Optimal MOSFET Design for Low Temperature Operation

by

Keith M. Jackson

Bachelor of Science in Electrical Engineering
Princeton University, June 1994

Master of Science in Electrical Engineering and Computer Science
Massachusetts Institute of Technology, September 1996

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

Doctor of Philosophy
In Electrical Engineering and Computer Science

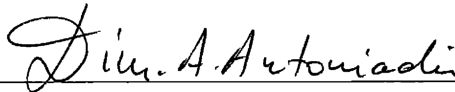
at the

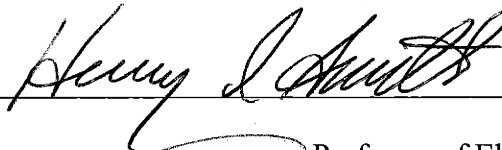
Massachusetts Institute of Technology

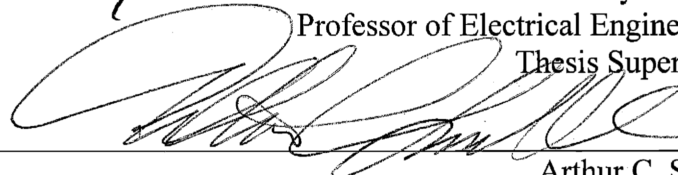
June 2001

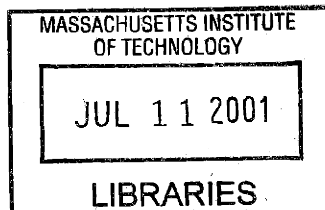
© 2001 Massachusetts Institute of Technology. All rights reserved.

Signature of Author
Department of Electrical Engineering and Computer Science
May 15, 2001

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

Certified by 
Henry I. Smith
Professor of Electrical Engineering
Thesis Supervisor

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


MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUL 11 2001
LIBRARIES
ARCHIVES

Optimal MOSFET Design for Low Temperature Operation

by

Keith M. Jackson

Submitted to the Department of Electrical Engineering and Computer Science
on May 15, 2001 in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

The phenomenal scaling of MOSFET feature size, two orders of magnitude in the past 30 years, has provided the gains in performance and packing density that underlie the GHz microprocessors and 256 MB DRAMs that exist today. Looking forward, the connection between increased performance and smaller devices faces significant challenges.

Lowering the operating temperature can help achieve the desired increases in performance as device size scales. Lowering the temperature reduces the off-state leakage of a MOSFET removing constraints on reducing the threshold voltage. In addition, lower temperatures increase the current drive via increased carrier mobility and saturation velocity. Equally as important, the parasitic resistances of the device and of the interconnect decrease as temperature decreases. The approach of this thesis is to use comparisons of optimal designs across channel lengths and across temperatures to accurately assess the performance increases and increased design flexibility that come with lowering the device operating temperature.

Using analytical equations, the tradeoff between fully scaled performance and maintaining reasonable off-current levels is clearly shown. As an alternative to allowing off-currents to rise, two possible temperature scaling scenarios, that either meet or exceed fully scaled performance, are explored.

Focusing on a nominal channel length of 90 nm (worst-case of 75 nm) operating at 200 K, a detailed analysis of channel doping profile design to achieve the highest on-current at the nominal channel length, while meeting the off-current limit for the worst-case channel length is performed. Using an inverse modeling approach, a 2-D numerical simulator is first calibrated at various temperatures to measured device data down to 80 nm channel lengths. Coupling the simulator with an optimizer, a range of different halo, retrograde, and uniform doping profiles are examined.

Halo doping is found to give the best device performance due to its lower threshold voltage, lower threshold voltage decrease with channel length, and lower body effect. The halo profiles become more abrupt for lower temperature designs. Comparing optimal designs for a 90 nm nominal device across temperature, on-current gains, and thus switching speed gains, of 3.5% for every 10 °C decrease in temperature can be achieved.

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

Thesis Supervisor: Henry I. Smith
Title: Professor of Electrical Engineering

Acknowledgements

Paraphrasing the words of Hillary Clinton (if she had been a PhD student): "It takes a village to raise a graduate student." I am truly thankful for all of the interactions with and help given by people at MIT and beyond. My thanks first go out to my advisor, Prof. Dimitri Antoniadis, for his technical guidance and for supporting this research as it explored various directions. To Prof. Hank Smith, thanks for sharing your enthusiasm and knowledge about lithography. To Prof. Judy Hoyt, thanks for your suggestions on my thesis and papers. I am very grateful to Dr. Bob Dennard of IBM for his interest in my project – your insightful questions have helped push this project along at key points.

This thesis would not have been possible without the use of devices fabricated at IBM. My sincere thanks to Dr. Lisa Su and the IBM SRDC for their generous sharing of devices that are the core data this thesis rests on.

Although fabrication appears as only a small appendix, the majority of my time here was spent in the fab, and my sincere thanks goes out to the staff of the ICL and TRL, past and present, for their help and hard work: Dan Adams, Paudley Zamora, Paul McGrath, Paul Tierney, Pat Burkhart, Joe DiMaria, Kurt Broderick, Wayne Price, Ron Stoute, Joe Walsh, Jim Bishop, and Barry Farnsworth, to name a few. A special thanks goes out to Bernard Alamariu for his help with all of the various diffusion process changes I worked on with him. Thanks also the MTL computer support team (Myron Freeman, Ralph Nevins, Tom Wingard, Mike Hobbs and Tom Lohman) for all of your help. Thanks also to Debb Hodges-Pabon for her help and wit.

Outside of MIT I am very grateful for the support of Wade Krull and Axcellis (Eaton SEO) that allowed me to use their development implanters to have very low energy Boron, Decaborane, and Arsenic implanted. Wade's trust in my ability to glue 4" wafers to 8" wafers made the project possible.

In addition, I am appreciative of the fabrication advice and help of Dr. Doug Buchanan of IBM (gate oxides), Dr. Shariar Ahmed of Intel (TEMs), and Aditya Agarwal of Axcellis (Shallow S/D implants). I thank DEC for donating their low-temperature measurement station that allowed the measurements in this thesis.

This work was initially funded by a grant from DARPA and was continued on generous support from IBM and Intel – thank you for making this project possible.

At MIT I was fortunate to be included in two different research groups. I thank the members of Dimitri's group for the MOSFET discussions and marathon fab sessions. Thanks to: Dr. Melanie Sherony, Dr. Andy Wei, Dr. Tony Lochtefeld, Dr. Isabel Yang, Dr. Mark Armstrong, Dr. Jarvis Jacobs, Dr. Zachary Lee, Jim Fiorenza, Hasan Nayfeh, Isaac Lauer, Corina Tasana and Ali Kakiforuz. A big thanks goes to Ihsan Djomehri for all of his help that made the inverse modeling and device optimization possible.

I thank Hank's group for including one of the "ICL crowd" in their midst and for exposing me to the fine art of lithography and microscopy. From the pizza parties to SEMing to X-ray masks to lunches at Redbones – thanks. In the NSL, thanks to Jimmy Carter, Mark Mondol, Ed Murphy, and James Daley for all of their lab support.

To my parents, thank you so much for your love, support and perspective over the past 7 years. Last, and most importantly, thank you to Stacy from the bottom of my heart for your devotion and love through all of the trials and tribulations and for the wonderful times we have had together in New England. I look forward to the new adventures we will face together as we both graduate from MIT.

Contents

Chapter 1 Introduction **13**
1.1 Motivation 13
1.2 Thesis Focus and Organization 15

Chapter 2 Impact of Lower Temperatures **16**
2.1 MOSFET Turn-on Characteristics 16
2.2 Short Channel Effects 21
2.3 Transport 25
2.3.1 Mobility 26
2.3.2 Saturation Velocity 29
2.4 Incomplete Ionization of Dopants 32
2.5 Parasitic Source and Drain Resistance 35
2.6 Interconnect 37

Chapter 3 Scaling Length, Scaling Temperature **40**
3.1 Previous Results 40
3.2 Scaling Theory 41
3.3 Scaling Scenarios 47
3.3.1 Room Temperature Scaling Scenarios 49
3.3.2 Low Temperature Scaling Scenarios 52
3.4 Conclusion 55

Chapter 4 Device Design Comparison at 200 K **56**
4.1 200 K Fundamental Performance Increase 57
4.2 Performance Comparison 58
4.3 Role of the depletion depth 63

Chapter 5 Optimal Device Design **70**
5.1 Inverse Modeling 70
5.2 Device Optimization Setup 76
5.3 Device Optimization Results 78
5.4 Optimal Designs Across Temperature 85
5.4.1 Device Performance 85
5.4.2 Device Designs 88
5.4.3 Room Temperature Comparison 89
5.5 Conclusion 90

Chapter 6 Conclusion **91**
6.1 Contributions 92
6.2 Suggestions for Future Work 93

References **94**

Appendix A Mobility Extraction **100**

Appendix B Fabrication Technology for 50nm MOSFETs **109**

List of Figures

Figure 1-1: Historical and projected scaling of device feature size, from [] 13

Figure 2-1: I_d-V_g char. of a L_{gate}=20 um NMOS device versus temperature. A: Logarithmic scale B: Linear scale; T=[350, 300, 250, 200, 150, 100 K]; V_{ds}= 50 mV.. 17

Figure 2-2: Subthreshold slope (A) and extrapolated threshold voltage (B) versus temperature for an L_{gate}=20 μm NMOS device; V_{ds} = 50 mV. 18

Figure 2-3: Shape of the Fermi function F(E) for different temperatures, from []. 19

Figure 2-4: Band-gap and Fermi-level position versus temperature, from [7]. The numbers label the doping (thus free carrier) concentration which corresponds to the Fermi-level position 20

Figure 2-5: Shift in delta V_{th} (V_{ds}=50mV, Extrapolated) vs. channel length at 297K and 200K.. 22

Figure 2-6: I_{ds}-V_{gs} of L_{eff} = 80 nm device. The parallel shift of the curves to the left with increasing drain voltage in subthreshold (low V_{gs}) is caused by drain induced barrier lowering (DIBL). V_{ds} = [0.01, 0.05, 0.1, 0.5, 1.0, 1.5, 1.8 V], T = 300K................ 23

Figure 2-7: Shift in constant current threshold voltage from V_{ds} = 0.1 V to V_{ds} = [1.0,1.5,1.8V] (a measure of DIBL) for a L_{eff} =80 nm device. As temperatures decrease, this shift decreases, showing the reduction of short channel effects......... 24

Figure 2-8: Simulated results of the surface potential of a 0.25 μm device with uniform $1x10^{17}$ cm^{-3} channel doping. V_{gs} is set to give I_d = 100 nA in all cases, from [10].. 25

Figure 2-9: Diagram of mobility vs. vertical effective field (E_{eff}) in a MOSFET inversion layer, showing the roles of the different scattering mechanisms, from []................ 26

Figure 2-10: Electron Mobility measured on a L_{gate}=20 μm NMOS device at T=[350,300,250,200,150,100] K. 28

Figure 2-11: Electron mobility versus temperature. The mobility at constant V_{gs}-V_{th} = 1.5 V increases less than the peak mobility due to the impact of surface roughness scattering. 29

Figure 2-12: Plot of velocity versus lateral electric field as described by equation (1.4) (β = 2, μ_s = 200). 30

Figure 2-13: Carrier velocity ($g_{m,max}$/WC_{ox}) versus Drain Induced Barrier Lowering. The 200 K values are significantly increased from the 300K values. Channel length is the implicit variable that is changing in this plot. Higher DIBL corresponds to a shorter L_{eff} 31

Figure 2-14: Comparison of percentage gain in measured velocity ($g_{m,max}$/WC_{ox}) versus temperature with % gain in mobility (at V_{gs}-V_{th} = 1.5 V) and the theoretical gain in saturation velocity, from equation (2.6).. 32

Figure 2-15: Percentage ionization of Boron and Indium versus temperature. For the equation, p+ are free hole, N_d+ are ionized donors, n- are free electrons, and N_A are ionized acceptors. N_c,N_v are density of states, E_f is Fermi level, band edges are E_c,

E_v, donor /acceptor ionization energy levels E_d, E_a, degeneracy of acceptors and donors are GCB,GVB. 33
Figure 2-16: C-V and accompanying band diagrams showing freeze-out at flatband (B) and ionization of the dopants in the depletion region (C), from [25]. 34
Figure 2-17: 35
Figure 2-18: Percentage change in source/drain series resistance (R_{sd}) versus temperature from published reports in the literature. [,,,] 35
Figure 2-19: Specific contact resistance as a function of temperature normalized to the T = 305K for W contact to Si:P. Surface doping concentration is labeled in cm^{-3}. Dashed curve is theoretical prediction for 2.3×10^{20} cm^{-3}, from []. 36
Figure 2-20: Thin film and bulk resistivity of Al versus temperature, from [34] 38
Figure 3-1: Schematic I_{ds}-V_{gs} characteristic showing the changes that occur as the device is scaled from L to L' at a constant temperature T. 44
Figure 3-2: Schematic I_d-V_g showing the addition of temperature scaling to the L' scaled device. The steeper subthreshold slope allows the off-current limit to be met........ 46
Figure 3-3: Input and output inverter waveforms for input high to low transistors, from []. 48
Figure 3-4: Device Performance (Nominal L) versus design node for the scaled-V_{th} and Constant-I_{off} Scenarios. 50
Figure 3-5: Off-current (Worst Case, 0.8*L device) versus design node for the Scaled-V_{th} and Constant-I_{off} scenarios. 51
Figure 3-6: Device design parameters at each node for the Scaled-V_{th} and Constant-I_{off} scenarios. A: Oxide thickness B: Uniform channel doping level 51
Figure 3-7: Device Performance (at nominal L) versus design node for the LT scaled-V_{th} and LT scaled-performance scenarios (I_{off}=10^{-9} A/μm for all points). Note that the LT scaled-performance case matches the performance of the RT scaled-V_{th} case (Figure 3-4). 53
Figure 3-8: SS of optimized designs. 54
Figure 3-9: Substrate Bias for the scaled-V_{th} and scaled-Performance scenarios as compared to the measured substrate bias needed for $I_s + I_d = 10^{-9}$ A/μm 55
Figure 4-1: Threshold voltage versus channel length for the High-V_{th} and Low-V_{th} device designs at 300 K. 57
Figure 4-2: I_{on}-I_{off} plot for the Low-V_{th} and High-V_{th} designs at 300K and 200K. (V_{ds} = 1.8 V, V_{bs} = 0 V). 58
Figure 4-3: A: Threshold voltage at V_{ds}=50 mV versus channel length for the Low –V_{th} design. Cooling the device raises the V_{th}, while applying a forward substrate bias lowers the V_{th} . B: Log I_{off} versus I_{on} at 200 K for the Low V_{th} design. V_{bs} = [0,0.25, 0.5, 0.7 V] 59
Figure 4-4: Substrate bias used at each L for each device design to reach I_{off}=10^{-9} A/μm. Note that all the values fall below the limit from junction leakage (I_s+I_d=10^{-9} A/μm) 60
Figure 4-5: On-current vs. L_{eff} for the two designs with I_{off} =1×10^{-9} A/μm at each point (T=200K). 61
Figure 4-6: V_{th}, SS, DIBL vs. L_{eff} at I_{off}=10^{-9} A/μm. (T=200 K). 62
Figure 4-7: Delta V_{th} from V_{bs} = 0 to 0.5 V for High-V_{th} design and Low-V_{th} design devices at 200 K. 63

Figure 4-8: Circuit diagram representation of the gate capacitance in subthreshold. ΔV_g is the incremental gate voltage, C_{ox} is the gate capacitance, C_d is the depletion region capacitance and $\Delta\psi_s$ is the incremental surface potential. 64
Figure 4-9: Simulated energy bands vs. depth in the silicon without substrate bias (black) and with +0.4 V substrate bias (gray). (V_{gs} = 1 V, $V_d = V_s$ = 0 V) The dotted line labeled R is the ratio of the hole concentration to channel doping. The depletion depth, which shrinks with V_{bs}, is at the point where R = 0.5. 65
Figure 4-10: Simulated energy bands vs. depth in the silicon with 1x10^{17} cm^{-3} doping (black) and 5x10^{17} cm^{-3} doping (gray). (V_{gs} = 1.0 V, $V_d = V_s = V_b$ = 0 V) The depletion depth is at R=0.5 (dotted line). Increasing the doping shrinks the depletion depth. 67
Figure 4-11: MOSFET schematic showing the geometry of the depletion depth (W_d) and the channel length. L_{eff} needs to be greater than x_d for SCE to be controlled, from [66]. 68
Figure 4-12: Plot of A: Subthreshold Slope (SS), B: Drain Induced Barrier Lowering (DIBL), and C: Threshold Voltage (V_{th}) vs. depletion depth using the analytical equations from Taur [] (L_{eff} = 80 nm, t_{ox}=41 Å). 68
Figure 5-1: Comparison of measured to simulated capacitance voltage characteristic of a L_{gate}=19.85 μm NMOS device. The discrepancies for $0<V_{gs}<-1$ are an artifact of the Van Dort model implementation in MEDICI []. 72
Figure 5-2: Long channel subthreshold data (+ and line) and inverse-modeled simulation (circles) showing match (V_{ds}= 50 mV). 73
Figure 5-3: Extracted vertical dopant profile of long channel device showing the different depletion depths of different V_{bs}. 73
Figure 5-4: Match between low V_{ds} long channel device data and simulation with the calibrated mobility model. 74
Figure 5-5: I_d-V_d at V_{gs}=1.8V comparing simulation (circles and line) and data (+ sign), showing the calibration of the velocity saturation model. 75
Figure 5-6: I_d-V_g of inverse modeled 80 nm device comparing the inverse modeled simulation results to the data at 300 K. V_{ds} = [0.01, 0.05, 0.1, 0.5, 1.0, 1.5, 1.8 V] 75
Figure 5-7: I_d-V_g of inverse modeled 80 nm device comparing the inverse modeled simulation results to the data at 200 K. V_{ds} = [0.01, 0.05, 0.1, 0.5, 1.0, 1.5, 1.8 V] 76
Figure 5-8: On-current of the nominal device versus the worst-case device's DIBL. All device designs have an $I_{off} = 10^{-9}$ A/μm for the worst-case device. T = 200 K. 78
Figure 5-9: On-current of the nominal device versus the threshold voltage (V_{ds} = 1.8V) of the nominal device. The labeled numbers are the worst case device DIBL values (mV) for the end points of each line. 79
Figure 5-10: Substrate bias used for each of the designs, versus the worst-case device DIBL. All values are less than the measured bias for 10^{-9} A/μm junction leakage current. 80
Figure 5-11: Measured electron mobility versus gate overdrive (V_{gs}-V_{th}) at 200 K. Applying a forward bias V_{bs} decreases the E_{eff} at a given gate overdrive, thus increasing the mobility. 80
Figure 5-12: On-current of the worst-case device versus V_{th} (V_{ds} = 1.8 V) of the worst-case device. 81

Figure 5-13: A: Shift in the nominal device threshold voltage (V_{ds} = 1.8 V) with ± 0.1 V V_{bs} from each design set-point versus the DIBL of the worst-case device. B: Shift in the worst-case device threshold voltage (V_{ds} = 1.8 V) with ± 0.1 V V_{bs} from each design set-point versus the DIBL of the worst-case device. 82
Figure 5-14: Threshold voltage (V_{ds} = 1.8 V) of both the nominal and worst-case devices versus the worst-case device DIBL. The Halo designs show the least amount of V_{th} roll-off from the nominal to the worst-case channel length. 83
Figure 5-15: Subthreshold slope (V_{ds} = 1.8 V) of the worst-case device versus the DIBL of the worst-case device. The steeper SS of the Halo designs yield the lowest V_{th}s, given the fixed $I_{off} = 10^{-9}$ A/μm for all designs. 84
Figure 5-16: On-current of the nominal device versus the threshold voltage (V_{ds} = 1.8V) of the nominal device. The labeled numbers are the worst case device DIBL values (mV) for the end points of each line. 85
Figure 5-17: A: On-current of the optimized devices at their operating temperatures. B: Percentage increase in on-current from the 300 K optimized design to lower temperature optimized designs. 86
Figure 5-18: Threshold voltage (V_{ds} = 1.8 V) of the nominal and worst-case devices for each of the optimized designs at their operating temperature. 87
Figure 5-19: Percentage change in optimized-nominal-device on-current from the 300 K design to lower temperature designs, as compared to the % gain in measured velocity of carriers in the 80 nm Low-V_{th} design device. 87
Figure 5-20: Worst-case device net-doping profiles of the different optimal designs at different temperatures. (Upper left: 300K, Upper right: 100K, Lower left: 200K, Lower right:1-D profile of absolute value of net doping at the oxide-silicon interface) For the 2-D profiles, p-type doping has solid lines, n-type has dashed lines. 89

List of Tables

Table 3-1: The impact of the changes in device characteristics described in Chapter 2 on the non-scaling scenarios discussed above. 46
Table 3-2: Setup of room temperature scaling scenarios 49
Table 3-3: Setup of low temperature scaling scenarios. 52
Table 5-1: Setup for Optimizations 77
Table 5-2: Comparison of room temperature to operating temperature device parameters for the worst case ($0.8L_{nom}$) device for the different designs. V_{th} and DIBL are at V_{ds}=1.8 V. 90
Table 6-1: Summary of performance improvements at L_{eff} =80 nm N-MOSFET with a 10 °C decrease in temperature. 91

Chapter 1
Introduction

1.1 Motivation

Since 1959 when Jack Kilby of Texas Instruments and Bob Noyce of Fairchild Semiconductor invented the integrated circuit [1], a hallmark of the semiconductor industry has been the continuous growth in chip performance and in the number of devices per chip. Only six years later in 1965, Gordon Moore, then of Fairchild, published his famous prediction that the number of devices on a chip would double every 12-18 months [2]. This prediction, known as "Moore's Law", has become the yardstick against which are measured achievements in performance and packing density that come from shrinking the size of the MOSFET. This phenomenal scaling of device size – surpassing two orders of magnitude in the last 30 years (Figure 1-1) – has provided the gains in performance and packing density that underlie the GHz microprocessors and 256 MB DRAMs that exist today.

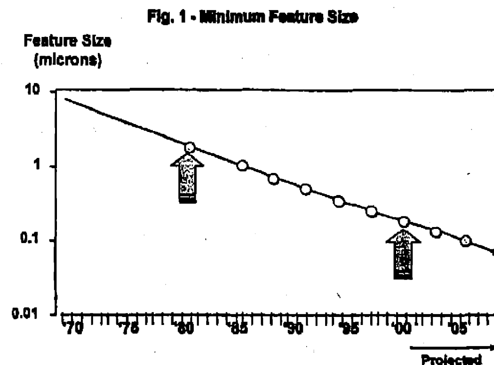


Figure 1-1: Historical and projected scaling of device feature size, from [3]

The connection between increased performance and smaller devices faces challenges at both the device and circuit level. Although shorter MOSFETs can provide higher current densities, unless the parasitic resistances and capacitances associated with the devices and with the metal lines that interconnect the devices scale similarly, the device performance gains will be hidden. This challenge has been continuously met though design innovations and material changes, but innovation will need to continue in future device designs.

At the device level, the reduction of a device's turn-on voltage (threshold voltage) is becoming constrained by leakage limits, i.e. how much current the device conducts when it is in its off-state. Not allowing the threshold voltage to be reduced as the devices shrink reduces the performance gains from shrinking the device dimensions. In addition, unless higher dielectric constant materials are introduced, scaling the gate oxide past the 8 Å thickness found in research devices today [4] poses challenges in controlling tunneling currents and maintaining reliability and manufacturability. Finally, the power density on chips is continuously rising and could reach unmanageable levels within a few generations.

In the past, there have been enough variables in the device design and the interconnect design to surmount such challenges so that the expected performance gain of 30% per device generation could be met. Looking into the future, however, the challenges mentioned above as well as the challenges of introducing new materials beg the question of what other design variables could be used to help achieve the desired performance gain.

Lowering the operating temperature in combination with using a forward substrate bias are two new design variables that can help achieve the desired increase in performance. Lowering the temperature reduces the off-state leakage of a MOSFET removing constraints on reducing the threshold voltage. In addition, lower temperatures increase the current drive of a MOSFET via increased mobility of the electrons and holes in the device. Equally importantly, the parasitic resistances of the device and the interconnect decrease as temperature decreases. The use of a forward substrate bias provides another design variable that helps achieve the desired lower threshold voltages

in smaller devices. The focus of this thesis is to look at the design space for achieving higher performance devices at lower operating temperatures.

1.2 Thesis Focus and Organization

In the process of examining high performance device design at low temperatures, this work will focus on three key questions:

What characterizes an optimal device design at lower temperatures?

What combination of channel doping and substrate bias gives an optimal design?

What are the performance gains achieved by an optimal device design?

Chapter 2 takes a broad look at the impact of lowering temperature on MOSFET operation. Changes in how the device turns on (threshold voltage and subthreshold slope) and in the current conduction (mobility and saturation velocity) will be discussed. In addition the improvements in parasitic device resistance and interconnect resistance will be touched upon.

Chapter 3 places the concept of lowering temperature within the scaling framework that guides device design today. How lower operating temperatures can alleviate off-current issues and help meet other scaling challenges will be discussed. The impact of integrating temperature into a scaling scenario will be evaluated with detailed analytical modeling.

Chapter 4 compares the measured performance of two different threshold voltage devices at 200 K. The correlation between the differences in performance and device characteristics will be analyzed. The role of channel depletion depth in setting the device characteristics will be explored.

Chapter 5 looks at the critical issue of how to design an optimal device for low temperature operation. Using calibrated 2-D numerical simulations, a framework for thinking about device design will be explored. The 2-D simulations will allow actual doping profiles and substrate bias combinations to be evaluated and optimized for a range of temperatures. The optimized designs will allow an accurate assessment of device performance gains as temperature decreases.

Chapter 6 concludes with a summary of this work's major contributions and with suggestions for future work.

Chapter 2

Impact of Lower Temperatures

Lowering the operating temperature of a MOSFET causes shifts in its device characteristics, but does not fundamentally change its behavior. Of particular interest is that the device turns on more abruptly in subthreshold. At the same time, the threshold voltage shifts higher. Carrier mobility and saturation velocities increase, resulting in higher currents. Parasitic resistances from interconnect and series resistance decrease. These changes provide both improved device performance, as well as design challenges.

Throughout this chapter, the measured data are from NMOS devices fabricated at IBM as part of the development of a 0.18 μm technology [5]. The devices have a 41 Å electrical t_{ox} at 1.8V (V_{dd}) in inversion and use a combination of retrograde channel doping and halo doping to control short channel effects.

2.1 MOSFET Turn-on Characteristics

At low gate-to-source voltages (V_{gs}) a MOSFET turns on with a characteristic exponential dependence of current on gate voltage. In this region of gate voltages below the threshold voltage, the subthreshold region, the exponential characteristic of the current is described by its slope. The subthreshold slope (SS) is the change in gate voltage needed to change the current by one decade. A smaller subthreshold slope gives a steeper turn-on of the device. This slope is visible in Figure 2-1A which shows the

subthreshold characteristics (V_{ds}=50 mV) of a L_{gate}=20 μm NMOS device versus temperature.

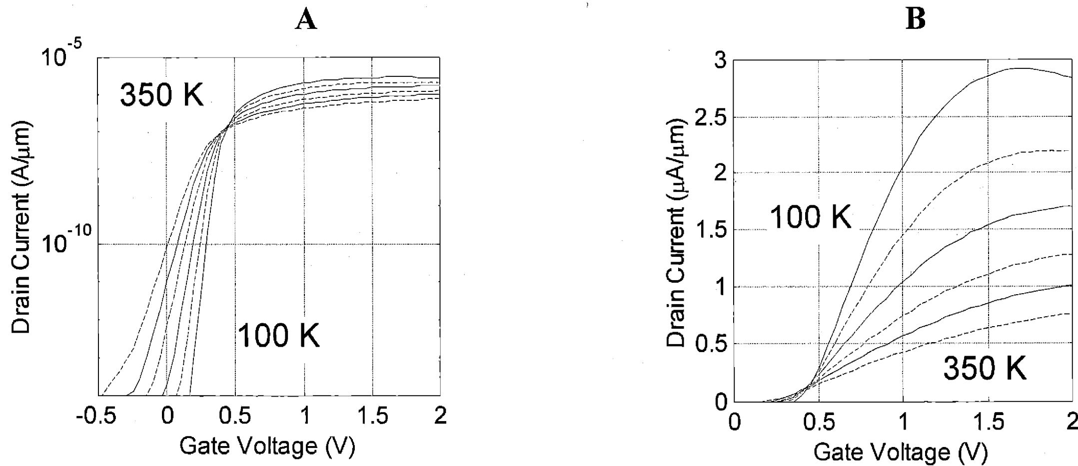


Figure 2-1: I_d-V_g char. of a L_{gate}=20 um NMOS device versus temperature. A: Logarithmic scale B: Linear scale; T=[350, 300, 250, 200, 150, 100 K]; V_{ds}= 50 mV.

As temperature is decreased the subthreshold slope becomes steeper, changing from 90 mV/decade at 350 K to 31 mV/decade at 100 K (Figure 2-2A). This decrease occurs almost directly proportionally to temperature, as is theoretically expected [6].

For very low temperatures, the subthreshold slope improvement is limited by interface states [6]. At lower temperatures, the electron quasi-Fermi level is much closer to the conduction band for a given inversion layer carrier concentration. In addition, the interface state density increases near the conduction and valence bands. Electrons have to fill these states before more free carriers can be added to the inversion layer. Thus with a high density of interface states around the Fermi-level position, a larger SS results than would be expected from the initial temperature dependence. This effect is starting to appear below 200 K for the data in Figure 2-2A.

Using a first order equation for the subthreshold slope [7], the impact of both decreasing the temperature (SS decreases) as well as increasing interface state density (D_{it}) (SS increases) can be seen:

$$SS = 2.3\frac{kT}{q}\left(1+\frac{C_d}{C_{ox}}+\frac{qD_{it}}{C_{ox}}\right) \text{ (mV/decade)} \qquad (2.1)$$

Where kT/q is the thermal potential, C_d and C_{ox} are respectively the oxide and depletion region capacitances per unit area, and q is the electron charge.

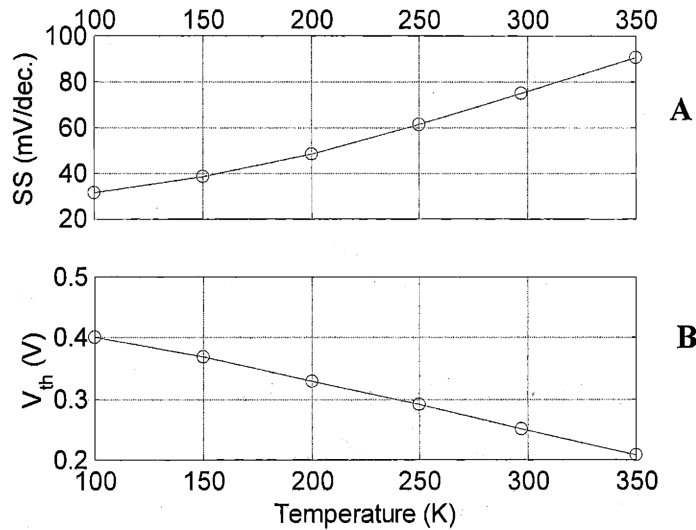


Figure 2-2: Subthreshold slope (A) and extrapolated threshold voltage (B) versus temperature for an L_{gate}=20 μm NMOS device; V_{ds} = 50 mV.

The threshold voltage, or voltage at which the device visibly begins to conduct current on a linear scale, increases as temperature decreases (Figure 2-2B). This shift is visible in Figure 2-1B where the point at which the I-V curve rises above the x-axis shifts to the right as temperature decreases. The rate of increase in threshold voltage versus temperature is 0.78 mV/K for the devices in Figure 2-2B. In general, this rate of change depends on the device design and lies in the 0.5-1.2 mV/K range.

The increase in threshold voltage as temperature is lowered (Figure 2-2B) can be tied directly to the narrowing of the Fermi probability function and the slightly increasing bandgap of Silicon. The threshold voltage is traditionally defined as the point where the

inversion charge density (cm^{-3}) at the silicon surface equals the bulk dopant density (cm^{-3}). The free-carrier density, for either the inversion layer or bulk free-charge, is computed from the overlap integral of the Fermi function and the density of states in the conduction (or valence) band, given a particular position of the Fermi level. The Fermi level is defined as the point where the Fermi probability function equals 0.5, as shown in Figure 2-3.

As temperature drops, the Fermi function becomes steeper (Figure 2-3), while the density of states changes very little. To achieve the same free-carrier density and thus the same value of the overlap integral, the Fermi function, and thus the Fermi level, must be closer to the conduction (or valence) band. On top of this, the band gap has become slightly larger, necessitating the Fermi level to move a little higher to achieve the required distance from the conduction (valence) band. Both of these behaviors are visible in Figure 2-4 where the position of the Fermi-level for a given free-carrier density is plotted in relation to the conduction and valence bands.

