Comprehensive Inverse Modeling for the Study of Carrier Transport Models in Sub-50 nm MOSFETs

by

Ihsan J. Djomehri

B. S. in Electrical Engineering and C. S., University of California at Berkeley, 1997
A. B. in Physics, University of California at Berkeley, 1997
S. M. in Electrical Engineering, Massachusetts Institute of Technology, 1998

Submitted to the Department of Electcrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

**Doctor of Philosophy
in Electrical Engineering and Computer Science**

at the

Massachusetts Institute of Technology

September 2002

Author ....
Department of Electrical Engineering and Computer Science
August 14, 2002

Certified by ....
Dimitri A. Antoniadis
Professor of Electrical Engineering
Thesis Supervisor

Accepted by ....
Arthur C. Smith
Professor of Electrical Engineering
Graduate Officer

Comprehensive Inverse Modeling for the Study of Carrier Transport Models in Sub-50 nm MOSFETs

by

Ihsan J. Djomehri

Submitted to the Department of Electrical Engineering and Computer Science on August 14, 2002 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

Direct quantitative 2-D characterization of sub-50 nm MOSFETs continues to be elusive. This research develops a comprehensive indirect inverse modeling technique for extracting 2-D device topology using combined log(I)-V and C-V data. An optimization loop minimizes the error between a broad range of simulated and measured electrical characteristics by adjusting parameterized profiles. The extracted profiles are reliable in that they exhibit decreased RMS error as the doping parameterization becomes increasingly comprehensive of doping features.

The inverse modeling methodology pieces together complementary MOSFET data sets such as capacitance of the gate stack, 1-D doping analysis, subthreshold I-V which is a strong function of 2-D doping, and C-V data which is especially sensitive to the source/drain. Combining the data sets enhances the extracted profiles. Such profiles serve as a basis for tuning diffusion coefficients in order to realistically calibrate modern process simulators.

The important application of this technique is in the calibration of carrier transport models. With an accurate device topology, the transport model parameters can be adjusted to predict the on-state behavior. Utilizing a mobility model that conforms to the experimental effective field dependence and including a correction for parasitic resistance, the transport model for an advanced NMOS generation at various gate lengths and voltages is calibrated. Employing the Energy Balance model yields an energy relaxation value valid over all devices examined in this work.

Furthermore, what has been learned from profile and transport calibration is used in investigating optimal paths for sub-20 nm MOSFET scaling. In a study of candidate architectures such as double-gate, single-gate, and bulk-Si, metrics for the power versus performance trade-off were developed. To conclude, the best trade-off was observed by scaling as a function of gate length with a single near-mid-gap workfunction.

Thesis Supervisor: Dimitri A. Antoniadis
Title: Professor of Electrical Engineering

Acknowledgements

I would like to acknowledge the assistance of many individuals during my Ph.D. thesis work. First and foremost, the guidance and support of my supervisor Prof. Dimitri Antoniadis were instrumental in carrying this project through. I am deeply appreciative of his general mentorship and insights into the research environment. Furthermore, the efforts of thesis readers and teachers Prof. Judy Hoyt and Prof. Jesus Del Alamo are very respectfully recognized.

I gratefully acknowledge all the help of my research group. Former students Dr. Zachary Lee, Dr. Tony Lochtefeld, Dr. Keith Jackson, Dr. Andy Wei, and Dr. Mark Armstrong provided my foundation. Graduate students Hasan Nayfeh, Ali Khakifirooz, Andy Ritenour, Isaac Lauer, Corina Tanasa, and Jim Fiorenza rounded out my collaboration. Important academic and industry contact with Dr. Hitoshi Wakabayashi, Prof. Ganesh Samudra of NUS, Prof. Mark Lundstrom of Purdue University, Prof. Robert Dutton of Stanford University, Dr. Jong Jung, Dr. Chris Bowen, Dr. Carlos Diaz, Dr. Jason Jenq, Dr. Phil Oldiges, Dr. Mei Kei Ieong, and especially Dr. Rafael Rios and Dr. Paul Solomon is gratefully recognized. Also, credit is due to Prof. Terry Orlando.

This work was supported largely by the SRC under contracts 98-NJ-625 and 2000-NJ-833.

I sincerely thank my parents, my wife, my family and friends for their patient support during my trials. In finality, I am reminded of the words "Read and your Lord is Most Honorable, He who taught the use of the pen, Taught man what he knew not." [Sura 96:3-5].

Table of Contents

List of Figures

= 3.3 nm device using “I-V & C-V” data versus “I-V Only” data. The methods give roughly the same junction depth but the S/D peak doping in “I-V Only” was arbitrarily set to 1×10^{20} cm^{-3}.

Figure 3.9: Comparison of the lateral doping profiles at the Si/SiO$_2$ surface for several t_{ox} = 3.3 nm devices with extracted physical L_{gate} = 100 nm, 130 nm, 160 nm using “I-V & C-V” data versus “I-V Only” data. The log(I)-V data provides sensitivity especially in the channel region while the addition of C-V data determines the S/D peak doping (which has two arbitrary settings of 1×10^{20} cm^{-3} and 5×10^{20} cm^{-3} for “I-V Only”) and slope.

Figure 3.10: Full C_{gg} characteristic extracting t_{ox} = 1.5 nm on this L = 10 μm device from an advanced NMOSFET technology. Due to leakage parasitic resistance, this C-V must be corrected [15] using two frequencies, here 800 kHz and 400 kHz. The fit exhibits good gate stack modeling of QM and polysilicon depletion effects.

Figure 3.11: Inverse modeling fit of the shortest t_{ox} = 1.5 nm device to careful C_{gds}-V measurements taken at 800 kHz and averaged between four samples per point for varying V_{BS} = 0.5 V, 0 V, and -1.5 V. The error is within the noise level of 0.025 fF/μm.

Figure 3.12: Extracted lateral profiles using the combined inverse modeling technique on t_{ox} = 1.5 nm devices; these short channel MOSFETs have L_{eff} ~ 35 nm, 45 nm, 55 nm, 80 nm, and 120 nm. The longer three lengths all fit to independent C_{gds}-V measurements. The shorter two lengths had no C-V data but their S/D peak values were fixed at the value obtained for the longer.

a decreased beta makes it harder for the device to reach velocity saturation.

Figure 4.3: Measured mobility at V_{DS} = 10 mV corrected for field above V_t for long channel devices with high bulk dopings extracted at 1×10^{17}, 8×10^{17}, 1.7×10^{18}, and 3.9×10^{18} cm^{-3}.

Figure 4.4: Optimization of coulombic mobility dependence as well as universal mobility coefficients using a range of I-V data for the long channel bulk devices with various doping levels.

Figure 4.5: Extracted curve of coulombic mobility versus doping level. The model assumes this mobility (some combination of impurity and phonon scattering) whenever it is under the universal curve. The mobility for minority carriers in bulk tracks this result.

Figure 4.6: Measured and calibrated mobility vs. effective field for a nitrided oxide NMOS technology with t_{ox} = 3.3 nm.

Figure 4.7: Measured and calibrated mobility vs. effective field for a nitrided oxide PMOS technology with t_{ox} = 3.3 nm.

Figure 4.8: A simulation of the effective short channel mobility at constant E_{eff} versus L_{eff} utilizing the calibrated coulombic mobility on the t_{ox} = 1.5 nm NMOS family. The good fit indicates that the merged halo doping likely degrades the mobility.

List of Tables

Chapter 1

Introduction

1.1 Motivation

As the semiconductor industry continues to shrink transistor sizes into the sub-100 nm regime, accurate characterization of the devices becomes essential. In order to obtain reliable models that will predict the performance of VLSI circuits, physical understanding and tools must be developed to determine both the physical structure of the devices and their transport behavior. This knowledge will then allow for an accurate analysis of the optimal path for sub-50 nm MOSFET scaling.

Unfortunately, direct characterization methods (e.g., scanning capacitance [1], scanning resistance [2], XTEM [3], or more recently 2-D SIMS [4]) have about one order of magnitude less spatial resolution and minimum doping sensitivity than needed for modern MOSFETs, not to mention that the measurement apparatus and sample preparation can be cumbersome. However, direct characterization may be deemed worth it to obtain an approximation to the doping.

On the other hand, indirect techniques [5] based on inverse modeling [6] using either C-V [7] or weak inversion log(I)-V [8] data have proven effective in extracting the 2-D profiles with required resolution. Whereas C_{gds}-V (where $C_{gds} = C_{gd} + C_{gs}$) data can provide good sensitivity for gate to source/drain (S/D) overlap doping features, subthreshold I-V data exhibit strong dependence on the channel profile, with weak dependence on the transport model and parasitic resistances and capacitances. Additionally, S/D to body diode capacitance has been used as another tool to detect aspects of the doping [9].

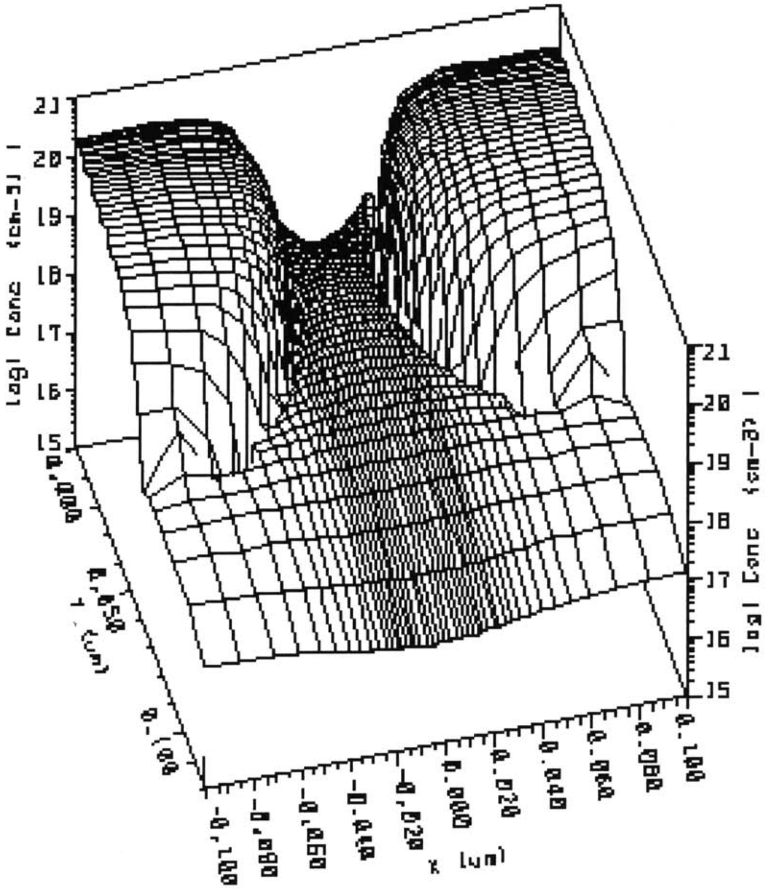

Figure 1.1: Complex 2-D doping distribution of an L_{eff} = 50 nm MOSFET generated in SUPREM exhibiting abrupt re-entrant source/drain regions, super-halo channel, and surface doping pile-up.

The complex doping distributions that appear after processing of modern MOSFETs [10] such as in Fig. 1.1 need all of their important features to be characterized. Hence, the thrust of this thesis will be to research ways to make the indirect, numerical technique of inverse modeling a comprehensive and reliable design evaluation tool for sub-100 nm devices. Furthermore, these results will be used to gain insight into the validity of short-channel transport models and the best path for future MOSFET scaling.

The project will progress through tasks summarized by the following: analysis of the reliability of inverse modeling, coding the non-linear optimization loop, performing accurate C-V and I-V measurements on advanced devices, developing an inverse modeling methodology that combines the data sets, inverse modeling of MOSFETs with a spread of channel lengths in multiple technology families, developing a methodology to calibrate and evaluate device transport models, study of MOSFET overscaling, designing sub-100 nm bulk and DG MOSFETs, and integration of the inverse modeling results to calibrate process simulation models.

1.2 Organization of the Thesis

The thesis goals and contributions to the field of device engineering are outlined in broad terms in the following paragraphs.

Chapter 2 will commence with a study of the computational techniques needed for the proposed device simulations. Understanding and appropriate choice of device physics models [11] and numerical solution methods, i.e. approximations, are required for accurate results. Then, methods of searching the parameter space, such as the design of experiments, are explored. It is important to evaluate the reliability of the inverse modeling optimizations to give confidence that the indirectly-obtained profiles are indeed unique.

Chapter 3 develops a technique for combining the log(I)-V weak inversion and C-V data that are most sensitive to the electrostatics of the 2-D gate-stack and doping structure. Fig. 3.1 provides a schematic for performing the idealized inverse modeling methodology. Each stage must have a detailed description of the experimental procedures for measuring the required data. By employing inverse modeling on a variety of sub-100 nm MOSFET technologies, its utility will be realized. The acquired profiles are useful for comparison and modification of process simulations.

Chapter 4 shows once the device structure has been acquired and fits the electrostatic behavior, it is then possible to investigate carrier transport. It is crucial to select transport models that will have the most physical significance. The methodology [12] used to calibrate these models must isolate the effects of low-field mobility, parasitics, and high-field transport to ensure the extraction of a unique set of transport parameters. The calibrated results can show excellent agreement over a family of device lengths. Also, the short-channel mobility can be explored.

Chapter 5 utilizes accurate models for both MOSFET topology and transport behavior to pursue new paths in scaling [13] to the sub-20 nm regime. For bulk-Si, relationships [14] between device integrity and doping as a function of scaling are developed. A technique of "overscaling" devices beyond the "well-tempered" electrostatic regime yields insight into the performance vs. power trade-off. From these analyses, an optimal double gate design can be infered.

Chapter 6 provides a conclusion to the thesis and suggestions for future work.

Chapter 2

Computational Techniques for Simulation

2.1 Relevant Device Physics

Before embarking on a project to characterize a device's electrostatic and transport properties, a critical review of the underlying physics relevant to the problem proves useful. The governing equations of electromagnetism [15] and the conservation laws must be solved as well as the statistical and quantum mechanics (QM) [16]. Consequently, analytical models for MOSFET operation are developed. Herein, the coordinate system is defined as Cartesian with X along the length and Y running depthwise of the MOSFET.

The core equations of motion in device simulation mix Maxwell's equations with Boltzmann transport. Because modern transistor logic is not optical, the Poisson's equation

$$0 = \varepsilon\nabla^2\psi + q(p - n + N_d - N_a) + \rho = F_0 \qquad \text{Equation 2.1}$$

contains the necessary electrodynamics where ε is the material permittivity, ψ is the potential, q is the elementary charge, ρ is extra fixed charge, and p, n, N_d, and N_a are the hole, electron, donor impurity, and acceptor impurity concentrations, respectively. For most purposes here, macroscopic approximations underpin the transport computations with the continuity equations

$$0 = \frac{1}{q}\nabla \bullet \bar{J}_n - U_n - \frac{\partial n}{\partial t} = F_{1n}$$ Equation 2.2

$$0 = \frac{-1}{q}\nabla \bullet \bar{J}_p - U_p - \frac{\partial p}{\partial t} = F_{1p}$$ Equation 2.3

where J_n and J_p represent the current density for electrons and holes, respectively. The quantities labeled as U indicate the net carrier recombination rates with most processes proportional to (np - n_i^2), where n_i is the intrinsic concentration. In static flow, the time derivates go to zero while

$$\bar{J}_n = -q\mu_n n \nabla \phi_n$$ Equation 2.4

$$\bar{J}_p = -q\mu_p p \nabla \phi_p$$ Equation 2.5

represent the densities as a function of mobility, μ, scattering terms and the gradient of quasi-Fermi potentials, ϕ, which include carrier diffusion and drift phenomena due to the applied field.

The interrelation between carrier concentration and potential introduces the role of statistical mechanics. The integral of the density of states multiplied by the population function yields

$$n = N_C F_{1/2}(\eta_n) \qquad \eta_n = \frac{-q\phi_n - E_C}{kT}$$ Equation 2.6

$$p = N_V F_{1/2}(\eta_p) \qquad \eta_p = \frac{q\phi_p + E_V}{kT}$$ Equation 2.7

with k as the Boltzmann constant and T as absolute temperature where N_C, E_C, N_V, and E_V are the effective density of states and energies for the conduction and valence bands edges, respectively, of the solid-state material (such as silicon). Using Fermi-Dirac statistics is necessary for highly doped semiconductors due to the dependence on the difference in quasi-Fermi and band energy in the integral of order 1/2

$$F_{1/2}(\eta_f) = \frac{2}{\sqrt{\pi}}\int_0^{\infty}\frac{\eta^{1/2}}{1+e^{\eta-\eta_f}}d\eta \qquad \text{Equation 2.8}$$

Furthermore, advanced MOSFETs with thin gate dielectrics must account for the QM effect of carrier quantization in the inversion layer. Although solving the Schroedinger equation [17] for the carrier wave function would be most accurate, for the devices in this study the Van Dort approximation [18] is sufficient and dramatically saves time.

Because the carrier distribution shifts away from the surface, the threshold voltage increases and the energy level splitting of the band gap, E_g, is modeled as

$$\Delta E_g = \frac{13}{9}\beta_{QM}\left(\frac{\varepsilon}{4kT}\right)^{1/3}|E_{\perp}|^{2/3} \qquad \text{Equation 2.9}$$

where $E_{\perp}$ is the electric field perpendicular to the surface at any grid point, and β_{QM} is a fitting factor roughly $4.1\times10^{-8}\ eV\cdot cm$. Also, the QM concentration shifts in the inversion layer by

$$n_{i,QM} = n_i e^{\frac{-\Delta E_g}{2kT}} \qquad \text{Equation 2.10}$$

To estimate the effect of further 2-D QM as the gate length decreases to sub-10 nm, consider [19] the potential V(x) between S/D as a harmonic oscillator barrier with peak x_0

$$qV(x) = \frac{m\omega^2}{2}(x-x_0)^2 \qquad kT = \frac{h\omega}{2\pi} \qquad \text{Equation 2.11}$$

where m is the effective mass and ω signifies the curvature. This ω can be translated into an effective temperature T. In the L = 6.5 nm simulation of Fig. 2.1, a V_{DD} of 0.6 V induces a curvature in the potential that is less than the curvature at T = 300 K, indicating that the S/D tunneling is still below the thermal off-current.

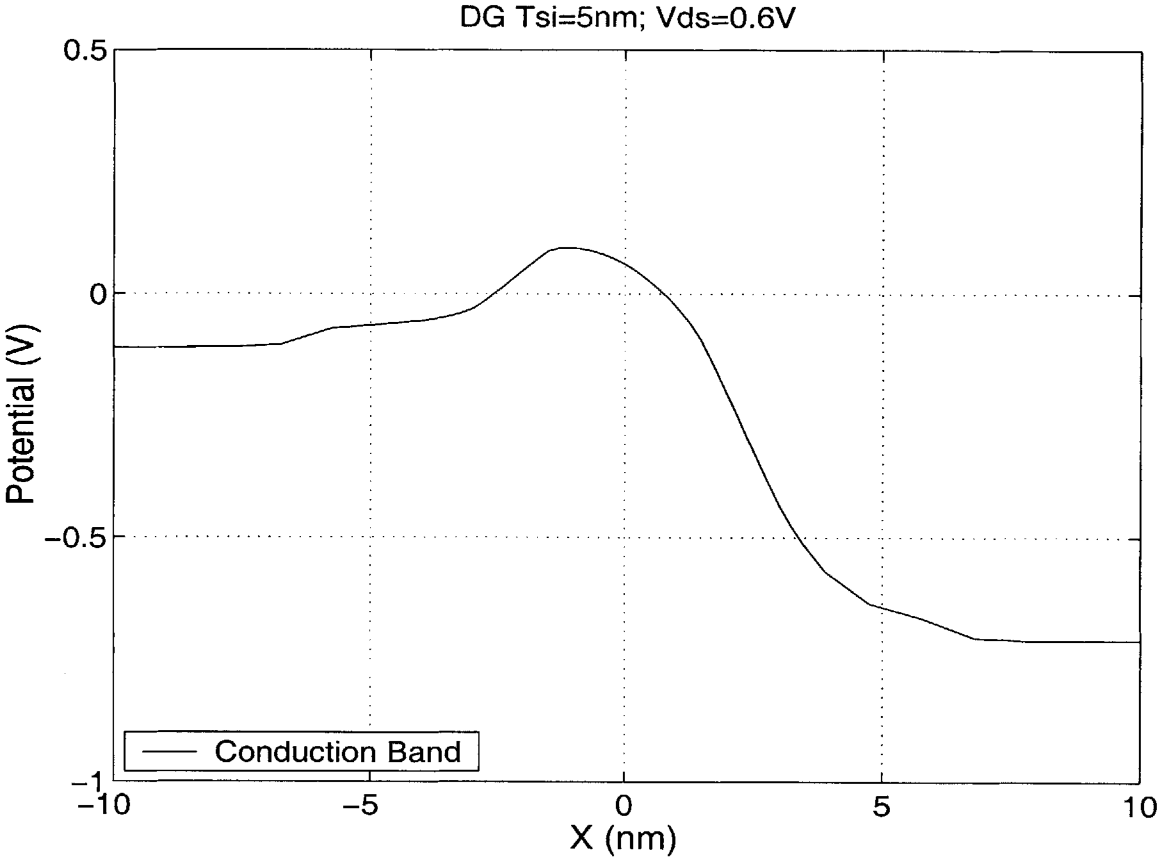

Figure 2.1: Classical shape of the conduction band in the lateral direction of an undoped L = 6.5 nm MOSFET at V_{GS} = 0 V and V_{DS} = 0.6 V. The band bending lacks enough curvature for the S/ D tunnelling to swamp the thermal current.

Another leakage path across material junctions is band tunneling which is described generically by a Fowler-Nordheim model where α and β are fitting parameters and absolute current density, J, is proportional to the applied electric field, E, as

$$J = \alpha E^2 e^{\frac{-\beta}{E}} \qquad \text{Equation 2.12}$$

Finally, using a few simplifications of these numerical equations results in an approximate yet handy analytic description of MOSFET operation meant for reference. The C-V characteristics will be discussed later and their measurement technique is described in Appendix B. In weak inversion, the drive current, I_D, is diffusion dominated and exponential with applied voltage yet linear with mobility. In strong inversion, drift is the dominant mechanism. For qualitative reference, the equations [20] for a MOSFET of width W and length L with uniform channel doping, N, in subthreshold condense to

$$I_D \approx WD\frac{dQ_I}{dx} = \frac{W}{L}I_x\exp\left(\frac{V_{GS}-V_x}{n\phi_t}\right)\left[1-\exp\left(-\frac{V_{DS}}{\phi_t}\right)\right]$$

Equation 2.13

$$I_x = \mu\phi_t^{\,2}\frac{\gamma_{eff}C_{ox}}{2\sqrt{1.5\phi_F - V_{BS}}}\exp\left(-\frac{\phi_F}{2\phi_t}\right)$$

Equation 2.14

$$V_x = V_{FB} + 1.5\phi_F + \gamma_{eff}\sqrt{1.5\phi_F - V_{BS}} - \sigma_{DIBL}(N, V_{BS})V_{DS}$$

Equation 2.15

$$n = 1 + \frac{\gamma_{eff}}{2\sqrt{1.5\phi_F - V_{BS}}} \qquad \gamma_{eff} = A_{SCE}\frac{\sqrt{2\varepsilon_s qN}}{C_{ox}}$$

Equation 2.16

where Q_I is the areal inversion charge, the Einstein relation gives $D = \mu\phi_t$, the thermal voltage ϕ_t = kT/q, ϕ_F is the Fermi potential, V_{FB} is the flat band voltage, C_{ox} is the oxide capacitance, and γ_{eff} the body factor along with the drain induced barrier lowering parameter σ_{DIBL} and a short-channel-effect fitting parameter A_{SCE} are highly dependent on the doping. For inverted operation, a relatively crude approximation for the current goes linear with voltage until the drain falls below the threshold voltage [21], V_t, at $V_{DS}^{(sat)}$

$$I_D \sim \int_0^L I_{Drift}dx = \frac{W}{L}\mu C_{ox}\left[(V_{GS}-V_t)V_{DS} - \frac{1}{2}V_{DS}^{\,2}\right]$$

Equation 2.17

$$V_{DS}^{(sat)} = V_{GS} - V_t \qquad V_t \sim V_{FB} + 2\phi_F$$

Equation 2.18

2.2 Numerical Solutions

If describing the solid-state physics as a set of relevant equations as in the previous section is accurate, then theory shall readily characterize the behavior of any device. However, as in most scientific cases, the complexity of the system demands numerical methods [22] to provide timely self-consistent solutions. These methods, along with techniques to optimize the devices at hand, require an overview to ensure their robustness.

Most sets of nonlinear partial differential equations can be solved accurately by linearizing into matrix form and iterating. When the electrostatics are relatively independent of the conservation laws, decoupled iterations that solve each equation individually are suitable. More generally,

$$\frac{\partial}{\partial \bar{x}} F_i(\bar{x}) \Delta \bar{x} = -F_i(\bar{x}) \qquad \text{Equation 2.19}$$

represents the coupled solution using Newton's method where the F_i are the governing equations (e.g., Eq. 2.1 to 2.3) and $\Delta\bar{x}$ is the update to the unknowns; the matrix in expanded form is

$$\begin{bmatrix} \frac{\partial F_0}{\partial \psi} & \frac{\partial F_0}{\partial n} & \frac{\partial F_0}{\partial p} \\ \frac{\partial F_{1n}}{\partial \psi} & \frac{\partial F_{1n}}{\partial n} & \frac{\partial F_{1n}}{\partial p} \\ \frac{\partial F_{1p}}{\partial \psi} & \frac{\partial F_{1p}}{\partial n} & \frac{\partial F_{1p}}{\partial p} \end{bmatrix} \begin{bmatrix} \Delta\psi \\ \Delta n \\ \Delta p \end{bmatrix} = - \begin{bmatrix} F_0 \\ F_{1n} \\ F_{1p} \end{bmatrix} \qquad \text{Equation 2.20}$$

This equation can itself be solved directly using Gaussian elimination, which unfortunately becomes computationally intractable for larger problems. Several iterative methods exist to solve the Ax = b matrix problem such as the conjugate gradient method. Fast convergence occurs if A has degenerate eigenvalues when essentially minimizing the residual

$$r = b - Ax \qquad \text{Equation 2.21}$$

For systems involving transient physical behavior, the time evolution is sufficiently approximated by first or second order difference equations. In the particular case of the small-signal response of a device, the vector x of unknowns is assumed to be of the form

$$\bar{x} = \overline{x_{DC}} + \overline{x_{AC}}e^{j\omega t} \qquad \text{Equation 2.22}$$

which leads to a DC solution and a set of AC parameters at frequency ω.

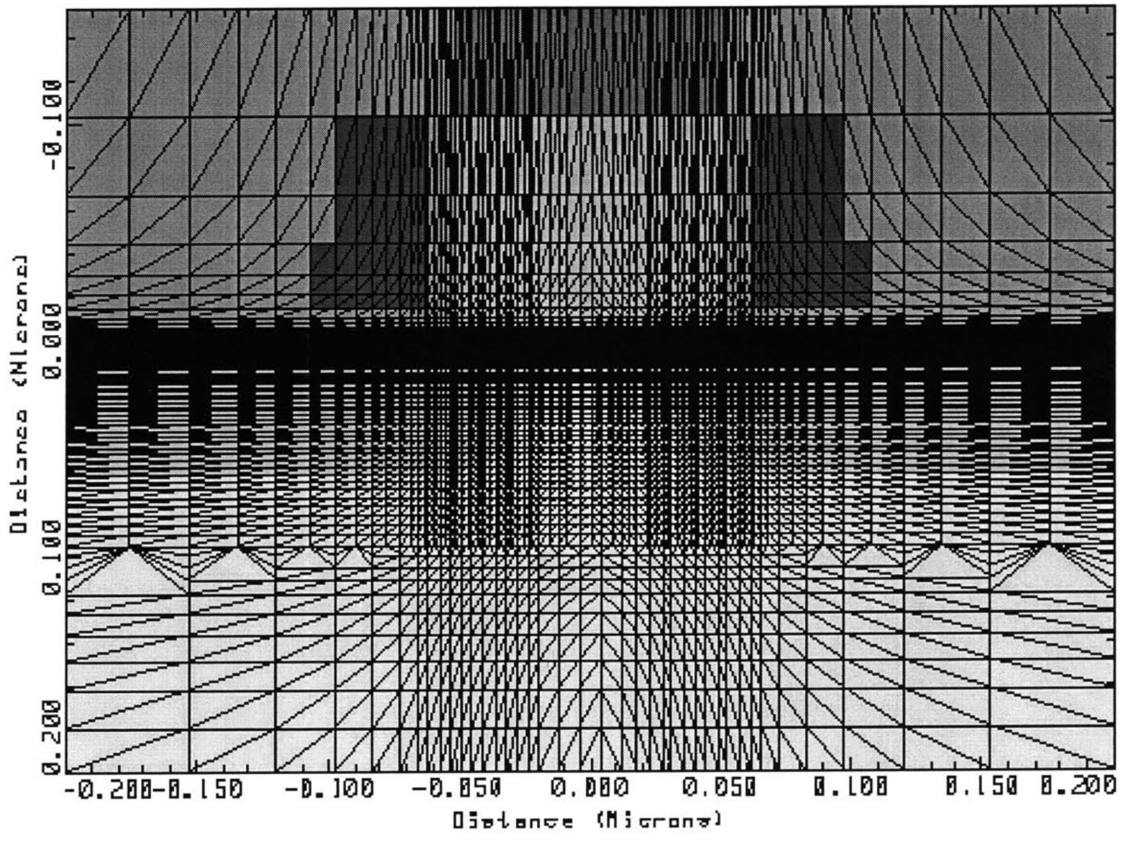

Figure 2.2: Plot delineating the 2-D rectangular gridding approach on a archetypal MOSFET with gate stack dielectrics and substrate doping distributions. The mesh becomes finer near regions of heavy transport and where structural details change rapidly.

Of course, the numerical apparatus must solve the aforementioned equations at *each* grid point in the problem. Thus, the finesse given to meshing issues often determines the ease of convergence. A typical gridding scheme for a MOSFET with a 2-D cross-section of the gate stack and doped substrate is rendered in Fig. 2.2. Although too coarse a mesh fails to capture rapid changes in the potential and carrier distributions, too fine a mesh can lead to unphysical values (e.g., if less than the lattice spacing ~ 0.5 nm). Moreover, the non-uniform mesh should be as smooth as possible: in the vertical dimension, a good choice grid spacing often varies geometrically from 1/3 of gate height at the top to 1/3 of t_{ox} at the oxide interface (which gives a few mesh lines in the inversion layer) to 1/10 of the substrate thickness at the bottom simulation edge; in the lateral direction, the grid is symmetric about mid-channel and typically ranges from 1/20 of the gate length to half the characteristic length of doping at the metallurgical junction to 1/5 of the spacer at the simulation edge. Since most device features are sufficiently rectangular, the governing equations are solved using finite difference between grid points rather than finite element.

Furthermore, good approximations used as the initial guess to the mesh solution often require projections of previous simulations and small increments to the bias points. A solution is determined to have converged when the error norm (difference between both sides of the equations) is below a specified tolerance. The non-contact materials observe the boundary condition

$$\varepsilon_1\frac{\partial\psi_1}{\partial\hat{n}} - \varepsilon_2\frac{\partial\psi_2}{\partial\hat{n}} = \sigma_s \qquad \text{Equation 2.23}$$

where the difference in permittivities times the derivatives of the potentials normal to the boundary realizes any existing surface charge, σ_s.

The last piece of structural information in standard device simulations is the impurity distribution. A simple yet powerful way to describe the 2-D doping is through a superposition of representation functions where the A_j are the peak doping values

$$N(x, y) = \sum_j A_j fX_j(x) fY_j(y) \qquad \text{Equation 2.24}$$

and the X and Y components are factored as exponentially varying functions such as gaussians

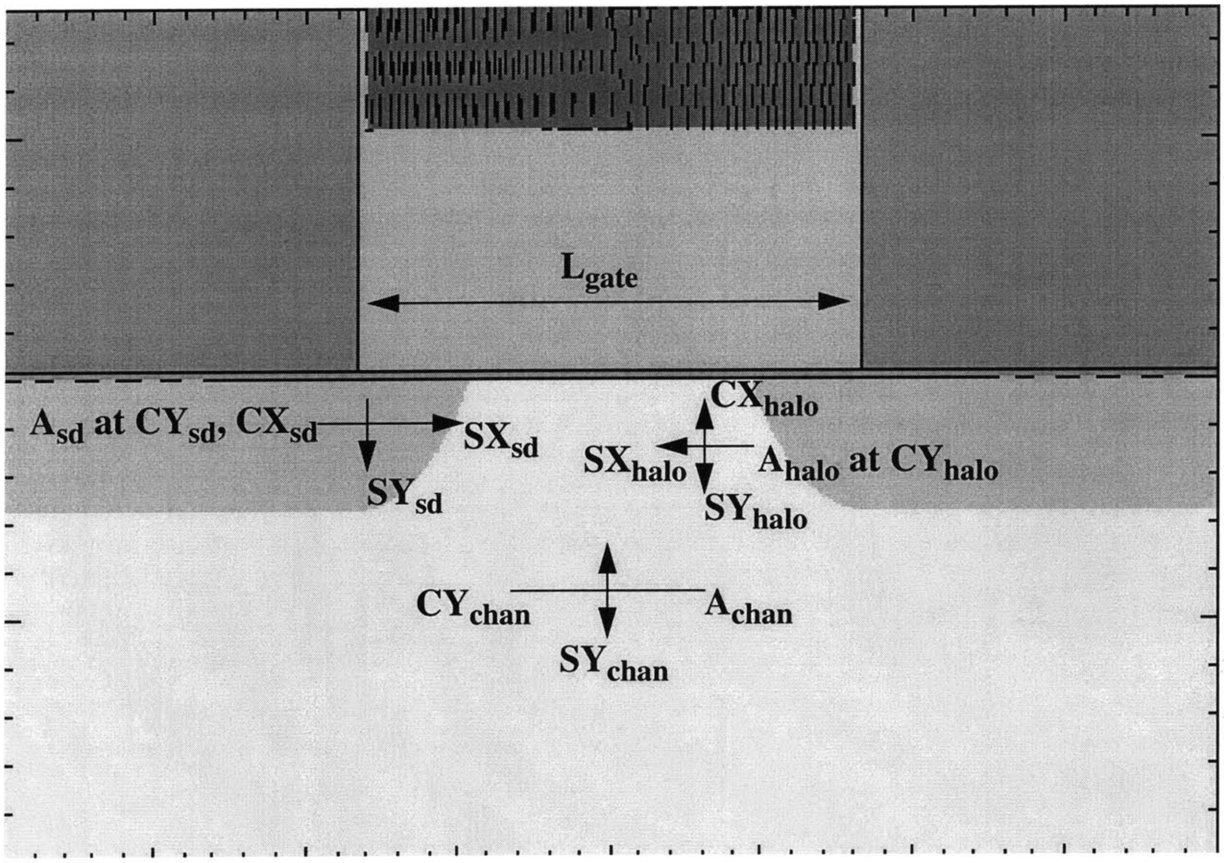

Figure 2.3: An illustration detailing the typical parameters used in analytically describing the complex 2-D doping profiles of a MOSFET. Each impurity has a peak concentration, X and/or Y center positions and associated sigma characteristic roll-off lengths.

$$fX_j(x) = \exp\left[-\left(\frac{x - C_{x,j}}{\sigma_{x,j}}\right)^2\right] \qquad \text{Equation 2.25}$$

$$fY_j(y) = \exp\left[-\left(\frac{y - C_{y,j}}{\sigma_{y,j}}\right)^2\right] \qquad \text{Equation 2.26}$$

As depicted in Fig. 2.3, the source/drain (S/D) and halos of the typical MOSFET are represented analytically by 2-D gaussians with center positions C_x and C_y, and characteristic lengths σ_x and σ_y; the channel implants usually take the form of 1-D gaussians.

Finally, with a reasonable device simulator in place, optimization of device parameters [23] becomes feasible. A non-linear optimizer such as Levenburg-Marquardt algorithm [24] takes

$$F(\bar{p}) = \sum_i h_i^2(\bar{p}) = \sum_i [y_i^{(sim)}(\bar{p}) - y_i^{(\exp t)}]^2 \qquad \text{Equation 2.27}$$

the sum, F, of squared errors h_i between simulated and experimental electrical data y_i and tries to minimize it with respect to the vector of parameters, $\bar{p}$, to be optimized with

$$\nabla_{\bar{p}} F = 2J^T\bar{h} = 0 \qquad \text{Equation 2.28}$$

where J signifies a special matrix containing the sensitivity of each data on each parameter

$$J_{ij} = \frac{\partial h_i}{\partial p_j} \qquad \text{Equation 2.29}$$

called the Jacobian. To find the zeros of the gradient of F, once again apply Newton's method

$$\left[2J^T J + 2\sum_i h_i \nabla^2 h_i\right] \Delta\bar{p} = H\Delta\bar{p} = -2J^T\bar{h} \qquad \text{Equation 2.30}$$

to reduce the problem to solving Ax = b for the parameter updates $\Delta\bar{p}$ using the matrix

$$H_{ij} = \frac{\partial^2 F}{\partial p_i \partial p_j} \qquad \text{Equation 2.31}$$

called the Hessian. The magnitude of the elements of the matrix H also indicate the sensitivity to a particular pair of parameters. For practicality, the algorithm approximates the Hessian as

$$H \approx 2J^T J + \lambda D \qquad \text{Equation 2.32}$$

where D is a matrix with only the diagonal of J. The optimization begins as a steepest descent method with large l which is then reduced every iteration. Until the desired data error tolerance is met, the parameters of the ith iteration are revised as

$$\bar{p}^{i+1} = \bar{p}^i + \Delta\bar{p}^i \qquad \text{Equation 2.33}$$

To conclude, an analysis of the time per iteration based on a computer's intrinsic speed τ_{sim} which is multiplied by the size of the mesh, data, and number of parameters is

$$\tau_{iteration} = (1 + n_{para}) \cdot n_{data} \cdot n_{node} \cdot \tau_{sim} \qquad \text{Equation 2.34}$$

2.3 Searching the Parameter Space

When designing or evaluating a device, it would be handy if one could accurately predict how changing certain device parameters [25] would affect certain characteristics and measures of performance. Inverse modeling is a technique which accomplishes these tasks through a numerical optimization procedure. Another mathematical method is the Design of Experiments which searches many parameters in as few as possible combinations, runs experiments to obtain the desired characteristics, and generates a multi-dimensional response surface fitted to these values.

A choice of parameters with both efficiency and accuracy for the Design of Experiments is the Central Composite Design (CCD). Each of N parameters is assigned a range of values including a minimum, midpoint, maximum. There are also two "2-Level" points equidistant from the midpoint by a distance $(maximum - midpoint)/(\sqrt[4]{2^N})$. Next, the following experiments are conducted: 1) runs of all combinations of parameters being at 2-Level points; 2) a centerpoint (all parameters at the midpoint) run counted about $3 \times \sqrt{2^N}$ times; and 3) axial runs of the minimum and maximum of a particular parameter with the midpoints of the others. A general purpose C++ program has been written to carry out this procedure. A standard device simulator performs the runs and the desired figure of merit output is then tabulated. Last, a statistical mathematics package such as SPLUS is utilized to compute a non-linear least squares fit of the data to a response manifold.

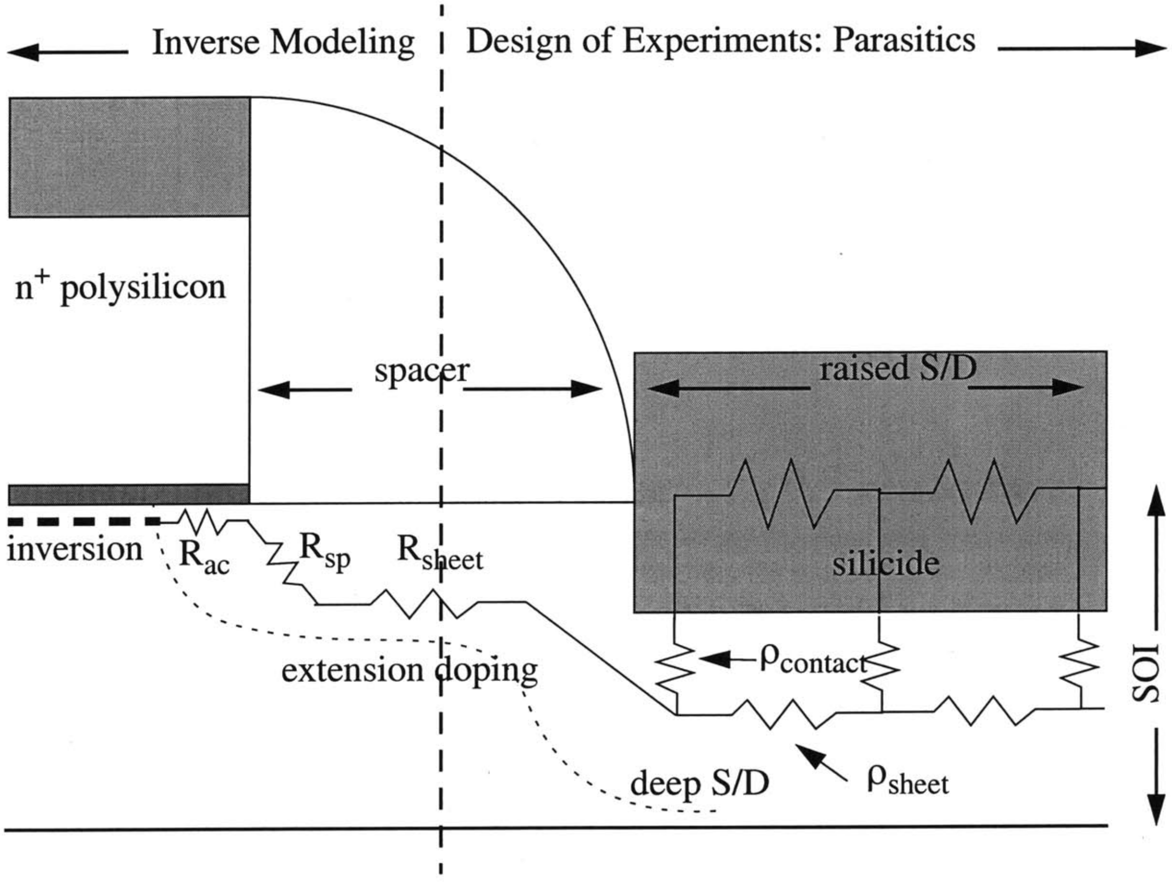

Figure 2.4: Example structure of the drain region of a MOSFET under investigation using Inverse Modeling and Design of Experiments. The external parasitic resistance here is determined by varying parameters such as S/D length, thickness, doping and contact material.

The simulated test structure of Fig. 2.4 depicts the drain region of a MOSFET that can be investigated by different methods. For example, the Design technique has been used to study the changes in the parasitic S/D resistance, R_{SD}, in a raised contact SOI NMOSFET. A small 0.1 V potential was applied from an artificial "contact" at the inversion layer to the raised contact; thus, R_{SD} can be calculated as twice the voltage divided by the drain current. The arbitrarily chosen features include a raised S/D height of 60 nm with a silicide contact extending down 40 nm from the top, and a donor concentration of 2×10^{20} cm^{-3} in the deep S/D region.

