circuits.

Static CMOS Inverter

The CMOS inverter is the most basic logic gate that performs a Boolean operation on a single input variable. For the CMOS inverter application we assumed the amount of hole direct tunneling for the PMOSFET is the same as that for the NMOSFET. The magnitude of the channel current is assumed to be identical, regardless of the gate oxide thickness, in order to focus on the circuit performance differences based on the oxide thickness dependent gate tunneling current contributions. Estimated gate current paths during the operation are shown schematically in Figure 2.51.

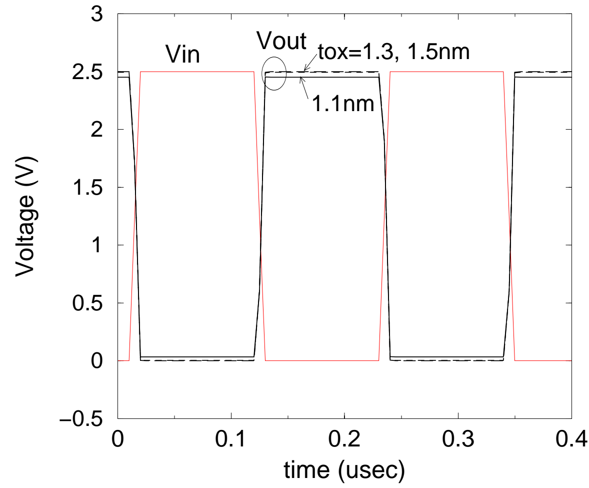


(a)

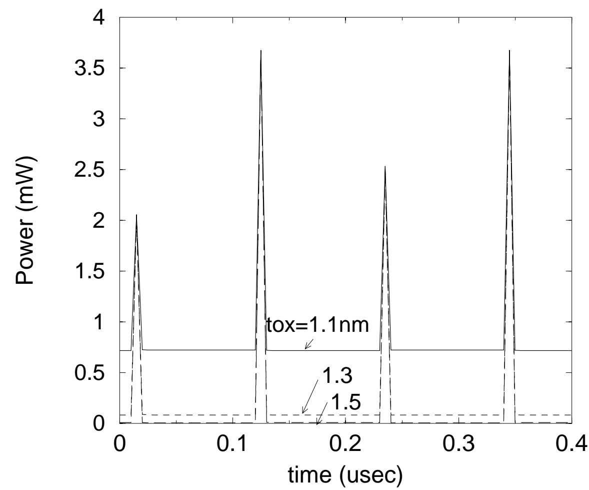


(b)

Figure 2.51: V_{OH} and V_{OL} modeling considering gate tunneling currents of a CMOS inverter. (a) V_{in} = logic-low = 0 V. (b) V_{in} = logic-high = V_{dd}.



(a)



(b)

Figure 2.52: CMOS inverter simulation results including gate direct tunneling for three gate oxide thicknesses, V_{dd} = 2.5 V (C_{out} = 0.08 pF (FO4) and $L_n = L_p$ = 50 nm). (a) input and output waveforms. (b) power consumptions.

When the input is *low* and the gate tunneling current is significant, V_{out} (i.e. V_{OH}) will not reach V_{dd} due to the leakage current that flows from the output node, as shown in Figure 2.51(a). Here, direct tunneling current components can be modeled as resistors and V_{OH} is approximated by the voltage divider:

$$V_{OH} \approx V_{dd} \times \frac{R_{gd,p}||R_{gd,n}}{R_p + (R_{gd,p}||R_{gd,n})} \tag{2.31}$$

where, R_p is the on-state channel resistance of the PMOS. $R_{gd,n}$ and $R_{gd,p}$ are gate-to-drain resistances of the N- and PMOSFET, respectively, each resistor models the respective gate direct tunneling effects. The resistor values are approximately $R_{gd,p} \approx V_{dd}/I_{gd,p}$ and $R_{gd,n} \approx V_{dd}/I_{gd,n}$. As an example, if the ratio of $R_{gd,n}$ (or $R_{gd,p}$) to R_p is 100, then V_{OH} will drop by 2 % from the V_{dd} level.

As shown in Figure 2.51(b), V_{OL} is again approximated using a voltage divider:

$$V_{OL} \approx V_{dd} \times \frac{R_n}{R_n + (R_{gd,n}||R_{gd,p})} \tag{2.32}$$

where, R_n is the on-state channel resistance of the NMOS.

As a result, when tunneling current is significant, V_{OL} will not fall to the GND level and the V_{out} swing (GND $< V_{out} < V_{dd}$) is reduced for very leaky, thin gate oxide CMOS inverters.

Again considering an example, assuming that V_{dd} = 2.5 V, $I_{ds,n}$ = 0.50 mA/μm, $I_{ds,p}$ = 0.25 mA/μm and $I_{gd,p}$ = $I_{gd,n}$ = 5.0$\times$ 10^{-6} A/μm (i.e. W_p = 2W_n = 20 μm), based on the simulations for t_{ox} = 1.1 nm and L_g = 50 nm, then $R_{gd,n}/R_n$ or $R_{gd,p}/R_p$ is about 100. In this case, the estimated V_{OH} and V_{OL} values are V_{OH} = 2.42 V and V_{OL} = 0.08 V, respectively,

from Eqs. (2.31) and (2.32), or a total reduced logic swing of 160 mV.

Figure 2.52 shows simulated waveforms of input/output and power consumption for a CMOS inverter using the macro-circuit model, parameterized with t_{ox} = 1.1, 1.3, and 1.5 nm. The output capacitance is an assumed value equivalent to a gate fanout of 4 (FO4). For V_{dd} = 2.5 V and t_{ox} = 1.1 nm, simulated V_{OH} and V_{OL} are 2.45 V and 0.04 V, respectively. The full logic-high (V_{dd}) and logic-low (GND) levels are achievable for t_{ox} = 1.3 and 1.5 nm. The DC power consumption of the CMOS inverter at V_{dd} = 2.5 V can be substantial for t_{ox} = 1.1 nm, as shown in Figure 2.52(b); the average power consumption is 0.75, 0.16, and 0.09 mW for t_{ox} = 1.1, 1.3, and 1.5 nm, respectively. The power consumption of 0.75 mW for t_{ox} = 1.1 nm, is about 10 times larger than for the case when the gate tunneling current is negligible (i.e. the 0.08 mW baseline).

For V_{dd} = 1.5 V, the output node of the inverter swings between full logic-high and logic-low (i.e. only 0.1 % of V_{OH} reduction even for t_{ox} = 1.1 nm). Since the channel current becomes much higher than the gate tunneling current, gate tunneling effects can be minimized when using lower voltage static-logic circuits. The power consumption for V_{dd} = 1.5 V is exponentially reduced compared to the case for V_{dd} = 2.5 V, due to the exponential decrease in gate current; basically a reduction of 20 $\sim$ 100 times is realized, compared to the V_{dd} = 2.5 V case. According to the ITRS roadmap, V_{dd} of 1.0 – 1.5 V is required for 70 nm CMOS technology. For the low V_{dd}, gate direct tunneling current effects on static-logic circuits will be less serious for oxide thicknesses down to 1.1 nm.

Dynamic AND Gate

Consider the domino CMOS two-input, AND-2 gate, shown in Figure 2.53. The circuit operation relies on first *precharging* the output node capacitance and subsequently, *evaluating* the output level according to the applied inputs, V_A and V_B. These operations are scheduled by a single clock signal, CK, which drives one NMOS and one PMOS transistor in each

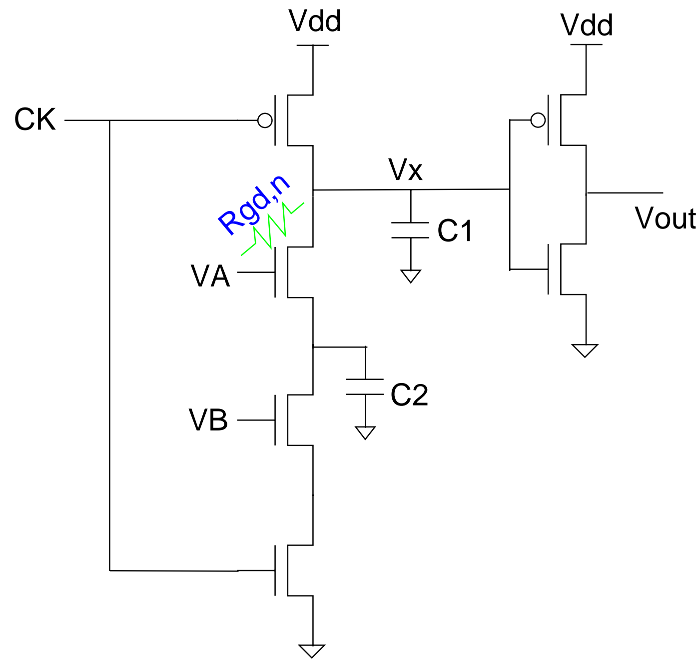
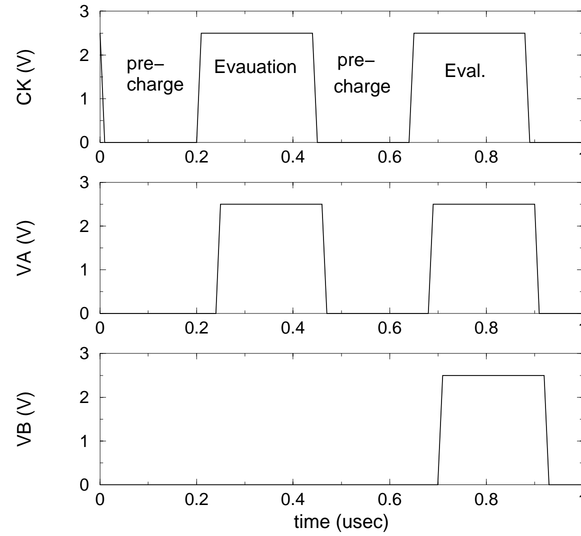
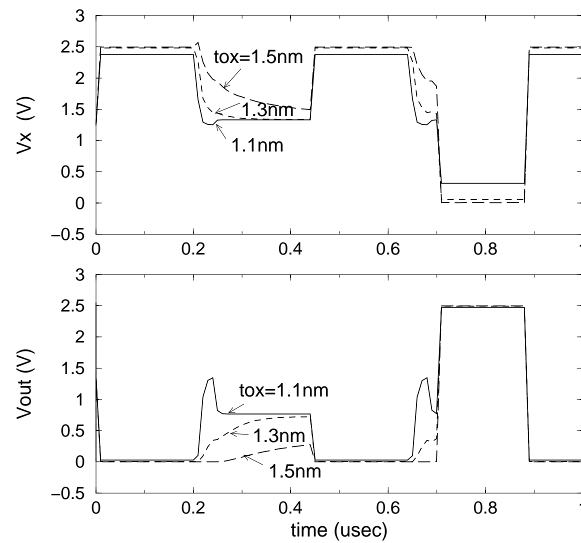


Figure 2.53: Domino CMOS AND-2 gate.

dynamic stage. Assume all inputs are low initially and the intermediate node voltage across C_2 has an initial value of 0 V. During the precharge phase, the output node capacitance (C_1) is charged up to its logic-high level of V_{dd} through the PMOS transistor. In the next phase, CK switches to logic-high and the evaluation begins. If the input V_A switches from low to high during the evaluation phase, charge stored on C_1 will be shared with C_2, and the node voltage V_X drops after the charge-sharing. If V_X is less than $V_{dd}/2$, the output of the inverter V_{out} erroneously switches to logic-high during the evaluation phase. When gate tunneling current becomes significant, V_X may be less than the logic-high level during the precharge phase, so that V_X can even drop to less than $V_{dd}/2$ during the first evaluation period. As a result, V_{out} will inadvertently switch to a logic-high, resulting in a logic error.

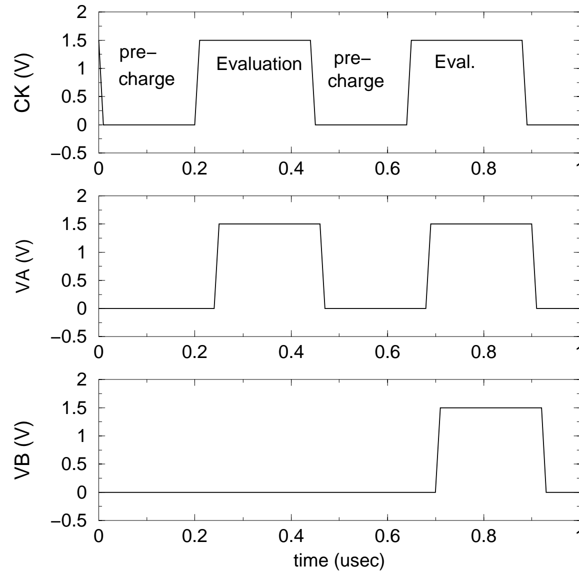


(a)

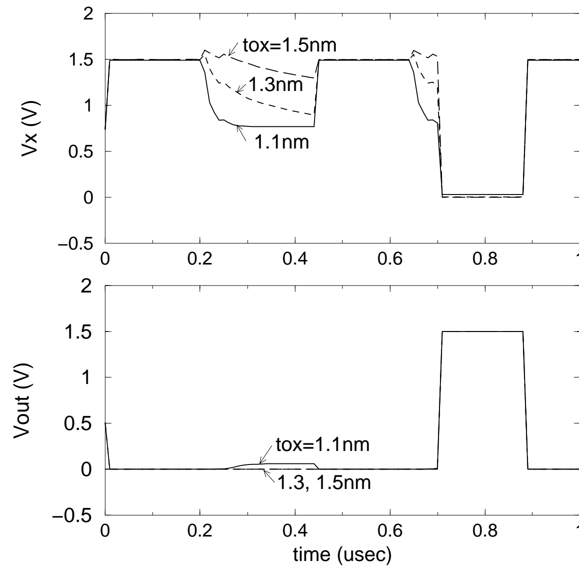


(b)

Figure 2.54: Simulated waveforms of the domino AND-2 gate. (a) clock and input signals. (b) output waveforms for V_{dd} = 2.5 V.



(a)



(b)

Figure 2.55: Simulated waveforms of the domino NAND-2 gate for V_{dd} = 1.5 V. (a) clock and inputs (b) output waveforms.

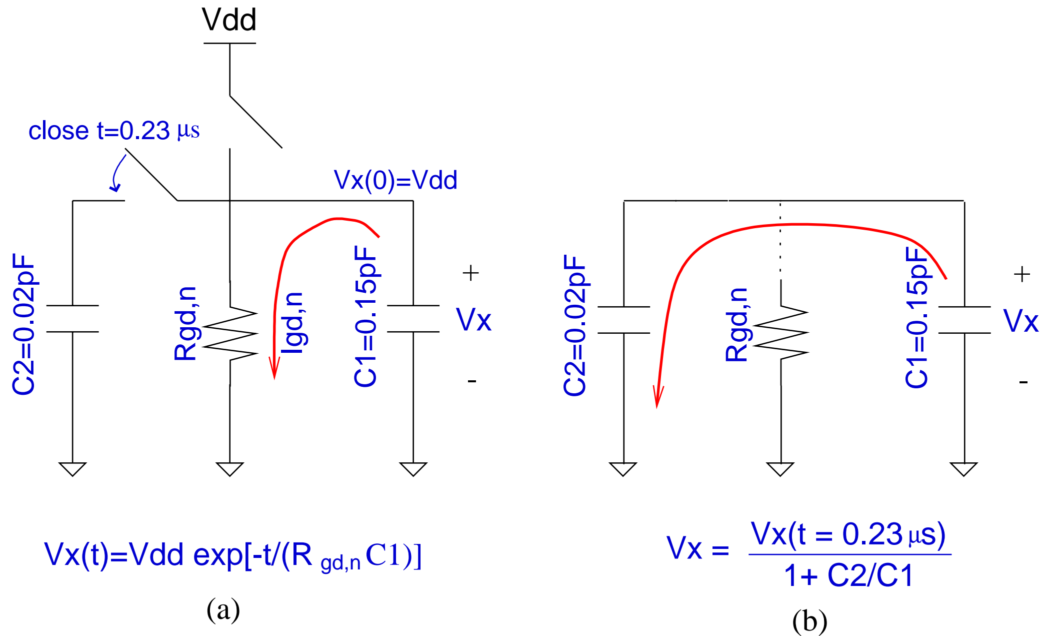


Figure 2.56: Modeling of discharge and charge sharing behaviors during first evaluation period of domino AND-2 gate. (a) discharge through tunneling resistance before V_A switches to logic-high (t = 0.20–0.23 μs). (b) charge sharing of C_1 with C_3 after the V_A switches to logic-high ($t > 0.23$ μs).

Figures 2.54 and 2.55 show simulated input and output waveforms of a domino AND-2 gate for V_{dd} = 2.5 and 1.5 V, respectively. When t_{ox} = 1.1 nm and V_{dd} = 2.5 V, V_X drops to about 1.2 V during the evaluation phase due to gate tunneling current effects. As a result, V_{out} erroneously switches during the first evaluation period and a glitch appears prior to the second evaluation phase. Even though these erroneous phenomena can be reduced for V_{dd} = 1.5 V, the dynamic logic circuit has potential problems due to the gate tunneling-induced off-state current during the precharge-and-evaluation stages.

These phenomena can be simply modeled as RC circuits, as shown in Figure 2.56. Initially, V_X has a value of V_{dd} during the precharge phase. When the evaluation begins at

t = 0.2 μs by switching of the CK signal to logic-high (i.e. V_A remains logic-low until t = 0.23 μs), charge stored in C_1 flows to ground through the tunneling resistance ($R_{gd,n}$), as illustrated in Figure 2.56(a). During the discharge process the V_X level drops as a function of time according to the following:

$$V_X(t) = V_X(t_0)\, e^{-\frac{t-t_0}{R_{gd,n}C_1}} \approx V_{dd}\, e^{-\frac{I_{gd,n}\ (t-t_0)}{V_{dd}C_1}} \tag{2.33}$$

where, t_0 = 0.20 μs in the case.

With a tunneling current of $I_{gd,n} = 4\times10^{-6}$ A, V_X drops to 1.60 V from its initial voltage of V_{dd} = 2.5 V during 0.03 μs (t = 0.20 – 0.23 μs), according to Eq. (2.33).

When V_A switches to logic-high at t = 0.23 μs, direct tunneling current is reduced due to a smaller voltage difference between V_X and V_A. Thus, instead of the discharge process, charge sharing begins; charge in C_1 is shared with C_2, as shown in Figure 2.56(b). The final V_X after the charge sharing is approximated as follows:

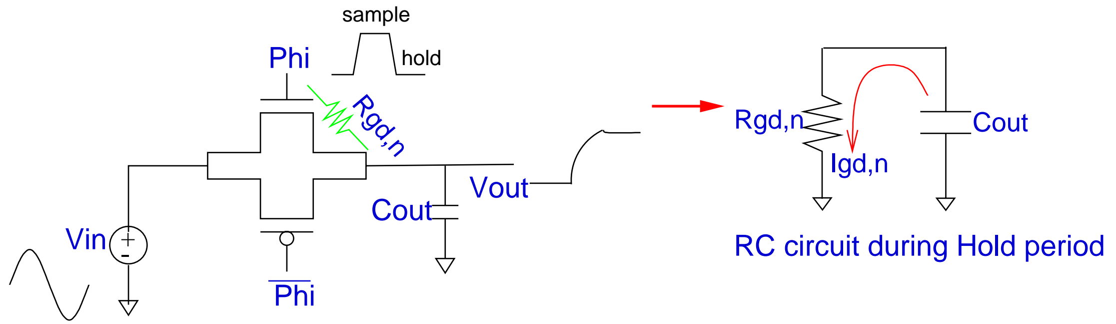
$$V_X \approx \frac{V_X(t = 0.23\mu s)}{1 + C_2/C_1} \tag{2.34}$$

When $V_X\ (t = 0.23\mu s)$ = 1.60 V, C_1 = 0.15 pF and C_2 = 0.04 pF, V_X after charge sharing estimated by Eq. (2.33) is about 1.26 V, which corresponds to $\sim V_{dd}/2$ level. Hence, V_{out} may erroneously switch during the evaluation phase, as shown in the simulation results of Figure 2.54(b). Even though these spurious results can be reduced by lowering V_{dd}, the dynamic logic circuit may have potential problems due to gate tunneling-induced off-state current during the precharge-and-evaluation phases of operation.

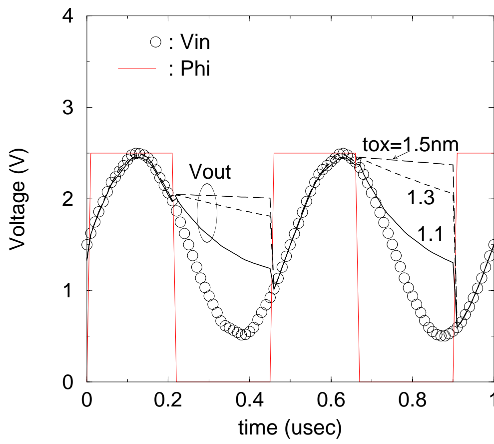
Sample and Hold Circuit

The sample and hold (S/H) circuit is an important analog building block in data-converter systems used to acquire analog signals and to store the value for some length of time. A simple S/H circuit is formed by a sampling CMOS switch followed by a hold capacitor, as shown in Figure 2.57(a). When the clock (*Phi*) is high, V_{out} follows V_{in}; when *Phi* goes low, V_{out} will ideally remain at a constant level. However, V_{out} will not hold this sampled value if leakage paths exist. This tunneling current-induced decay in V_{out} during the hold period can be modeled using the RC circuit shown in Figure 2.57(a). As for the previous dynamic AND gate, V_{out} decays as a function of time, again using the expression given in Eq. (2.33).