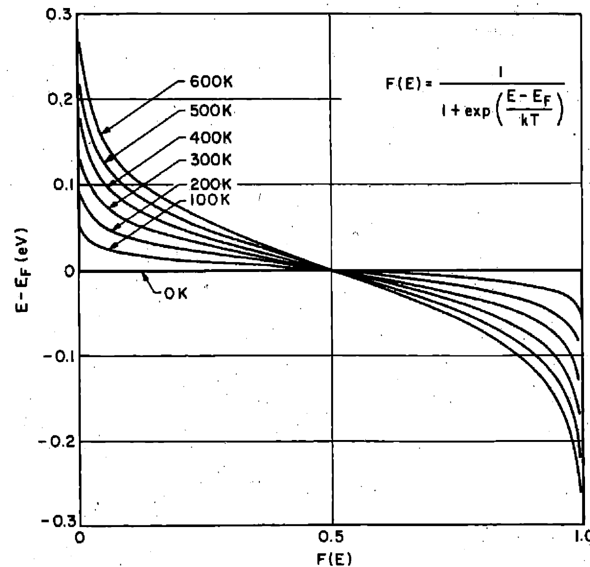


Figure 2-3: Shape of the Fermi function F(E) for different temperatures, from [8].

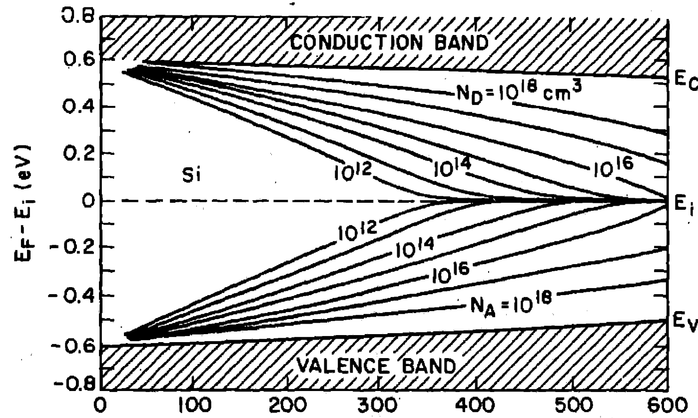


Figure 2-4: Band-gap and Fermi-level position versus temperature, from [7]. The numbers label the doping (thus free carrier) concentration which corresponds to the Fermi-level position

The threshold voltage definition of the inversion charge density (cm^{-3}) being equal to the bulk doping (cm^{-3}) is equivalent to saying that the band bending at the silicon surface, as referenced to the bulk of the MOSFET, is $2\phi_b$.

$$\phi_b = \frac{kT}{q}\log\left(\frac{N_b}{n_i}\right) \tag{2.2}$$

where kT/q is the thermal potential of the system, N_b is the bulk doping of the MOSFET body, and n_i is the intrinsic carrier concentration. ϕ_b is the distance of the Fermi level from the intrinsic level (mid-gap) and is set by the channel doping (N_b).

Since, at lower temperatures, the Fermi function has to be closer to the edges of the expanded bandgap to achieve the same carrier density, the total band bending to shift from the bulk carrier concentration to an inversion layer of the same concentration increases. This increase in $2\phi_b$ is visible in Figure 2-4 by following the difference (in energy) between the N_D=1x10^{17} cm^{-3} Fermi-level position near the valence band to its corresponding one near the conduction band versus temperature.

The bending of bands is directly related, by Poisson's equation, to the exposed charge in the body of the MOSFET. A larger band bending corresponds to a greater exposed (depleted) bulk charge density (Q'_{bulk}, C/cm^2). For a given V_{gs}, the charge on

the gate (Q'_{gate}, C/cm^2) is fixed and must be equal to the sum of the inversion (Q'_{inv}, C/cm^2) and bulk charges (Q'_{bulk}). Since Q'_{gate} is fixed, an increase in Q'_{bulk} results in a decrease in Q'_{inv}, thus necessitating a higher V_{gs} to reach the threshold point.

Examining the standard equation for long channel threshold voltage for a NMOS device with an n^+ doped gate (2.3), the major impact of changing temperatures occurs in the last term which represents the bulk charge.

$$
\begin{aligned}
V_{th} &= V_g\big|_{inv.} = \phi_s\big|_{inv.} + V_{fb} - \frac{Q'_{bulk}}{C_{ox}} \\
V_{th} &= +(2\phi_b) - \left(\frac{E_g}{2} + \phi_b\right) - \frac{Q'_{bulk}}{C_{ox}} \\
V_{th} &= \left(-\frac{E_g}{2} + \phi_b\right) + \frac{\sqrt{2q\varepsilon_s N_b (2\phi_b - V_{bs})}}{C_{ox}}
\end{aligned}
\tag{2.3}
$$

Where ϕ_s is the surface potential, V_{fb} is the gate voltage at flat-band, C_{ox} is the gate capacitance per unit area, ε_s is the dielectric constant, E_g is the bandgap, and V_{bs} is the bulk-to-source bias.

As temperature is lowered, ϕ_b approaches $E_g/2$, so the first term becomes less negative The increase in Q'_{bulk}, due to the increases in ϕ_b, is much more significant. As will be explored in Chapter 3, this increase in Q'_{bulk} is tied to an increase in the depletion width as temperature is lowered.

2.2 Short Channel Effects

Short channel effects refer to the reduction of the threshold voltage as channel length decreases within a particular device design. Short channel effects (SCE) also refer to the dependence of the threshold voltage on the drain voltage for short devices. Both experiments and simulations have shown that the short channel effects of a device improve slightly as temperature is lowered [9,10]. The drain bias dependence of the threshold voltage (DIBL) decreases, while the channel length dependence of short channel effects at low drain bias improves slightly.

The threshold voltage roll-off with decreasing channel length does not change much as a device is cooled. The NMOS devices in Figure 2-5 have similar changes in threshold voltage versus channel length at 300 K and 200 K. These devices exhibit an increase of V_{th} as length decreases, often termed a reverse short channel effect, which arises from the laterally non-uniform doping in the channel due to halos. The slight reduction in reverse short channel effect at lower temperatures has been previously observed [11,12]. Overall, the devices exhibit only slight changes in their V_{th} vs. L characteristics.

Scott [10] explains this similarity versus temperature by noting that the potentials (and thus electric fields) in the bulk of the depletion region of the device remain relatively constant versus temperature. Thus the reduction in V_{th} at short channel lengths due to depletion of channel charge by the source and drain is similar at room temperature and low temperatures.

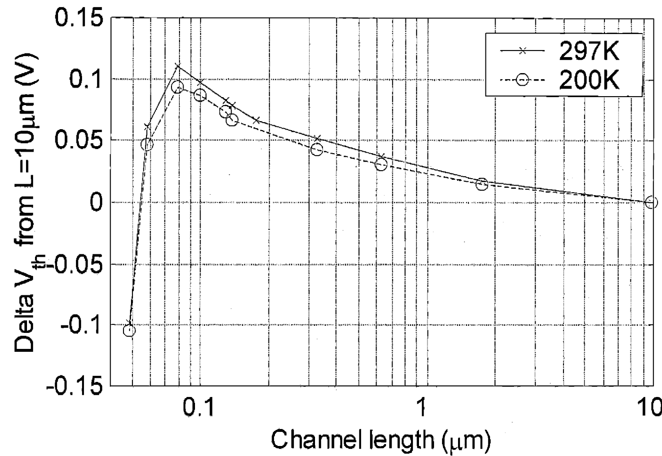


Figure 2-5: Shift in delta V_{th} (V_{ds}=50mV, Extrapolated) vs. channel length at 297K and 200K

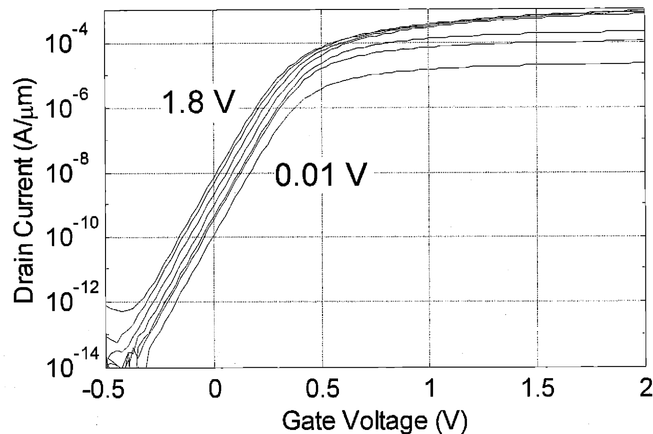


Figure 2-6: I_{ds}-V_{gs} of L_{eff} = 80 nm device. The parallel shift of the curves to the left with increasing drain voltage in subthreshold (low V_{gs}) is caused by drain induced barrier lowering (DIBL). V_{ds} = [0.01, 0.05, 0.1, 0.5, 1.0, 1.5, 1.8 V], T = 300K.

Drain induced barrier lowering (DIBL) is visible in subthreshold as a reduction in the threshold voltage as the drain bias is increased (Figure 2-6). DIBL is generally measured as the shift in the gate voltage to reach a particular current level, 10^{-10} A/μm in this case. The increase in subthreshold current observed with DIBL is the result of a small reduction in the source-to-channel barrier, which is due to the large lateral electric fields caused by the drain-to-source voltage (V_{ds}). Plotting this shift in V_{th} at a given V_{ds} for a 90 nm channel length NMOS device as a function of temperature shows a significant decrease of 30-40% at high V_{ds} as temperature is decreased (Figure 2-7).

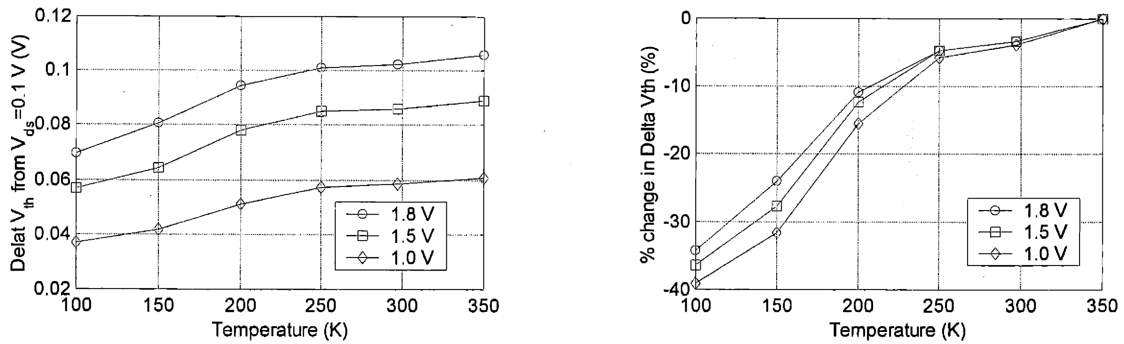


Figure 2-7: Shift in constant current threshold voltage from V_{ds} = 0.1 V to V_{ds} = [1.0,1.5,1.8V] (a measure of DIBL) for a L_{eff} =80 nm device. As temperatures decrease, this shift decreases, showing the reduction of short channel effects.

Both Woo and Scott [9,10] have observed lower DIBL as devices are cooled. Woo ties this reduction in DIBL to the reduction of the channel barrier for a given subthreshold current level at lower temperatures [9]. Lowering this barrier decreases the potential difference between the drain and the top of this barrier ($\psi_{db}=|V_{ds}|+\phi_{barrier}$). The impact of V_{ds} on the channel barrier is proportional to the lateral electric field it causes. To first order, this field is just ψ_{db} divided by the distance from the drain to the barrier. Thus, reducing ψ_{db} by reducing $\phi_{barrier}$ reduces the impact that the drain has on the $\phi_{barrier}$ [9]. Although this may be a bit of a simplistic view given the 2-D nature of the fields in the channel, certainly the change in surface potential with drain bias reaches smaller distances towards the source at lower temperatures (Figure 2-8). In addition, keeping the same subthreshold current as temperature drops requires a larger V_{gs} and thus increasing the vertical field and decreasing the relative magnitude of any lateral field.

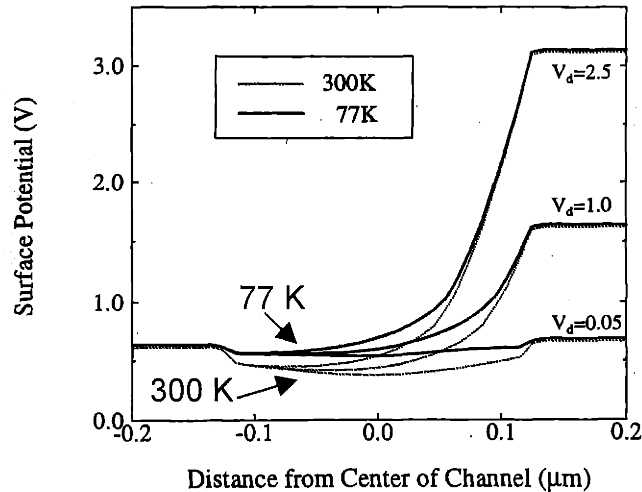


Figure 2-8: Simulated results of the surface potential of a 0.25 μm device with uniform $1x10^{17}$ cm^{-3} channel doping. V_{gs} is set to give I_d = 100 nA in all cases, from [10].

Overall, temperature causes a small reduction in short channel effects. The fundamental causes of SCE do not change, but changes in the potentials inside the device reduce their magnitude. This behavior suggests that designing a lower V_{th} device with controlled short channel effects may be easier at lower temperatures than at room temperature.

2.3 Transport

In a general sense, the current flow in a device is equal to the product of charge times velocity. The velocity of the carriers is related to the applied electric field via mobility for low electric fields and low velocities. This mobility captures the steady-state balance of forward momentum gained from the electric field and momentum lost via scattering mechanisms. At very high electric fields, carriers lose their forward momentum by optical-phonon emission at the same rate they gain it from the electric field, resulting in velocity saturation [13]. Lowering the operating temperature of a

device increases significantly the carrier mobility and slightly increases the saturation velocity.

2.3.1 Mobility

Carrier mobility (μ) in a MOSFET inversion layer is characterized by three different scattering mechanisms. These three mechanisms, phonon, coulomb, and surface scattering, are commonly combined using Mattheisson's [14] rule:

$$\frac{1}{\mu_{total}} = \frac{1}{\mu_c} + \frac{1}{\mu_p} + \frac{1}{\mu_{sr}} \tag{2.4}$$

where μ_c is the coulomb limited mobility, μ_p is the phonon limited mobility, and μ_{sr} is the surface roughness limited mobility.

Because of this formulation, the lower mobility at any instance will dominate. Thus it is useful to graphically look at the dependence of each component versus the vertical field in a MOSFET inversion layer (Figure 2-9). The effective vertical field (E_{eff}) refers to the average vertical field the inversion layer electrons experience and is approximately the vertical electric field at the centroid of the inversion layer.

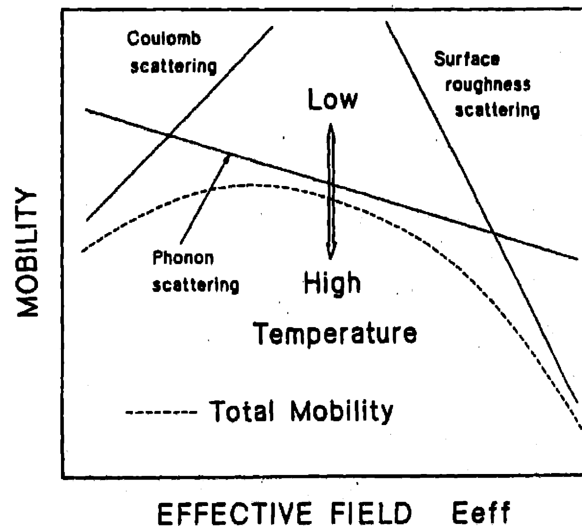


Figure 2-9: Diagram of mobility vs. vertical effective field (E_{eff}) in a MOSFET inversion layer, showing the roles of the different scattering mechanisms, from [15]

The first two mechanisms, phonon scattering and coulomb scattering, describe the mobility for bulk silicon. Phonon scattering, or momentum exchange with the silicon lattice, decreases with temperature, thus the phonon-limited mobility increases as temperature decreases. Temperature dependencies are often expressed as $\mu=T^{-n}$, where for electrons n ranges from 1.1 to 1.75 in the literature [15,16,17] with the theoretical value at 1.5 [18]. Coulomb scattering, or scattering off of charged impurities, generally is described as having almost no temperature dependence [6].

Within a MOSFET inversion layer, the analytical descriptions of these mechanisms are slightly modified. The phonon scattering is empirically found to have a vertical electric field dependence (E_{eff}^{-m}) described by a power-law with an experimentally fit exponent for electrons of m=0 to 0.3 [15,16,17]. Although experimental data are not unequivocal, in theory coulomb scattering decreases as the inversion carrier concentration increases because the carriers screen-out the charged impurities [15]. At room temperature, the charged impurities are the ionized dopants in the depletion layers. At lower temperatures interface states and fixed oxide charge also appear to impact the coulomb-limited mobility [15,6].

The third scattering mechanism, surface scattering, is due to imperfections in the $Si\text{-}Si0_2$ interface and is the key mechanism that gives MOSFETs a universal mobility curve [19]. This scattering mechanism has a large effective field dependence, and is generally considered independent of temperature [6]. Assuming that the gate oxide interface is optimized, this scattering mechanism should be independent of device design, and it also dominates at higher E_{eff}, thus removing any dependence on substrate doping for high E_{eff}. In general it is expressed as $\mu= E_{eff}^{-m}$ where m is approximately 2.6 [15,16]. In general the surface roughness component has little to no temperature dependence.

As can be seen in Figure 2-9, as the temperature is reduced, the phonon mobility will increase, which raises the carrier mobility significantly at lower E_{eff}, but at very high E_{eff}, the role of surface roughness scattering will dominate and much less of a mobility increase should be observed.

Data for long channel (L_{gate}=20 μm) NMOS device show both the expected increase in mobility magnitude as well as a change in vertical electric field dependence

versus temperature. As temperature decreases, the surface roughness, which has a strong dependence on field, becomes more pronounced and although the peak mobility rises considerably, the rise at higher fields is smaller.

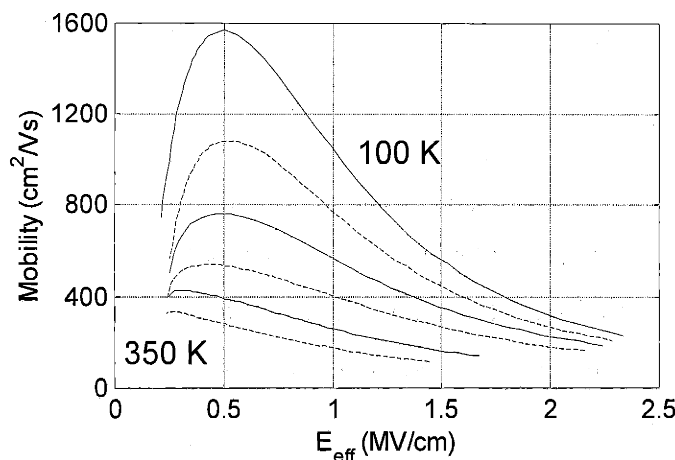


Figure 2-10: Electron Mobility measured on a L_{gate}=20 μm NMOS device at T=[350,300,250,200,150,100] K.

Although plotting mobility vs. E_{eff} allows an accurate comparison of mobility versus temperature independent of the device it is measured on, it does not give a clear picture of the mobility gain in the on-state ($V_{gs}=V_{ds}=V_{dd}$) of the device. Because the threshold voltage of the device used for the mobility measurements will be too high at lower temperatures, using a constant gate overdrive ($V_{gs}-V_{th}$=1.5V) should give a more accurate comparison of what the mobility will be at high V_{gs} for a device with an optimal threshold voltage. Figure 2-11 shows the mobility at a constant gate overdrive increasing about 3x from 300 K to 100 K. This is a significant gain, but much less than the 5x increase in peak mobility versus temperature. For comparison to Figure 2-10, the E_{eff}=1 and 1.4 MV/cm points are also shown.

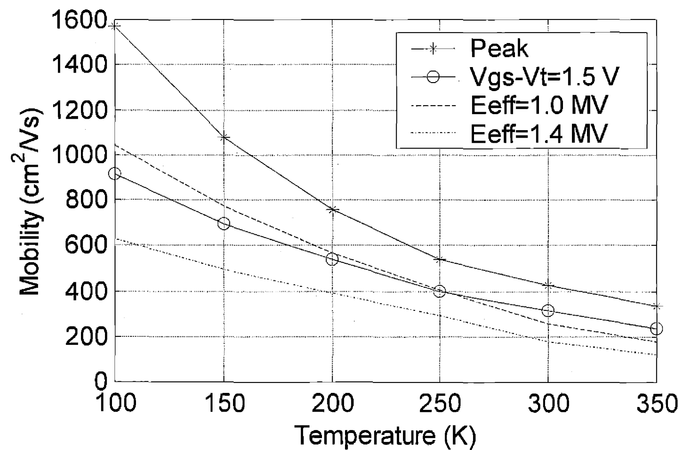
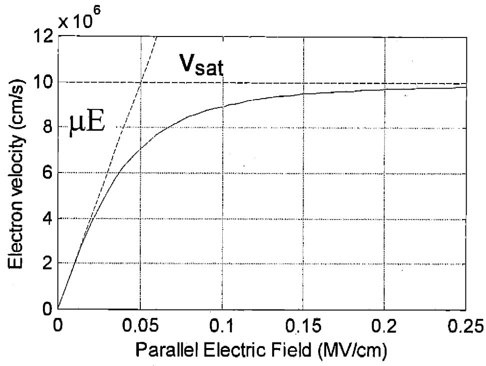


Figure 2-11: Electron mobility versus temperature. The mobility at constant V_{gs}-V_{th} = 1.5 V increases less than the peak mobility due to the impact of surface roughness scattering.

2.3.2 Saturation Velocity

In general, the saturation current of a MOSFET is far lower than would be predicted by mobility alone regulating the velocity of the carriers. The concept of velocity saturation captures the empirical decrease in velocity gain with an increase in driving field. In reality, however, electrons in short devices exceed such a velocity, but the numerical simplicity of this concept and its ability to model the I-V characteristics (analytically or in 2-D numerical simulation) make it a useful tool. In general the saturation velocity used in device modeling is regarded as an empirical parameter that is fit for a particular channel length and a particular design. A common form assumed for the velocity vs. lateral electric field is [20]:



$$\mu = \frac{\mu_s}{\left[1+\left(\frac{\mu_s E_{//}}{v_{sat}}\right)^{\beta}\right]^{1/\beta}} \qquad (2.5)$$

Figure 2-12: Plot of velocity versus lateral electric field as described by equation (1.4) ($\beta = 2$, $\mu_s = 200$).

At low lateral electric fields, the velocity increases linearly with field with a slope equal to the mobility. At higher fields the velocity grows more slowly with E-field and eventually saturates. This behavior of a slower rise in carrier velocity with electric field as the electric fields get larger is what is commonly referred to as velocity saturation. Theoretically, this saturation velocity increases slowly with temperature according to[7]:

$$v_{sat}(T) = \frac{2.4*10^7}{1+0.8\exp\left(\frac{T}{600}\right)} \qquad (2.6)$$

One commonly used measure of the velocity of carriers near the source of the device is $g_{m,max}/WC_{ox}$ [21]. This metric has been shown to display a unique tradeoff with the magnitude of DIBL that is insensitive to oxide thickness and threshold voltage. Plotting the $g_{m,max}/WC_{ox}$ vs. DIBL for a rage of lengths down to 50 nm for the NMOS devices used in this thesis shows, irrespective of channel length, a significant increase in velocity with lower temperature as expected.

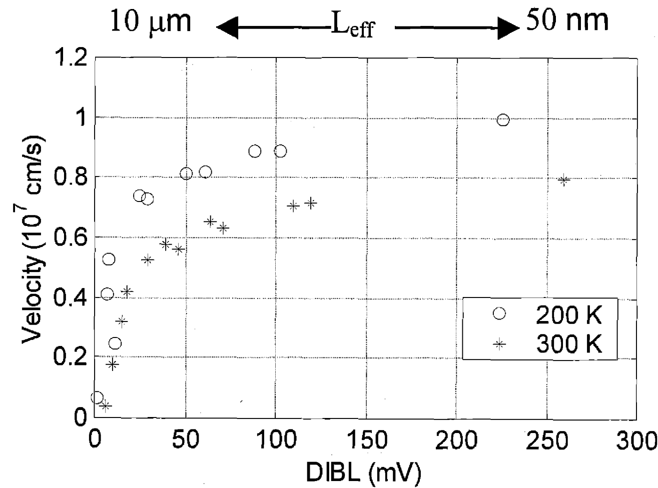


Figure 2-13: Carrier velocity ($g_{m,max}/WC_{ox}$) versus Drain Induced Barrier Lowering. The 200 K values are significantly increased from the 300K values. Channel length is the implicit variable that is changing in this plot. Higher DIBL corresponds to a shorter L_{eff}.

The measured gain in velocity for the 80 nm device (DIBL of 120 mV at 300 K) from Figure 2-13 is much less than the long-channel mobility gain, but almost twice the theoretical gain in saturation velocity from equation (2.6) (Figure 2-14). The fact that the measured velocity gains are higher than the theoretical gain in saturation velocity suggests that mobility continues to play a role at such short channel lengths and lower temperatures.

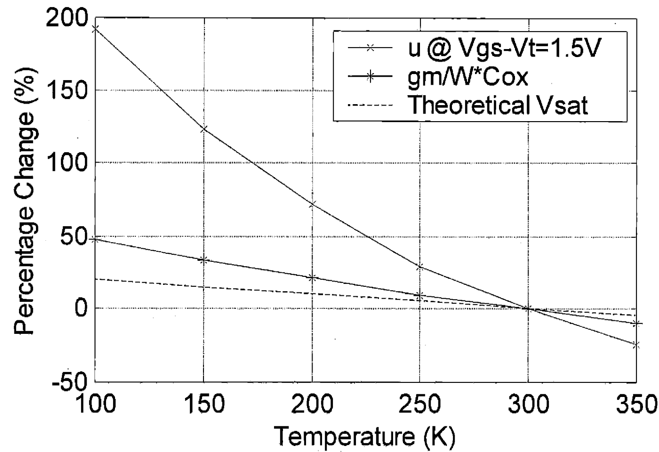
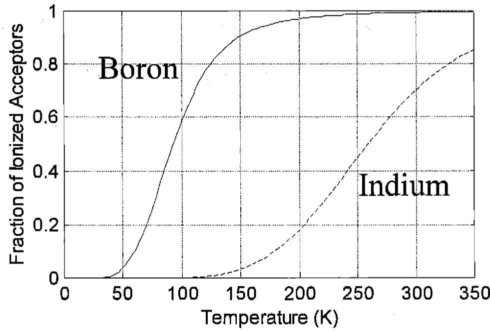


Figure 2-14: Comparison of percentage gain in measured velocity ($g_{m,max}/WC_{ox}$) versus temperature with % gain in mobility (at V_{gs}-V_{th} = 1.5 V) and the theoretical gain in saturation velocity, from equation (2.6).

2.4 Incomplete Ionization of Dopants

Impurities in silicon contribute electrons or holes by becoming ionized – that is, their extra electron or hole gain enough thermal energy to become free carriers and can contribute to current conduction. In an energy band perspective, the dopants sit at a level close to the valence band (for holes) and conduction band (for electrons). When there is enough thermal energy, their extra electron or hole becomes ionized and moves into the conduction/valence band. This process is generally described by equations (2.7). At room temperature, most of the dopants are ionized.



$$p^{+} + N_D^{+} = n^{-} + N_A^{-}$$

$$p = N_V \exp\left(\frac{E_f - E_c}{kT}\right) \tag{2.7}$$

$$N_A^{-} = \frac{N_A}{1 + G_{VB} \exp\left(\frac{E_a - E_f}{kT}\right)}$$

Figure 2-15: Percentage ionization of Boron and Indium versus temperature. For the equation, p+ are free hole, N_d+ are ionized donors, n- are free electrons, and N_A are ionized acceptors. N_c,N_v are density of states, E_f is Fermi level, band edges are E_c, E_v, donor /acceptor ionization energy levels E_d, E_a, degeneracy of acceptors and donors are GCB,GVB.

As is visible in Figure 2-15, Indium freezes out at higher temperatures than boron due to its deeper energy level.

An important exception to this behavior is when dopant concentrations rise above the 2-4x10^{18} cm^{-3} range. At these concentrations, the donor/acceptor levels merge with the conduction/valence bands, and the dopants are permanently ionized. Known as the Mott transition [22], this shift in behavior for dopant concentrations above 2-4x10^{18} cm^{-3} indicates that freeze-out is not an issue for the heavily doped source, drain, and gate polysilicon of a MOSFET.

Although the dopant concentrations in the depletion region of a MOSFET are generally below the Mott transition, the fermi level is above the dopant energy levels in the depletion region and thus the dopants are fully ionized according to equation (2.7) (see Figure 2-16, band diagram C) [23,24]. The ionization of the dopants in the depletion region has been experimentally demonstrated for MOSFETs by comparing the doping extracted from capacitance voltage characteristics at 290 K, 50 K, and 10 K [6].

The one case where dopant freeze-out at lower temperatures becomes an issue is around the flatband of the device (in the range of V_{gb}=–1 V for a NMOS device with an n^+ polysilicon gate). Although not important for the operation of a MOSFET since it almost always stays in inversion, it is important for modeling the full capacitance-voltage characteristic (C-V) of the device. At flatband, the fermi level is now below the acceptor

level (for p-type doping) and the dopants are no longer ionized. This freeze-out at flat-band (Figure 2-16 right) is visible in the C-V as a dip (point B, Figure 2-16 left). At this point the Debye length, which determines the capacitance here, has grown much larger since there are fewer ionized dopants [25].

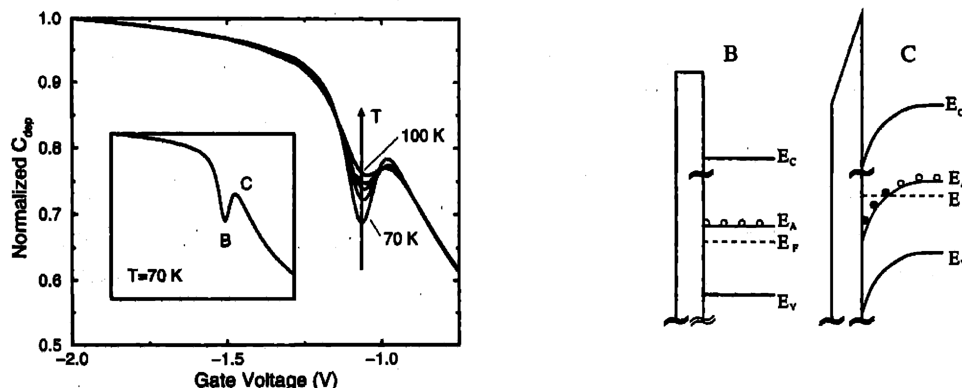


Figure 2-16: C-V and accompanying band diagrams showing freeze-out at flatband (B) and ionization of the dopants in the depletion region (C), from [25].

The measured results from the 20 μm NMOS device show a similar behavior with temperature (Figure 2-17A). As expected, the inversion capacitance is constant versus temperature (Figure 2-17B). Just as the subthreshold slope is steeper at lower temperatures, the inversion layer forms more quickly, thus the faster turn-on of the capacitance.

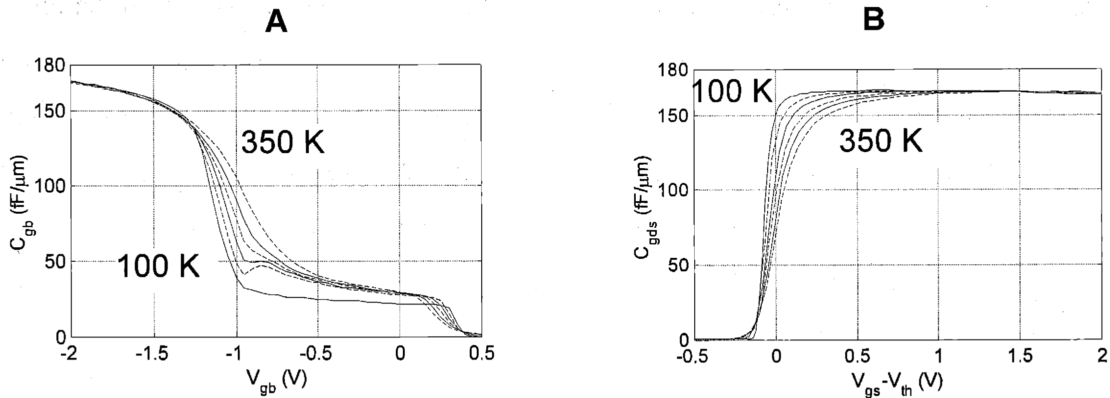


Figure 2-17:

A: Gate to substrate capacitance curves from the 20 μm NMOS device. As temperature decreases, dopants begin to freeze out around the flatband voltage, causing a decrease in the capacitance. T = [350, 300, 250, 200, 150, 100 K].