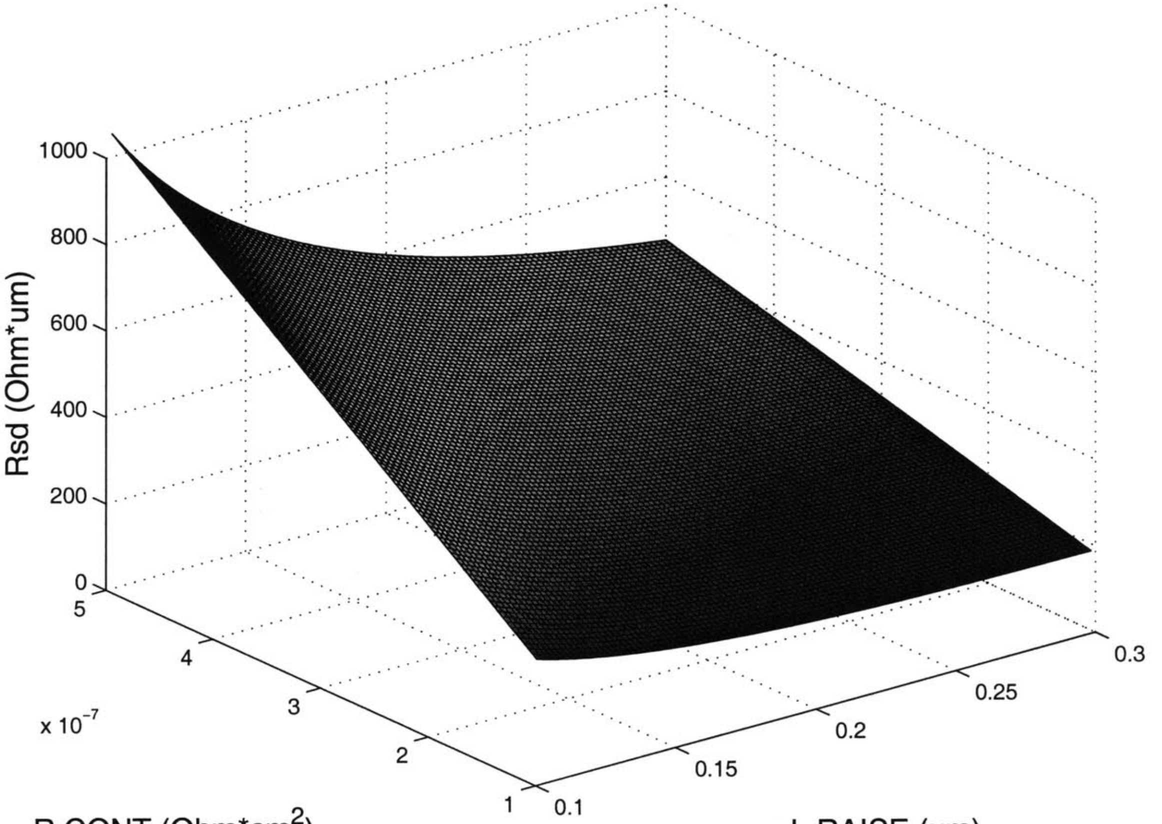

Figure 2.5: Design of experiments parameter space surface for R_{SD} as a function of the dominant components of contact resistance and length of the raised S/D.

Several experiments were then conducted varying the following parameters: length of the S/D, L.EXT, from 13 to 37 nm; length of the raised region, L.RAISE, from 100 to 300 nm; SOI thickness, Y.SOI, from 10 to 30 nm; doping concentration in the S/D extension, D.EXT, from 5×10^{19} to 1.5×10^{20} cm^{-3}; and contact resistivity, R.CONT, from 1×10^{-7} to 5×10^{-7} Ωcm^2.

Picking the midpoints for L.EXT, Y.SOI and D.EXT, one can render the response surface for R_{SD} in Fig. 2.5 as a function of the two dominant parameters R.CONT and L.RAISE. Most of the resistance components follow the physically intuitive dependence

$$R = \rho\frac{Length}{Area} \qquad \text{Equation 2.35}$$

where ρ is resistivity and the proportionality is linear with length and inverse with contact area.

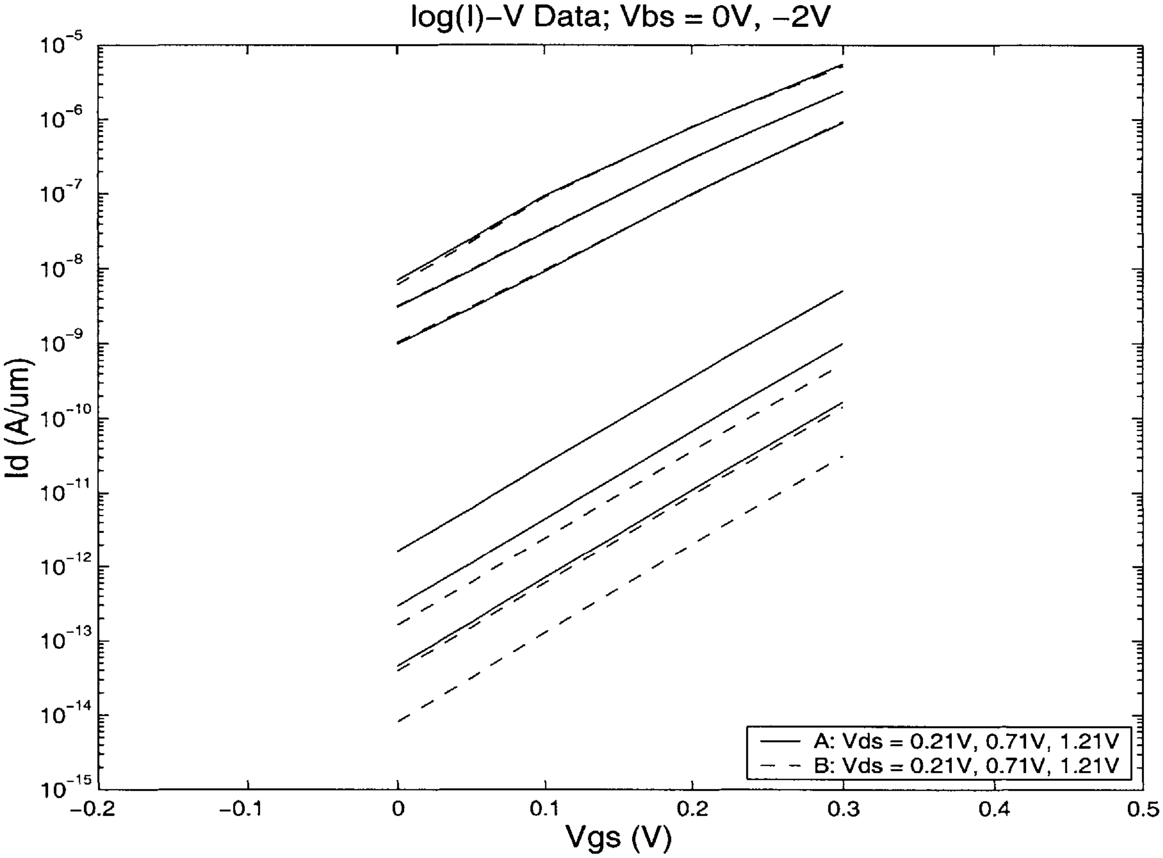

Figure 2.6: Subthreshold I-V curves for two L_{eff} = 50 nm NMOS devices with different 2-D doping that exhibit the same I_{off}, V_t and DIBL at V_{BS} = 0 V. The electrostatics diverge at V_{BS} = -2 V.

The trickiest step of this method is coming up with an appropriate fitting function that minimizes the error. In general, this formula consists of linear, inverse linear, higher order, and cross terms. In practice, it requires knowledge of the underlying physics as

$$R_{SD} = k_0 + k_{10}L_{EXT} + \frac{k_{i1}}{L_{RAISE}} + k_{11}L_{RAISE} + \frac{k_{i2}}{Y_{SOI}} + k_{12}Y_{SOI} + \frac{k_{i3}}{D_{EXT}} + k_{13}D_{EXT} + k_{14}R_{CONT} + k_{24}R_{CONT}^{\ 2} + k_c\frac{R_{CONT}}{L_{RAISE}}$$

Equation 2.36

where the k's are fitted coefficients. From Eq. 2.X, one justifies the linear dependence on L.EXT and the mostly inverse linear depedence on Y.SOI, D.EXT (resistivity varies inversely with doping), and L.RAISE (proportional to contact area).

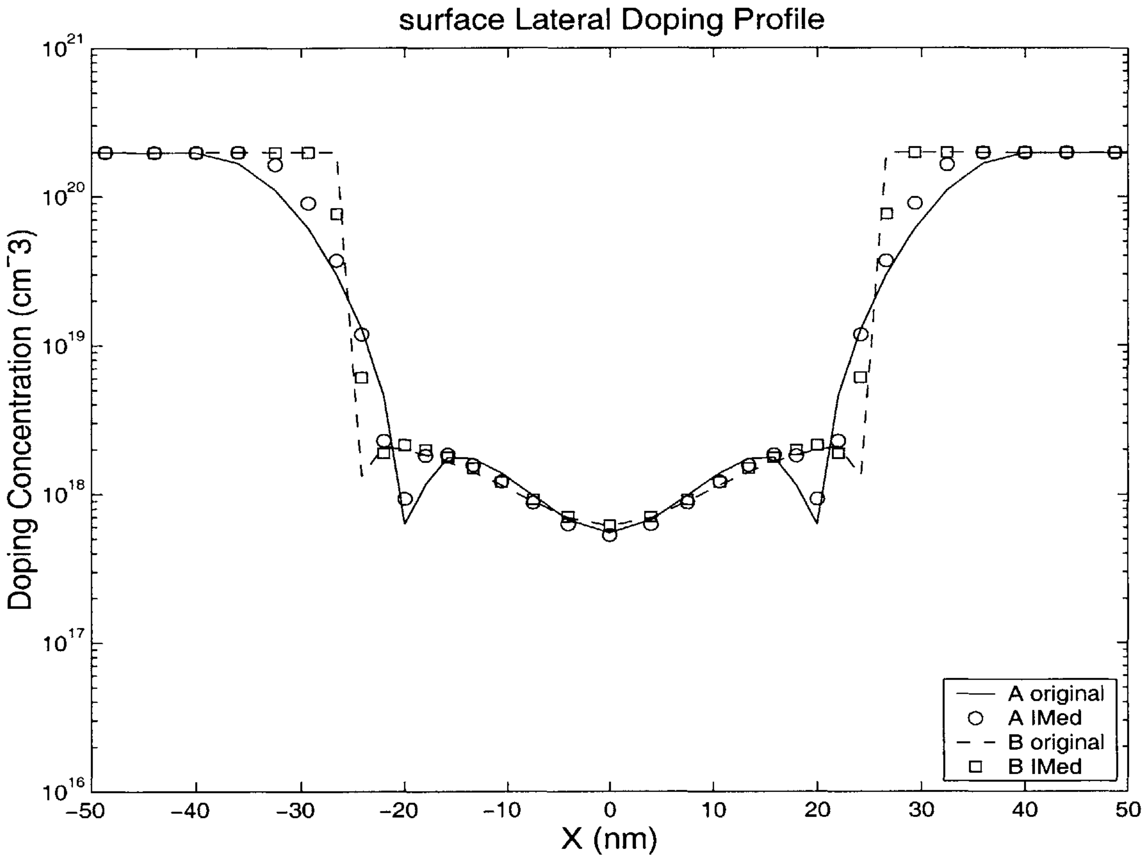

Figure 2.7: Inverse modeling fits using the entire data range showing the lateral doping profiles at the surface for the two $L_{eff} = 50$ nm NMOS devices with same electrostatics at $V_{BS} = 0$ V.

Even with good means of searching parameter space, one must verify how broad a range of electrical data is necessary to ensure a unique profile. As an exercise, two simulated $L_{eff} = 50$ nm NMOS devices have been constructed with identical electrostatic qualities (i.e., I_{off}, V_t and DIBL) at a back bias, $V_{BS} = 0$ V, as in Fig. 2.6. Device A trades off the electrostatic influence of a gradual S/D with a short junction depth, x_j, of 25 nm while B has an abrupt fall-off and an x_j of 40 nm; the halos are modified slightly to maintain the desired I-V.

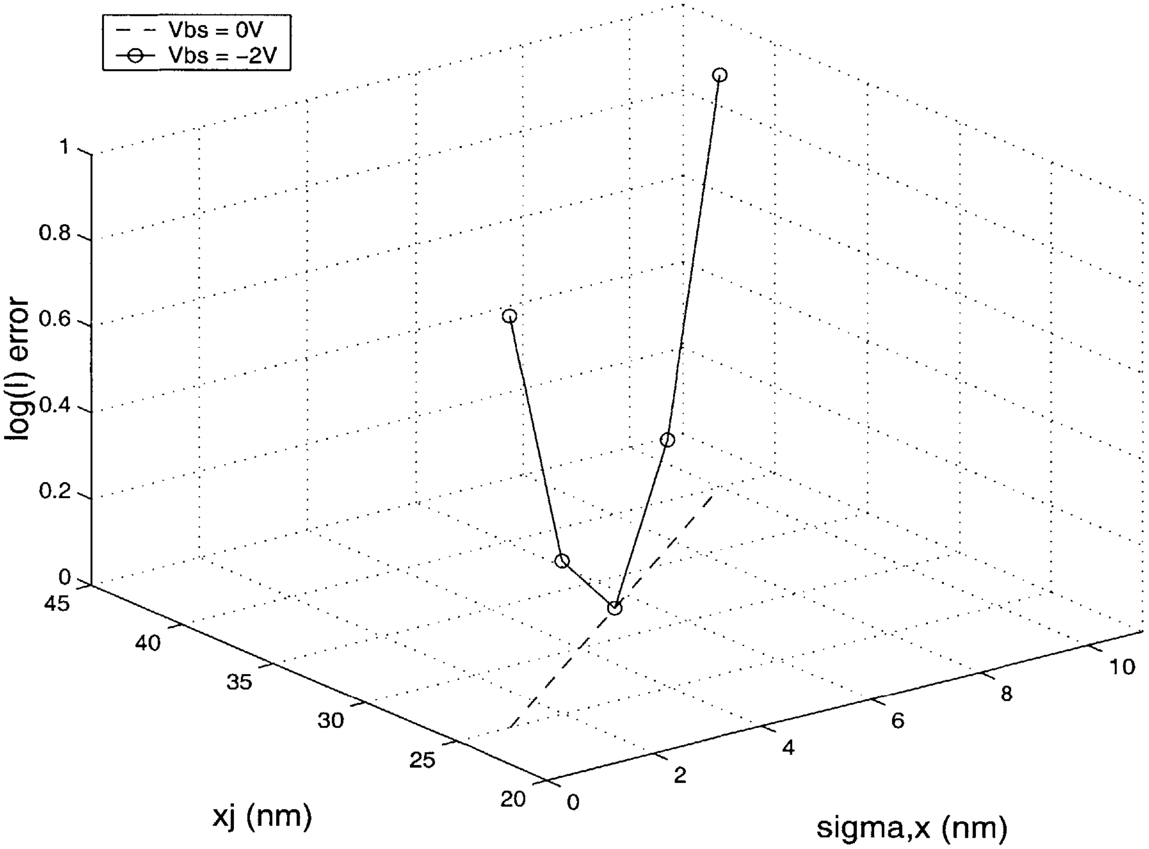

Figure 2.8: Order of magnitude error for I-V in parameter space as the example $L_{eff} = 50$ nm NMOSFET trades off junction depth and lateral roll-off. There is no error for $V_{BS} = 0$ V but a global minimum appears for $V_{BS} = -2$ V.

However, as soon as a broad range of bias is included in the log(I)-V data, the difference in electrical signature becomes apparent. In fact, it is only through utilizing this difference in electrical data that the optimization loop was able to distinguish between the two dissimilar doping profiles as in Fig. 2.7. The lateral cut of the 2-D profile shows that a wide range of data, in this case

V_{BS} = -2 V which detects doping deeper, is vital to ensure that the inverse modeling method captures all existing topological device features. If there are no electrical data sensitive to a particular feature, detection cannot be expected.

To further highlight the importance of selecting a broad range of data, the error is plotted as a function of the parameter space variables x_j and σ_x in Fig. 2.8. The line of zero error represents a continuum of devices that trade-off between depthwise and lateral junction parameters. Going along the same multi-variable path but looking at the error for V_{BS} = -2 V produces an error curve relative to the I-V data for the device with the midpoint doping parameterization. Clearly, using more data is a means of obtaining a true global minimum. The graphical depiction of error accentuates how the optimizer calculates the updated search direction and verifies that enough sensitivity exists to extract the parameters.

2.4 Reliability of Optimizations

While the potential power of inverse modeling in 2-D profiling of sub-100 nm devices is evident, it would be ideal to have a means of characterizing the reliability of this technique. However, since no direct approach can currently verify the modeling results, we resort to a heuristic assessment. The most obvious question is how close the simulated electrical data from the inverse modeled device agree with the experimental electrical data. From a wide array of inverse modeling experience, it is found that the majority of converged results typically exhibit relative RMS errors below 0.1.

All positions reference mid-channel as X = 0 and the surface as Y = 0. The starting value for the S/D extension peak doping, A_{sd}, is 2×10^{20} cm^{-3} which extends down uniformly to a center Y position, $C_{y,sd}$, of 5 nm beyond which it starts to roll-off. As for the halo centers, Cx,h is usually placed at the lateral metallurgical junction at $C_{x,sd}$ - $1.5 \times \sigma_{x,\,sd}$, and $C_{y,h}$ is located at half of

x_j, which is tracked by $C_{y,c}$.

Node L_{gate} (nm)	$C_{x,sd}$ (nm)	$\sigma_{x,\,sd}$ (nm)	$\sigma_{y,\,sd}$ (nm)	A_h $\frac{10^{18}}{cm^3}$	$\sigma_{x,\,h}$ (nm)	$\sigma_{y,\,h}$ (nm)	A_c $\frac{10^{18}}{cm^3}$	$C_{y,c}$ (nm)	$\sigma_{y,\,c}$ (nm)
130	65	7	25	1	14	50	1	55	50
100	50	6	20	2	12	40	2	45	40
70	35	5	15	3	10	30	2	35	30
50	25	4	10	4	8	20	1	25	25
30	15	3	5	5	6	10	“	15	“
20	10	2	3	6	4	6	“	11	“

Table 2.1: For each Road Map node, these initial guess specifications for the MOSFET source/drain and channel 2-D gaussian doping parameterizations should lead to inverse modeling convergence given a broad enough range of electrical data.

How easy is it to achieve convergence to the global minimum error with a given parameterization? Taking the aforementioned sum of two 2-D gaussians for the S/D and halos and 1-D gaussian for the channel as the doping representation functions, Table 2.1 quantifies inital guess parameter values for each L_{gate} node that should lead to convergence. The estimates for characteristic roll-off lengths derive from experience in modeling industry devices: the S/D is typically twice as abrupt as the halos. Also, the super-steep retrograde channel doping is not very significant for the shortest nodes which are essentially super-halo devices. If numbers for an intermediate node are desired, simple interpolation of the doping level trends will suffice.

More challenging questions present themselves regarding the ideality of the final set of doping function parameters. Does the converged parameter set actually represent the original profile? An equivalent statement is whether the solution becomes unique using the given doping representation functions. Furthermore, how closely does the inverse modeled parameterization approximate the real 2-D doping distribution? Perhaps even more significant, how complex should the parameterization be to obtain the best fit and to what extent can this choice of representation functions be applied generally? A promising strategy to attack these questions is via qualitative numerical studies because direct methods are too inaccurate.

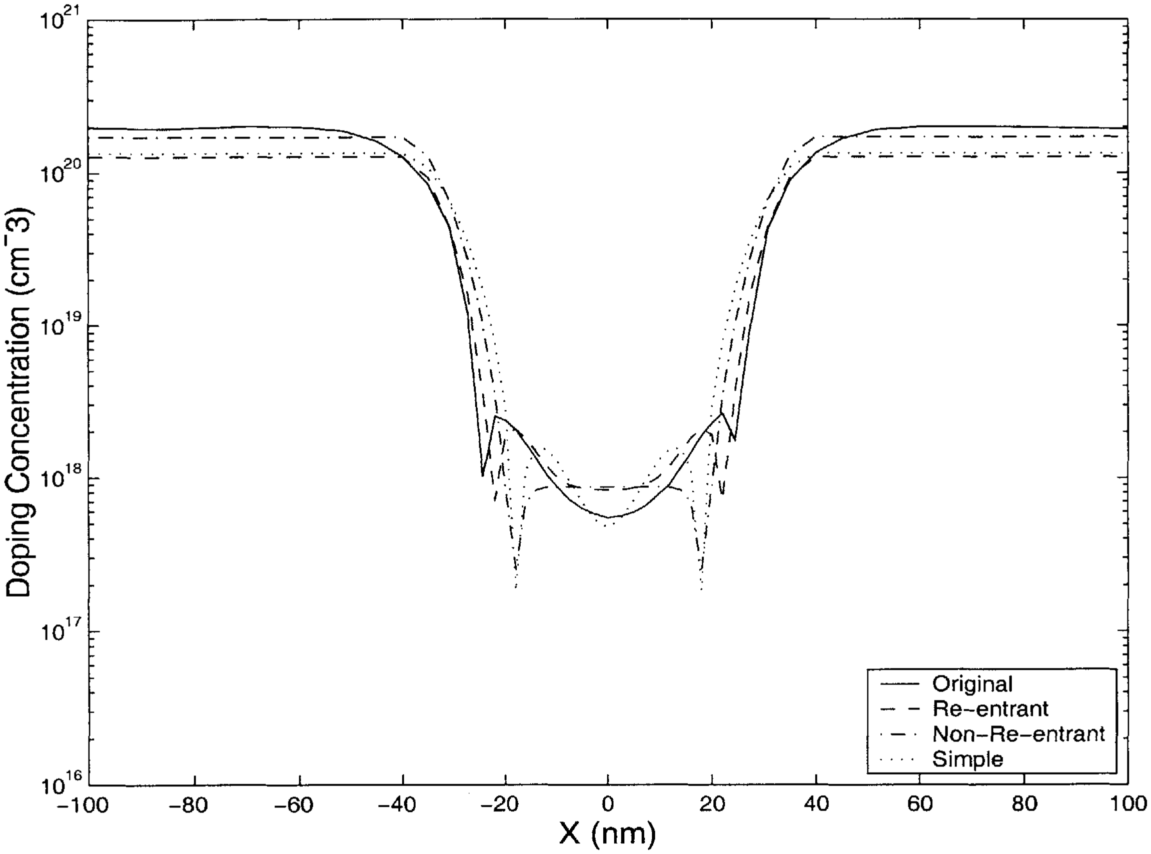

Figure 2.9: Comparison of the original SUPREM profile to the "re-entrant", "non-re- entrant", and "simple" doping representations: lateral profiles at depth Y = 0 nm show that re-entry is necessary to match the L_{eff}.

To evaluate the uniqueness of inverse modeling solutions, a virtual nFET that exhibits the super-halo characteristic [26] with 50 nm Leff and 2 nm physical t_{ox} was generated in TSUPREM4 using Monte Carlo implantation and a RTA with a transient enhanced diffusion model. The doping profiles for this virtual symmetrical device are quite complex, exhibiting re-entrant and box-like S/D features with halos spiking prominently at the surface and washing together deeper in the channel, being quite far from simple gaussians. In this approach, knowing the original 2-D doping allows for a qualification of the inverse modeling. Three inverse modeling representations of the virtual device of decreasing complexity were used: "re-entrant" (a 2-D gaussian for each S/D and halo with peak at depth Y > 0, plus a 1-D gaussian background), "non-re-entrant" (same but with peak at Y = 0), and "simple" (uses the non-re-entrant 2-D gaussians only). Fig. 2.9 and Fig. 2.10 compare the original with the extracted profiles at two depths. The corresponding converged log(I)-V RMS error is 0.01, 0.02, and 0.12; futhermore, the C-V RMS error is 0.003, 0.010, and 0.019 fF/μm, respectively.

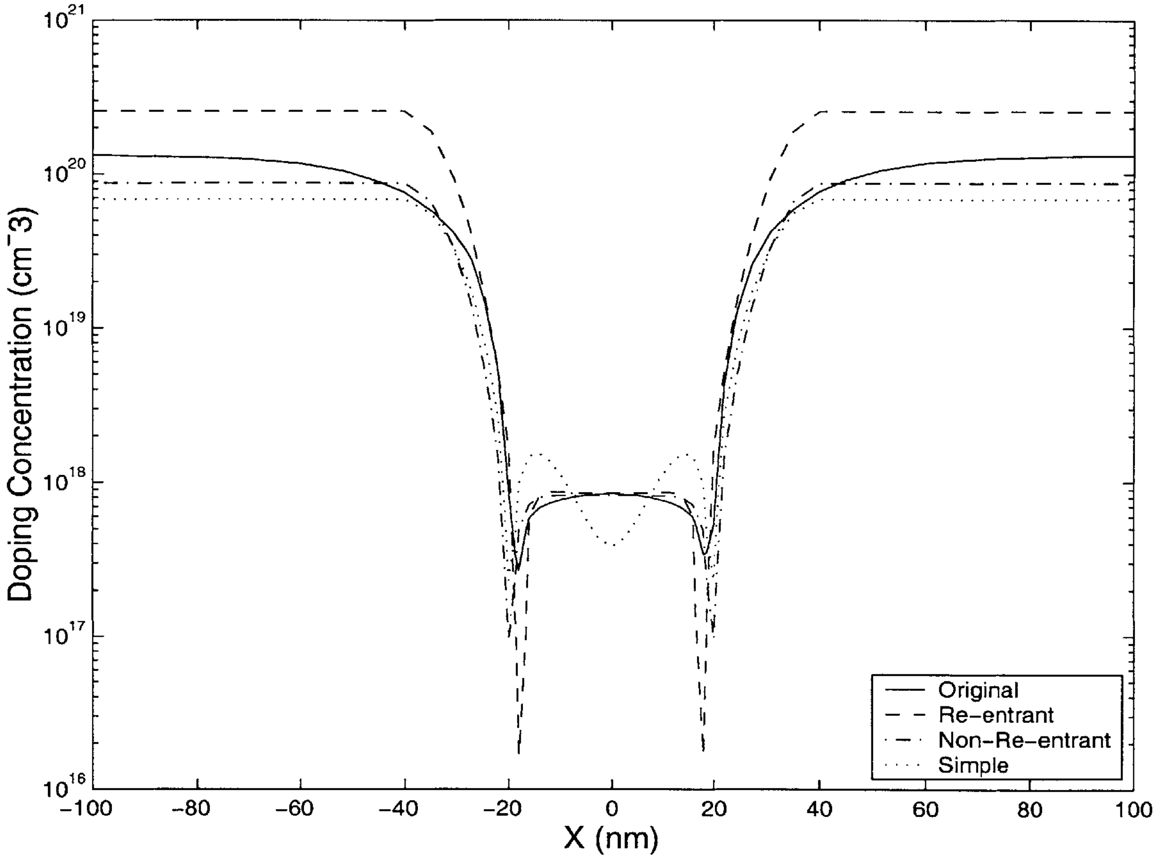

Figure 2.10: Comparison of the original SUPREM profile to the "re-entrant", "non-re- entrant", and "simple" doping representations: lateral profiles at Y = 20 nm show that the "simple" profile is not complex enough to capture the doping pile-up in the channel (and hence has more error in its fit to the electrical data).

For each parameterization, the solution converges as close as possible to the real doping. Evidently, the better the parameterization, the better the fit to data (i.e., the smaller the RMS error), and the better the fit to doping. This main conclusion regarding the uniqueness and ideality of possible parameterizations can be phrased independently of which electrical data sets are used in the optimization loop. The "re-entrant" profile is needed to match the longer channel length at the surface, and agrees to better than a factor of two with the original profile except at the peak doping where the re-entry has diminished the peak level. While the log(I)-V data has tried to cause the representation functions to match the L_{eff} at all depths, the inverse modeling has made a compromise by maintaining a S/D peak doping that fits the C-V data as close as possible. While even the "non-re-entrant" doping fits the original after a certain depth, the "simple" representation cannot account for the background created by the merging of the halos in the channel.

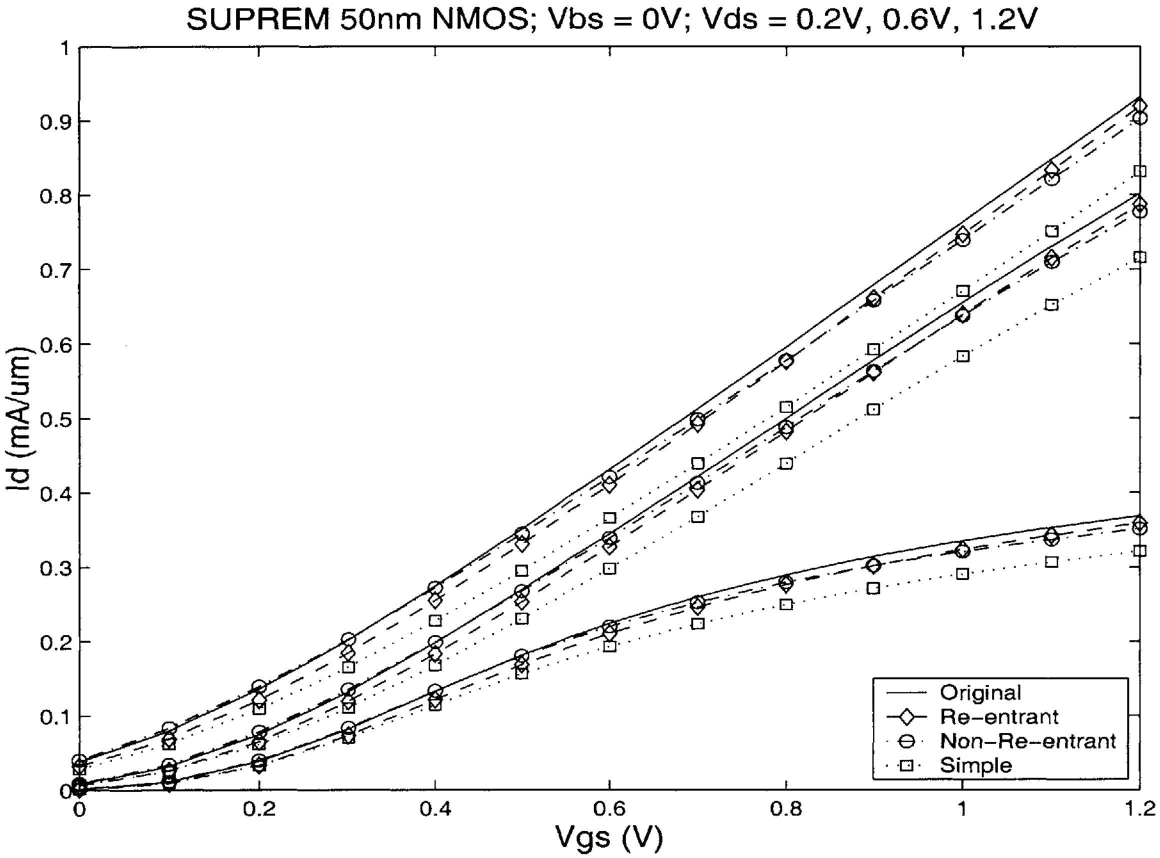

Figure 2.11: Comparison of the strong inversion I-V characteristics for V_{DS} = 0.2 V, 0.6 V, and 1.2 V on the original SUPREM profile, the "re-entrant", "non-re- entrant", and "simple" doping representations. The more complex doping solutions give less than 5% error.

Furthermore, concerns regarding how the differences in doping representation translate into possible errors in the strong inversion data are important if one were to use inverse modeled profiles in transport studies. In Fig. 2.11, the more complex dopings exhibit less than 5% error in the output I-V characteristics while the "simple" profile has about 12% error at several voltages. The important conclusion from this exercise is that the electrical and doping fits improve as the representation functions are given increasing degrees of freedom to approximate the profile. A modified parameterization with lateral S/D extension fall-off as a function of Y and variable peak to capture the L_{eff} and the S/D doping level at all depths would be required. The most general representation function is theoretically a sufficiently fine grid 2-D spline function, but simulation time for the optimization would be prohibitive.

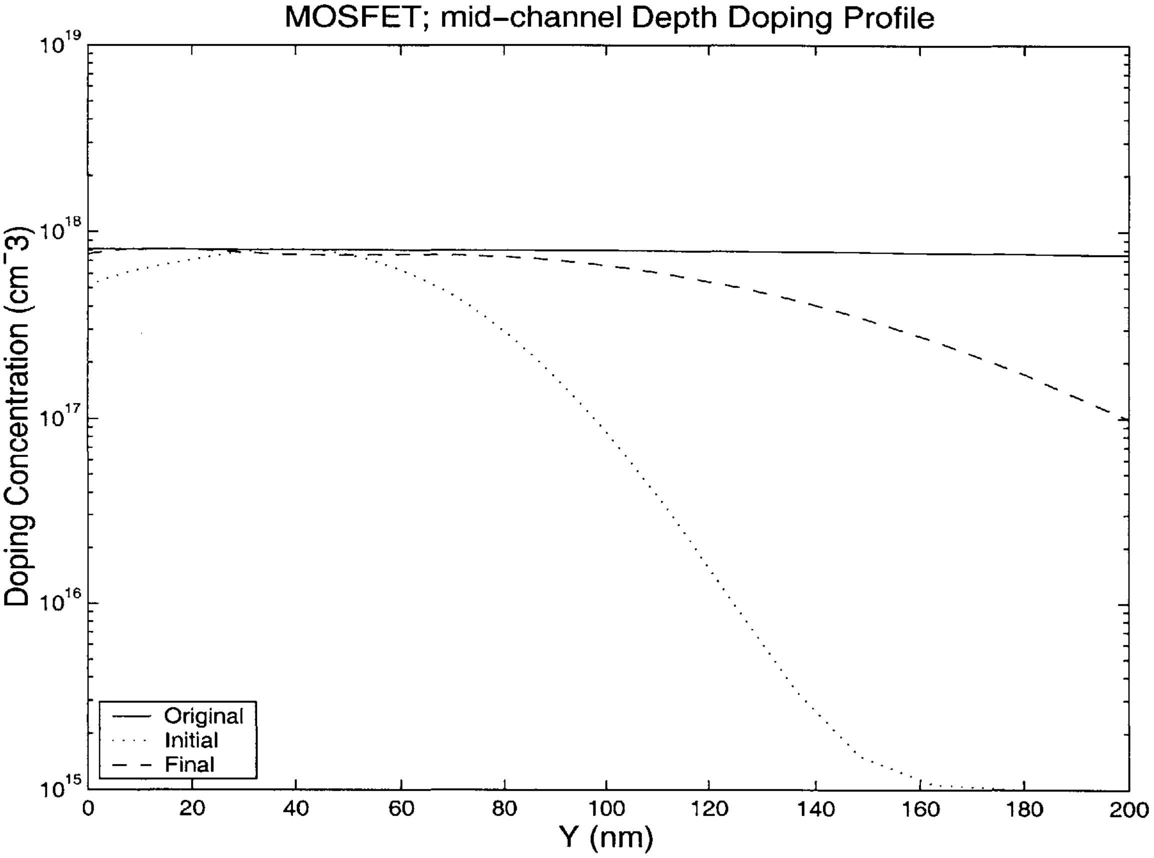

Figure 2.12: Convergence of an "initial" to "final" parameterized (allowing for 2-D S/D, halo, and two 1-D channel doping features) inverse modeling profile to the original 8e17 cm-3 uniform doping of a 90 nm L_{eff} nFET: in the mid-channel depth profile, the "final" profile matches the original below the depletion depth.

One must emphasize that it is not intuitive that simply adding more doping parameters and making the representation more and more general will always lead to a more realistic profile. Will a very detailed parameterization (e.g., "re-entrant") work in the general case of trying to inverse model an unknown profile? Towards answering this question a device with minimal doping features such as uniform channel doping was inverse modeled starting from a complicated "initial" guess (2-D gaussian S/D and halos, plus two 1-D gaussian channel profiles); the results are displayed in Fig. 2.12 and Fig. 2.13. As anticipated for a robust algorithm, the "final" profile converges to the uniform doping within the limit of depletion depth, which defines the region of log(I)-V sensitivity; any extraneous functions (e.g., halos) are suppressed as shown in the lateral profile.

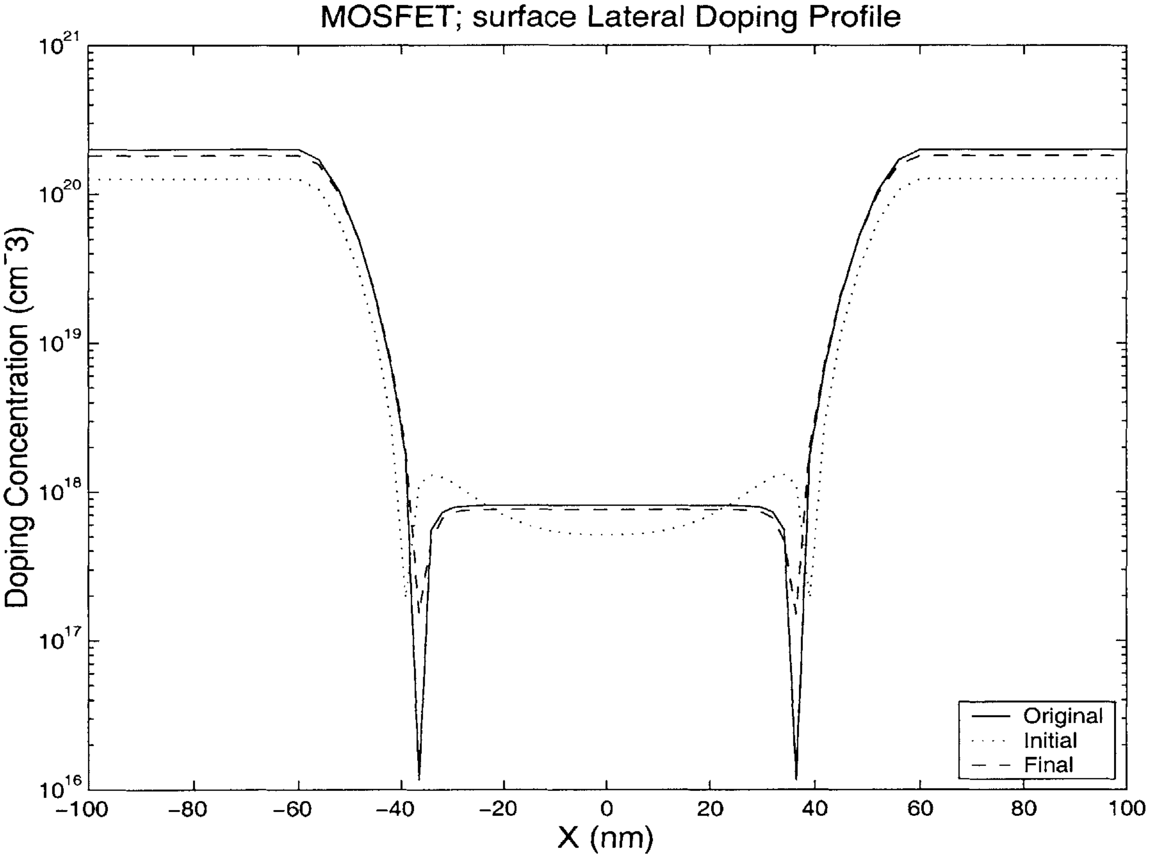

Figure 2.13: Convergence of an "initial" to "final" parameterized (allowing for 2-D S/D, halo, and two 1-D channel doping features) inverse modeling profile to the original 8e17 cm-3 uniform doping of a 90 nm L_{eff} nFET: from the surface lateral profile it is clear that the extraneous halos are suppressed.

In conclusion, the simulations performed in this thesis draws on rigorous computational techniques that describe the device physics and allow for optimization. A broad range of data exhibiting electrical signatures to the actual 2-D features will reliably extract the MOSFET topology given a parameterization with the corresponding doping representation functions.

Chapter 3

Combined I-V and C-V Inverse Modeling

3.1 Comprehensive Methodology

As MOSFETs scale into the sub-100 nm regime, knowledge of the two-dimensional (2-D) doping distribution is critical for accurate device analysis, process calibration, and compact circuit modeling. This work demonstrates and evaluates a comprehensive inverse modeling technique [27] that combines the sensitivities of log(I)-V and C-V data. As discussed, the highlights of employing log(I)-V data are high dependence on both lateral (through V_{DS} variation to bring out short channel effects) and depthwise (through varying V_{BS} and depletion depth) doping features including the S/D to bulk junction with associated doping gradients as well as the non-uniform 2-D channel doping distributions. The 2-D cross-section in Fig. 3.1 plots the edge of the channel depletion region in the off-state as a function of multiple V_{BS} and V_{DS}; the 2-D sensitivity is quite apparent.

The main advantage of adding the C-V data [28] to the optimization is their sensitivity to the physical gate length and to the shape of the S/D overlap region, including detection of the peak doping (through depletion of part of the highly doped extension). With this more coherent and complete methodology, the indirect characterization of 2-D MOSFET topography may prove itself a dominant device engineering tool in the deep sub-100 nm regime.

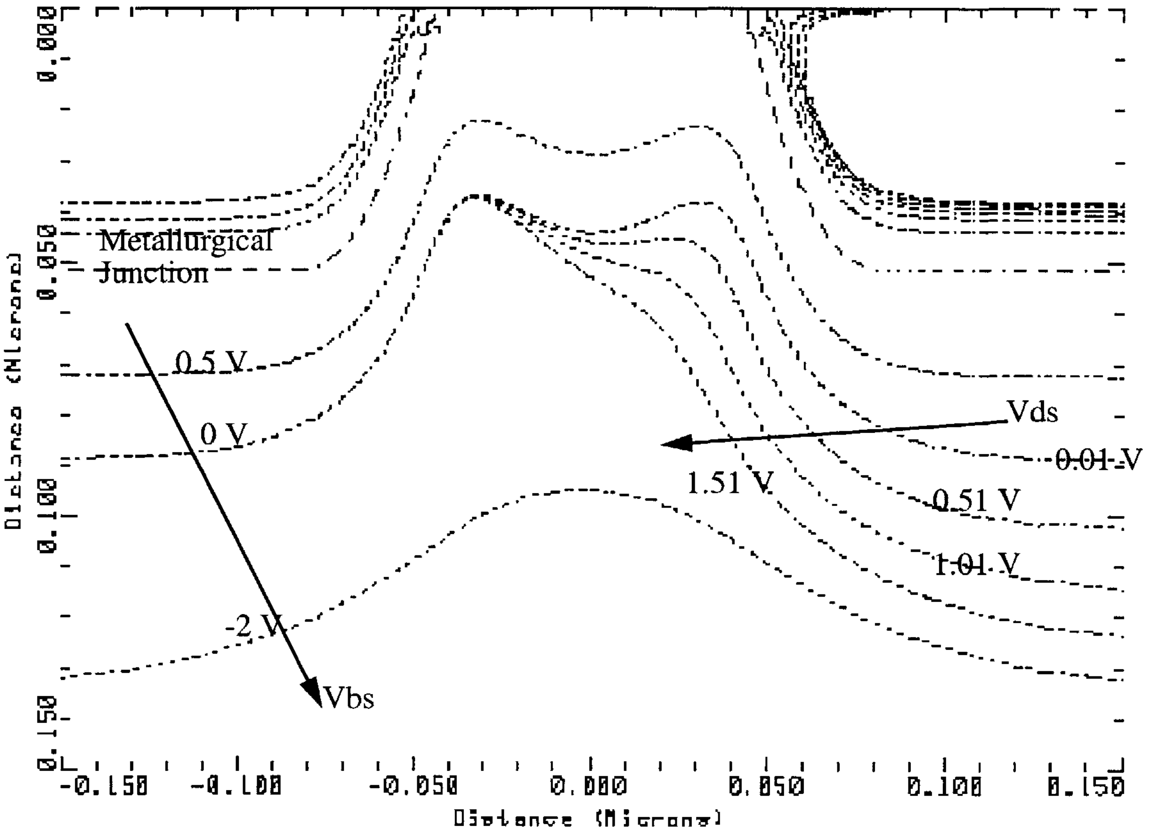

Figure 3.1: Cross-section of a t_{ox} = 3.3 nm NMOSFET with L_{eff} ~ 110 nm depicting I-V sensitivity via movement of the edge of the depletion region at V_{GS} = 0 V in the substrate with increasing junction reverse bias: depthwise as V_{BS} decreases and laterally as V_{DS} increases.