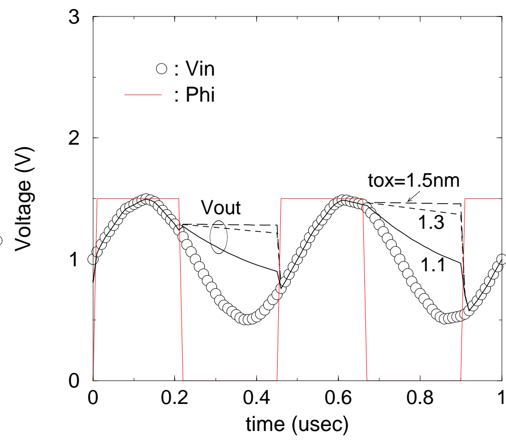
Figure 2.57(b) and (c) show simulation results of a S/H switch for three gate oxide thicknesses. During the holding period, the output node does not maintain the sampled value due to gate leakage current, and degradation becomes increasingly severe as the oxide thickness is scaled down. These gate tunneling effects appear even at low V_{dd}, as shown in Figure 2.57(c), implying that the S/H circuit has poor robustness in the face of gate leakage current, limiting its operation to oxide scaling down to the 1.5 nm regime.



(a)



(b)



(c)

Figure 2.57: CMOS sample and hold (S/H) circuit and simulated waveforms for different t_{ox}'s. (a) CMOS S/H circuit schematic and RC circuit model during hold period. (b) waveforms for V_{dd} = 2.5 V. (c) waveforms for V_{dd} = 1.5 V.

Alternative Gate Dielectrics

Alternate insulating materials with a dielectric constant larger than that of SiO_2 are under evaluation to replace the conventional SiO_2 gate stack, thereby overcoming the tunneling problems cited above; remote-plasma nitrided oxide (RPNO) [72][73], JVD-Si_3N_4 [74] and nitride/oxide (N/O) [45] are all possible replacement materials. Using such gate dielectrics, devices with lower gate-leakage current can be achieved as a result of the increased film thickness achieved with the dielectric constant of nitride-based layers ($\epsilon_{nitride} \sim$ 7.8).

Figure 2.58 illustrates the simulated gate tunneling current for an L_g = 50 nm NMOSFET with alternative gate dielectric of Si_3N_4 shown as a solid line, assuming a dielectric constant of $\kappa = 2\epsilon_{ox}$ and thickness of 2.6 nm ($t_{ox,eq}$ = 1.3 nm). Much lower gate current is produced by using the Si_3N_4 device, compared to the pure oxide device with the same equivalent oxide thickness of t_{ox} = 1.3 nm. Secondary effects such as surface roughness and interface traps are not considered.

Voltage bootstrapping is used to overcome threshold voltage drops in digital circuits. Figure 2.59(a) shows a schematic of the bootstrapping circuit, including the bootstrap MOS capacitor; the voltage V_X is increased during the V_{in} switching event. As a result, the threshold voltage drop can be compensated at the output node, V_{out}.

When V_{in} switches to logic-low, V_{out} and V_X are approximated as follows [75]:

$$V_{out} \approx V_X - V_{th,M2} \tag{2.35}$$

where

$$V_X \approx (V_{dd} - V_{th,M3}) + V_{dd}\frac{C_{boot}}{C_s + C_{boot}} \tag{2.36}$$

(V_{dd}-$V_{th,M3}$) is the initial condition of V_X and the second term of Eq. (2.36) represents the increase in V_X after V_{in} switches to 0 V. However, this V_X expression should be modified to account for the gate tunneling current. First, the initial V_X is reduced due to the discharge via the gate leakage resistor (i.e. $R_{g,cboot}$), such that

$$(V_{dd} - V_{th,M3}) \longrightarrow (V_{dd} - V_{th,M3})\, e^{-\frac{t}{R_{g,cboot}C_{boot}}} \tag{2.37}$$

In addition, the second term of Eq. (2.36) is modified due to the gate-to-drain resistance effects of M1 such that,

$$V_{dd}\frac{C_{boot}}{C_s + C_{boot}} \longrightarrow \left(V_{dd}\frac{R_{gd,M1}}{R_{M2} + R_{gd,M1}}\right) \cdot \frac{C_{boot}}{C_s + C_{boot}} \tag{2.38}$$

Thus, both V_X and V_{out} will be reduced in the presence of substantial gate tunneling current.

Figure 2.59(b) illustrates the simulated input and output waveforms of the bootstrapping circuit. For t_{ox} = 1.1 and 1.3 nm, the final V_{out} does not reach V_{dd} because of the leakage in both the MOS capacitor (C_{boot}) and the driving transistor (M1). In contrast, a full V_{dd} output voltage can be achieved by adopting a Si_3N_4 gate dielectric due to the substantial reduction in leakage current. Note that the equivalent oxide thickness of $t_{Si_3N_4}$ = 2.6 nm is t_{ox} = 1.3 nm. Alternative gate dielectrics will be necessary to replace leaky gate oxides in MOS circuits, especially where charge conservation or charge bootstrapping techniques are required.

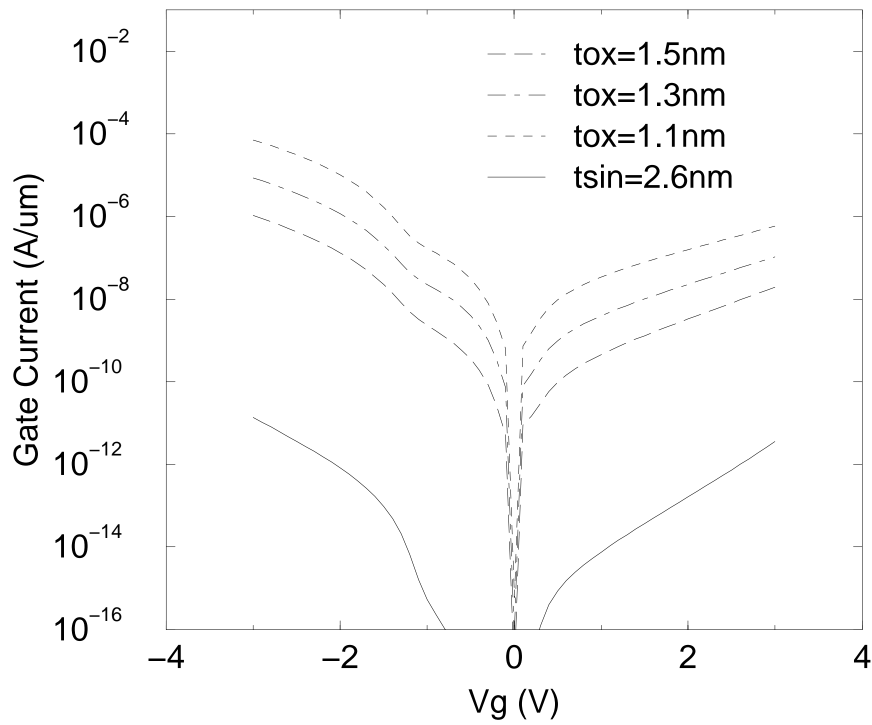
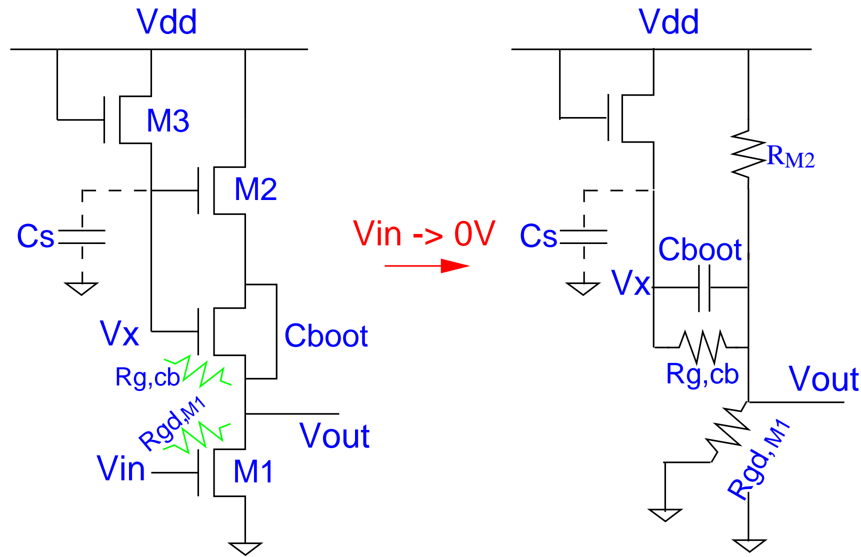
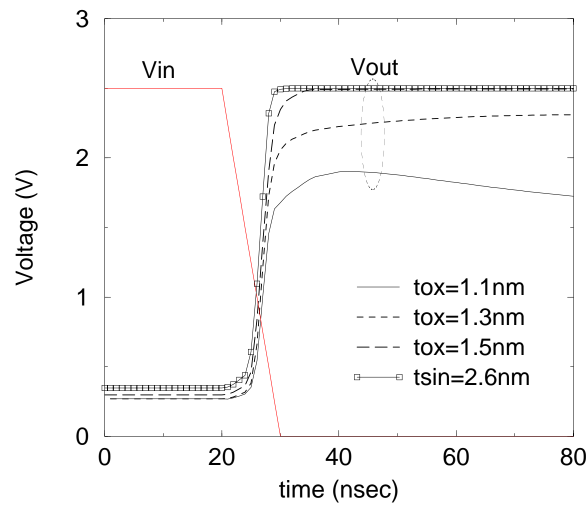


Figure 2.58: Simulated gate tunneling current of L_g = 50 nm NMOSFET with an alternative gate dielectric of Si_3N_4, κ = $2\epsilon_{ox}$ and thickness is 2.6 nm.



(a)



(b)

Figure 2.59: Voltage bootstrapping circuit and simulation results. (a) circuit schematic and its equivalent circuit when V_{in} switches to 0 V. (b) simulated waveforms with an alternative gate dielectric (Si_3N_4) and pure oxides.

2.6.4 Review of ITRS Oxide Scaling Roadmap

It is instructive to estimate how the gate direct tunneling current (i.e. edge direct tunneling current) will affect oxide scaling in the technology roadmap. According to the latest ITRS logic technology roadmap shown in Table 2.2, the range of equivalent t_{ox} from 120 to 70 nm node (1999 – 2004) is chosen to keep $V_{dd}/t_{ox} \leq 8$ MV/cm, based on oxide reliability concerns. From the 65 nm node (2005) the use of high κ materials other than SiO_2 is assumed (i.e. $\kappa = 2\epsilon_{ox}$), and $V_{dd}/t_{ox} \leq 8$ MV/cm requirement is no longer followed [3]. In the roadmap, the requirement of maximum gate leakage current (I_{gate}) limit (Table 2.2 – rows 4 and 5) is roughly taken to be less than or equal to only 1% of the transistor off-state leakage when $V_s = V_g = 0$ V and $V_d = V_{dd}$. This implies that the edge tunneling current effect is not seriously considered relative to the sum of the subthreshold current and the junction leakage current in the specification of the equivalent t_{ox} requirement. In reality, the portion of the edge direct tunneling current relative to the total off-state leakage will be a large fraction in the off-state bias condition.

Comparison has been made between the I_{gate} limit of the roadmap and the device simulation results. I_{gate} simulation for the low-power device was performed with a combination of the lower value of the V_{dd} range and a thinner oxide across the equivalent t_{ox} range to produce a maximum gate current value. The higher value of the V_{dd} range and a thinner oxide for the equivalent t_{ox} range were used for the high-performance case. As a result, the maximum I_{gate} limit of the technology roadmap in the 1999 – 2004 nodes is far below the simulated I_{gate}, as shown in Table 2.2. In particular, the simulated I_{gate} for the low-power device is 3 – 5 orders of magnitude higher than the required limit. Namely, the maximum I_{gate} limits in Table 2.2 – rows 4 and 5 are too strict for the oxide thickness range in the technology roadmap. The simulated I_{gate} beyond the 65 nm node (2005) is rather close to the roadmap requirement owing to the use of a nitride-based dielectric with $\kappa = 2\epsilon_{ox}$, but

dielectrics with $\kappa > 2\epsilon_{ox}$ will be necessary beyond 45 nm node technology, especially for low-power devices.

In order to satisfy the roadmap requirements for low-power operation, the use of oxides thicker than the equivalent t_{ox} range of the roadmap is desirable. For example, an oxide thickness of about 2.0 nm is required to ensure the I_{gate} limit for the 70 nm technology node (2004), as shown in Table 2.2 – row 8. For the high-performance case, the equivalent t_{ox} range ensures the I_{gate} limit except for the 70 nm node, as indicated in Table 2.2 – row 9.

In conclusion, the oxide scaling of the roadmap may be a little too aggressive, especially for the low-power devices, due to neglecting the edge direct tunneling current effects. In order to ensure the I_{gate} limit of the technology roadmap, an early use of high κ dielectrics or a conservative oxide scaling (Tabel 2.2– 8 and 9) is necessary.

2.6.5 Summary

CMOS circuit robustness in the presence of gate tunneling currents has been studied using circuit simulation, combined with a two-lump (macro) circuit model of gate tunneling current and analytic estimation of the effects. CMOS static inverters at V_{dd} = 1.5 V show acceptable noise margins with low power consumption for the oxide thicknesses down to 1.1 nm, while dynamic AND gates have a potential weakness in the presence of gate current during the precharge and evaluation phases. For circuits that require charge-conservation or charge-bootstrapping, including the S/H circuit, significant performance degradation can be expected for $t_{ox} < 1.5$ nm, even considering low voltage operation. A dual-gate oxide process or use of high-κ dielectric materials will be necessary for these circuits to continue device scaling.

Based on the simulation study, expected values of oxide thickness, needed to ensure the

Table 2.2: ITRS logic technology roadmap (2000 edition) [3] and oxide thicknesses to ensure edge direct tunneling leakage (dielectrics with $\kappa = 2\epsilon_{ox}$ are assumed for 2005–2011 node).

	Year	1999	2001	2004	2005	2008	2011
1	MPU gate length (nm)	120	100	70	65	45	30
2	Logic V_{dd} (V)	1.5–1.8	1.2–1.5	0.9–1.2	0.9–1.2	0.6–0.9	0.5–0.6
3	t_{ox} equivalent (nm)	1.9–2.5	1.5–1.9	1.2–1.5	1.0–1.5	0.8– 1.2	0.6–0.8
4	I_{gate} limit (A/μm) for low-power device	0.07p	0.10p	0.16p	0.2p	0.4p	0.8p
5	I_{gate} limit (A/μm) for high-performance	70p	100p	160p	200p	400p	800p
6	Simulated I_{gate} (A/μm) for low-power device	150p	2.5n	19n	2p	65p	4.1n
7	Simulated I_{gate} (A/μm) for high-performance	450p	9.5n	50n	6p	240p	6n
8	t_{ox} (nm) to ensure low-power I_{gate} limit	$\sim$2.6	$\sim$2.4	$\sim$2.0	$\sim$1.1	$\sim$1.0	$\sim$0.9
9	t_{ox} (nm) to ensure high-perform. I_{gate} limit	$\sim$2.2	$\sim$1.9	$\sim$1.7	$\sim$1.0	$\sim$0.8	$\sim$0.7

off-state gate leakage requirement of the ITRS roadmap, are outlined.

Chapter 3

Source and Drain Modeling

3.1 Introduction

Shallow junctions in the Source / Drain Extension (SDE) regions are needed to minimize the short channel effects in scaling of sub-100 nm MOSFETs. The ITRS roadmap predicts that the SDE depths (X_j) should be as low as 10 nm to achieve an L_{eff} of 50 nm, however, too shallow an X_j can lead to high external resistance, resulting in current degradation [8]. This implies that further improvements in I_{dsat} and MOSFET performance cannot be expected when the channel resistance becomes comparable to the source and drain resistance. In other words, further down-scaling without improvement of MOSFET performance is meaningless, even if short-channel effects are completely suppressed by introducing the shallow SDE. Accordingly, to minimize external resistance components it is necessary to predict them accurately for shallow SDE MOSFETs using an appropriate mobility model for the inversion and accumulation layers as they relate to different process conditions as well as device structures [76].

To determine an exact external resistance experimentally, the L_m-array (L_m:gate length)

method is most commonly used. However, the method is based on linear regression which potentially can extract incorrectly; negative resistance values can result from the nonlinear $R_{tot} - L_m$ characteristics due to process variations [77]. Moreover, the device designer cannot directly extract the different resistance components contributing to R_s and R_d by using such methods, since experimentally extracted resistance values are inherently lumped. For device optimization, understanding of the details among different resistance components with respect to the doping profiles, dimensions, and material parameters is of great importance.

In this chapter, the effects of shallow High-Doped Drain (HDD) SDE to the external resistance are discussed based on device simulations that exploit a unified mobility model for both inversion and accumulation layer regions [78].

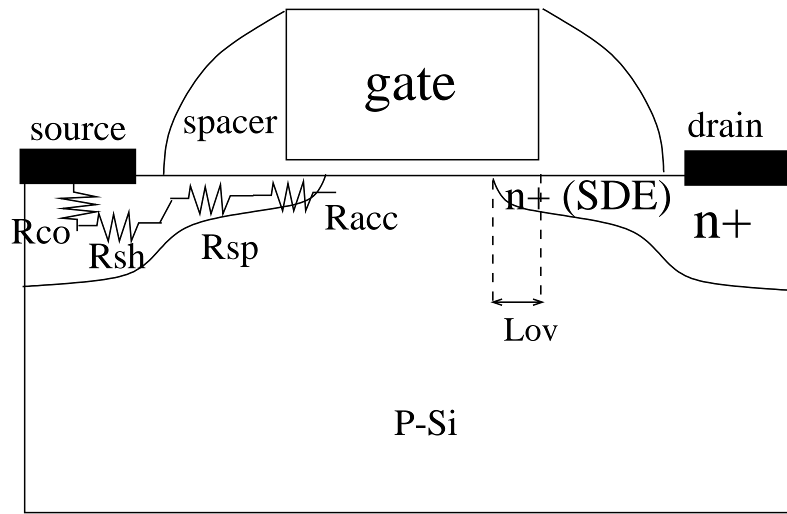
3.2 Shallow SDE Effects on External Resistance

3.2.1 Source/Drain Resistance

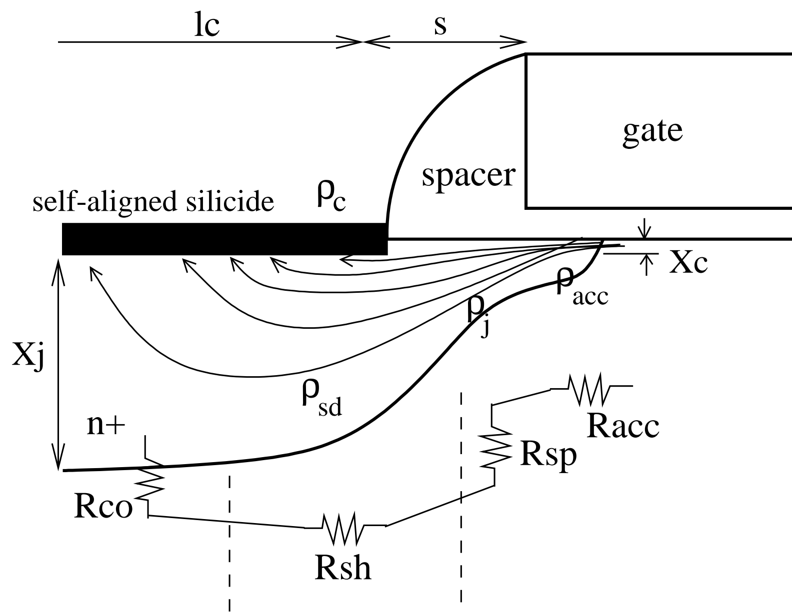
In reality, there is a voltage drop in the source and drain regions as the current flows from the channel to the terminal contact, due to the finite silicon resistivity and metal contact resistance. In a long-channel device, the source/drain parasitic resistance is negligible compared with the channel resistance. In a short-channel device, the source/drain resistance is an appreciable fraction of the channel resistance and therefore causes significant current degradation. The most severe current degradation due to series resistance occurs in the linear region (low V_{ds}) when the gate voltage is high, because the MOSFET channel resistance is the lowest under such bias conditions. Channel resistance is simply determined based on the inverse of the well-known current equation:

$$R_{ch} = \frac{V_{ds}}{I_{ds}} = \frac{L}{\mu_{eff} C_{ox} W (V_g - V_{on} - V_{ds}/2)} \tag{3.1}$$

The MOSFET current in the saturation region is the least affected by the resistance of the source/drain, since drain current is essentially independent of V_{ds} in saturation. Figure 3.1 shows several resistance components of the external source/drain resistance in a conventional HDD NMOS transistor [79]: (i) R_{co} is the contact resistance in the region where the current flows into the metal contacts, (ii) R_{sh} is the sheet resistance of the deep source/drain junction where the current flows uniformly, (iii) R_{sp} is the spreading resistance of the SDE, associated with current spreading from the surface into a uniform pattern across the depth of the source/drain, (iv) R_{acc} is the accumulation-layer resistance under the gate overlap where the current mainly stays at the surface.