B: Gate to source/drain capacitance curve versus $V_{gs} - V_{th}$ for the 20μm NMOS device. The inversion capacitance is constant versus temperature. T = [350, 300, 250, 200, 150, 100 K].

2.5 Parasitic Source and Drain Resistance

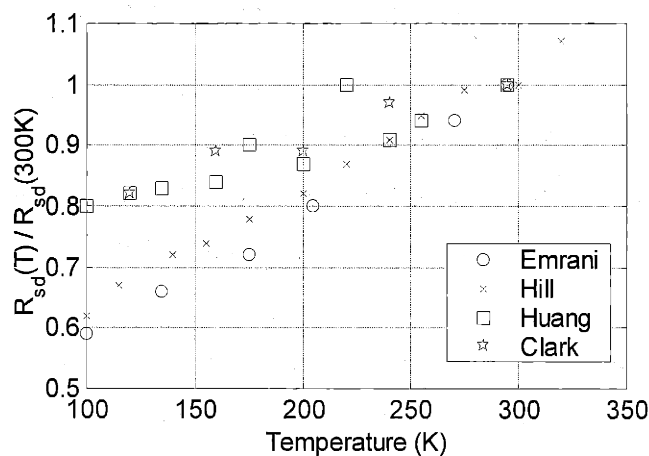


Figure 2-18: Percentage change in source/drain series resistance (R_{sd}) versus temperature from published reports in the literature. [26,27,28,29]

The parasitic source/drain resistance of a device is found to decrease 20-40% as temperature is decreased from 300 K to 100 K. The series resistance is a combination of the contact resistance, the sheet resistance of the source/drain, a spreading resistance, and an accumulation layer resistance near the channel [30]. The increase in mobility at lower temperatures should improve the sheet resistance, spreading resistance, and accumulation layer resistance [6]. The change in contact resistance with temperature is a function of the surface doping (Figure 2-19). For the heavily doped (~2x10^{20} cm^{-3}) source/drains in modern devices, the contact resistance should be at worst constant with temperature.

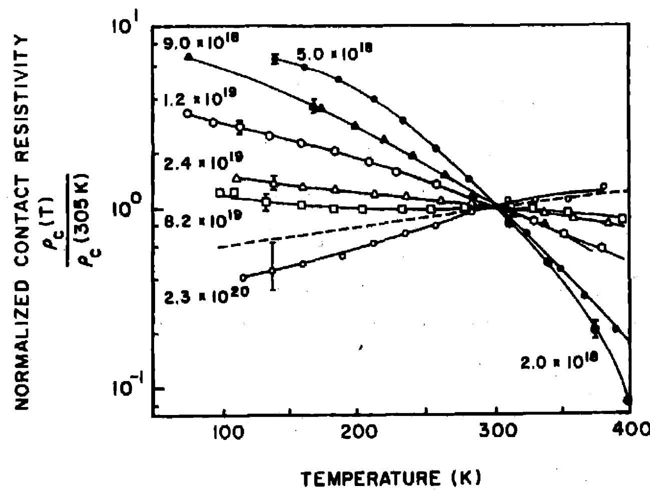


Figure 2-19: Specific contact resistance as a function of temperature normalized to the T = 305K for W contact to Si:P. Surface doping concentration is labeled in cm^{-3}. Dashed curve is theoretical prediction for 2.3 x 10^{20} cm^{-3}, from [31].

Finally, the silicide resistance, which accounts for some of the sheet resistance, is also found to decrease as temperature is decreased. Results for titanium silicided ($TiSi_2$) polysilicon resistors show a 4x reduction in sheet resistance between 300 K and 77 K [32].

This reduction in the parasitic series resistance with temperature is an important part of the overall performance gains of MOSFETs at low temperatures. Without it, the decreased resistance of the intrinsic MOSFET, from higher mobility and saturation velocity, could be largely hidden.

2.6 Interconnect

The interconnect between devices represents another key component of circuit delays. Decreasing, or at least maintaining, propagation delays as feature lengths scale continues to be a challenge [33]. Changes in materials and design strategies have so far provided the reductions in resistance and capacitance needed. Lower temperatures can significantly reduce the interconnect resistance, providing another way to address the issue of propagation delay.

A simple metric to capture the propagation delay of a signal along an interconnect line is to calculate the RC time constant of the interconnect line. Assuming an interconnect line surrounded in all four directions by similar interconnects (thus four parallel plate capacitors), one expression of this time constant is:

$$RC = 2\rho\varepsilon\varepsilon_0\left(4\frac{L^2}{P^2}+\frac{L^2}{T^2}\right) \tag{2.8}$$

where L is line length, P is the minimum metal pitch (the line-width is assumed to be ½ of P), and T is the metal thickness [33]. This delay is the performance limiter only in the case where the time constant for the output node of the driver transistor is much shorter than this propagation delay; a situation true only for long interconnect lines on a chip [34]. Unlike device scaling, straight scaling of interconnect in which the pitch (P) decreases, increases the RC delay.

Techniques to at least maintain this delay with scaling have addressed both the materials part of this equation ($\rho\varepsilon\varepsilon_0$) and the dimensional design side of this equation ($4L^2/P^2 + L^2/T^2$). Materials changes have included the move from Aluminum to Copper, which reduced the resistivity of the lines from about 3.0 μΩ-cm for Al-0.5% Cu to 1.7 μΩ-cm for pure Copper [33]. In addition, low dielectric constant (ε) materials are being introduced, with the best of these reducing ε by a factor of 2 [33]. On the design side, strategies for using large or "fat" interconnect lines on higher metal levels and fully scaled lines on lower metal levels have allowed the average pitch to decrease slowly with

scaling and thus line resistance to increase more slowly than in a straight scaling scenario[35].

Lowering the operating temperature decreases the resistivity (ρ) of the interconnect. Measured results on pure Aluminum thin films (Figure 2-20) show the resistivity decreasing by about 0.01 μΩ-cm/°C for a 250 nm thick film.

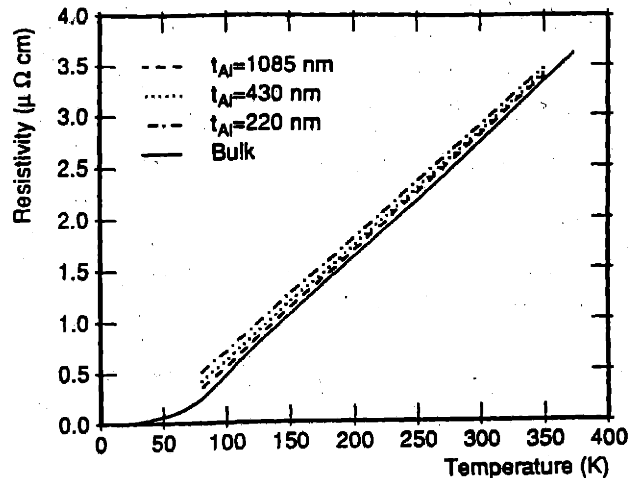


Figure 2-20: Thin film and bulk resistivity of Al versus temperature, from [34]

A shift from 300 K to 200 K would reduce the resistivity by 1 μΩ-cm from 3.96 μΩ-cm to 2.96 μΩ-cm – a 25 % improvement. This would directly translate into a 25% reduction in RC delay. This reduction in RC delay with temperature suggests that little redesign of the interconnect will be necessary to take advantage of the increase in device performance as temperatures decreases.

Measurements on copper interconnect lines show copper resistivity similarly decreasing with temperature. Over the 400 K to 100 K range, copper maintains a resistivity 1.3 to 2.2 times smaller than aluminum at the same temperature [36].

One of the other issues constraining the scaling of interconnect is the degradation of the metal lines that occurs at high current densities known as electromigration. This process, where the metal moves at grain boundaries in ways that can narrow the line-width and increase the interconnect resistance, is thermally activated (i.e. the time till a metal line fails is exponentially dependent on 1/T). Thus, as temperatures are lowered,

the process slows down, allowing higher current densities to be carried with the same level of reliability [37]. The reduction in electromigration could be especially important in terms of taking advantage of the larger current drive of transistors at lower temperatures.

Chapter 3
Scaling Length, Scaling Temperature

The relentless push for faster switching devices and greater packing densities embodied in Moore's Law [2] has historically been satisfied by scaling the dimensions of the device and adjusting the doping and voltages accordingly. Looking forward, the non-scaling of the subthreshold slope stands squarely in the way of either scaling the threshold voltage or meeting reasonable off-current levels. Lower temperature operation allows the full performance gain of scaling to be realized by scaling the subthreshold slope. In addition, the higher mobility and lower parasitic resistances in the device and interconnect that occur at lower temperatures can help maintain the expected performance increases even if some of the device parameters are not able to be fully scaled.

3.1 Previous Results

Lowering temperature to improve device performance has been an active area of research for about 25 years. Studies have shown for various generations of devices – 1 μm [38], 0.5 μm [39], 0.25 μm [40,41,42], 0.18 μm [43], 0.1 μm [44], 0.05 μm [45] - that devices can be designed and operated at liquid nitrogen temperatures and achieve much higher performance. For longer channel lengths (2 μm to 0.5 μm), approximately a 2x improvement in ring oscillator switching frequency has been observed with a change

from 300 K to 77 K [39,40,6]. Experimental investigations and analytical modeling have been used to suggest that for 0.1 μm devices the gain from 300 K to 77 K in ring oscillator switching frequency is in the 1.6x –1.8x range [44,46].

The concept of using a forward substrate bias to lower the threshold voltage has been previously shown to work and also to help with scaling by reducing the built-in potential between the substrate and source/drain [40,44,6].

Theories for scaling temperature at a fixed channel length have been explored that aim to scale the current and voltage proportionally to temperature while keeping constant the distribution of mobile carriers in the device.[47,48] Like the standard length-scaling theory described in the next section, this approach starts with a working design and then systematically adjusts it to achieve another optimal design point. This approach reduces all of the voltages (and thus electric fields) and doping in the device by the same factor as the temperature is reduced. In addition, a forward substrate bias is applied to scale the source/drain to substrate built-in potential by this same factor. The result is a device which has the same short channel effects, slightly higher drive currents, and much lower power dissipation when operated at the scaled temperature. This approach has been combined with a traditional length-scaling theory to produce 0.18 μm MOSFETs operating at 77 K that were based off a 300K 0.8 μm design.[49,50]

In contrast to previous work, this work views lowering temperature and using a substrate bias as tools to achieve the full performance gains that scaling can bring. Within this perspective, temperatures well above 77 K could be quite useful.

3.2 Scaling Theory

Scaling devices to improve their performance centers around the idea of reducing their channel length, which simultaneously reduces the resistance of the channel and decreases the capacitance of the device. Theories of scaling provide guidelines about how to move from a good device design to the design of the next generation or next smaller device. Pioneered when devices were moving from 5 μm to 1 μm channel lengths, the scaling concept was to maintain constant electric field patterns and magnitudes in the device by scaling the voltages, dimensions, and doping by the same

factor [51]. Termed constant field scaling, this approach was derived from first order MOSFET equations and Poisson's equation [51,52]. It makes the assumption that the initial device is optimal and that larger electric fields (giving larger currents) were ruled out by material or reliability issues.

Two key changes have been made to this approach. First, because devices and materials were able to stand much higher internal electric fields, the power supply was scaled more slowly than physical dimensions [53]. Second, since the built-in potentials in the device do not scale, channel doping has increased more slowly than predicted to allow the threshold voltage to scale. However, the spirit of scaling continues to be followed as thinner gate oxides and shallower source/drains are still key steps to create shorter channel devices with controlled short channel effects. Today, the internal fields of the device have been pushed very close to reliability limits and scaling has almost returned to a constant field scaling approach.

The framework proposed by the constant field scaling theory clearly identifies the key changes that occur as devices are scaled. With the standard change of a 30% reduction in length per generation (k=0.7), constant field scaling requires:

Lengths:	$L' = 0.7L$	$W'=0.7W$	$t_{ox}'=0.7t_{ox}$
Voltages:	$V_{dd}'=0.7V_{dd}$	$V_{th}'=0.7V_{th}$	
Doping:	$N_b' \approx N_b/0.7$		
Currents:	$I_{on}'=I_{on}$ (A/μm)	$I'=W'I_{on}=0.7WI_{on}$ (A)	

Where L is the channel length, W is the device width, t_{ox} is the oxide thickness, V_{dd} is the power supply, V_{th} is the threshold voltage, N_b is the doping the bulk of the device, and I_{on} is the current at $V_{gs}=V_{ds}=V_{dd}$.

Using a delay metric for an inverter that will be further explored in section 3.3, the switching delay (τ) can be shown to scale by the same factor (k=0.7):

$$\tau' = \frac{C_{Load}' V_{dd}'}{I'} = \frac{0.7V_{dd}}{0.7W} * \frac{0.7C_{Load}}{I_{on}\,(A/\mu m)} = 0.7\frac{V_{dd}C_{Load}}{I} = 0.7\tau \qquad (3.1)$$

For the switching delay to scale the load capacitance (C_{Load}) must scale by 0.7 and the on-current (I_{on}) must stay constant. The major components of the load capacitance are the gate capacitance (C_{gate}) of the next device and the junction and overlap capacitance ($C_{parasitic}$)[54]:

$$C'_{Load} \approx C'_{gate} + C'_{parasitic} = \frac{(0.7W)(0.7L)}{0.7t_{ox}} + 0.7C_{parasitic} \quad (3.2)$$

The gate capacitance naturally scales with dimensional scaling, but scaling the parasitic capacitance requires careful design.

The on-current can be written as a product of charge times velocity. Maintaining a constant on-current (A/μm) requires that both the charge density and carrier velocity remain constant:

$$I_{on}(A/\mu m) = (charge\ density) * (velocity) \quad (3.3)$$

Picking a point near the source of the MOSFET:

$$charge\ density \approx \left(\frac{\varepsilon_{ox}}{0.7t_{ox}}\right)(0.7V_{dd} - 0.7V_{th}) \quad (3.4)$$

$$velocity \approx \mu E_{\parallel} \leftrightarrow v_{sat} \quad (3.5)$$

Maintaining a constant charge density requires t_{ox} as well as V_{dd} and V_{th} to scale fully for each generation. The velocities in a device are between mobility times the parallel electric field ($E_{\parallel}$) for a long channel device and the saturation velocity (v_{sat}) for a very short channel device. The velocities of the carriers should remain constant if the electric fields in the device stay constant.

Looking at an I_{ds}-V_{gs} plot of an original versus scaled device demonstrates the changes due to scaling (Figure 3-1). The combination of shorter L, thinner t_{ox}, and lower V_{th} is balanced by a lower V_{dd} and gives the same I_{on} (in A/μm). A critical parameter that is not remaining constant is the off-current (I_{off} is the drain current (I_{ds}) at a gate voltage (V_{gs}) equal to 0). Because the subthreshold slope is not scaling, the decrease in V_{th} and L has causes an exponential increase in the off-current.

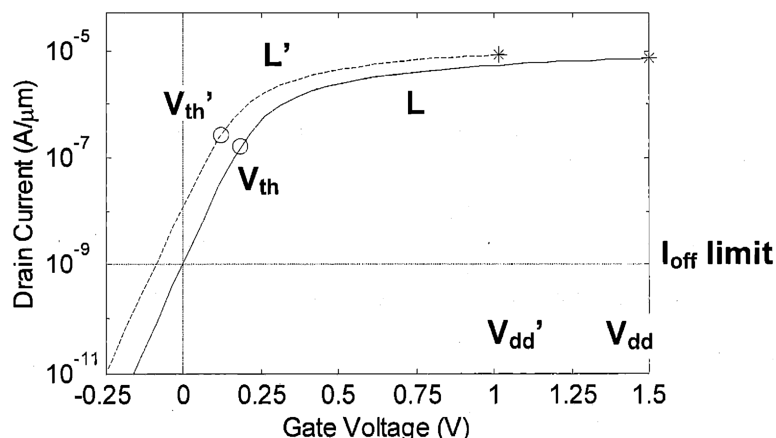


Figure 3-1: Schematic I_{ds}-V_{gs} characteristic showing the changes that occur as the device is scaled from L to L' at a constant temperature T.

Ideally a MOSFET would be fully off (i.e. conducting no current) at $V_{gs}=0$. Any off current that does exist contributes to power dissipation ($I_{off}V_{dd}$) and is undesirable in terms of limiting the total power dissipation of the chip. In addition, any off current decreases the length of time charge is stored on a node before leaking away. The reality is that a MOSFET turns off exponentially, as visible in Figure 3-1, which means it conducts a finite current at $V_{gs}=0$. As explored in Chapter 2, the rate at which it turns off, the subthreshold slope, is relatively insensitive to device design.

Limits on the off-current of a device have gradually been pushed upwards, with values around 10^{-9} to 10^{-7} A/μm (of width) common set points today. The effect of these I_{off} limits, since the subthreshold slope does not scale, is to limit the threshold voltage scaling. This means that as the power supply voltage scales, the amount of inversion charge is reduced (eqn. (3.4)), which decreases the on-current of the device (eqn. (3.3)).

Other possible deviations from the constant field scaling theory could occur if the oxide thickness or parasitic resistance do not scale. If the oxide thickness scales less than expected due to limits on tunneling currents or from reliability and manufacturability concerns, the charge density will decrease (eqn. (3.4)), thus the on-current will decrease (eqn. (3.3)), and the load capacitance will increase (eqn. (3.2)). Although not directly visible in the above equations, the parasitic resistance of the device causes a voltage drop such that the power supply voltage that appears at the MOSFET channel edges (the one

in the above equations) is less than the applied power supply voltage. If this resistance doesn't scale, then this effective power supply voltage decreases further than scaling would suggest and the charge density and velocity (due to a lower parallel electric field) decrease, causing a decrease in the on-current. Both the non-scaling of the gate oxide or the parasitic resistances would cause the switching delay to scale less than the expected 30%.

Finally, improvements in circuit speed require that both the device delay (τ) and the interconnect delay ($R_{int}C_{int}$ – see eqn. (2.8)) to scale similarly. As was explored in section 2.6, the RC delay increases with straight scaling of the interconnect, thus requiring clever design approaches or changes in materials.

The reality of device scaling is that the constant field scaling rules provide a starting point from which non-scaling parameters are compensated for by adjusting other parameters to achieve the full 30% performance gain expected each generation. A common approach, historically, has been to scale the power supply voltage more slowly to compensate for such issues as the threshold voltage or load capacitance not fully scaling. A consequence of this is that the active power density of devices increases with each generation:

$$P_{device}(W/cm^2) \approx \frac{C_{Load}V_{dd}^2}{W * L}\frac{1}{\tau} \quad \text{(active power density)} \qquad (3.6)$$

The power density faces limits within a few generations from reliability issues, and reducing the power supply voltage (V_{dd}) would be the easiest way to reduce the power density. Reducing the power supply voltage would reduce the on-current of the device and not allow the full 30% performance increase.

Reducing the operating temperature provides an approach to address each of these possible non-scaling scenarios. The non-scaling of the subthreshold slope can be directly addressed by lowering the operating temperature. The subthreshold slope is directly proportional to temperature (Figure 2-2A), so if the operating temperature is scaled by the same factor as the length (T'=0.7T), the threshold voltage can be fully scaled (V_{th}') while meeting the off-current limit and the full benefits of scaling can be realized. This scaling of temperature would allow a constant I_{off} to be maintained for the next generation of

devices (Figure 3-2). Note that at the same time the on-current of the device has actually increased due to higher carrier velocities that occur at lower temperatures.

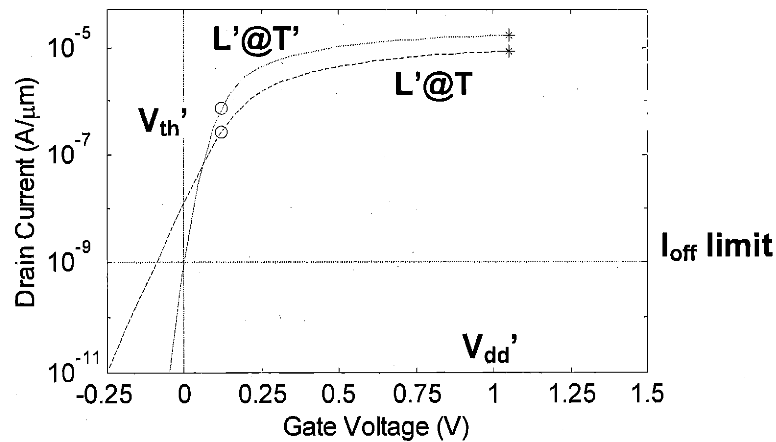


Figure 3-2: Schematic I_d-V_g showing the addition of temperature scaling to the L' scaled device. The steeper subthreshold slope allows the off-current limit to be met.

In addition, reducing the operating temperature decreases the parasitic device and interconnect resistances that need to decrease with scaling. Even more importantly, the increases in carrier mobility and saturation velocity that result in higher on-currents can be used to compensate for other non-scaling parameters and maintain the expected 30% decrease in delay with scaling. This increase in on-current could allow the power supply voltage to be reduced, thus decreasing the power density while maintaining the performance of the device.

Change as temperature is decreased	Impact on the non-scaling scenarios
Subthreshold slope becomes steeper	V_{th} can scale fully while I_{off} remains constant
Device parasitic resistance decreases	Effective power supply voltage scales fully
Interconnect resistance decreases	RC delay of interconnect scales fully
Carrier velocity increases	Increases I_{on} to compensate for other parameters that don't fully scale

Table 3-1: The impact of the changes in device characteristics described in Chapter 2 on the non-scaling scenarios discussed above.

As will be further explored in the next section, gradually lowering the operating temperature can help future device designs to meet the performance gains expected with scaling.

3.3 Scaling Scenarios

Focusing on the issue of off-current limits, different scaling scenarios will be modeled to clearly illustrate the impact of limiting off-current on the device performance. In addition this modeling will allow the impact of including temperature in a scaling scenario to be examined and will give estimates of how much the temperature would need to change.

Comparing performance at different scaling points requires optimal devices for each channel length. The device optimization will determine device design parameters (uniform channel doping, oxide thickness, and power supply voltage) for each device design and compute its threshold voltage, on-current, and off-current from those values. Using optimized devices for each channel length is key to allowing a fair comparison of performance gains between the different scenarios. By basing the performance calculations on actual device design parameters, the impact of the scenario constraints can be accurately expressed in the resulting performance calculation.

Pairing an optimizer with a set of analytical equations and a set of limits from device and material issues, the device design values came quite close to those suggested by the SIA roadmap [55], validating the approach. The set of analytical equations for short channel device characteristics used below have been shown to match a family of N-MOSFETs from L_{eff} =5 μm to 100 nm, and thus model short channel effects and velocity saturation well [56].

The goal of these optimizations is to maximize the switching speed of an inverter, as represented by an I/CV figure of merit (FOM) [57]. An inverter is chosen because it is a simple, easily modeled, and very common switching element in an integrated circuit. The rate of change of the output node of an inverter (dV/dt) is proportional to the capacitance of the output node (C) and the drive current (I_{on}) of the inverter that is

charging or discharging that capacitance. Defining $V_{dd}/2$ as the output voltage at which the inverter has switched, the switching *delay* is $(CV_{dd}/2)/I_{on}$. The case of output high to low, (NMOS device turning on) is shown in Figure 3-3.

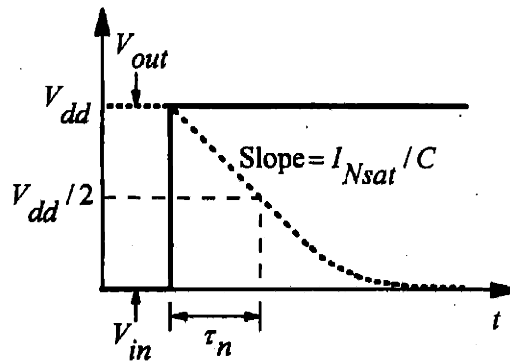


Figure 3-3: Input and output inverter waveforms for input high to low transistors, from [58].

Assuming the input to the inverter is instantaneously switched, then the NMOS device is initially biased (V_{gs}=V_{ds}=V_{dd}) such that the full on-current flows through the device and discharges the capacitance on the output node. The current stays at this level until V_{out} drops below $V_{ds,sat}$, so the on-current is the correct current to use in the delay equation. The FOM used in the optimizations below is for the switching *speed*, reciprocal of the delay, and thus is $I_{on}/(CV_{dd}/2)$.

For a chain of inverters, the capacitance at the output node is a sum of the gate capacitance of the next stage (C_{ox}), and the drain junction (C_j) and overlap capacitance of this stage. Parasitic capacitance from interconnect could also be added in, but will be ignored for the following optimizations. The performance of an actual set of inverters can best be fit by increasing the numerical multiplier of CV/I from 0.5 to closer to 0.75 which accounts for the fact that the input voltage does not instantaneously change [59]. Since only relative comparisons, not absolute values matter here, this numerical factor will be dropped and a straight I/CV FOM will be used for these optimizations.

The definition of off-current is complicated by the statistical variation of channel length that occurs in manufacturing. For a given chip or wafer, a variation of ±20% of the nominal channel length (L_{nom}) is a standard value used for a 3σ deviation from the mean or nominal device. The off-current for the device designs below is measured at the

worst-case or strongest device ($0.8L_{nom}$) since the greater short channel effects for the shorter channel length cause a lower V_{th} and thus higher off-currents. The on-current for the design, however, is measured at the nominal length (L_{nom}), the mean length of the distribution.

3.3.1 Room Temperature Scaling Scenarios

Two scenarios for scaling at a constant temperature are a scaled-V_{th} scenario and a constant-I_{off} scenario (Table 3-2). The scaled-V_{th} scenario aims to optimize the switching speed at each channel length with the constraint that the nominal device (L_{nom}) have a threshold voltage fixed at 1/5 of the power supply ($V_{dd}/5$). The constant-I_{off} scenario aims to optimize the switching speed at each channel length with the constraint that the worst case device ($0.8L_{nom}$) have an I_{off} fixed at 10^{-9} A/μm [55]. In all cases, the operating temperature of the device is picked as 80 °C (353 K) and the substrate bias (V_{bs}) is fixed at 0 V. Devices in each of these scenarios were individually optimized to maximize the nominal device switching speed FOM (I/CV) while satisfying the basic constraints of controlled short channel effects (DIBL of the worst case device is 100 mV) and the E-field across the oxide was set at the estimated material limit of 5.5 MV/cm [55]. For each channel length, the uniform substrate doping (N_b), oxide thickness (t_{ox}), and power supply (V_{dd}) are allowed to vary to meet the various constraints.

Scenario	Inputs	Variables	Constraints	Goal
Scaled V_{th}	L_{eff} T=80 °C (353 K)	N_b t_{ox} (V_{dd}, V_{th})	$V_{th} = V_{dd}/5$	Maximize FOM: $\frac{I_{on}}{V_{dd}(C_{junction} + C_{gate})}$
Constant I_{off}	L_{eff} T=80 °C (353 K)	N_b t_{ox} (V_{dd}, V_{th})	$I_{off@0.8Lnom} = 10^{-9}$ A/μm	
All Scenarios: SCE @ $0.8L_{nom}$: DIBL = 100 mV $V_{dd} = t_{ox} * 5.5$ MV/cm				

Table 3-2: Setup of room temperature scaling scenarios

Fixing I_{off}, and thus constraining V_{th}, yields a device switching speed that scales more slowly than 1/L (Figure 3-4) because the I_{on} (A/μm) is decreasing due to the non-scaling of the V_{th}. Scaling V_{th} with V_{dd} (constant field scaling) yields a FOM that grows as 1/L as would be expected from constant field scaling, but gives unacceptably high off-currents (Figure 3-5). The tradeoff between fully-scaled performance and increased off-current as L scales at a constant temperature is clearly visible in these figures. Fully scaled performance requires the V_{th} to scale, yet because the SS is fixed at constant temperature, scaling the V_{th} will cause off-currents to quickly rise to unacceptable levels.

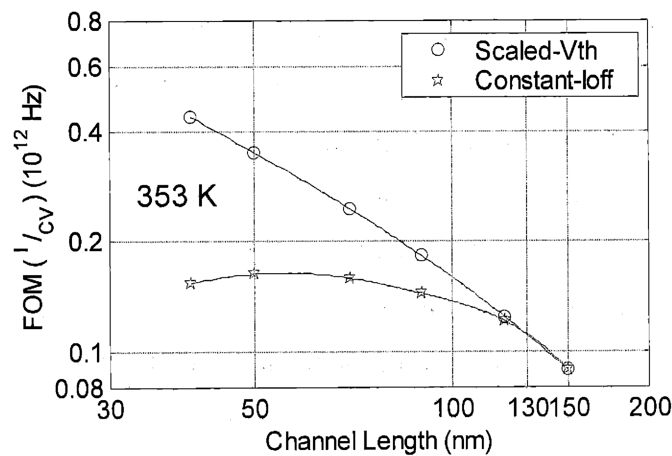


Figure 3-4: Device Performance (Nominal L) versus design node for the scaled-V_{th} and Constant-I_{off} Scenarios.

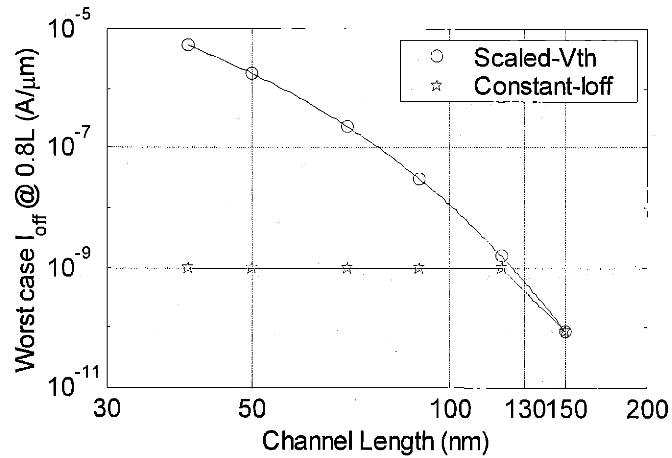


Figure 3-5: Off-current (Worst Case, 0.8*L device) versus design node for the Scaled-V_{th} and Constant-I_{off} scenarios.

The device parameters that underlie each of the design points in Figure 3-4 and Figure 3-5 fall in a reasonable range of values. The t_{ox} (Figure 3-6A) comes close to the SIA numbers [55] and shows a steady decline, especially for the scaled V_{th} case. As expected from the constant field scaling framework, doping levels climb as L is reduced (Figure 3-6B).

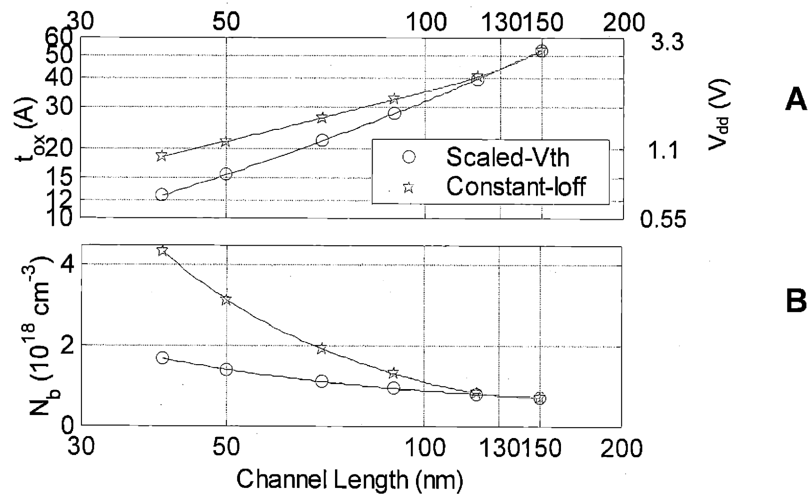


Figure 3-6: Device design parameters at each node for the Scaled-V_{th} and Constant-I_{off} scenarios. A: Oxide thickness B: Uniform channel doping level

3.3.2 Low Temperature Scaling Scenarios

Lowering the operating temperature as the device scales allows the subthreshold slope to scale and removes the tradeoff between the performance gains and increased off-currents that occurs at a constant temperature. Two different scenarios that include temperature (T) in the scaling are a low temperature scaled-V_{th} (LT scaled-V_{th}) scenario and a low temperature scaled performance (LT scaled-performance) scenario (Table 3-3). Similar to the 353K case, the LT scaled-V_{th} scenario aims to maximize the switching speed (I/CV) at each channel length. The constraints are that the nominal device has a V_{th} equal to $V_{dd}/5$, plus the additional constraint that I_{off} remain constant at 10^{-9} A/μm. The LT scaled-FOM scenario aims to match the 1/L performance gain that the 353K scaled-V_{th} case shows but with the constraint that the I_{off} remain constant.

Scenario	Inputs	Variables	Constraints	Goal
LT Scaled V_{th}	L_{eff} t_{ox}=RT Scaled-V_{th} #	N_b T V_{bs}	$I_{off@0.8Lnom}=10^{-9}$ A/μm $V_{th}=V_{dd}/5$	Maximize FOM: $\frac{I_{on}}{V_{dd}(C_{junction}+C_{gate})}$
LT Scaled Performance	L_{eff} t_{ox}=RT Scaled-V_{th} #	N_b T V_{bs}	$I_{off@0.8Lnom}=10^{-9}$ A/μm	Meet 80 °C scaled-V_{th} FOM
All Scenarios: SCE @ 0.8L_{nom}: DIBL = 100 mV $V_{dd}=t_{ox}$*5.5 MV/cm				

Table 3-3: Setup of low temperature scaling scenarios.