C_{gds} [29] as a function of V_{GS} and V_{BS} (source and drain are shorted in this measurement) is then used to extract S/D extension properties. Although the electrostatics of this measurement are complex in 2-D, it is obvious that both the doping distribution and the physical gate length play a crucial role in the spatial charge configuration. For example, to determine the slope of C_{gds} vs. V_{GS} as V_{GS} becomes increasingly negative (thus depleting the S/D), examine the change in depletion region edge in the schematic of Fig. 3.2; for more positive V_{BS} the decrease of S/D to bulk depletion permits a stronger gate-controlled depletion, resulting in a steeper slope versus V_{GS}. In addition, the effect of the channel accumulation layer as a function of V_{GS} and V_{BS} screening the internal fringing charge determines the magnitude of that component of C_{gds}.

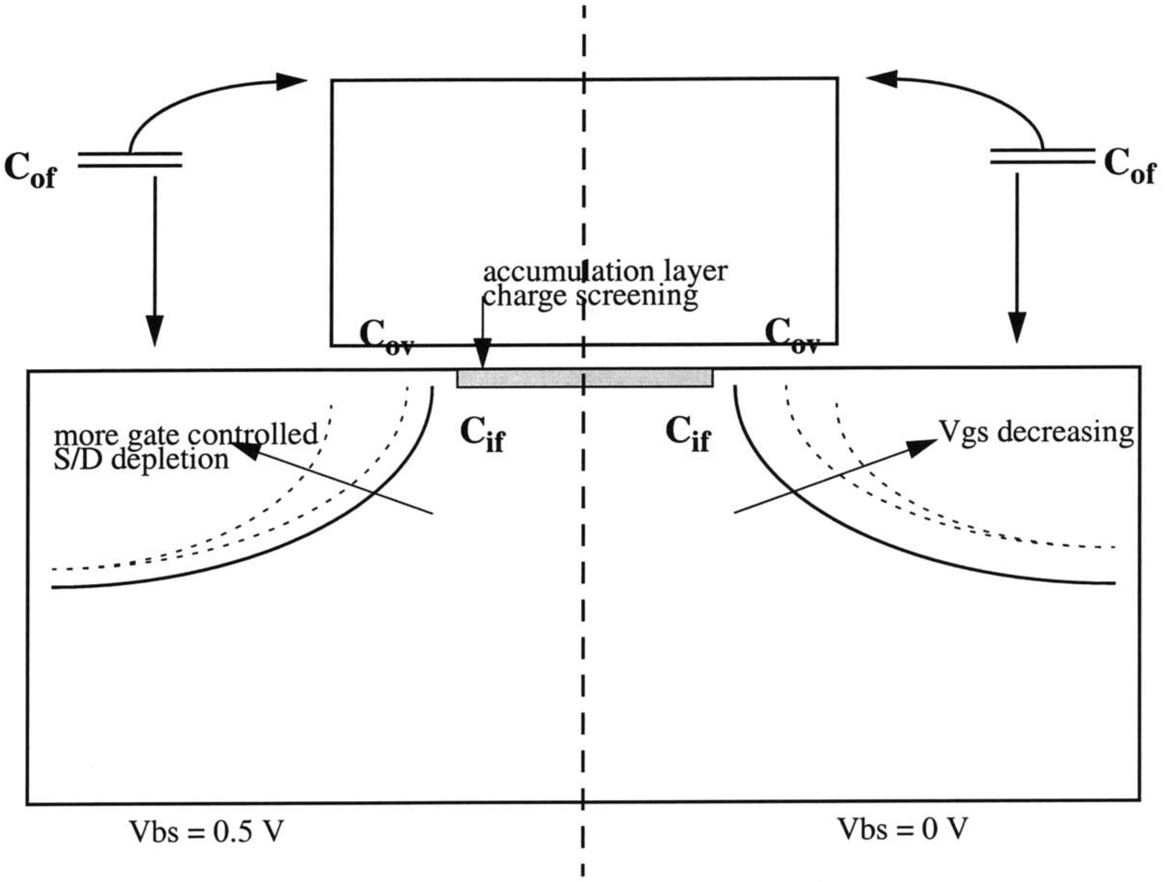

Figure 3.2: Typical MOSFET 2-D cross-section illustrating C_{gds} sensitivity to depletion edge in the source/drain. With applied forward V_{BS} (left half) there is more gate controlled depletion than at $V_{BS} = 0$ V (right half). An accumulation layer will screen the internal fringing capacitance.

Because of the deeply scaled device regime of interest, the issues involved in integrating the C-V data in the methodology have been given careful treatment. For example, the presence of significant gate leakage necessitates a correction to the data before it is used as the target in inverse modeling, which currently does not have a precise and computationally efficient model for gate tunneling. Also, because only 2-D device simulation is performed, which cannot account for a shallow-trench-isolation (STI) edge-FET (which is an inversion layer at the extremities of the MOSFET width that can turn on quickly), a problem would arise if the edge-FET leakage dominated a subthreshold I-V with high V_t; however, this is typically negligible in well engineered modern technologies.

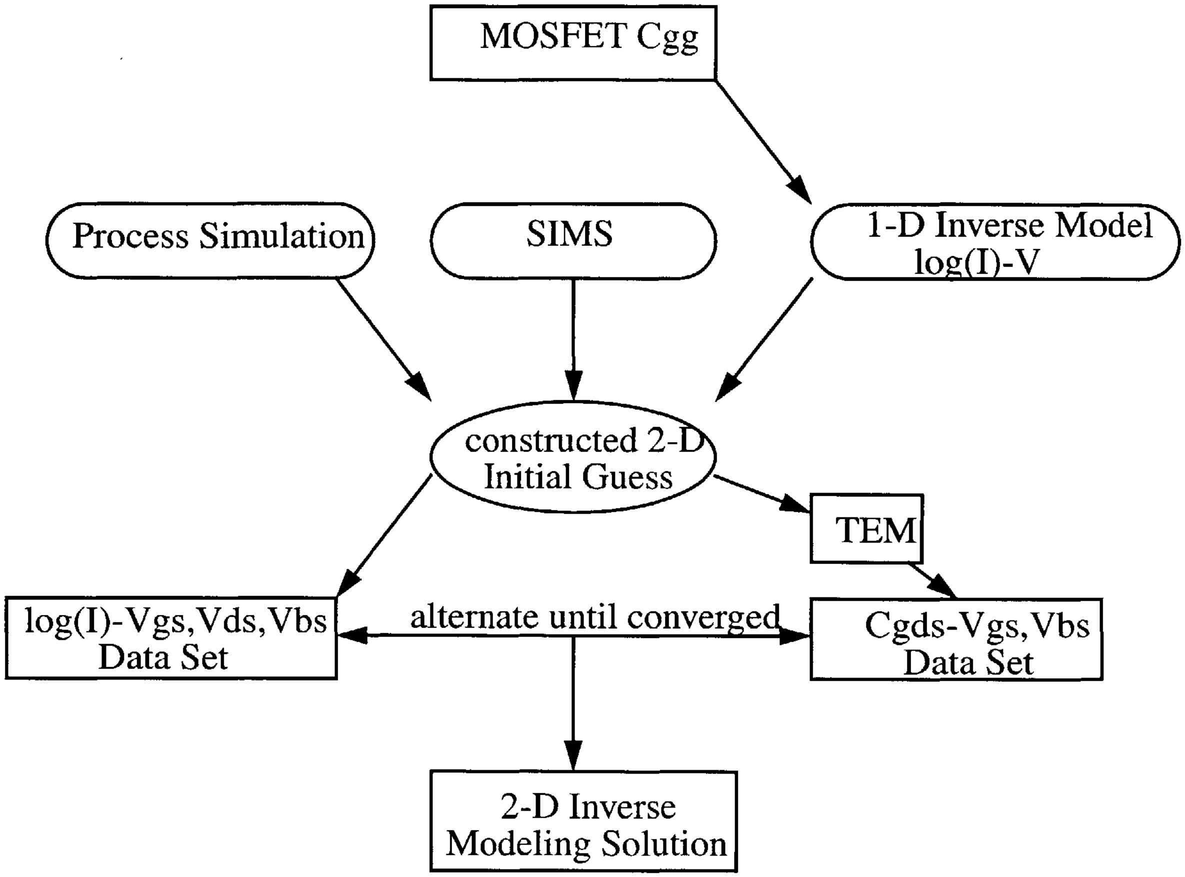

Figure 3.3: A flowchart that delineates the ideal comprehensive inverse modeling methodology as a step-by-step procedure from gate stack analysis to determination of a plausible initial guess to a combined log(I)-V & C-V optimization by alternating between data sets to achieve the final 2-D profile.

Integrating the different data sets at the disposal of the device engineer into a coherent description of MOSFET structure and behavior is like piecing together a puzzle: each type of data should reveal new characteristics for which it is most suited but at the same time must overlap with the other information to make physical sense. The idealized procedure for the inverse modeling method is sketched in the flowchart on Fig. 3.3: beginning with gate stack information, the technique then uses various complementary means to obtain an initial guess to the 2-D structure, i.e., process simulation and 1-D characterization using long-channel MOSFET subthreshold I-V, and SIMS data if available; the optimal solution profile is then acquired through simultaneous inverse modeling using log(I)-V and C_{gds}-V (where the simulation mesh conforms to the device structure as evidenced in cross-sectional TEM). While this flow is the best way to perform inverse modeling in that the extracted profiles are checked against a lot of experimental data, the methodology is flexible enough to omit process simulation and SIMS (as done in the results section).

Because the electrical characteristics of modern MOSFETs depend heavily on their gate stack configuration, it is logical to first obtain numbers for gate dielectric thickness, t_{ox}, and gate electrode doping. Starting with a full measured $C_{gg} = \frac{\partial Q_G}{\partial V_G}$ curve (corrected for parasitic current leakage if needed) and invoking appropriate quantum mechanical models to account for carrier quantization in the channel, a physical t_{ox} is extracted by matching a simulated electrical tox to the C_{gg} in accumulation and using an appropriate dielectric permitivity. Next, the polysilicon depletion [30] decrease in C_{ox} at higher inversion bias is used to extract an assumed uniform active polysilicon doping value by comparing the simulated long-channel MOSFET to the measurements. Hence, the poly-depletion and quantum effects for both NMOS and PMOS devices are accounted for in the same manner.

Due to the non-linear dependence of the device electrostatics on a specific 2-D distribution, the inverse modeling optimization technique can be sensitive to the initial guess of the doping parameterization. Therefore, the engineer should generate a simulated profile that is in the "ball park": one that exhibits the major doping features such that the optimization will not get stuck on a parameterization with high error corresponding to a local minimum. This resulting profile can be acquired by inserting the relevant impurity implantation and diffusion/heat steps of the device process traveller into a standard 2-D process simulator (e.g., SUPREM [31]); this discretized profile must then be converted to the analytical profiles used in inverse modeling by hand or via optimization. If the information necessary to perform this step is incomplete, then an educated guess using analytic formula should be made.

3.2 I-V and C-V Optimization

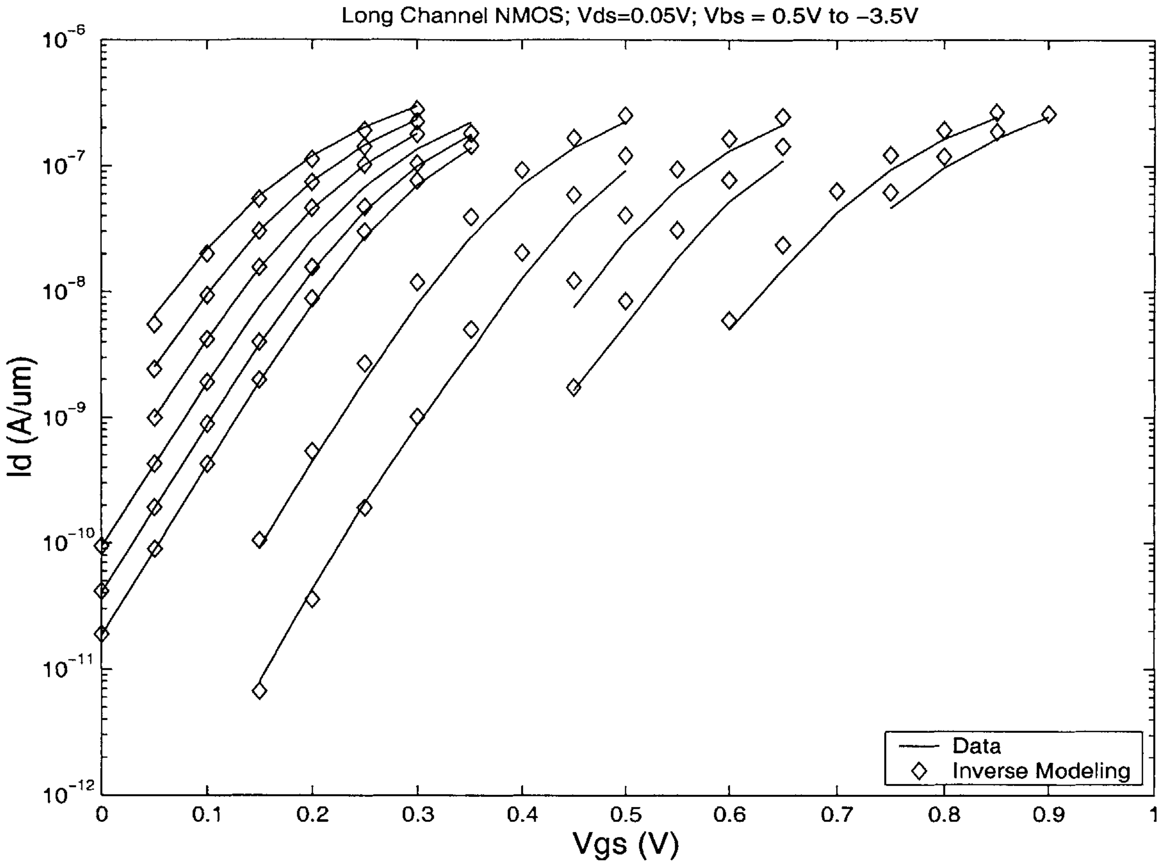

Figure 3.4: Fit from 1-D inverse modeling of long channel subthreshold I-V data for a t_{ox} = 3.3 nm NMOSFET with V_{BS} ranging from 0.5 V down to -3.5 V.

The other initial modeling steps to be undertaken in this methodology involve independently extracting particular 1-D doping profiles and using these as starting values in the final 2-D extraction. Results from 1-D SIMS analysis can be used for depth profiles of purely 1-D doping features in the device such as super-steep retrograde (SSR) implants or S/D extensions. The 1-D channel doping is then verified, or if no direct data are available it is fully extracted by log(I)-V inverse modeling of a long and wide channel device (making an educated guess to the S/D profile which in this case is not critical) with sufficient source/body voltage, V_{BS}, steps to sweep over the widest possible range of the depletion region and hence attain best bulk charge sensitivity. Best results are obtained for V_{BS} ranging from forward S/D diode bias to near reverse bias breakdown; for example, see the fit to data for the device of Fig. 3.4.

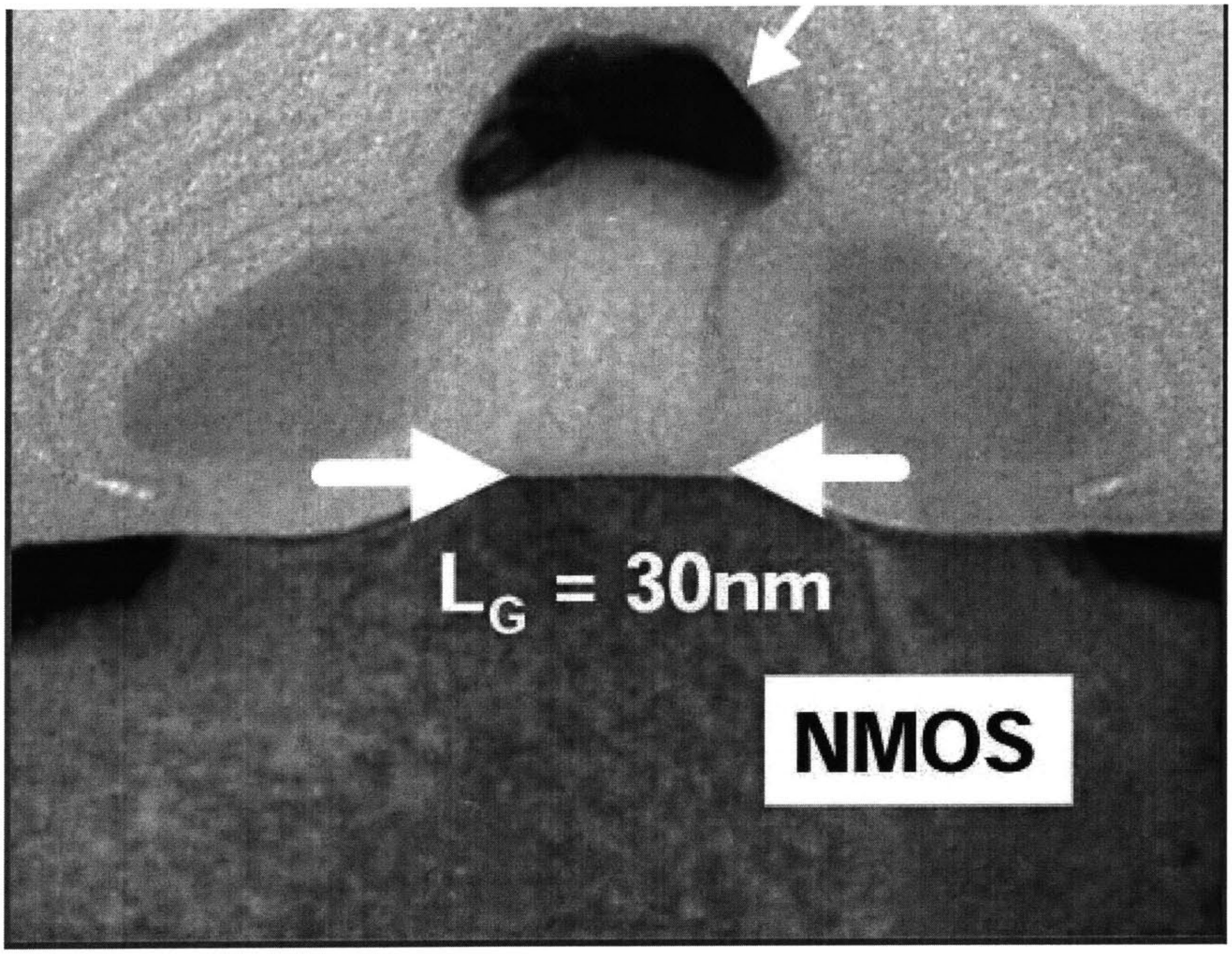

Figure 3.5: An example [32] cross-sectional transmission electron micrograph (TEM) of a L = 30 nm NMOS device fabricated in industry. All non-uniformities such as the configurations of spacer dielectrics and non-planar interfaces must be accounted for in simulations.

Before combining the C_{gds}-V data set with the log(I)-V in the full blown optimization of 2-D doping parameters, many details of the MOSFET structure must be obtained as precisely as possible via cross-sectional TEM analysis such as in Fig. 3.5. Because the topography of the short-channel gate stack including spacers, silicide, etc... contributes significantly (even up to half) of the overlap capacitance, a representative TEM image and knowledge of the dielectrics and their permitivities in the process flow must be used to create a simulation mesh that captures all these features. The requirement for TEM analysis of the gate stack seems to create a drawback to the “indirect”-ness of inverse modeling. However, such analysis is currently routine and it can be safely assumed that all the similarly fabricated gate structures will have similar features, regardless of gate length (which is a parameter in the optimization); hence, only one TEM analysis per technology is sufficient to inverse model an entire family of devices.

In order to optimize the extracted topography, the device must be represented by a set of parameters that describe analytically the doping profile. The most flexible way to construct this parameterization is by superposition of doping representation functions. Appropriate analytical expressions must be both versatile enough to apply to any doping feature (i.e., implant, doping pile-up, etc.) and able to closely track the shape of exponentially changing distributions. Experience shows that for inverse modeling of MOSFETs it is sufficient to utilize one gaussian per lateral and depthwise doping feature: typically, a 2-D gaussian for each of the symmetric S/D extensions and halos, and 1-D gaussians for the channel and well implants. The parameters that are varied include the lateral and depth placement of the peak of the S/D and halo dopings, the peak values, and their associated characteristic fall-off lengths. As long as multiple functions are used to describe the multiple doping features, these functions will accurately describe both NMOS and PMOS devices. The initial guess for the 2-D distribution is formed by selecting values of these parameters that result in profiles that correspond to the process simulation and intuition. The initial guess must also include the 1-D inverse modeling profiles by fixing them in the 2-D device structure. In short devices where point defect densities from the S/D regions may enhance the diffusion of the 1-D channel doping, then the 1-D gaussians will also be varied in the 2-D inverse modeling.

In the optimization loop each parameter is varied by a small fraction (usually 0.5 to 5 percent) while keeping the others at their values at the start of the iteration. Each slightly altered 2-D profiles is fed into a device simulator and the desired simulated electrical characteristics are obtained. Next, the error between the simulated and experimental data is calculated: for C-V, an absolute error is used; for log(I)-V, the difference of log(I) is used. Absolute error was chosen as opposed to relative error in order to treat the absolute shifts in electrical data at the various bias points with the same statistical significance. Thus, the change in error with respect to change in each parameter is tabulated and used to solve (e.g., using the Levenburg-Marquardt algorithm) for the update in parameter space that will minimize the RMS error. This entire sequence has been automated in software.

Finally, the 2-D doping distribution of short channel devices is obtained by implementing a standard optimization loop on the most broad range of available electrical data (i.e., log(I)-V and C-V). The total simulation time is dominated by the time to do a bias sweep after changing each parameter. The parameters are updated to achieve the best fit. It has been observed that alternating between small numbers of iterations fitting log(I)-V and then C-V data provides a fast and accurate convergence to the final profile. On the other hand, trying to fit the data sets simultaneously can sometimes hinder convergence since the data sets may suggest opposing search directions for a given parameter. This effect is more pronounced for very short devices; for example, with the greater fractional mismatch between physical gate and effective channel length, L_{eff}, the C-V data might pull the S/D edge out to fit the gate capacitance while the log(I)-V data might push the S/D in to fit the short channel behavior. In this case, alternating between the data sets allows the I-V to set the L_{eff} and then the C-V iterations pull in the physical gate to get the right overlap.

3.3 Inverse Modeling Results

In order to demonstrate the effectiveness of the aforementioned inverse modeling methodology, case studies are presented. In each instance, the idealized procedure is followed to the extent possible given the available data, thus validating the modeling flow. First, the new technique has been applied to a recent NMOS generation utilizing a Levenburg-Marquardt non-linear optimizer and a standard device simulator [33]. The device models used include Drift-Diffusion transport with approximate Fermi-Dirac statistics and bandgap narrowing. A generalized mobility model that accounts for impurity, phonon, and surface scattering is used but its accuracy has minimal impact on the simulation of the chosen data sets. While simultaneously accounting for poly depletion and quantum mechanical effects [34], the C_{gg}-V curve is used to characterize the gate stack, giving a physical t_{ox} of 3.3 nm. Next, with neither information on the process steps nor SIMS results available, the 1-D channel doping for this device family is determined to be a simple well with a concentration of about 2×10^{17} cm^{-3} by measuring long channel I-V behavior.

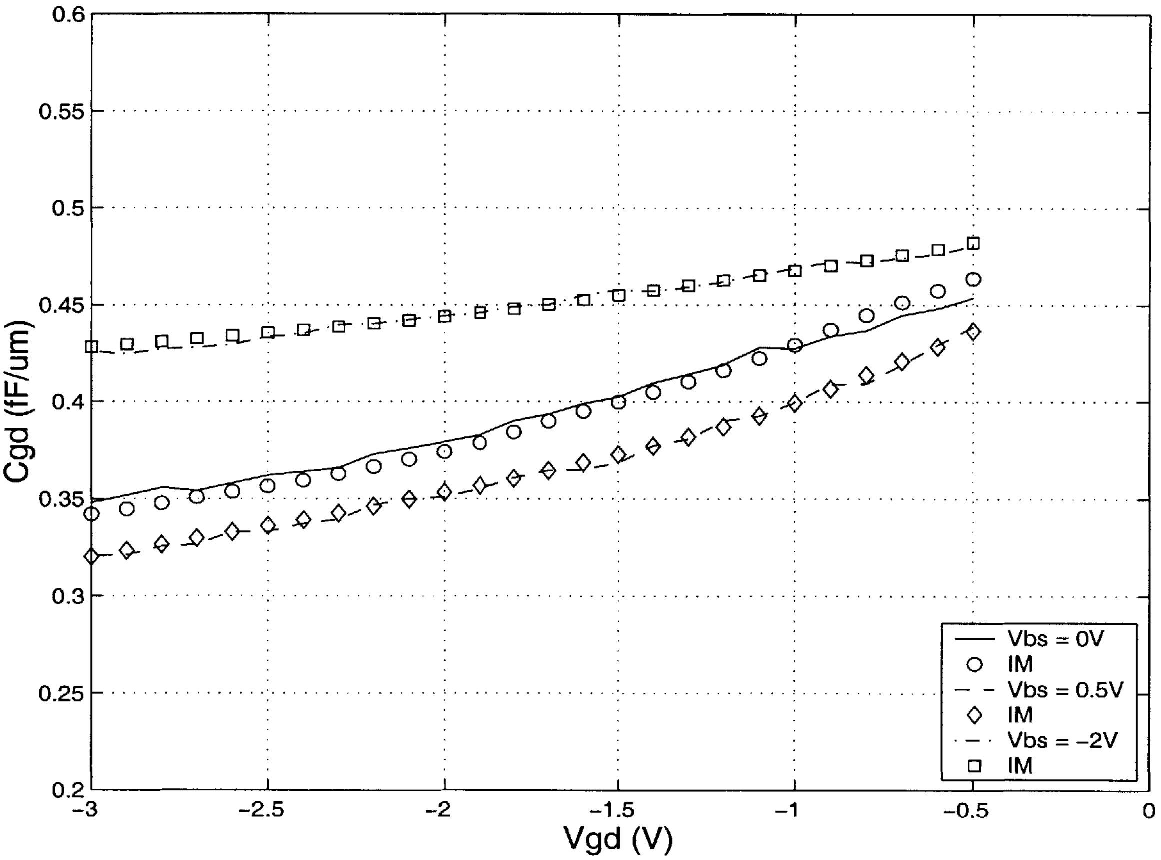

Figure 3.6: Inverse modeling fit of the shortest t_{ox} = 3.3 nm device to careful C_{gds}-V measurements taken at 800 kHz and averaged between four samples per point for varying V_{BS} = 0.5 V, 0 V, and -2 V. The overlap capacitance data for this technology was taken from the stand-alone L = 1 μm device. The error is within the noise level of 0.025 fF/μm.

The 2-D doping profile is represented by the sum of one 2-D gaussian each for the S/D and the halo profiles and their parameters are optimized by alternating more than five times between three independent iterations using the C-V and then the I-V data set. An extended simulation mesh that takes into consideration the industry devices' gate stack, dielectric and spacers from cross-sectional TEM fits the experimental C_{gd} of Fig. 3.6 to within the equipment noise range of

0.025 fF/μm. For this technology, there is a 10 nm thick "L"-shaped layer of oxide running down the sides of the polysilicon gate and extending about 50 nm along the Si/SiO$_2$ interface; nested within the "L" is the nitride spacer. These C-V measurements came from an L = 1 μm device because it was the only device with independent gate contact. It is assumed that the overlap capacitance values taken at least 0.5 V below the onset of inversion are mainly sensitive to the S/D which matches shorter devices. For more negative V_{BS} the gate has less control over C_{gds}, while the magnitude is affected by the onset of accumulation layer screening of the inner fringing.

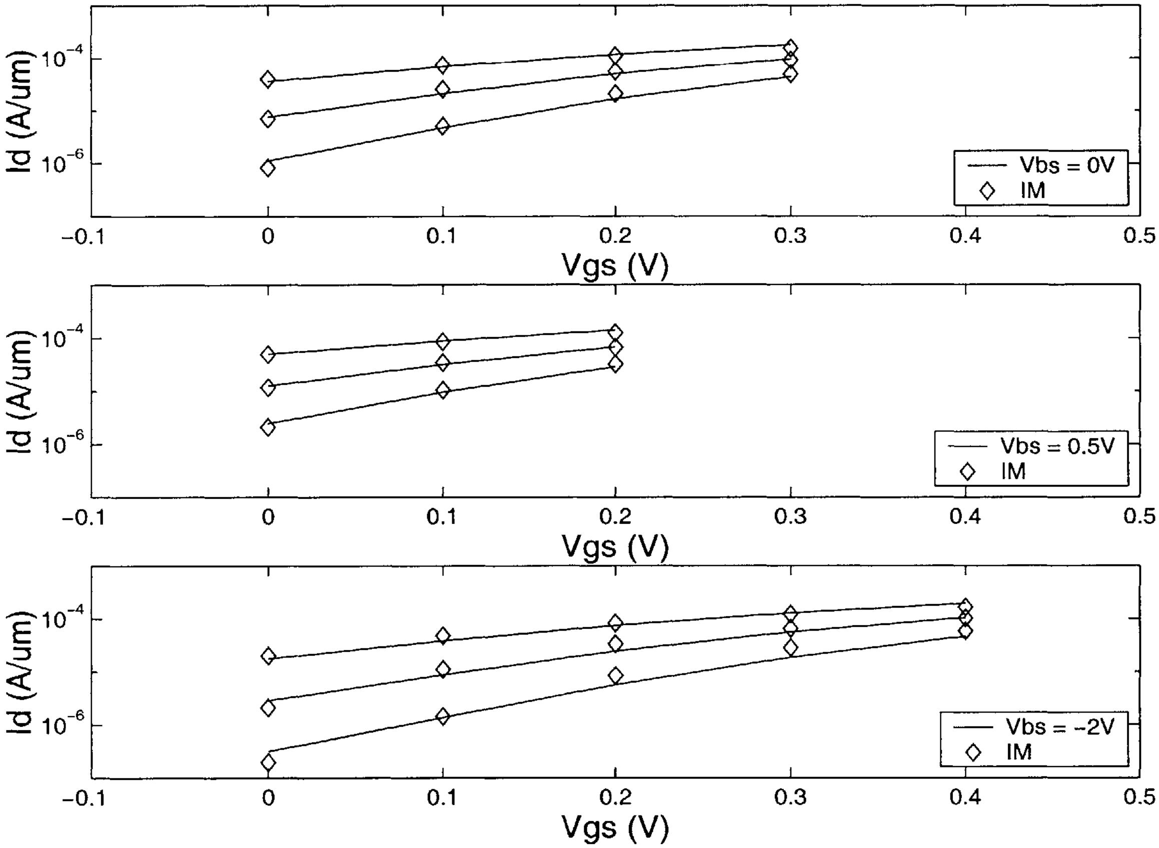

Figure 3.7: Inverse modeling fit of the shortest t_{ox} = 3.3 nm device to subthreshold I-V data for various biases (V_{DS} = 0.21 V, 0.61 V, 1.21 V) at V_{BS} = 0.5 V, 0 V, and -2 V. The better than 0.08 relative RMS error indicates a converged solution to the 2-D doping profile.

On the other hand, the RMS error is 0.08 for a broad range of log(I)-V data as depicted in Fig. 3.7; varying V_{DS} reveals DIBL and threshold voltage (V_t) roll-off, giving lateral sensitivity to the 2-D doping, while varying V_{BS} sweeps the doping dependent depletion depth.

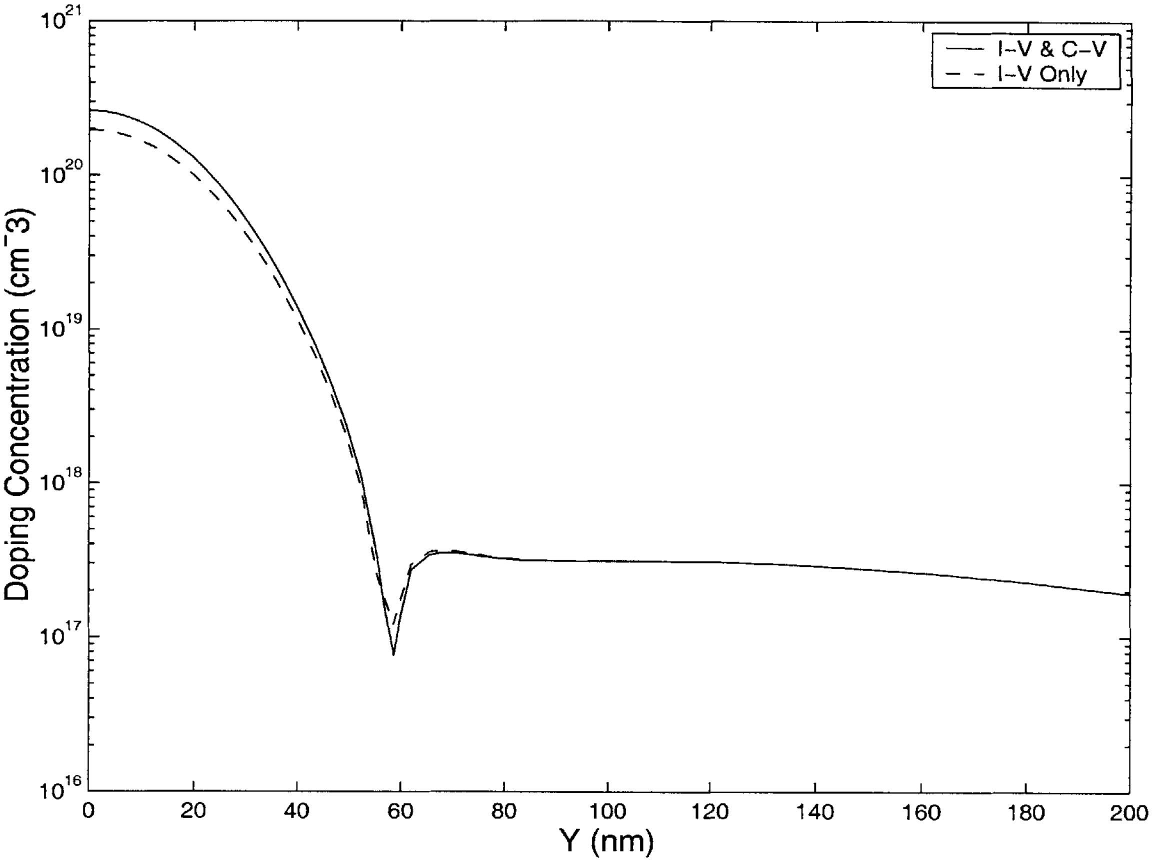

Figure 3.8: Comparison of the depth doping profile extracted at the gate edge for the shortest t_{ox} = 3.3 nm device using "I-V & C-V" data versus "I-V Only" data. The methods give roughly the same junction depth but the S/D peak doping in "I-V Only" was arbitrarily set to 1×10^{20} cm^{-3}.

How has utilizing the new combined log(I)-V and C-V technique [35] improved the converged solutions? First, consider the resulting depthwise cross-section of the S/D extension of the shortest t_{ox} = 3.3 nm device shown in Fig. 3.8. When comparing the inverse modeling technique using log(I)-V and C-V data versus using log(I)-V only data, the junction depths are observed to match due to strong I-V dependence. The primary advantage of C-V is the determination of the S/D peak doping (which was arbitrarily set to 1×10^{20} cm^{-3} for I-V only), while at the same time the overall confidence in the profile is increased due to agreement with another data set.

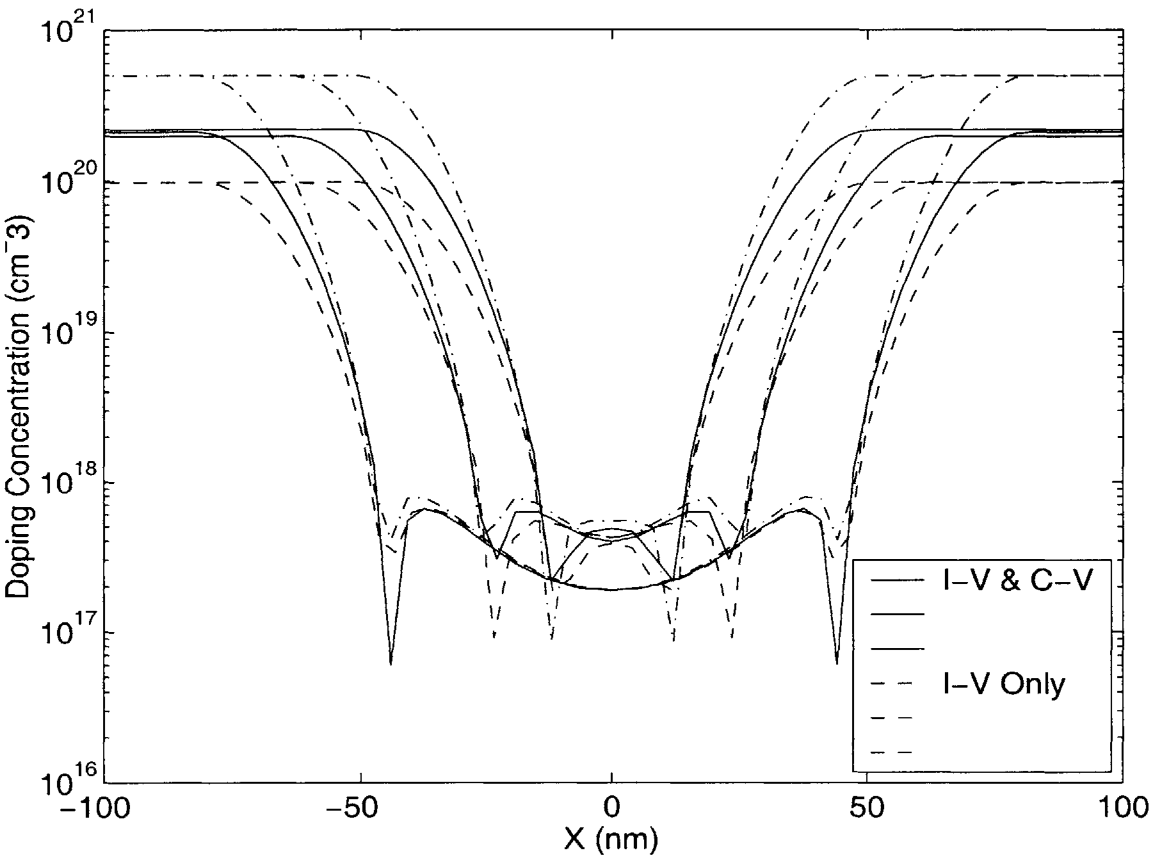

Figure 3.9: Comparison of the lateral doping profiles at the Si/SiO$_2$ surface for several t_{ox} = 3.3 nm devices with extracted physical L_{gate} = 100 nm, 130 nm, 160 nm using "I-V & C-V" data versus "I-V Only" data. The log(I)-V data provides sensitivity especially in the channel region while the addition of C-V data determines the S/D peak doping (which has two arbitrary settings of 1×10^{20} cm^{-3} and 5×10^{20} cm^{-3} for "I-V Only") and slope.

Now examine the lateral profiles for several short channel devices in Fig. 3.9. Utilizing the C-V data detects S/D extension under-diffusion and fall-off steepness while S/D concentrations on the order of 1×10^{20} cm^{-3} are depleted with negative gate bias. Furthermore, in the I-V & C-V run a physical gate length, L_{gate}, is extracted for each MOSFET and hence the overlap characteristic provides enhanced sensitivity to the S/D peak doping; however, this level is fairly arbitrary past the gate edge where the capacitance sensitivity is diminished. Since the method with I-V only relies on an arbitrary pre-fixed value of the S/D peak, it is easy to understand the significant errors at high doping in Fig. 3.9. Also, it is apparent that when including C-V the inverse modeling optimizer adjusts the halo channel doping slightly to maintain the proper V_t and subthreshold characteristics. These halo profiles are also expected to be more reliable.

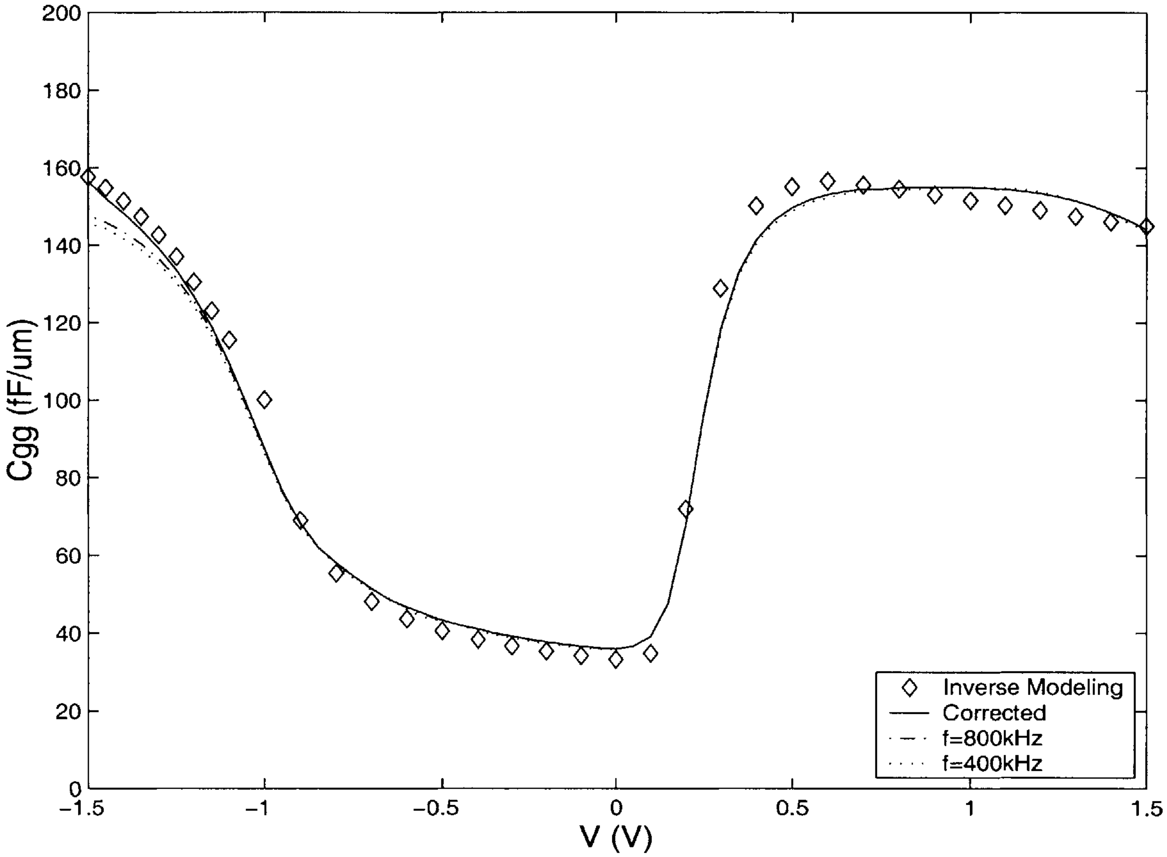

Figure 3.10: Full C_{gg} characteristic extracting t_{ox} = 1.5 nm on this L = 10 μm device from an advanced NMOSFET technology. Due to leakage parasitic resistance, this C-V must be corrected [15] using two frequencies, here 800 kHz and 400 kHz. The fit exhibits good gate stack modeling of QM and polysilicon depletion effects.