(a)



(b)

Figure 3.1: External resistance components in source/drain extension (SDE) in an NMOSFET. (a) resistance components (R_{acc} : accumulation, R_{sp} : spreading, R_{co} : contact, R_{sh} : sheet resistances.) (b) current flow and resistivity components.

Accumulation-layer Resistance

The accumulation-layer resistance, R_{acc}, depends on the gate voltage. The sheet resistivity of the accumulation layer, ρ_{acc}, can be simply estimated in the overlap region between gate and source/drain, as [27]:

$$\rho_{acc} = \frac{1}{\mu_{acc} Q_{acc}} = \frac{1}{\mu_{acc} C_{ox} V_g} \tag{3.2}$$

where μ_{acc} is the average electron mobility in the accumulation layer.

The gate voltage in the gate overlap region can be expressed as:

$$V_g = V_{fb} + \psi_s - \frac{Q_{acc}}{C_{ox}} \tag{3.3}$$

where V_{fb} is the flat-band voltage determined by the work-function difference between the gate electrode and the n-type silicon (i.e. SDE region) given by,

$$V_{fb} = -\frac{E_g}{2q} + \psi_B = -\frac{E_g}{2q} + \frac{kT}{q} ln(\frac{N_d}{n_i}) \tag{3.4}$$

Therefore, R_{acc} can be expressed approximately as:

$$R_{acc} = \frac{\rho_{acc} L_{ov}}{X_c} = \frac{L_{ov}}{\mu_{acc} C_{ox} V_g X_c} \tag{3.5}$$

where, L_{ov} is the gate ovelap length and X_c is the accumulation-layer thickness, represented in Figure 3.1.

The electron mobility in the accumulation layer has a similar field dependence to that of the inversion layer. It is not limited entirely by the impurity scattering for moderately high doping levels at the surface; the Coulombic scattering limits the mobility, because the

carrier concentration in the accumulation region greatly exceeds the donor concentration at the gate overlap region, since screening of Coulombic scattering occurs in this case.

Sheet resistance

The sheet resistance of the source/drain region is simply,

$$R_{sh} = \rho_{sd}\frac{S}{W} \tag{3.6}$$

where W is the device width, S is the spacing between the gate edge to the contact, and ρ_{sd} is the sheet resistance of the source/drain diffusion, as illustrated in Figure 3.1(b).

Spreading resistance

An analytical expression has been derived for the spreading resistance, R_{sp}, assuming an idealized case when the current spreading takes place in a uniformly doped medium with resistivity ρ_j:

$$R_{sp} = \frac{2\rho_j}{\pi W}\, ln(0.75\frac{X_j}{X_c}) \tag{3.7}$$

where W is the device width and X_j and X_c are the junction depth and the accumulation layer thickness, respectively.

Contact Resistance

The contact resistance can be expressed as:

$$R_{co} = \frac{\sqrt{\rho_{sd}\rho_c}}{W} coth(l_c\sqrt{\frac{\rho_{sd}}{\rho_c}}) \tag{3.8}$$

where l_c is the width of the contact window and ρ_c is the interface contact resistivity (in Ω-cm^2) of the ohmic contact between the metal and silicon. Equation (3.8) has two limiting cases: the short contact and long contact limits. In the short contact case, $l_c \ll \sqrt{\rho_c/\rho_{sd}}$, thereby

$$R_{co} = \frac{\rho_c}{Wl_c} \tag{3.9}$$

is dominated by the inter-facial contact resistance. In the long contact limit, $l_c \gg \sqrt{\rho_c/\rho_{sd}}$, thereby

$$R_{co} = \frac{\sqrt{\rho_{sd}\rho_c}}{W} \tag{3.10}$$

This is independent of l_c, since most of current flows into the front edge of the contact.

In practice, however, it is difficult to apply these equations, since current flows in the region where the local resistivity is highly nonuniform due to the lateral source/drain doping gradient.

3.2.2 Calculation of External Resistance

From the conventional drift-diffusion theory of electron transport in semiconductors, the sheet resistivity (R_{sh}) in non-uniformly doped regions is calculated from an incremental

form of Ohm's law [81]:

$$I_{ds} = \int_0^\infty J(y)dy = \frac{d}{dx}\phi_n(x)\int_0^\infty qn(x,y)\mu(x,y)dy = \frac{\frac{d}{dx}\phi_n(x)}{R_{sh}(x)} \quad (3.11)$$

where x is the direction parallel to the Si surface, y is the direction perpendicular to the interface, and ϕ_n is the electron quasi-Fermi level obtained from two-dimensional device simulation. I_{ds} is independent of x due to current continuity; the sheet resistivity is simply proportional to the lateral gradient of $\phi_n(x)$. This expression is valid as long as the device is sufficiently wide and the current flow is largely parallel to the x-direction [27]. Also, the total resistance (R_{tot}) is calculated based on the integration of $R_{sh}(x)$ along the direction from source to drain. While the accumulation layer in the HDD region is similar in many respects to that of the inversion layer in the channel, certain fundamental differences exist because of the nature of scattering mechanisms as carriers enter the accumulation region. Hence, a unified mobility model for the inversion and accumulation layers is needed for precise resistance calculations as previously reported [78].

Due to the two-dimensional nature of the electron gas in the accumulation layer, mobility degradation occurs in a manner similar to that in the inversion layer owing to the transverse electric field. For an accurate accumulation region resistance (R_{acc}) calculation, however, the mobility model should account for screening due to Coulombic scattering. This is because the carrier mobility at low to moderate fields in the accumulation layer is different from that in the inversion layer; the carrier concentration in the accumulation layer is always higher than that in the inversion layer due to the higher background dopant concentrations [82]. Another principal difference between Coulombic scattering in the accumulation and inversion layers is that electrons are scattered by positively charged donor

atoms in an accumulation layer, whereas electrons are scattered by negatively charged acceptor atoms in the inversion layer for a NMOS transistor. Thus, Coulombic scattering is stronger in the accumulation layer compared to the inversion layer. Without proper mobility modeling, lower external resistance is predicted, resulting in an under estimation of degradation in device performance due to R_{acc}.

Coulombic scattering consists of two parts: a screened part that varies with electron concentration and an unscreened part that is independent of electron concentration. In the limit of low electron concentration, the screened part approaches the unscreened part. Unscreened Coulombic effects are important in weak inversion, where they directly affect off-current. As carrier density in the channel increases, screened Coulombic scattering becomes important – the effect is mostly in altering the I-V curves near threshold [78]. In this work, the mobility model considers both screened and unscreened Coulombic contributions as well as the phonon and surface roughness scattering terms expressed as [81]:

$$\frac{1}{\mu_{total}} = \frac{1}{\mu_{phonon}} + \frac{1}{\mu_{surface}} + \frac{1}{\mu_{coulomb}} \tag{3.12}$$

where μ_{total} is total carrier mobility, μ_{phonon} and $\mu_{surface}$ are mobility terms that account for acoustic phonon and surface roughness, respectively. The Coulombic scattering term of Equation (3.12) considers both screened and unscreened scattering given by

$$\mu_{coulomb} = max\left[D_1 \frac{n^{\kappa}}{{N_a}^{\beta_1}}, \frac{D_2}{N_a^{\beta_2}}\right] \tag{3.13}$$

where n is the local carrier concentration, N_a is the background dopant density, and other Coulombic scattering parameters are determined empirically as reflected in ref. [30]. In Equation (3.13) the first term represents the screened contribution and the second term represents the unscreened Coulombic scattering. As the local carrier density (n) in either the

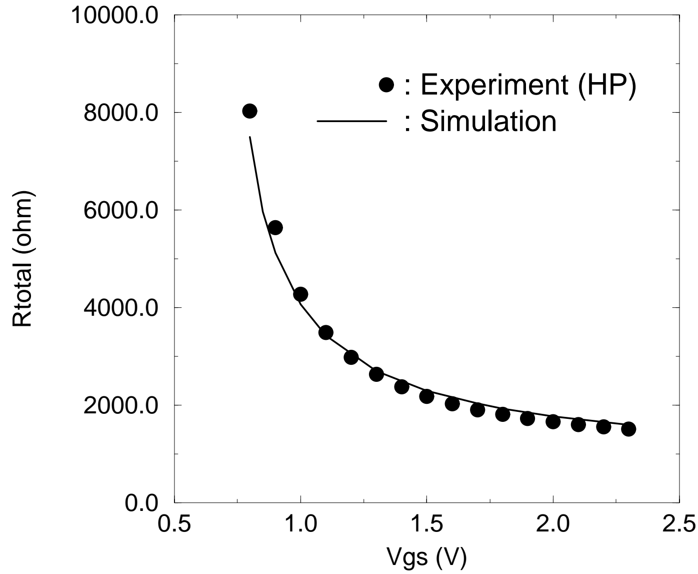


Figure 3.2: Comparisons of experimental (HP Labs.) and simulated total resistance (R_{tot} = R_{sd} + R_{ch}) of L_{eff} = 0.25 μm NMOS for V_{ds} = 0.1 V, the unified mobility model for inversion / accumulation layers is used in device simulation.

accumulation or inversion layer increases, screened Coulombic scattering term becomes dominant, causing the mobility to increase until the phonon scattering term becomes dominant. Conversely, as electron density decreases, unscreened mobility becomes dominant, retaining the mobility at a constant value.

Figure 3.2 shows measured and simulated total resistance (R_{tot}) using the mobility model as a function of the V_{gs} for the NMOS with L_{eff} = 0.25 μm, which was fabricated in the Hewlett-Packard Laboratories.

With the process and device models used in simulation for an L_{eff} = 0.25 μm experimental transistor, the L_{eff} = 0.08 μm NMOS simulated structure was created using 2D process simulation to investigate the external resistance effects for sub-0.1 μm MOSFETs.

The simulated shallow SDE uses an arsenic implantation with dose of $1.0 \times 10^{15}\ cm^{-2}$, and energies in the range of 5 – 30 KeV, producing SDE depths ranging from 30 to 70 nm.

3.2.3 Shallow SDE and Gate Overlap Effects

Figure 3.3(a) shows the computed sheet resistivity along the channel direction for different V_{gs} biases in the linear operation region (i.e., V_{ds} = 0.05 V).

It should be noted that the sheet resistivity in the channel and accumulation regions (outside the metallurgical channel) is highly dependent on gate bias, and the peak resistivity occurs in the accumulation region under the gate overlap (L_{ov}) for $V_{gs} > V_{th}$ = 0.45 V. This gate-modulated series resistance is associated with the laterally finite source/drain doping gradient. Because of this finite lateral gradient, current injection from the surface inversion layer into the source/drain does not occur immediately at the metallurgical junction. This implies that a significant voltage drop occurs in the extrinsic region. Namely, the sheet resistance in the accumulation region and current-crowding/spreading region is comparable to the sheet resistance in the inversion layer in sub-0.1 μm MOSFET's. The contact resistance (R_{co}) for simulation is assumed to be 4 $\Omega/\Box$, which is a typical value achieved in silicide processes.

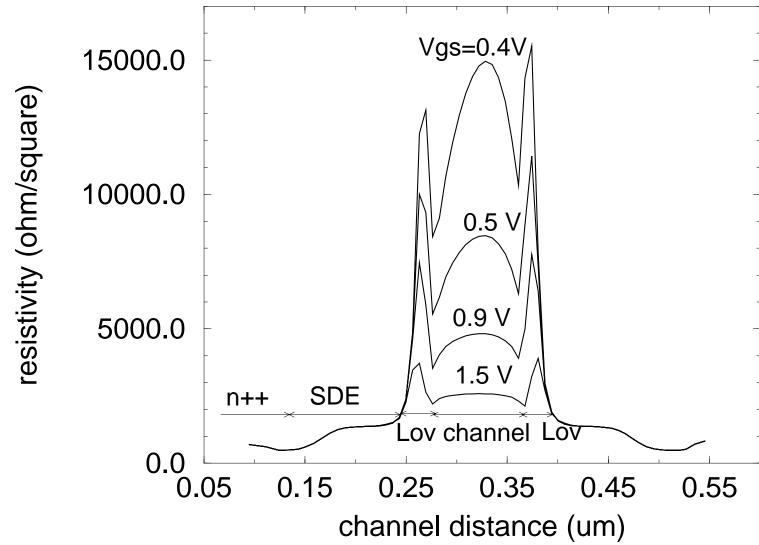
Figure 3.3(b) shows R_{ext} with respect to the gate bias for two gate oxide thicknesses (2.0 and 4.0 nm). Quantum mechanical effects due to the thin gate oxides are not considered in this work. The accumulation resistance (R_{acc}) decreases as the gate oxide thicknesses are reduced, because R_{acc} is proportional to t_{ox}.

The portions of external resistance (R_{ext}) and accumulation resistance (R_{acc}) relative to the total resistance (R_{tot}) become dominant as the device sizes are scaled down, as shown in Figure 3.4.

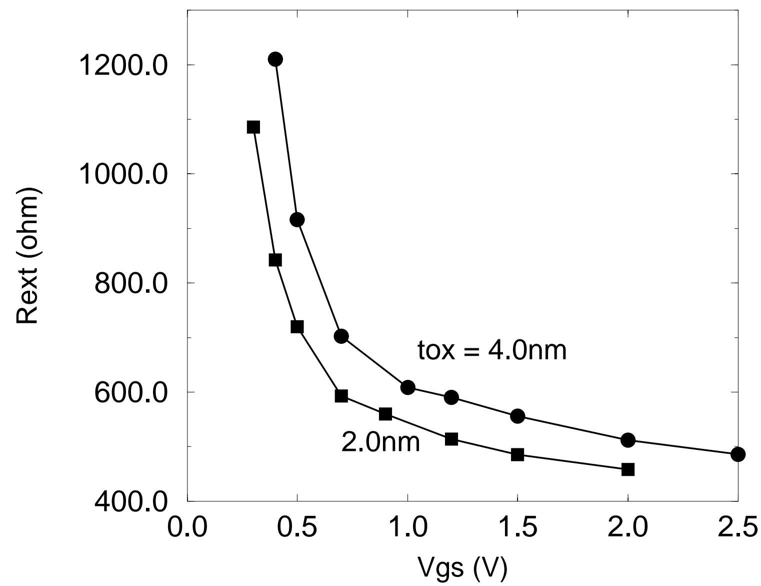
As can be seen, the simulated ratios of R_{ext}/R_{tot} and R_{acc}/R_{tot} in L_{gate} = 0.07 μm are

more than twice that observed for L_{gate} = 0.25 μm.

In addition, R_{ext} increases as the SDE depth (X_j) becomes shallower than 50 nm, as shown in Figure 3.5(a). Also, R_{ext} increases sharply when L_{ov} is shorter than 20 nm because the gate coupling to the SDE becomes poor, as shown in Figure 3.5(b). Thus, careful optimization of X_j and L_{ov} is necessary to minimize the R_{ext}. Otherwise, I_{dsat} and the SDE to gate coupling can be degraded even though the short channel effects are reduced.



(a)



(b)

Figure 3.3: Simulated external resistance with respect to V_{gs} bias (V_{ds} = 0.05 V). L_{eff} = 0.08 μm and X_j = 40 nm. (a) Resistivity along the channel direction, t_{ox} =2.0 nm. (b) R_{ext} ($= R_s + R_d$) with respect to V_{gs} when t_{ox}'s are 2.0 and 4.0 nm, respectively.

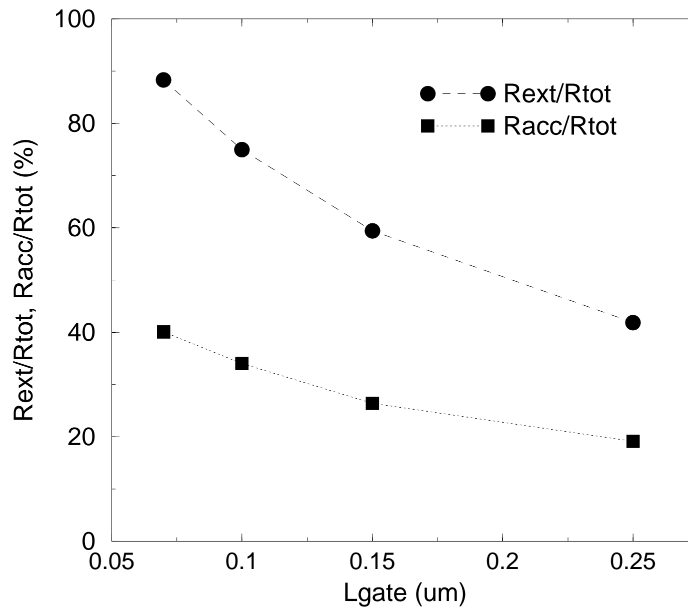
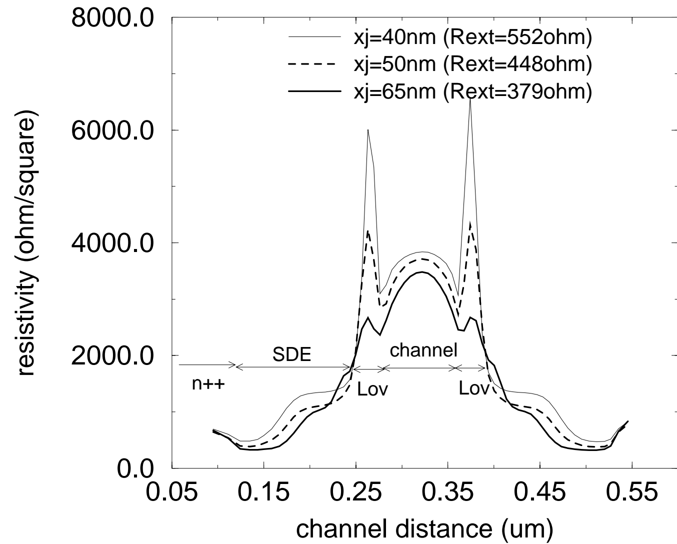
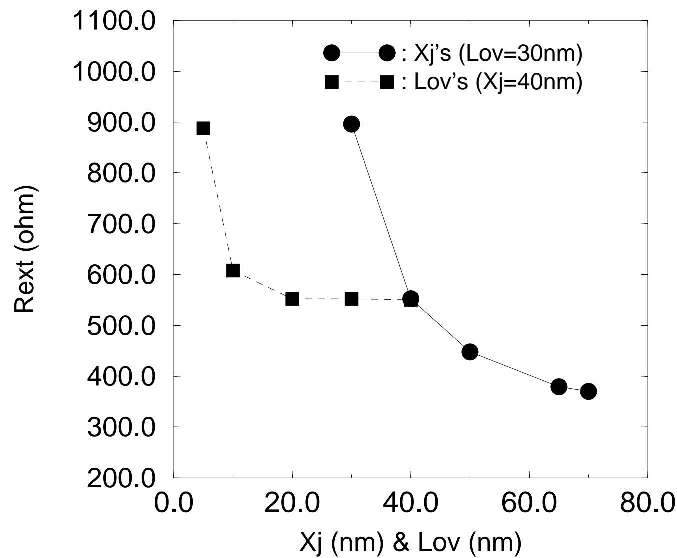


Figure 3.4: The ratios of R_{ext}/R_{tot} and R_{acc}/R_{tot} for different drawn gate lengths (L_{gate}) ranging from 0.25 to 0.07 μm, V_{gs} = 1.5 V and V_{ds} = 0.05 V

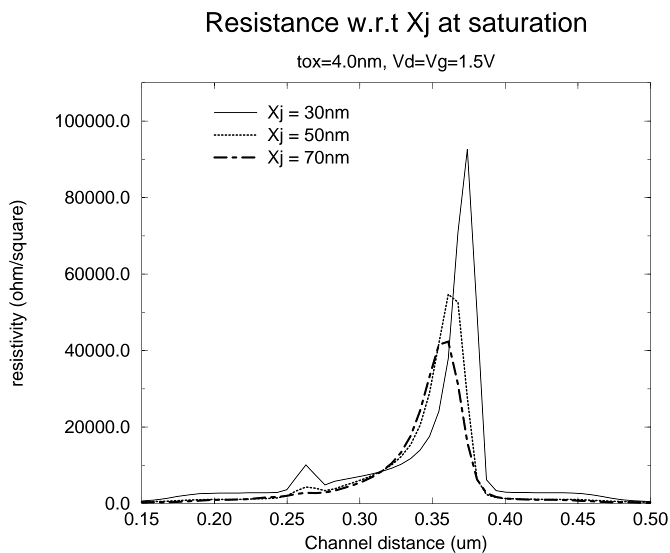


(a)

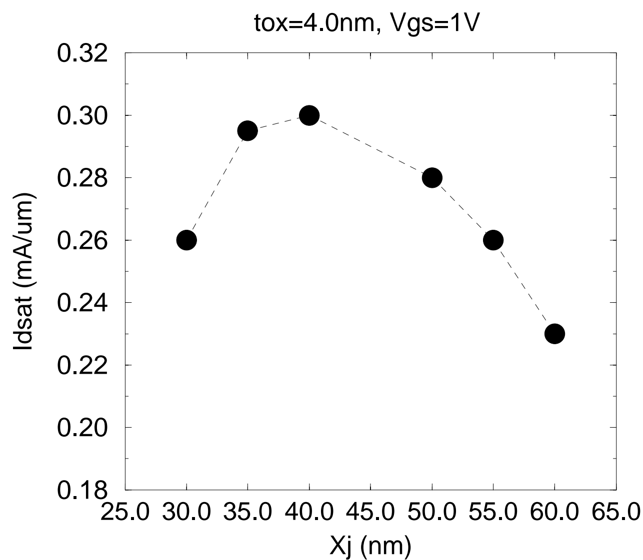


(b)

Figure 3.5: Simulated external resistance with respect to the source-drain extension (SDE) depth (X_j) when V_{gs} = 1.5 V and V_{ds} = 0.05 V (L_{eff} = 0.08 μm and t_{ox} = 4.0 nm). (a) Resistivity along the channel direction for X_j = 40, 50, and 65 nm. (b) R_{ext} $(= R_s + R_d)$ with respect to different X_j and L_{ov}.