Similar to the room temperature scenarios, in each case, device design parameters were optimized for each channel length within the constraints of the scenarios (Table 3-3) using the variables of Doping (N_b), Temperature (T), and substrate bias (V_{bs}). The additional design variable of a forward substrate bias (V_{bs}>0 for NMOS) will allow the V_{th} to be lowered while maintaining control of short channel effects.

Scaling the V_{th} and keeping the off-current constant requires the subthreshold slope to scale rapidly (Figure 3-8) and thus the temperature drops quickly as the length is scaled. The overall performance exceeds that of the RT scaled-V_{th} case because of the increased mobility and saturation velocity at lower temperatures (Figure 3-7). Matching the RT Scaled-V_{th} performance while keeping the I_{off} constant requires the SS to scale more slowly (Figure 3-7) (V_{th} is not fully scaling) and temperature drops more slowly. Because mobility and saturation velocity have increased, the V_{th} does not have to scale fully to achieve the RT performance.

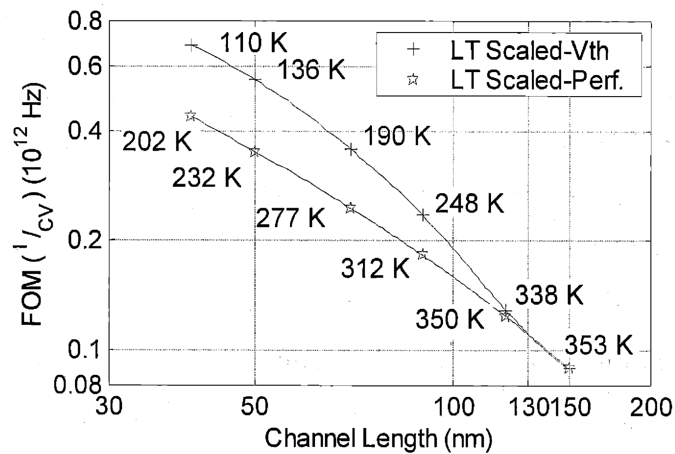


Figure 3-7: Device Performance (at nominal L) versus design node for the LT scaled-V_{th} and LT scaled-performance scenarios (I_{off}=10^{-9} A/µm for all points). Note that the LT scaled-performance case matches the performance of the RT scaled-V_{th} case (Figure 3-4).

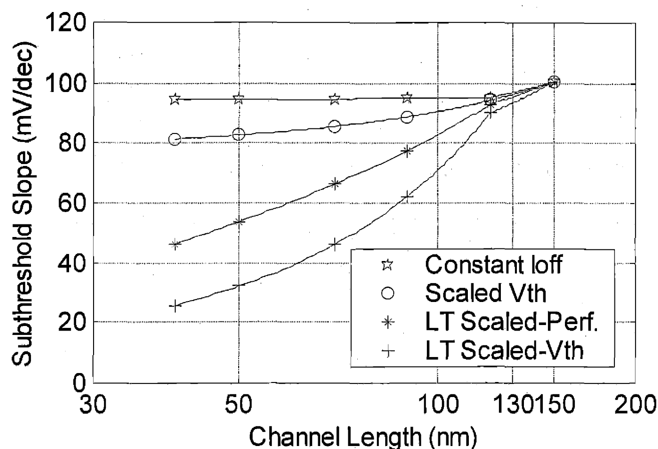


Figure 3-8: SS of optimized designs.

In each of the low temperature scenarios, a substrate bias is used as part of the design. Allowing too large of a forward substrate bias will cause the source/drain junctions to turn on and increase the off-current of the device. The range is quite large, especially as temperature decreases. Figure 3-9 compares the substrate bias suggested in the LT Scaled-V_{th} scenarios with measured data from the 0.18 μm technology from Chapter 2. The substrate bias values fall well below the 10^{-9} A/μm (I_{off}) current limit. The substrate biases for the LT scaled-performance designs are actually slightly negative (-0.15 V $< V_{bs} <$ 0 V), suggesting that for this scenario, $V_{bs} = 0$ designs should be possible.

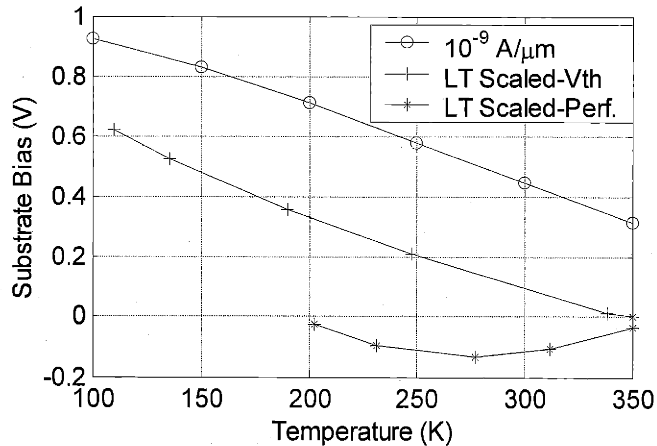


Figure 3-9: Substrate Bias for the scaled-V_{th} and scaled-Performance scenarios as compared to the measured substrate bias needed for $I_s + I_d = 10^{-9}$ A/μm .

3.4 Conclusion

Integrating temperature into a scaling scenario allows the full performance gain from scaling to be realized while maintaining a constant off-current. The additional variable of forward substrate bias adds flexibility to the design and is not overly constrained by leakage current limits. The scaling scenarios explored in this chapter suggest how temperature could be gradually reduced to either match or exceed room-temperature scaled performance.

Looking more closely at one of the design nodes requires an examination of what an optimal design for a real device would be at this temperature. In particular what balance of halo doping, retrograde channel doping, and substrate bias would give the highest I_{on} for a given I_{off}? These issues will be explored in detail in the next chapter.

Chapter 4
Device Design Comparison at 200 K

Given the motivation towards using lower operating temperatures to meet or exceed fully-scaled device performance, a device design needs to be optimized for a specific operating temperature. The analysis in this section will focus on NMOS data at 200 K (-73 °C). 200 K is near the temperature suggested by the analysis in Chapter 3 (Figure 3-7) for maximum performance at the 80 nm L_{eff} node ($I_{off}=10^{-9}$ A/μm), the design point of the NMOS devices from IBM used in this thesis.

As explored in Chapter 3, one metric for device switching speed is the reciprocal of the delay of an inverter (I/CV). Maximizing the switching speed points towards achieving the maximum current for the minimum possible capacitance. For a given device technology, the minimum channel length, the oxide thickness, and source drain resistance are generally constrained by process issues. These elements set the key capacitances and the power supply voltage. In contrast, the channel doping design and substrate bias, which influence the threshold voltage, short channel effects, and on-current, are constrained more by a designer's imagination of how to adjust them to meet the overall device criteria. The focus of this chapter will be on optimizing the on-current part of the switching speed.

The devices examined in this section were made at IBM as part of the development of a 0.18 μm technology[5]. The devices have a 33 Å physical N_2O grown oxide that yields a 41 Å electrical t_{ox} at 1.8 V (V_{dd}) in inversion. Devices with L_{eff} from 20 μm down to 50 nm exist on the wafers. The devices use a combination of retrograde channel doping and halos/pockets to achieve two different threshold voltage designs. The only difference between the two designs is in their respective channel dopings. The

"High-V_{th}" design has a long channel V_{th} about 0.42 V, the "Low-V_{th}" design has a long channel V_{th} of 0.25 V. Both designs show well behaved short channel effects down to channel lengths of about 70 nm (Figure 4-1).

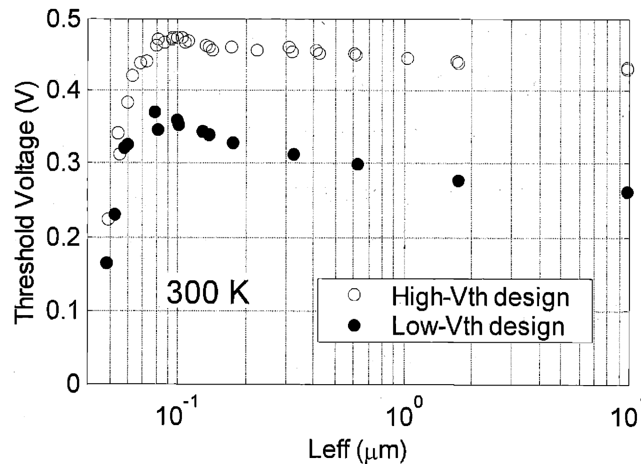


Figure 4-1: Threshold voltage versus channel length for the High-V_{th} and Low-V_{th} device designs at 300 K.

Using two different threshold devices with otherwise the same device design will allow the performance of two different combinations of doping and substrate bias to be compared.

4.1 200 K Fundamental Performance Increase

Simply cooling the devices cases a fundamental increase in their performance. One of the metrics used in industry today to compare different designs in development is an on-current versus log-scale off-current plot [5]. This plot contains data for each of a whole range of channel lengths of the same device design, i.e. having seen the exact same fabrication process. Figure 4-2 shows an I_{on}-I_{off} plot for the High-V_{th} and Low-V_{th} designs at 300 K and 200 K. At a given temperature, the difference in I_{off} at lower I_{on} (longer L_{eff}) is due to the different threshold voltages of the different designs. The turn up of the plot at higher I_{on} (shorter L_{eff}) is due to short channel effects that cause the

threshold voltage to decrease (and thus I_{off} increases) as the channel length is reduced. The merging of the two designs (for a given temperature) suggests that they have similar short channel characteristics.

The power of this graph is that shifting this whole curve to higher on-currents is only possible through fundamental device changes such as improvements in mobility, reduction of source/drain resistance, and decreasing the oxide thickness [60]. Since the only difference between the High-V_{th} and Low-V_{th} designs is their channel doping, and thus V_{th}, they fall on the same curve at short channel lengths (high I_{on}). However, cooling the devices to 200 K significantly shifts the curves right, showing a fundamental improvement in device performance through mobility/saturation velocity increases and source/drain resistance (R_{sd}) decreases. Although a similar shift would occur if the devices were re-designed for a shorter channel length (thinner t_{ox} and lower R_{sd}), the mobility and saturation velocity gains are unique to lowering the operating temperature.

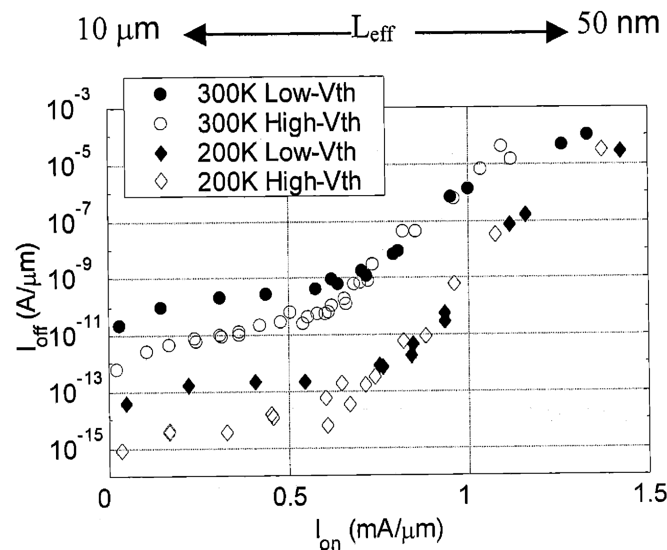


Figure 4-2: I_{on}-I_{off} plot for the Low-V_{th} and High-V_{th} designs at 300K and 200K. (V_{ds} = 1.8 V, V_{bs} = 0 V).

4.2 Performance Comparison

Determining an optimal design requires focusing on the particular channel length that will be used in circuits for a particular generation of devices. A meaningful

comparison of the performance of the High-V_{th} and Low-V_{th} designs at 200 K requires comparing them at the same off-current and channel length.

Decreasing the temperature from 300 K to 200 K shifts the threshold voltages of the devices higher (open symbols in Figure 4-3). However, applying a forward substrate bias can shift the threshold voltage lower again (filled symbols in Figure 4-3). The goal of having the lowest possible V_{th} on a device implies pushing the off-current to its maximum allowed value. Thus, by using a forward substrate bias, a given device's V_{th} can be lowered to the point where the device's I_{off} hits the limit, 1×10^{-9} A/µm in this case. The shift in I_{off} with substrate bias can be seen directly on a log I_{off} versus I_{on} graph (Figure 4-3B). In this way, each of the devices for the Low-V_{th} and High-V_{th} designs can be adjusted to have an $I_{off}=1\times10^{-9}$ A/µm.

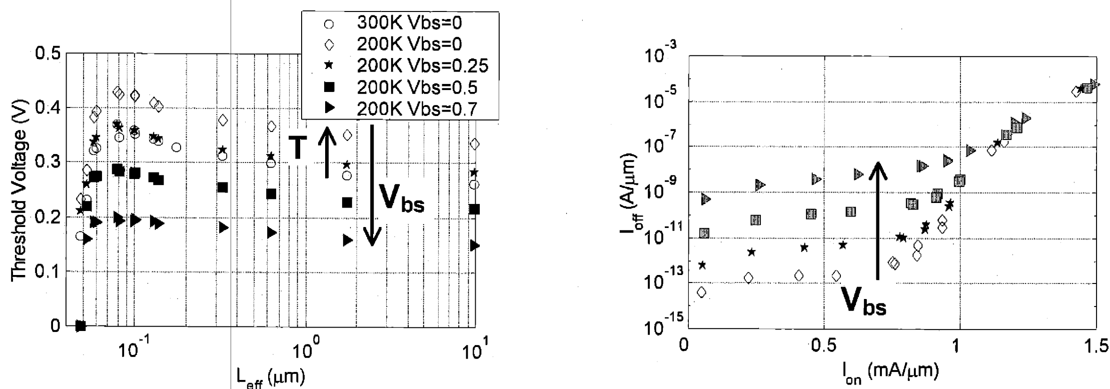


Figure 4-3:
A: Threshold voltage at V_{ds}=50 mV versus channel length for the Low –V_{th} design. Cooling the device raises the V_{th}, while applying a forward substrate bias lowers the V_{th} .
B: Log I_{off} versus I_{on} at 200 K for the Low V_{th} design. V_{bs} = [0,0.25, 0.5, 0.7 V]

Adjusting the off-currents to the same value at 200 K allows a fair comparison of the two designs. For each L_{eff} of the High-V_{th} and Low-V_{th} designs, a forward substrate bias is used to lower the threshold voltage until the device meets the I_{off} criteria (10^{-9} A/µm) at 200 K. The substrate biases applied are within the limits set by junction leakage as explored in Chapter 3 (Figure 4-3).

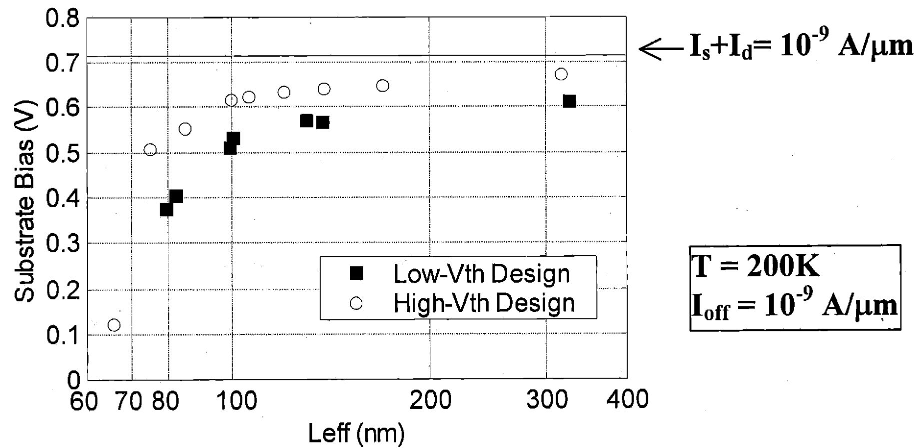


Figure 4-4: Substrate bias used at each L for each device design to reach $I_{off}=10^{-9}$ A/μm. Note that all the values fall below the limit from junction leakage ($I_s+I_d=10^{-9}$ A/μm)

With the proper substrate bias applied, plotting each device's I_{on} versus each devices' L_{eff} [61] (Figure 4-5) clearly shows that the Low-V_{th} design consistently yields a higher I_{on} for a given L_{eff}. Since a device's switching speed ($\propto$ I/CV) depends on both current and capacitance, comparing I_{on} at the same L_{eff} (approximately the same C_{gate}) allows I_{on} to give a direct measure of device switching speed. Since the two different device designs have the same source/drain design and only different channel or halo implants, it is assumed that other capacitances like the overlap and junction capacitances will be the same.

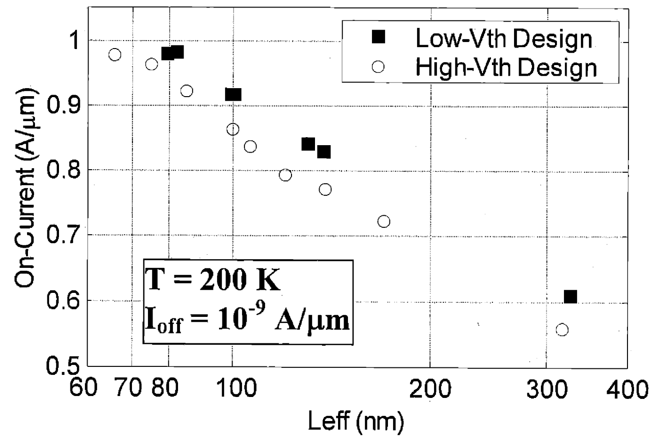


Figure 4-5: On-current vs. L_{eff} for the two designs with $I_{off} = 1x10^{-9}$ A/μm at each point (T=200K).

The origins of this difference in on-current correlate with the different threshold voltages of the two designs once they are adjusted to the same off-current. Examining device characteristics (at the V_{bs} needed to set I_{off}), the Low-V_{th} devices have a steeper subthreshold slope than the High-V_{th} devices (Figure 4-6A). Given the fixed I_{off}, the steeper SS results in a lower V_{th} (Figure 4-6B) and thus a higher I_{on} for the Low-V_{th} design. The threshold voltage at 1.8 V is used because it relates closely to the on-current. The High-V_{th} devices have both a shallower subthreshold slope and have less short channel effects (SCE) as evidenced by their lower drain induced barrier lowering (DIBL) at a given L_{eff} (Figure 4-6C). The lower SCE and worse SS indicate that the channel depletion depth (x_d) is smaller for the High-V_{th} design than for the Low-V_{th} design. The data indicate that optimizing the channel doping to maximize the on-current of a device, while meeting an off-current goal, requires the subthreshold slope to be as steep as possible to minimize the threshold voltage.

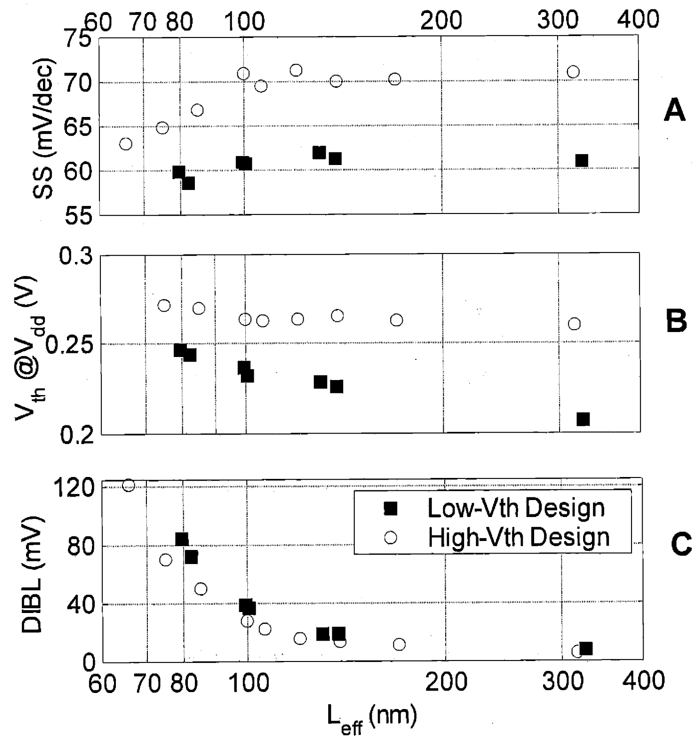


Figure 4-6: V_{th}, SS, DIBL vs. L_{eff} at I_{off}=10^{-9} A/μm. (T=200 K).

The difference in the threshold voltages does not seem to be enough on its own to cause the observed increase in the on-current. For the L_{eff} =80 nm device, the low-V_{th} design has a 4.2% higher I_{on} that the High-V_{th} design. To first order, the impact of the threshold voltage is expected to be between $(V_{gs}-V_{th})^2$ for the case of no velocity saturation and $(V_{gs}-V_{th})$ for the fully velocity saturated case. The 25 mV lower V_{th} for the Low-V_{th} design should then yield between a 1.7% to 3.3% increase in the on-current (V_{gs}=1.8 V). The expectation with such a short device is that the impact of the threshold voltage would fall much closer to the velocity saturated increase of 1.7%. The 4.2% increase in on-current observed is too high to have come purely from a change in threshold voltage.

Another difference in the designs is that the Low-V_{th} device has a much lower body effect that the High-V_{th} device. This is visible in Figure 4-7 from the larger shift in threshold voltage the High-V_{th} design shows from 0 to 0.5 V substrate bias. A larger body effect causes the inversion charge to be reduced more quickly as drain bias

increases, resulting in lower drive currents [62]. The question of what causes this on-current gain between different designs at the same channel length will be further examined in Chapter 5.

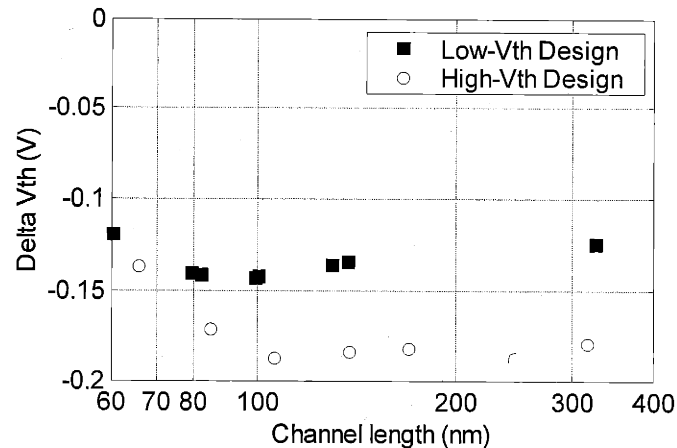


Figure 4-7: Delta V_{th} from $V_{bs} = 0$ to 0.5 V for High-V_{th} design and Low-V_{th} design devices at 200 K.

The Low-V_{th} design yields higher on-currents with its lower body effect and deeper depletion depth that the High-V_{th} design at the same off-current. This suggests that an optimal design for low temperature will require a combination of lower doping (lower room temperature V_{th}) and substrate bias. In addition, the depletion depth should be as large as possible to give a steep subthreshold slope and lower body effect, yet no so large as to cause too large a DIBL.

4.3 Role of the depletion depth

The higher I_{on} of the Low-V_{th} design correlates with a steeper subthreshold slope, a lower V_{th}, and higher DIBL. Each of these point to the Low-V_{th} design having a larger depletion depth in the channel (x_d). This suggests that beyond meeting an off-current limit, optimizing a device requires finding a balance between a steep subthreshold slope (large x_d) and controlled short channel effects (small x_d).

The impact of the depletion depth on the subthreshold slope can be explained through the perspective of a capacitive divider. The subthreshold slope occurs in weak inversion, where the depletion charge is much larger than the inversion charge [63] and can be considered equal to the total charge in the silicon. Thus the gate voltage is dropped across two regions; the oxide, then the band bending (the depletion charge) in the silicon [64].

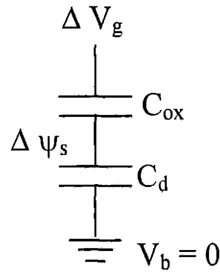


Figure 4-8: Circuit diagram representation of the gate capacitance in subthreshold. ΔV_g is the incremental gate voltage, C_{ox} is the gate capacitance, C_d is the depletion region capacitance and $\Delta\psi_s$ is the incremental surface potential.

A change in the gate voltage (ΔV_g) is divided such that the change in surface potential is [65]:

$$\Delta\psi_s = \Delta V_g\left(\frac{C_{ox}}{C_d + C_{ox}}\right) \tag{4.1}$$

Although the inversion charge, and thus the subthreshold current, is an exponential function of the surface potential (ψ_s), the subthreshold slope reflects the change of the logarithm of inversion charge. Thus the subthreshold slope (SS) is directly proportional to this $\Delta\psi_s$ expressed in equation (4.1) and the same capacitive voltage divider shows up directly in the equation for the subthreshold slope:

$$SS = \frac{\Delta V_g}{\Delta \log(I_d)} \propto \frac{\Delta V_g}{\Delta\psi_s}$$
$$SS = 2.3 * \frac{kT}{q} * \left(\frac{C_d + C_{ox}}{C_{ox}}\right) \quad \text{where} \quad C_d = \frac{\varepsilon_{si}}{x_d} \text{ and } C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \tag{4.2}$$

To be exact, equation (4.2) actually describes the inverse subthreshold slope [23], but it is common practice to use the definition in (4.2).

Changing the depletion capacitance (C_d) can have a large impact on the subthreshold slope. Substituting in the equations for C_d and C_{ox}, the SS can be written as:

$$SS = 2.3 * \frac{kT}{q} * \left[1 + \frac{\varepsilon_{si}}{\varepsilon_{ox}} \left(\frac{t_{ox}}{x_d} \right) \right] \tag{4.3}$$

Using a positive substrate bias (with respect to the source) on an NMOS device decreases the depletion depth and degrades the subthreshold slope (makes it larger). A positive substrate bias decreases the amount of band bending needed to invert the surface (Figure 4-9). The depletion depth will be smaller (less band-bending = less depleted charge = less depletion depth) thus increasing C_d and making the subthreshold slope worse (larger) (Figure 4-9).

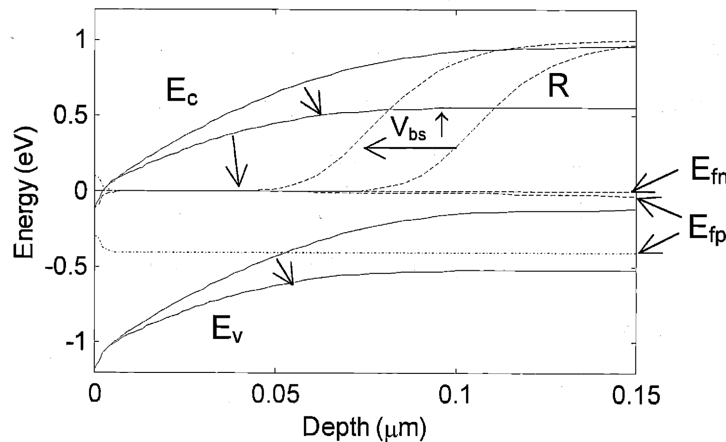


Figure 4-9: Simulated energy bands vs. depth in the silicon without substrate bias (black) and with +0.4 V substrate bias (gray). (V_{gs} = 1 V, V_d = V_s = 0 V) The dotted line labeled R is the ratio of the hole concentration to channel doping. The depletion depth, which shrinks with V_{bs}, is at the point where R = 0.5.

The impact of a substrate bias on the threshold voltage can also be seen in Figure 4-9. Because in a MOSFET the inversion layer (electrons in this case) is in

communication with the source (n^+ doped), the electron Fermi level is not shifted, but the hole Fermi level, which communicates directly with the substrate (p doped) contact, is shifted down by the positive substrate bias. Thus the hole concentration in the bulk of the substrate remains the same, yet the amount of band bending needed to invert the device decreases. The depletion depth is decreased and less depletion charge is exposed. However, both the gate voltage and the electron Fermi-level in the channel have not changed, so the same voltage drop exists across the oxide. Thus, the same charge is on the gate. Given that the gate charge must equal the sum of the inversion and bulk charge, the fact that there is less bulk charge must be balanced by more inversion charge, resulting in a lower threshold voltage. Applying a forward substrate bias reduces the threshold voltage at the same time it reduces the depletion depth.

The sensitivity of the threshold voltage to the substrate bias, the body effect, is also related to the depletion depth. Looking back at the capacitive divider (Figure 4-8), for a given ΔV_{bs} (ΔV_b - $\Delta\psi_s$), a larger capacitance (smaller x_d) causes a larger shift in the depleted charge. Thus, the smaller the depletion width, the larger the body factor.

This change in threshold voltage and subthreshold slope due to a substrate bias can be contrasted to the case where the substrate doping is increased (Figure 4-10). With a higher substrate doping, more band bending is required to invert the surface, and the threshold voltage increases (for the same V_{gs} you get more bulk charge and less inversion charge). Yet because the substrate is doped more heavily, the depletion depth will actually shrink. Thus a shallower maximum depletion depth can be achieved with higher doping giving a larger SS, but in contrast to a forward substrate bias, the threshold voltage will be increased.

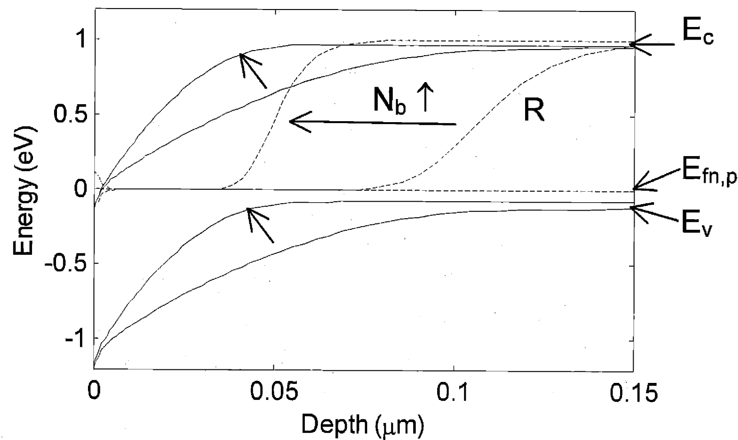


Figure 4-10: Simulated energy bands vs. depth in the silicon with 1x10^{17} cm^{-3} doping (black) and 5x10^{17} cm^{-3} doping (gray). (V_{gs} = 1.0 V, $V_d = V_s = V_b$ = 0 V) The depletion depth is at R=0.5 (dotted line). Increasing the doping shrinks the depletion depth.

For short channel effects (SCE), the depletion depth at inversion (x_d) is a critical indicator of the electrostatic integrity of the device. The ideal situation is for the gate to completely control the inversion and bulk charge in the channel, thus turning the electrostatics into a one-dimensional vertical problem with the source and drain having negligible effect. In a geometric sense, this happens when the box defined on top by the gate-oxide, bottom by the maximum depletion depth, and on the sides by the source and drain is much wider than it is tall (Box A-B-E-F-A in Figure 4-11). As the channel length decreases and the box moves towards a square shape, the source and drain start to have a big impact on the charge in the channel and the problem becomes more two-dimensional – thus short channel effects. In this sense, the ratio of the x_d to the channel length should be definitely less than one for short channel effects to be controlled. Given a particular channel length, the smaller the depletion depth, the less the short channel effects (of which DIBL is one). This picture is one used by Taur et al. in their analytical model for short channel effects, and for which the SCE depend on the ratio of the channel length to a combination of the maximum depletion depth and the oxide thickness [66].

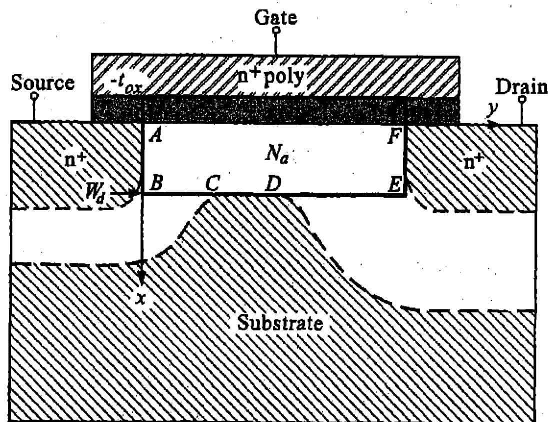


Figure 4-11: MOSFET schematic showing the geometry of the depletion depth (W_d) and the channel length. L_{eff} needs to be greater than x_d for SCE to be controlled, from [66].