In order to push this new methodology closer to its limits, another advanced NMOS family with thin oxide was employed. Starting by measuring the C_{gg} vs. V_{GS} in Fig. 3.10 with an HP4192 impedance analyzer set to the parallel capacitance and resistance model, the gate is swept. For each bias point the gate has a 25 mV small signal excitation voltage at the set frequency (there were sweeps at both 400 kHz and 800 kHz) while the small signal current is read out of the tied source-drain-body. Then correcting for series resistance using multiple high frequencies [36] in the presence of gate leakage extracts a poly doping of 7.9×10^{19} cm^{-3} and a 1.5 nm equivalent SiO_2 physical t_{ox} (assuming $\varepsilon_{ox} = 3.45\times10^{-13}$ F/cm) is extracted which corresponds to the value extracted from TEM. Next, using subthreshold I-V data with V_{BS} swept from 0.5 V to -4 V on a long channel device results in an extracted 1-D SSR and well profile (composed of three 1-D gaussians) with RMS error ~ 0.1.

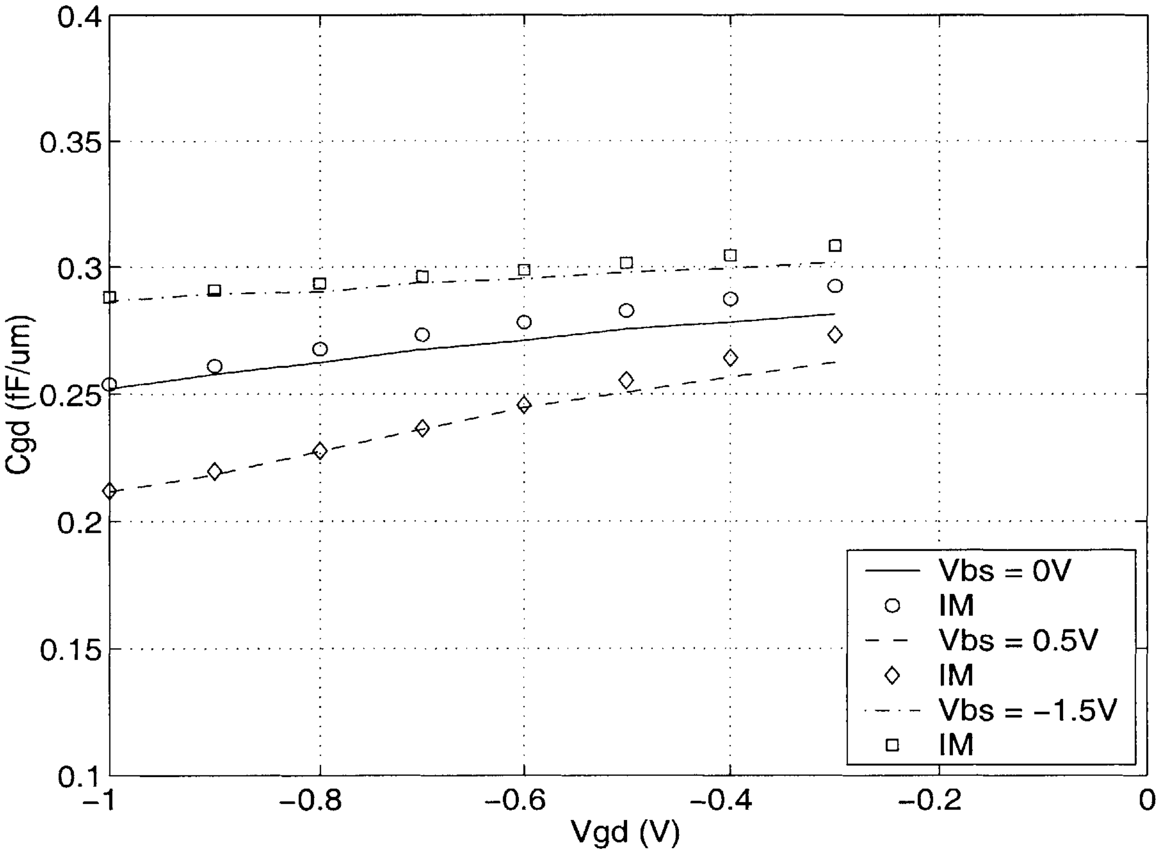

Figure 3.11: Inverse modeling fit of the shortest t_{ox} = 1.5 nm device to careful C_{gds}-V measurements taken at 800 kHz and averaged between four samples per point for varying V_{BS} = 0.5 V, 0 V, and -1.5 V. The error is within the noise level of 0.025 fF/μm.

The thin oxide has sufficient leakage current at large bias (inversion or accumulation) that the measured parasitic resistance limits the sensitivity of the C_{gds}-V measurements; to enhance the accuracy for this short-channel data, a small signal voltage of 100 mV is used along with an average of four measurements per bias point. Finally, alternating between iterations of C-V (RMS error ~ 0.01 fF/um in Fig. 3.11) and log(I)-V, with a broad range of biases measured with an HP4156 semiconductor parameter analyzer (relative RMS error < 0.1), the short channel S/D and halo profiles are extracted.

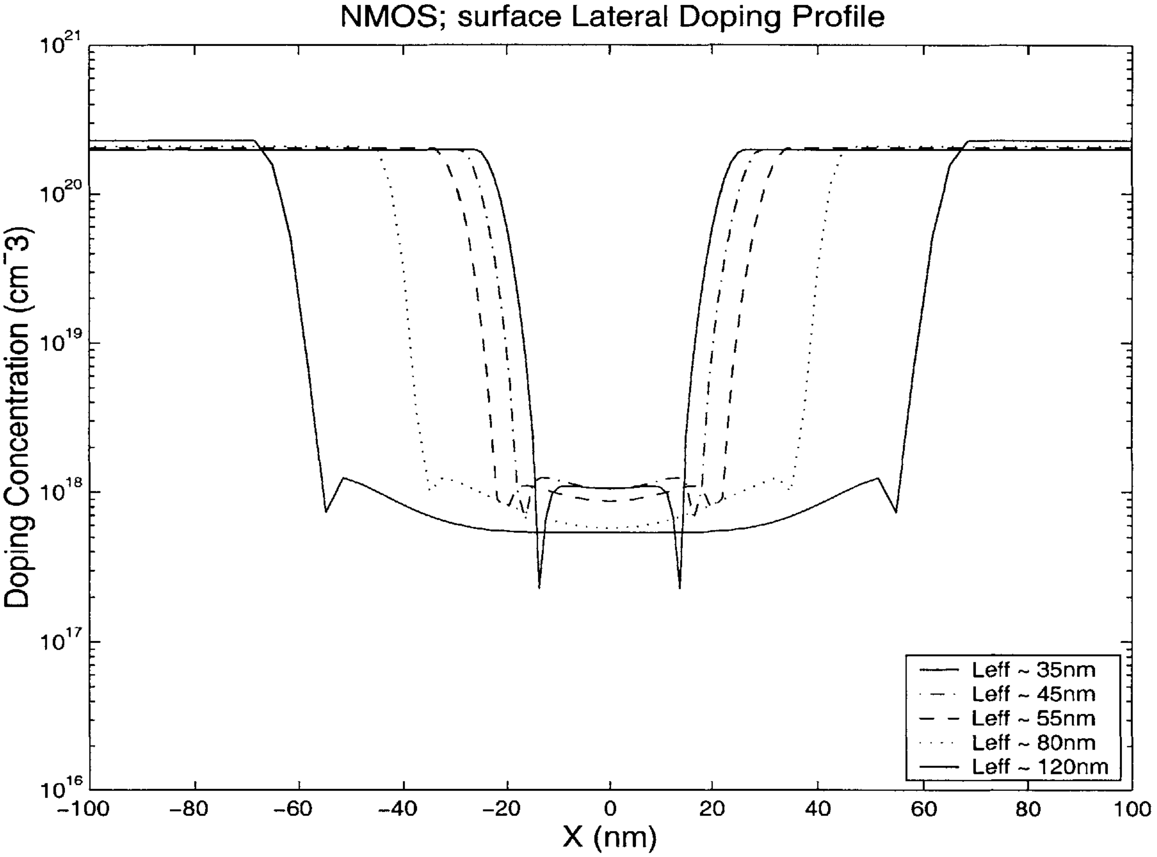

Figure 3.12: Extracted lateral profiles using the combined inverse modeling technique on t_{ox} = 1.5 nm devices; these short channel MOSFETs have L_{eff} ~ 35 nm, 45 nm, 55 nm, 80 nm, and 120 nm. The longer three lengths all fit to independent C_{gds}-V measurements. The shorter two lengths had no C-V data but their S/D peak values were fixed at the value obtained for the longer.

The doping representation functions used, 2-D gaussians, indicate a junction depth, x_j ~ 35 nm for this device technology, in line with proper scaling. To obtain reasonable values of extracted gate overlap for this technology, the MOSFET simulation structure included a 5 nm layer of over-oxidation at the Si/SiO$_2$ interface just beyond the gate edge (effectively decreasing the C_{gds}). Using all these details in the C-V inverse modeling, L_{gate} - L_{eff} = 15 nm is extracted for this technology, where L_{eff} is defined as the distance between S/D extension dopings of 2×10^{19} cm^{-3} [37]. The extracted surface lateral net doping profiles of the NMOS family are plotted in Fig. 3.12, marking a steeper S/D lateral roll-off and a more aggressive halo configuration. The effective increase in doping at mid-channel as L_{eff} shrinks and the halos merge is depicted in Fig. 3.13; this increase is necessary to control device electrostatics and process variation at short channels. Importantly, the combination of log(I)-V and C-V data enhances as well as provides confidence in the accuracy of the S/D profiles.

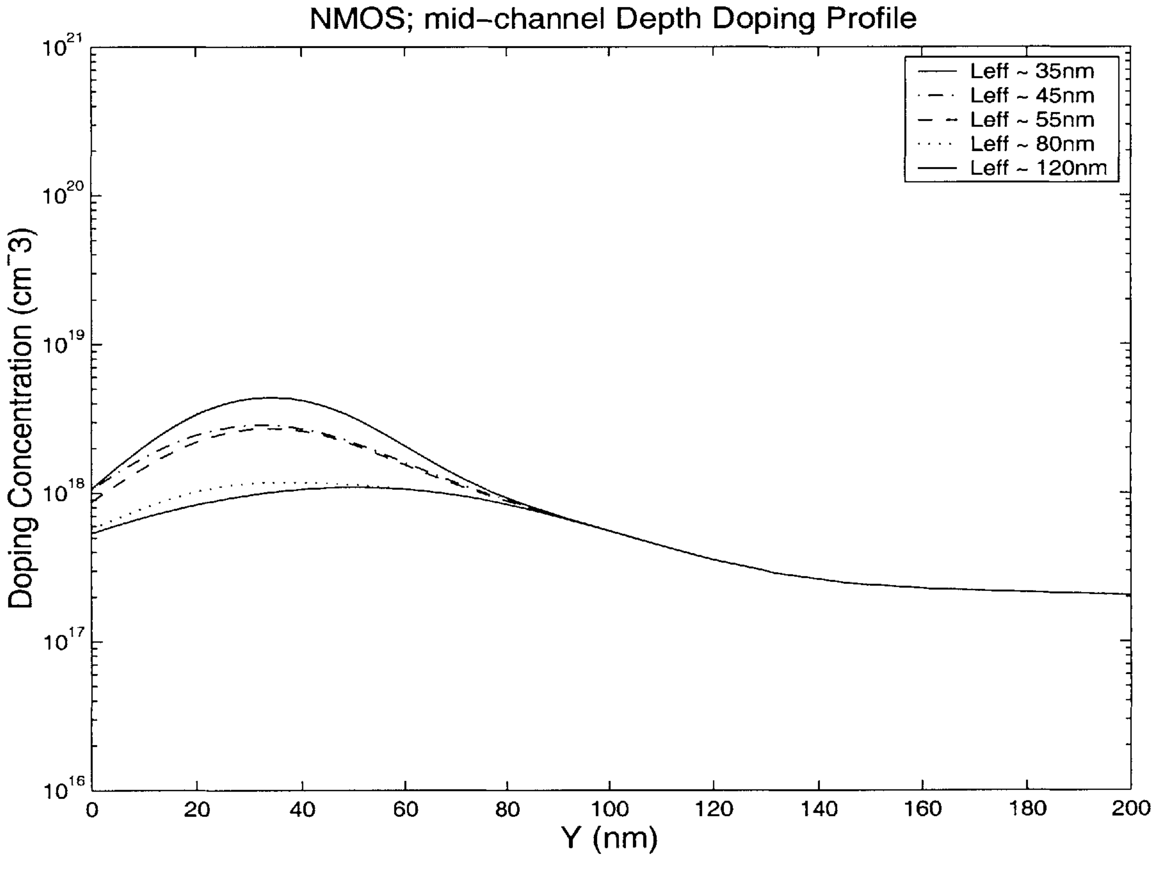

Figure 3.13: Extracted depthwise mid-channel doping profiles of the t_{ox} = 1.5 nm technology devices. The profile approximates the 1-D long channel profile but increases due to merging halos at shorter channels to control short channel effects.

To summarize, this work describes a step-by-step implementation of a comprehensive indirect methodology of inverse modeling to characterize sub-100 nm MOSFETs. After piecing together several levels of experimental device and initial guess information, a formal optimization algorithm provides a 2-D doping representation that best fits the combined log(I)-V & C-V data. The technique must utilize a sufficiently broad range of data (to achieve strong electrical signatures of any doping features) and one parameterized gaussian per doping feature to yield profiles both unique and closely descriptive of the real doping to within experimental tolerance.

3.4 Calibration of Process Simulation

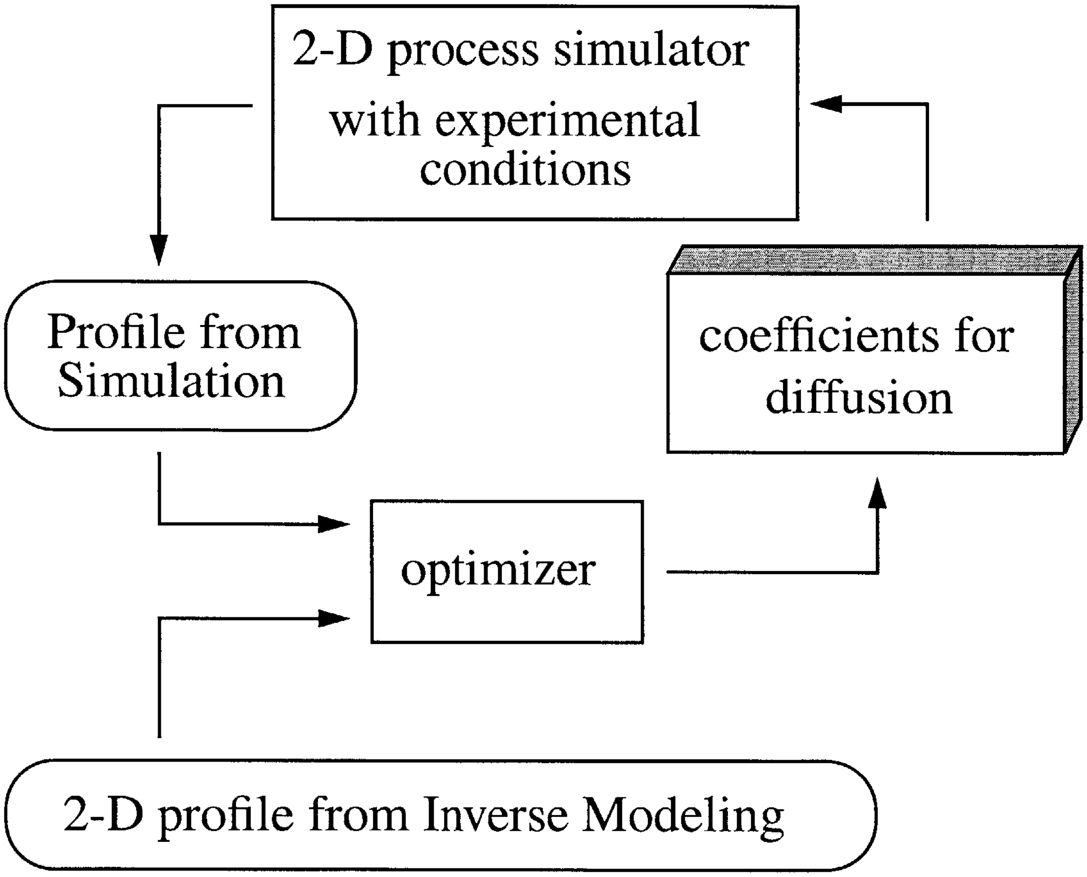
2-D process simulator with experimental conditions
Profile from Simulation
coefficients for diffusion
optimizer
2-D profile from Inverse Modeling

Figure 3.14: Flowchart for optimization of coefficients for processing steps such as diffusion that are simulated with experimental conditions to give a 2-D profile that matches inverse modeling.

A promising extension of the inverse modeling methodology lies in improving the results from standard process simulators. While a process simulation may give a device topology that exhibits the proper features, the exact shape is often poorly captured. Thus, having a way to tune the process model parameters as in Fig. 3.14 is highly beneficial. A standard optimizer is utilized again to find the best match between process and inverse modeled 2-D doping profiles. For instance, the updated diffusion coefficients along with the detailed experimental conditions should yield the real profile.

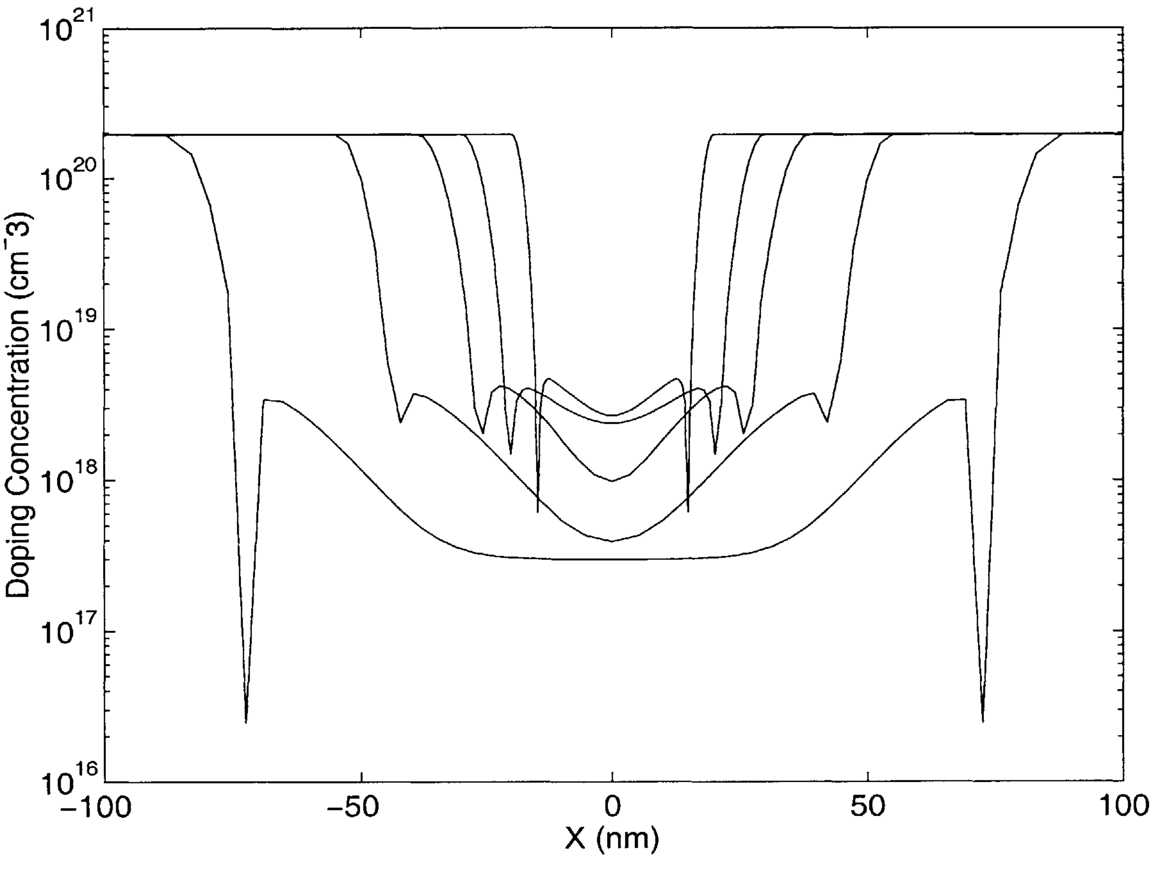

Figure 3.15: Lateral doping profiles at the surface from 2-D inverse modeling of t_{ox} = 1.7 nm NMOSFETs with effective channel lengths of about 30 nm, 45 nm, 60 nm, 95 nm, and 150 nm.

Inverse modeling solutions provide the trustworthy 2-D profile data necessary to tune the models. In the following investigation, the doping profiles [38] of the family of NMOS devices in

Fig. 3.15 with t_{ox} = 1.7 nm have been extracted. Because all the relevant dopant and thermal details of this process technology are known, the simulated profile can only be corrected by adjusting the model parameters of the diffusion equation:

$$\frac{\partial C}{\partial t} = -\nabla \bullet (\vec{J}_i + \vec{J}_v) \qquad \text{Equation 3.1}$$

where the time derivative of concentration of a specific dopant, C, goes as the flux due to interstitial and vacancy point defects.

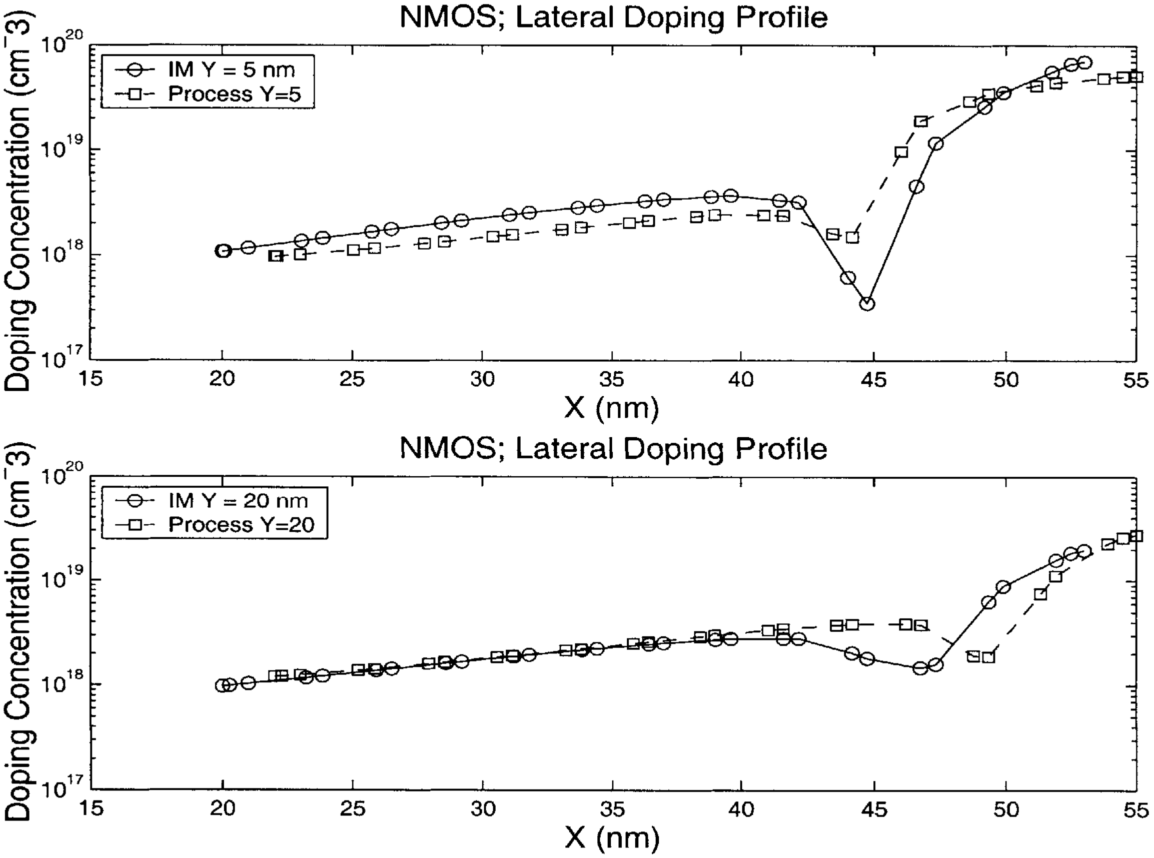

Figure 3.16: Fit of lateral doping profiles of the L = 95 nm device using calibrated "Fermi" point defect diffusion. While decent, the process simulation has some mismatch in metallurgical junction between the surface and Y = 20 nm deep.

In the interests of saving optimization time, somewhat simplistic diffusion models were employed. For example, the "Fermi" model takes under an hour per simulation for the flux

$$\vec{J}_i = factor_i \cdot -D_i\left[\nabla(C_m) - (C_a)Z\frac{q\vec{E}}{kT}\right] \qquad \text{Equation 3.2}$$

where the diffusivity D multiplies the gradient of the mobile carrier concentration C_m and the electric field $\vec{E}$ set up by the active concentration C_a of impurities. The drawback is that there is no dependence on point defect supersaturation; this ratio of defects to their equilibrium concentrations are on the order of 10 to 100. However, this effective value is extracted by allowing the multiplicative diffusivity *factor* to vary. The lateral profile fits in Fig. 3.16 carry factors ~ 40 for the arsenic S/D extension doping and factors ~ 2 for the angled boron halo doping.

A more complex model that accounts for the transient enhanced diffusion (TED) [39] with defect supersaturation (e.g., I/I*) but takes about 6 hours to run per iteration is

$$\vec{J}_i = -D_i\left[\nabla\left(C_m\frac{I}{I*}\right) - \left(C_a\frac{I}{I*}\right)Z\frac{q\vec{E}}{kT}\right] \qquad \text{Equation 3.3}$$

$$D_i = \sum_s (D_{i,s}\eta^{-s}) \qquad \eta = \frac{n}{n_i} \qquad \text{Equation 3.4}$$

$$D_{i,s} = Diffusivity_{i,s} \cdot \exp\left(\frac{-Energy_{i,s}}{kT}\right) \qquad \text{Equation 3.5}$$

The overall diffusivity D_i is a sum of components $D_{i,s}$ for charged diffusion with either interstitials or vacancies. These individual processes still need to have their diffusivities and activation energies associated with each dopant and point defect calibrated. For the simulations displayed in Fig. 3.17, the diffusivity prefactors were used as the fitting parameters and the junctions line up very well. In this case, the prefactors were changed from the SUPREM defaults by 15 times for interstitials and vacancies with As which is sensitive to both, and 0.5 fold for the intersitials with B.

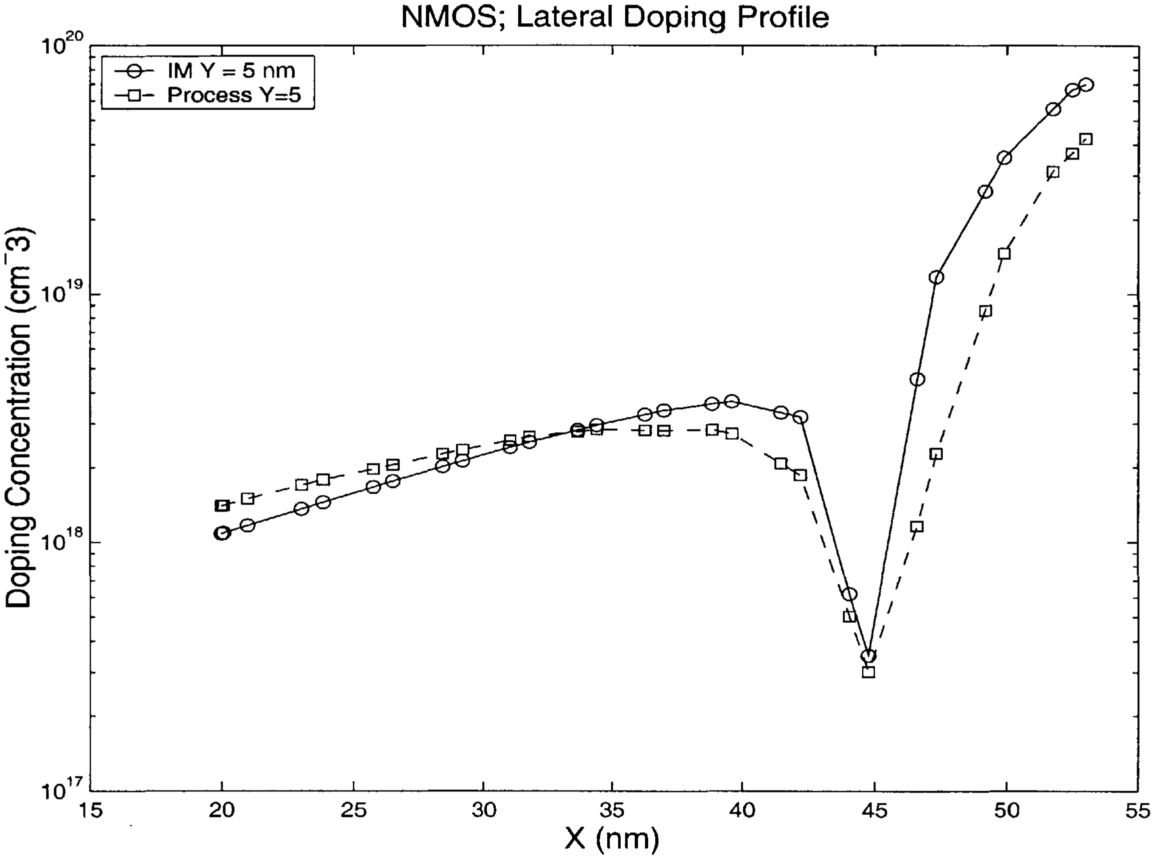

Figure 3.17: Fit of lateral doping profiles of the L_{eff} = 95 nm NMOS device using a calibrated TED diffusion model.

In conclusion, the inverse modeled solutions offer a window into better tuning diffusion parameters to ensure accurate process modeling of modern devices. As expected, using the more physical TED model outperforms a simple effective calibration of the "Fermi" method; however, the more robust model takes longer to converge. While the moderate adjustments to the default parameters give decent fits, they are still only valid for this particular process technology. A universal calibration will require data for dopant distributions annealled at various spike temperatures to resolve adjustments between effective diffusivity and activation energy.

Chapter 4

Transport Model Calibration

4.1 Transport Model Selection

As MOSFETs continue to scale into the sub-50 nm regime [40], it has become a subject of debate [41] as to how long the macroscopic approximations to the Boltzmann transport equation currently employed in popular device simulators would remain valid in describing carrier behavior. Because short channel device transport is highly dependent on the 2-D impurity distribution, which is difficult to quantify, it has been difficult to evaluate quantitatively the various transport models over a wide range of device structures. In this study, the comprehensive inverse modeling scheme obtains the 2-D doping profiles [42] of three advanced MOSFET technologies which are then used as a foundation for calibrating and evaluating transport models, both Drift-Diffusion (DD) [43] and Energy Balance (EB) [44], over a range of devices.

As famous as it is fundamental, the Boltzmann transport equation [45] describes

$$\frac{\partial f}{\partial t} + \bar{v} \cdot \nabla_r f + \frac{q\bar{E}}{h} \cdot \nabla_k f = \left.\frac{\partial f}{\partial t}\right|_{collision} \sim \frac{f(r, k, t) - f_{eq}}{\tau} \qquad \text{Equation 4.1}$$

the microscopic [46] and macroscopic motion of particles with a distribution function f(r, k, t). The sum of the time derivate, velocity $\bar{v}$ dot product with the position (r) gradient, and the phasespace (k) gradient product with the applied field $\bar{E}$ divided by h = Plank's constant / 2π should equal the distribution change due to collisions, which is approximately equal to the deviation from equilibrium f_{eq} over a time constant τ. The macroscopic approximation then takes the first two moments of this equation to arrive at the continuity Eq. 2.2 to 2.3 and the energy balance [47]

$$0 = \nabla \bullet \bar{S}_n - \frac{1}{q}\bar{J}_n \bullet \bar{E} + \frac{3k}{2q}\left[\frac{n}{\tau_{wn}}(T_n - T_0) + \frac{\partial}{\partial t}(nT_n)\right] - H_n = F_{2n} \qquad \text{Equation 4.2}$$

$$0 = \nabla \bullet \bar{S}_p - \frac{1}{q}\bar{J}_p \bullet \bar{E} + \frac{3k}{2q}\left[\frac{p}{\tau_{wp}}(T_p - T_0) + \frac{\partial}{\partial t}(pT_p)\right] - H_p = F_{2p} \qquad \text{Equation 4.3}$$

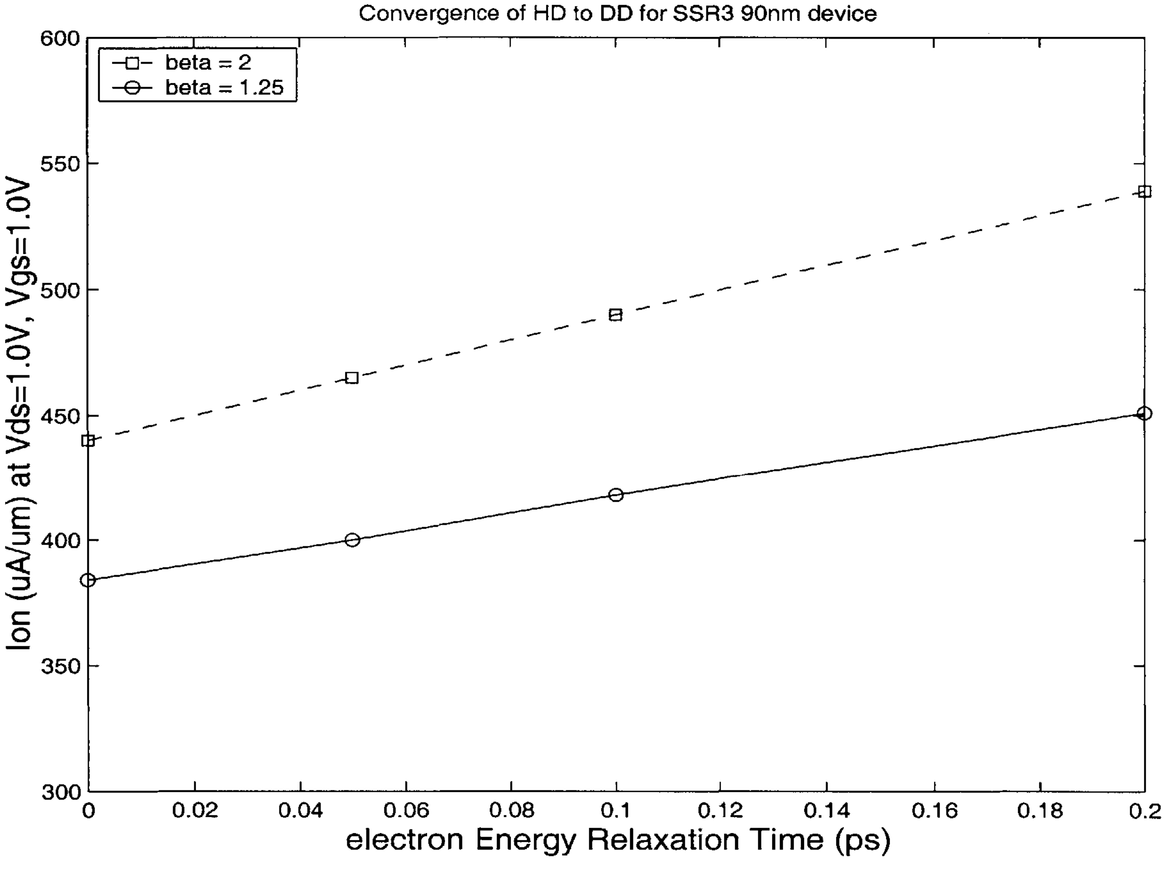

Figure 4.1: Confirmation of the convergence of the simulated drive current of a NMOSFET using EB to the DD model as energy relaxation time goes to zero.

where $\bar{S}$ represents energy flux, H is net energy dissipation, and τ_{wn}, τ_{wp}, T_n, T_p, and T_0 signify the energy relaxation time constants for electrons and holes, the carrier temperatures, and the lattice temperature, respectively. The divergence of energy flux is offset by the energy of carriers moving in the field, their scattering rate, and the dissipation. The case of perfect energy scattering drives the EB current of a device in Fig. 4.1 to the DD current symbolized at τ_w goes to zero. Here, the temperature diffusion is included in the DD terms, revising Eq. 2.4 to 2.5 as

$$\bar{J}_n = q\mu_n\left[n\bar{E} + \nabla\left(\frac{kT_n}{q}n\right)\right] \qquad \text{Equation 4.4}$$

$$\bar{J}_p = q\mu_p\left[p\bar{E} - \nabla\left(\frac{kT_p}{q}p\right)\right] \qquad \text{Equation 4.5}$$

Furthermore, the energy flux itself depends on the internal energy of the carriers and the variation

$$\bar{S}_n = -\frac{5}{2}\frac{kT_n}{q}\left[\frac{1}{q}\bar{J}_n + C\mu_n n\nabla\left(\frac{kT_n}{q}\right)\right] \qquad \text{Equation 4.6}$$

$$\bar{S}_p = \frac{5}{2}\frac{kT_p}{q}\left[\frac{1}{q}\bar{J}_p - C\mu_p p\nabla\left(\frac{kT_p}{q}\right)\right] \qquad \text{Equation 4.7}$$

in carrier temperature multiplied by the heat capacity, C.

Armed with the machinery to simulate carrier transport, deriving theory based targets for the model parameters becomes instructive. In particular, it would be useful to have an intuitive feel for the validity of using certain values as time constants. Using an effective *lateral* field

$$\mu_n E_{efflat}{}^2 = \frac{3}{2q}\frac{k(T_n - T_0)}{\tau_{wn}} \qquad \text{Equation 4.8}$$

yields a relation for the spacially steady state behavior from Eq. 4.2 and 4.4. Moreover, the momentum scattering captured by the mobility can be expressed with a time constant

$$\tau = \frac{m}{q}\frac{v}{E_{efflat}}$$ Equation 4.9

which is the momentum of the effective mass, m times velocity, divided by the field force on the carriers. Substituting into Eq. 4.8 gives a relation for the energy time constant

$$\tau_{wn} = \frac{3}{2q}\frac{k(T_n - T_0)}{E_{efflat}^{2}}\frac{m}{q\tau}$$ Equation 4.10

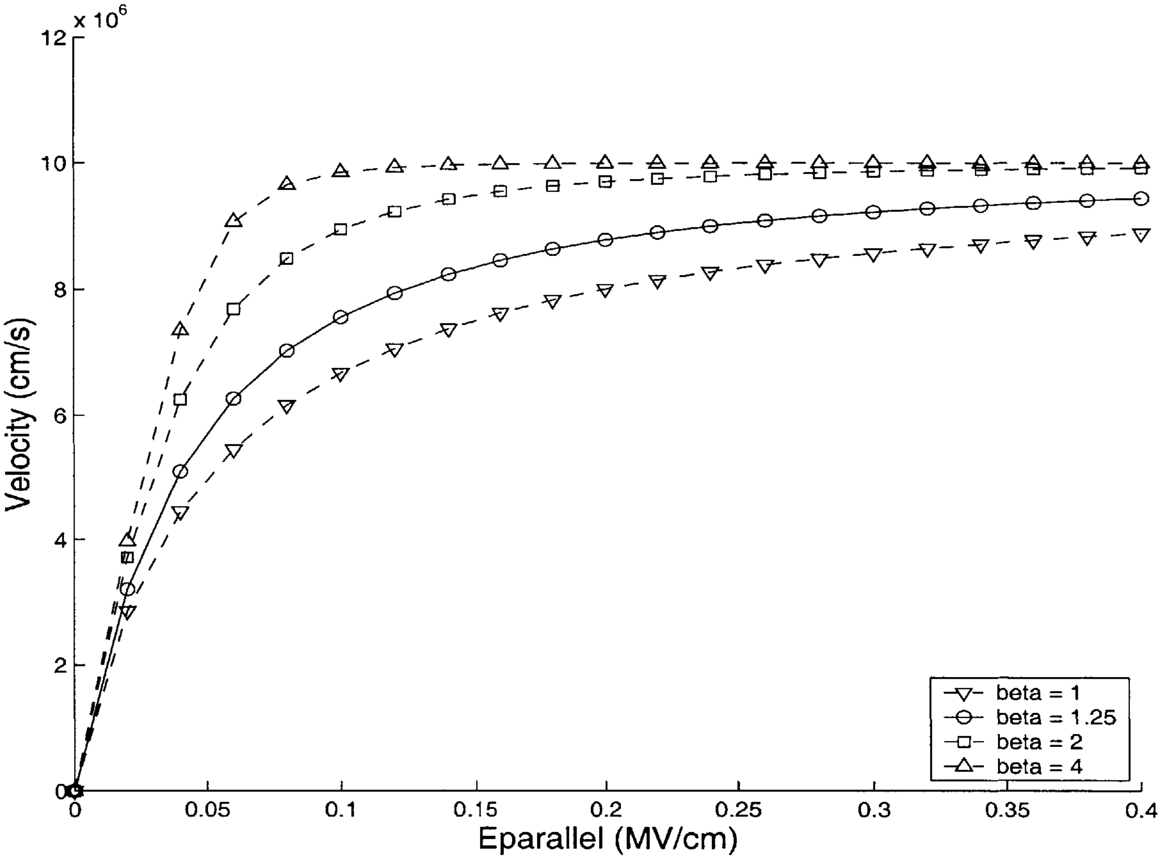

Figure 4.2: Plot of the effective velocity (the Caughey-Thomas mobility times the electric field) as a function of the effective field assuming μ_{gen} = 200 cm^2/Vs and v_{sat} = 10^7 cm/s. The effect of a decreased beta makes it harder for the device to reach velocity saturation.

A few more pieces of information promise a quantitative estimate for the key transport model constants. Empirically, the Caughey-Thomas expression gives a mobility field dependence

$$\mu = \frac{\mu_{gen}}{\left[1 + \left(\frac{\mu_{gen} E_{efflat}}{v_{sat}}\right)^{\beta}\right]^{1/\beta}} \qquad \text{Equation 4.11}$$

where μ_{gen} is the low-field generalized mobility and v_{sat} is the saturation velocity. The fitting parameter β modulates the rapidity at which the device velocity (effectively μ times $E_{eff,lat}$) reaches saturation. The plot of Fig. 4.2 graphically displays that decreasing β causes the velocity to saturate at higher lateral fields. In addition, the carrier mobility is related to the low-field mobility through the Einstein relation where the diffusion rates must be equivalent such that

$$\mu_n = \frac{T_0}{T_n} \mu_{gen} \qquad \text{Equation 4.12}$$

Plugging [48] the knowledge of Eq. 4.11 and 4.12 into the original estimate of Eq. 4.10 yields

$$\tau_{wn} = \frac{3}{2} \frac{kT_0}{q} \frac{\mu_{gen}}{v_{sat}^{\ 2}} \frac{\left(\frac{T_n}{T_0} - 1\right)}{\frac{T_0}{T_n}\left(\left(\frac{T_n}{T_0}\right)^{\beta} - 1\right)^{2/\beta}} \qquad \text{Equation 4.13}$$

Thus, for transport situations in which energy scattering matters $T_n \gg T_0$ and assuming $T_0 = 300$ K, $v_{sat} = 10^7$ cm/s, and $\mu_{gen} = 250$ cm^2/Vs results in a predicition for τ_{wn} of 0.1 ps. Carrying this hand calculation further, the mean free path between scattering events goes as velocity 10^7 cm/s multiplied by time 0.1 ps, giving an estimate of 10 nm for a non-ballistic transport length above which the macroscopic EB models remain valid [49]. For β in the range of 1.25 to 2, there is small dependence of τ_w on energy where the carriers are at least a couple times hotter than the lattice. Also, improved transport behavior is implied for substrates with enhanced mobility.