(a)



(b)

Figure 3.6: External resistance with respect to X_j in the saturation region ($L_{eff} = 0.08\mu$m and t_{ox} = 4.0 nm). (a) R_{ext} with respect to X_j in the saturation region, V_{ds} = V_{gs} = 1.5 V. (b) Simulated on-current (I_{on}) for different SDE X_j's at V_{gs} = 1.0 V, each V_{dsat} for different X_j's is chosen at the V_{ds} when 1 nA/μm of off-leakage current is produced.

The R_{sh} distributions for various values of X_j in the saturation region where V_{ds} = V_{gs} = 1.5 V are shown in Figure 3.6(a), where the maximum resistivity appears in the drain accumulation region. Here, R_d increases as X_j becomes shallower, which implies I_{dsat} degradation for the SDE depths less than 40 nm. In order to observe the R_{ext} effects on the on-current (I_{on}), simulations were performed for the different values of X_j as in Figure 3.6(b). To consider the short channel effects (i.e., DIBL) induced leakage current, the V_{dsat} for each X_j condition was chosen as the drain bias that produces 1 nA/μm of off-state leakage current. As a result, I_{on} degradation is observed for X_j values less than 40 nm, which corresponds to the experimental trend in ref. [8].

3.2.4 Summary

Accurate determination of external resistance in scaled MOS devices can be addressed using proper mobility modeling. External resistance in the accumulation layer (R_{acc}) becomes dominant for the shallow SDE depths less than 40 nm. R_{acc} is highly dependent on the junction depth and overlap length, which should be minimized for improvement of current driving capability and SDE coupling to the gate, contrary to future trends that project using raised source and drain structure [9]. An NMOS transistor with L_{eff} of 0.08 μm shows optimum on-current when L_{ov} is 20 nm and X_j is 40 nm.

3.3 Analysis and Optimization of 70 nm Laser Thermal Processed MOSFETs

As MOSFET scaling continues, ultra-shallow and highly-activated junctions are essential for device performance and control of short-channel effects. In addition, lateral abruptness of the source/drain extension regions is indispensable to minimize short-channel effects, because external resistances play an important role in determining electrical performance as devices are further scaled to sub-100 nm dimensions [83]. As shown in Figure 3.7, abrupt source and drain junction profiles are much more desirable than graded profiles in terms of drive current and short-channel effects. Super-doped extension regions, instead of the conventional LDD or HDD, will be required along with the aggressive scaling of junction depth and abruptness in the near future. Recent low-energy ion implantation has achieved implant energies as low as 200 eV. However, as junction depth drops below 100 nm, Rapid Thermal Process (RTP) faces problems of undesired thermal diffusion and low activation that is limited by solid solubility, so that the minimum junction depth is no longer governed by the ion energy alone in the low energy regime. Instead, channeling effects and transient enhanced diffusion (TED) contribute substantially to the junction formation, therefore activating the dopant without diffusion is a significant challenge. Recently, new techniques have been developed to reduce the short-channel effects due to the source and drain field effects. For example, low temperature activation of source/drain impurities by using recrystallization was introduced to keep an abrupt lateral source/drain junction profile, but it has drawbacks such as long thermal cycle and relatively high level of gate-induced drain leakage (GIDL) [84].

Laser Thermal Process (LTP) has been actively studied as an alternative to RTP in order to achieve ultra-shallow SDE and salicide contact formation [85]. Since LTP provides very short and spatially localized thermal budgets in activating the source/drain dopant up to the

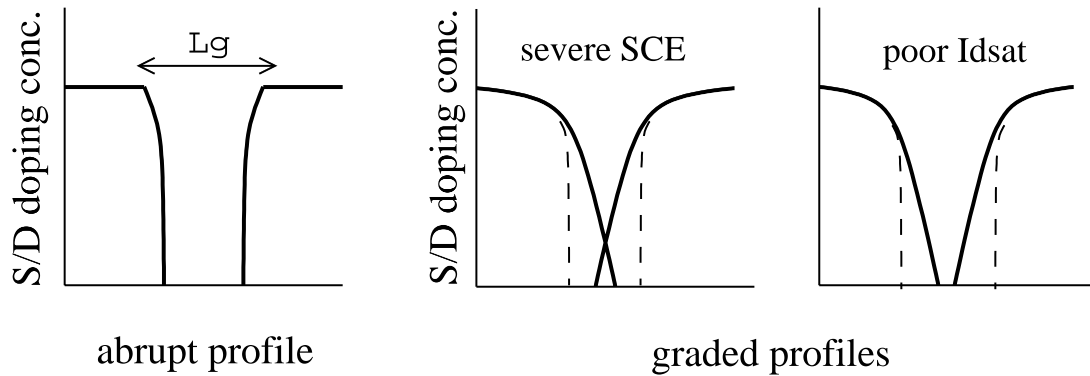


Figure 3.7: Comparison of lateral source/drain profiles (After [84]).

solid solubility, more aggressive junction depth scaling is allowed by reducing undesired dopant redistributions induced by the implant damage.

In this section, comparisons of device characteristics for LTP- and conventional RTP-processed 70 nm NMOSFETs are presented. Also, optimization of LTP-processed MOSFETs to improve on-current, while minimizing parasitic capacitance and resistance effects in the SDE region is discussed based on process and device simulation results.

3.3.1 Dopant Profiles of LTP and RTP

The fabrication flow of an LTP-processed 70 nm transistor is shown in Figure 3.8 [85]. After the gate patterning, a Si_3N_4 spacer was formed, followed by a deep source/drain implant. A high-temperature RTP was used for dopant activation in the source/drain and the polysilicon gate. After removing the spacer, shallow source and drain junctions were formed with a low-KeV, high dose source/drain implant. Then, the dopants were electrically activated by LTP with a pulsed 308 nm XeCl excimer, ultraviolate laser beam, scanning across the wafer field by field.

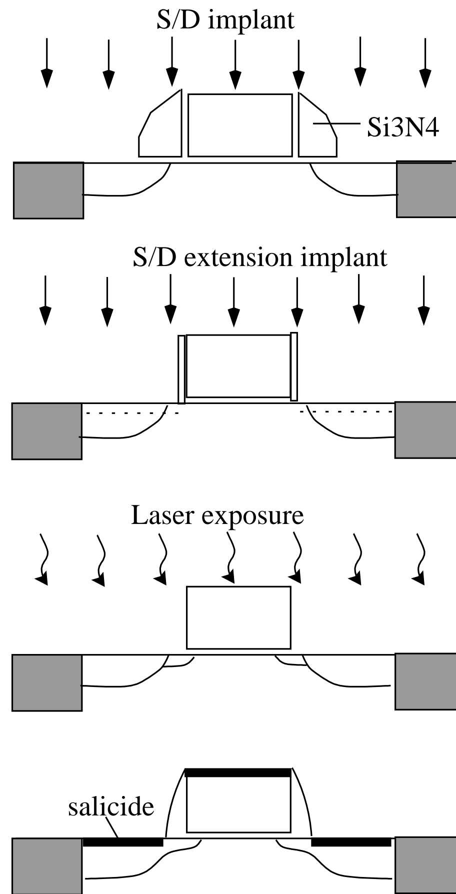


Figure 3.8: Process flow for an LTP NMOS transistor (After [85]).

Figure 3.9 shows the current-voltage characteristics of an LTP NMOS with L_g = 70 nm. The relatively high off-current in the I_{ds} vs. V_{gs} may be caused by the substrate punch-through leakage due to the absence of halo or retrograde channel implants. Room exists for further device performance improvement; device and process parameters should be carefully optimized to take full advantage of the LTP process.

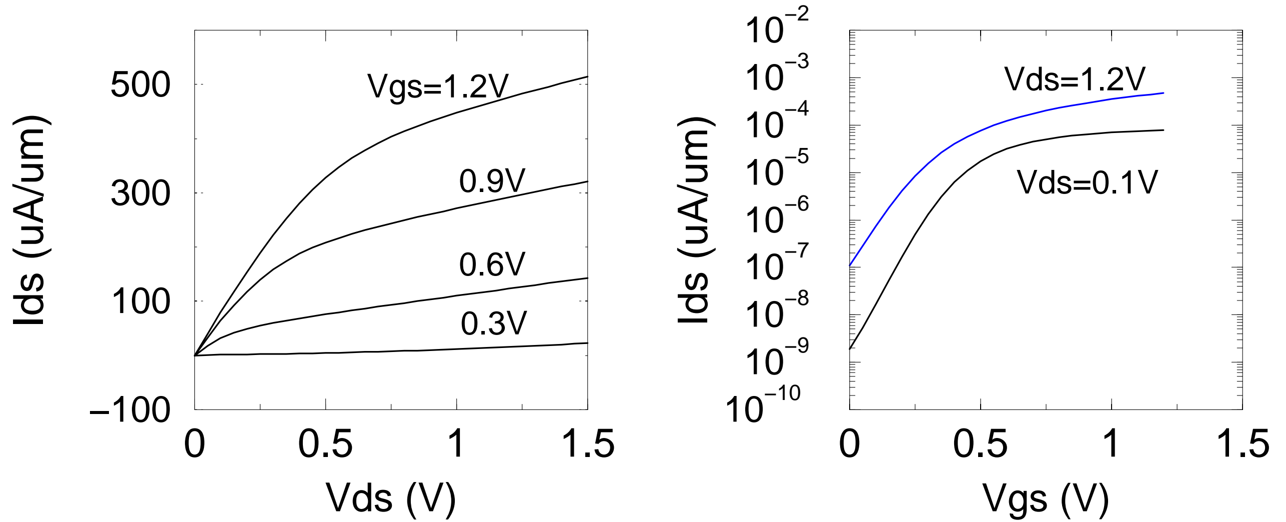


Figure 3.9: Measured I_{ds}–V_{ds} and I_{ds}–V_{gs} of the AMD 70 nm LTP NMOSFETs.

Figure 3.10 shows comparisons of arsenic dopant profiles obtained using ion mass spectroscopy (SIMS) and simulated LTP. The simulated LTP profile was obtained by adjusting moment-related parameters of the analytic ion implantation model in TSUPREM4 – those which determine the shape of the dopant distribution. Specifically parameters control range, standard deviation and skewness of profiles were adjusted. The simulated RTP profile for an implant dose of $6\times10^{14}/cm^2$ was obtained by considering the TED effects with the processing conditions of 1010 °C and 5 seconds. In this figure, the surface dopant concentration of the LTP is higher than that of the RTP due to highly activated dopants. In addition, the junction depth of the LTP is shallower than that for RTP due to minimized TED effects. It was reported that the thin molten silicon layer, generated by a few seconds of pulsed laser, is grown back to crystalline silicon during the LTP process. As a result, about 8 orders of magnitude higher dopant diffusivity are achieved, leading to almost uniform and abrupt dopant profiles [86]. Furthermore, since the melting and recrystallization

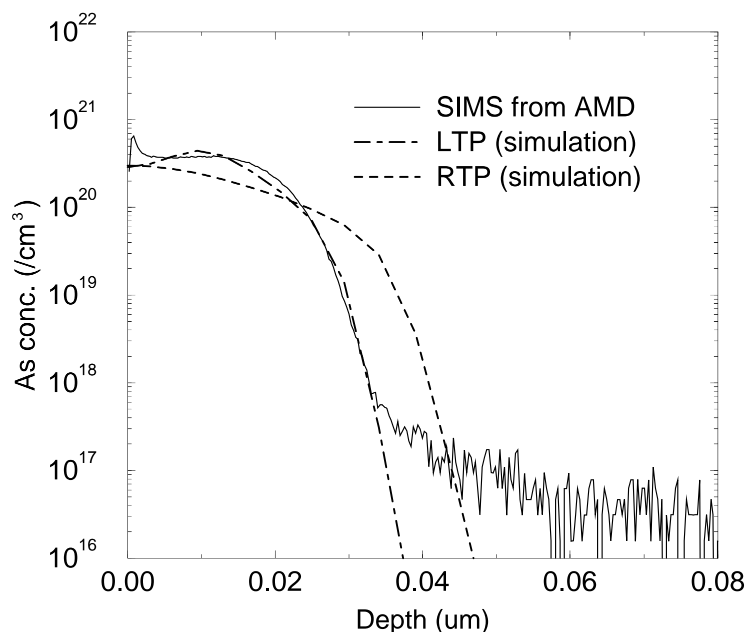
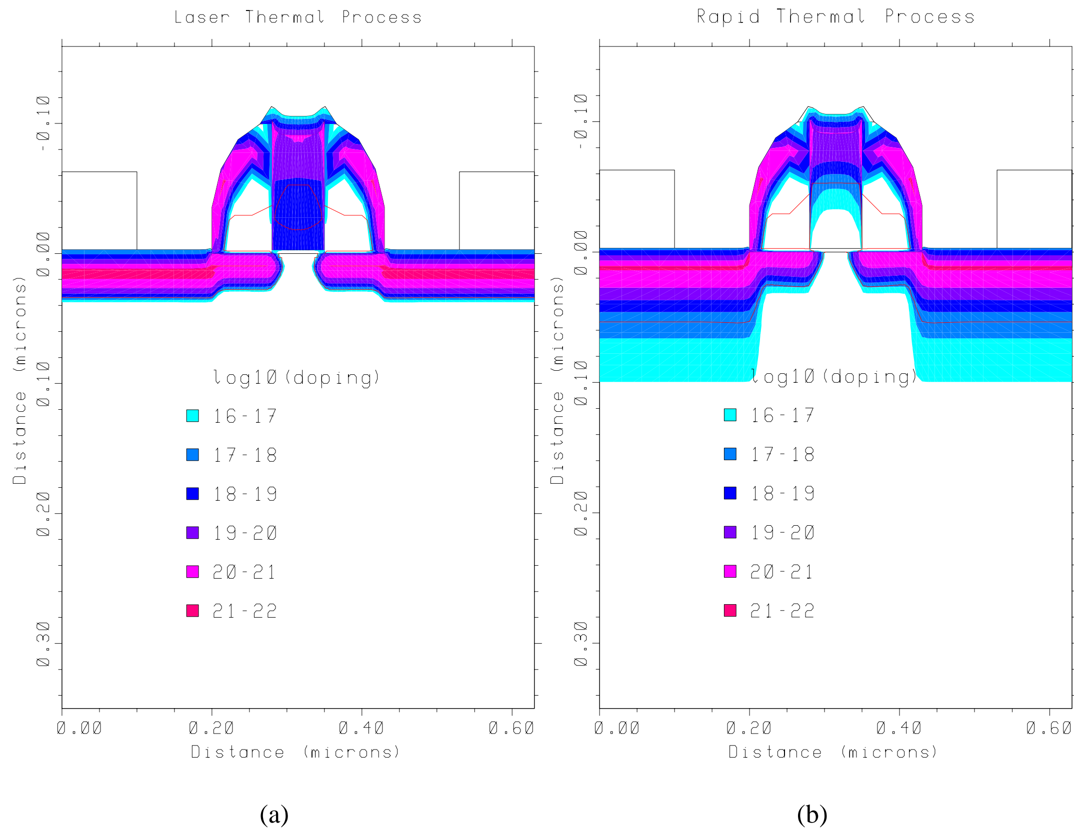


Figure 3.10: Arsenic dopant profiles for LTP (SIMS and process simulation) and RTP (process simulation), Arsenic implant dose and energy are $6 \times 10^{14}/cm^2$ and 2 KeV, RTP temperature and time are 1010°C and 5 seconds, respectively.

time is less than 100 ns, the activated dopant concentration reaches the solid solubility limit.

Figure 3.11 shows simulated dopant distribution contours for 70 nm NMOSFETs in both the LTP and RTP cases. The same ion implantation model parameters are used; they were previously calibrated to the SIMS data. Figure 3.12 shows net doping profiles of LTP- and RTP-processed devices along the silicon surface. From these results it can be observed that the SDE junction, using RTP, is deeper ($X_j \sim$ 40 nm) than for the LTP case ($X_j \sim$ 30 nm), and the lateral dopant diffusion for RTP is more severe than that of LTP. By using the LTP process, higher arsenic dopant concentrations with steeper dopant abruptness in the SDE region are achievable – these results are more desirable for improving device



(a) (b)

Figure 3.11: Comparisons of simulated arsenic dopant profiles (contour) of 70 nm NMOS transistor, (a) LTP case, (b) RTP case.

performance in sub-100 nm MOSFETs.

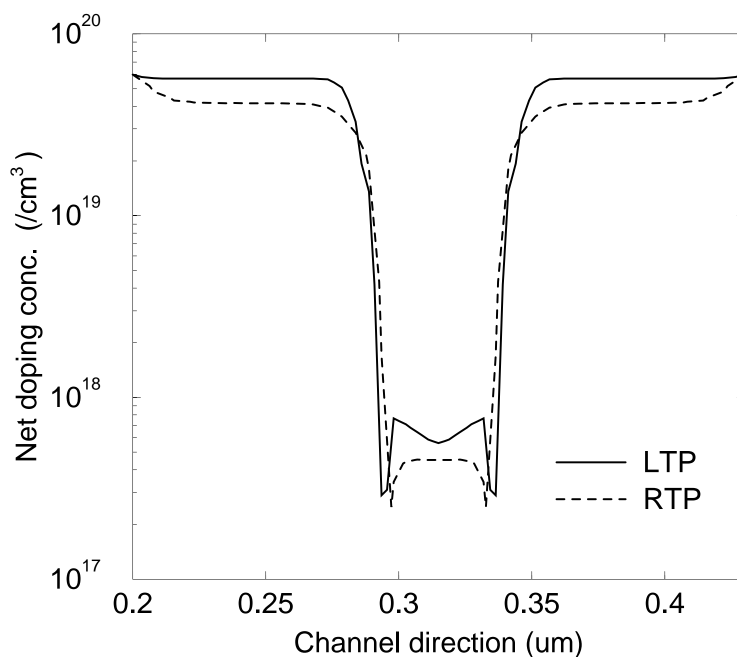


Figure 3.12: Simulated net dopant profiles along the silicon surface for LTP and RTP.

3.3.2 Calculation of External Resistance

For accurate calculation of the source and drain resistance for LTP-processed devices, the previously mentioned unified mobility model was used. Prior to the calculation, simulations were performed and model parameters were tuned to the measured I-V data. Figure 3.13 shows the computed and measured total resistance (R_{tot}) versus V_{gs} for an NMOS with an L_g = 70 nm.