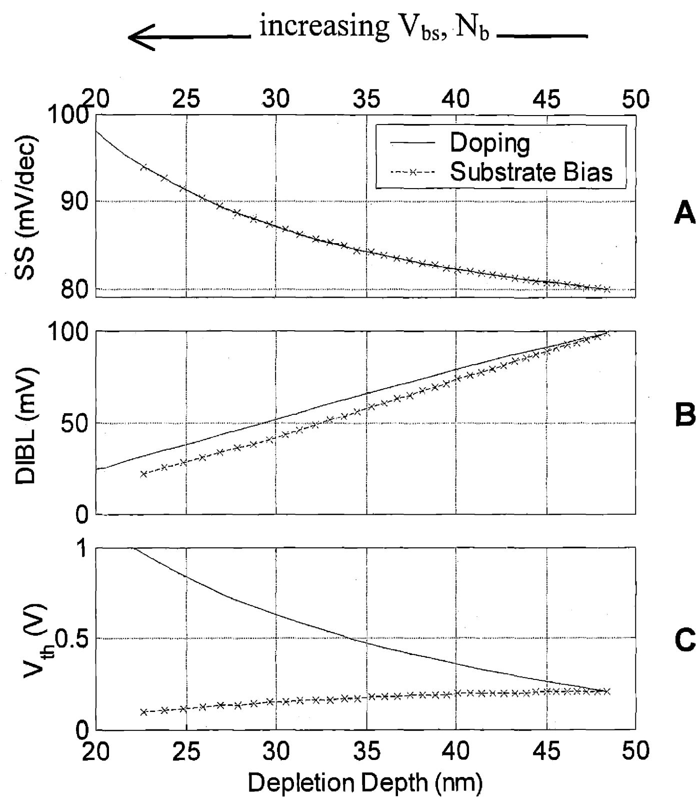


Figure 4-12: Plot of A: Subthreshold Slope (SS), B: Drain Induced Barrier Lowering (DIBL), and C: Threshold Voltage (V_{th}) vs. depletion depth using the analytical equations from Taur [67] (L_{eff} = 80 nm, t_{ox}=41 Å).

The depletion depth is directly related to both the subthreshold slope and the magnitude of the short channel effects. Using Taur's analytical equations, the impact of changing the depletion depth (x_d) by changing the doping or substrate bias is shown in Figure 4-12. Increasing either the forward substrate bias or channel doping reduces x_d and thus increases the SS and decreases the DIBL (a measure of the short channel effects) (Figure 4-12A,B). However, a key difference is that using a forward substrate bias reduces the V_{th} as the depletion depth decreases (Figure 4-12C). This is the opposite effect of using doping to reduce the depletion depth, where the threshold voltage increases.

By combining an increase in forward substrate bias with a decrease in substrate doping, the resulting device can have a reasonable depletion depth and a normal subthreshold slope. The degraded SS often associated with a forward substrate bias is really an indication of too shallow of a the depletion depth.

A forward substrate bias is attractive at lower temperatures where the optimal design has a lower threshold voltage than one at room temperature, yet the short channel effect control needs to be the same. The higher on-currents of the Low-V_{th} design point out that in addition to meeting off-current limits, the design needs to have a depletion depth that strikes a balance between being large enough to yield a steep subthreshold slope, yet shallow enough to have controlled short channel effects. The impact of both doping and substrate bias on the threshold voltage and depletion depth mean that both parameters will be important in achieving an optimal design.

Chapter 5
Optimal Device Design

The analysis of the two different device designs in section 4.2 left two key issues unresolved. Although the Low-V_{th} design has higher currents than the High-V_{th} design (Figure 4-5) it is not necessarily the most optimal design. Also, although the lower threshold voltages correlate with higher on-currents at a given L_{eff}, the difference in threshold voltage does not seem to be enough to account for the on-current difference.

The focus of this chapter is to explore these issues with calibrated simulations of optimal designs. Simulating a range of different designs can help identify the optimal design, but the results will only be relevant if the simulator can match measured data. Using an inverse modeling approach, a 2-D numerical simulator (MEDICI [68]) will be calibrated to the 80 nm Low-V_{th} design device, allowing an accurate exploration of the optimal design space at this channel length.

The end goal will be to compare the performance of optimized designs at different temperatures to gauge the performance increase that lower temperatures can give.

5.1 Inverse Modeling

Inverse modeling the Low-V_{th} design devices refers to the process of matching the output of a 2-D numerical simulator [68] to device data and extracting the oxide thickness, poly doping, and doping profiles of the device. Given this information, the simulator's mobility model and velocity saturation model can then be fit to the device data. The simplification of the doping that allows this process to proceed in a relatively short time is the assumption that the doping profiles of the device (vertical retrograde,

source/drain, and halos) can be described by two-dimensional Gaussian distributions. This allows the number of parameters that have to be optimized to be shrunk to a reasonable number yet allows excellent fits between the measured and simulated current-voltage characteristics [69].

A key part of the inverse modeling technique is to extract the different components of the device structure and transport models in cases where only a subset of the device parameters have an impact. In practice, this means using a combination of long channel and short channel data, both I-V and C-V, and carefully choosing the operating regimes in which to fit different parameters.

The oxide thickness (t_{ox}) and polysilicon gate doping (N_{poly}), respectively, are extracted by fitting the accumulation and inversion regions of the capacitance voltage curve of a large MOSFET. Using a 20 μm long device ensures that small errors in the gate length do not impact the t_{ox} extraction. The Van Dort model for quantum mechanical effects in the inversion and accumulation layers was used in the channel and source/drain regions [70]. By using only data from strong inversion and strong accumulation, any error in the channel doping (an initial guess) is almost eliminated [71]. The capacitance in the strong accumulation region, which is relatively insensitive to the poly doping, since both channel and poly are in accumulation, is used to fit the oxide thickness. The capacitance in the strong inversion region, where the polysilicon gate begins to deplete, is used to fit the doping in the polysilicon gate. After adding the extracted channel doping, the measured and simulated capacitance voltage characteristics match quite well (Figure 5-1).

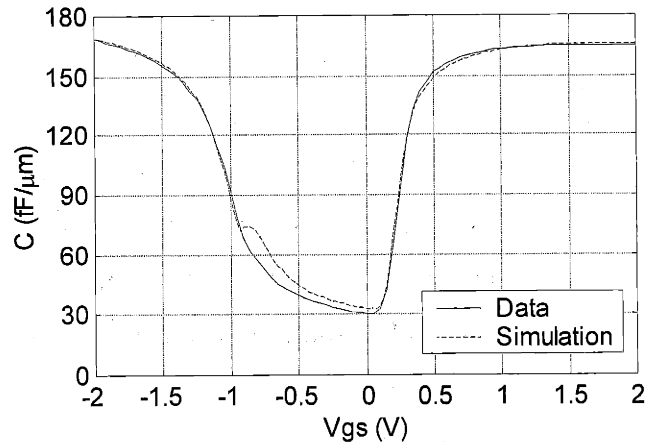


Figure 5-1: Comparison of measured to simulated capacitance voltage characteristic of a L_{gate}=19.85 μm NMOS device. The discrepancies for $0<V_{gs}<-1$ are an artifact of the Van Dort model implementation in MEDICI [72].

Once the poly doping and t_{ox} are pinned down, long channel subthreshold current-voltage (I-V) characteristics vs. V_{bs} are used to extract the 1-D doping profile to capture the retrograde, well, and/or V_{th} implants that are common to all lengths of the design and that dominate the channel doping at long channels. Guesses for the halo doping were used, although they had little effect given the length of the channel.

The subthreshold characteristics have previously been shown to be very sensitive to the two-dimensional potential distribution (and thus doping distribution) in the depletion region of the device [69]. In addition the subthreshold current has been found to be relatively insensitive to errors in mobility or device width [69]. The low currents in subthreshold remove any sensitivity to the external resistance or velocity saturation models, although approximate numbers were used in the models. The use of a range of substrate bias (V_{bs}) and drain bias (V_{ds}) allows the depletion region to average over different depths of the channel, which makes the current-voltage characteristics sensitive to the doping profile vs. depth.

Figure 5-2 shows the match between data and simulated I-V characteristics and Figure 5-3 shows the extracted vertical dopant profile in the channel, with the depletion depth at different V_{bs} superimposed. The sensitivity of this extraction to mobility was

examined and it was found that a couple of iterations between the doping and mobility extractions gave an RMS error in subthreshold of less than 6%.

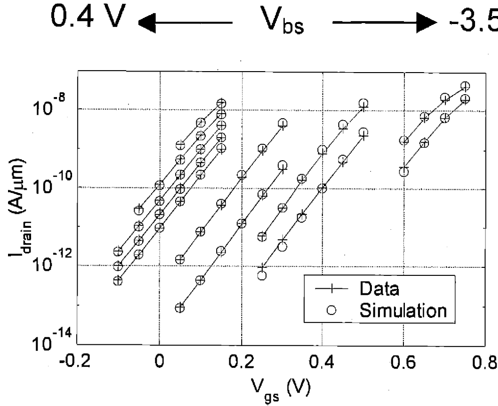


Figure 5-2: Long channel subthreshold data (+ and line) and inverse-modeled simulation (circles) showing match (V_{ds}= 50 mV).

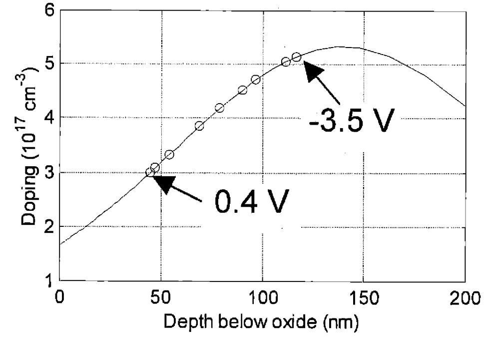


Figure 5-3: Extracted vertical dopant profile of long channel device showing the different depletion depths of different V_{bs}.

Once the doping has been extracted, fitting the mobility can be done with the long-channel device at low V_{ds}. A mobility model that is applied only to the inversion layer and based on the universal mobility dependence was used:

$$\mu_n = \frac{\mu_0}{\left[1 + \left(\frac{E_{eff}}{E_0}\right)\alpha\right]} \tag{5.1}$$

Where μ_0, α, and E_o are fitting parameters.

After initially matching the analytic form of the mobility to the measured mobility by hand, the simulator was able to match the low V_{ds} vs. V_{gs} current with an RMS error of about 2 nA (Figure 5-4). Because it is a long-channel device at low V_{ds} (thus low currents and low parallel electric fields) the details of the external resistance and saturation velocity model do not come into play.

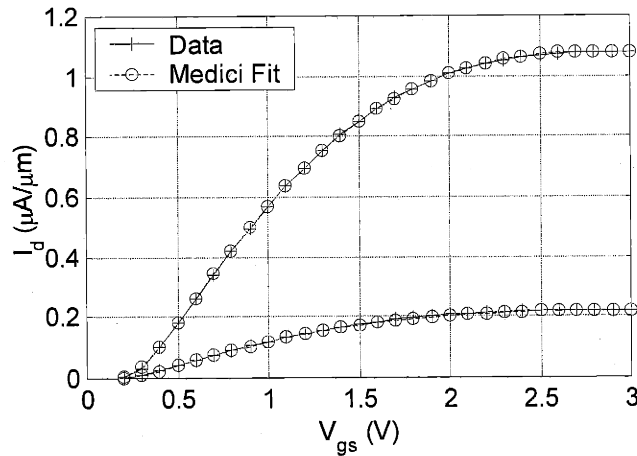


Figure 5-4: Match between low V_{ds} long channel device data and simulation with the calibrated mobility model.

Having determined the background vertical doping profile, t_{ox}, and N_{poly} using the long channel device, the next step is to extract the 2-D doping of the short channel device (L_{eff} = 80 nm). Using the vertical channel doping from the long channel device, the short channel device can be fit by adding a symmetric pair of 2-D Gaussian distributions to represent the halo doping and by another pair of 2-D Gaussian distributions representing the source and drain. Both pairs of Gaussians will be allowed to vary to meet the subthreshold I-V characteristics. The result is similar to Figure 5-2 for the long channel device with I-Vs at various V_{bs} and, in this case, V_{ds}. For the short channel device, the V_{ds} plays a key role in the ability to determine the 2-D nature of the doping. The V_{ds} changes the drain depletion region and the short channel effects of the device. The peak doping level of the source and drain was fit by matching C_{gd} overlap capacitance vs. V_{gs} measurements which are sensitive to S/D doping [73].

Once the doping was extracted, the above threshold I-V characteristics were used to extract the R_{ext} (at low V_{ds}) and velocity saturation model (at medium and high V_{ds}), using the mobility model fit at long channels. Because only the extensions of the source/drain are modeled, matching the total source/drain resistance requires the addition of a fixed external resistor to match the device characteristics. The resulting fit to an I_d-V_d curve at V_{gs}=1.8 V is quite good (Figure 5-5).

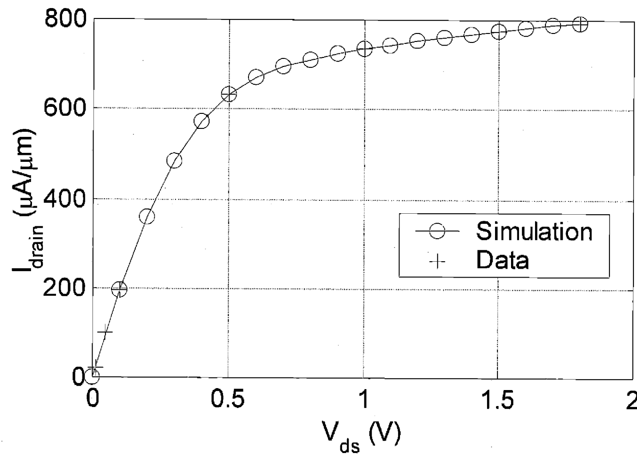


Figure 5-5: I_d-V_d at V_{gs}=1.8V comparing simulation (circles and line) and data (+ sign), showing the calibration of the velocity saturation model.

Pulling all the pieces together, the full I-V characteristics of the device and simulation can be compared (Figure 5-6). The process of calibrating the transport models can be repeated at other temperatures using the same doping profiles, e.g. 200 K as shown in Figure 5-7. The mismatch of the shape of the I_d-V_g characteristics at high V_{ds} results from the inaccuracies of the velocity saturation model used in these drift-diffusion simulations.

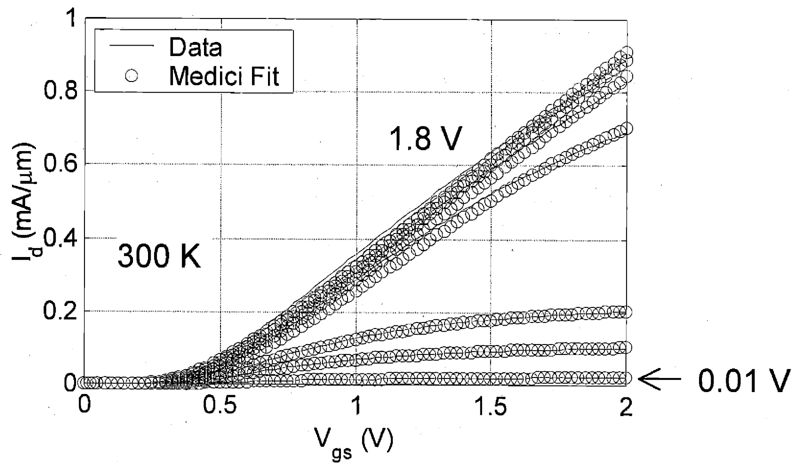


Figure 5-6: I_d-V_g of inverse modeled 80 nm device comparing the inverse modeled simulation results to the data at 300 K. V_{ds} = [0.01, 0.05, 0.1, 0.5, 1.0, 1.5, 1.8 V]

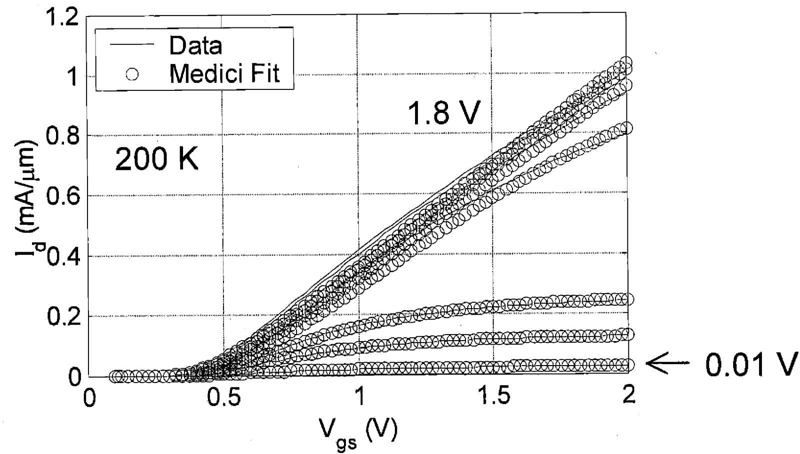


Figure 5-7: I_d-V_g of inverse modeled 80 nm device comparing the inverse modeled simulation results to the data at 200 K. V_{ds} = [0.01, 0.05, 0.1, 0.5, 1.0, 1.5, 1.8 V]

The end result is that MEDICI can model the behavior of the device quite accurately and will allow the simulation results of any perturbation of the device design to come quite close to the results of a fabricated design.

5.2 Device Optimization Setup

The calibrated simulator can now be used to address the issues of what device design gives the highest on-currents for the nominal device and of how well the highest on-currents correlate with lower threshold voltages. Using the same optimizer employed in the inverse modeling, the device design will be optimized to have an off-current of 10^{-9}A/μm and various DIBL levels. Three different types of channel doping profiles will be explored: uniform, halos, and vertically retrograded doping.

As explored in Section 3-3, in order to properly take manufacturing variations in channel length into account when designing the device, the on-currents are evaluated at a nominal channel length, while the off-currents are evaluated at around a 20% smaller channel length. For these set of simulations, the worst-case device ($0.8L_{nom}$) will have an L_{eff} =75 nm, while the nominal device (L_{nom}) will have an L_{eff} =90 nm. The characteristics of the nominal device will be simulated by using the same profiles and

substrate bias as the $0.8L_{nom}$ device but with the S/D and halo profiles moved out and the gate length increased.

To simplify the optimization process, individual optimizations to achieve an off-current of 10^{-9} A/μm were performed for a range of fixed DIBL and profile depth points for the worst-case device ($0.8L_{nom}$) and then were evaluated to find the maximum on-current (drain current at $V_{gs}=V_{ds}=1.8V$) for the nominal device This approach is motivated by the key role of the depletion depth in device design (Section 4.3) which suggests that that depth of the retrograde or halo doping will play a key role in setting the device characteristics. Given this approach, the results will be presented for different DIBL design points.

In all cases, the oxide thickness, polysilicon gate doping, source/drain doping, and transport parameters will be used from the 80 nm inverse modeled device. The halos will be represented by a symmetric pair of 2-D Gaussian profiles, while a 1-D vertical Gaussian will be used to represent the retrograde doping. Both the doping and the substrate bias will be allowed to vary to meet the optimization goals of $I_{off}=10^{-9}$ A/μm and fixed DIBL.

Scenarios	Variables	Fixed parameters	Goals
1) Uniform 2) Retrograde 3) Halos	Doping Profile Substrate Bias	t_{ox} & N_{poly} Source/Drain doping L_{gate} and S/D spacing T=200 K, V_{dd}=1.8V Halo/Retrograde depth (range of values)	I_{off}@$0.8L_{nom}$= 10^{-9} A/μm DIBL (range of values)

Table 5-1: Setup for Optimizations

Throughout the chapter, the threshold voltages are reported for the case where V_{ds}=1.8 V. This V_{th} is extracted by finding the linear extrapolated V_{th} at V_{ds}=50 mV and subtracting the drain induced barrier lowering from V_{ds}=50 mV to 1.8 V. This approach is used to provide a threshold voltage that is directly associated with the on-current.

5.3 *Device Optimization Results*

Comparing the results of the optimizations, the Halo designs give the higher on-current for the nominal device at a fixed off-current for the worst case device. Figure 5-8 shows the on-currents for the nominal device versus the DIBL design points for the worst-case device. The depth of the peak doping of the halos for each of the design points is 20 nm, the depth that gave the best results, while the depth of the retrograde peak doping is 50 nm.

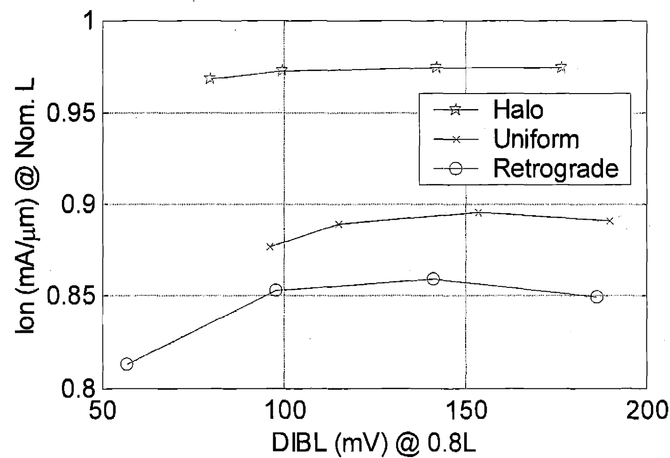


Figure 5-8: On-current of the nominal device versus the worst-case device's DIBL. All device designs have an $I_{off} = 10^{-9}$ A/μm for the worst-case device. T = 200 K.

Plotting the on-current of the nominal devices versus their threshold voltage (at V_{ds}=1.8 V) shows that the higher I_{on} of the halo design correlates strongly with a much lower threshold voltage (Figure 5-9). However, this trend is reversed for the Retrograde and Uniform designs where the Retrograde designs have lower threshold voltages but have less on-current than the Uniform designs. One part of this discrepancy lies in the different amount substrate bias applied to the different designs (Figure 5-10).

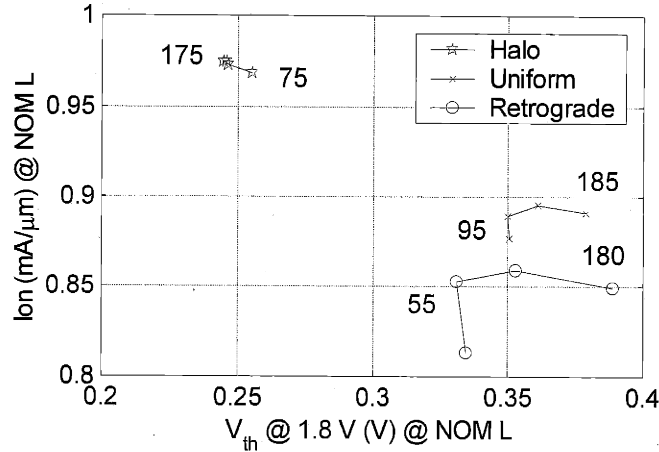


Figure 5-9: On-current of the nominal device versus the threshold voltage (V_{ds} = 1.8V) of the nominal device. The labeled numbers are the worst case device DIBL values (mV) for the end points of each line.

Applying a positive substrate bias can reduce the vertical electric field and thus increase the mobility at a given gate voltage. Figure 5-11 shows the significant impact of a positive and a negative substrate bias on the mobility of a the 20 μm Low-V_{th} design device, plotted against V_{gs}-V_{th}. The Uniform designs, with the largest substrate biases, probably have a highest mobility of any of the designs, thus at the same V_{th} they have a higher on-current. Since the Uniform designs have higher on-currents than the Retrograde designs at both the worst-case and nominal channel lengths, it is likely that the mobility difference plays a role at both channel lengths.

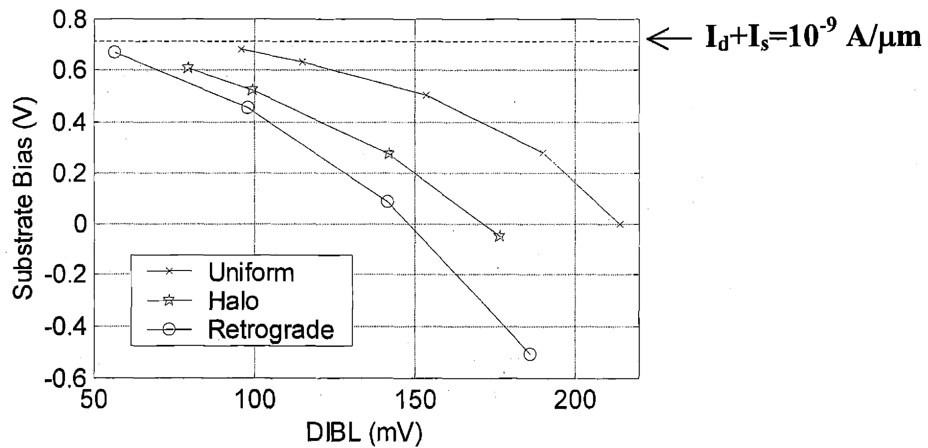


Figure 5-10: Substrate bias used for each of the designs, versus the worst-case device DIBL. All values are less than the measured bias for 10^{-9} A/μm junction leakage current.

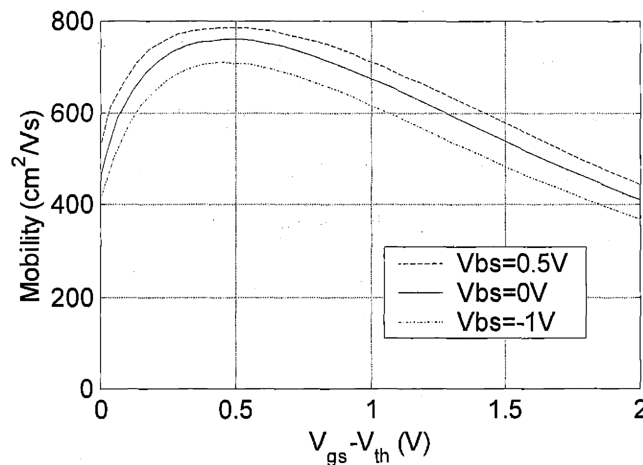


Figure 5-11: Measured electron mobility versus gate overdrive (V_{gs}-V_{th}) at 200 K. Applying a forward bias V_{bs} decreases the E_{eff} at a given gate overdrive, thus increasing the mobility.

Although the specific mobility model used in Medici (Unimob) for these simulations does not account for degradation of mobility from impurity scattering, the same trends for on-current versus design were observed when using an (uncalibrated) model that has been found to over-penalize the mobility due to impurity scattering (Lombardi) [74].

Looking at the substrate biases used for the different designs (Figure 5-10), each type of design has a point where no substrate bias is used (V_{bs}=0) yet the off-current criteria is met. The power of using a substrate bias is that the design can be shifted from this one fixed DIBL point and by re-optimizing the doping, again meet the off-current limit.

The worst case devices also demonstrate that other factors besides the threshold voltage and mobility are influencing the on-current. Figure 5-12 shows the on-current of the worst case devices versus their threshold voltage (at V_{ds}=1.8 V). The order of the currents has changed, with the Halo designs having lower currents than the Uniform designs and with the Retrograde designs still the lowest. Correlating with this shift, the relative magnitudes of their body effects have changed from the nominal case (Figure 5-13 A&B).

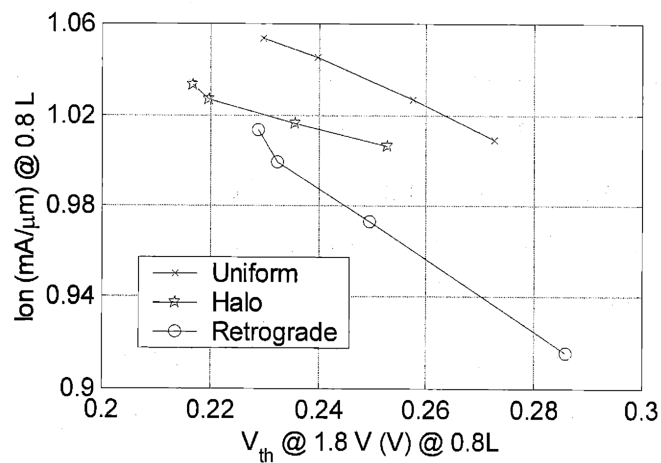


Figure 5-12: On-current of the worst-case device versus V_{th} (V_{ds} = 1.8 V) of the worst-case device.

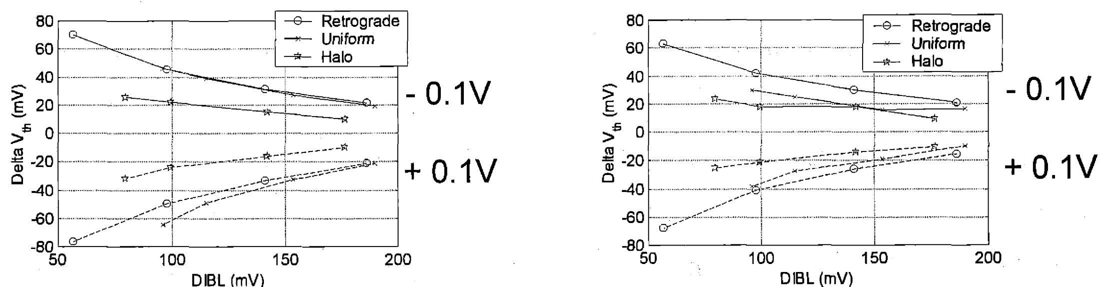


Figure 5-13:
A: Shift in the nominal device threshold voltage (V_{ds} = 1.8 V) with ± 0.1 V V_{bs} from each design set-point versus the DIBL of the worst-case device.
B: Shift in the worst-case device threshold voltage (V_{ds} = 1.8 V) with ± 0.1 V V_{bs} from each design set-point versus the DIBL of the worst-case device.

The different shifts in device threshold voltage with the application of ± 0.1 V V_{bs} from the design's set point shows the impact of the different body effects of the three designs (Figure 5-13 A&B). The uniform designs at the worst case device have lower body effect than at the nominal device, while the halo and retrograde designs' body effect remains relatively constant for both devices. This correlates with the shift of the uniform design on-currents relative to the halo design such that the halo design falls between the uniform and retrograde designs for the worst case device, but is higher than both for the nominal device. It has previously been reported that a larger body effect can degrade the on-current of the device by decreasing the drain voltage at which the device saturates [75].

Even though the mobility and body effect of a device play a role in setting the on-current of the device, the reason the halo designs have a much higher on-current at the nominal design is primarily due to their much lower threshold voltages (Figure 5-14). The very low threshold voltage of the nominal device for the Halo designs is due to the small threshold voltage reduction, due to SCE, from the nominal to worst-case devices that is visible in Figure 5-14.

This behavior versus channel length is due to the lateral non-uniformity of the channel doping. The halos place the majority of their doping close to the source and

drain, reducing the impact of the source and drain on the channel and controlling short channel effects. At the same time they create a lower doping in the center of the channel, reducing the overall device threshold voltage from the case of a uniform doping at the level of the halos. In addition the halos themselves are slightly retrograded, which can help reduce the threshold voltage even further while not impacting the short channel effects of the device [76]. The halos also create, at the same DIBL, a steeper subthreshold slope, suggesting a deeper depletion depth (Figure 5-15).

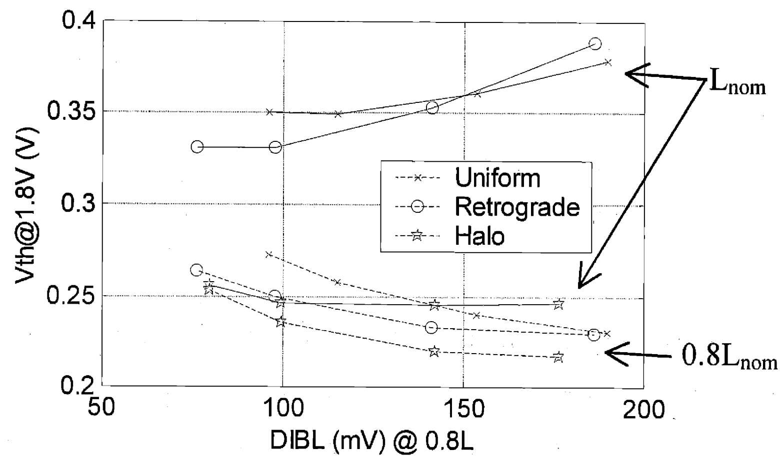


Figure 5-14: Threshold voltage (V_{ds} = 1.8 V) of both the nominal and worst-case devices versus the worst-case device DIBL. The Halo designs show the least amount of V_{th} roll-off from the nominal to the worst-case channel length.

Larger DIBL correlates with an increased threshold voltage change with channel length as is visible for the uniform and retrograde designs in Figure 5-14. The resulting higher threshold voltage of the nominal device at higher DIBL results in a nominal device on-current that decreases at very high DIBL, as visible for the uniform and retrograde designs in Figure 5-8.

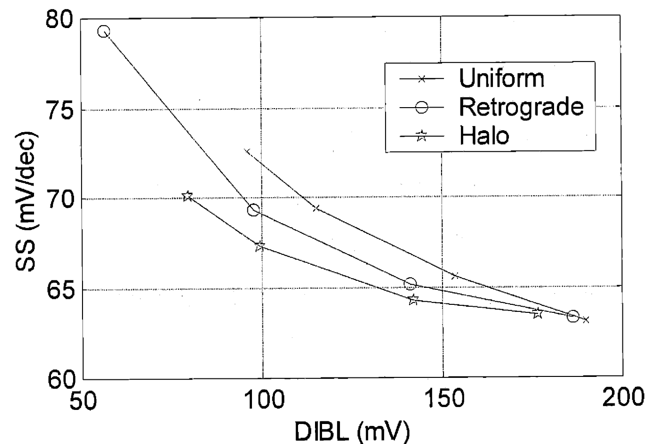


Figure 5-15: Subthreshold slope (V_{ds} = 1.8 V) of the worst-case device versus the DIBL of the worst-case device. The steeper SS of the Halo designs yield the lowest V_{th}s, given the fixed $I_{off} = 10^{-9}$ A/μm for all designs.