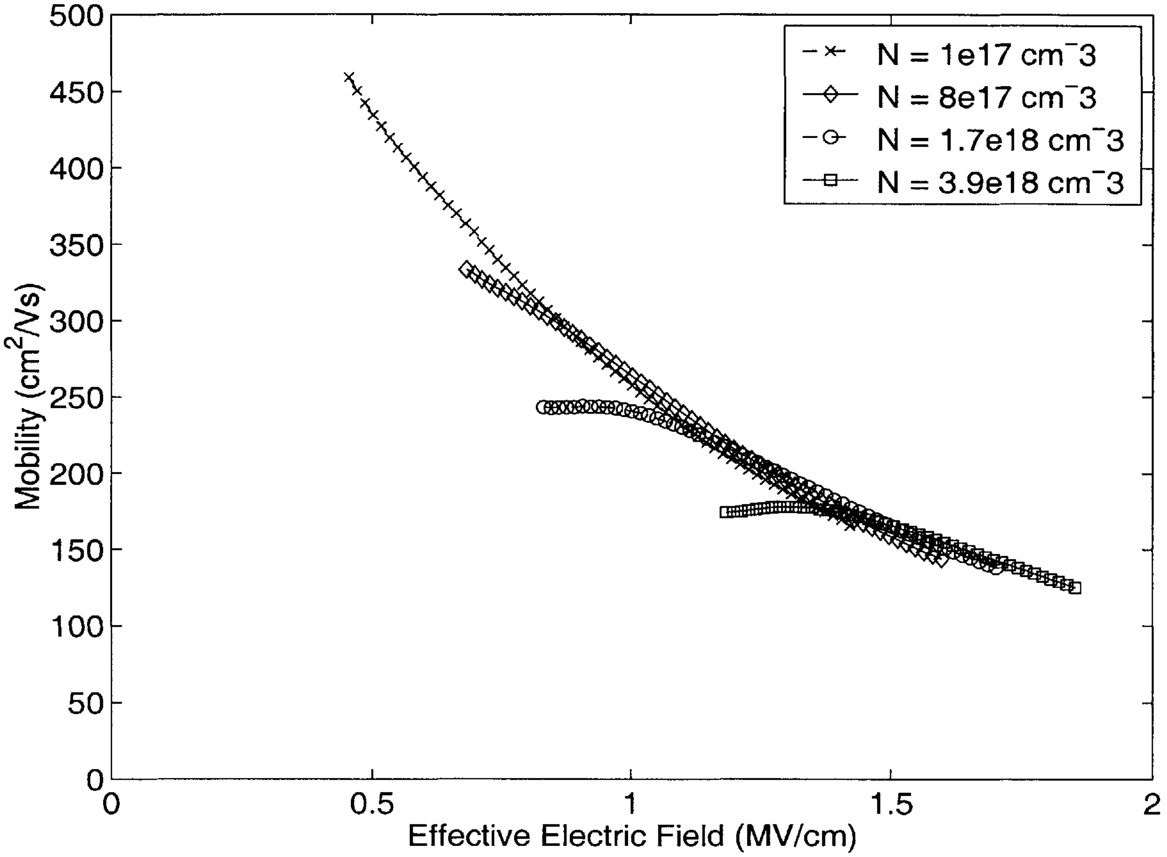

Figure 4.3: Measured mobility at V_{DS} = 10 mV corrected [51] for field above V_t for long channel devices with high bulk dopings extracted at 1×10^{17}, 8×10^{17}, 1.7×10^{18}, and 3.9×10^{18} cm^{-3}.

Last but not least, the low-field mobility itself must be well characterized. The principle scattering mechanisms involve interactions with Coulomb impurity centers, phonon and surface vertical-field-dependent "universal" mobility. The long channel devices [50] of Fig. 4.3 exhibit dependence on both vertical field and multiple high channel dopings.

4.2 MOSFET Mobility Model

Before evaluating the energy dependent transport behavior of deep submicron MOSFETs, a well calibrated model for the momentum scattering dependent mobility must be established. The challenge of separating the coulombic scattering from the field dependent mobility requires caution.

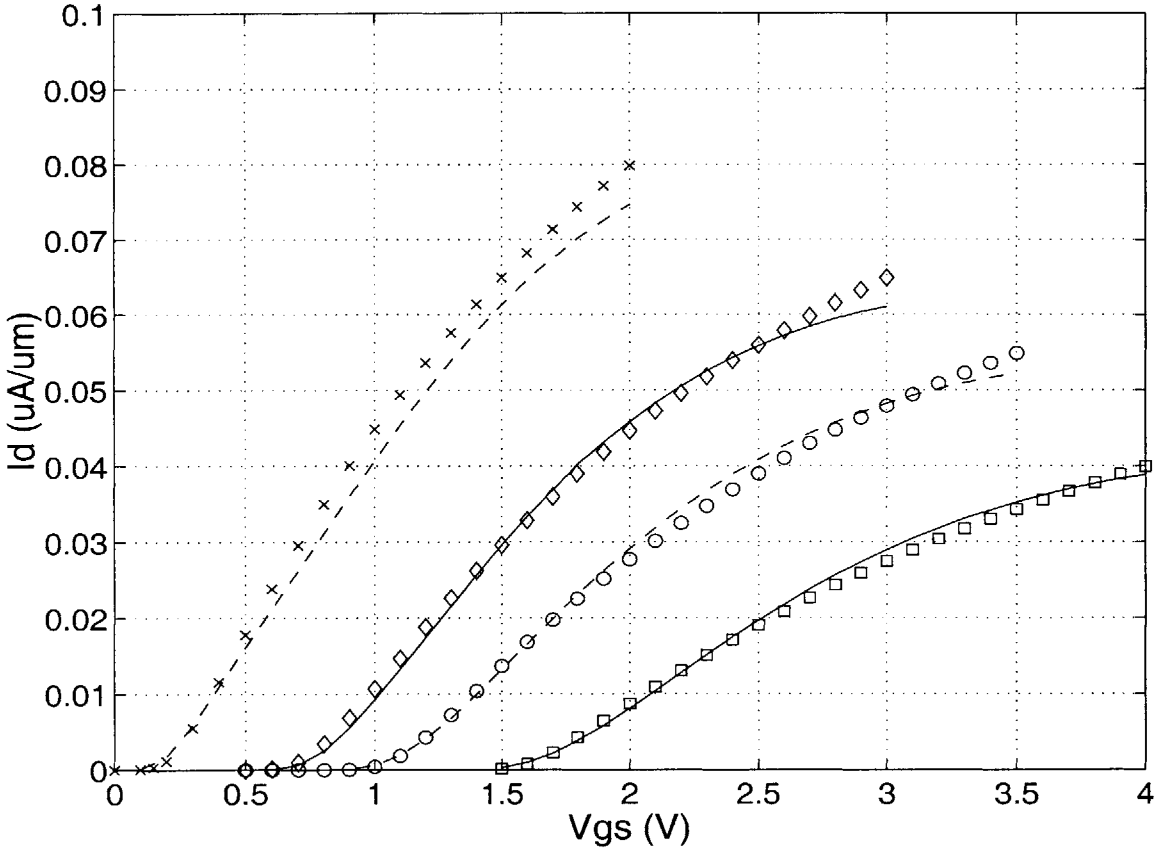

Figure 4.4: Optimization of coulombic mobility dependence as well as universal mobility coefficients using a range of I-V data for the long channel bulk devices with various doping levels.

An opportune method of extracting the coulombic component takes the aforementioned long channel bulk devices for each of which a uniform doping value was extracted by fitting to the split C-V measurements. These NMOSFETs with a range of high dopings then undergo an optimization of the bulk mobility by fitting to moderate inversion I-V data displayed in Fig. 4.4. Because the mobilities of Fig. 4.3 depart from the universal curve μ_{uni} into a doping dependent plateau μ_{bulk}, an appropriate choice of a generalized mobility model μ_{gen} takes minimum [52]

$$\mu_{gen} = min(\mu_{uni}, \mu_{bulk})$$ Equation 4.14

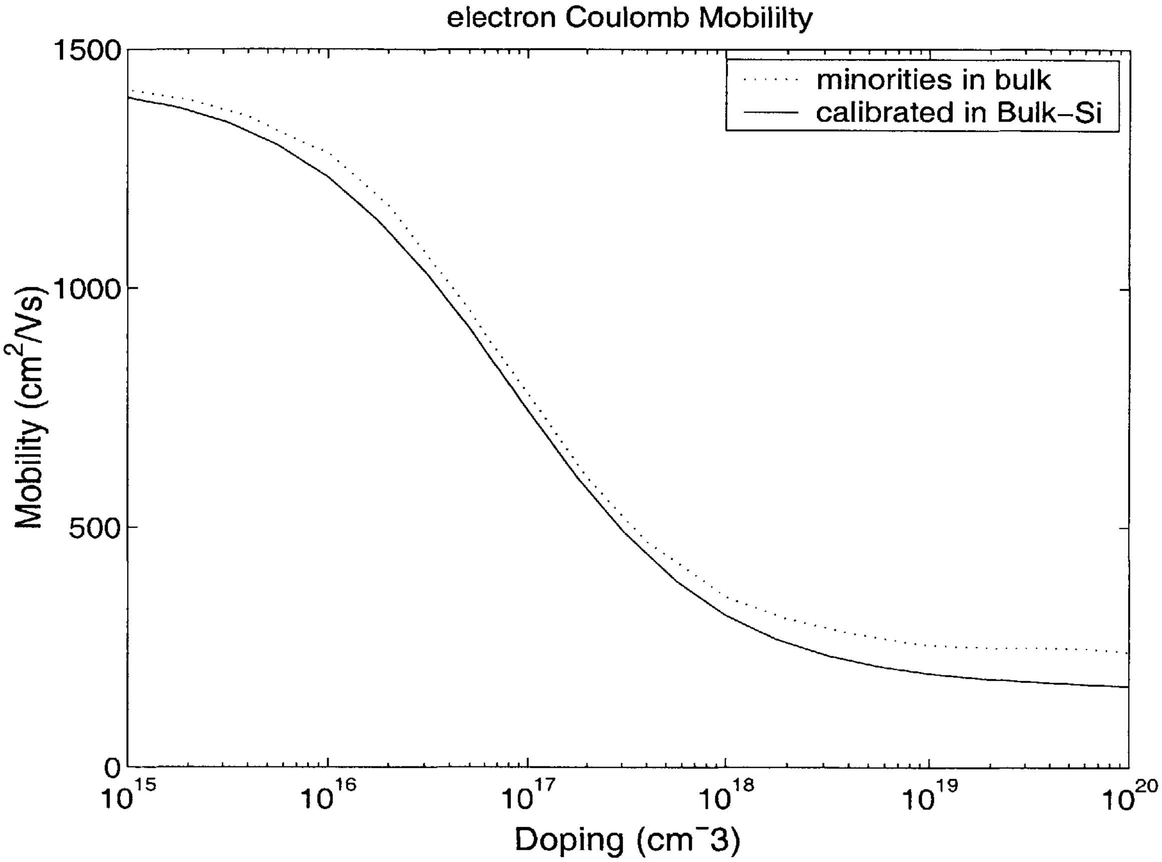

Figure 4.5: Extracted curve of coulombic mobility versus doping level. The model assumes this mobility (some combination of impurity and phonon scattering) whenever it is under the universal curve. The mobility for minority carriers in bulk tracks this result.

The I-V optimization provides four coulombic mobility [53] data points as in Fig. 4.5 between 1×10^{17} and 1×10^{19} cm^{-3}. An analytical expression for the mobility of minority carriers in bulk, as is the case for carriers in an inversion layer, was modified from

$$\mu_{bulk} = \mu_0 + \frac{\mu_{max0}(T/300)^{-\gamma} - \mu_0}{1 + (N_{total}/C_r)^{\alpha_0}} - \frac{\mu_1}{1 + (C_s/N_{total})^{\alpha_1}}$$ Equation 4.15

by lowering the minimum mobility, μ_0, to 170 cm^2/Vs and adjusting slightly the power, α_0, dependence on total doping N_{total} over a default critical concentration $C_r = 7.92\times10^{16}$ cm^{-3}. Experimentally, the μ_{max0} is about 1440 cm^2/Vs while μ_1 is negligibly less than 10 cm^2/Vs.

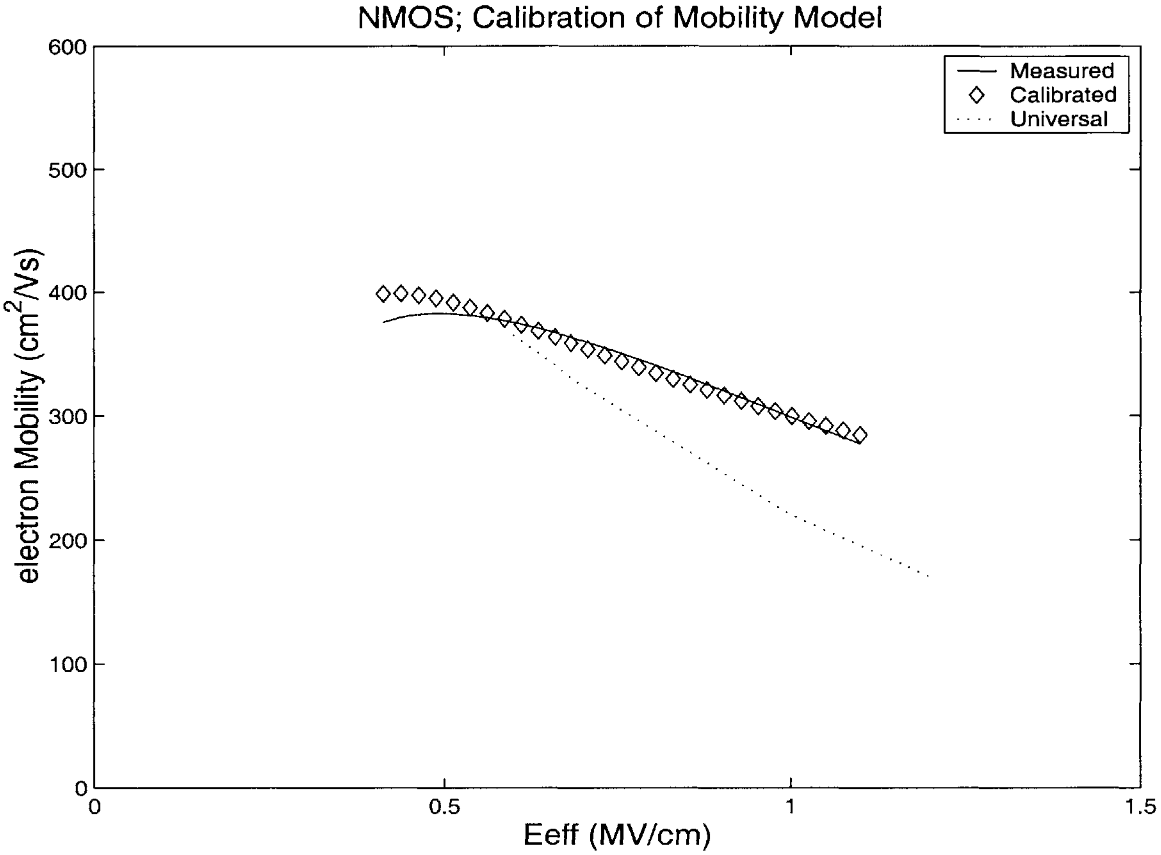

Figure 4.6: Measured and calibrated mobility vs. effective field for a nitrided oxide NMOS technology with t_{ox} = 3.3 nm.

Next, the universal mobility [54] needs a separate calibration for each technology because

$$\frac{1}{\mu_{uni}} = \frac{1}{\mu_{ph}} + \frac{1}{\mu_{sr}} \qquad \text{Equation 4.16}$$

$$\mu_{ph} = \frac{B}{E_{\perp}} + \frac{C \cdot N_{total}^{\ \ 0.0284}}{T \cdot E_{\perp}^{\ 1/3}}$$ Equation 4.17

$$\mu_{sr} = \frac{D}{E_{\perp}^{\ \delta}}$$ Equation 4.18

depends on scattering [55] for phonons μ_{ph} and the surface roughness μ_{sr} where coefficients B, C, D and δ are vertical field, $E_{\perp}$, fitting parameters to the long channel strong inversion I-V data.

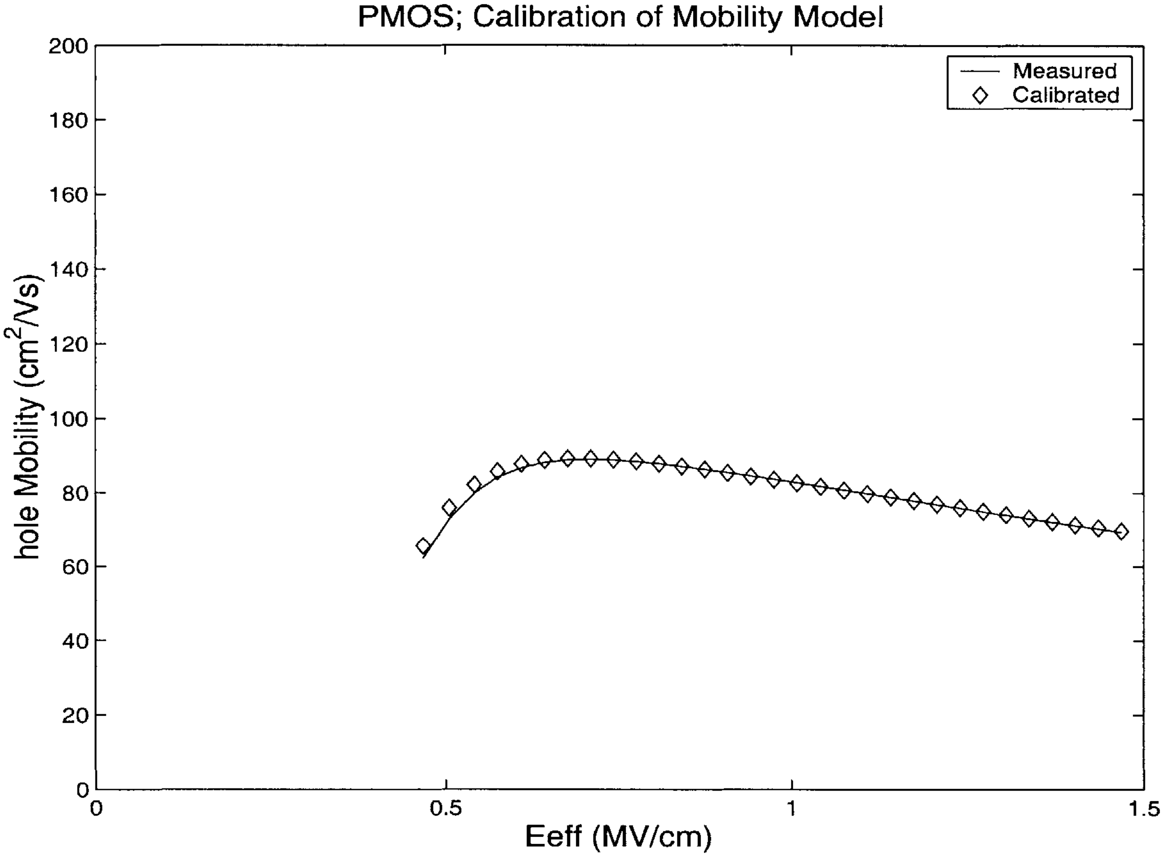

Figure 4.7: Measured and calibrated mobility vs. effective field for a nitrided oxide PMOS technology with t_{ox} = 3.3 nm.

Plots of the extracted universal mobilities for devices with *nitrided* [56] gate oxide of a NMOS family in Fig. 4.6 and PMOS in Fig. 4.7 illustrate the quality of fits obtained where the effective electric field, E_{eff}, is defined as the field across the gate minus half the inversion field

$$E_{eff} = \frac{V_{GS}}{3t_{ox}} - \frac{Q_i}{2\varepsilon_s} \qquad \text{Equation 4.19}$$

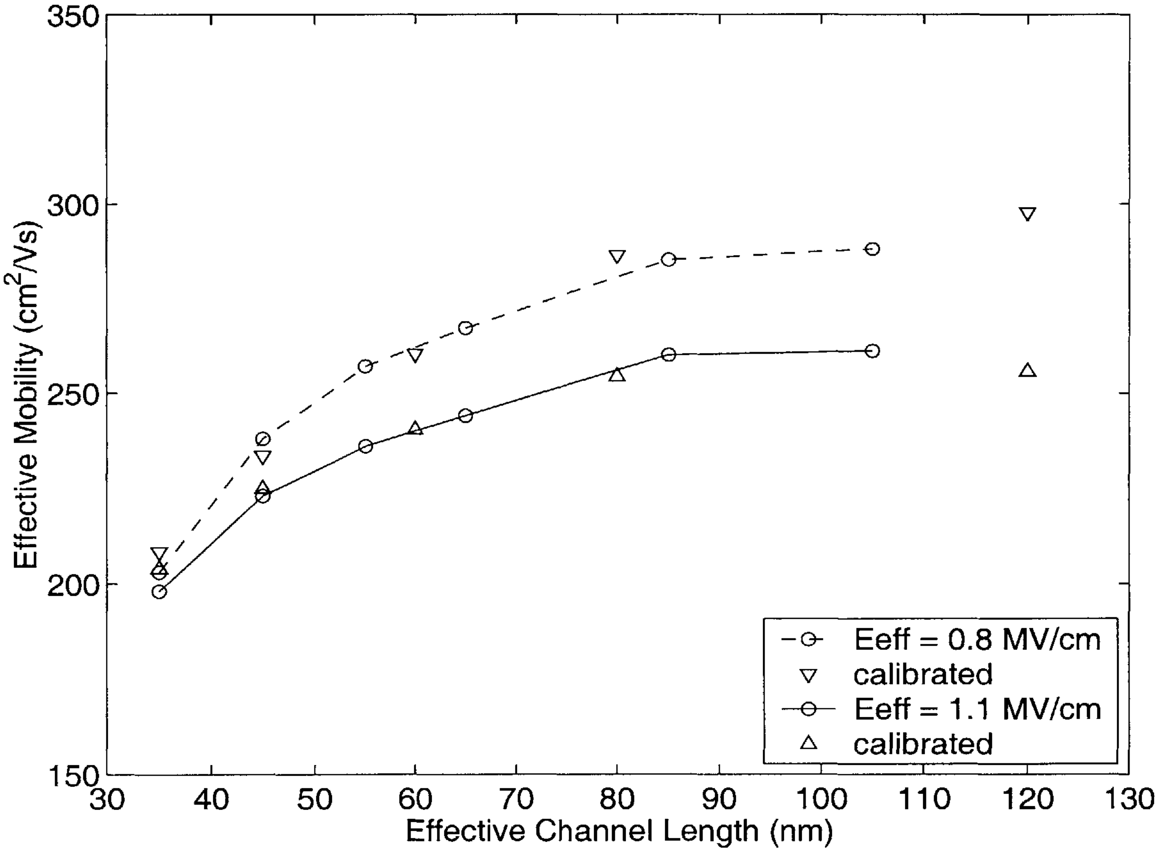

Figure 4.8: A simulation of the effective short channel mobility at constant E_{eff} versus L_{eff} utilizing the calibrated coulombic mobility on the t_{ox} = 1.5 nm NMOS family. The good fit indicates that the merged halo doping likely degrades the mobility.

An important corraboration of the dominance of coulombic mobility arises when examining curves of mobility versus effective channel length [57] at constant E_{eff} as for the t_{ox} = 1.5 nm family of Fig. 4.8. The curves were experimentally extracted [58] using integrated split C-V (shifted by the V_t changes for each short device) of a long channel device to get Q_i, with the short channel drive corrected by resistance R where L_{eff} is extracted by inverse modeling

$$\mu_{eff} = \frac{L_{eff}}{(V_{DS} - I_D R)} \frac{I_D}{W Q_i}$$ Equation 4.20

Convergence occurs for both E_{eff} because the mobility plateaus for a particular doping. Because simulation using a coulombic mobility offers a good fit, the mobility degradation is largely explained as a rise in the merging halo dopings. However, long-range Coulomb scattering [59] (e.g., from S/D) might contribute at sub-50 nm. With the coulomb term removed, the simulated points in Fig. 4.9 follow a nearly universal E_{eff} dependence for both 0.8 MV/cm and 1.1 MV/cm.

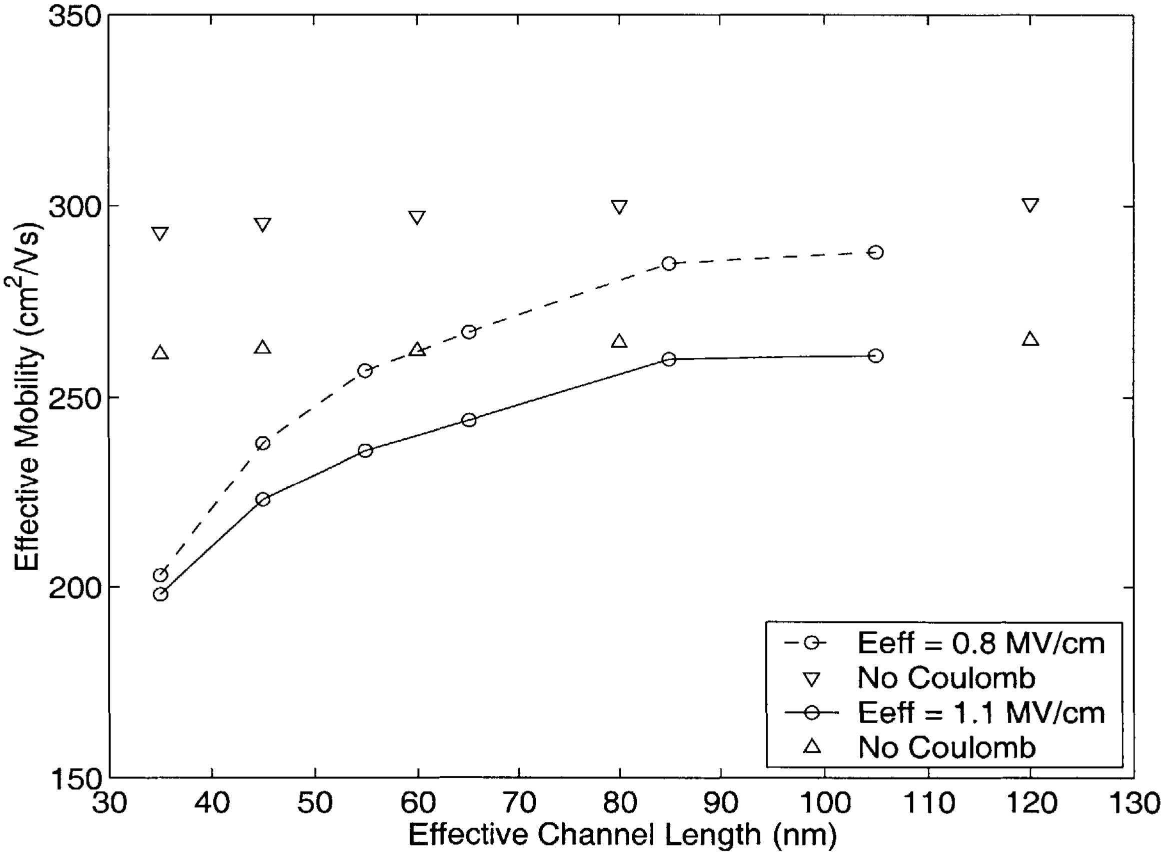

Figure 4.9: Experimentally observed mobility degradation versus effective channel length in a t_{ox} = 1.5 nm NMOS device family for constant effective field. The triangle symbols represent simulations at various L_{eff} for E_{eff} of 0.8 MV/cm and 1.1 MV/cm without a Coulomb mobility model.

4.3 Calibration Methodology

A method of calibrating macroscopic transport models is presented utilizing MOSFET dopings from 1-D and 2-D inverse modeling. An independent calibration of mobility model, parasitic resistance, and transport parameters for each technology studied (with oxide thicknesses of 3.3 nm, 1.5 nm, and 1.7 nm) accounts for the strong inversion characteristics over different voltages and channel lengths.

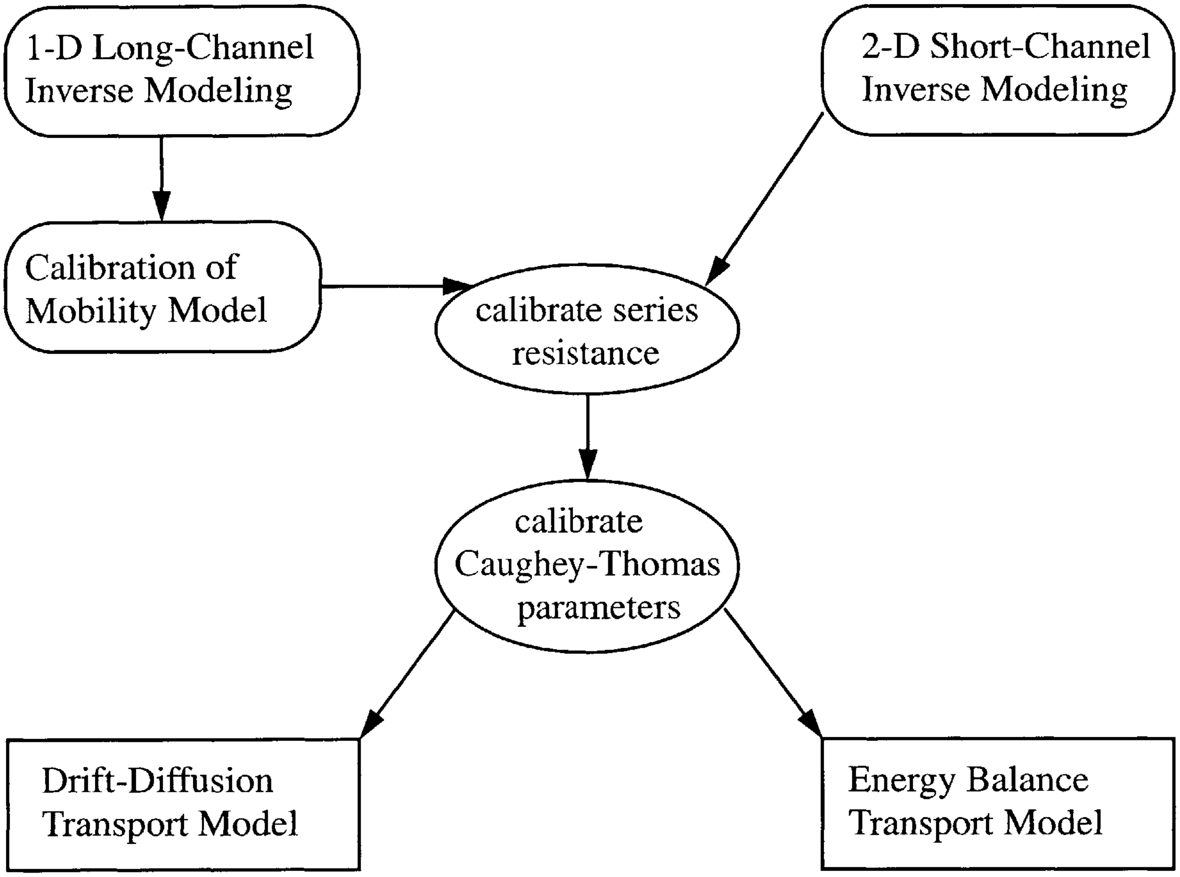
1-D Long-Channel Inverse Modeling
2-D Short-Channel Inverse Modeling
Calibration of Mobility Model
calibrate series resistance
calibrate Caughey-Thomas parameters
Drift-Diffusion Transport Model
Energy Balance Transport Model

Figure 4.10: Flowchart outlining the transport model calibration procedure from inverse modeling, mobility, parasitics, and transport parameters.

Fig. 4.10 outlines the calibration procedure. After obtaining realistic 2-D doping profiles (which are relatively insensitve to mobility), a channel mobility model [60] is calibrated for each MOSFET technology to account for different gate-stack dielectric fabrication processes. Thin oxide quantum effects are approximated by the Van Dort model. The optimization is performed on long-channel I-V data which have negligible dependence on parasitics and mobility degradation due to velocity saturation. The mobility model is chosen to include the effects of both impurity scattering and the surface mobility (a Mathiessen's combination of two terms with inverse dependence on vertical field).

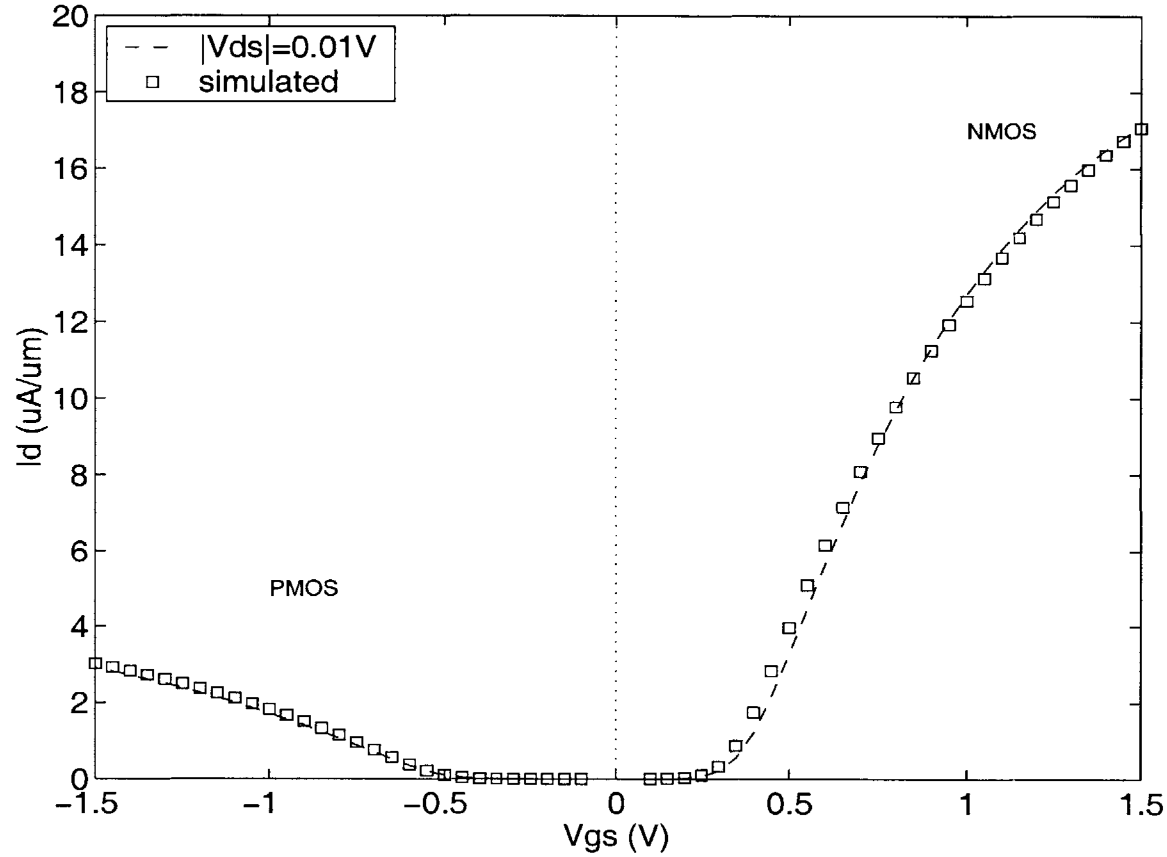

Figure 4.11: Calibration of parasitic resistance using strong inversion I-V at low V_{DS} = 10 mV for a t_{ox} = 3.3 nm family with a L_{eff} ~ 110 nm NMOS and L_{eff} ~ 150 nm PMOS device.

Having fit the mobility parameters on the long channel devices, the short channel transport phenomena are calibrated. Using strong inversion I_D vs. V_{GS} data at V_{DS} low enough, e.g. < 50 mV, to avoid velocity saturation effects, a lumped S/D resistance is extracted to account for the series resistance components arising from the contacts and sheet resistance of carriers travelling through the S/D. For instance, the inversion I-V fits of Fig. 4.11 at V_{DS} = 10 mV for a L_{eff} ~ 110 nm NMOS and a L_{eff} ~ 150 nm PMOS device of the tox = 3.3 nm family exhibit good calibration using the lumped parasitic model. Because the 2-D topography is known, the simulation structure already includes the voltage-dependent accumulation and spreading resistance of the S/D extensions. If desired, this internal resistance can be simply extracted in inversion by dividing the I_D by the drop in S/D potential up to edge of the L_{eff}.

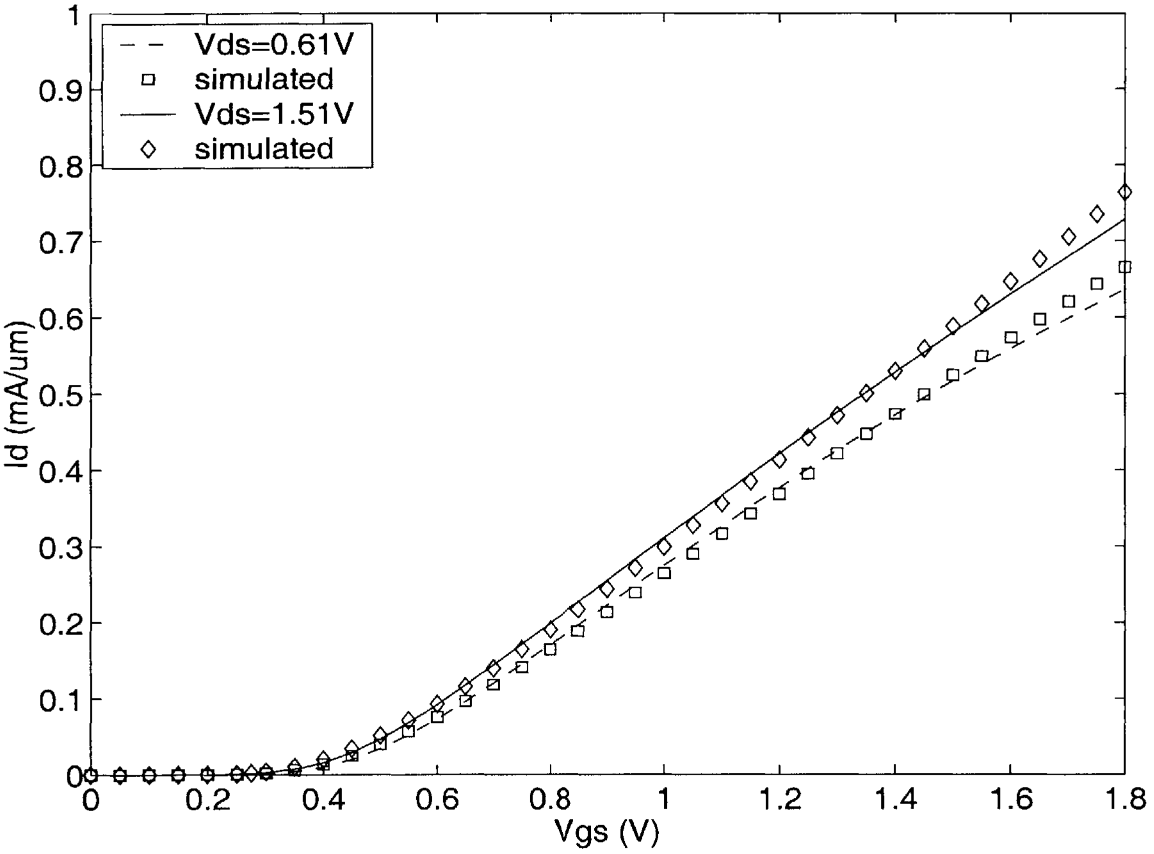

Figure 4.12: Fit to strong inversion data for NMOS L_{eff} ~ 110 nm t_{ox} = 3.3 nm device with extracted R_{SD} = 245 $\Omega\mu m$ at V_{GS} = 1.8 V.

Next, the empirical Caughey-Thomas (C-T) expression [61] (uniform E-field to mobility relation) is calibrated. Note that this relation is needed in both DD and EB models. Fig. 4.12

shows the strong inversion I-V optimized fits for the NMOS t_{ox} = 3.3 nm technology. For this calibration it is important to use devices that are short enough so the C-T coefficients can have an effect but long enough that the DD model still applies. In other words, the simulated I_D is nearly independent of the EB model energy relaxation time, τ_w, for reasonalbe values, e.g. < 0.5 ps. The extracted C-T exponent (β) parameter for all technologies was nearly 1.25 while the model parameter v_{sat} was 9.5×10^6 cm/s [62]. The corresponding PMOS t_{ox} = 3.3 nm technology I-V fits in Fig. 4.13 also exhibit very good agreement out to V_{GS} = -1.8 V for drain voltages of -0.61 V and -1.51 V.

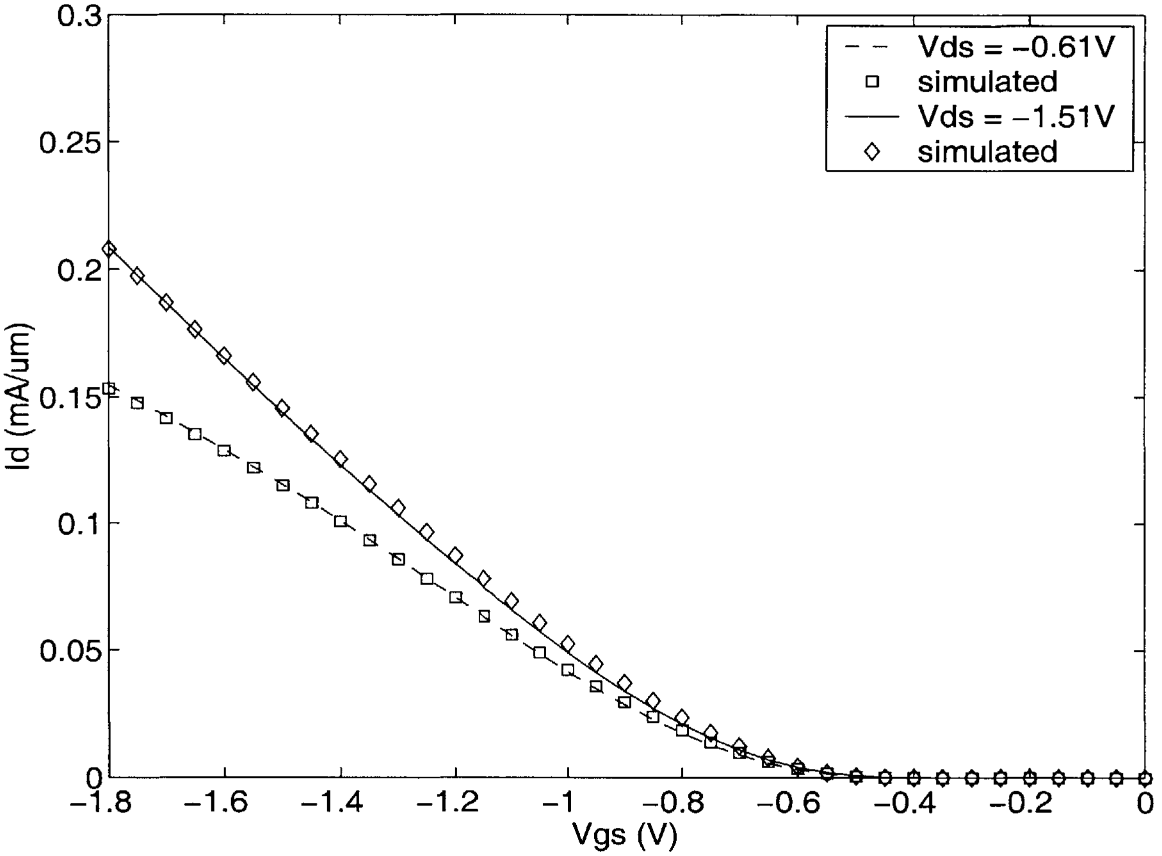

Figure 4.13: Fit to strong inversion data for PMOS L_{eff} ~ 150 nm t_{ox} = 3.3 nm device with extracted R_{SD} = 600 $\Omega\mu m$ at V_{GS} = -1.8 V.