Figure 3.14(a) shows the computed sheet resistivity along the channel direction for LTP- and RTP-processed transistors. The resistivity in the LTP case is lower than for the RTP case due to the higher dopant concentration in the SDE, even though the SDE junction depth of the LTP is shallower than for the RTP. Figure 3.14(b) shows sheet resistivity of

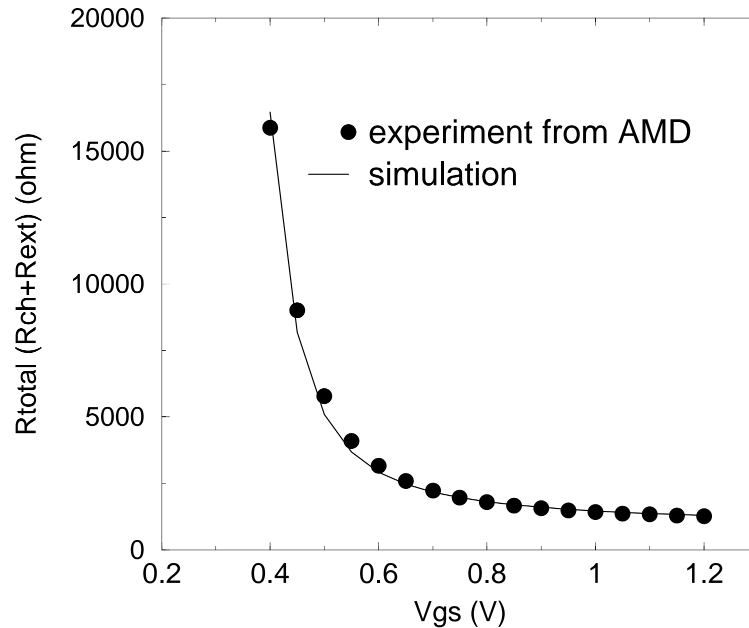
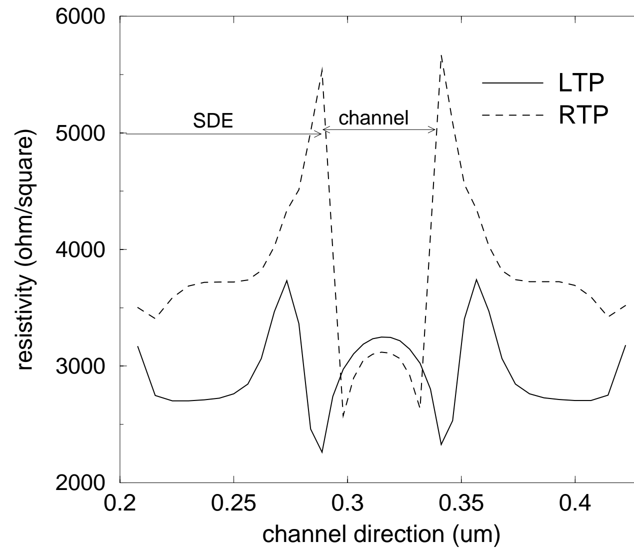
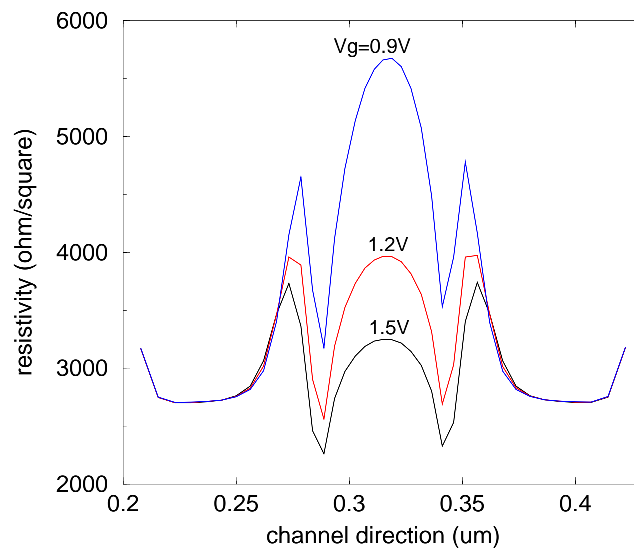


Figure 3.13: Measured (AMD) and simulated total resistance (R_{tot}=R_{ch}+R_{sd}) vs. V_{gs} using the unified mobility model, V_{ds} = 0.1 V and L_g = 70 nm NMOS.

the LTP for different V_{gs} biases in the linear region (i.e. V_{ds} = 0.1 V). It should be noted that the sheet resistivity in the channel and accumulation regions (outside the metallurgical channel) is highly dependent on gate bias, thus the high resistivity occurs in the accumulation region under the gate overlap (L_{ov}), since this gate-modulated series resistance is associated with the laterally finite source/drain doping gradient. Because of the finite lateral gradient, current injection from the surface inversion layer into the source/drain does not occur immediately at the metallurgical junction, leading to a voltage drop in the extrinsic region. The contact resistance (R_{co}) for the simulations is again assumed to be 4 $\Omega/\square$, which is a typical value for a silicide process.



(a)



(b)

Figure 3.14: Computed sheet resistivity along the channel direction for LTP and RTP devices, (a) comparisons of resistivity distributions for LTP and RTP, V_{gs} = 1.5 V and V_{ds} = 0.1 V (b) variations of sheet resistivity vs. gate biases for LTP case, V_{ds} = 0.1 V.

3.3.3 Device Optimization in LTP Transistor

For high performance LTP-processed transistors, careful choice of junction depth (X_j) and overlap length (L_{ov}) is necessary. Otherwise, I_{on} and coupling of the SDE to the gate will be degraded even though short channel effects are reduced.

Figure 3.15(a) shows a comparison of on-current (I_{on}) versus different gate overlaps (L_{ov}) for LTP- and RTP-processed devices. Here, I_{on} was extracted with respect to the drain bias ($V_{ds,on}$) that produces 1 nA/μm of I_{off} at V_{gs} = 0 V. Overall values of I_{on} for LTP are higher than those for the RTP and the peak appears at $L_{ov} \sim$ 10 nm. Figure 3.15(b) illustrates the normalized I_{on} and gate overlap capacitance (C_{ov}) for different values of L_{ov}. Note that a very short L_{ov} produces lower on-current, because it results in poor gate control on the channel, leading to high external resistance. Meanwhile, a long overlap reduces the effective channel length and causes severe short channel effects and lower I_{on}. Also, the longer L_{ov} produces higher overlap capacitance in the accumulation region; 5–10 nm of L_{ov} shows the highest I_{on} while maintaining relatively lower overlap capacitance in 70 nm transistors.

Figure 3.16 shows external resistance ($R_{ext} = R_s + R_d$) and I_{on} for various SDE junction depths (X_j). R_{ext} abruptly increases as X_j becomes shallower than 20 nm; I_{on} is degraded due to the high external resistance for $X_j <$ 20 nm. I_{on} is also degraded for $X_j >$ 30 nm due to severe short channel effects. Note that I_{off} becomes higher due to severe short channel effects, thus both $V_{ds,on}$ and I_{on} become lower. By considering the trade-offs between R_{ext} and the short channel effects, an optimum X_j is 20–25 nm. Resistivity distributions for various X_j's at V_{ds} = 0.1 V and V_{gs} = 1.2 V are shown in Figure 3.17. Here, the maximum resistivity appears in the SDE region, which implies substantial I_{on} degradation for a SDE depth less than 20 nm. The impact of spacer length on external resistance becomes significant as the spacer length increases, as shown in Figure 3.18. I_{on} is less

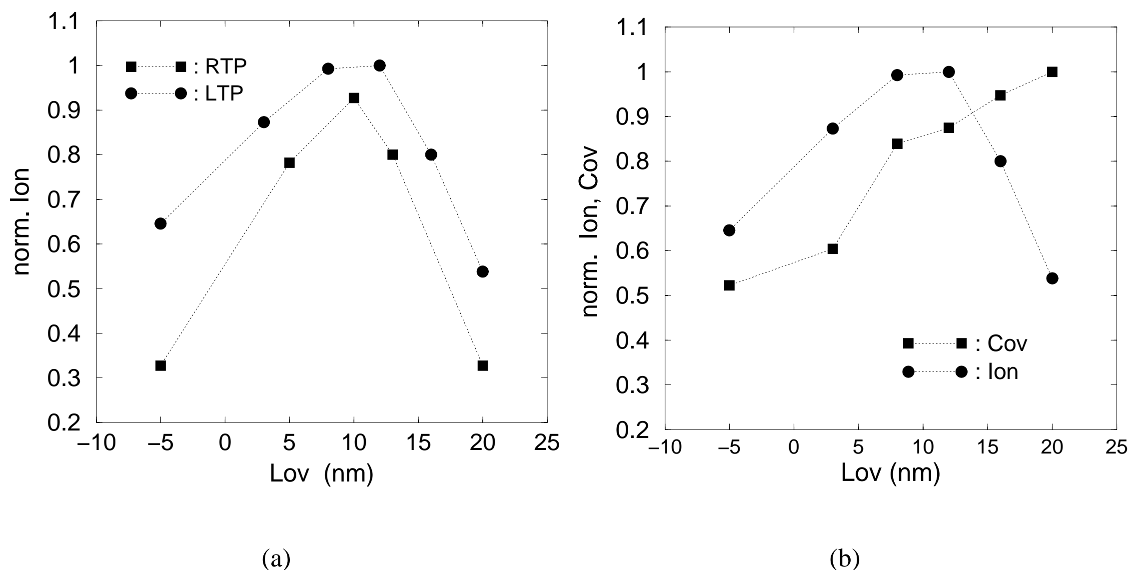


Figure 3.15: I_{on} and C_{ov} versus L_{ov} for LTP and RTP devices. (a) I_{on} of LTP and RTP for different gate overlaps (L_{ov}), I_{on} was chosen at the drain bias (V$_{ds,on}$), producing off-current (I_{off}) of 1 nA/μm (V_{gs} = 0.9 V). (b) Normalized I_{on} and gate overlap capacitance (C_{ov}) for various overlap lengths (L_{ov}).

dependent on the spacer length for X_j = 20 nm, since the resistance of shallow junctions is the dominant external resistance. However, I_{on} is degraded as spacer length increases for X_j = 30 nm, because high external resistance in the long spacer length becomes dominant in this case.

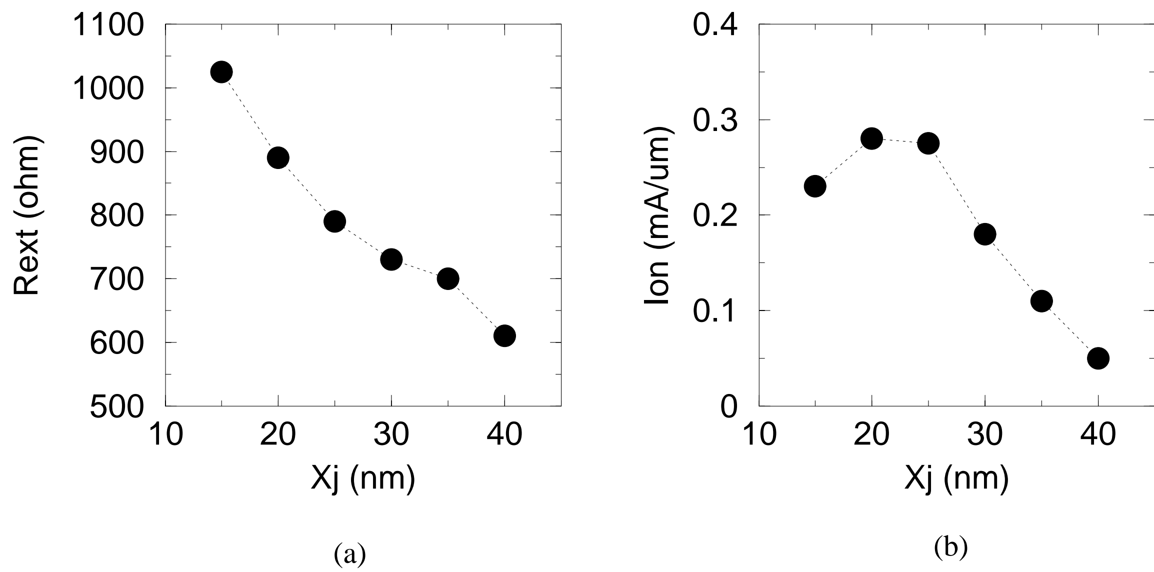


Figure 3.16: R_{ext} and I_{on} current for various SDE junction depths (X_j) in LTP process, (a) R_{ext} vs. X_j (V_{ds} = 0.1 V and V_{gs} = 1.2 V), (b) I_{on} vs. X_j (V_{ds}'s at I_{off} = 1 n A/μm and V_{gs} = 0.9 V).

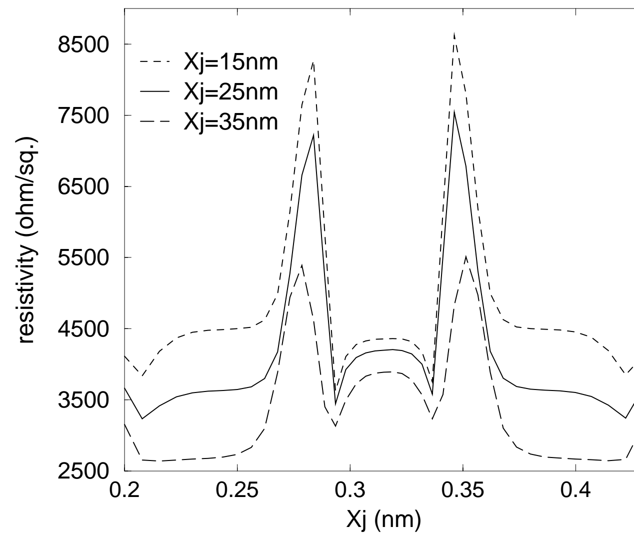


Figure 3.17: Computed sheet resistivity along the channel direction for different SDE depths (X_j).

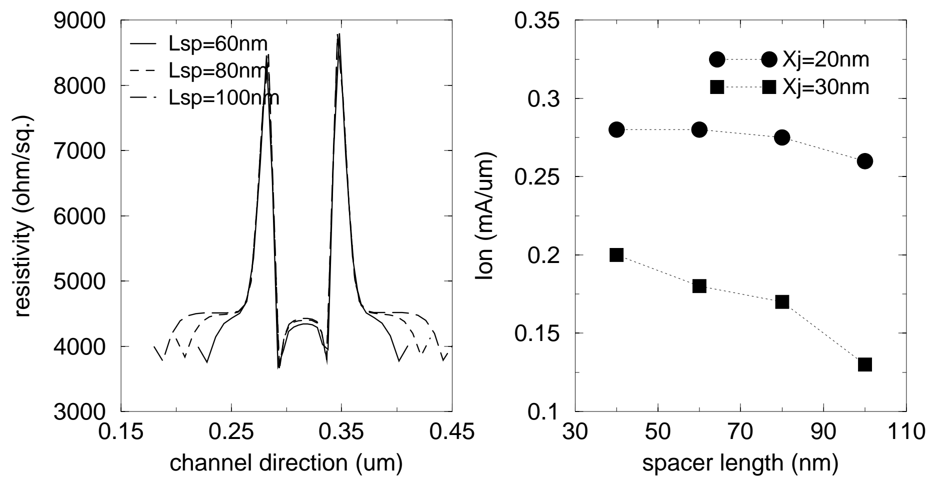


Figure 3.18: Impact of spacer length on external resistance and on-current.

3.3.4 Summary

The trade-offs between the external resistance and the short channel effects in 70 nm LTP MOSFETs fabricated by AMD are studied based on extrinsic resistance calculations. Proper mobility modeling, especially in the accumulation region, is essential for consideration of future MOSFET scaling.

Chapter 4

Channel Modeling

4.1 Introduction

Carrier transport in nanoscale transistors involves off-equilibrium transport, where the beginning of the channel is a quasi-equilibrium point, populated by carriers at thermal equilibrium [87][88]. Electrons are injected from the source into the channel across a potential barrier and its height is modulated by the gate voltage. The positively-directed velocity of the degenerate Fermi gas depends on the sub-band occupancies and the maximum velocity at the source approaches the thermal injection velocity when very few thermal carriers injected into the channel are back-scattered to their entry point into the channel. The on-current of a nanoscale MOSFET has a strong sensitivity to the low-field mobility because it determines the back-scattering in the critical region at the beginning of the channel [87]. Thus, MOSFET on-current depends on the thermal injection velocity near the source according to fundamental transport theory. Off-equilibrium carrier transport effects in nanoscale MOSFETs alter the electric field and thermal injection velocity at the source, which can be observed by using either Monte-Carlo or hydrodynamic transport models.

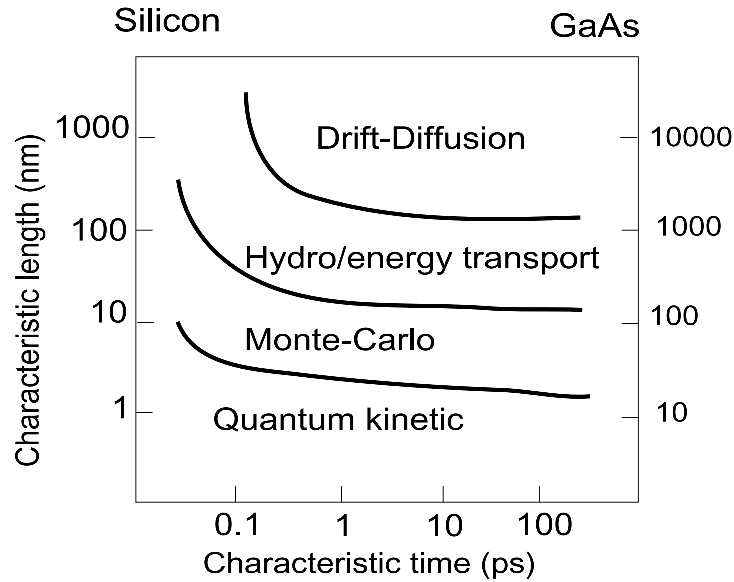


Figure 4.1: Regions of validity for various device simulation models (After [89]).

Hence, source injection and channel backscattering at the source end and velocity overshoot in the channel need to be considered in current balancing approaches. In practice, however, device performance will fall below the ballistic limit because series resistance becomes increasingly important as the channel resistance decreases.

Figure 4.1 illustrates the regions of validity for various device simulation models [89]; for deep sub-100 nm devices that switch very quickly, using drift-diffusion equation only is not adequate. A more rigorous treatment of carrier transport under spatially non-uniform high-field conditions has been traditionally carried out using a Monte Carlo solution of the Boltzmann transport equation for the electron distribution function [90][91]. Monte Carlo device simulation has been established as a powerful tool for the investigation of ballistic effects occurring in deep submicron MOSFETs. One major concern that prevents its widespread use is the large computation time required [92]. Velocity overshoot is clearly

an off-equilibrium effect and cannot be predicted with simple drift-diffusion simulations, thus it is necessary to use more sophisticated simulations. However, the same results may be achieved using a drift-diffusion model, augmented to account for velocity overshoot by means of the spatial gradient of electric field along with mobility models that include relevant scattering terms near the source.

In this chapter, a simple on-current model considering the off-equilibrium and external resistance effects to capture physical insight regarding carrier transport of nanoscale MOSFETs is discussed; computed results are compared with hydrodynamic device simulations.

4.2 Non-local Effects in Small Devices

As carriers move in the crystal under the action of the external force (i.e., the electric field caused by the voltage applied to the transistor contacts), they suffer various collisions as a result of : thermal lattice vibrations, dopant ions (both positively and negatively charged), and other carriers. If the mean distance between two successive collisions (typically, a few nanometers) is much smaller than the dimensions of the transistor, carriers collide many times while transiting the device. On the one hand, this prevents the carriers from gaining kinetic energy much in excess of the thermal energy they would have in the absence of a driving force. On the other hand, these collisions can be treated statistically, by lumping their effects as a kind of average friction on the current, without worrying too much about the details of what happens to electrons or holes in any single collision. As a useful by-product of these simplifications, one has to worry only about what happens in a very small neighborhood of each carrier. The problem now becomes "local" in the sense that only the driving force (i.e. electric field) at a given position in the device is needed to describe charge transport at that particular position. Device simulation programs that employ drift-diffusion models rely heavily on these localized approximations. They use the fact that carriers do

not acquire sufficient kinetic energy to violate the use of a simplified band structure of the crystal, hence the use of an effective mass to handle the motion of the carriers remains valid. They also use lumped concepts, such as mobility and the diffusion constant to account for carrier collisions in a grand-averaged and local way. Now that transistors have reached dimensions approaching the mean distance between collisions, these simplifying approximations fail: carriers gain significant kinetic energy, in excess of 50 to 100 times larger than their thermal energy. Therefore, the details of the electronic band-structure become important. At the same time, only a few collisions occur as carriers move across the device. Thus, it becomes important to look at single collisions with more attention and to account for the driving force everywhere they occur in the device when studying charge transport at a given position in the transistor – so-called non-local effects [93]. Finally, as devices shrink even more, their dimensions approach the wavelength of the electron. Electrons must then be treated as full-fledged quantum mechanical particles. Rather than picturing them as tiny billiard balls, they must be regarded as waves traveling across the device, reflecting off boundaries and contacts, interfering with other waves. The following discussion will not consider the rigorous quantum mechanical limits to carrier transport and scaling.

4.3 Fundamental Drain Current Model

Figure 4.2(a) represents the essential physical picture, showing that carriers are injected into the channel from a thermally equilibrated reservoir (the source), across a potential energy barrier whose height is modulated by the gate bias, into the channel, which begins at the top of the barrier. The beginning of the channel is populated by carriers injected from the source region which is at thermal equilibrium [94]. The maximum value of the average carrier velocity at the beginning of the channel is approximately the uni-directional

thermal velocity, v_T, since the positive velocity of carriers at the beginning of the channel comes from the source reservoir which is at thermal equilibrium. Backscattering from the channel determines how close to this upper limit the device operates; velocity overshoot occurs within the channel and determines the carrier density profile. Figure 4.2(b) shows a schematic fluid flow model for a MOSFET under high gate and drain bias conditions. Here, carrier transport through the drain end of the channel is rapid because velocity overshoot occurs. As a result, the current is controlled by how rapidly carriers are transported across a short low-field region near the source. The length of the current-limiting region at the beginning of the channel is denoted as l; its value is about equal to one mean-free-path for carriers.