Although separating the three doping design types allows a clear comparison of different doping shapes, a combination of the halo and retrograde designs are often used in practice. Figure 5-16 adds the results of three different combinations of halo and retrograde doping to the previous results in Figure 5-9. As before, all of the designs have an $I_{off}=10^{-9}$ A/μm for the worst case device. The addition of a broad retrograde (large σ, depth=50 nm – squares), steep retrograde (small σ, depth=50 nm – diamonds), or well implant (large σ, depth=140 nm - triangles) to the halos (depth=20 nm) shifts the nominal device on-currents down from the halo design results towards the uniform and retrograde results. Designing the device with only halo doping continues to give the highest on-current for the nominal device.

Using only halo doping allows the least change in threshold voltage with channel length, giving the lowest threshold voltage and highest on-current at the nominal channel length. This concept has previously been noted for a simulated 25 nm device design that uses only halos [76]. The lateral non-uniformity of the halo doping causes an increase in the threshold voltage as the channel length decreases, such that at short channel lengths it helps slow the onset of threshold reduction due to short channel effects [11]. This approach works well if only short channel lengths are to be used, but if much longer channel lengths are also needed, then the addition of a retrograde doping would probably

be required to keep the threshold voltage at a high enough value for longer channel lengths.

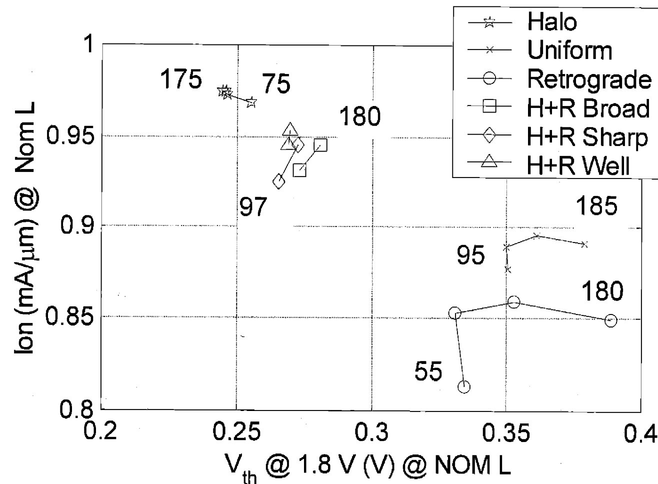


Figure 5-16: On-current of the nominal device versus the threshold voltage (V_{ds} = 1.8V) of the nominal device. The labeled numbers are the worst case device DIBL values (mV) for the end points of each line.

5.4 Optimal Designs Across Temperature

Having identified the halo doping as the optimal design approach, the same process of finding an optimal design can be repeated at 300 K and 100 K. As was done for the 200 K case, these optimal designs points have the maximum I_{on} at the nominal channel length, having met $I_{off}=10^{-9}$ A/μm at the worst case channel length. The mobility and saturation velocity models for each temperature were calibrated using the inverse modeled devices. As was found at 200 K, the optimal halo designs were found to occur at the largest DIBL.

5.4.1 Device Performance

The results in Figure 5-17 show that with optimized designs, the device current can be increased about 3.5% for every 10 K (10 °C) that temperature is decreased. This

translates into a 35% gain with a shift from 300 K to 200 K and a 70% gain from 300 K to 100 K.

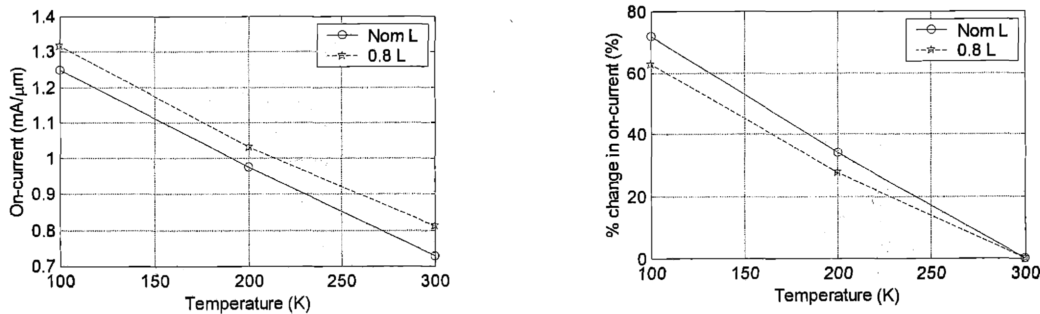


Figure 5-17: A: On-current of the optimized devices at their operating temperatures. B: Percentage increase in on-current from the 300 K optimized design to lower temperature optimized designs.

At the same time the threshold voltage of these devices has been reduced significantly (Figure 5-18). A reduction in threshold voltage causes an increase in the inversion charge density at a given gate bias. The inversion charge is most easily found by integrating a long-channel gate to source/drain capacitance-voltage curve. Since the threshold voltages for the long and short devices are different, the inversion charge at the source for the short device can be found by plotting the long channel C-V versus V_{gs}-V_{th} and then integrating up to the short channel V_{gs}-V_{th}. Figure 5-18 shows the percentage increase in inversion charge density (Q_i) at V_{gs}=1.8V versus temperature for the optimized nominal devices.

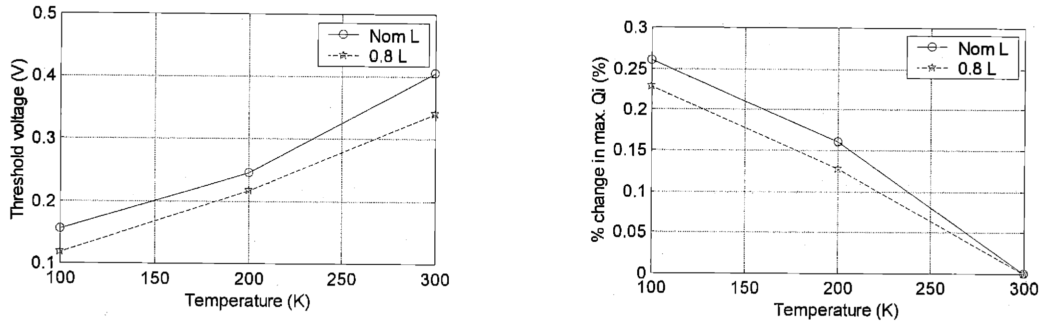


Figure 5-18: Threshold voltage (V_{ds} = 1.8 V) of the nominal and worst-case devices for each of the optimized designs at their operating temperature.

Figure 5-19 compares the percentage increase in current of the optimal device at each temperature from the 300 K optimal device with the percentage increase in measured velocity gains of the 80 nm low-V_{th} design (see section 2.3.2). Returning to the fundamental concept of current as a product of the change and carrier velocity, the increases in velocity are indicated by the increase in g_m/WC_{ox}. The difference between the velocity gains and the on-current gains is due to the increased charge from the lower threshold voltage allowed by the steeper subthreshold slope. Both the increased charge and velocity can help to maintain the performance increases associated with scaling.

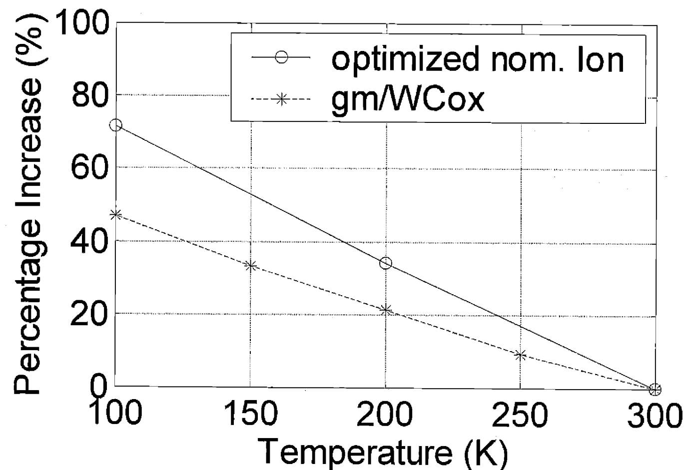


Figure 5-19: Percentage change in optimized-nominal-device on-current from the 300 K design to lower temperature designs, as compared to the % gain in measured velocity of carriers in the 80 nm Low-V_{th} design device.

5.4.2 Device Designs

The doping profiles in these optimal devices in general become steeper as the temperature is decreased. Figure 5-20 shows the 2-D profiles and the 1-D lateral cuts of the surface doping for the optimal designs at each temperature. The peak doping increases slightly as temperature decreases, from $1.1\text{x}10^{18}$ at 300K to $2.5\text{x}10^{18}$ at 100K. In addition the lateral sigmas of the Gaussian distribution decrease as is visible in the increase in the number of lateral contours (Figure 5-20). The optimal peak depth in all cases was 20 nm.

The large difference in the vertical sigma of the 200 K design as compared to either the 100 K or 300 K design suggest with such a variety of parameters to adjust, the halo doping does not have one unique solution.

Looking at the lateral cut of the doping at the surface (Figure 5-20) the trend towards stepper halos laterally is clearly visible. This shape exploits the ability of the halo to reduce short channel effects due to its higher doping near the source and drain, yet maintain a low V_{th} due to the low doping in the center of the channel. Overall, the doping profiles are quite broad and show only minimal retrograding for the 300 K and 200 K designs. This behavior is consistent with the optimization results which showed that the retrograded profiles gave consistently lower on-currents than the uniformly doped profiles.

The substrate biases for the 300 K and 200 K designs (0 and –0.04 V) are near zero, while for the 100K design, it is 0.64 V.

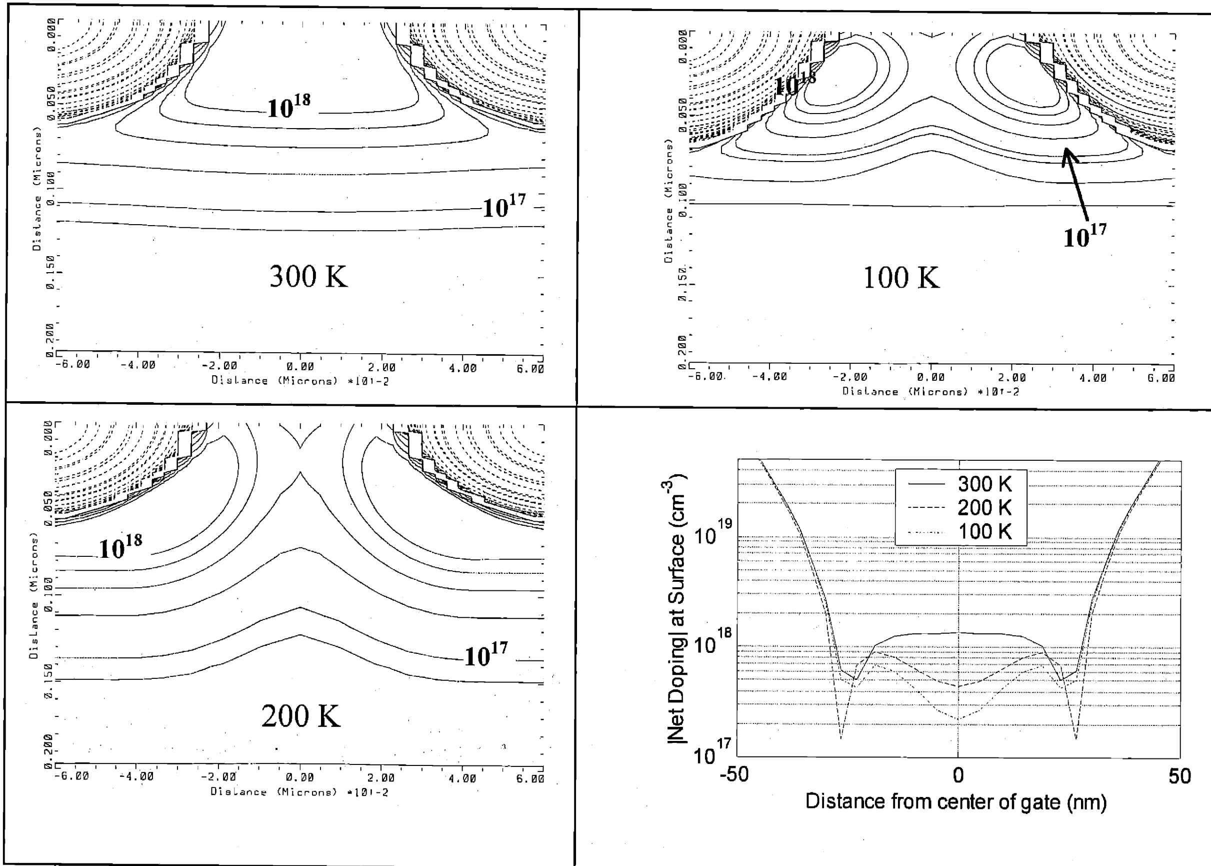


Figure 5-20: Worst-case device net-doping profiles of the different optimal designs at different temperatures. (Upper left: 300K, Upper right: 100K, Lower left: 200K, Lower right:1-D profile of absolute value of net doping at the oxide-silicon interface) For the 2-D profiles, p-type doping has solid lines, n-type has dashed lines.

5.4.3 Room Temperature Comparison

Comparing the devices designed for three different temperatures all at 300 K with no substrate bias applied shows the shift towards lower V_{th} designs for lower temperature operation (Table 5-2). Also, the DIBL of these devices at 300 K and V_{bs}=0 is much higher than the 300 K design since the decrease in temperature and increase in forward substrate bias reduces the DIBL at their design points.

Design	V_{th} @ T	V_{th} @ 300 K	DIBL @ T	DIBL @ 300K
300 K	0.34 V	0.34 V	205 mV	205 mV
200 K	0.22 V	0.13 V	177 mV	192 mV
100 K	0.12 V	-0.06 V	178 mV	292 mV

Table 5-2: Comparison of room temperature to operating temperature device parameters for the worst case ($0.8L_{nom}$) device for the different designs. V_{th} and DIBL are at V_{ds}=1.8 V.

5.5 Conclusion

The increasing currents of the optimal nominal devices as temperature is decreased directly translates into a faster switching speed as defined by a I/CV metric. Given the identical power supply voltage, gate length, source/drain doping, oxide thickness and poly doping of each of the designs, the capacitance times voltage (CV) term should be quite similar versus temperature.

In addition to the large gains in on-current that can be achieved by designing a device for operating temperatures 100 degrees lower, the results suggest that even small changes in the operating point could give measurable performance increases with a re-designed device. Decreasing the operating temperature of a 90 nm device by 10 °C yields a 3.5% gain in the on-current, and thus switching speed, for a re-optimized device. The 70% performance gain from 300 K to 77 K is a similar to the 50-75% gains literature data suggest [44,46].

Across temperature, the Halo design gives the highest on-currents at the nominal device due to the combination of small threshold roll-off with channel length variation and low body factor. In all cases, the higher DIBL designs give higher currents, although too high a DIBL correlates with a larger threshold roll-off, pushing the nominal V_{th} higher and reducing the nominal I_{on}.

The change in device design as temperature is lowered is towards steeper halos and higher DIBL at room temperature (same at operating temperature). Optimal designs with reasonable DIBL and no substrate bias are possible down to 200K, but at lower temperatures, a positive substrate bias is necessary.

Chapter 6
Conclusion

The results of the data analysis and calibrated simulations in chapters 3, 4, and 5 allow the questions posed in the introduction to be answered.

An optimal design at any temperature meets the off-current limit for the worst-case device, while having the maximum on-current for the nominal device. Having a low threshold voltage as well as a small body effect are key to achieving an optimal design. In addition higher DIBL designs, within reason, give higher on-currents. Maximizing the on-current is a key step in maximizing the switching speed (I/CV).

Halo doping with the addition of a forward substrate bias for temperatures below 200 K gives the best device performance. The halo profiles become steeper for lower temperature designs.

Comparing optimal designs for a 90 nm nominal device across temperature, on-current gains, and thus switching speed gains of 3.5% for a 10 °C decrease in temperature can be achieved. Table 6-1 summarizes the improvements in device characteristics of a L_{eff} =80 nm N-MOSFET that decreasing the operating temperature by 10 °C and re-optimizing the device design, to maintain an $I_{off}=10^{-9}$ A/μm, can bring. These gains clearly show that lowering temperature is a valuable tool to help maintain the expected performance gains with device scaling.

On-Current	+ 3.5 %
Inversion Charge	+ 2.4 %
Carrier Velocity (near source)	+ 1.25 %
Source & Drain Resistance	- 1-2 %
Interconnect Resistance (Al)	- 2.5 %

Table 6-1: Summary of performance improvements at L_{eff} =80 nm N-MOSFET with a 10 °C decrease in temperature.

6.1 Contributions

The consistent approach of this thesis has been to use comparisons of optimal designs across channel lengths and across temperatures to accurately assess the performance increases and increased design flexibility that come with lowering the operating temperature of the device. Optimizations of device parameters in analytical equations and 2-D doping profiles in a numerical simulator were used to identify the optimal designs.

The impact of lowering the operating temperature on MOSFET device operation was explored by combining data from the literature and a variety of measurements of long and short devices from a set of 0.18 μm generation NMOS devices.

Using a performance metric related to inverter switching speed (I/CV), the impact of off-current limits on performance gains with length scaling at constant temperature were examined. The tradeoff between fully scaled performance and maintaining reasonable off-current levels was clearly shown. As an alternative to allowing off-currents to rise, two possible temperature scaling scenarios, with the goal of meeting or exceeding the fully scaled performance at constant temperature, were explored. The use of analytical equations based on optimized device parameters allowed the performance estimates to be consistent with the constraints of each scaling scenario.

Data from two different device designs showed that the lower-V_{th} design yielded higher on-currents at 200 K when the designs were compared at the same channel length and off-current. The role of depletion depth in the device performance and the impact of substrate bias and doping on the depletion depth were explained.

Using a 2-D numerical simulator the doping profiles of the Low-V_{th} NMOS devices were inverse modeled and the transport models of the simulator were calibrated at a range of temperatures. Focusing on a nominal channel length of 90 nm (worst-case of 75 nm) a detailed analysis of channel doping profile design to achieve the highest on-current at the nominal channel length, while meeting the off-current limit for the worst-case channel length was performed. The halo doping consistently gave higher on-currents than the either of the retrograde or uniform doping designs. This optimal

performance was linked to the lower threshold voltage, and lower threshold voltage decrease with channel length, and lower body effect achievable with the halo designs. As the operating temperature is reduced, the halo designs showed steeper lateral doping profiles and a forward substrate bias was needed below 200 K.

6.2 Suggestions for Future Work

Measuring the performance of devices fabricated with the simulated optimal designs would close the loop on the performance analysis. In addition, evaluating the performance gains of ring oscillators with separate biases for the NMOS and PMOS would provide data to confirm the performance estimates of the I/CV metric used in this thesis.

An interesting additional low temperature scaling scenario to explore would be the case where the substrate of a device is biased by tying it to the opposite power supply (NMOS substrate tied to positive V_{dd}, PMOS substrate tied to ground) [77]. This could work well for a future device generation with a power supply near 0.5 V and would eliminate the need to generate a separate substrate bias. In this case in would be interesting to explore the design approaches for dual-V_{th} circuits, which could involve either substrate bias or doping changes.

The use of lower temperatures to improve the inverse modeling of the device could be interesting to explore. The ability to go to larger forward biases at lower temperatures could help better extract the doping near the surface of the MOSFET. In addition, changing the temperature, like changing the substrate or drain biases, could provide more data about the doping profiles by providing another operating condition for the profiles to fit.

References

[1] M. Riordan and L. Hoddeson. *Crystal Fire: The Invention of the Transistor and the Birth of the Information Age.* 1997.

[2] Moore, G.E. "Cramming More Components onto Integrated Circuits." *Electronics* , April 19, 1965. pp. 114-117.

[3] Y. A. El-Mansy, "VLSI Symposium and Silicon Technology: A Twenty Year Perspective." *2000 Symposium on VLSI Technology Digest of Technical Papers*, 2000.

[4] R. Chau, et al. "30nm Physical Gate Length CMOS Transistors with 1.0 ps n-MOS and 1.7 ps p-MOS Gate Delays." *International Electron Devices Meeting Technical Digest*, 2000.

[5] L. Su et al. "A High-Performance 0.08 μm CMOS." *1996 Symposium on VLSI Technology Digest of Technical Papers*, 1996, pp. 12-13.

[6] G. SH. Gildenblat, "Low-Temperature CMOS" in *VLSI Electronics : Microstructure Science, Vol 18*, Academic Press, 1989, pp. 191-236.

[7] S.M. Sze, *Physics of Semiconductor Devices*, second edition. Wiley, New York, 1981.

[8] S.M. Sze, *Semiconductor Devices: Physics and Technology.* Wiley, New York, 1985.

[9] Jason C. Woo and James D. Plummer. "Short-Channel Effects in MOSFET's at Liquid-Nitrogen Temperature." *IEEE Transactions on Electron Devices*, Vol. ED-33, No. 7, July 1986, pp. 1012-1019.

[10] John A. Scott. *MOS Device Design For Reduced Temperature Operation.* EECS Ph.D. Thesis, Stanford University Technical Report No. ICL 96-060. 1996.

[11] Keith M. Jackson. *Laterally Non-Uniform Doping Profiles in MOSFETs: Modeling and Analysis.* EECS Master Thesis, Massachusetts Institute of Technology, 1996.

[12] B. Szelag, F. Balestra, and G. Ghibaudo. "Comprehensive Analysis of Reverse Short-Channel Effect in Silicon MOSFET's from Low-Temperature Operation." *IEEE Electron Device Letters*, Vol. 19, No. 12, December 1998, pp. 511-513.

[13] Y. Taur and T.H. Ning. *Fundamentals of Modern VLSI Devices.* Cambridge University Press, USA, 1998. p. 22.

[14] P. P. Debye and E. M. Conwell, "Electrical properties of n-type germanium." *Physical Review*, Vo. 93, No. 4, 1954, pp. 693-706.

[15] S. Takagi, A. Toriumi, M. Iwase, H. Tango. "On the Universality of Inversion Layer Mobility in Si MOSFET's: Part I – Effects of Substrate Impurity Concentration." *IEEE Transactions on Electron Devices*, Vol. 41, No. 12, December 1994, pp. 2357-2362.

[16] B. Cheng and J. Woo. "A Temperature Dependent MOSFET Inversion Layer Carrier Mobility Model for Device and Circuit Simulation." *IEEE Transactions on Electron Devices,* Vol.44, No. 2, February 1997, pp. 343-345..

[17] N. D. Arora and G. SH. Gildenblat. "A Semi-Empirical Model of the MOSFET Inversion Layer Mobility for Low-Temperature Operation." *IEEE Transactions on Electron Devices*, Vol. ED-34, No. 1, January 1987, pp. 89-93.

[18] Taur, *Fundamentals*, p. 20.

[19] A.G. Sabins and J.T. Clemens, "Characterization of the electron mobility in the weakly inverted <100>silicon surface," *International Electron Devices Meeting Technical Digest*, 1979, p.18.

[20] D.M. Caughey and R.F. Thomas, "Carrier Mobilities in Silicon Empirically Related to Doping and Field," *Proceedings of the IEEE*, Vol. 55,1957, pp.2192-2193.

[21] H. Hu et al. "Channel and Source/Drain Engineering in High-performance sub-0.1 μm NMOSFETs Using Xray Lithography", *1994 Symposium on VLSI Technology*, pp. 17-18.

[22] C. Kittel, *Introduction to Solid State Physics*, 6th edition, Wiley, N.Y, 1986, pp. 268-270.

[23] F. H. Gaensslen, V. L. Rideout, E. J. Walker, and J. J. Walker. "Very Small MOSFET's for Low-Temperature Operation." *IEEE Transactions on Electron Devices*, Vol. ED-24, No. 3, March 1977, pp. 218-229.

[24] D. P. Foty, "Impurity Ionization in MOSFETs at very low temperatures." *Cryogenics*, Vol 30, December 1990, pp. 1056-1063.

[25] A. Pirovano, A. Lacaita, A. Pacelli, A. Benvenuti. "Novel Low-Temperature C-V Technique for MOS Doping Profile Determination Near the Si/Si02 Interface." *IEEE Transactions on Electron Devices*, Vol. 48, No. 4, April 2001, pp.750 –757.

[26] A. Emrani, G. Ghibaudo, and F. Balestra. "Generalised Method for Extraction of MOSFET Source-Drain Series Resistance Against Temperature." *Electronics Letters*, Vol. 29, No, 9, April 29, 1993, pp. 786-788.

[27] J. L. Hill and R. L. Anderson. "The Extraction of Electrical Parameters for MOSFET's with Applications to Low Temperature." *Proceedings of the Workshop on Low Temperature Semiconductor Electronics*, August 1989, pp. 58-62.

[28] C.-L. Huang and G. SH. Gildenblat. "*An Accurate Engineering Model of an n-Channel MOSFET for 60-300 K Temperature Range*." Solid-State Electronics, Vol. 33, No. 10, 1990, pp. 1309-1318.

[29] W. F. Clark, et al. "Low Temperature CMOS – A Brief Review." *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, Vol. 15, No. 3, June 1992, pp. 397-404.

[30] K.W. Ng and W.T. Lynch, "Analysis of the Gate-Voltage Dependent Series Resistance of MOSFET's." *IEEE Transactions on Electron Devices*, Vol. ED-33, No. 7, July 1986, pp. 965-972.

[31] S.E. Swirhun and R.M. Swanson, "Temperature Dependence of Specific Contact Resistivity." *IEEE Electron Device Letters,* Vol. EDL-7, No. 3, March1986, p.155-7.

[32] L. Krusin-Elbaum, J. Y.-C. Sun, and C.-Y. Ting, "On the Resistivity of Ti-Si_2: The Implication for Low Temperature Applications." *IEEE Transactions on Electron Devices*, Vol. ED-34, No. 1, January 1987, p 58-63.

[33] Mark T. Bohr, "Interconnect Scaling – The Real Limiter to High Performance ULSI", *International Electron Devices Meeting Technical Digest*, 1995, p.241-244.

[34] Jeffery T. Watt and James D. Plummer. "The Effect of Interconnection Resistance on the Performance Enhancement of Liquid-Nitrogen-Cooled CMOS Circuits." *IEEE Transactions on Electron Devices,* Vol. 36, No. 8, August 1989, pp 1510-1520.

[35] G.A. Sai-Halasz, "Performance Trends in High-End Processors." *Proceedings of the IEEE*, Vol. 83, 1995, p. 20.

[36] I. Aller, et al. "CMOS Circuit Technology for Sub-Ambient Temperature Operation." *ISSCC Digest of Technical Papers*, 2000, pp. 214-214, pp. 168-169,441.

[37] F. Fischer and F. Neppl. "Sputtered Ti-doped Al-Si for Enhanced Interconnect Reliability." *Proceedings of the International Reliability Physics Symposium*, 1984, p. 190-192.

[38] F.H. Gaensslen, V.L. Rideout, E.J. Walker, and J.J. Walker. "Very Small MOSFET's for Low-Temperature Operation." *IEEE Transactions on Electron Devices*, Vol. ED-24, No. 3, March 1977, pp. 218-229.

[39] J.Y. Sun, Y. Taur, R.H. Dennard, S.P. Klepner. "Submicrometer-Channel CMOS for Low-Temperature Operation." *IEEE Transactions on Electron Devices*, Vol. ED-34, No. 1, January 1987, pp. 19-26.

[40] G. Baccarani, M.R. Wordeman, R.H. Dennard. "Generalized Scaling Theory and Its Application to a ¼ Micrometer MOSFET Design." *IEEE Transactions on Electron Devices*, Vol ED-31, No. 4, April 1984, pp. 452-462.

[41] J. Koga, M. Takahashi, H. Naiiyama, M. Iwase, M. Fujisaki, and A. Toriumi. "0.25 μm Gate Length CMOS Devices for Cryogenic Operation." *IEEE Transactions on Electron Devices*, Vol. 41, No. 7, July 1994, pp. 1179-1183.

[42] M. Kakumu, D. Peters, H.-Y. Liu, and K.-Y. Chiu. "Design Optimization Methodology for Deep-Submicrometer CMOS Device at Low-Temperature Operation." *IEEE Transactions on Electron Devices*, Vol. 39, No. 2, February 1992. pp. 370-377.

[43] J. Xu and M.-C. Cheng "Design Optimization of High Performance Low-Temperature 0.18 μm MOSFET's with Low-Impurity-Density Channels at Supply Voltage Below 1V." *IEEE Transactions on Electron Devices*, Vol. 47, No. 4, April 2000, pp. 813-821.

[44] G.A. Sai-Halaz, M.R. Wordeman, D.P. Kern, S.A. Rishton, E. Ganin, T.H.P. Chang, and R.H. Dennard. "Experimental technology and performance of 0.1-μm-gate-length FETs operated at liquid-nitrogen temperature." *IBM Journal of Research and Development*. Vol. 34, No. 4, July 1990, pp. 452-464.

[45] S.J. Wind, et al. "Very High Performance 50 nm CMOS at Low Temperature." *International Electron Devices Meeting Technical Digest*, 1999, pp. 928-930.

[46] Taur, *Fundamentals*, p. 288.

[47] Y.-W. Yi, K. Masu, K. Tsubouchi, and N. Mikoshiba. "Temperature-Scaling Theory or Low-Temperature-Operated MOSFET with Deep-Submicron Channel." *Japanese Journal of Applied Physics*, Vol. 27, No. 10., October 1988, pp. L1958-L1961.

[48] M. Yokoyama, T. Hidaka, Y.-W. Yi, K. Maus, and K. Tsubouchi. "Low Temperature MOSFET Operation by the Temperature Scaling Theory." *Extended Abstracts of the 1992 International Conference on Solid Sate Devices and Materials*, 1992, pp. 499-501.

[49] K. Masu, M. Yokoyama, and K. Tsubouchi. "Dimension-Temperature Combination Scaling for Low-Temperature 0.1 μm CMOS." SPIE Vol. 2636, 1995, pp. 62-73.

[50] M. Yokoyama, T. Hidaka, K. Sasaki, K. Masu, and K. Tsubouchi. "Short-Channel-Effect Free 0.18 μm MOSFET by Temperature-Dimension Combination Scaling Theory: Design and Experiment." *IEEE Electron Device Letters*, Vol. 15, No. 6, June 1994, pp. 202-205.

[51] R. H. Dennard et al. "Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions", *IEEE Journal of Solid State Circuits*, Vol. SC-9, No. 5, October 1976, pp. 256-267.

[52] G. Bacarrini et al. "Generalized Scaling Theory and Its Application to a ¼ Micrometer MOSFET Design", *IEEE Transactions on Electron Devices*, Vol. ED-31, No. 4, April 1984, pp. 452-462.

[53] P.K. Chatterjee et al. "The Impact of Scaling Laws on the choice of n-channel or p-channel for MOS VLSI", *IEEE Electron Device Letters*, Vol. EDL-1, No. 10, October 1980, pp. 220-223.

[54] Taur, *Fundamental*s, pp. 263,275.

[55] Semiconductor Industry Association, *The International Technology Roadmap for Semiconductors*, 1997.

[56] J. A. del Alamo et al. "An Analytical Framework for First-Order CMOS Device Design." *1998 International Conference on Characterization and Metrology for ULSI Technology.*

[57] M. Rodder et al. "A 0.10 μm Gate Length CMOS Technology with 30 Å Gate Dielectric for 1.0 V-1.5 V applications." *International Electron Devices Meeting Technical Digest*. 1997, pp. 223-226.

[58] Taur, *Fundamentals*, p. 229.

[59] Taur, *Fundamentals*, p. 263.

[60] Personal Communications with Dr. Mark Armstrong (Intel) and Dr. Melanie Sherony (IBM)

[61] L_{eff} extraction uses Shift and Ratio method: Y. Taur, et al. "A New Shift and Ration Method for MOSFET Channel Length Extraction." *IEEE Electron Device Letters*, Vol. EDL-13, p. 267.

[62] Narain Arora. *MOSFET Models for VLSI Circuit Simulation: Theory and Practice*. Springer-Verlag, 1993, p. 257.

[63] Taur, *Fundamentals*, p. 126.

[64] Taur, *Fundamentals*, p. 70.

[65] Taur, *Fundamentals*, p. 128.

[66] Taur, *Fundamentals*, p. 429.

[67] Taur, *Fundamentals*, p. 427-433.

[68] MEDICI™ 1999.2 , Avant! Corporation.

[69] Z.K. Lee, M.B. MacIlrath, D.A. Antoniadis, "Two-Dimensional Doping Profile Characterization of MOSFET's by Inverse Modeling using I-V Characteristics in the Subthreshold Region." *IEEE Transactions on Electron Devices*, Vol. 46, No. 8, August 1999, pp. 1640-1649.