4.4 Calibrated Results

To test the validity of the method and of the calibrated models, simulated versus measured I_{on} vs. I_{off} are compared over a broad range of device lengths and voltages for the three technologies from three different companies.

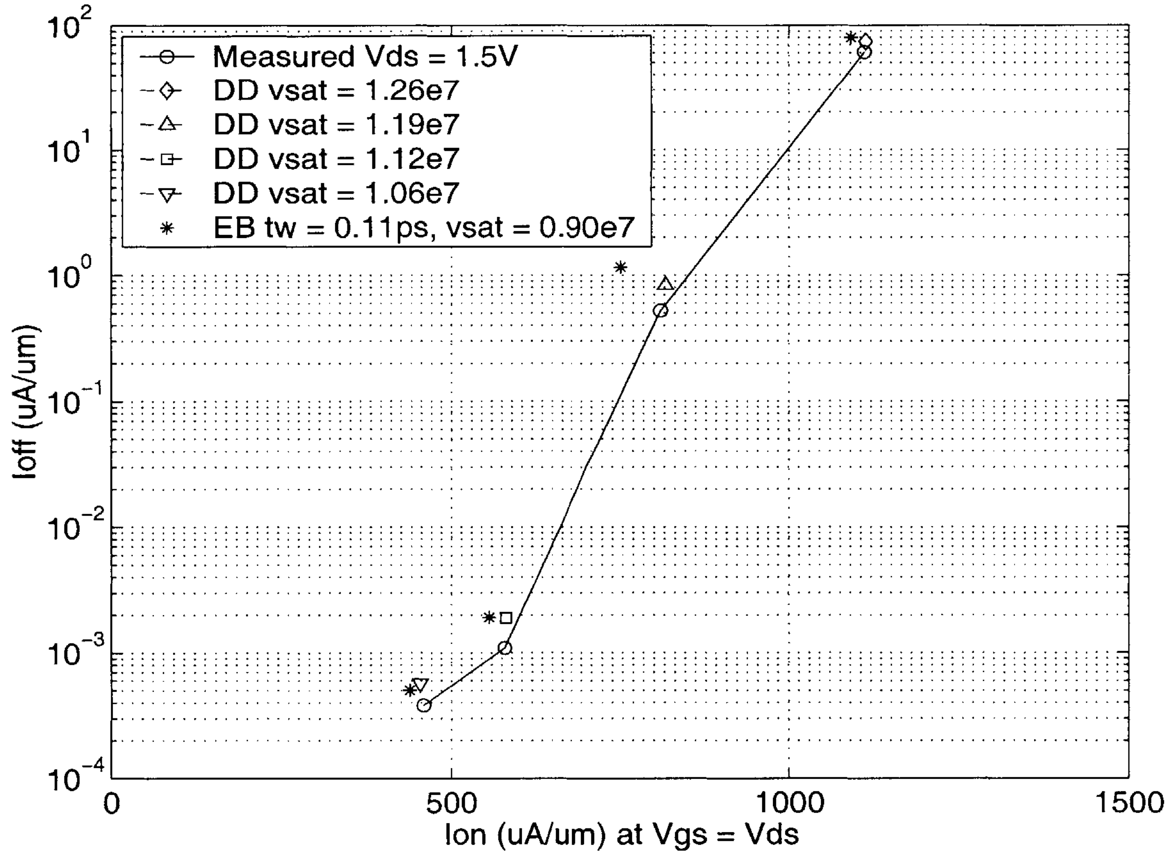

Figure 4.14: Measured vs. DD and EB I_{on} vs. I_{off} for NMOS t_{ox} = 3.3 nm family of L_{eff} ~ 50 nm, 80 nm, 110 nm, 150 nm with V_{DS} = 1.5 V.

As is well known, another way to fit the drive currents on a range of channel lengths is to use the DD approximation but scale v_{sat} as dimensions shrink (as an indication of the effective overshoot). The corresponding DD v_{sat} values provide decent fits for the t_{ox} = 3.3 nm NMOS family as shown in Fig. 4.14 assuming a V_{DD} of 1.5 V. On the other hand, in employing the EB approach, one must find the characterization of carrier energy relaxation that best fits the experimental data. Choosing a constant v_{sat} and locating a value for the energy relaxation time (τ_w) of 0.11 ps, the EB results are still in good agreement with measurements.

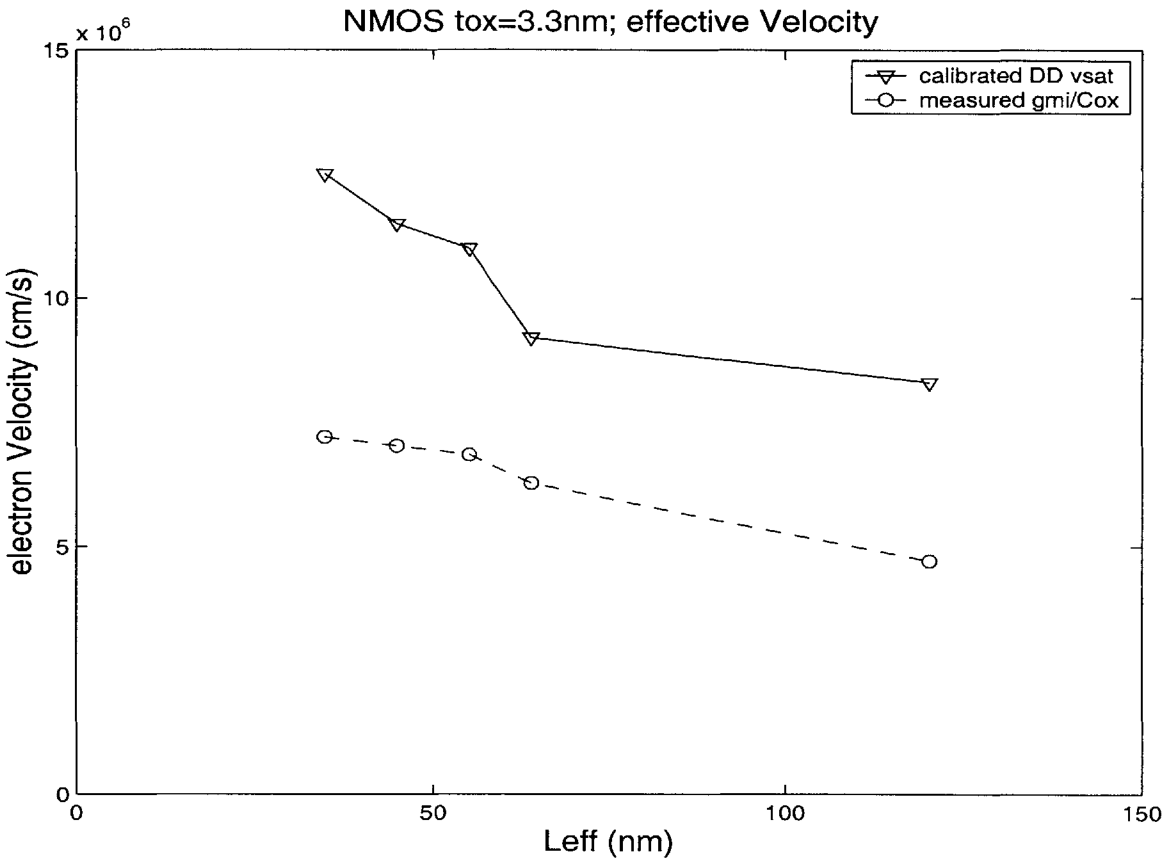

Figure 4.15: Comparison of the scaling trends of effective velocities defined as being extracted using the g_{mi} method and the calibrated DD v_{sat} for the NMOS t_{ox} = 3.3 nm family.

An interesting exercise is to compare the scaling behavior of the effective device velocity as in Fig. 4.15 extracted using the calibrated DD v_{sat} numbers against the measured g_{mi} where

$$v_{eff} = \frac{g_{mi}}{WC_{ox}} \qquad \text{Equation 4.21}$$

The overall trend in both cases presents a monotonically increasing velocity as channel length decreases. Although more data would be necessary to build confidence in an empirical model for velocity overshoot as a function of L_{eff}, this finding suggests its importance.

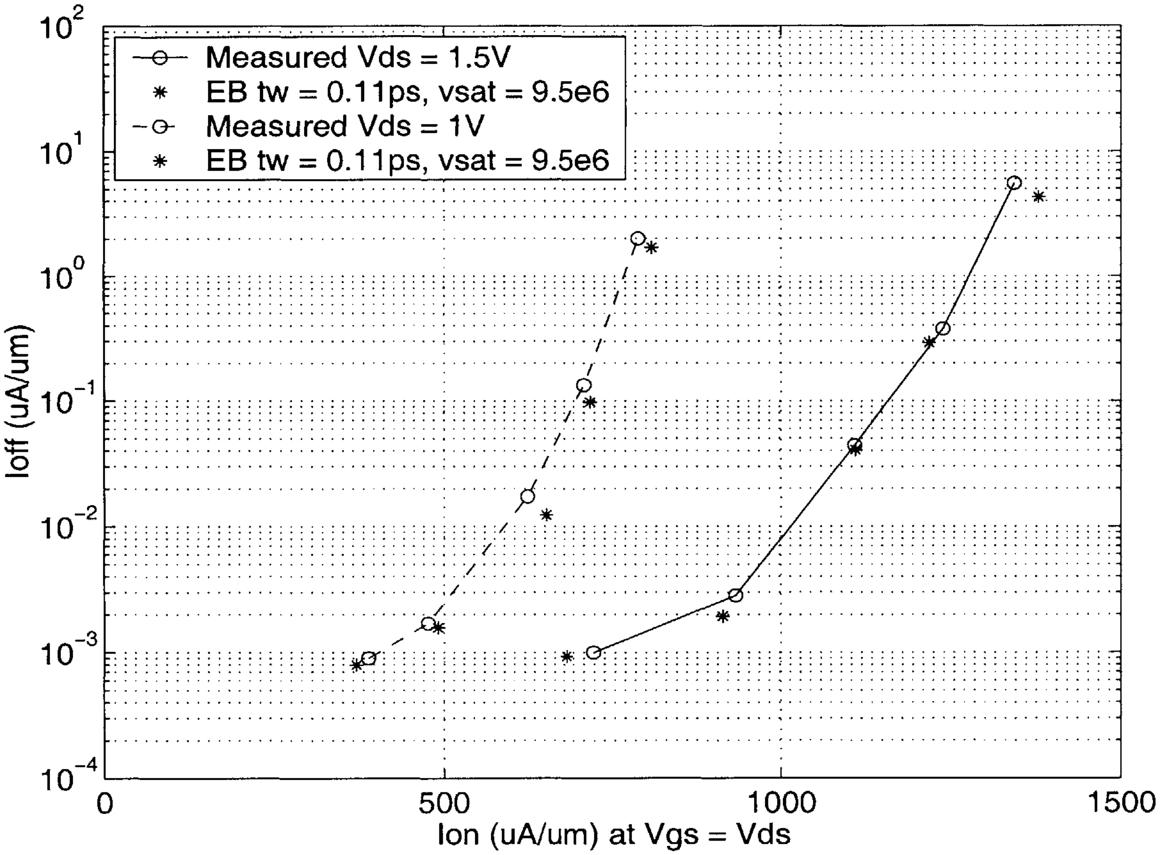

Figure 4.16: Measured vs. EB I_{on} vs. I_{off} for NMOS t_{ox} = 1.5 nm family of L_{eff} ~ 35 nm, 45 nm, 55 nm, 80 nm, 120 nm with V_{DS} = 1.5 V and 1 V.

Using the EB model and again adjusting *once* the constant τ_w to 0.11 ps, Fig. 4.16 shows very good agreement to the NMOS I_{on} vs. I_{off} curves for the t_{ox} = 1.5 nm family of devices for two different V_{DS} values. Due to slow QM convergence, the EB simulations sometimes approximated it with ~ 0.5 nm increment to the physical t_{ox}. Since the effective t_{ox} varies slightly between weak and strong inversion, a slight V_t shift is introduced and the values for current were read at an adjusted V_{GS}.

Furthermore, it is reassuring to observe that the EB model with the same energy transport parameters also fits the t_{ox} = 1.7 nm family very well. The results are consistent even down to an effective channel length of 30 nm, as evidenced by Fig. 4.17, for the V_{DD} values of 1.5 V and 1 V. The key contribution here is verifying experimentally that the macroscopic transport models do predict the device drive characteristics. An improvement to the model would incorporate an energy dependent τ_w to account for the errors from hotter carriers as the L_{eff} scales in the very short regime. To accurately develop such a calibration would require many sub-50 nm devices with full 2-D profiles from inverse modeling.

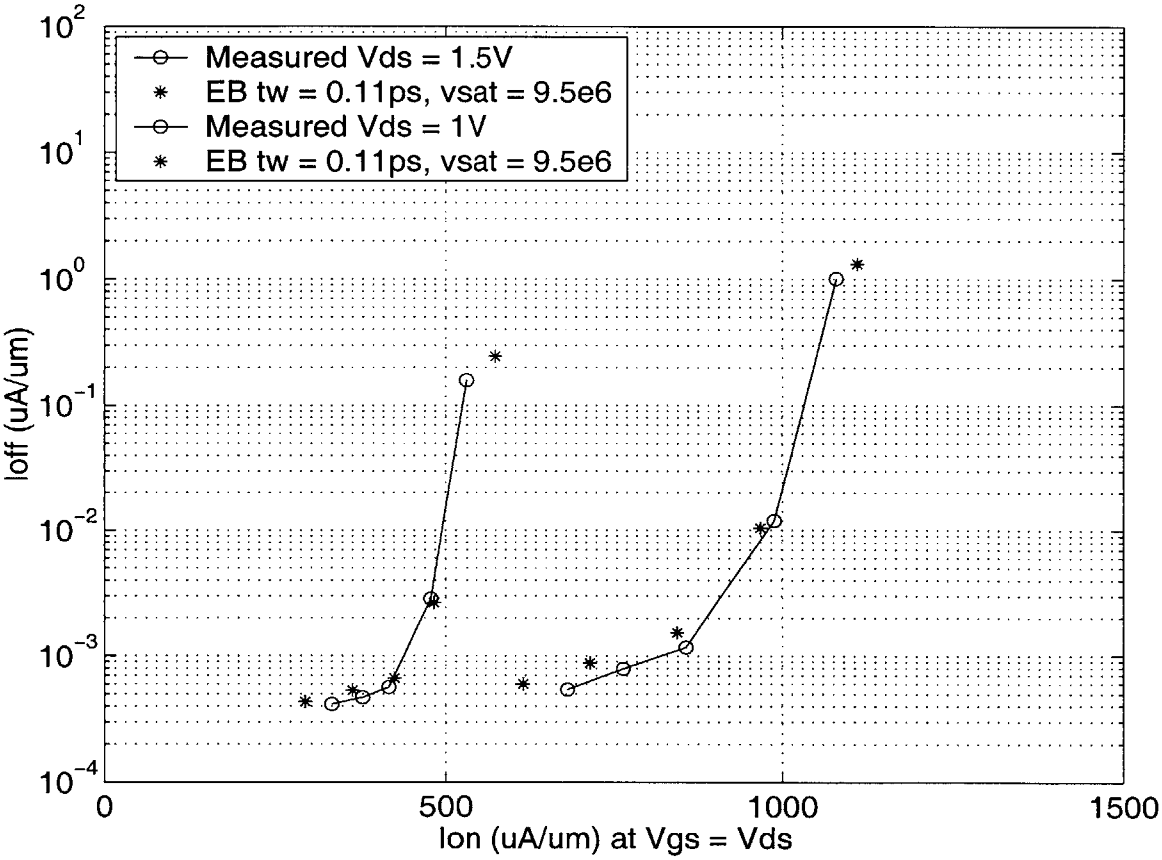

Figure 4.17: Measured vs. EB I_{on} vs. I_{off} for NMOS t_{ox} = 1.7 nm family of L_{eff} ~ 30 nm, 45 nm, 65 nm, 95 nm, 150 nm with V_{DS} = 1.5 V and 1 V.

The constant v_{sat} and τ_w = 0.11 ps EB results are still in good agreement with measurements. The t_{ox} = 3.3 nm PMOS family in Fig. 6 gives good agreement for a constant velocity saturation value in DD, which coincides with the EB independent of τ_w.

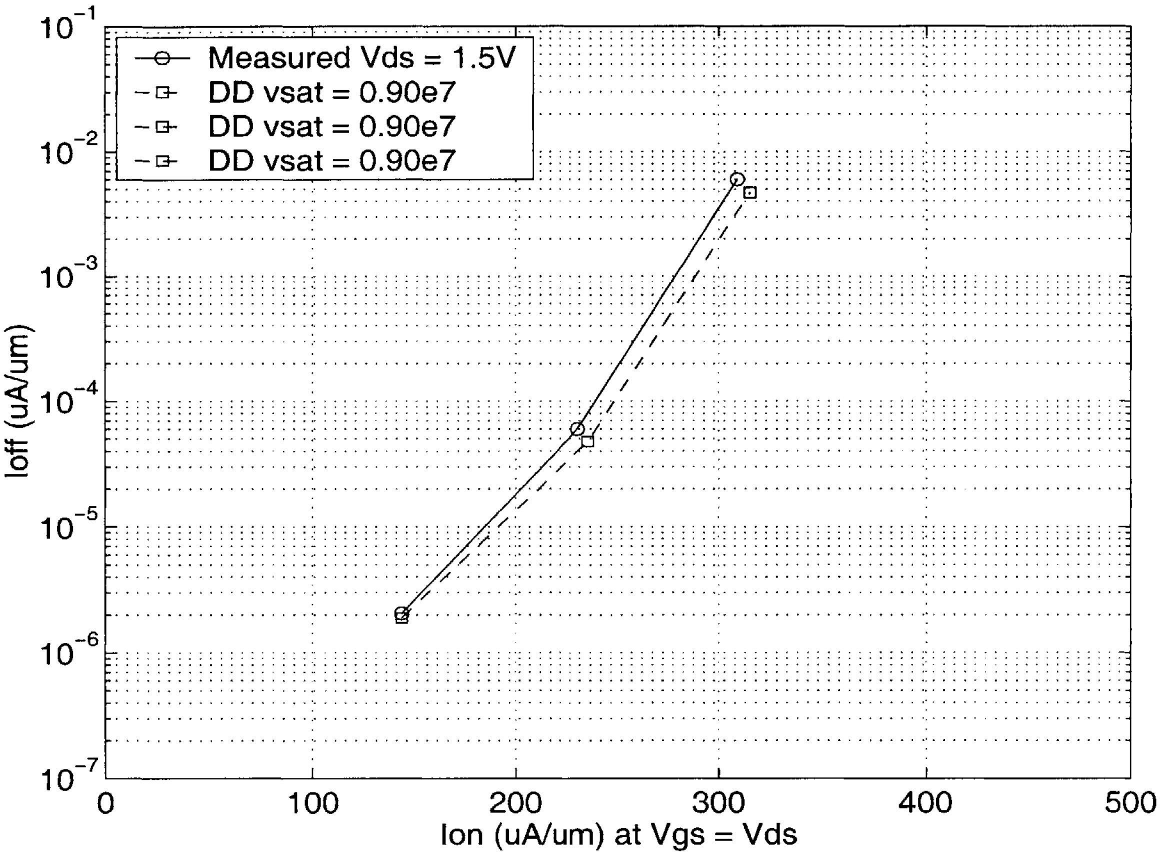

Figure 4.18: Measured vs. DD I_{on} vs. I_{off} for PMOS t_{ox} = 3.3 nm family of L_{eff} ~ 70 nm, 90 nm, 150 nm with V_{DS} = -1.5 V; EB calibration coincides (not shown).

In conclusion, a comprehensive inverse modeling scheme based on fitting the electrostatic device characteristics provides the details of 2-D doping configuration needed to calibrate mobility, parasitic resistance, and transport model parameters in the strong inversion regime. The fact that one set of scattering related constants fits the experimental data well for multiple technology families suggests that the EB simulation approach will continue to be useful well below 50 nm.

Chapter 5

Scaling to Sub-20 nm MOSFETs

5.1 Scaling of Device Topology

Harnessing the power of silicon at the limits of scaling [63] will require major changes [64] in device architecture, materials, or both. The bulk-Si (Bulk) MOSFET, which has been the workhorse of the semiconductor industry for a quarter century, becomes diseased when scaled into the sub-20 nm regime. At these gate lengths, short channel effects play so heavily with the threshold voltage behavior that reasonable off-currents are difficult to obtain.

A "well-tempered" device is defined as one that meets a certain set of electrostatic criteria. In general, the off-state current must not exceed a certain limit for the technology which will depend on the acceptable power dissipation of the entire integrated circuit. The typical V_t should be high enough to minimize I_{off} but low enough to provide appropriate gate overdrive (V_{GS} - V_t) voltage. Moreover, the gate should be the terminal that dominates the transistor turn-on. This requirement can be expressed as limiting the DIBL to about 150 mV/V.

Keeping the "well-tempered" ideal in mind, the 2-D doping of a Bulk NMOSFET was tailored with sufficiently peaked halos to meet the specifications of the ITRS 1999 Road Map for the L_{gate} = 13 nm node with effective t_{ox} = 1.03 nm. As displayed in Fig. 5.1, the S/D peak of 2×10^{20}

cm^{-3} rolls off sharply at the gate edge at 1 nm/dec and the 1×10^{19} cm^{-3} halo doping drops at about 2 nm/dec in the lateral direction. The steepness of the doping profile assumes a best case annealing technology that gives a nearly as-implanted activation; hence, there is little S/D overlap and the device exhibits short channel effects with L_{eff} ~ 11 nm.

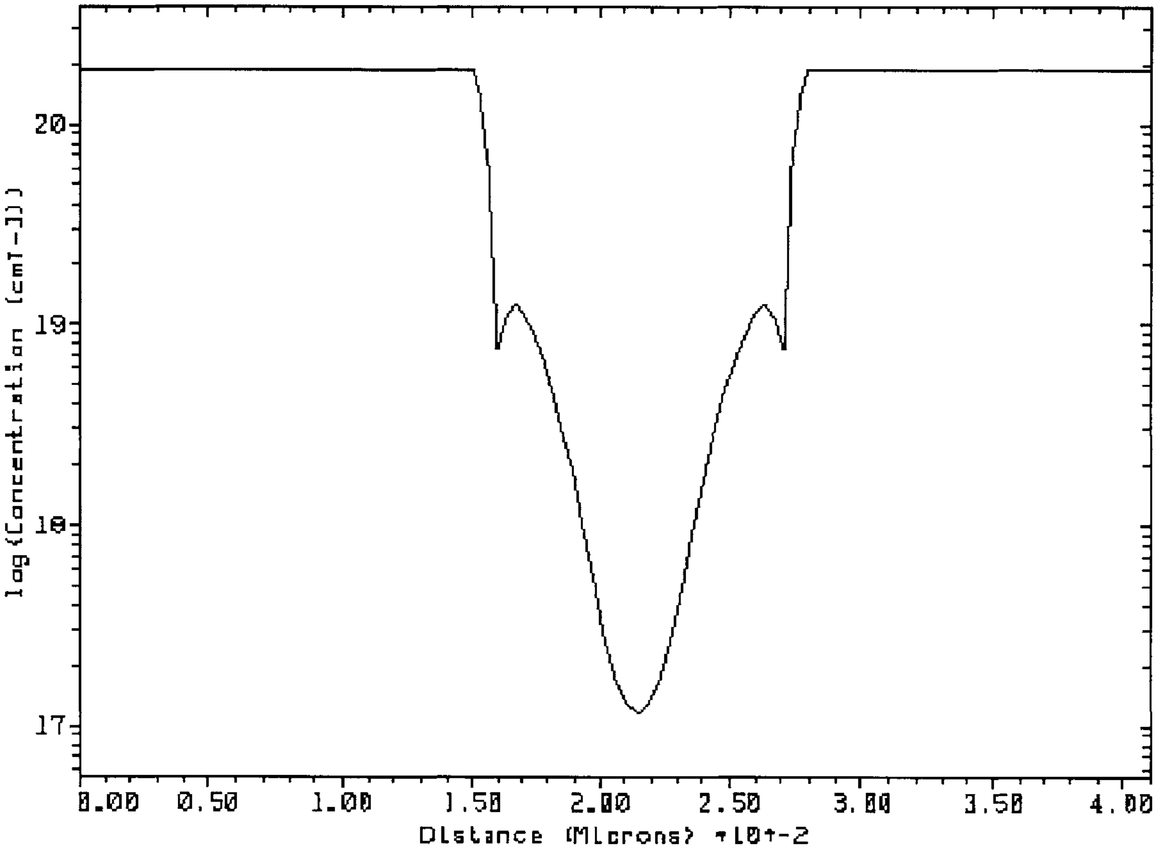

Figure 5.1: “Well-tempered” Bulk NMOSFET designed at the Lgate = 13 nm node on the Road Map. The abrupt lateral doping profile at the surface is shown.

While simulation (neglecting tunnelling) of the I_{off} is below the limit of 7 $\mu A/\mu m$ at the maximum drive voltage (V_{DD}) of 0.5 V, the DIBL was constrained to about 200 mV/V. It was found that regardless of the shape of the halo doping, a fixed amount of p-type impurity in the channel integrated in a box bounded by the metallurgical junctions with length L_{eff} and depth to x_j ~ 15 nm resulted in nearly the same I_{off} at V_{DS} = 0.1 V. To minimize the DIBL, it is best to slosh that doping into the most abrupt halos possible; however, there was no improvement for sharper than 2 nm/dec because there is a limit to how much the p+ halo can ameliorate the drain induced band bending at the source. Indeed, even if the p-type solid solubility were large enough, an even steeper halo would merely increase the amount of band-to-band leakage.

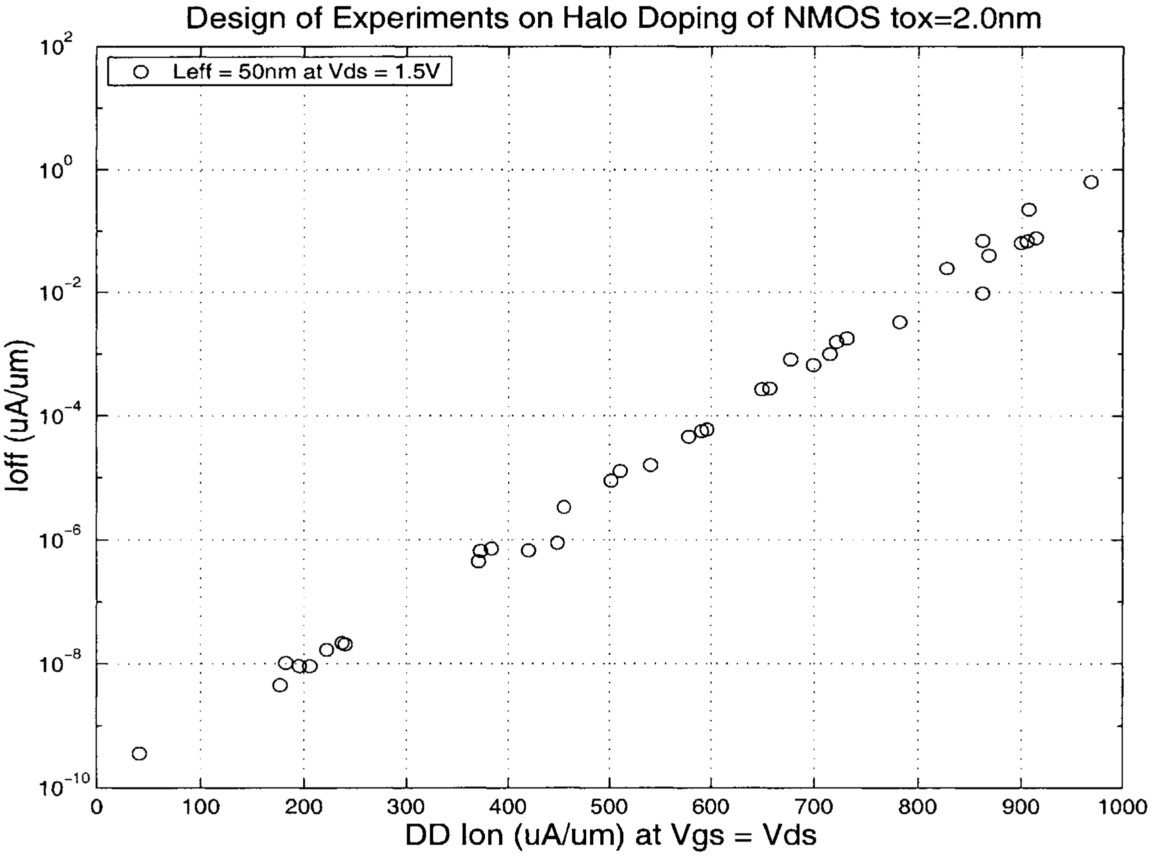

Figure 5.2: Design of experiments by varying the halo doping of a L_{eff} = 50 nm NMOSFET with effective t_{ox} ~ 2.4 nm. The simulated I_{on} vs. I_{off} curve was generated at $V_{DS} = V_{GS}$ = 1.5 V.

The conclusion from this study is that it is difficult to maintain the electrostatic integrity of sub-20 nm Bulk devices. On the other hand, one must also evaluate what will happen to the on-state device behavior as the doping is varied. The easiest way to capture this trend is to examine the I_{on} vs. I_{off} curve produced in a design of experiments as shown in Fig. 5.2 where a L_{eff} = 50 nm node was chosen with effective t_{ox} ~ 2.4 nm. The currents were obtained via DD simulation for speed, with v_{sat} arbitrarily set at 10^7 cm/s. Each point on the curve represents a particular permutation of 2-D gaussian halo doping parameters; the range of the peak is from 1×10^{18} to 1.9×10^{19} cm^{-3}, the halo overlap spans ±10 nm, the peak center is from the surface to 40 nm deep (like a SSR), the σ_x is from 10 to 30 nm, and the σ_y is from 10 to 40 nm.

Assuming $V_{DS} = V_{DD}$ and $V_{BS} = 0$ V, the subthreshold current of Eq. 2.13 reduces to

$$I_{off} = \frac{W}{L}\mu\phi_t^{\,2}\frac{\gamma C_{ox}}{2\sqrt{1.5\phi_F}}e^{\frac{-\phi_F}{2\phi_t}}e^{\left(\frac{-V_x}{n(\gamma)\phi_t}\right)} \qquad \text{Equation 5.1}$$

Taking the logarithm of each side relates $\log(I_{off})$ to V_x, which can be approximated here by V_t. With I_{on} defined as the product of carrier velocity and density, one obtains the expression

$$I_{on} = vC_{ox}[V_{DD} - V_t] = vC_{ox}[V_{DD} + n(\gamma)\phi_t\log(I_{off}) - \ldots] \qquad \text{Equation 5.2}$$

where '...' are terms independent of I_{off}. Thus, the slope of the I_{on} vs. I_{off} curve depends on 1) v, a velocity that depends on scattering events; 2) t_{ox} through the capacitance; 3) temperature through ϕ_t; and 4) γ, the doping dependent body factor. Going through the algebra reveals a weak relationship of I_{on} to doping configuration; the improvement in short channel effects is likely offset by doping dependent mobility degradation. While aggressive channel doping will continue to stabilize the electrostatics of Bulk devices, advances in the I_{on} vs. I_{off} slope in successive technologies will require thinner t_{ox} or improved transport through the semiconductor material.

Having examined some of the power and performance requirements that any device topology must meet in the sub-20 nm regime, it is possible to construct template MOSFETs for investigating the scaling of electrical behavior. The templates chosen for this study come in three viable candidate architectures: the continuation of Bulk, the double-gate (DG) and single-gate (SG) which both have thin Si film substrates.

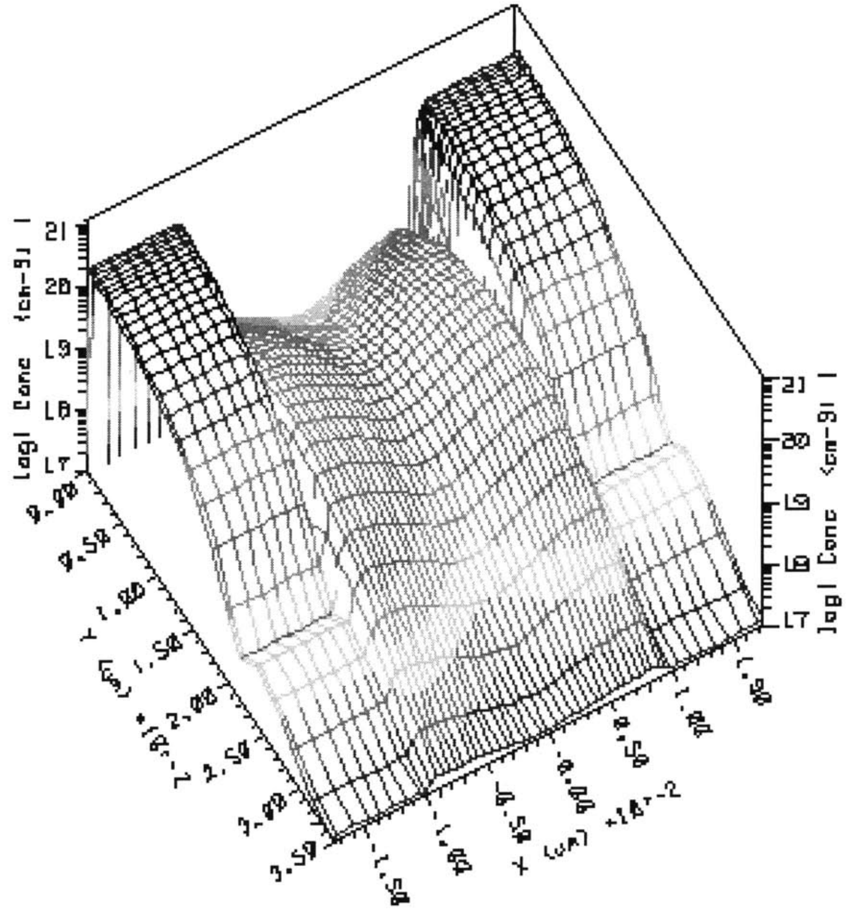

Figure 5.3: Template Bulk MOSFET topology showing complicated net doping on the z-axis as a function of the cross-section of the device, here with L = 20 nm.

The Bulk structure, which requires doping to maintain electrostatic integrity, employs a super-halo clustered around the S/D extensions. In Fig. 5.3, the S/D extensions begin falling laterally at the gate edge at an assumed 1 nm/dec and exhibit junction depths of 20 nm. Furthermore, the 2×10^{19} cm^{-3} super-halo is set, with $\sigma_x = 6$ nm (about 9 nm/dec) and $\sigma_y = 9$ nm, to peak at roughly half of x_j to impact electrostatics yet minimize the impurity density in the inversion layer. As the gate length is changed by ΔL in the scaling study, both the S/D and corresponding halos retain their assumed best-case characteristic fall-off values while only their lateral positions with respect to the middle of the channel are shifted by $\Delta L/2$ per side.

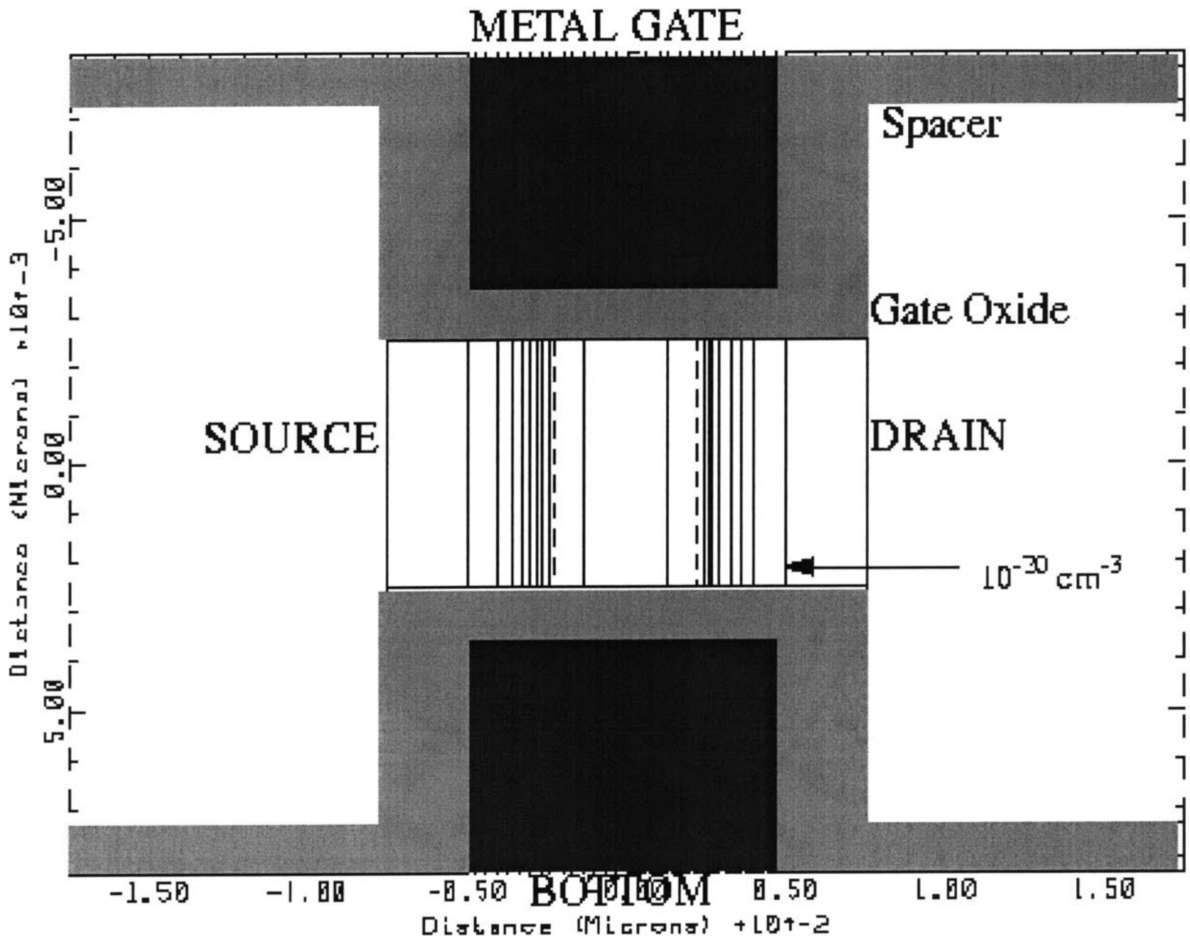

Figure 5.4: Template MOSFET topology for DG and SG architectures exhibiting raised S/D, spacer, gates, and undoped substrate 2-D cross-section of the device, here with L = 10 nm.

For the template DG [65] and SG, a “hourglass” structure is employed with undoped body. As in the Bulk case, the simulation structure itself includes internal parasitic capacitance and resistance components, here determined by the 2.5 nm spacer and 10 nm extended S/D region per side. Again, the S/D extension doping is assumed to fall-off at the gate edge at 1 nm/dec as shown in the contours of Fig. 5.4. In choosing the vertical dimensions, it is important to realize that all film thicknesses will eventually be limited by the variation control over the atomic sized layers. Thus, an effective t_{ox} value of 0.825 nm that is reasonable to process was used. The silicon film thickness, T_{Si} (which is essentially SOI [66] for the SG case), was set at 5 nm which is thin enough [67] to provide good gate control without worrying about thin-Si changes in carrier effective mass [68].

5.2 Metrics for Power and Performance

To recapitulate, one may assume that the vertical dimensions of the gate dielectric and substrate film in the final generation transistor will likely be fixed due to finite thickness variation control. With a given set of process capabilities, the only handles remaining to improve the device electrical properties will be shrinking the gate length, L, and modifying the gate workfunction, ϕ, of the gate material. The technique of "overscaling" [69] explained herein has been developed to probe the merit of NMOSFET designs that deviate from the "well-tempered" ideal. PMOS designs can be similarly analyzed using the analogous voltages.

Before launching into the scaling study, it should be clear that all charge and current data will be acquired through device simulation. Because the transport length scales are nearly ballistic, only direct solutions of the Boltzman Transport Equation will have accurate physical meaning. However, because direct solutions take prohibitive time to run, simulations with the macroscopic EB model are tuned to give realistic data. In particular, a carrier τ_w of 0.15 ps was used and agrees with I-V characteristics for DG NMOS of Monte Carlo simulation [70]. Also in the interest of speed, no quantum mechanical effects were included but were instead modeled into the effective gate dielectric.

To illustrate the operating conditions of the sub-20 nm MOSFETs under investigation, Fig. 5.5 plots the drop in the threshold voltage of the template DG. The definition used is

$$V_t = V_{GS} \qquad where \qquad I_D(V_{GS}) = \frac{W}{L}10^{-7}(A) \qquad \text{Equation 5.3}$$

Despite the thin substrate, it is clear that the devices will fall from the well-tempered regime into the overscaling range. The conventional thinking has been that there exists a generalized scale length [71] (approximately L = 16 nm for the DG device with the stated vertical dimensions) beyond which short channel effects prevent proper device operation. However, in the overscaling range where threshold voltage loses its phyical meaning, a change in viewpoint is necessary.

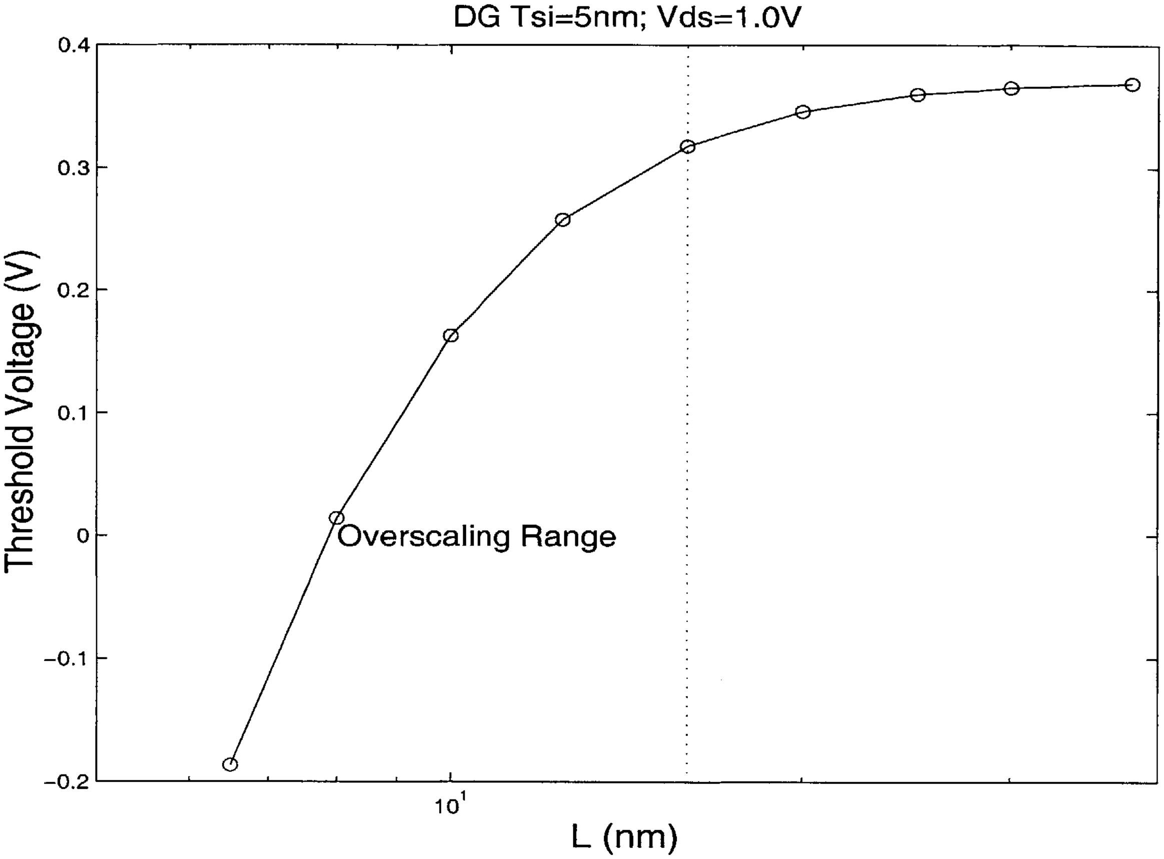

Figure 5.5: The threshold voltage roll-off versus channel length for the template DG device with T_{Si} = 5 nm at V_{DS} = 1.0 V becomes severe shorter than L = 16 nm.