While the carrier velocity reaches high values in the high-field region near the drain, MOSFET drain current is mainly controlled by the average carrier velocity near the source end of the channel, where the inversion charge density Q_i is determined by $C_{ox}(V_{gs} - V_{th})$, independent of the drain voltage. Carrier velocity near the source is determined by the thermal velocity as the carriers are injected from the source. Velocity overshoot near the drain helps to increase the overall MOSFET current by increasing the electric field driving force near the source. Once the device approaches the ballistic limit, the current depends only on the injection velocity from the source.

The fundamental limit in determining on-current for a MOSFET is described as [95]:

$$I_{ds} = C_{eff} W v(0)(V_{gs} - V_{th}) \tag{4.1}$$

where C_{eff} is the total gate capacitance and $v(0)$ is the average carrier velocity near the source.

There is an upper limit on the MOSFET current set by thermal injection from the source. Such a limiting current takes the same form as Equation (4.1), except that the parameter

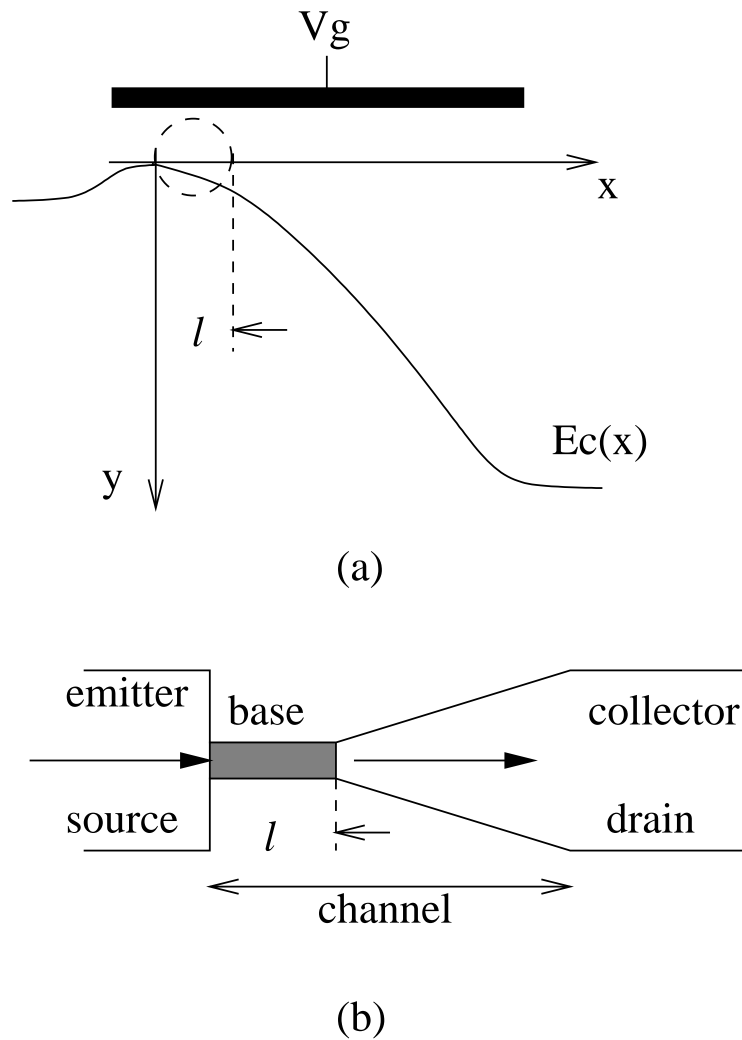


Figure 4.2: Essential physical picture for carrier transport in MOSFETs, after [94]. (a) conduction band from the source and drain, (b) fluid flow analogy under high drain and gate bias.

$v(0)$ should be replaced with the source thermal velocity v_T, which for electrons can be 1.5–2 times the velocity saturation (v_{sat}) in the high doped degenerate n^+ source. In small transistors, strong velocity overshoot occurs along the channel, affecting the electric field at

the source as well as carrier backscattering. However, these effects cannot extend the upper limit on MOSFET current, beyond the source thermal velocity v_T. $v(0)$ can be expressed as a function of both the field and scattering rates (mobility) in the channel region near the source:

$$v(0) = \mu^0 E_X(0^+) \tag{4.2}$$

where μ^0 and $E_X(0^+)$ are the low field mobility and lateral electric field near the source, respectively.

However, the total gate capacitance (C_{eff}) in Equation (4.1) should be replaced with (C_{gs}) because I_{ds} is determined by charge near the source [96]. It has been experimentally demonstrated that electron velocity exceeds the saturation velocity of bulk silicon (1×10^7 cm/s) at the room temperature, where the electron velocity (v_e) was obtained by dividing the drain current (I_d) by the source to gate capacitance (C_{gs}) [97]. In addition, the external source/drain resistance plays an important role, as it becomes comparable to the channel resistance in small devices. Thus, Equation (4.1) can be modified as follows:

$$I_{ds} = C_{gs} W v(0)(V_{gs'} - V_{th}) \tag{4.3}$$

where $V_{gs'} = V_{gs} - I_{ds}R_s$, as shown in Figure 4.3.

Validation of Equation (4.3) based on device simulations using a hydrodynamic model was performed for the structure shown in Figure 4.4(a), which has a junction depth for the source/drain extension of 30 nm and a gate oxide thickness of 2.5 nm. To observe off-equilibrium carrier transport in nanoscale devices Monte-Carlo simulations [98][99] are more useful, since they do not lump microscopic physics into the process-dependent macroscopic transport parameters. In addition, the computation of the carrier distribution

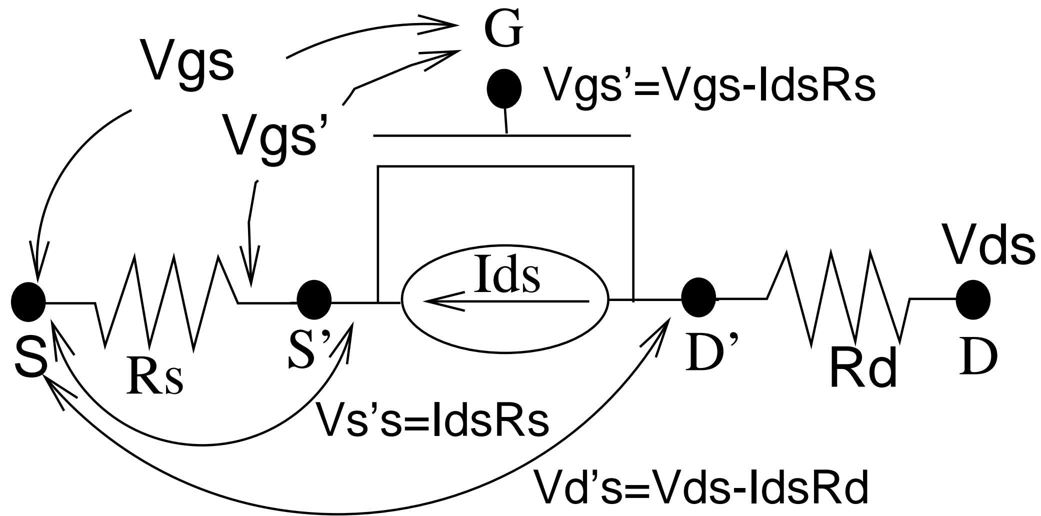
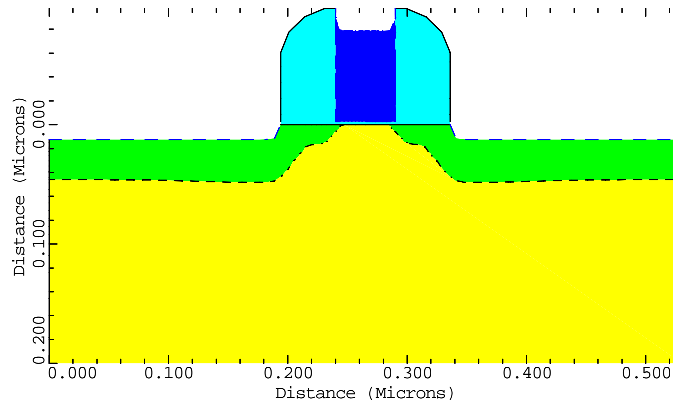


Figure 4.3: MOSFET showing source and drain external resistance, effective (intrinsic) gate voltage is denoted by $V_{gs'}$.

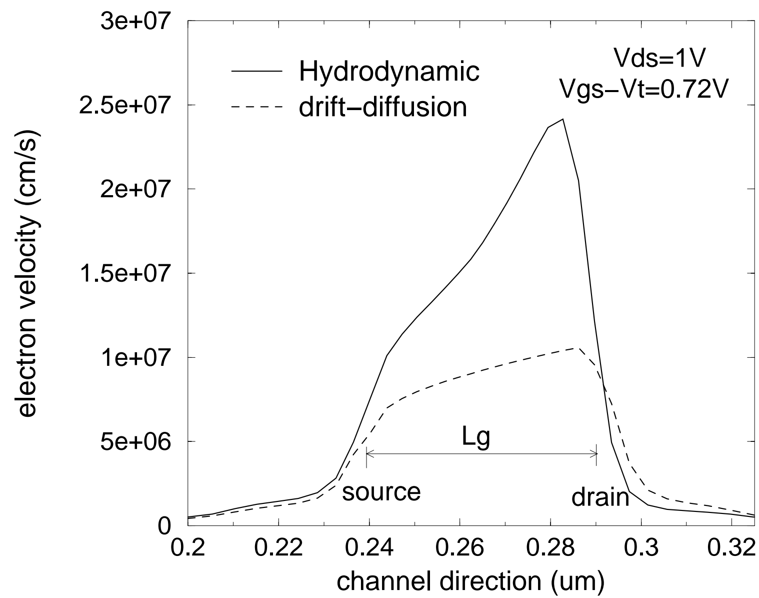
function provided by the Monte Carlo method allows one to consider energy threshold phenomena such as impact ionization and oxide injection on a solid physical basis. However, the computational burden limits its use for many device engineering applications.

The hydrodynamic (HD) model is derived by taking moments of the Boltzmann equation; however, simplifying assumptions are needed to formulate a tractable set of equations. In the HD (or alternatively the energy transport formulation) model the transport equations are derived from moments of the Boltzmann transport equation (BTE), which are similar to the hydrodynamic flow equations of used in fluid dynamics [89]. The drift-diffusion model is a first-order approximation to the momentum balance equation; it ignores the spatial variation of the average carrier energy and assumes that the mobility is uniquely specified by the local electric field. The HD model includes the energy gradient in the current equation and assumes that the mobility is a function of the average carrier energy; the first three moments of the BTE are solved in the HD model rather than two as is done in

the drift-diffusion approach. The effects of scattering in the HD formulation are described by macroscopic relaxation times that involve averages of the microscopic relaxation time over the distribution function. Although HD device modeling is limited by simplified representations of band structure and carrier transport assumptions, it provides an intuitive and computationally efficient means of analyzing physical trends.



(a)



(b)

Figure 4.4: (a) NMOSFET structure with L_g = 50 nm, V_T = 0.28 V, X_j = 30 nm, and t_{ox} = 2.5 nm. (b) Electron velocities along the channel obtained from HD and DD transport models for the L_g = 50 nm NMOSFET, HD has been calibrated according to the Monte-Carlo results in ref. [90].

To treat the non–local transport effects that occur in small devices, hydrodynamic (HD) device simulations [100][101] were used in this work. Drift-diffusion (DD) sets the maximum velocity of carriers to the saturated velocity (i.e. $\sim$1.0$\times$10^{7}cm/s), thus, velocity overshoot cannot be observed as in Figure 4.4(b). By contrast, the HD formulation reflects off-equilibrium behavior near the drain (i.e. strong velocity overshoot). As a result, higher I_{ds} is produced in the HD model in comparison to the DD model (Figure 4.5(a)). $v(0)$ versus V_{ds} near the source ($\sim$7 nm away from the source) is shown in Figure 4.5(b). In spite of the strong velocity overshoot near the drain, the carrier velocity near the source does not exceed the $\sim$1.0$\times$10^{7}cm/s, which corresponds to Lundstrom's prediction, where velocity overshoot affects the carrier velocity near the source, but it does not change the upper limit of $v(0)$, the thermal injection velocity (i.e. $v_T \sim$ 1.4$\times$10^7 cm/sec for L_g= 50 nm in ref. [95]).

Figure 4.5(c) shows a comparison of the analytical expression in Equation (4.3) with HD simulation, where C_{gs}, $v(0)$ and R_s are extracted from HD simulation results.

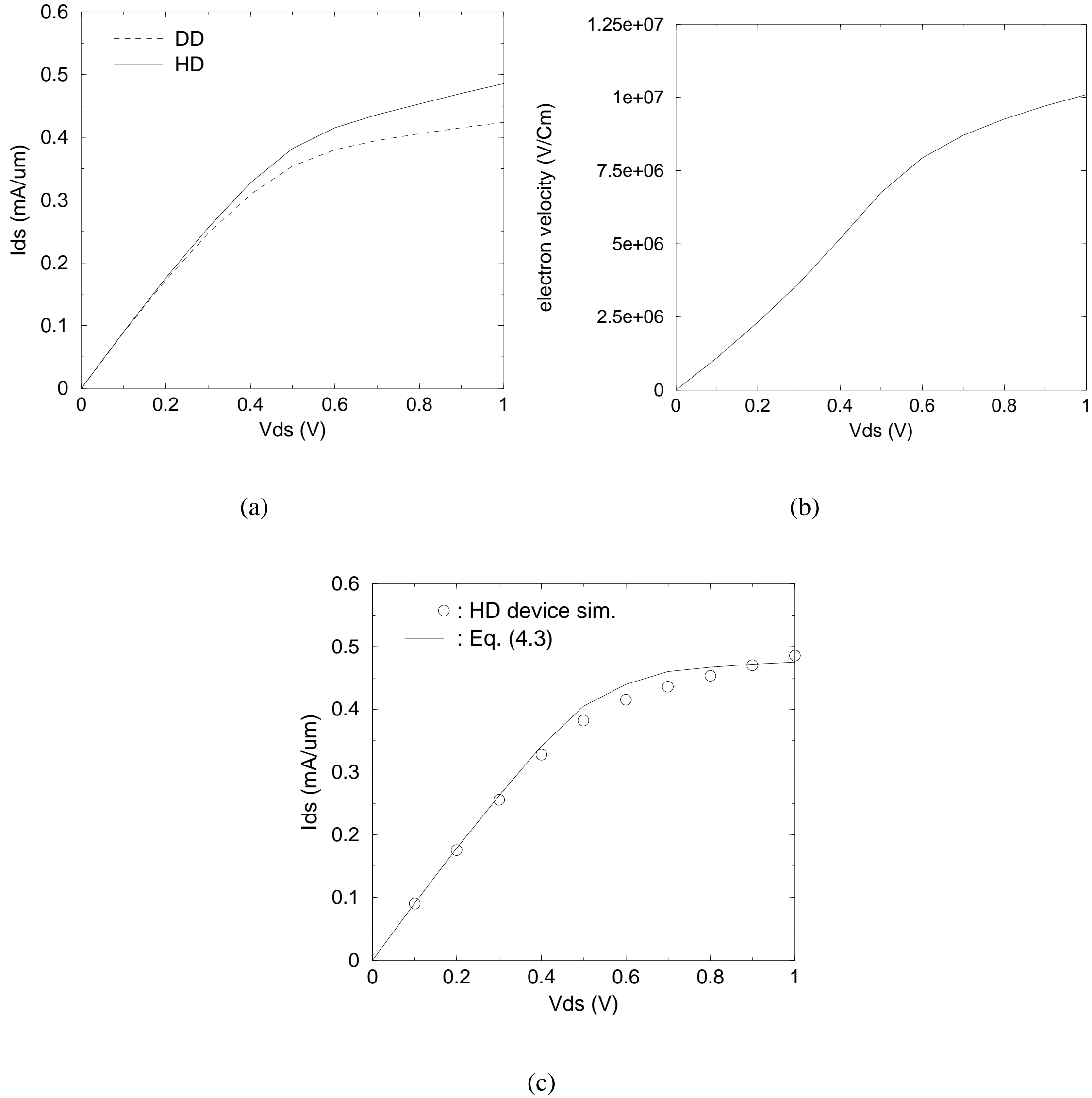


Figure 4.5: Simulated drain current and carrier velocity. (a) Simulated drain current between HD and DD models, V_{gs} = 1 V. (b) Electron velocity near the source extracted from hydrodynamic simulation result, V_{gs} = 1 V. (c) I_{ds} comparison between HD simulation and Equation (4.3), V_{gs} = 1 V.

4.4 Mobility and External Resistance Modeling

The channel mobility in MOSFETs is the most important physical quantity used to describe drain current. Moreover, it is an essential probe used to study the electrical properties of a two-dimensional carrier system. To extract an accurate low field mobility (μ^0) a local formulation which accounts for the Coulombic scattering, as well as phonon and surface roughness terms is used [78]:

$$1/\mu^0 = 1/\mu_{phonon} + 1/\mu_{surface} + 1/\mu_{coulomb} \tag{4.4}$$

where μ_{phonon}, $\mu_{surface}$, and $\mu_{coulomb}$ are respectively:

$\mu_{coulomb} = max(1.35 \times 10^{11} \frac{n^{1.5}}{{N_a}^{2.0}}, \frac{2.9 \times 10^8}{{N_a}^{2.0}});$

$\mu_{surface} = 8.3 \times 10^{14}/{E_{eff}}^{2.0};$

$\mu_{phonon} = 8.95 \times 10^5/E_{eff} + 3.23 \times 10^6 N_a^{0.03}/(T \cdot {E_{eff}}^{0.333});$

The quantity n is the local carrier concentration near the source and N_a is the background dopant density. The normal field, E_{eff}, is given as, $E_{eff} = (V_T - V_{FB} - 2\psi_B)/3t_{ox} + (V_{gs} - V_{FB})/6t_{ox}$ [27].

Due to the lateral gradient of source/drain doping in scaled devices under the gate overlap region, the portion of accumulation resistance (R_{acc}) relative to the total external resistance ($R_S = R_{acc} + R_{SDE} + R_{contact}$) becomes significant. R_{acc} is expressed as:

$$R_{acc} = \frac{L_{ov}}{X_j \mu_{acc} C_{ox} V_{gs}} \tag{4.5}$$

where μ_{acc} is the mobility under the gate overlap. Then, R_{acc} is added to the source/drain extension resistance (R_{SDE}) as source resistance.

4.5 Potential and E-field Modeling

In small devices, the step-like increase of lateral electric field in the channel region near the source is responsible for the nonlocal effects and contributes significantly in determining in drain current. Thus, drain current in the non-local channel region can be predicted by proper modeling the electric field (i.e. derivative of the potential) near the source [80].

Figure 4.6(a) shows the quasi-Fermi potential differences along the channel direction, comparing HD and DD simulation results. The potential at the drain edge, extracted from DD simulations is lower than that obtained using HD simulations due to the velocity saturation effect that occur in the DD model (i.e. $v_{sat} \sim 1 \times 10^7$ cm/sec). In addition, the total potential drop at the source and drain is equal to nearly half of the applied drain bias (V_{ds}) due to external source/drain resistance contributions. The electric-field near the source ($E_X(0^+)$), derivative of the potential, for the HD model is higher than that for the DD model, as shown in Figure 4.6(b). To model the off-equilibrium potential and electric field distributions along the channel (x-direction), the following expression is proposed, using the boundary conditions at the source and drain edges:

$$V(x) = \frac{V_{s's}}{1 - \frac{x}{L_g}(1 - \frac{\delta V_{s's}}{V_{d's}})} \tag{4.6}$$

where $V_{s's}$=$I_{ds}R_s$ and $V_{d's}$=V_{ds}-$I_{ds}R_s$ by assuming $R_S \simeq R_D$; δ is a fitting parameter in the range 0.8$\sim$1.0.