[70] 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. 27, No. 3, 1994, pp. 411-414. (Default implementation in MEDICI was used)

[71] R. Rios and N. D. Arora. "Determination of Ultra-Thin Gate Oxide Thickness for CMOS Structures using Quantum Effects." *International Electron Devices Meeting Technical Digest*, 1994, pp. 613-616.

[72] MEDICI™ 2000.4 manual, Avant! Corporation.

[73] I. J. Djomehri and D.A. Antoniadis. "Inverse Modeling of Sub-100nm MOSFETs Using log(I)-V and C-V," to be published.

[74] Private Communication between Prof. Dimitri Antoniadis and Dr. Dan Connelly (Motorola).

[75] S. Venkatesan et al. "Device Drive Current Degradation Observed with Retrograde Channel Profiles." *International Electron Devices Meeting Technical Digest*, 1995, pp. 419-422.

[76] Y. Taur, C. H. Wann, D. J. Frank. "25 nm CMOS Design Considerations." *International Electron Devices Meeting Technical Digest*, 1998, pp. 789-792.

[77] D. J. Frank, et al. "Device Scaling Limits of Si MOSFETs and Their Application Dependencies." *Proceedings of the IEEE*, Vol. 89, No. 3, March 2001, pp. 259-288.

Appendix A
Mobility Extraction

A.1 Mobility Extraction

Extraction of the mobility of carriers in a MOSFET inversion layer is accomplished by measuring the inversion charge and current separately and using these in the equation for the MOSFET current. At the highest level, the current flow in a long-channel MOSFET follows the general equation for drift-diffusion current for electrons[1]:

$$J(x) = q\mu(x)N(x)E_x(x) + \left(\frac{kT}{q}\mu(x)\right)q\frac{dN(x)}{dx} \qquad (0.1)$$

where J is the current density (A/cm^2), x is the distance along the channel from source to drain (cm), q is the magnitude of the electron charge (C), $\mu(x)$ is the carrier mobility (cm^2/Vs), N is the carrier density (cm^{-3}), and kT/q is the thermal potential (V). Although not explicitly stated, each of the variables is also a function of the gate voltage. The first term in equation (0.1) is the drift component of the current while the second term is the diffusion component of the current. Integrating equation (0.1) with depth and using the fact that the current (I) is constant both along the length and width of the device:

$$\frac{I}{W} = \mu(x)Q_i'(x)E_x(x) + \left(\frac{kT}{q}\mu(x)\right)\frac{dQ_i'(x)}{dx} \qquad (0.2)$$

where $Q'_i(x)$ is the inversion charge density per area (C/cm^2), I is the current flowing in the device (A), and W is the width of the device (cm). As explicitly stated, each of these variables can be functions of x except for the current which is constant versus x. This equation can then be inverted to solve for $\mu(x)$:

$$\mu(x) = \frac{I}{W\left[E_x(x)Q_i'(x) + \left(\frac{kT}{q}\right)\frac{dQ_i'(x)}{dx}\right]} \tag{0.3}$$

Sodini et al. [2] have shown that for small source to drain voltage (V_{ds}) on a long channel device, that $E_x(x)$ is constant over much of the channel and is equal to:

$$E_x = \frac{V_{ds}}{L}F(V_g) \text{ where } F(V_g) = \frac{C_{inv}}{C_{inv,\max}} \tag{0.4}$$

For the particular L=20 μm device they simulated, this value of E_x was valid starting about 3 μm in from the source or drain. Sodini shows that the $F(V_g)$ empirically gives the correct behavior of the electric field in the channel. The existence of $F(V_g)$ can be motivated by noting that near the threshold voltage, the inversion and bulk charges are of similar magnitude, so that some of the drain voltage will be dropped across bulk charge near the source and drain and not appear across the inversion charge at the center of the channel. Inserting E_x into (0.3) leaves $\mu(x)$ depending only on the x-dependence of the inversion charge ($Q'_i(x)$).

A.2 Measurement

The key variables to measure are the current (I) and the charge (Q'_i). Both quantities are generally measured on a long and wide gate device, where small inaccuracies in length or width will have little impact on the mobility extraction. In addition, a wide device provides a larger capacitance, which is easier to measure. The

current is usually measured with a low V_{ds} (10-50 mV) [3] applied. The measurements in this thesis used V_{ds}=50 mV, a small signal of 25 mV amplitude at 800 kHz, and voltage step size of 50 mV (except for C_{gds} at T < 250K where 25mV was used). The inversion charge (Q'_i) is found by integrating the measured gate to source/drain capacitance (C_{gds}) versus gate voltage (Figure A-1) and dividing it by the area (W*L).[2]

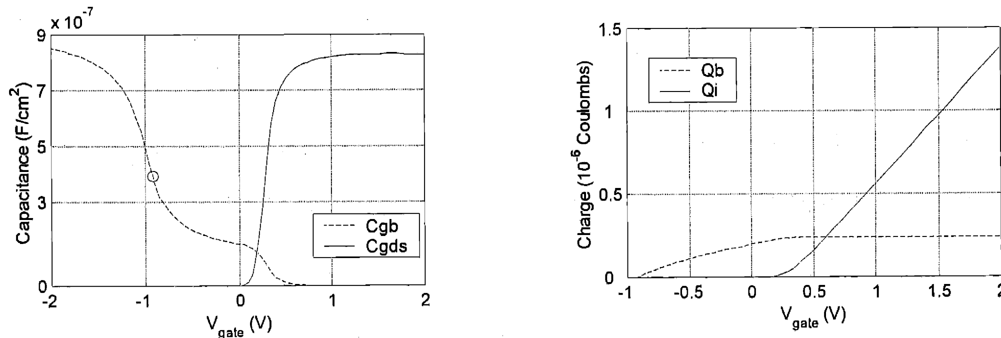
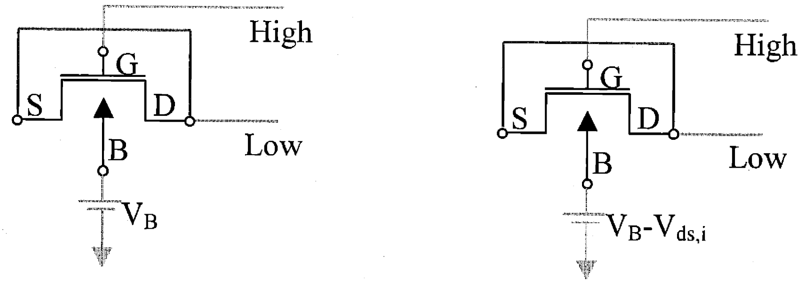


Figure A-1: The C_{gds} measurement (A) is integrated to give the inversion charge (B). The C_{gb} measurement (A) is integrated from the V_{fb} (circle) to give the bulk charge (B).

The inversion layer charge extracted from the C_{gds} in Figure A-1 is the charge for the case where V_d=V_s=0 V, while the current was measured at with V_d>V_s. Applying a source to drain voltage (V_{ds}) will reduce the charge in the channel since V_g-V_d will be less than V_g-V_s. At low V_{ds} and long channels, the charge sheet approximation should hold and the charge will decrease linearly from source to drain [4]. Thus the charge extracted from the measurement with V_d=V_s=0 V is only accurate near the source.

Given the almost linear shape of the charge distribution along the channel, by measuring the charge at the drain, the charge at any point in the channel can be determined. In theory this can be accomplished by repeating the same C_{gds} measurement with V_d=V_s=$V_{ds,i}$ where $V_{ds,i}$ is the source to drain voltage used when measuring the current. Practically, however it is easier to shift all the voltages such that the V_s=V_d=0 V again (Figure A-2). Note that once the measurement is taken, the voltages all have to be shifted back (V_g=V_{High}+$V_{ds,i}$) to achieve the initially desired data for charge at the drain.



A: Standard C_{gds} which yields the charge near the source (V_g=V_{High})

B: Modified C_{gds} which yields the charge near the drain. (V_g=V_{High}+$V_{ds,i}$)

Figure A-2: Measurement setup for Q'$_i$(source) and Q'$_i$(drain). V_B is the substrate bias applied during the measurement of the I-V.

The charge at any point in the channel can then be written as:

$$Q_i^{'}(x) = Q_i^{'}(s) - \left(\frac{Q_i^{'}(s) - Q_i^{'}(d)}{L} \right) x \tag{0.5}$$

where Q'$_i$(s/d) is the inversion charge at the source/drain. The $\frac{dQ_i^{'}(x)}{dx}$ term in equation (0.3) can now be easily evaluated using (0.5). Since the value of E_x in equation (0.4) is only valid away from the source and drain, it is easiest to evaluate the mobility at the center of the channel (x=L/2) where the equation for the mobility is:

$$\mu|_{x=L/2} = \frac{I}{\frac{W}{L}\left[V_{ds}F(V_g) \left(\frac{Q_i^{'}(s) + Q_i^{'}(d)}{2} \right) + \left(\frac{kT}{q} \right) \left(Q_i^{'}(s) - Q_i^{'}(d) \right) \right]} \tag{0.6}$$

In contrast, the standard approach for extracting mobility [3] uses the equation:

$$\mu = \frac{I_d}{\frac{W}{L} V_{ds} Q_i^{'}} \quad (0.7)$$

where $Q'_i = Q'_i(s)$ and is the integral of C_{gds} with $V_d = V_s = 0$ V. This is a similar form to (0.6) but ignores three key issues, which impact the mobility extraction near the threshold voltage of the device (Figure A-3 & Figure A-4). First, eqn. (0.7) ignores the diffusion current (2nd term in the denominator of (0.6)) which will cause the mobility to be slightly over-estimated near V_{th} where the drift current becomes small. Second, (0.7) ignores the impact of V_{ds} on Q'_i which results in an over-estimation of Q'_i and thus an under-estimation of the mobility. Even worse, this under-estimation is V_{ds} dependent (Figure A-5). Finally, (0.7) overestimates the lateral electric field (E_x) by not including the empirical $F(V_g)$ (Figure A-4) .

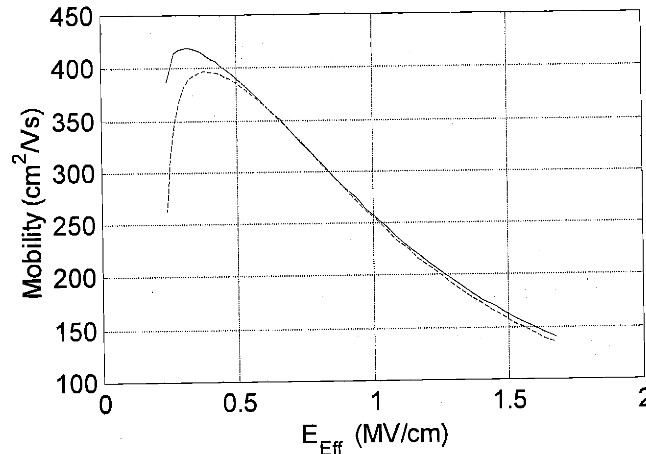


Figure A-3: Comparison of mobility extracted by equations (0.6), solid line, & (0.7), dashed line, at 300K from a L=20 μm NMOS device with V_{ds}=50 mV. Note the significant difference that occurs in the mobility near the threshold voltage of the device.

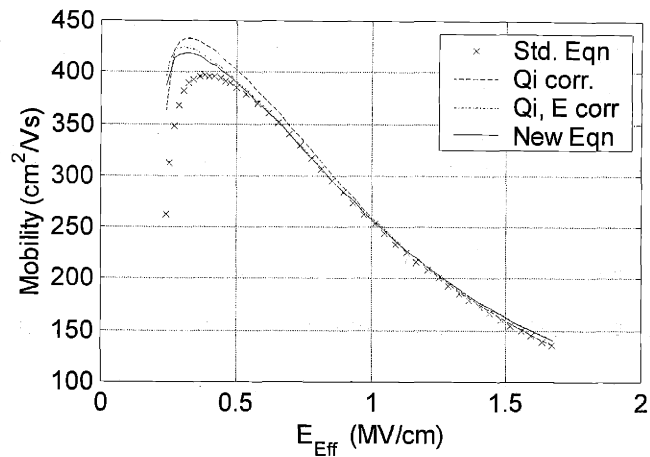


Figure A-4: Illustration of the impact of each of the 3 corrections to equation (0.7) for the same device as Figure A-3. A: eqn. (0.7); B: Q'_i correction; C: Q'_i and E_x correction; D: Q'_i , E_x , and diffusion current correction = eqn. (0.6).

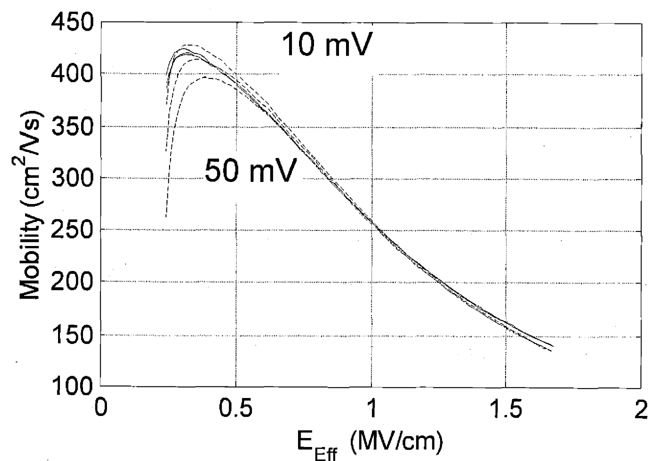


Figure A-5: Comparison of mobility vs. V_{ds} as extracted by equations (0.6), solid line & (0.7), dashed line, at 300K from a L=20 μm NMOS device with V_{ds}= 10, 25, 50 mV. Note the removal of the V_{ds} dependence by using eqn. (0.6).

As is visible in Figure A-3, Figure A-4, and Figure A-5, the corrections included in equation (0.6) impact the low V_{gs} (low E_{eff}) mobility values, but at higher V_{gs} they have little impact [5]. The difference in mobility calculated from (0.6) and (0.7) at large E_{eff} visible in Figure A-3 is due to poly depletion which reduces the C_{gsd}, thus reducing

the $F(V_g)$ and increasing the extracted mobility. Note that equation (0.6) is only valid down to near the threshold voltage of the device where the assumption of the linear variation of charge along the channel breaks down. Thus, all of the mobility data is reported for gate voltages down to the linearly extrapolated threshold voltage of the device.

Others have addressed the shortfalls of the standard equation for mobility (0.7) by measuring the capacitance of a device under bias [6], or calculating a correction factor to the extracted mobilities from analytical modeling [5] and simulation [7].

A.2.1 Effective Vertical Field

The universal mobility that occurs in a MOSFET inversion layer is a function of the effective vertical field in the inversion layer. This effective field is generally written as [3]:

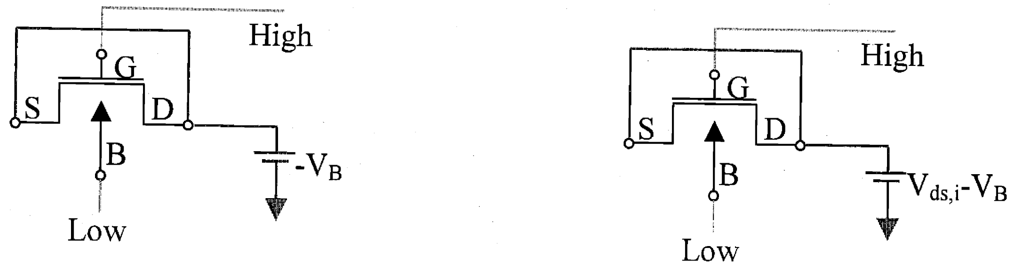
$$E_{eff} = \frac{Q_b' + \eta Q_i'}{\varepsilon_{si}} \qquad (0.8)$$

where Q'_b is the bulk depleted charge, Q'_i is the inversion charge, ε_{si} is the dielectric constant of silicon, and η is a fitting parameter generally ½ for electrons and $^1/_3$ for holes at room temperature. At other temperatures, η is fit by choosing the value that makes the mobility curves overlap for different V_{bs} [8]. For this thesis, curves at V_{bs}=0.5,0,-1 V were compared. The values for η used were:

Temperature (K)	η
350	0.42
300	0.50
250	0.67
200	0.70
150	0.72
100	0.76

Table A-1: Values used for η versus temperature for the electron mobility measured on the L=20 μm NMOS device in this thesis.

The bulk charge (Q'_b) is found by integrating the measured gate to bulk capacitance (C_{gb}), with source and drain grounded, and dividing it by the area (W*L) (Figure A-1). Similar to the case for Q_i, the impact of finite V_{ds} is accounted for by averaging the C-V at $V_d = V_s = 0$ and $V_d = V_s = V_{ds,i}$ (Figure A-6).



A: Standard C_{gb} which yields the charge near the source. V_g=(V_{High}+V_B)

B: Modified C_{gb} which yields the charge near the drain. (V_g=V_{High}+V_B)

Figure A-6: Measurement setup for Q'_b(source) and Q'_b(drain). V_B is the substrate bias applied during the measurement of the I-V.

The starting point for integration of the Q'_b is the flat band voltage (V_{fb}). Since V_{fb} is very difficult to analytically model and extract for non-uniformly doped channels, careful 2-D numerical simulations were done to find V_{fb}. Using the inverse modeled long device from Chapter 5, the V_{fb} was chosen at the gate voltage in the simulation that minimized the net charge at the surface of the device in the middle of the channel. The values used for the NMOS devices in this thesis were:

Temperature (K)	V_{fb} (V)
350	-0.885
300	-0.920
250	-0.950
200	-0.980
150	-1.010
100	-1.055

Table A-2: Values used for V_{fb} versus temperature for the electron mobility measured on the L=20 μm NMOS device in this thesis.

References

[1] S.M. Sze, *Physics of Semiconductor Devices*, second edition. Wiley, New York, 1981, p.51.

[2] C. G. Sodini, T. W. Ekstedt, and J. L. Moll. "Charge Accumulation and Mobility in Thin Dielectric MOS Transistors." *Solid-State Electronics*, Vol. 25, No. 9, 1982, pp. 933-841.

[3] S. Takagi, A. Toriumi, M. Iwase, H. Tango. "On the Universality of Inversion Layer Mobility in Si MOSFET's: Part I – Effects of Substrate Impurity Concentration." *IEEE Transactions on Electron Devices*, Vol. 41, No. 12, December 1994, pp. 2357-2362.

[4] Y. Taur and T.H. Ning. *Fundamentals of Modern VLSI Devices*. Cambridge University Press, USA, 1998, p. 121.

[5] C.-L. Huang, and G. SH. Gildenblat. "Correction Factor in the Split C-V Method for Mobility Measurements." *Solid-State Electronics*, Vol. 36, No. 4, 1993, pp. 611-615.

[6] C.-L. Haung, J. V. Faricelli, and N. D. Arora. "A New Technique for Measuring MOSFET Inversion Layer Mobility." *IEEE Transactions on Electron Devices*, Vol. 40, No. 6, June 1993, pp. 1134-1139.

[7] J.-G. Ahn, W.-S. Choi, Y.-K. Park, and H.-S. Min, *Proceedings: 1991 International Semiconductor Device Research Symposium (ISRDS)*,1991, p. 123

[8] C.-L. Huang and G. SH. Gildenblat. "Measurements and Modeling of the n-Channel MOSFET Inversion Layer Mobility and Device Characteristics in the Temperature Range 60-300 K." *IEEE Transactions on Electron Devices*, Vol. 37, No. 5, May 1990, pp. 1289-1300.

Appendix B
Fabrication Technology for 50 nm MOSFETs

Translating an optimal design for lower temperature operation into a real device poses a series of fabrication challenges. The thin gate oxides, shallow yet low resistance source and drain, and the abruptly changing doping profiles required significant process development to implement. The following appendix will examine the challenges faced in each of the sections of the process flow and the solutions that were found for processes possible in MTL's Integrated Circuit Laboratory (ICL). Both NMOS and PMOS devices were built on separate wafers to allow parallel processing, rather than doing both on one wafer (CMOS), which would require each doping step to be done twice.

The first section uses process and device simulations to explore halo doping options and to motivate the thermal budget control required. The following sections examine the challenges and solutions for each of the sub-processes used in the fabrication process. Many of these challenges reflect the tighter tolerances that are required as the dimensions of the MOSFET scale.

B.1 Device Design Simulations

Similar to the approach in Chapter 5, the goal for these set of fabricated devices is to achieve a range of low threshold voltages (0-200 mV at 300 K), at channel lengths near 50 nm, so that at lower temperatures they would map out a design space of threshold voltage and short channel effects.

The set of devices focused on varying the halo doping to achieve these different threshold voltages, keeping the oxide thickness and source & drain implants fixed. To decide on the halo implants to use, an extensive set of process simulations were done

using TOMCAT (2-D monte carlo implant simulator from UT-Austin [1]) to simulate the implants and TSUPREM4 [2] to simulate the dopant diffusion. The results of these process simulations were fed into a 2-D numerical device simulator, MEDICI [3], to generate the I-V characteristics and extract the threshold voltage and short channel effects.

To validate this approach, process simulations were performed (Figure B-1) for 100nm devices previously fabricated at MTL[4]. MEDICI simulations using the simulated doping profiles compare favorably with the real device data (Figure B-2). Although the V_{th} of the TSUPREM4/MEDICI simulated device is 80mV lower, its DIBL is similar to the real device data.

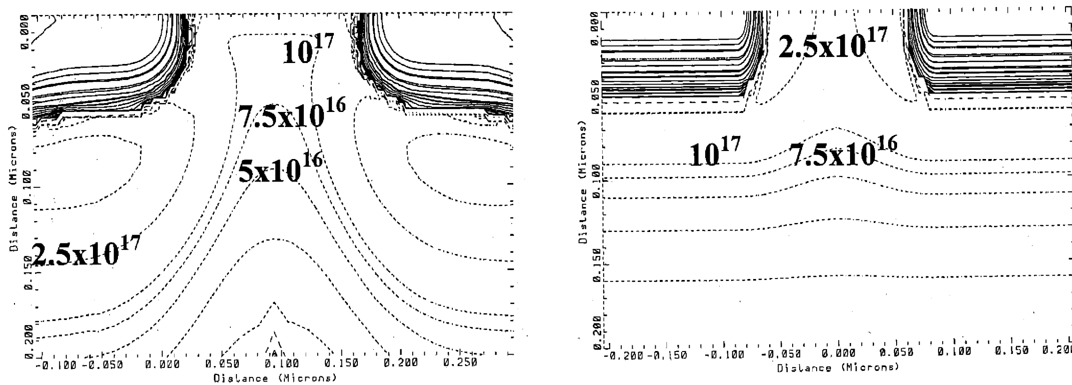


Figure B-1: TSUPREM4 simulation (Left) and Inverse Modeled doping profiles (Right).

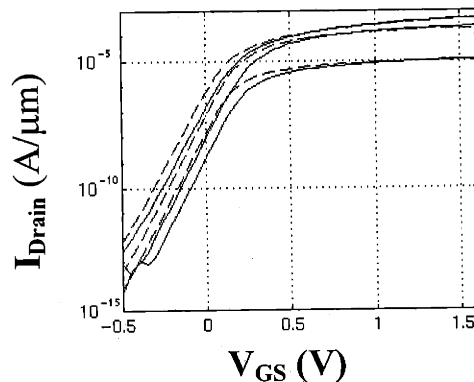


Figure B-2: Comparison of measured device data (solid lines) to Simulated data using the process simulated profiles (dashed lines). V_{ds}=0.01, 0.21, 1.61 V.

Comparing the simulated doping profiles with profiles extracted from inverse modeling of the real device data shows that TSUPREM4 seems to have underestimated the diffusion of dopants in the channel. The similarity of the I-Vs of the process simulated device and real device, however, suggest that the marked differences in doping profiles deeper in the device have little impact on the device characteristics. Thus the inverse-modeling may not have extracted these parts of the profiles and only the channel area can truly be compared.

The similarities of the I-V characteristics give validity to this simulation approach and also suggest the need for a wide range of design. Because of the range of V_{th}s designed, even if the fabricated devices' V_{th}s are shifted from the simulation devices, some of the designs should still have low threshold voltages with controlled short channel effects. The doses, angles, and energies were also picked to cover a range of designs with the aim of examining the impact of changes in each parameter.

For both NMOS and PMOS, two different halo dopants were used. For the NMOS devices, Indium and Boron were used. The NMOS designs, with Arsenic source and drains, used implant angles near 0° due to the low diffusivity and low standard deviation of the peak of the arsenic implants coupled with the higher diffusivity of both Boron and Indium. An 80 nm gate length was used to achieve L_{eff} =50 nm.

For PMOS devices, Arsenic and Phosphorus were used. The PMOS designs, with Boron source and drains, had halo implants angles that ranged from 10-30°. The higher diffusivity of the Boron used in the source and drain required the halo doping to be put further out in front of the shallow and deep source/drains. In addition a longer gate length of 105 nm was needed to achieve the same L_{eff} =50 nm. Although Antimony had been previously explored for PMOS halos, it was not used in this case because the halo dopant levels needed were higher than the reported activation levels for Antimony.

Dopant	Dose (cm^{-2})	Angle (° off vert.)	Energy (keV)	Quad/ Single	300 K V_{th} (V)	90 K V_{th}	L_{eff} (nm)	300 K DIBL(mV)
In	6e13	0	35	S	0.073	0.19	45	131
In	2e13	0	35	S	-0.06	0.11	44	190
In	8e13	0	35	S	0.13	0.32	54	52
In	2.5e13	15	35	Q	0.16	0.29	50	85
B	4e13	0	15	S	0.081	0.21	46	141
B	2e13	0	15	S	0.032	0.16	50	111
B	4e13	0	10	S	0.12	0.24	45	118
B	1e13	20	15	Q	0.22	0.36	45	99
As	2e13	30	35	Q	-0.12	-0.25	74	61
As	2e13	20	50	Q	-0.015	-0.15	69	90
As	1.3e13	30	45	Q	-0.06	-0.18	72	100
As	0.7e13	40	55	Q	-0.14	-0.26	59	132
As	1e13	30	80	Q	-0.22	-0.35	58	89
P	1.5e13	20	35	Q	-0.09	-0.22	69	113
P	2e13	7	25	Q	-0.001	-0.14	87	110
P	2.5e13	7	25	Q	-0.052	-0.22	82	59
P	2e13	15	35	Q	-0.14	-0.28	85	78
P	2e13	25	35	Q	-0.2	-0.4	70	75

Table B-1: Summary of halo implant conditions used for NMOS (In and B) and PMOS (As and P) devices. V_{th} and DIBL numbers are from MEDICI simulations using the TSUPREM4/TOMCAT simulated profiles.

Because of the ring shape of some of the devices, the angled halo implants were performed as quad implants. A quad implant starts with a rotation 45° from the device orientation and then does 4 implants at 90° rotations, all the while at the same tilt. To convert the quad implants back to standard double tilt (+,-) implants, the dose would need to be doubled and the tilt angle slightly increased.

B.2 Fabrication Challenges

The 20 Å physical oxide thickness, 30 nm source/drain extension depth, and thermal cycles chosen for these devices were initially drawn from the National Technology Roadmap for Semiconductors (NTRS)[5] and various papers in the literature. The process and device simulations then confirmed that these choices would allow the creation of 50 nm L_{eff} devices with low threshold voltages and controlled short channel

effects. These three parameters were the key drivers behind much of the process development.

The 20 Å t_{ox} used for these devices required both the development of a growth process and a very selective etching process. The tolerances on over-etching the gate polysilicon, so as not to etch into the source-drain areas, were extremely tight.

Data from various papers in the literature suggested [6] that the combination of a 1050 °C spike anneal and very low energy implants (0.5-2 keV) would achieve the 30 nm design depth for the source/drain extensions. The very shallow extension implants put very tight requirements on both the amount of oxide on the silicon during implantation as well as the amount of oxide and silicon that could be lost during subsequent cleans. The need to really limit the thermal budget outside of the spike anneal forced the spacer process to be moved from using a 780 °C LPCVD nitride to a 450 °C PECVD nitride.

In addition to meeting the tighter tolerances and reducing the thermal budget, the process was carefully examined to try to reduce the number of steps needed. Since only individual devices were being fabricated, not circuits, the process areas of device isolation and back-end - pieces not related to the intrinsic MOSFET device – were targeted. Separate from process modifications, a one-mask (one photolithography step) MOSFET design was examined in an attempt to drastically shorten the process. The advantages and disadvantages of this concept are presented in section B.16.

B.3 Overview of the process steps

The major sections of the process flow are:

- Isolation
- Threshold Voltage Implant
- Gate Formation
- Shallow source/drain extension implant
- Spacer formation
- Backside implant
- Deep source/drain implant
- Anneal
- Silicide
- Contact cuts
- Metal deposition and patterning

Putting these steps together, the final device looks like:

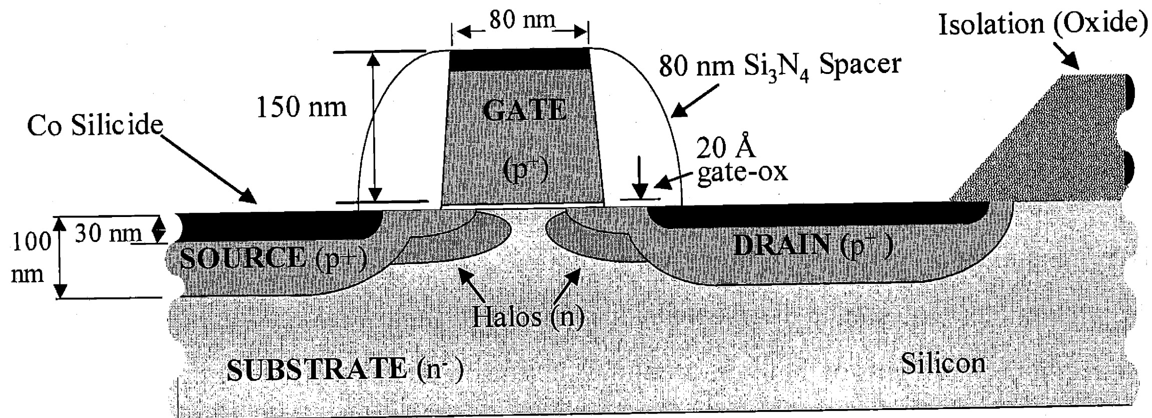


Figure B-3: Sketch of fabricated P-MOSFET cross-section.

The focus of the next sections is to look at specific challenges and important details for each of the parts of the process flow outlined above.

B.4 Isolation

Two common methods of isolating devices are LOCOS (local oxidation of silicon) isolation and for more modern processes, STI (shallow trench isolation). In this case, a simplified approach was taken to attempt to reduce the number of steps required to complete this sub-process. For these devices, a thick wet oxide was thermally grown and then active areas were etched into it. The two key pieces that allowed this to work were starting with heavily enough doped wafers and sloping the sidewalls of the oxide.

Generally a field implant is used to insure that the threshold voltage of the parasitic MOSFET on the isolation will not turn on in the voltage range of interest. In this case, wafers doped around $1x10^{16}$ cm^{-3} coupled with an isolation oxide thickness of 4500 Å satisfied this condition without the need for an additional field implant. Note that besides the parasitic MOSFET, the doping needs to be high enough that the concentration of positive fixed charge in the oxide does not invert the surface (only an issue for NMOS which has a p-type substrate). In the one case that lower doped wafers (~ $3x10^{15}$ cm^{-3}) were used to test NMOS source/drain diode formation, extensive diode leakage was

observed that which was attributed to the silicon surface underneath the isolation inverting from the fixed charge in the oxide.

The sidewalls of the oxide isolation at the active area edge were sloped to about a 45 degree angle to make sure that the poly would smoothly transition from the active area to the isolation and a poly spacer would not form when the gate was etched (Figure B-16). If the sidewalls were perfectly vertical, then later in the process when the gate is etched, the directional nature of the RIE gate etch would leave some polysilicon in the corner between the Isolation wall and the source or drain in the same way a spacer is formed before the deep source/drain implant (See Section B.8). By sloping the sidewalls the corner where poly might remain was removed. In addition, the Isolation step height here was about 4000 Å, while the polysilicon thickness was about 1400 Å, so the slope of the sidewalls was critical to guarantee continuity of the polysilicon gate from the active area onto the isolation.

Using too shallow of an angle will cause problems later in the process after the deep source and drain implant. The wet processing and HF dips that occur after the deep source drain implant will pull the isolation back away from the implanted (and annealed) source and drain areas. Since a silicide will cover the entire exposed area, if the isolation recedes past the source drain edge, the silicide will short the source/drain to the substrate creating large edge leakage for the source/drain diode. The reason the angle matters is that for every 50 Å that the isolation is etched, the edge of the isolation will move 50 Å/$\sin(\theta)$ = 71 Å at 45°. An isolation slope closer to 60 ° would make this much less of an issue, while still satisfying the above reasoning for using a slope.