In fact, the qualities of the device with the most physical relevance are the I_{off} and I_{on}. The logical procedure is then to develop reasonable quantitative metrics for power consumption and device performance to later examine how they will trade-off with scaling L and ϕ. As displayed in Fig. 5.6, the I-V characteristics at a particular V_{DD} with a mid-gap gate were simulated for a range of L from 38 nm to 6.5 nm, well into the overscaling range. Instead of repeating the simulations for different workfunctions, the I-V curve for a ϕ of +0.1 V relative to mid-gap (where a polysilicon gate doped at n+ is -0.5 V and p+ is +0.5V) was extracted by adding +0.1 V to the V_{GS} axis.

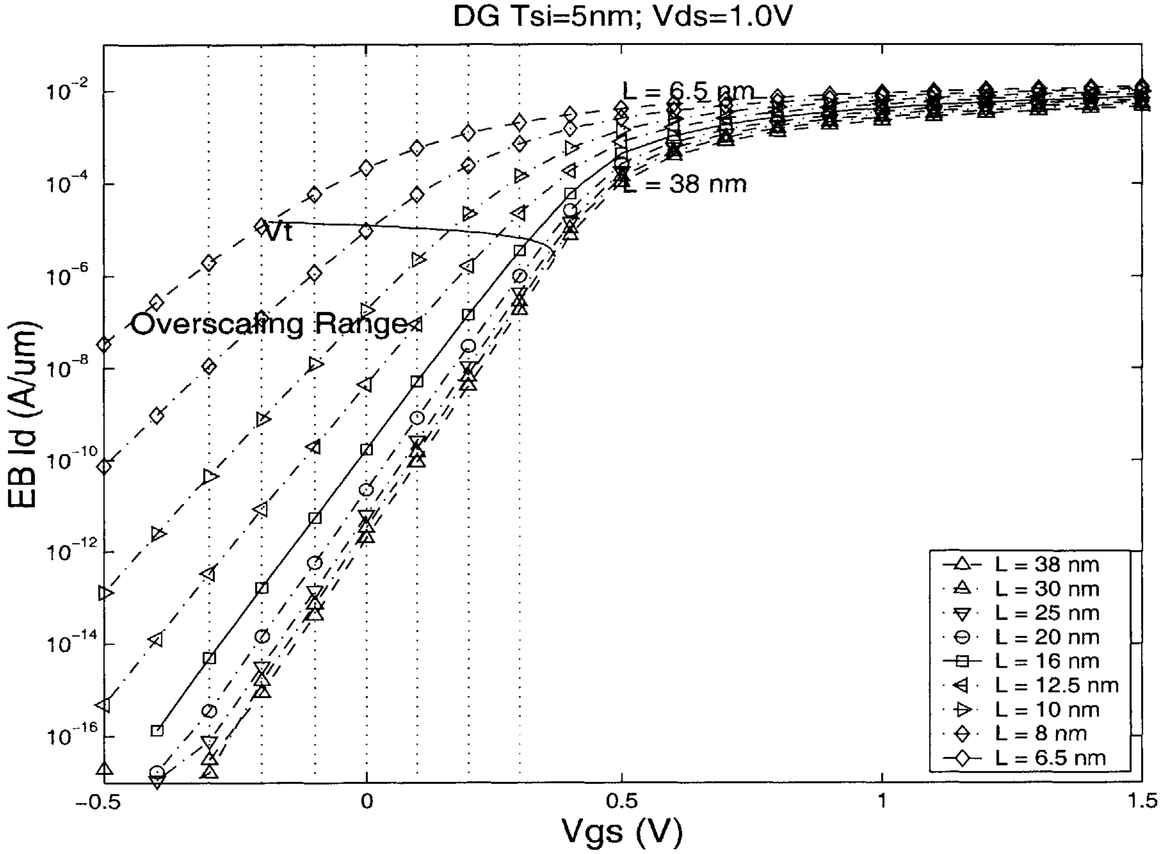

Figure 5.6: I-V characteristics at V_{DS} = 1.0 V for the template DG device with mid-gap gates. Gate lengths vary from 38 nm down into the overscaling range to L = 6.5 nm. Multiple workfunctions are extracted by shifting V_{GS}.

Before creating an estimate for the power, the issue of process variation must be addressed. In this study, an absolute rather than percentage gate variation, ΔL, proves more relevant because devices of different nominal lengths will likely be fabricated on the same chip and subjected to the same engineering level of process control. A fair assumption describes the density of variations by a gaussian with $3\sigma = \Delta L$ as in Fig. 5.7. Thus,

$$Power_{\mathrm{StandBy}} = \frac{V_{DD}}{\sqrt{2\pi\sigma}} \cdot \int_{-\infty}^{\infty} I_{off}(L) e^{-\left(\frac{L - L_{nominal}}{\sqrt{2}\sigma}\right)^2} dL \qquad \text{Equation 5.4}$$

Despite a low density of devices in the tail of the gaussian, including these shorter lengths becomes significant as they clearly contribute most of the power dissipation because of their higher off currents.

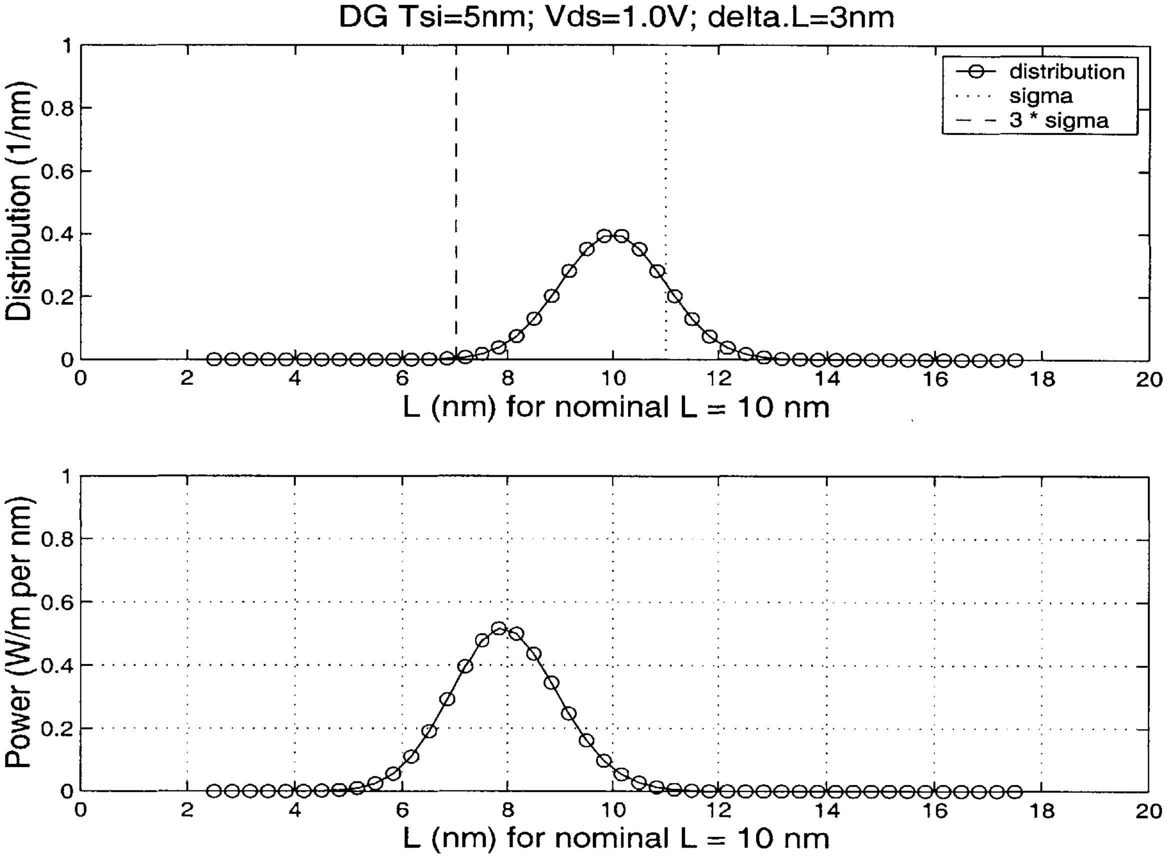

Figure 5.7: The density of devices with variation $\Delta L = 3$ nm around a nominal L = 10 nm is assumed to be a gaussian distribution with $3\sigma = \Delta L$. The peak of the power distribution is skewed due to the rapidly increasing I_{off} at shorter L.

Assuming a low fractional time usage of any particular device, the stand-by power will dominate over the dynamic power dissipation,

$$Power_{Dynamic} = C_{ox} \cdot V_{DD}^2 \cdot f \qquad \text{Equation 5.5}$$

For L ~ 10 nm, with approximately C_{ox} of order 1 fF/μm, frequency of order 10 GHz and 10% usage gives 1 $\mu A / \mu m$ of dynamic current which is lower than road-map I_{off} values.

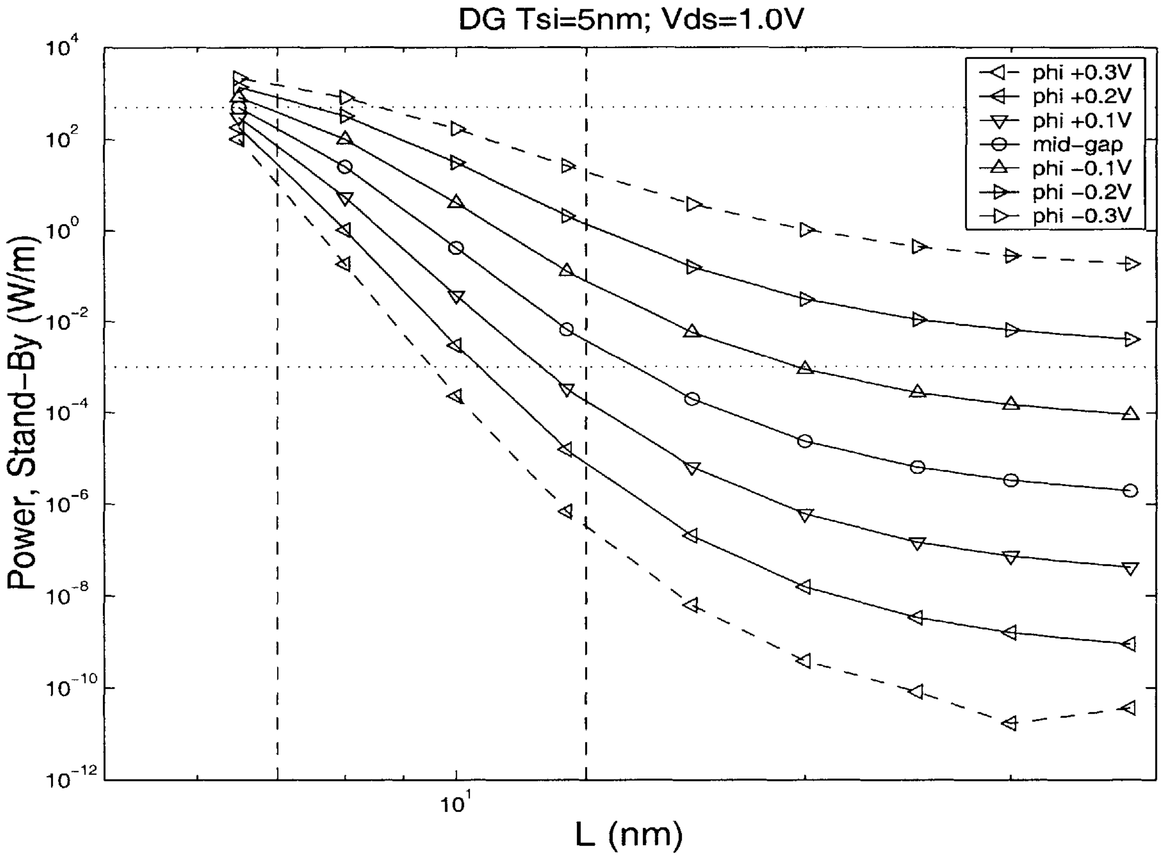

Figure 5.8: Stand-by power versus gate length of the template DG device at V_{DD} = 1.0 V for various workfunctions from -0.3 V to +0.3 V tracks the exponential rise in off current.

Having developed this metric for power, its behavior versus L is generated in Fig. 5.8 for various ϕ around mid-gap. The slope of the increased dissipation at shorter lengths becomes steeper as the workfunction increases; higher ϕ devices start at lower I_{off} and are more susceptible to short channel effect penalties. Also, the horizontal dotted lines indicate the useful range of power operation. The upper boundary is set at the point that the transistor will barely switch where off-current is about 10% of on-current. The lower boundary of 1 mW/m corresponds to MOSFETs operating at V_{DD} = 1 V with about 1 nA/μm of leakage.

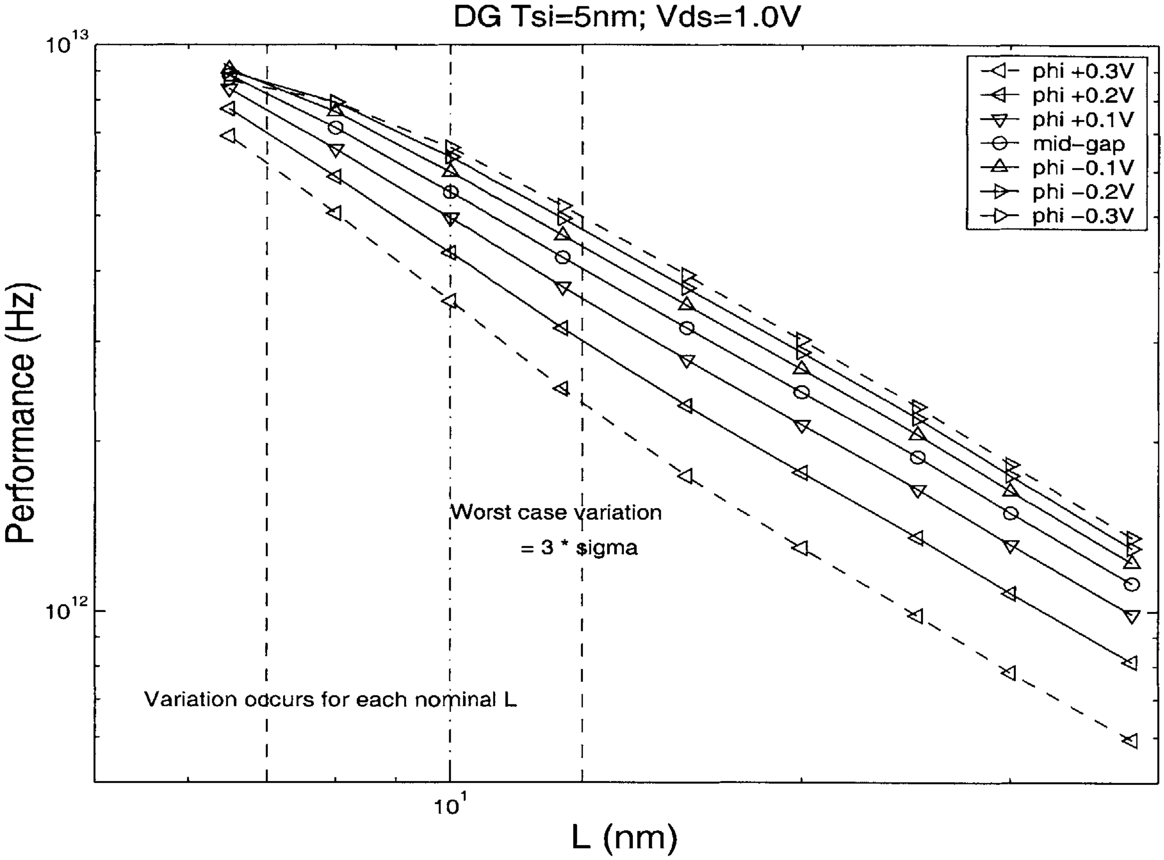

Figure 5.9: Performance versus gate length of the template DG device at V_{DD} = 1.0 V for various workfunctions from -0.3 V to +0.3 V.

Furthermore, having obtained the data for all permutations of L and ϕ, a performance metric is created that reflects the I/CV "frequency" of a CMOS inverter in a ring oscillator,

$$F = \frac{(I_{on} - I_{off})}{(Q_{on} - Q_{off})} \qquad \text{Equation 5.6}$$

where the I_{on} that carries the charge Q_{on} off the output node until Q_{off} remains is mitigated by the static leakage I_{off}. The net change in charge is taken as the difference in gate charge from on (V_{GS} high, V_{DS} low) to off (V_{GS} low, V_{DS} high). Finally, this performance is plotted in Fig. 5.9 at V_{DD} = 1.0 V.

5.3 Overscaling Trade-offs

With the overscaling technique in hand, it becomes only a matter of running simulations to produce data that will reveal the way the chosen performance and power metrics will trade-off as a function of L and ϕ. A chief advantage of examining the data in this format is that many studies can be run with comparisons between various process and operating conditions for the device architectures (DG, SG, or Bulk) in question.

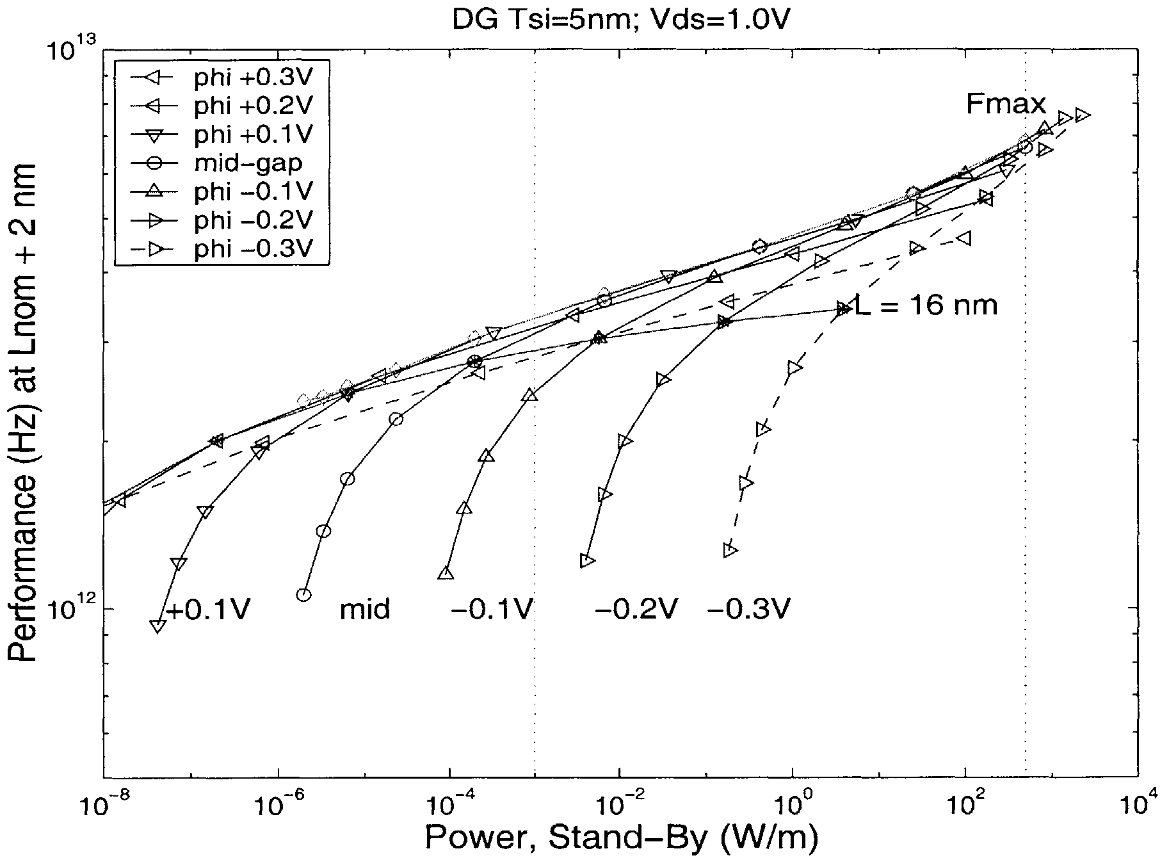

Figure 5.10: The trade-off with ΔL = 2 nm of the template DG MOSFET at V_{DD} = 1.0 V for various ϕ. For each ϕ curve, the performance (as in Fig. 5.9) and power (as in Fig. 5.8) associated with each L ranging from 38 nm to 6.5 nm is plotted. F_{max} is the performance envelope.

The results of an overscaling study on the template DG MOSFET are displayed in Fig. 5.10 for an assumed ΔL = 2 nm and V_{DD} = 1.0 V. The line described as F_{max} represents the maximum attainable performance in the power range via any combination of gate size and material. Within the useful power range (bounded by the vertical lines), it seems that the trade-off curve for a single workfunction near mid-gap closely tracks the performance envelope. On the other hand, if one were to stop scaling at a particular node (such as L = 16 nm for DG) to preserve electrostatic integrity and successively lower ϕ to increase the overdrive, the resulting trade-off curve would fall below the overscaling line.

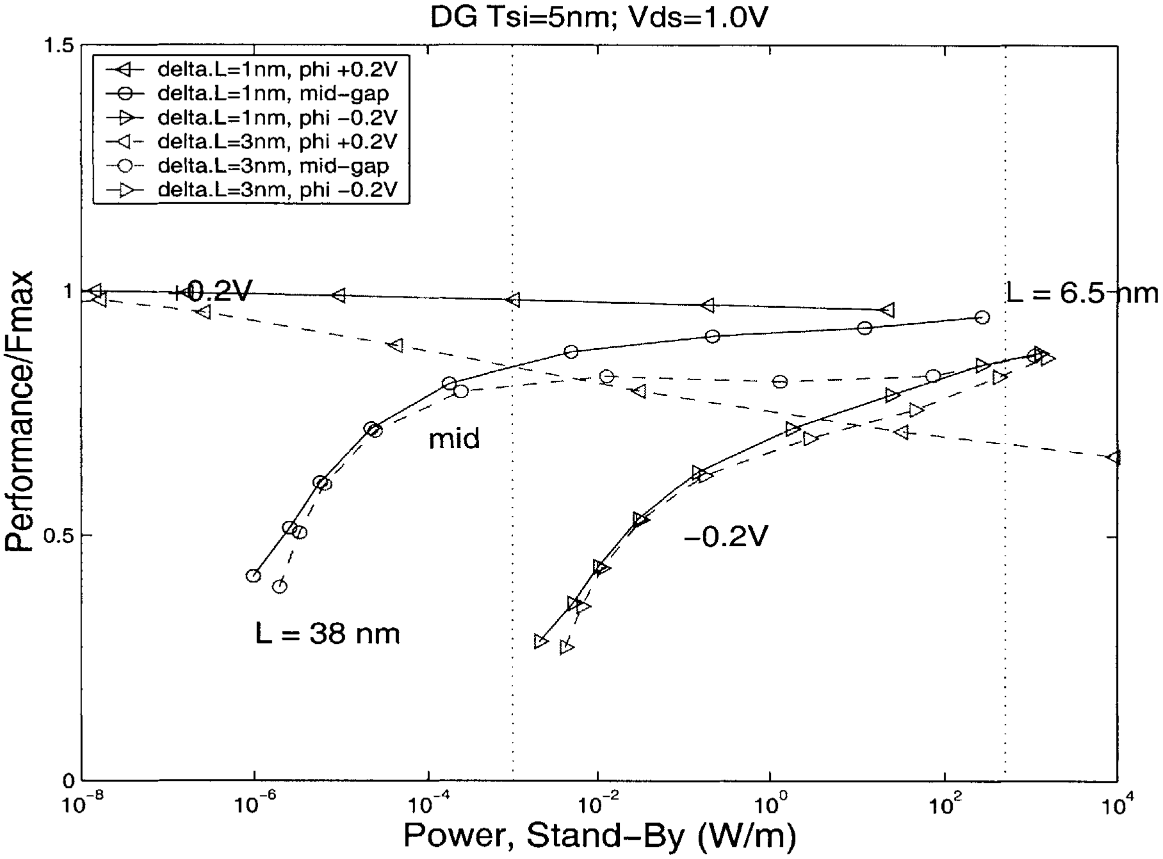

Figure 5.11: Ratio of performance to F_{max} vs. power of the template DG MOSFET at V_{DD} = 1.0 V for various ϕ plotted as a function of L, with ΔL of 1 nm (solid lines) and 3 nm (dashed).

The significance of comparing trade-offs under different conditions is best illustrated when normalizing to the performance envelope (assuming no process variation) as in Fig. 5.11. The plot verifies the essential result that one workfunction ensures the best possible performance. The ramifications of this result are compelling; industry can achieve optimal trade-off with only having to develop a single gate material. Another interesting trend rears itself in the sensitivity of performance to process variation. Apparently, the reductive effects of bigger variation (ΔL = 3 nm as opposed to 1 nm in Fig. 5.11) is more prominent in the more positive ϕ devices where short channel effects were initially more controlled.

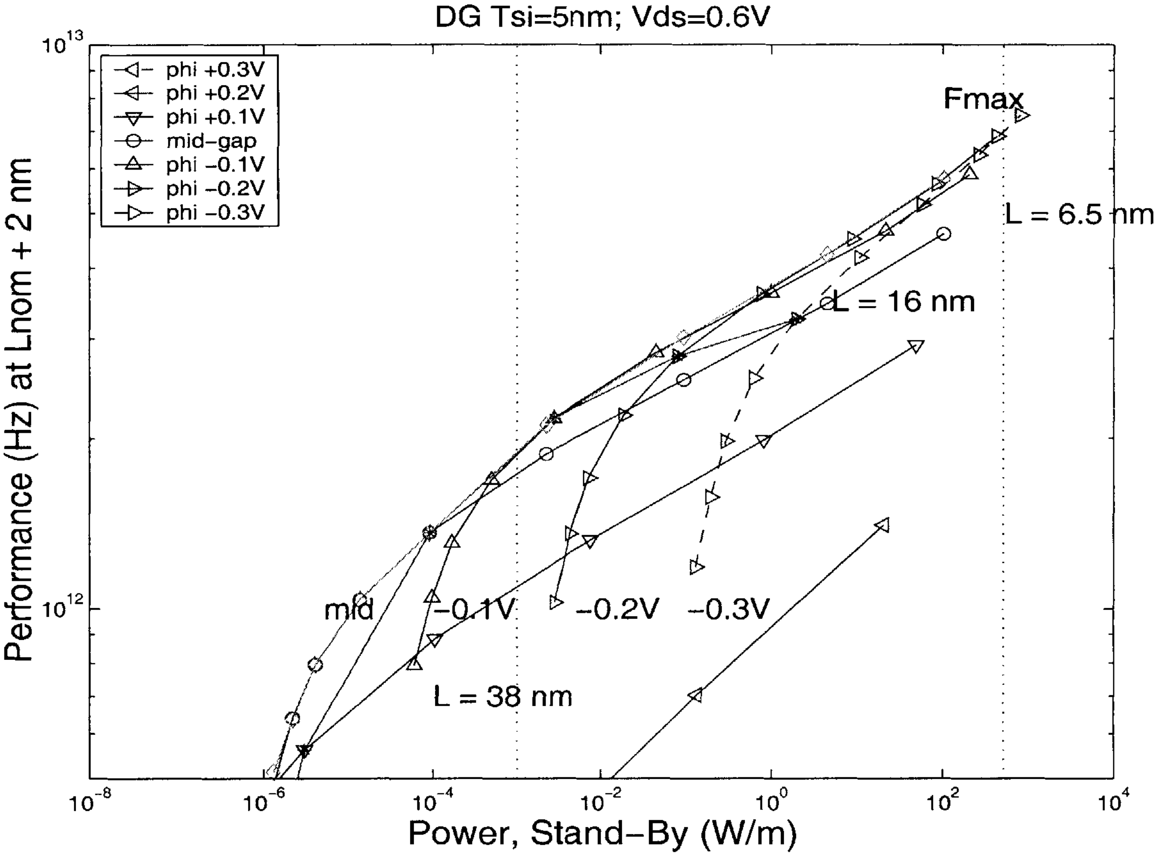

Figure 5.12: The trade-off of performance vs. power of the template DG MOSFET at V_{DD} = 0.6 V for various ϕ plotted as a function of L, with ΔL = 2 nm.

Yet another important variable in any overscaling study is the operating bias. The results for the template DG at V_{DD} = 0.6 V are plotted on the same scale in Fig. 5.12. Not surprisingly, the performance for any L becomes severely degraded above a certain ϕ level because the higher

V_t yields a lower overdrive voltage. However, the overscaling principle stills seems to apply in that choosing ϕ = -0.1 V has better performance versus power characteristics than utilizing a fixed L. Using a non-mid-gap workfunction to optimize the scaling behavior of this NMOS technology will require the development of a complementary ϕ of opposite sign for the PMOS.

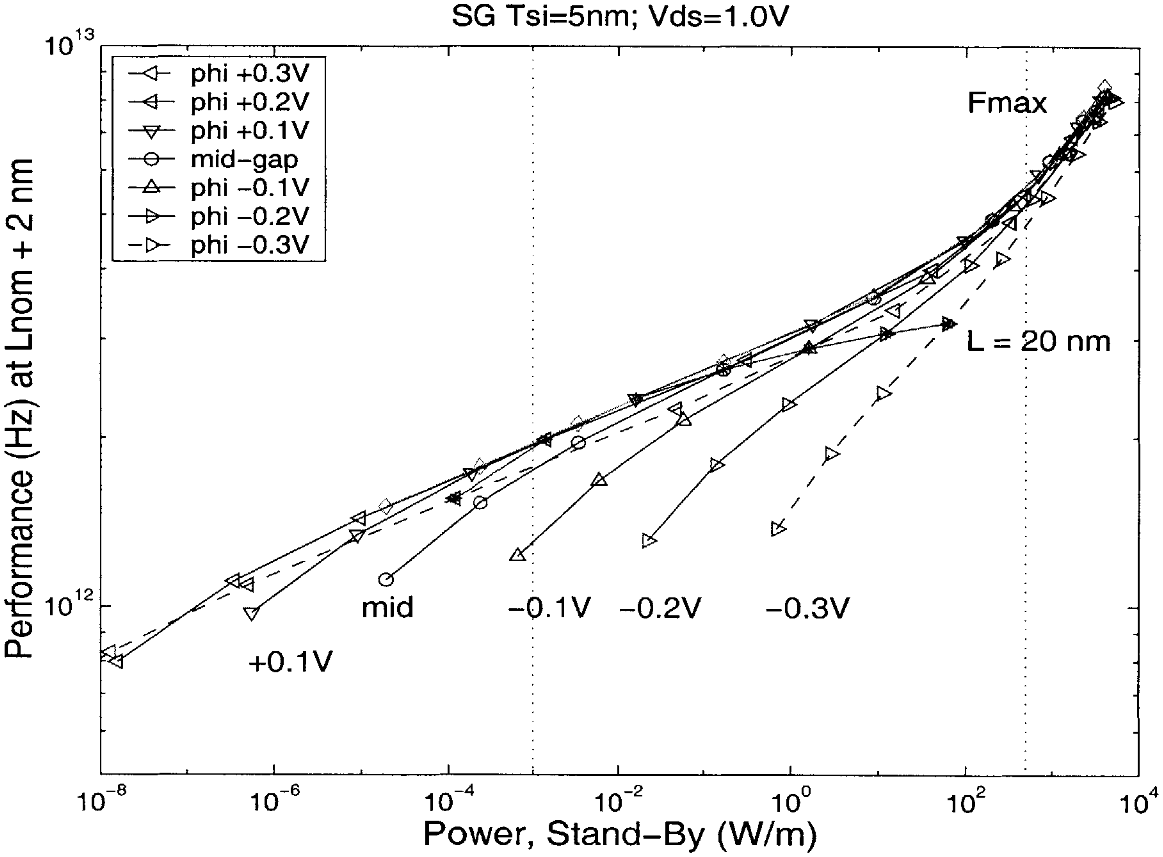

Figure 5.13: The trade-off of performance vs. power of the template SG MOSFET at V_{DD} = 1.0 V for various ϕ plotted as a function of L, with ΔL = 2 nm.

While the electrical characteristics of a back gate MOSFET remain very controlled for thin enough substrates, Fig 5.13 presents a valuable comparison with the template SG fully-depleted SOI (FDSOI) at V_{DD} = 1.0 V and ΔL = 2 nm. Again, overscaling near mid-gap mimics following the F_{max} curve. The main difference here is that all the curves are bent/shifted towards the higher power dissipation values because of the reduced gate control. Also, changing only ϕ

for the minimum well-tempered node for planar devices with single-gate, around L = 20 nm, does not give the best trade-off.

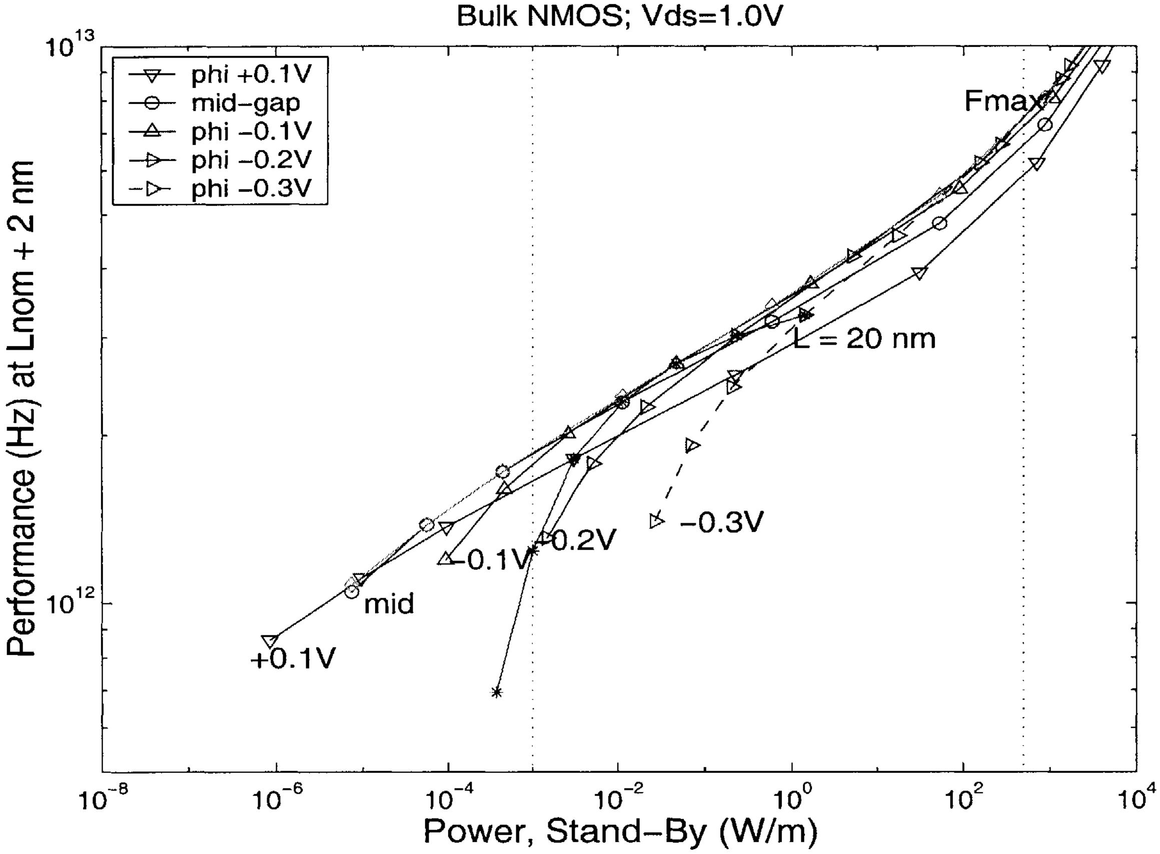

Figure 5.14: The trade-off of performance vs. power of the template Bulk MOSFET at V_{DD} = 1.0 V for various ϕ plotted as a function of L, with ΔL = 2 nm.

To round out the study of sub-20 nm scaling, the same trade-off curves are plotted for the template Bulk device for the same process and operating conditions as in Fig. 5.14. As expected, the values for stand-by power appear similar to the SG case, since the 2-D doping was chosen to elicit a similar level of off-current control. The fact that just one workfunction for NMOS provides near optimal performance as the devices of different architectures scale offers relief from cxcessive materials development for future technologies.

5.4 Optimal Double-Gate Considerations

The simulation data incorporated into the overscaling technique propose an interesting direction for the fabrication of sub-20 nm MOSFETs. Namely, picking a mid-gap workfunction and using our expected handle on lithography to scale the DG device to the limits wins. Of course, the previous investigation has made a number of assumptions that deserve more critical analysis before embarking on a mission to optimize the DG. A direct comparison should clarify the level of dominance of one architecture over another. Moreover, one must account for the sensitivity of the template structure to variations in vertical dimensions, S/D doping, and parasitic resistance.

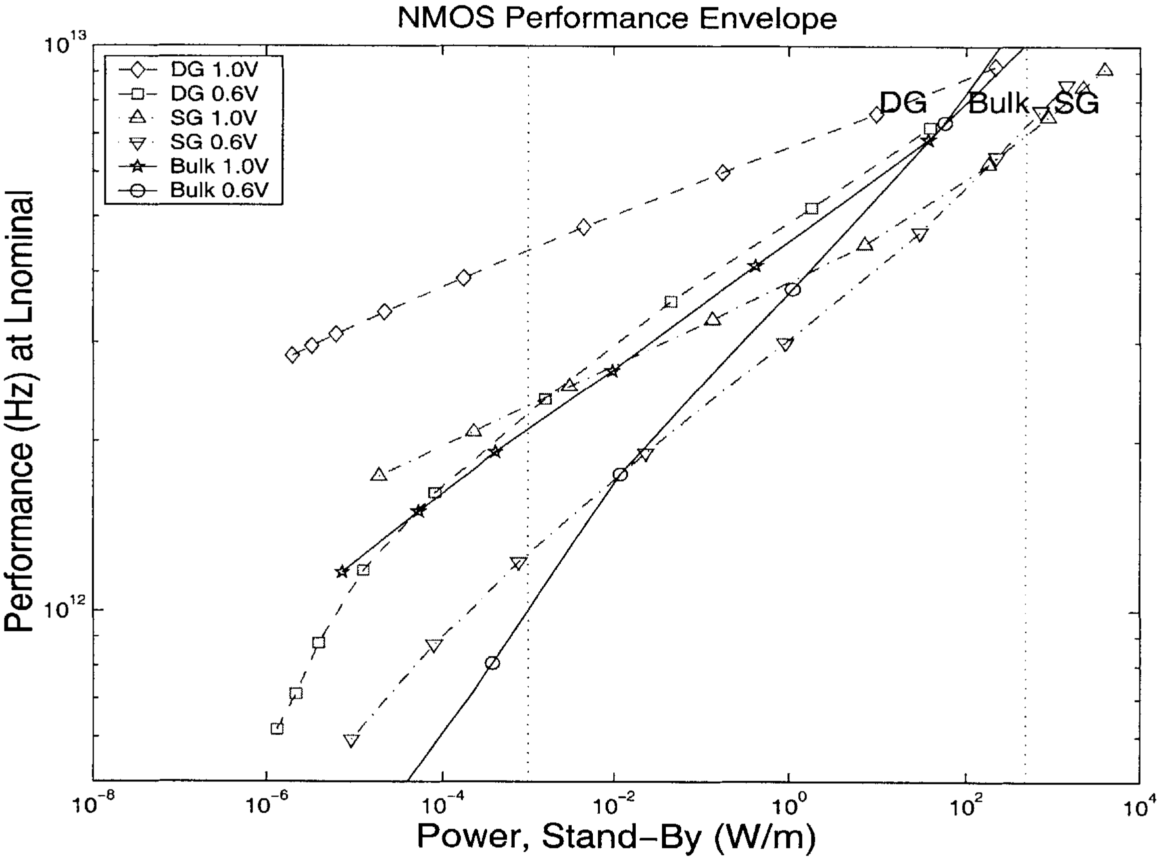

Figure 5.15: The performance envelope trade-off (assuming no ΔL) versus stand-by power for the template DG, SG, and Bulk devices at operating biases of 1.0 V and 0.6 V.

Fig. 5.15 summarizes the effects of varying MOSFET architecture and operating points by evaluating the performance envelopes for each case. One might argue that DG has double the current drive but twice the gate capacitance and hence the CV/I delay will be the same as for a SG. However, the DG still wins when contrasting the performance for a *specific* power level because it possesses greater electrostatic control. Consideration of gate length variation does not alter this finding significantly. In determining the best V_{DD} voltage, the trend begins by delivering higher performance for higher overdrive. However, as each architecture nears the upper end of the useful power range, the example of the 0.6 V bias converges to and even surpasses the 1.0 V where the large DIBL forces the I_{off} so high that it degrades the frequency metric. Although the tradition of scaling down the drive voltage helps the high end devices, it does not have to be done so aggressively.

No analysis of sub-20 nm devices would be complete without a discussion of the role of quantum leakage mechanisms. While the undoped FDSOI devices would not exhibit appreciable band-to-band tunnelling, the strongly doped Bulk device would sustain higher levels of I_{off}. Thus, including this effect on the Bulk in Fig. 5.15 would increase the power consumption and shift its trade-off curves close to the SG results. Next, the direct tunnelling mechanism from source to drain poses a threat as the gate length shrinks below 10 nm; but theory [19] suggests that thermal current noise still washes out the S/D tunnelling (predicted via simulation of the conduction band barrier) magnitudes at or less than order 1 $\mu A / \mu m$. Finally, gate current leakage will play a similarly muted role as long as the effective dielectric thicknesses stabilize.

Taking a step back to see the big picture, why should the scaled mid-gap DG outperform its competition? Abstractly visualizing the simulation structure yields regions of both differing geometry and materials (equivalently, instrinic voltages). It is precisely the manner in which these factors contribute to the solution of the Poisson and transport equations that determines the optimal. If the ideal DG becomes too costly to mass produce, a number of alternate architectures become desirable. For instance, technologies such as Fin-FET are analyzed as giving hybrid characteristics between the DG and SG. Another promising scheme is to allow for misalignment with a back gate length larger than the top gate; in this case, the detrimental boost in capacitance would trade-off with ease of fabrication.