Hence,

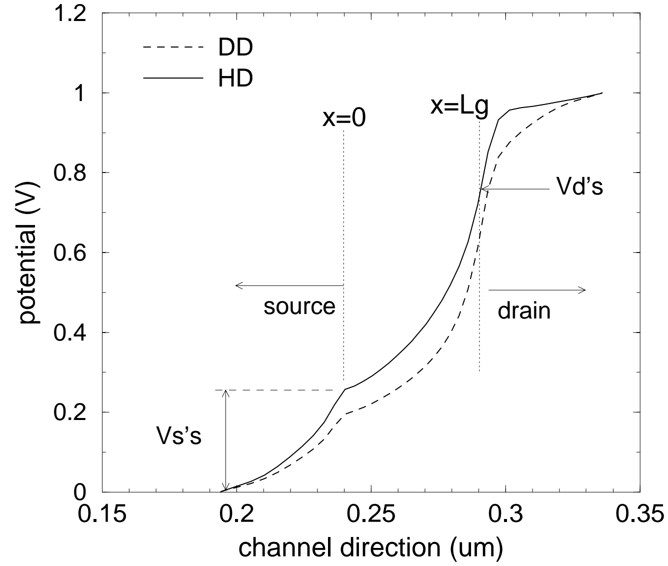
$$\frac{dV(x)}{dx} = \frac{V_{s's}(1 - \frac{\delta V_{s's}}{V_{d's}})}{L_g\Big[1 - \frac{x}{L_g}(1 - \frac{\delta V_{s's}}{V_{d's}})\Big]^2} \tag{4.7}$$

Thus $V(x)$ = $V_{s's}$ at x = 0 and $V(x)$ = ($V_{ds} - V_{s's}$)/δ at x = L_g. Figure 4.7(a) shows

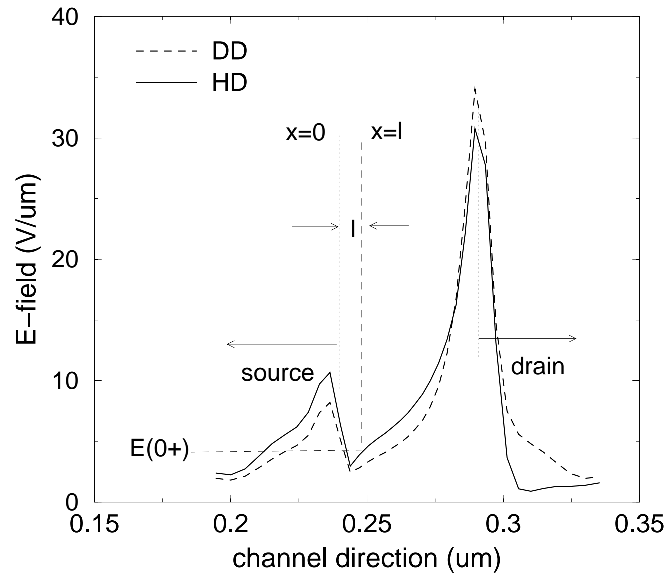
the potential distribution obtained from Equation (4.6) along with HD simulation results. The electric field distribution is calculated by taking the derivative of $V(x)$ with respect to channel position, as shown in Figure 4.7(b).

The critical distance (l) from the source in Figure 4.6(b) where the potential changes by k_BT/q is determined as $E_X(0^+)$ [87]:

$$l = (k_BT/q)/E_X(0^+) = (k_BT/q)/\frac{dV(x)|_{x=l}}{dx} \tag{4.8}$$

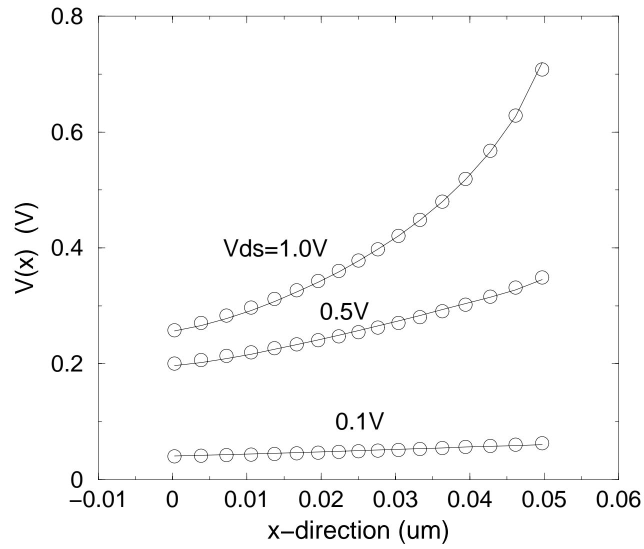


(a)

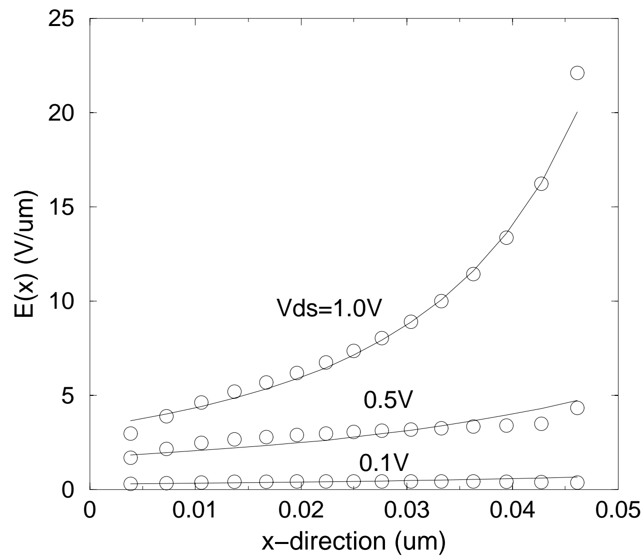


(b)

Figure 4.6: Potential and electric field distribution for HD and DD device simulation (a) Quasi-Fermi potential distribution along the channel for HD and DD models, L_g = 50 nm and $V_{gs} = V_{ds}$= 1 V. (b) Lateral electric field distribution along the channel between HD and DD models, and relationship l and $E_X(0^+)$, $V_{gs} = V_{ds}$= 1 V.



(a)



(b)

Figure 4.7: Potential and electric field distributions calculated by using Equation (4.6) and dV/dx. V_{gs} = 1 V, symbols represent HD simulation results. (a) potential (b) e-field.

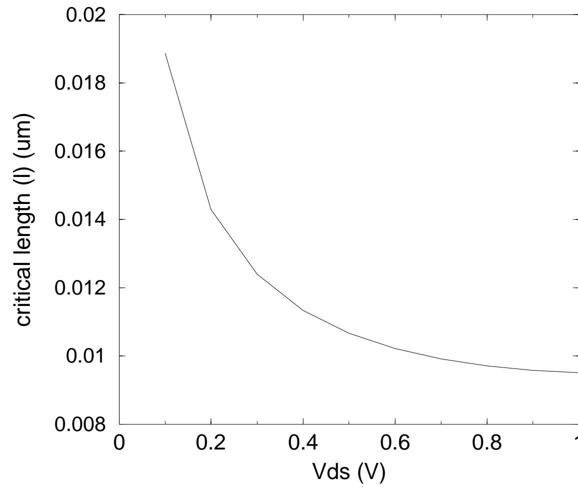


Figure 4.8: Critical distance (l) from the source calculated by using Equation (4.8).

Figure 4.8 illustrates the calculated l based on using Equation (4.8), which decreases as V_{ds} increases because $E_X(0^+)$ increases. Finally, I_{ds} is obtained by substituting the $v(0) = \mu^0 E_X(0^+)$ into Equation (4.3) and iterating to a solution, assuming that C_{gs} is independent of V_{ds} bias. Figure 4.9 shows the calculated electric field at l ($E_X(0^+)$) for different drain bias. Figure 4.10 shows the calculated I_{ds} based on the proposed model compared with HD simulation. The curve shows some small differences with simulation which can be attributed to the use of constant C_{gs} values with respect to V_{ds}. For more accurate prediction for drain current, accurate modeling of C_{gs} is required since it may vary as the ultra-short channel MOSFET enters into the saturation region.

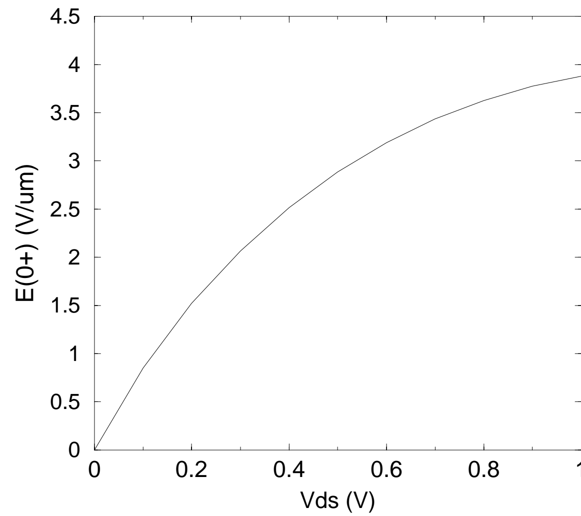


Figure 4.9: Calculated $E_X(0^+)$ for drain bias based on Eqs. (4.6) and (4.8), V_{gs}=1 V.

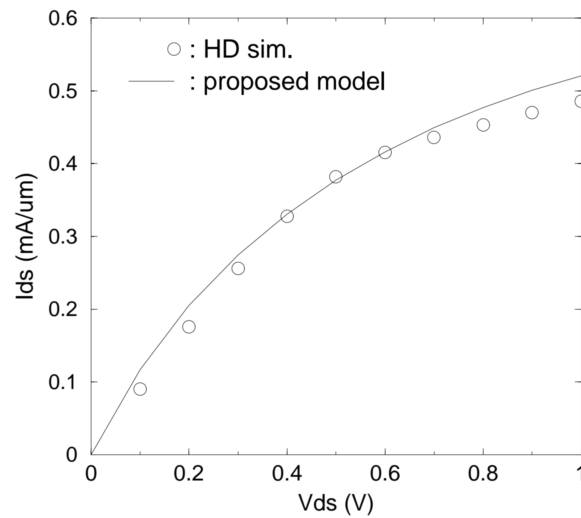


Figure 4.10: Drain current by using the proposed $v(0)$ model, symbols represent HD simulation, $V_{gs} = V_{ds} = 1$ V.

4.6 Summary

A simple MOSFET on-current model, taking into account the off-equilibrium transport is proposed based on use of Lundstrom's formulation, augmented with analytical calculations that are calibrated using HD simulations. The model shows good agreement with the hydrodynamic device simulation results for 50 nm MOSFETs.

Chapter 5

Conclusions

5.1 Introduction

There are a number of issues in scaling MOSFET devices, particularly for the sub-100 nm technology evolution. The most critical issue is the gate dielectric, because very thin gate oxides are required for sub-100 nm generations. For such thin oxides, gate leakage current due to direct tunneling becomes unacceptably large. In order to decrease the leakage current due to tunneling, the physical thickness of the dielectric must increase, while the equivalent oxide thickness must continue to be reduced. Use of alternate gate materials with dielectric constant higher than that of silicon dioxide is the leading projected solution to reduce the gate leakage current to more tolerable levels. Therefore, modeling of the thin oxide related issues, such as gate direct tunneling current, gate capacitance, and capacitance reconstruction are crucial for further gate oxide scaling. On the other hand, ultra-shallow source/drain extensions are also required, and technologies for ultra-low energy implant ($<$ 0.5 keV) and laser annealed dopant activation are projected solutions needed to achieve the ultra-shallow junctions. However, the requirements of low external resistance and shallow

junction depth will become increasingly difficult without technology innovations, because of the trade-off between good short-channel characteristics and high-drive current. Moreover, conventional low-energy ion implantation tends to result in high sheet resistance due to the degradation of carrier activation efficiency. In addition, to maintain the on-current and deal with short-channel effects, advanced channel engineering techniques such as optimized halo implants and the use of high mobility silicon/germanium epitaxial layers are envisioned.

Modeling and simulation can play a critical role in overcoming difficult challenges for future technology; in this thesis, device physics issues on nanoscale MOSFETs are discussed and new modeling approaches are proposed in the scope of gate, source/drain, and channel regions.

5.2 Gate Modeling

Accurate characterization and modeling of ultra-thin oxides in the direct tunneling regime is essential as the MOSFET gate oxide thickness rapidly approaches the direct tunneling limit that ultimately leads to an intolerable increase in leakage current and standby power. MOS C–V characteristics of gate oxide thickness less than 2.0 nm are modeled with an empirical, hybrid QM corrections and implemented in a 2D device simulator. The C-V characteristics of ultra-thin MOS capacitors ranging between 1.3–1.8 nm, which show a sharp decrease in capacitance for gate oxides below 2.0 nm, are modeled using a distributed RC network that includes a QM capacitance model. A numerically calculated gate tunneling current is used to set the shunt conductance in the gate. This combined numerical/lumped model maintains excellent (and physical) accuracy compared to experiments. It combines gate tunneling current in both the accumulation and inversion regions, which is calculated using a Green's function solver. A reconstruction technique to correct for anomalous $C-V$

behavior has also been proposed using the distributed RC network to account for both the QM and distributed RC effects in MOSFETs. The intrinsic gate capacitance is simultaneously extracted from the anomalous measured $C-V$ curves. The reconstructed $C-V$ data is comparable to the theoretical QM calculations for very-thin nitride/oxide gate dielectric ($t_{ox,eq-qm} \sim$ 1.4 nm) MOSFETs.

A direct tunneling model for circuit simulation is developed that incorporates an explicit surface potential model and quantum-mechanical corrections. Simulated gate currents from this model demonstrate good agreement with the results from a numerical solver and measured data for gate oxides with thickness ranging between 1.3 – 1.8 nm. The impact of gate direct tunneling current on VLSI circuits has been studied for $t_{ox} <$ 1.8 nm, L_g = 50 nm MOS circuits based on circuit simulations. The gate current effects on t_{ox} = 1.5 nm MOS circuits are shown to be negligible for low V_{dd} static-logic circuits, while for analog and dynamic-logic circuits they are shown to be vulnerable to the off-leakage current due to direct tunneling, even for low V_{dd} operation.

For scaling of sub-100 nm CMOS, a performance-based delay model that takes into account the severe effects of gate leakage has been introduced. Using the delay model, technology trade-offs for scaling CMOS below 100 nm are quantized; optimization of circuit speed as a function of technology parameters is achievable based on the consideration of intrinsic physical effects as well as inclusion of parasitic components.

5.3 Source/Drain Modeling

For MOSFET scaling, the concurrent suppression of the short-channel effects and the improvement of drain current and transconductance are necessary conditions. The ITRS roadmap predicts that the SDE depths necessary to suppress short-channel effects will be

as low as 10 nm for a channel length of 50 nm. Due to the high external resistance, further improvements in driving current for sub-100 nm device regime is extremely difficult, as the channel resistance becomes comparable to the source and drain external resistance. Since further down-scaling without improvements in MOSFET performance is meaningless, reduction of the external resistance components and accurate prediction their values for shallow SDE MOSFETs is of great importance. To achieve an accurate calculation for external source/drain resistance, device simulation with an unified mobility model has been applied to 100 nm devices. The trade-offs between the external resistance and the short channel effects in sub-100 nm nMOSFETs are studied based on the accurate resistance calculation. As a result, R_{acc} is shown to be highly dependent on the junction depth and overlap length, which should be minimized (or optimized) for improvement in current driving capability.

The external resistance for laser thermal processed 70 nm NMOSFETs have been optimized using process and device simulations. The 70 nm NMOS transistor of AMD technology shows the highest I_{on} when the L_{ov} is 5–10 nm for a X_j is 20–25 nm and a spacer length of 60–70 nm.

5.4 Channel Modeling

Limitations of gate length reduction in terms of the suppression of short channel effects are estimated to be around 25 nm [1]. In the ultra-short-channel device regime, the drift-diffusion model breaks down, especially where rapid spatial variations of potential are observed. In such cases, the scattering events are no longer localized, and carriers may acquire excess thermal energies near the drain. These carriers are not in thermal equilibrium with the silicon lattice which is referred to hot carrier effects. Under these circumstances, it is possible for the carrier velocity to exceed the saturation velocity, which is so called

velocity overshoot. Even though Monte Carlo device simulation has been established as a powerful tool for the investigation of ballistic effects in deep submicron MOSFETs, its excessive computation time is a burden for practical design applications.

A simple on-current model considering the off-equilibrium and external resistance effects is presented. To model the off-equilibrium potential and electric distributions along the channel region, a new expression is developed using simulation-based boundary conditions at the source and drain edges. This approach provides improved physical insight of carrier transport in nanoscale devices and shows good agreement with the hydrodynamic device simulation results for L_g = 50 nm MOSFETs.

5.5 Recommendations for Future Work

There are various modeling issues that await exploration in nanoscale MOSFETs. First, numerical device simulation poses several ongoing challenges. As transistors have reached dimensions approaching the mean distance between collisions, it is strongly desirable to fully exploit a multi-dimensional Poisson-Schrödinger solver that treats electrons as waves traveling across the device. Modeling and characterization of high gate field effects – gate direct tunneling (DT), polysilicon depletion, and gate-induced drain leakage (GIDL) – are crucial. The gate DT current effects pose significant effort in understanding the physics in the context of practical 2-dimensional devices. Hence, further numerical analysis is needed, based on Poisson-Schrödinger solutions in order to calculate tunneling coefficients. In addition to ongoing use of NEMO, Non-Equilibrium Greens Function calculations will support the development and demonstration of a model suitable for implementation in a conventional device simulator, based on the Density Gradient (DG) formulation. Modeling of 2-dimensional dopant distributions including dopant diffusion mechanisms inside

polysilicon gate and electrical simulation depending on 2/3D gate geometry are key considerations in the modeling process for polysilicon depletion effects. GIDL is a primary concern in nanoscale MOSFETS along with the gate edge DT current in terms of off-state leakage current. Since GIDL is determined by both vertical and lateral electric fields along the gate and drain overlap region, the relationship between GIDL and DT currents must be understood. Along with the existing commercial device simulators that use band-to-band tunneling formulations, it is necessary to determine appropriate models for trap-assistant tunneling and its dependence on processing, bias and ambient conditions. Analog and RF design typically requires an accurate description of not only currents and capacitance, but also of the small-signal behavior. Impact of all the above effects on small-signal behavior is no longer negligible in analog and RF application, especially for high frequency noise performance.

The demand for new material and technologies becomes increasing in the nanometer CMOS regime; understanding of mobility enhancement in the germanium strained silicon, reliability and mobility degradation in the high-κ gate dielectrics, and various device parameters in the metal electrodes are required. Scaling CMOS towards the 25 nm channel length generation requires innovative device structures to circumvent barriers due to the fundamental physics in the conventional MOSFETs. These are SOI, back-gate FET, double-gate FET and FinFET. Fundamental issues for these structures such as the physics of carrier transport in very thin silicon channel must be further understood.

Bibliography

[1] H. Iwai, “Current Status and Future of Advanced CMOS Technologies: Digital and Analog Aspects,” *International Conference on Advanced Semiconductor Devices and Microsystems*, pp. 1–10, 1998.

[2] Y. Taur, D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S. H. Lo, G. Sai-Halasz, R. Viswanathan, and *et al.*, “CMOS scaling into nanometer regime,” *Proceedings of IEEE*, vol. 85, pp. 486–504, Apr. 1997.

[3] “International Technology Roadmap for Semiconductors 2000 Edition,” *http://public.itrs.net/, Semiconductor Industry Assoc.,* Austin, TX, 2000.

[4] E. J. Lerner, “The End of the Road for Moore’s Law,” *Journal of IBM research*, no. 4, pp. 6–11, 1999.

[5] P. A. Packan, “Pushing the Limits,” *Science*, vol. 285, pp. 2079–2080, Sep. 1999.

[6] C. Fiegna and A. Abramo “Analysis of Quantum Effects in Nonuniformly Doped MOS Structures,” *IEEE Trans. Electron Devices*, vol. 45, pp. 877–880, Apr. 1998.

[7] S.-H. Lo, D. A. Buchanan, and Y. Taur, “Modeling and Characterization of Quantization, Polysilicon Depletion, and Direct Tunneling Effects in MOSFETs with Ultrathin Oxides,” *IBM J. Res. Develop*, vol. 43, pp. 327–337, May 1999.

[8] S. Thompson, P. Packan, M. Alavi, I. Post, S. Tyagi, S. Ahmed, S. Yang, and M. Bohr, "Source/Drain Extension Scaling for 0.1μm and Below Channel Length MOSFETs," *IEEE Symp. on VLSI Technology Digest*, pp. 132–133, 1998.

[9] H. Iwai, "CMOS Technology – Year 2010 and Beyond," *IEEE J. of Solid-State Circuits*, vol. 34, pp. 357–366, Mar. 1999.

[10] S. Deleonibus, C. Caillat, G. Guegan, M. Heitzmann, M. E. Nier, S. Tedesco, B. Dal'zotto, F. Martin, P. Mur, A. M. Papon, G. Lecarval, S. Biswas, and D. Souil, "A 20-nm Physical Gate Length NMOSFET Featuring 1.2 nm Gate Oxide, Shallow Implanted Source and Drain and BF_2 Pockets," *IEEE Electron Device Lett.*, vol. 21, pp. 173–175, Apr. 2000.

[11] H. Kawaura, T. Sakamoto, T. Baba, Y. Ochiai, J. Fujita, and J. Sone, "Transistor Characteristics of 14-nm-Gate-length EJ-MOSFET's," *IEEE Trans. Electron Devices*, vol. 47, pp. 856–860, Apr. 2000.

[12] P. Keys and C. Rafferty, "Serise Resistance Limits for 0.05μm MOSFETs," *3rd NASA Workshop on Device Modeling,* Aug. 1999.

[13] M. T. Bohr and Y. A. El-Mansy, "Technology for Advanced High-Performance Microprocessors," *IEEE Trans. Electron Devices*, vol. 45, pp. 620–633, Mar. 1998.

[14] G. Timp and $et\ al.$, "The Ballistic Nano-transistor," in *IEDM Tech. Dig.,* pp. 55–58, 1999.

[15] H-S. P. Wong, D. J. Frank, P. M. Solomon, C. H. J. Wann, and J. F. Welser, "Nanoscale CMOS," *Proceedings of the IEEE,* vol. 87, pp. 537–570, Apr. 1999.

[16] A. Hori and B. Mizuno, "CMOS Device Technology toward 50 nm Region," in *IEDM Tech. Dig.,* pp. 641–644, 1999.

[17] E. Morifuji and $et\ al.$, "An 80 nm Dual-gate CMOS with Shallow Extensions Formed after Activation Annealing and SALICIDE," in *IEDM Tech. Dig.,* pp. 649–652, 1999.

[18] S. Song and $et\ al.$, "High Performance Transistors with State-of-the-Art CMOS Technologies," in *IEDM Tech. Dig.,* pp. 427–430, 1999.

[19] B. Yu, H. Wang, O. Milic, Q. Xiang, W. Wang, J. X. An, and M-R. Lin, "50 Nm Gate-Length CMOS Transistor with Super-Halo: Design, Process, and Reliability," in *IEDM Tech. Dig.,* pp. 653–656, 1999.

[20] P. A. Stolk, F. P. Widdershoven, and D. B. Klaassen, "Modeling Statistical Dopant Fluctuation in MOS Transistors," *IEEE Trans. Electron Devices*, vol. 45, pp. 1960–1971, Sep. 1998.

[21] S. Mudanai, Y. Fan, Q. Quyang, A. Tasch, and S. K. Banerjee, "Modeling of Direct Tunneling Current Through Gate Dielectric Stack," *IEEE Trans. Electron Devices*, vol. 47, pp. 1851–1859, Oct. 2000.

[22] H. F. Luan and $et\ al.$, "High Quality Ta_2O_3 Gate Dielectrics with $T_{ox,eq} < 10$ Å," in *IEDM Tech. Dig.,* pp. 141–144, 2000.

[23] R. W. Dutton and Z. Yu, *Technology CAD: computer simulation of IC processes and devices*, Kluwer Academic, Boston, 1995.

[24] K. J. Yang and C. Hu, "MOS Capacitance Measurements for High-Leakage Thin Dielectrics," *IEEE Trans. Electron Devices*, vol. 46, pp. 1500–1501, July 1999.

[25] C.-H. Choi, J.-S. Goo, T.-Y. Oh, Z. Yu, R. W. Dutton, A. Bayoumi, M. Cao, P. V. Voorde, D. Vook, and C. H. Diaz "MOS C–V Characterization of Ultra-Thin Gate Oxide Thickness (1.3–1.8 nm)," *IEEE Electron Device Lett.*, vol. 20, pp. 292–294, June 1999.

[26] K. Ahmed, E. Ibok, G. C. F. Yeap, Q. Xiang, B. Ogle, J. J. Wortman, and J. R. Hauser, “Impact of Tunnel Currents and Channel Resistance on the Characterization of Channel Inversion Layer Charge and Polysilicon-Gate Depletion of Sub-20-$\AA$ Gate Oxide MOSFET’s,” *IEEE Trans. Electron Devices*, vol. 46, pp. 1650–1655, Aug. 1999.

[27] Y. Taur and T. Ning, *Fundamental of modern VLSI device.* Cambridge University Press, 1998.

[28] R. Muller and T. Kamins, *Device Electronics for Integrated Circuits*, Wiley & Sons, Inc., 1986.

[29] N. D. Arora, *MOSFET Model for VLSI Circuit Simulation.* New York: Springer-Verlag, 1993.

[30] *MEDICI : Two-Dimensional Semiconductor Device Simulation.* Technology Modeling Associates, p. 2-27, 1998.

[31] A. P. Gnadinger and H. E. Tally, “Quantum Mechanical Calculation of the Carrier Distribution and The Thickness of The Inversion Layer of A MOSFET,” *Solid State Elec.*, vol. 13, pp. 1301–1309, 1970.

[32] S. Takagi, M. Takayanagi, and A. Toriumi, “Impact of Electron and Hole Inversion-Layer Capacitance on Low Voltage Operation of Scaled n- and p-MOSFETS’s,” *IEEE Trans. Electron Devices*, vol. 47, pp. 999–1005, May 2000.

[33] E. M. Vogel, W. K. Henson, C. A. Richter, and J. S. Suehle, “Limitations of Conductance to the Measurement of the Interface State Density of MOS Capacitors with Tunneling Gate Dielectrics,” *IEEE Trans. Electron Devices*, vol. 47, pp. 601–608, Mar. 2000.

[34] W. K. Henson, K. Z. Ahmed, E. M. Vogel, J. R. Hauser, J. J. Wortman, R. D. Venables, M. Xu, and D. Venables, "Estimating Oxide Thickness of Tunnel Oxides Down to 1.4 nm Using Conventional Capacitance-Voltage Measurements on MOS Capacitors," *IEEE Electron Device Lett.*, vol. 20, pp. 179–181, Apr. 1999.

[35] M. J. van Dort, P. H. Woerlee, and A. J. Walker, "A simple model for quantisation effects in heavily-doped silicon MOSFETs at inversion conditions," *Solid-State Elec.*, vol. 37, p. 411, 1994.

[36] W. Hänsch, T. Vogelsang, R. Kircher, and M. Orlowski, "Carrier transport near the Si/SiO_2 interface of a MOSFET," *Solid-State Elec.*, vol. 32, p. 839, 1989.

[37] P. Vande Voorde, P. B. Griffin, Z. Yu, S.-Y. Oh, and R. W. Dutton, "Accurate doping profile determination using TED/QM models extensible to sub-quarter micron nMOSFETs," in *IEDM Tech. Dig.*, pp. 811–814, 1996.

[38] C. Bowen, C. Fernando, G. Klimeck, A. Chatterjee, D. Blanks, R. Lake, J. Hu, J. Davis, M. Kulkarni, S. Hattangady and I-H Chen, "Physical oxide thickness extraction and verification using quantum mechanical simulation," in *IEDM Tech. Dig.*, pp. 869–872, 1997.

[39] S. Nagano, and M. Tsukiji, E. Hasegawa, and A. Ishitani, "Mechanism of leakage current through the nano-scale SiO_2 layer," *J. of Applied Physics*, vol. 75, p. 3530, 1994.

[40] C. S. Rafferty, B. Biegel, Z. Yu, M.G. Ancona, J. Bude, and R. W. Dutton, "Multi-Dimensional Quantum Effect Simulation Using a Density-Gradient Model and Script-Level Programming Techniques," in *Proc. Simulation Semiconductor Processes and Devices (SISPAD)*, pp. 137–140, 1999.

[41] H. S. Momose, M. Ono, T. Yoshitomo, T. Ohguro, S. Nakamura, M. Saito, and H. Iwai, "1.5 nm Direct-Tunneling Gate Oxide Si MOSFET's," *IEEE Trans. Electron Devices*, vol. 43, pp. 1233–1242, Aug. 1996.

[42] C.-H. Choi, J.-S. Goo, T.-Y. Oh, Z. Yu, R. W. Dutton, A. Bayoumi, M. Cao, P. V. Voorde, and D. Vook, "C–V and Gate Tunneling Current Characterization of Ultra-Thin Gate Oxide MOS (t_{ox}=1.3–1.8 nm)," in *Proc. Symp. VLSI Technol.*, pp. 151–152, 1999.

[43] H. S. Momose, H. Kimijima, S. Ishizuka, Y. Miyahara, T. Ohguro, T. Yoshitomi, E. Morifuji, and $et\ al.$, "A Study of Flicker Noise in N^- and P^- MOSFETs with Ultra-thin Gate Oxide in the Direct-tunneling Regime," in *IEDM Tech. Dig.,* pp. 923–926, 1998.

[44] Y. Wu and G. Lucovsky, "Ultrathin Nitride/Oxide (N/O) Gate Dielectrics for p^+-Polysilicon Gated PMOSFET's Prepared by a Combined Remote Plasma Enhanced CVD/Thermal Oxidation Process," *IEEE Electron Device Lett.*, vol. 19, pp. 367–369, Oct. 1998.

[45] Y. Wu, Y-M. Lee, and G. Lucovsky, "1.6 nm Oxide Equivalent Gate Dielectrics Using Nitride/Oxide (N/O) Composites Prepared by RPECVD/Oxidation Process," *IEEE Electron Device Lett.*, vol. 21, pp. 116–118, Mar. 2000.

[46] C.-H. Choi, Y. Wu, J.-S. Goo, Z. Yu, and R. W. Dutton, "Capacitance Reconstruction from Measured C-V in High Leakage, Nitride/Oxide MOS," *IEEE Trans. Electron Devices,* vol. 47, pp. 1843–1850, Oct. 2000.

[47] K. Kano, *Semiconductor Devices.* Prentice-Hall, 1998.

[48] *HSPICE User's Manual,* AVANT! Corp., 1998.

[49] K. Doganis, D. L. Scharfetter, "General Optimization and Extraction of IC Device Model Parameters," *IEEE Trans. CAD*, vol. 30, pp. 1219–1228, Sep. 1983.

[50] R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. Leblanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE J. Soild-State Circuits*, vol. SC-9, no. 5, p. 256, 1974.

[51] T. Ohguro, S. Nakamura, M. Saito, and *et al.*, "Ultra-shallow junction and salicide techniques for advanced CMOS devices," *Electrochemical Society Proceedings of the IEEE*, vol. 3, p. 275, 1997.

[52] K. Chen, H. C. Wann, P. K. Ko, and C. Hu, "The impact of device scaling and power supply change on CMOS gate performance," *IEEE Electron Device Lett.*, vol. 17, pp. 202–204, May 1996.

[53] M. Lundstrom, "Scattering theory of the short channel MOSFET," in *IEDM Tech. Dig.*, pp. 387–390, 1997.

[54] Z. Yu, *Private communication,* 1997.

[55] K. Chen, C. Hu, P. Fang, M. R. Lin, and D. L. Wollesen, "Predicting CMOS Speed with Gate Oxide and Voltage Scaling and Interconnect Loading Effects," *IEEE Trans. Electron Devices,* vol. 44, pp. 1951–1957, Nov. 1997.

[56] N. Shigyo, T. Shimane, M. Suda, T. Enda, and S. Fukuda, "Verification of Saturation Velocity Lowering in MOSFET's Inversion Layer," *IEEE Tran. Electron Devices*, vol. 45, pp. 460–464, Feb. 1998.

[57] G. J. Hu, C. Chang, and Y. Chia, "Gate-Voltage-Dependent Effective Channel Length and Series Resistance of LDD MOSFET's," *IEEE Tran. Electron Devices*, ED-34, p. 2469, Dec. 1987.

[58] S. H. Lo, D. A. Buchanan, Y. Taur, L. K. Han, and E. Wu, "Modeling and Characterization of n^+ and p^+ Polysilicon-Gated Ultra Thin Oxides (21–26 Å)," in *Proc. Symp. VLSI Technol.*, pp. 149–150, 1997.

[59] C.-H. Choi, K.-H. Oh, J.-S. Goo, Z. Yu, and R. W. Dutton, "Direct Tunneling Current Model for Circuit Simulation," in *IEDM Tech. Dig.*, pp. 735–738, 1999.

[60] K. F. Schuegraf and C. Hu, "Hole Injection SiO_2 Breakdown Model for Very Low Voltage Lifetime Extrapolation," *IEEE Trans. Electron Devices*, vol. 41, pp. 761–767, May 1994.

[61] R. V. Langevelde, "A Compact MOSFET Model for Distortion Analysis in Analog Circuit Design," *PhD dissertation, Eindhoven Univ. of Tech.*, 1998.

[62] M. Depas, B. Vermeire, P. W. Mertens, R. L. Van Meirhaeghe, and M. M. Heyns, "Determination of Tunneling Parameters in Ultra-Thin Oxide Layer Poly-Si/SiO_2/Si Structures," *Solid-State Electron.*, vol. 38, pp. 1465–1471, 1995.

[63] H. Iwai and H. S. Momose, "Ultra-Thin Gate Oxides – Performance and Reliability," in *IEDM Tech. Dig.,* pp. 163–166, 1998.

[64] C. Hu, "Gate Oxide Scaling Limits and Projection," in *IEDM Tech. Dig.,* pp. 319–322, 1996.

[65] P. J. Wright and K. Saraswat, "Thickness Limitations of SiO_2 Gate Dielectrics for MOS ULSI," *IEEE Trans. Electron Devices*, vol. 37, pp. 1884–1892, Aug. 1990.

[66] C.-H. Choi, K.-Y. Nam, Z. Yu, and R. W. Dutton, "Impact of Gate Direct Tunneling Current on Circuit Performance: A Simulation Study," *IEEE Trans. Electron Devices,* vol. 48, pp. 2823–2829, Dec. 2001.

[67] W. K. Henson, N. Yang, E. M. Vogel, J. Wortman, and A. Naem, “Analysis of Leakage Currents and Impact on Off-State Power Consumption for CMOS Technology in 100-nm Regime,” *IEEE Trans. Electron Devices*, vol. 47, pp. 1393–1400, July 2000.

[68] K. N. Yang, H. T. Huang, M. J. Chen, Y. M. Lin, M. C. Yu, S. M. Jang, C. H. Yu, and M. S. Liang, “Edge Hole Direct Tunneling in Off-State Ultrathin Gate Oxide p-Channel MOSFETs,” in *IEDM Tech. Dig.*, pp. 679–682, 2000.

[69] N. Yang, W. K. Henson, and J. Wortman, “A Comparative Study of Gate Direct Tunneling and Drain Leakage Currents in N-MOSFET’s with Sub-2 nm Gate Oxides,” *IEEE Trans. Electron Devices*, vol. 47, pp. 1636–1644, Aug. 2000.

[70] Z. Yu, R. W. Dutton, and R. A. Kiehl, “Circuit/Device Modeling at Quantum Level,” *IEEE Trans. Electron Devices*, vol. 47, pp. 1819–1825, Oct. 2000.

[71] A. Shanware, J. P. Shiely, and H. Z. Massoud, “Extraction of the Gate Oxide Thickness of N- and P-Channel MOSFETs Below 20Å from Substrate Current Resulting from Valence-Band Electron Tunneling,” in *IEDM Tech. Dig.,* pp. 815–818, 1999.

[72] M. Rodder, S. Hattangady, N. Yu, W. Shiau, P. Nicollian, T. Laaksonen, M. Mehrotra, C. Lee, and S. Aur, “A 1.2V, 0.1μm Gate Length CMOS Technology : Design and Process Issues,” in *IEDM Tech. Dig.,* pp. 623–626, 1998.

[73] T. Ghani, K. Mistry, P. Packan, S. Thompson, M. Stettler, S. Tyagi, and M. Bohr, “Scaling Challenges and Device Design Requirement for High Performance Sub-50 nm Gate Length Planar CMOS Transistors,” in *Proc. Symp. VLSI Technol.*, pp. 174–175, 1998.

[74] Y. C. Yeo, Q. Lu, W. C. Lee, T-J. King, C. Hu, A. Wang, X. Guo, and T. P. Mam, "Direct Tunneling Gate Leakage Current in Transistors with Ultrathin Silicon Nitride Gate Dielectric," *IEEE Electron Device Lett.*, vol. 21, pp. 540–542, Nov. 2000.

[75] S. M. Kang and Y. Leblebici, *CMOS Digital Integrated Circuits.* The McGraw-Hill Companies, Inc., 1996.

[76] C.-H. Choi, J.-S. Goo, Z. Yu, and R. W. Dutton, "Shallow Source/Drain Extension Effects on External Resistance in Sub-0.1 μm MOSFET's," *IEEE Trans. Electron Devices*, vol. 47, pp. 655–658, Mar. 2000.

[77] S. Biesemans, M. Hendricks, S. Kubicek, and K. De Meyer, "Practical Accuracy Analysis of Some Existing Effective Channel Length and Series Resistance Extraction Methods for MOSFET's," *IEEE Trans. Electron Devices*, vol. 45, pp. 1310–1316, June 1998.

[78] S. A. Mujtaba, S. Takagi, and R. W. Dutton, "Accurate Modeling of Coulombic Scattering, and Its Impact on Scaled MOSFETs," in *Proc. Symp. VLSI Technol.*, pp. 99–100, 1995.

[79] K. Ng and W. T. Lynch, "Analysis of the Gate-Voltage Dependent Series Resistance of MOSFETs," *IEEE Trans. Electron Devices,* vol. 33, p. 965, 1986.

[80] J.-H. Song, Y.-J. Park, and H.-S. Min, "Drain Current Enhancement Due to Velocity Overshoot Effects and Its Analytic Modeling," *IEEE Trans. Electron Devices,* vol. 43, pp. 1870–1875, Nov. 1996.

[81] S. A. Mujtaba, "Advanced Mobility Models for Design and Simulation of Deep Submicrometer MOSFETs," *Ph.D dissertation, Stanford Univ.*, 1995.

[82] S. Mudanai, G. L. Chindalore, W. K. Shin, H. Wang, Q. Ouyang, A. F. Tasch, C. M. Maziar, and S. K. Banerjee, "Models for Electron and Hole Mobilities in MOS Accumulation Layers," *IEEE Trans. Electron Devices*, vol. 46, pp. 1749–1759, Aug. 1999.

[83] Y. Taur, C. H. Wann, and D. J. Frank, "25 nm CMOS Design Considerations," in *IEDM Tech. Dig.*, pp. 789–792, 1998.

[84] K. Tsuji, K. Takeuchi, and T. Mogami, "High Performance 50-nm Physical Gate Length pMOSFETs by Using Low Temperature Activation by Re-Crystallization Scheme," in *Proc. Symp. VLSI Technol.*, pp. 9–10, 1999.

[85] B. Yu, Y. Wang, H. Wang, Q. Xiang, C. Riccobene, S. Talwar, and M. R. Lin, "70 nm MOSFET with Ultra-Shallow, Abrupt, and Super-Doped S/D Extension Implemented by Laser Thermal Process (LTP)," in *IEDM Tech. Dig.*, pp. 509–512, 1999.

[86] S. Talwar, G. Verma, K. Weiner, and C. Gelatos, "Laser thermal processing for shallow junction and silicide formation," *proceeding SPIE*, pp. 74–81, 1998.

[87] M. Lundstrom, "Elementary Scattering Theory of the MOSFET," *IEEE Electron Device Lett.*, vol. 18, pp. 361–363, July 1997.

[88] F. Assad, Z. Ren, S. Datta, M. Lundstrom, and P. Bendix, "Performance Limits of Silicon MOSFET's," in *IEDM Tech. Dig.*, pp. 547–550, 1999.

[89] M. Lundstrom, *Fundamentals of Carrier Transport,* The University Press, Cambridge, 2000.

[90] M. Pinto, E. Sangiorgi, and J. Bude, "Silicon MOS Transconductance Scaling into the Overshoot Region," *IEEE Electron Device Lett.*, vol. 14, pp. 375–378, Aug. 1993.

[91] J. D. Bude, "MOSFET Modeling Into the Ballistic Regime," in *Proc. Simulation Semiconductor Processes and Devices (SISPAD)*, pp. 23–26, 2000.

[92] F. M. Bufler, A. Schenk, and W. Fichtner, "Efficient Monte Carlo Device Simulation with Automatic Error Control," in *Proc. Simulation Semiconductor Processes and Devices (SISPAD)*, pp. 27–30, 2000.

[93] "DAMOCRES: Monte Carlo Simulation of Semiconductor Devices," *http://www.research.ibm.com/0.1um/laux/dam.html*.

[94] M. Lundstrom, Z. Ren, and S. Datta, "Essential Physics of Carrier Transport in Nanoscale MOSFETs," in *Proc. Simulation Semiconductor Processes and Devices (SISPAD)*, pp. 1–5, 2000.

[95] F. Assad, Z. Ren, D. Vasileska, S. Datta, and M. Lundstrom, "On the Performance Limits for Si MOSFET's: A Theoretical Study," *IEEE Trans. Electron Devices*, vol. 47, pp. 232–240, Jan. 2000.

[96] T. Mizuno, "New Channel Engineering for Sub-100 nm MOS Devices Considering Both Carrier Overshoot and Statistical Performance Fluctuations," *IEEE Trans. Electron Devices*, vol. 47, pp. 756–761, Apr. 2000.

[97] T. Mizuno and R. Ohba, "Experimental Study of Carrier Velocity Overshoot in Sub-0.1μm Devices," in *IEDM Tech. Dig.*, pp. 109–112, 1996.

[98] S. E. Laux and M. V. Fischetti, "Monte Carlo Study of Velocity Overshoot in Switching a 0.1-Micron CMOS Inverter," in *IEDM Tech. Dig.*, pp. 877–880, 1997.

[99] D. J. Frank, S. E. Laux, and M. V. Fischetti, "Monte Carlo simulation of a 30 nm Dual-Gate MOSFET: How Short Can Si Go ?" in *IEDM Tech. Dig.*, pp. 553–556, 1992.

[100] M. Ieong and T.-W. Tang, “Influence of Hydrodynamic Models on the Prediction of Submicrometer Device Characteristics,” *IEEE Trans. Electron Devices*, vol. 44, pp. 2242–2251, Dec. 1997.

[101] W.-S. Choi, F. Assaderaghi, Y.-J. Park, H.-S. Min, C. Hu, and R. W. Dutton, *IEEE Electron Device Lett.*, vol. 16., pp. 333–335, July 1995.