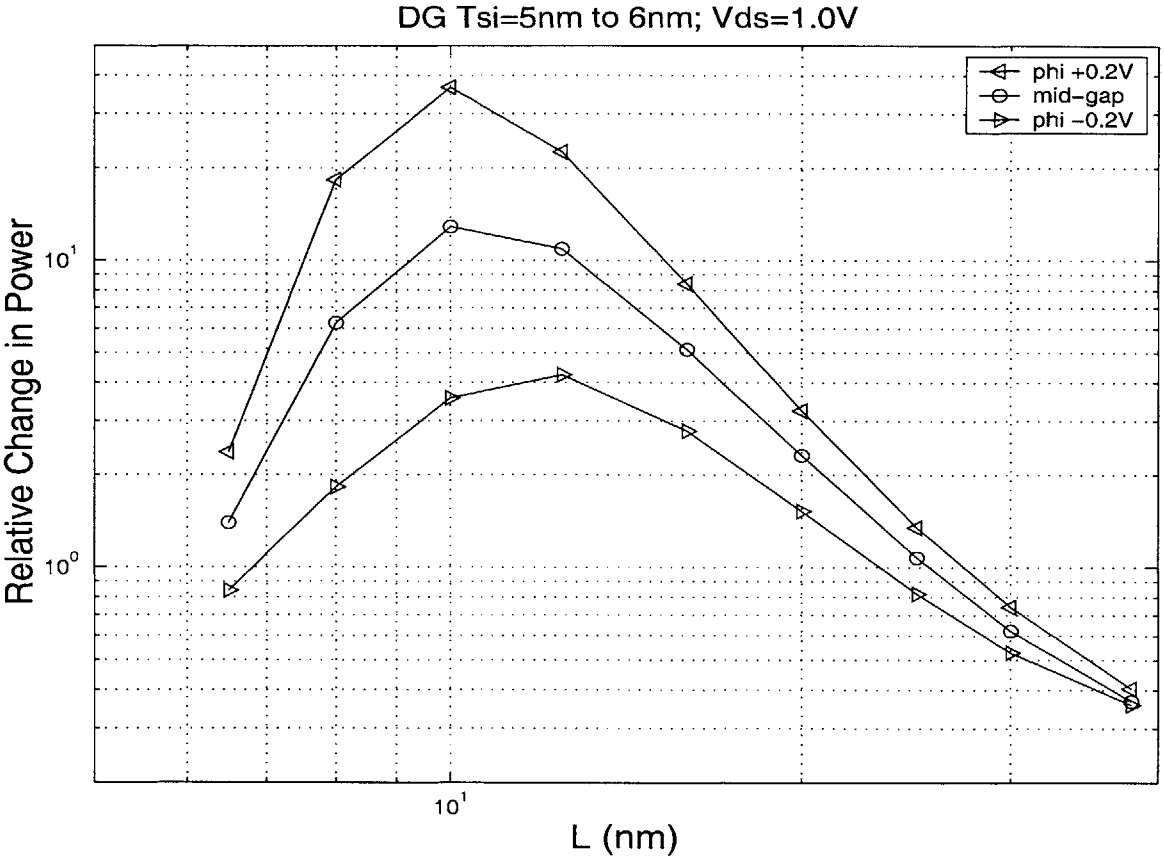

Figure 5.16: Relative change in stand-by power versus gate length for a 1 nm process variation in T_{Si} in the template DG at V_{DD} = 1.0 V for various ϕ from -0.2 V to +0.2 V.

A peculiar yet appropriate assumption throughout the study has been the constant vertical dimensions. Although process repeatability prevents traditional scaling of S/D thickness, the engineer must account for the variance in the substrate. As depicted in Fig. 5.16, a 1 nm increase of T_{Si} in the template DG causes an order of magnitude increase in power over the entire range of nodes. The longer devices suffer less from 2-D effects and are less sensitive to T_{Si} fluctuations while the shortest devices are less dependent due to already high off-current. Also, the initially higher V_t, ϕ = +0.2 V device shows a harder hit in relative power change in the mid-range L.

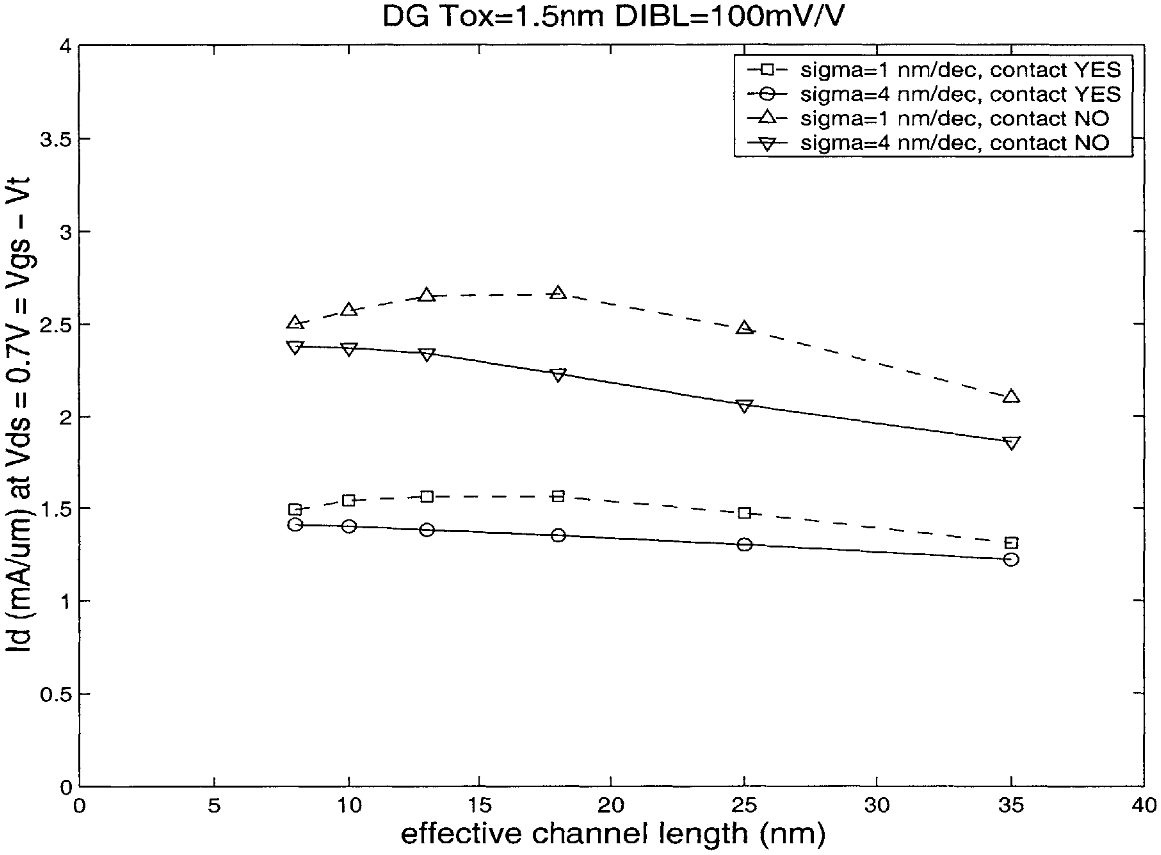

Figure 5.17: Impact on drive current of changing the abruptness σ_x of the S/D extensions as a DG device scales with constant DIBL and overdrive, V_{GS} - V_t = 0.7 V, for conditions of having YES/NO contact resistance.

Another test for the DG MOSFET is whether it can handle a S/D doping less abrupt than 1 nm/dec. The study of Fig. 5.17 assumes a fixed electrostatic quantity as a DIBL (to isolate the impact on the carrier velocity) of 100 mV/V as the length scales by simultaneously thinning T_{Si}. Simulating at constant overdrive, the primary effect of less abrupt S/D junctions is to depress I_{on} since they have a longer region of resistance. The effect is more pronounced when contact resistance is neglible such that it does not mask the intrinsic resistance from the more extended S/D. Further, the drive current of the devices with more abrupt σ_x scale up as transistor size shrinks but appear to plateau earlier because they are more sensitive to the increased resistance of the thin substrate.

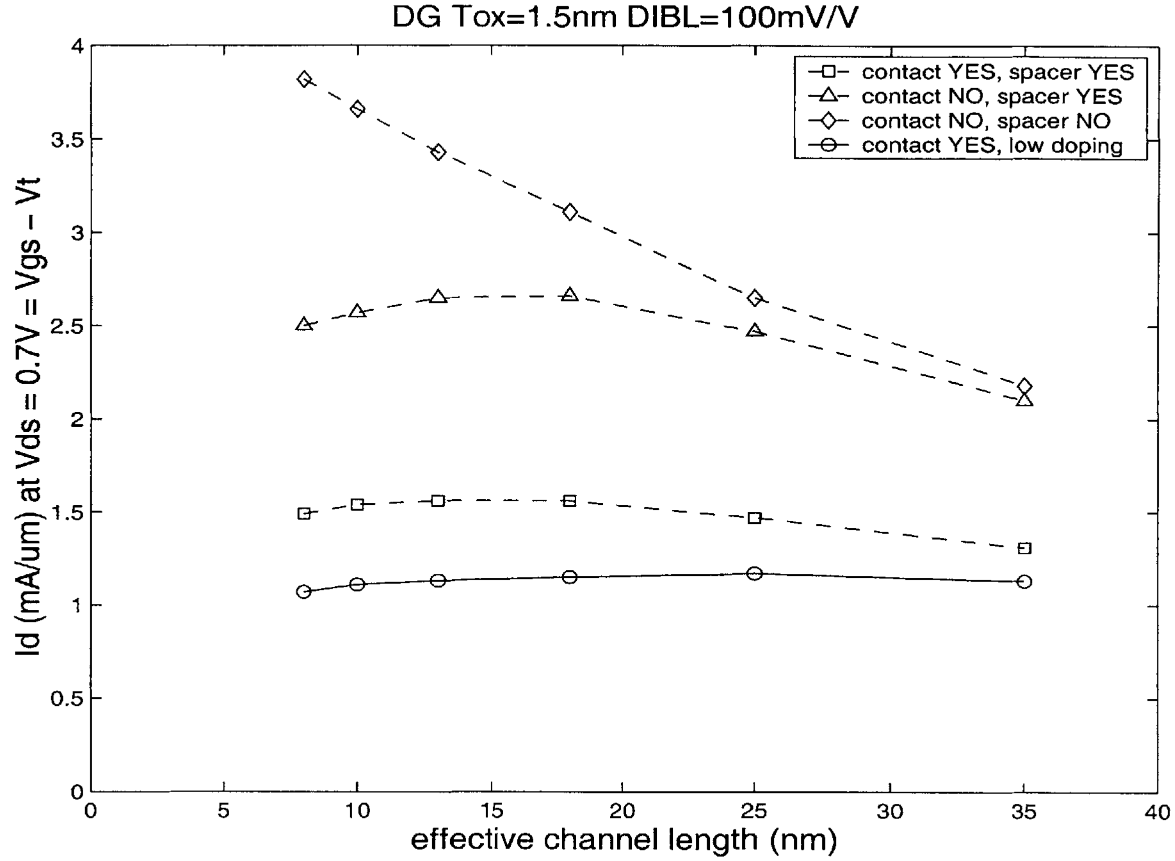

Figure 5.18: Impact on I_{on} at constant overdrive and $V_{DS} = 0.7$ V as a DG device scales with constant DIBL of having 4×10^{-7} Ωcm^2 (assuming a 10 nm long S/D) or NO contact resistance, YES/NO to having a 10 nm spacer, or having a 2×10^{19} cm^{-3} rather than 1×10^{20} cm^{-3} S/D.

More generally, the problem of parasitics hindering the optimal device drive demands probing. Fig. 5.18 demonstrates the effects of different types of increased resistance on a DG kept at constant DIBL. Attempting to eliminate the contact resistance, the predominant parasitic (currently on the order of 4×10^{-7} Ωcm^2), would drastically improve I_{on}. To accentuate the effect of the resistance of carrier flow through the S/D, taking away the spacer region showed improvement while lowering the doping level to 2×10^{19} cm^{-3} exhibited suppressed drive. For the overscaling study, the parasitic capacitance was modeled into the fringing fields of the MOS structure, assuming that it dominates over the interconnect.

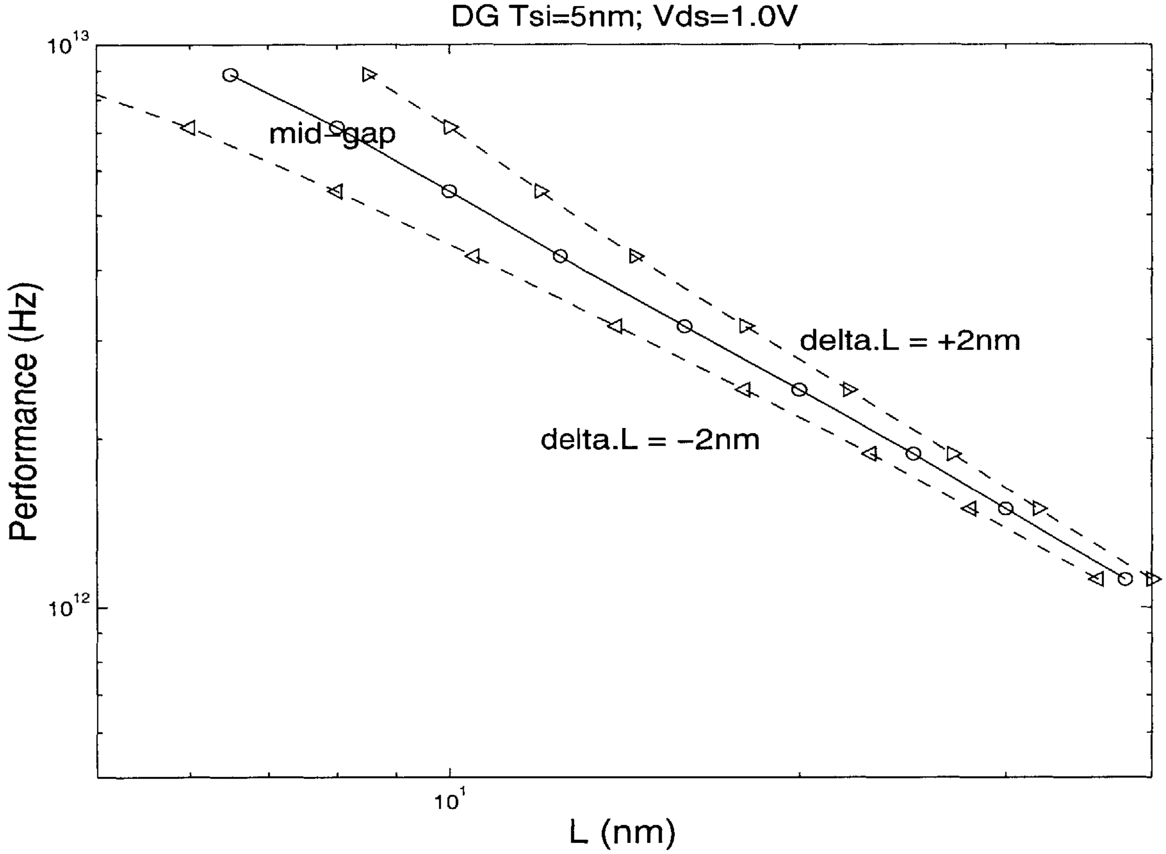

Figure 5.19: Performance vs. L for the template DG at V_{DD} = 1.0 V bounded by the curves for ΔL of -2 nm and +2 nm which leads to clock skew.

In conclusion, this foray into the sub-20 nm regime offers more than a path towards optimally scaled MOSFETs: it provide a simulation methodology for tuning this path as new estimates of processing capability become available. Assuming that integrating a DG with low parasitics on the S/D will become feasible, Fig. 5.19 plots the predicted switching speed. The advantages of this architecture are 1) the substantial gate control over the carriers should improve the ultimate performance and 2) the minimal channel doping should alleviate Coulomb scattering and dismiss dopant fluctuation issues. A compact modeling drawback that must be designed around is that inevitable process variation leads to larger skew between the high and low end of the shortest devices.

Chapter 6
Conclusions

6.1 Summary

This project began as a follow-up to the work done on inverse modeling in the subthreshold I-V regime. Since then it has evolved into investigations of various transport and device issues. Behind it all has been a core of physical and mathematical principles that actually form the basis for many scientific disciplines. Here in the electrical engineering of computer devices, a mix of quantum mechanical energy bands, Maxwell's equations of electrodynamics and the laws of particle motion characterize the relevant behavior. The analytical castings of these phenomena engender their own numerical formalisms for tangible solutions. In order to verify the robustness of the inverse modeling approach, the sensitivity of data as a function of parameter space placed requirements on the electrical signature. Furthermore, investigations demonstrated that the optimizations improve as the parameterizations become more comprehensive of structural features of the devices.

A major goal of the project was to develop inverse modeling into a comprehensive methodology for characterizing the device structure of even sub-50 nm MOSFETs. This accomplishment ties together direct probing methods, accounting for the gate stack and surrounding dielectrics, analysis of I-V data that reflect the 2-D channel doping, and C-V data sensitive to the S/D extension regions. Of course, several generations of advanced industry devices were instrumental in obtaining careful measurements and verifying the work over several experiments. Moreover, the inverse modeling results provide many device profiles that form a basis for testing. For instance, progress has led to the calibration of diffusion models in a standard process simulator.

The culmination of the prior device characterization work has contributed to the understanding of carrier transport. Theoretical underpinnings suggest that macroscopic transport approximations remain valid into the sub-50 nm regime. First, using measured data the mobility due to impurity and surface scattering mechanisms were extracted. Additionally, a methodology for calibrating the transport models was developed that accounts for parasitics in the strong inversion I-V data. The crowning achievement of this study was the realization that for several different t_{ox} technologies operating at various biases, a single set of Energy Balance parameters produced good agreement over a wide experimental range of gate length nodes.

In conclusion, applying relevant physics to the characterization of VLSI devices and their transport behavior serves itself as a basis for investigating the best path for scaling to sub-20 nm. An analysis reveals the difficulties of achieving well-tempered MOSFET operation among architectures such as DG, SG and Bulk. To study the design trade-offs, realistic metrics for power dissipation and expected performance were developed. By examining these characteristics for different combinations of L and ϕ, it appears that overscaling with a fixed workfunction gives the best results. Thus, one can optimize and discuss the features of a future DG device.

6.2 Suggestions for Future Work

A number of avenues of potential or tangential study have opened as a result of this project. For the extreme sub-50 nm devices of the future, a more exact treatment of quantum mechanical carrier confinement must be utilized such as direct implementation of the Schroedinger equation. With regards to making the optimization loop fully automated, setting up a design of experiments running on parallel processors would be a better way of coming up with an initial guess.

On the inverse modeling front, in terms of choosing the 2-D doping representation functions themselves, gaussians with tails rotated as per implant angle would be most realistic and hence have the most sensitive optimization. Also, with more specific dopant and thermal details of processing technologies, one might derive a universal calibration for process simulation.

In the realm of transport physics, there are several possible improvements. As new generations of deeply scaled MOSFETs become available, it will become increasingly important to calibrate more exact transport solutions such as Monte Carlo to the experimentally observed data. In addtion, the various mobility phenomena that emerge for alternate semiconductor substrates should be explored and quantified.

Lastly, charting the path for scaling transistors to the end of the Road Map with the optimal mix of performance and cost efficiency presents a fruitful challenge. Conducting further overscaling investigations while accounting for expected circuit problems promises advances in compact modeling.

References

[1] J. S. McMurray, J. Kim, and C. C. Williams, “Quantitative measurement of two-dimensional dopant profile by cross-sectional scanning capacitance microscopy”, J. Vac. Sci. Technol. B 15(4), p. 1011-1014, 1997.

[2] P. De Wolf, W. Vandervorst, H. Smith, and N. Khalil, “Comparison of two-dimensional carrier profiles in metal-oxide-semiconductor field-effect transistor structures obtained with scanning spreading resistance microscopy and inverse modeling”, J. Vac. Sci. Technol. B 18(1), p. 540-544, 2000.

[3] R. Alvis, S. Luning, L. Thompson, R. Sinclair, P. Griffin, ‘Physical characterization of two-dimensional doping profiles for process modeling”, J. Vac. Sci. Technol. B 14(1), p. 231-235, 1996.

[4] W. Vandervorst, T. Clarysse, N. Duhayon, P. Eyben, T. Hantschel, M. Xu, T. Janssens, H. De Witte, T. Conard, J. Deleu, G. Badenes, “Ultra shallow junction profiling”, IEDM proceedings, p. 429-432, 2000.

[5] G. Ouwerling, “Physical Parameter Extraction by Inverse Device Modeling: Application to One- and Two-Dimensional Doping Profiling,” Solid-State Electronics 3(6), p. 757, 1990.

[6] A. Das, D. Newmark, I. Clejan, M. Foisy, M. Sharma, S. Venkatesan, S. Veeraraghavan, V. Misra, B. Gadepally, L. Parrillo, “An Advanced MOSFET Design Approach and a Calibration

Methodology using Inverse Modeling that Accurately Predicts Device Characteristics", IEDM proceedings, p. 687-690, 1997.

[7] N. Khalil, J. Faricelli, D. Bell, and S. Selberherr, "The Extraction of Two-Dimensional MOS Transistor Doping via Inverse Modeling", IEEE EDL-16 (1), p. 17-19, 1995.

[8] Z. K. Lee, M. B. McIlrath, and D. A. Antoniadis, "Inverse Modeling of MOSFETs using I-V Characteristics in the Subthreshold Region", IEDM proceedings, p. 683-686, 1997.

[9] C. Y. T. Chiang, Y. T. Yeow, and R. Ghodsi, "Inverse Modeling of Two-Dimensional MOSFET Dopant Profile via Capacitance of the Source/Drain Gated Diode", IEEE TED-47 (7), p. 1385-1392, 2000.

[10] B. Yu, H. Wang, O. Milic, Q. Xiang, W. Wang, J. X. An, M-R. Lin, "50nm Gate-Length CMOS Transistor with Super-Halo: Design, Process, and Reliability," IEDM proceedings, p. 653-656, 1999.

[11] N. W. Ashcroft and N. D. Mermin, Solid State Physics, 1998.

[12] K. M. Jackson, "Optimal MOSFET Design for Low Temperature Operation", Ph. D. Thesis, 2001.

[13] I. J. Djomehri, T. A. Savas, and H. I. Smith, "Zone-plate-array lithography in the deep ultra-violet", J. Vac. Sci. Technol. B 16, 3426, 1998.

[14] H. Hu, J. B. Jacobs, L. T. Su, and D. A. Antoniadis, "A Study of Deep-Submicron MOSFET Scaling Based on Experiment and Simulation", IEEE TED-42 (4), p. 669-677, 1995.

[15] J. D. Jackson, Classical Electrodynamics, 1998.

[16] D. J. Griffiths, Introduction to Quantum Mechanics, 1995.

[17] T. Janik and B. Majkusiak, "Analysis of the MOS Transistor Based on the Self-Consistent Solution to the Schrodinger and Poisson Equations and on the Local Mobility Model", IEEE

TED-45(6), p. 1263-1271, 1998.

[18] 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 Electronics, Vol. 37, No. 3, p. 411-414, 1994.

[19] K. Likharev, "Sub-20-nm Electron Devices", manuscript, 2001.

[20] Y. Tsividis, Operation and Modeling of The MOS Transistor, 1999.

[21] Z.-H. Liu, C. Hu, J.-H. Huang, T.-Y. Chan, M.-C. Jeng, P. K. Ko, and Y. C. Cheng, "Threshold Voltage Model for Deep-Submicrometer MOSFET's", IEEE TED-40(1), p. 86-94, 1993.

[22] W. H. Press, et al, Numerical Recipes in C, 1998.

[23] M. S. Sharma and N. D. Arora, "OPTIMA: A Nonlinear Model Parameter Extraction Program with Statistical Confidence Region Algorithms", IEEE TCAD-12(7), p. 982-986, 1993.

[24] R. Meyer and P. Roth, "Modified Damped Least Squares: an algorithm for nonlinear optimization," Journal of the Institute of Mathematics and its Applications, vol. 9, p. 218, 1972.

[25] M. J. Sherony, "Design, Process, and Reliability Considerations in Silicon-On-Insulator (SOI) MOSFETs", Ph. D. Thesis, 1998.

[26] Y. Taur, C. H. Wann, and D. J. Frank, "25 nm CMOS Design Considerations", IEDM proceedings, p. 789-793, 1998.

[27] I. J. Djomehri and D. A. Antoniadis, "Inverse Modeling of Sub-100 nm MOSFETs Using I-V and C-V", IEEE TED-49(4), p. 568-575, 2002.

[28] C.-L. Huang, J. V. Faricelli, D. A. Antoniadis, N. A. Khalil, and R. A. Rios, "An Accurate Gate Length Extraction Method for Sub- Quarter Micron MOSFET's", IEEE TED-43 (6), p. 958-963, 1996.

[29] R. Shrivastava and K. Fitzpatrick, "A Simple Model for the Overlap Capacitance of a VLSI

MOS Device", IEEE TED-29(12), p. 1870-1875, 1982.

[30] N. D. Arora, R. Rios, and C. Huang, "Modeling the Polysilicon Depletion Effect and Its Impact on Submicrometer CMOS Circuit Performance", IEEE TED-42 (5), p. 935, 1995.

[31] TSUPREM4 manual, Avant!, 1999.

[32] R. Chau, et al, IEDM proceedings, 2000.

[33] MEDICI manual, Avant!, 1999.

[34] R. Rios, N. D. Arora, C. Huang, N. Khalil, J. Faricelli, L. Gruber, "A Physical Compact MOSFET Model, Including Quantum Mechanical Effects, for Statistical Circuit Design Applications", IEDM proceedings, p. 937-940, 1995.

[35] B. Agrawal, V. K. De, and J. D. Meindl, "Device parameter optimization for reduced short channel effects in retrograde doping MOSFET's", IEEE TED-43(2), p. 365-368, 1996.

[36] K. Yang and C. Hu, "MOS Capacitance Measurements for High-Leakage Thin Dielectrics", IEEE TED-46 (7), p. 1500-1501, 1999.

[37] R. Logan, Y. Taur, E. Crabbe, "Asymmetry in effective-channel length of n- and p-MOSFETs", SISPAD proceedings, p. 21, 1997.

[38] C. S. Rafferty, H.-H. Vuong, S. A. Eshraghi, M. D. Giles, M. R. Pinto, S. J. Hillenius, "Explanation of Reverse Short Channel Effect by Defect Gradients", IEDM proceedings, p. 311-314, 1993.

[39] J. M. Poate, D. J. Eaglesham, G. H. Gilmer, H.-J. Gossmann, M. Jaraiz, C. S. Rafferty, and P. A. Stolk, "Ion Implantation and Transient Enhanced Diffusion", IEDM proceedings, p. 77-80, 1995.

[40] I. J. Djomehri, H. Wakabayashi, and D. A. Antoniadis, "Transport Model Calibration in sub-100 nm MOSFETs via Inverse Modeling", IEEE EDL, submitted 2002.

[41] T. Grasser, H. Kosina, and S. Selberherr, “Investigation of Spurious Velocity Overshoot Using Monte Carlo Data”, SISPAD proceedings, p. 54-57, 2001.

[42] Z. K. Lee, M. B. McIlrath, and D. A. Antoniadis, “Two-Dimensional Doping Profile Characterization of MOSFETs by Inverse Modeling using I-V Characteristics in the Subthreshold Region”, IEEE TED-46 (8), p. 1640-1649, 1999.

[43] F. Assad et al., “The Drift-Diffusion Equation Revisited,” Solid-State Electronics 42(3), p. 283, 1998.

[44] A. Forghieri, R. Guerreri, P. Ciampolini, A. Gnudi, M. Rudan, and G. Baccarani, “A New Discretization Strategy of the Semiconductor Equations Comprising Momentum and Energy Balance”, IEEE Trans. CAD, Vol. 7, No. 2, Feb. 1988.

[45] K. Rahmat, J. White, and D. A. Antoniadis, “Solution of the Boltzmann Transport Equation in Two Real-Space Dimensions using a Spherical Harmonic Expansion in Momentum Space”, IEDM proceedings, p. 359-362, 1994.

[46] H. Kosina, M. Nedjalkov, and S. Selberherr, “Theory of the Monte Carlo Method for Semiconductor Device Simulation”, IEEE TED-47(10), p. 1898-1908, 2000.

[47] P. Ciampolini, A. Pierantoni, and G. Baccarani, “An Energy-Balance Model for non-Isothermal Device Simulation”, IEDM proceedings, p. 733-736, 1992.

[48] P. F. Bagwell, D. A. Antoniadis, T. P. Orlando, “Nanostructured Silicon Inversion Layers”.

[49] F. Assad, Z. Ren, D. Vasileska, S. Datta, and M. Lundstrom, “On the Performance Limits for Si MOSFET’s: A Theoretical Study”, IEEE TED-47(1), p. 232-240, 2000.

[50] H. M. Nayfeh, personal communication, 2002.

[51] C.-L. Huang, J. V. Faricelli, and N. D. Arora, “A New Technique for Measuring MOSFET Inversion Layer Mobility”, IEEE TED-40(6), p. 1134-1139, 1993.

[52] J. B. Jacobs, "Modeling of Electron Transport in Sub-100 nm Channel Length Silicon MOSFETs", Ph. D. Thesis, 1995.

[53] M. Kondo and H. Tanimoto, "An Accurate Coulomb Mobility Model for MOS Inversion Layer and Its Applicatin to NO-Oxynitride Devices", IEEE TED-48(2), p. 265-270, 2001.

[54] M. N. Darwish, J. L. Lentz, M. R. Pinto, P. M. Zeitzoff, T. J. Krutsick, and H. H. Vuong, "An Improved Electron and Hole Mobility Model for General Purpose Device Simulation," IEEE TED-44 (9), p. 1529-1537, 1997.

[55] S. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the Universality of Inversion Layer Mobility in Si MOSFET's: Part I - Effects of Substrate Impurity Concentration", IEEE TED-41(12), p. 2357-2362, 1994.

[56] H. Fang, K. S. Krisch, C. G. Sodini, J. E. Chung, and D. A. Antoniadis, "Ultrathin Furnace Reoxidized Nitrided Oxide Gate Dielectrics For Extreme Submicrometer CMOS Technology", IEDM proceedings, p. 621-624, 1992.

[57] D. A. Antoniadis, I. Djomehri, A. Lochtefeld, "Electron Velocity in sub-50-nm Channel MOSFETs", SISPAD proceedings, p. 156-161, 2001.

[58] A. Lochtefeld and D. A. Antoniadis, "On Experimental Determination of Carrier Velocity in Deeply Scaled NMOS: How Close to the Thermal Limit?", IEEE EDL-22(2), p. 95-97, 2001.

[59] M. V. Fischetti and S. E. Laux, "Long-range Coulomb interactions in small Si devices. Part I: Performance and reliability", Journal of Applied Physics, Vol. 89, No. 2, p. 1205-1231, 2001.

[60] S. A. Mujtaba, R. W. Dutton, and D. L. Scharfetter, "Semi-Empirical Local NMOS Mobility Model for 2-D Device Simulation Incorporating Screened Minority Impurity Scattering", NUPAD V, Hawaii, June 5-6, 1994.

[61] D. M. Caughey and R. E. Thomas, "Carrier Mobilities in Silicon Empirically Related to Dop-

ing and Field", Proc. IEEE, Vol. 55, p. 2192-2193, 1967.

[62] J. A. Cooper and D. F. Nelson, IEEE EDL-2 (7), p. 171-173, 1981.

[63] B. Yu, H. Wang, A. Joshi, Q. Xiang, E. Ibok, M.-R. Lin, "15nm Gate Length Planar CMOS Transistor", IEDM proceedings, 2001.

[64] H. P. Wong, D. J. Frank, and P. M. Solomon, "Device Design Considerations for Double-Gate, Ground-Plane, and Single-Gated Ultra-Thin SOI MOSFET's at the 25 nm Channel Length Generation", IEDM proceedings, p. 407-410, 1998.

[65] Y. Taur, "An Analytical Solution to a Double-Gate MOSFET with Undoped Body", IEEE EDL-21(5), p. 245-247, 2000.

[66] A. Wei, "Device Design and Process Technology for Sub-100 nm SOI MOSFETs", Ph. D. Thesis, 2000.

[67] L. Chang, S. Tang, T. King, J. Bokor, and C. Hu, "Gate Length Scaling and Threshold Voltage Control of Double-Gate MOSFETs", IEDM proceedings, 2000.

[68] S. Takagi, J. Koga, and A. Toriumi, "Subband Structure Engineering for Performance Enhancement of Si MOSFETs", IEDM proceedings, p. 219-222, 1997.

[69] P. Solomon, personal communication, 2001.

[70] 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?", IEDM proceedings, p. 553-556, 1992.

[71] D. J. Frank, Y. Taur, and H. P. Wong, "Generalized Scale Length for Two-Dimensional Effects in MOSFET's", IEEE EDL-19 (10), p. 385-387, 1998.

[72] Y. Taur and T. H. Ning, Fundamentals of Modern VLSI Devices, 1998.

Appendix A

Brief Inverse Modeling Manual

This text describes how to use the Inverse Modeling Package, a numerical software tool developed to reverse engineer a parameterization of the morphology of a device given its electrical characteristics.

In addition to the provided files, the user will need a C++ compiler and a device simulator of choice. The “example” directory contains a walk-through of inverse modeling a real sub-100 nm MOSFET. The “source” directory holds the source code of the optimization loop; the accompanying “Makefile” (it assumes usage of the g++ compiler, and the HERE macro must be changed to the current working directory) is in the current directory. Type “make” at the prompt to build the executable named “im”; then simply running “im” will display a brief help screen.

This section comments on the optimization loop as implemented by each of the following source code files in the “source” directory. The program uses the common Levenburg-Marquardt algorithm. In summary, the sum of squares error between the simulated and experimental data is minimized by finding the derivatives of the error with respect to changes in the parameterization of the topology, and then solving for an improved parameterization. The code files are

“immain.cc”: Main optimization loop

“help.cc”: Contents of the help screen

“optim_read.cc”: Functions that read the input file and data comparison file, and read the experimental and simulated data

“optim_run.cc”: Functions that run executables, determine loop termination, and update search conditions

“optim_fit.cc”: Functions that compute error, derivatives, solution matricies, new parameterizations, and damping

“optim_stat.cc”: Functions that print log, output, and warning files

“optim.h”: Header file of parameter, data, solution, optimization classes

“globals.h”: Header of global variables; modify and re-compile as desired

Some important global variables, their default values, and functionalities are

max_iter	25	terminate on maximum number of iterations
tolerance	1.0e-5	terminate on RMS error below tolerance
frac_improve	1.0e-4	terminate on this fractional improvement in error
frac_deriv	0.01	fractional change in parameters for derivatives
small_deriv	1.0e-5	smallest addition to parameters for derivatives

This section illustrates a practical inverse modeling run, similar to those recently published, on a bulk-Si NMOSFET. A recommended preliminary step is to enter the process flow of the desired device into a process simulator and obtain an initial guess to the topology and two-dimensional (2-D) doping profile.

The first step in MOSFET modeling is to characterize the gate stack (e.g., using Cgg data). After determining the poly doping and physical Tox, describe the device in a standard 2-D device simulator. In this example, pdope = 7e19 cm^-3, Tox = 0.0045 um, the poly was modeled as silicon to account for poly depletion, and a quantum mechanical effects model was turned on using the MEDICI device simulation package.

The next step if a channel doping profile exists, in this case an SSR and well profile, is to use Id(Vgs, Vbs) data of a long channel device to extract it, as in the “example/long” directory. The details of the inverse modeling methodology here are similar to those of the short channel device, which are discussed in the next paragraph. The resulting parameterized 1-D channel doping (using three 1-D gaussian doping representation functions), "chan.dope" was copied and used on the desired device.

The final step is to perform inverse modeling using Id(Vgs, Vbs, Vds) on the desired short channel device. The "example/ssr3_start" directory contains all the files necessary before running the optimizer:

data:
Here labeled as "d<Vds>b<Vbs>", each data file must have two columns: the first being Vgs, the second being Id. Only subthreshold I-V data should be used since it has a strong dependence on doping features. A broad range of Vbs bias sweeps the extent of the depletion region and hence has sensitivity at different depths. A broad range of Vds bias detects short channel effects and is sensitive to lateral doping (e.g., halos).

"inputfile":
This main input file can have arbitrary name but must be composed as follows, where <number of data sets> = 1 since only I-V data is used here; the actual initial guess value of each parameter is the <guess value> times the <factor> and is bounded by <min> and <max>, this is done to keep the solution matricies stable; the <weight> = 1 since the parameter has full weight on the I-V data.
Format:
#topology <topology executable>
#simulate <simulation executable>
#merit <number of data sets> <number of data files> <data compare file name>
#fit <number of parameters>
<guess value> <min> <max> <factor> <weight1 ...>

"Merit":
This arbitrarily named data compare file has one row per data file, with the columns meaning: <set #> = 1 since all data belong to I-V; the second column allows for taking the difference of the data or the logarithm of the data (e.g., log(Id,simulated) - log(Id,experimental) as is used for subthresh-

old I-V); the last two columns give a 1:1 correspondence between experimental and simulated data files.
Format:
<set #> <merit [lin|log]> <raw data file> <sim data file>

"imParameter":
Updated whenever a new parameterization must be simulated, this file has a single column of the actual parameter values. The fractional change in a parameter used for calculating derivatives can be modified as per part 2.

topology executable:
Compile the program "mdg.cc" via the command line "g++ -o mdg mdg.cc". In general, the parameters in "imParameter" are read in and are reformatted and dumped to a file "sdha.dope" that can be read by the device simulator, which is MEDICI in this case. The parameters to optimize for this device include the peaks, centers, characteristic lengths for 2-D gaussians for source/drain and halos. Also, some default parameters have been set within this program such as Asd = 2e20 cm^-3, S/D peak doping, since the I-V data is insensitive at that level.

simulation executable:
The appropriate command to run the device simulation of given parameterization and to convert the output into acceptable data format (for MEDICI, this is "md10000 <run file>", followed by a shell script "genim" to do conversion). These files may also call further helper scripts.

"SSR3mesh":
This file acts as a NMOSFET template structure for MEDICI, in which is defined the device geometry, numerical mesh, material regions, contacts, and doping (which loads "chan.dope" and "sdha.dope")

"ssr3init":
Before starting the simulation runs, it's a good idea to create initial solution files for the varying biases. Create by hand an "imParameter" with the initial guess parameters, run "mdg", and then

run “md10000 ssr3init”.

“ssr3run”:

As the main run file, the structure file is first called and put through loops of varying bias on Vgs, Vds, and Vbs. The I-V outputs are saved in the MEDICI TIF format. Also, device simulation models such as Fermi-Dirac statistics, QM effects, and mobility (not too critical since subthreshold I-V has weak dependence) are specified here.

“genim”:

In order to convert the device simulator output to columnar format, this script picks off each data point and appends them to the correct files (in case the simulator had to solve a bias point between the expected ones); a simpler script is “genim_simple”. “genim” also calls the UNIX awk based script “md2iv_dec” to convert the MEDICI output; if using a Sun workstation try “md2iv_sun”.

Now, to begin the inverse modeling run “nohup im inputfile > /dev/null &” and make sure “im” is in the path; in UNIX nohup keeps the job running even if you exit and & lets it run in the background; funnel the screen output to /dev/null if desired. The contents of the “example/ssr3” directory are of a completed run. Look for the following output files to be generated:

“imIterate.log”:

For each Iteration N, the RMS error and associated updated parameters are logged. The run will terminate based upon three conditions discussed in part 2 above.

“imError”:

For each iteration, this is the error vector (simulated - experimental) where data points from the data files are concatenated per their order in “Merit”. This is useful to check which data are having a hard time converging.

“imStatus.log”:

For each iteration, this outputs the solution matricies for debugging.

“imResult”:

gives the final parameterization!

Appendix B

Measurements for C-V

This appendix provides further explanation of the technique for measurements. The C-V data used in the inverse modeling procedure is $C_{gds} = C_{ov} + C_{if} + C_{of}$ which is the sum of the capacitance of the overlap (L_{ov}), internal fringing, and outer fringing [72] with gate height t_{gate}

$$C_{ov} = \frac{\varepsilon_{ox} W L_{ov}}{t_{ox}} \qquad \text{Equation B.1}$$

$$C_{of} = \frac{2\varepsilon_{ox} W}{\pi} \ln\left(1 + \frac{t_{gate}}{t_{ox}}\right) \qquad \text{Equation B.2}$$

Thus, expect the fringing as a significant contributor to the total capacitance. The width of the MOSFETs measured should be large (preferably > 50 μm) to increase the detectivity since the overlap is very small.

A standard measurement set-up as illustrated in Fig. B.1 is used to apply bias to the four MOSFET terminals for either the I-V or C-V readings. In the C-V configuration, BNC cables (as short as possible to reduce cable parasitics) tie the drain to the source which is connected to the "low" potential on the C-V meter, while the gate is connected to the "high" potential. The "low" provides the ground node through which the small signal current is measured; the "high" imposes a V_{BIAS} voltage and a small signal V_{SS}. The back terminal is connected to a voltage source

V_{BACK} whose ground is connected to the *case* ground of the C-V meter. Also, the BNC sheaths are connected to the metal frame of the probe station.

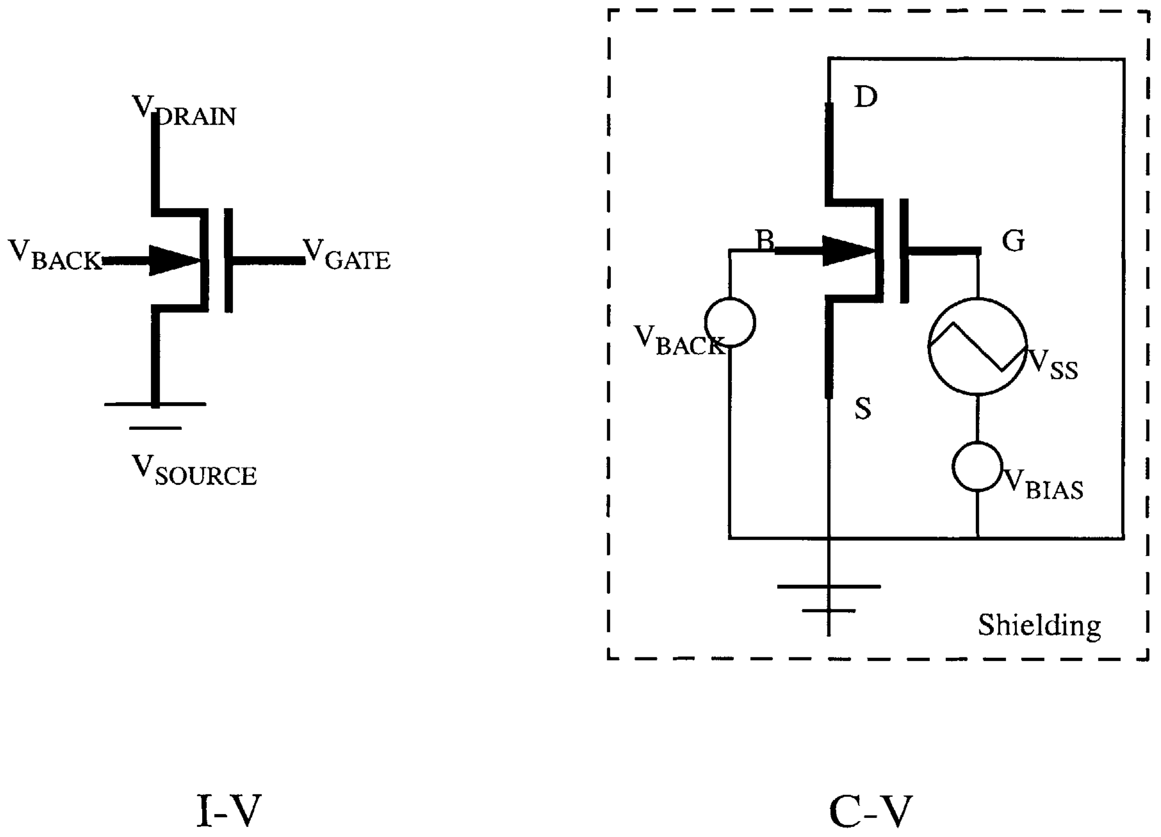

Figure B.1: Measurement set-up for short channel MOSFETs in order to obtain the log(I)-V and C_{gds}-V data necessary for inverse modeling.

Before proceeding with the measurement, the probes must be up and the C-V meter calibrated as both an open circuit and a closed (“low” connected to “high”) circuit to zero out parasitics from the apparatus. Next, it is very important that the wafer be stationary, the standard probe tips have firm contact with the device pads, and the probe station frame is closed gently to fully surround and shield the devices. Presumably, the C-V meter is set in a mode to detect the capacitance in parallel with a parasitic resistance R (which may exist due to gate leakage). The small signal amplitude (which should be less than the bias step size) of V_{SS} can be raised to increase signal strength. Likewise, since the small signal voltage of frequency ω induces varying charges and hence currents at the terminals, and since the impedance goes as $j\omega C$, higher frequencies like 800 kHz are good. Then V_{BACK} is set to the desired bias and the V_{BIAS} is swept from inversion to negative V_{GS} to deplete the S/D.

4935-24