The sloped sidewalls were fabricated by adding oxygen to the CHF_3 RIE etch to erode the photoresist mask at the same time that the oxide was being etched [7,8]. In theory, to achieve 45° sidewalls, the lateral etch rate of the photoresist had to be made equal to the vertical etch rate of the oxide. The final profile of the photoresist looks similar to the profile in the oxide, suggesting the corners of the resist etched more quickly, as might be expected for an isotropic etch which will round off any corners. Also it is possible that some sputtering of the resist was happening, which has an angular dependence that can also produce this sloped shape in the resist [9]. This shape in combination with the observation that while the process gave sloped sidewalls for 4000^+

Å oxide, sloped sidewalls were not observed for 1000 Å oxides, suggests it took time for a sloped resist to form and then be transferred into the oxide.

To achieve the 45 ° sidewalls, a ratio of 1:3 O_2:CHF_3 gas flows was used. The full RIE recipe used on the AME P5000 is listed in Table B-2. In general, the etch was stopped with the thinnest point on the wafer having about 100 Å of oxide left. The remaining oxide was removed using 50:1 HF during a subsequent RCA clean. This approach made sure that the active are silicon surface was not damaged by the RIE etch.

Parameter	Main Etch
CHF_3 Flow (sccm)	30
O_2 Flow (sccm)	10
Pressure (mTorr)	25
RF Power (W)	350
Magnetic Field (Gauss)	90
Oxide etch rate (Å/sec)	30.5
Photoresist etch rate	42
Sidewall Angle	~ 45 °

Table B-2: Isolation Oxide Etch Recipe

B.5 Threshold Voltage Implant

With the goal of creating low threshold voltage devices, the majority of the channel doping was planned to be done by the halo implant. However, the arsenic (and the boron) used in the deep source/drain implant showed extremely long tails, that were not being compensated by the ~$1x10^{16}$ cm^{-3} doping of the wafer. To cut off these tails and thus prevent any possibility of a sub-surface short from source to drain at short channel lengths, a deep V_{th} implant of 35 keV, 3e12 cm^{-2}, Tilt=7 Boron for NMOS and 120 keV, 4e12 cm^{-2}, Tilt=7 Phosphorus for NMOS.

B.6 Gate Stack

Having a thin gate oxide is critical part of controlling short channel effects. With the goal of achieving 50 nm channel lengths, a 20 Å gate oxide was fabricated. Having

such a thin gate oxide requires a very high etch selectivity between polysilicon and oxide in order to stop on the oxide when the gate is etched. Finally, short gate lengths were attempted using a shrink of optically defined lines, although x-ray lithography could have provided better results. Figure B-4 shows the general gate stack fabrication process described in this section.

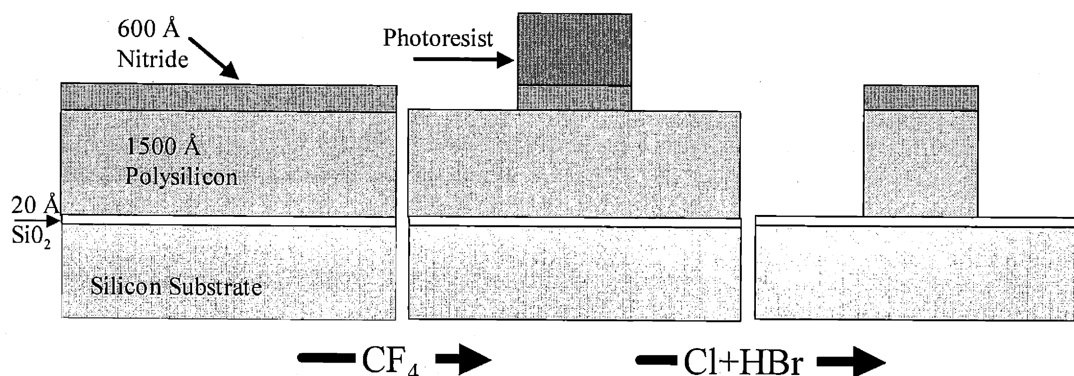


Figure B-4: Gate Stack Fabrication Sequence

B.6.1 Gate oxide growth

The MOSFETs in this thesis use a 20 Å physical gate oxide grown in an N_2O ambient. As well as being demonstrated in the literature to give good quality gate oxides[10], nitrous oxide provides a slight nitridatation of the gate oxide, thus slowing Boron penetration from the gate in PMOS devices.

An initial set of devices were fabricated to examine different oxide thicknesses and gate doping conditions. MOSFETs with 20 Å, 25 Å, and 45 Å (pure O_2) oxides were fabricated:

Fabrication Run	Temperature, Gas	Time	Thickness (Ellipsometer)
Initial Set	800 °C, O_2	30 min.	45 Å
	800 °C, N_2O	12 min.	26 Å (26 Å)
	750 °C, N20	8 min.	19 Å (19 Å)
Final Devices	750 °C, N20	6 min.	22 Å (18 Å)

Table B-3: Gate oxide growth conditions. Numbers in parentheses are from Berkley's fabrication facility.

These results are similar to those achieved in Berkley's fabrication facility (numbers in parenthesis above)

Such thin oxides are very sensitive to the initial oxide starting thickness, as well as any interactions with heated oxygen outside of the oxidation step. Measuring wafers right after the rca clean (SC1: 10min 73 °C, Rinse to ~ 7 MΩ, 50:1 HF 30sec (clear any oxide), Rinse to ~ 7 MΩ, SC2: 15 min ~80 °C, Rinse to ~ 7 MΩ, Spin Dry 300s) using a UV1280 Spectroscopic Ellipsometer gave oxide thicknesses in the 7-10 Å range, similar to what others have reported. An HF last process was not attempted because of metallic contamination that HF can easily cause due to plating of positive ions onto the wafer surface as the oxide is etched [11]. Having dry N_2 in the spin dryer was critical to maintaining this oxide thickness. In one case a small amount of water was leaking through a valve and mixing with the heated N_2 that was drying the wafers and increased the final (after oxidation) thickness considerably. In addition, loading the wafers into the tubes at 450 °C (as opposed to 800 °C) reduced the oxide thickness by 5-10 Å. It was assumed that during the load (or unload steps) oxygen from the lab air was being heated and oxidizing the wafers.

Results from the initial set of devices show working capacitors for each of the gate oxide thicknesses with no gate-oxide shorts found, even for 1 mm^2 capacitors.

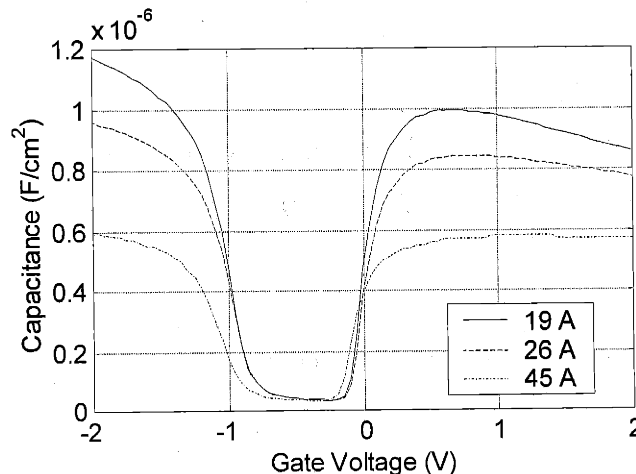


Figure B-5: Capacitance voltage characteristics for the three different oxide thicknesses (ellipsometer values) measured on a 100x100 μm N-MOSFET.

B.6.2 Polysilicon and Nitride Deposition

Immediately following the gate oxidation, the wafers are loaded into the polysilicon LPCVD tube and amorphous silicon is deposited at 560 °C. Following previous fabrication runs [12], amorphous silicon was used, with the goal of achieving straighter sidewalls when the very short gates were etched. However, the subsequent Nitride deposition crystallizes the amorphous silicon, so the gate is actually polysilicon when etched.

Next the wafers are loaded into the Nitride LPCVD tube and 600 Å of stoichometric nitride is deposited at 780 °C. This material will serve as the hard-mask for the gate etch. In the case where the hard-mask was shrunk before the gates were etched (see section B.6.3.1.1), the wafers were transferred to the LTO (low temperature oxide) LPCVD tube and 1000 Å of LTO was deposited.

B.6.3 Gate Definition

This step involves using lithography to pattern the resist, transfer of that image into a nitride hard mask, removing the resist, and then etching the polysilicon using the nitride as a mask. The nitride hard mask is used for two reasons. During the gate etch it allows better selectivity by removing carbon, found in photoresist, from the etch plasma, which would degrade the selectivity between the polysilicon and the oxide [13]. In addition, the hard mask has been found to etch more slowly than photoresist, which gives a better gate profile.

A nitride hard mask, rather than oxide, is used because the hard mask needs to be removed before the gate can be doped, silicided, and contacted with metal. Removing the nitride hard mask right after the wafer is re-oxidized is possible without damaging the exposed gate-oxide edges and with large process latitude. An oxide hard mask could not be removed at this point as the gate oxide edges would be attacked during the oxide etch. An oxide hard mask has previously been used and removed after the spacers were formed. In this case, there is a very small process window between removing the oxide on top of the gate, and etching away the LTO that is underneath the nitride spacers [4].

An oxide hard mask has be successfully used in the case of using polysilicon doped before the gate was etched and where not siliciding the gate was allowed. In this case, the oxide was etched through during the contact cuts so that the metal contacted the poly [14].

B.6.3.1 Lithography

For MOSFETs fabricated in the ICL, the only dimension that is truly scaled is the gate length. For all levels besides the gate level , a G-line stepper was always used. Two approaches were available to achieve small linewidths for the gate level, x-ray lithography and a shrink of g-line lithography using a wet etch.

B.6.3.1.1 Hard Mask Shrink

Gate lengths of 0.5 μm were initially patterned in photoresist by overexposure from 1.0 μm lines on the mask. In order to achieve a nitride hard mask of less than 100 nm, lateral wet etching was used (Figure B-6). After etching the 0.5 μm photoresist line into the LTO/nitride stack, the photoresist was removed and the wafer was etched in hot phosphoric acid (175 °C), which etched the nitride, but only slightly etched the LTO. The oxide cap allows the nitride to be laterally etched, but prevents it from being vertically etched. Figure B-6 shows the modified hard-mask process where the layer of LTO is added, patterned, and removed before the gate is etched. The lateral etch rate of the nitride was carefully calibrated by measuring etched linewidths in a SEM.

As will be explored in section B.15, this process was found to not result in a perfectly straight hard mask which caused serious problems later in the process. This result was in contrast to the case where a nitride cap was used to help shrink a densified LTO hardmask in BOE, which showed very straight hardmask sidewalls [15].

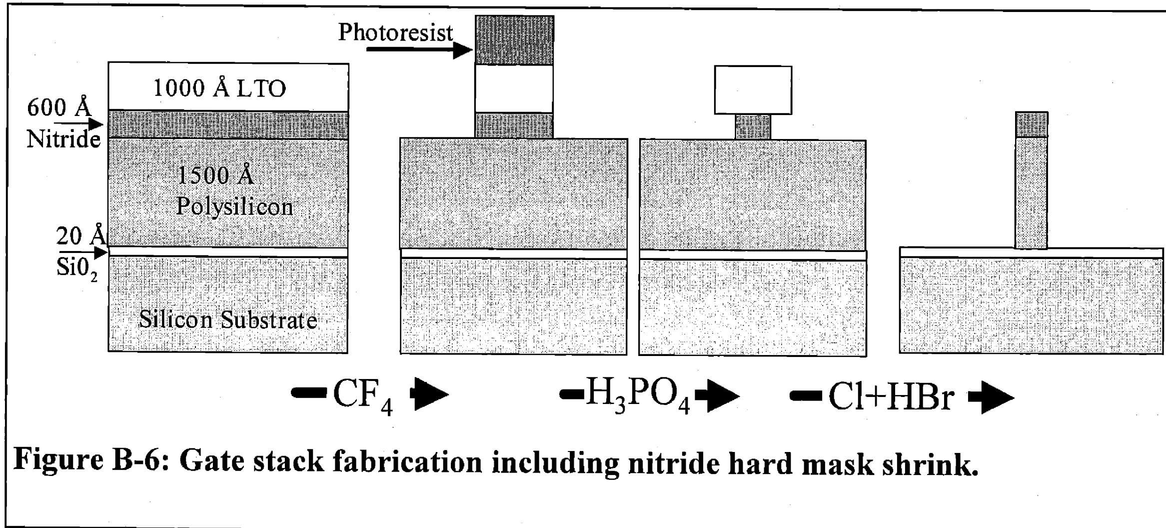


Figure B-6: Gate stack fabrication including nitride hard mask shrink.

B.6.3.1.2 X-ray

Although X-ray lithography was not used for the fabricated devices, it has the capability to define linewidths down to 30 nm and has previously been successfully used for the gate level lithography for fabricating MOSFETS in the ICL [4].

B.6.3.2 Nitride Etch

The process recipe used in the AME P5000 to etch the nitride hard mask is:

Parameter	Main Etch
CF_4 Flow (sccm)	15
O_2 Flow (sccm)	5
Pressure (mTorr)	50
RF Power (Watts)	250
Magnetic Field (Gauss)	100
Nitride Etch Rate (Å/s)	34
Thermal Oxide Etch Rate (Å/s)	12
Polysilicon Etch Rate (Å/s)	13

Table B-4: Nitride Etch Recipe

B.6.3.3 Polysilicon Etch

The gate etch has two main requirements. First the etch must replicate the hardmask pattern in the polysilicon layer with straight sidewalls. Second, the etch must be very selective between silicon and oxide and thus be able to stop on the gate oxide. These two issues were separated by etching most of poly-silicon thickness using one set of conditions [16], and etching the last few hundred angstroms (usually 500 Å) using a different set of conditions. All of these etches were performed in an Applied Materials P5000 MERIE (magnetically enhanced reactive ion etcher) using a combination of chlorine and hydrogen bromide gases. The conditions are listed in Table B-5.

Parameter	Main Etch	Over etch
Cl_2 Flow (sccm)	20	0
HBr Flow (sccm)	20	40
Pressure (mTorr)	200	100
RF Power (Watts)	350	75
Magnetic Field (Gauss)	50	50
Polysilicon etch rate (Å/s)	55	10
Thermal oxide selectivity	-	100:1
Nitride selectivity	5.4:1	5.3:1

Table B-5: Gate Etch Recipe on AME P5000.

Achieving straight sidewalls requires both an anisotropic etch and a mask material that does not erode quickly with the etch. Prior experience with patterning small gate lengths found that a layer of nitride on top of the polysilicon worked far better than resist as a mask for the polysilicon gate etch. Thus, as described above, a 600 Å layer of LPCVD nitride is deposited as part of the gate stack. The nitride is patterned in an RIE with CF_4 and O_2 using the photoresist as a mask. The resist is then stripped and the gate is patterned using the nitride as a hard-mask. The selectivity of the etch is better than 5:1 (polysilicon : nitride) which is slow enough that the edges of the nitride mask do not significantly change during the gate etch.

In addition to having a durable masking layer, achieving straight sidewalls requires the vertical etch rate to be much higher that the lateral etch rate, which is the definition of an anisotropic etch. In an RIE system this is generally achieved by having a strong DC bias between the wafer and plasma as well as having a low chamber pressure.

The high DC bias accelerates the ionized species vertically towards the wafer. The low chamber pressure means that there are less gas molecules for the ionized species to collide with and thus get knocked off path. In addition, one would like the physical (or physically activated) etching due to bombardment of the surface to dominate over the chemical etching as the ionized species interacts with the silicon, since the straight chemical component tends to be isotropic. For the main part of the etch, a high RF power (which causes a high DC bias) was used along with a low chamber pressure.

Achieving a selectivity to the gate oxide allows an over-etch of the polysilicon. Overetching the polysilicon allows one to clear the entire wafer of poly, since the etch, and probably the poly deposition, are bound to be non-uniform. This means that areas of the wafer will clear before others, so while the poly continues to etch in some areas, the gate oxide needs to resist the etch in other areas. Since a selectivity of around 100:1 polysilicon to oxide was achieved, a long over etch, thus ensuring the poly has cleared both in the field and at the base of the gate features, was possible.

The selectivity of the over-etch was achieved using lower DC biases, via lower RF power, and pure HBr as the etch gas [17,18]. As the DC bias is lowered, both the silicon and oxide etch rates decrease. However there is a region where the oxide etch rate has almost stopped, but the silicon is still etching. By picking conditions around this point, very high selectivities are possible.

A last detail of the etch is that both steps are selective to oxide, which means that all native oxide must be removed before starting the etch. This was done with a 15 second BOE dip right before the etch. After dump-rinse and spin-dry, the wafers are immediately loaded into the load-lock of the etcher and the load-lock is pumped down.

B.6.3.4 Re-oxidation and Hard Mask Removal

After the gate is etched, the next step is to grow a thermal oxide to help repair any damage the etch may have caused to the gate oxide at the edge of the gate. For this particular process, this additional oxide is important to protect the gate oxide during the wet etch to remove the nitride hard mask. A 850 °C, 15 min anneal in dry O_2 grew an additional 30 Å, such that the total oxide on the source and drain regions was 49-52 Å.

This re-oxidation is performed before the nitride is removed to protect the gate-oxide edges from the slow etching of oxide that occurs in the hot phosphoric acid used to remove the nitride hard mask (see section B.6.4). The re-oxidation grows only a few angstroms of oxide on top of the nitride which causes a short delay in the etching of the nitride while the hot phosphoric acid slowly etches this thin oxide.

B.6.4 Wet Cleans post gate-etch

The thin gate oxide and extremely shallow implants used in this fabrication process necessitated a very tight control on the amount of silicon and oxide etched during the process steps between gate etch and spacer deposition, when the gate oxide edge and shallow source/drain area are exposed.

The results of an initial set of fabricated devices showed a diode-like connection between source and drain that scaled mostly with width, but decreased for very short length devices, possibly due to gate resistance. This behavior suggests a leakage path at the edge of the gate.

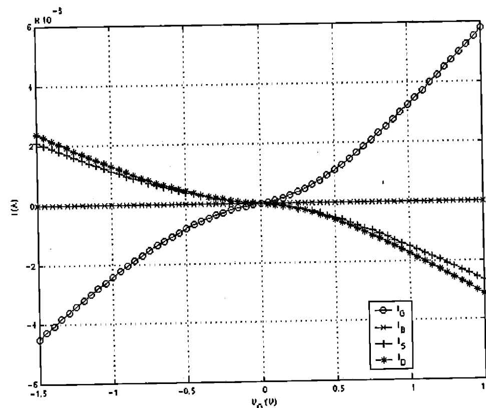


Figure B-7: Current Voltage Characteristics of the source/drain leakage for a 100 μm x 100 μm PMOSFET with t_{ox}=22 Å. Since I_B is negligible, the current flows only from gate to source and gate to drain.

TEM analysis showed that about 15 nm of silicon had been etched away from the source and drain area (Figure B-8). This loss of silicon is unlikely to have come from the gate overetch step given the extremely high selectivity and care with which that step was

performed. The depth was found to match the etch depth predicted by summing up all of the wet etches and cleans that the device was exposed to.

Figure B-8: TEM of a finished device from the initial device runs showing the etched silicon film in the source/drain area. The "bump-up" in the silicon next to the left edge of the gate is an artifact of the picture due to gate not being perpendicular, with depth, to the TEM slice.

Using an approach described in the literature [19], the etch rates of silicon, oxide, and nitride were measured in the various wet process steps. The results show that during the normal SC-1 clean, up to 55 Å of silicon could be lost, and that during a long nitride etch, 10s of angstroms of oxide could be etched.

RCA clean		Thermal $Si0_2$ Etch Rate (Å/min)	Silicon Etch Rate (Å/min)	Remaining Oxide (Å) (final oxide on exposed Silicon)
SC-1 (1:1:5) 70C	Literature	~2 [20,21,22]	~5 [19]	11.5 [23,24]
	Experiment	1.56	3.67	11.4
SC-2 (1:1:6) 70C	Literature	Negligible	?	5-8 [22,25], 9.5 [26] (w/HF dip after SC1)
	Experiment	0- 0.27	?	11.4 (no HF dip)

The SC-1 Solution ($NH_4OH:H_2O_2:H_2O$, 1:1:5, 70-73 °C) etches both Silicon and $Si0_2$. The Silicon is oxidized by the Hydrogen Peroxide and the oxide is etched by the

Ammonium Hydroxide. Etch rate of Si [19] and SiO_2 [20,21] increases significantly as temperature increases. Silicon is not etched until the t_{ox} is reduced down to about 10 Å [21] (oxygen can then diffuse & react with the surface). The silicon etch rate has been reported to decrease with time [19] which means that the initial etch rate may be higher than the number reported above for the SC-1 solution.

Pirhanna Clean		Thermal SiO_2 Etch Rate (Å/min)	Silicon Etch Rate (Å/min)	Remaining Oxide (Å)
Pirhanna (3:1, H_2SO_4:H_2O_2)	Literature	?	?	14 [27] (<111> Silicon)
	Experiment	<0.25 (0.15 ?)	0.3	19.2

175 °C Phosphoric Acid		Thermal SiO_2 Etch Rate (Å/min)	Silicon Etch Rate (Å/min)	LPCVD Nitride Etch Rate (Å/min)	Remaining Oxide (Å)
175 °C Phosphoric Acid	Literature (160 °C)	0.8	0.3 [28]	42	?
	Experiment	1.1	0.8	86.5	4.9

Heavily doping silicon or silicon dioxide increases their etch rate dramatically. Etch rates for BPSG in a SC-1 solution have been measured at 88.5A/min [29]. In hot phosphoric acid, etch rates of n^+ doped polysilicon are > 7 Å/min [30] and PSG about 24 Å/min [30].

Given these results, a pirhanna clean was substituted for the SC-1 clean in the RCA cleaning steps performed after the gate is etched (the SC-2 solution was used as before). The cleaning method of the pirhanna solution, attacking particles on the surface, is different and less effective than the SC-1 solution which oxidizes then etches away the silicon underneath particles, lifting them off into the solution [31,32]. Because of this, any case where there was resist on prior step an additional oxygen plasma ash was performed on the wafers to make sure all organics were removed. All HF dips during the RCA clean were skipped until after the deep source/drain anneal was performed to avoid etching away the oxide at the gate edges, as well as over the source drain area. In addition, the amount of re-oxidation after the gate etch was increased to allow for the 10^+

Å loss that occurs during the hardmask removal in the 175 °C phosphoric acid. In addition to eliminating the short between gate and source/drain, retaining the silicon film is very important to retaining the very shallow implants performed for the source/drain extensions, as is explored in the next section.

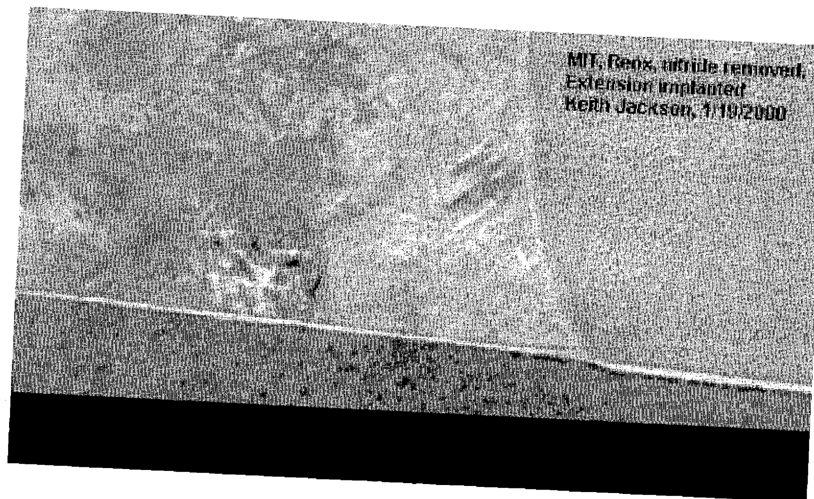


Figure B-9: TEM of a device fabricated with the new approach to cleans and wet processing.

TEM performed on devices fabricated with the new approach to cleans and wet processing show minimal etching of the silicon over the source and drain. In addition, measurements of the final devices show no source/drain to gate leakage (see section B.15).

B.7 Shallow Source/Drain Implant

To achieve the 30nm deep (at 10^{18} cm^{-3}) source and drain extensions needed for short channel effect control, very low energy implants were used. For the NMOS, a split between the 1 and 2 keV As with a 10^{15} cm^{-2} dose was used and for the PMOS, a split between 0.5 keV, $5x10^{14}$ cm^{-2} Boron (B^{11}) and 5.6 keV, $5x10^{13}$ cm^{-2} Decaborane (equivalent implant to the B^{11} one) was used. Coupling these with a 1050 °C spike anneal achieves the goal of a 30 nm junction depth [6].

Crucial to using such low energy implants is controlling the oxide thickness before implant and silicon loss during clean after the implant, but before the anneal. Figure B-10 shows the results of TOMCAT simulations of a 0.5 keV, $5x10^{14}$ cm^{-2} Boron implant into silicon with a variable oxide thickness on top of the silicon. A 30 Å oxide would cause less than 50 % of the Boron dose to end up in the silicon. Similar results occur for the Arsenic implants. Before the device was sent to have the shallow source and drain extensions implanted, the oxide thickness on the source and drain was carefully reduced to about 15 Å using a calibrated dip in 50:1 HF.

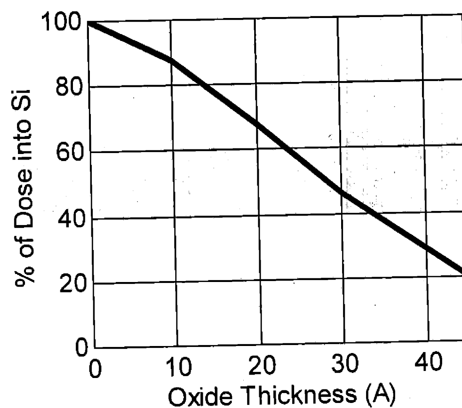


Figure B-10: Impact of oxide thickness on resulting dose in the silicon for a 0.5 keV $5x10^{14}$ Boron implant (simulated using TOMCAT).

B.8 Halo Implant

Using the simulated implant conditions from section B.1 the halo dopants were implanted after the shallow source and drain implants. Within wafer splits were achieved by coating the wafer with photoresist and clearing half of the dies for a given implant.

B.9 Backside Clean and Implant

Clearing the backside of all layers is important to provide a backside contact to the devices. A SF_6 RIE etch was used to etch through all of the layers. The wafers were then heavily implanted ($3x10^{15}$ Boron or Arsenic) to provide a good ohmic contact for

measurements at lower temperatures. This high doping is necessary to make sure the contact depends on tunneling, which has little temperature dependence, rather than the thermionic emission that would dominate for low doping, and which gives a very high contact resistance for measurements at low temperatures.

B.10 Spacer Formation

The spacer process was shifted from using an LPCVD nitride (780 °C, 2^+ hours) to using a PECVD nitride (450 °C, 10 minutes) that allows the spike anneal to be the only high temperature step the shallow extensions see.

The general spacer process involves depositing 200 Å of LTO and then 1400 Å (CHECK) of nitride. The nitride is then blanket etched back (see Table B-4 for recipe) with about a 20 % overetch. The anisotropic nature of the RIE etch leaves spacers of nitride along the vertical walls of the polysilicon gates. The 200 Å LTO thickness is picked so that the overetch of the nitride will etch ½ to ¾ of the oxide away but stay safely away from the silicon surface. (The nitride etch is not very selective to oxide ~ 3:1 nitride:oxide and not at all selective to silicon).

Unexpectedly, using the PECVD nitride caused the spacer not to appear and almost all of the nitride was etched off next to the polysilicon gates. PECVD oxide has 20-25% hydrogen (by # of molecules) incorporated into the film from the gas precursors. The hydrogen caused the RIE etch to become more isotropic and thus completely etch away the nitride. Annealing the nitride film at 600 °C for 1 hour in nitrogen evaporated a significant percentage of the hydrogen out of the film. Etching this annealed film gave acceptable results for the spacers, indicating that the extra hydrogen had been removed and that the anisotropic nature of the RIE etch was working correctly.

B.11 Deep Source/Drain Implant and Anneal

With the goal of achieving 80 – 100 nm junction depths for the NMOS and PMOS device, 20 keV, $3\text{x}10^{15}$ cm^{-2} Arsenic and 1-2 keV, $3\text{x}10^{15}$ cm^{-2} Boron implants

were used. Low energy Boron was used instead of BF_2 to avoid the fluorine which has been shown to cause increased Boron penetration in the PMOS devices. The most important criteria for the deep source and drains is that they form low leakage diodes. Results for the PMOS (Figure B-11) as well as the NMOS were consistently good.

A 1050 °C spike anneal was used to activate both the deep source/drains and extensions. A spike anneal consists of using an RTA to ramp to 1050 °C at a rate of about 80 °C/sec and then an immediate decrease in temperature at about 70 °C/sec after 1050 °C is reached. This anneal is the only temperature above 450 °C the device sees between the extensions are implanted and the deep source drains are annealed. This required the changes in the spacer process detailed in the last section .

The implant for the deep source and drain doping is also used to dope the polysilicon gate. A such, it is important that it have a high enough dose to dope the gate to a high level, and also have a low enough energy that the tail of the implant does not go through the gate oxide and counter dope the channel. The 1050 °C spike anneal, which activates the deep and shallow source and drains, also diffuses the dopants in the poly. This anneal was found to be enough to adequately diffuse the doping through the poly, but not so much as to cause significant boron penetration through the gate oxide for the PMOS devices.

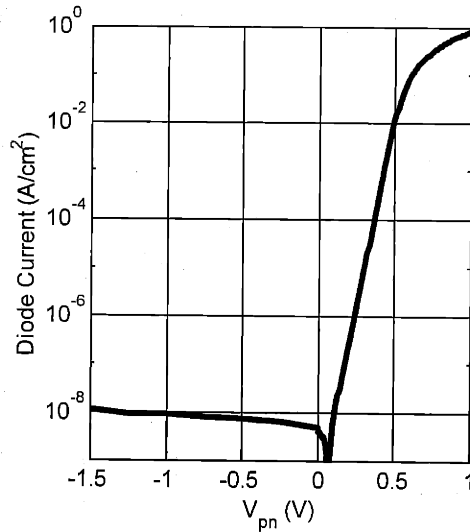
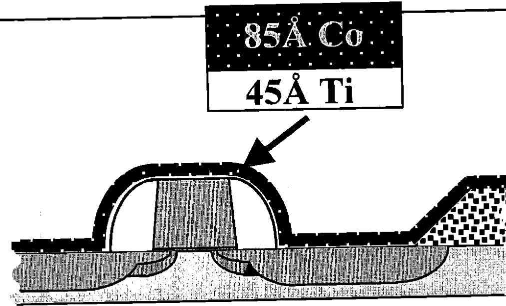


Figure B-11: Diode I-V characteristic for a PMOS deep source/drain showing the low leakage in reverse bias and near ideal subthreshold slope in forward bias. This measurement is from a test run that had only 10^{16} cm^{-3} substrate doping and had Ti-Al contacts with no silicide.

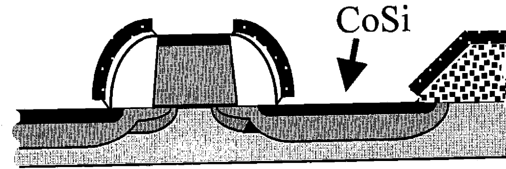
B.12 Cobalt Silicide

Although the cobalt silicide process in the ICL has previously been well documented and explored [33], it proved to be sensitive to changes in equipment status. The basic process, illustrated in Figure B-12 consists of depositing a layer of titanium then cobalt without breaking vacuum in an e-beam evaporator, performing a first step anneal at low temperatures to form CoSi, stripping the remaining unreacted Ti and Co, then using a high temperature anneal to transform the CoSi into the lower resistance $CoSi_2$ phase.

1. RCA Clean, HF dip, Rinse, Dry
2. Deposit 45A Ti, 85A, Co



3. 1st Step Anneal: ~550 °C, 30s



4. Strip remaining Ti/Co in Pirhanna.
5. 2nd Step Anneal: 700 °C, 60s

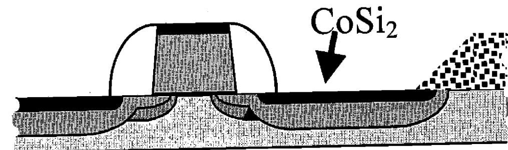


Figure B-12: Self-aligned Cobalt Silicide process.

The titanium film acts as an intermediary layer that helps getter any oxide on the silicon surface and prevent the cobalt silicide from agglomerating. However, it was found that depositing the Ti and Co at very low pressure ($\sim 2\times10^{-7}$ Torr) was necessary to avoid agglomeration of the $CoSi_2$ films. The thickness of the cobalt was chosen such that the resulting $CoSi_2$ film would be 30 nm deep, about one third the depth of the deep source and drains. In general the thickness of silicon consumed in forming the $CoSi_2$ has been found to be 3.6x the thickness of the metal cobalt film deposited [33].

The concept behind the two step anneal is that at the higher temperature (700 °C) needed to form the $CoSi_2$, the silicon diffuses rapidly into and along the cobalt film and forms voids at the source/drain edges and stringers onto the isolation and spacers. The lower temperature first step anneal allows the unreacted Ti and Co on the isolation and spacers to be removed while the CoSi is unaffected, and little diffusion of the silicon has occurred at the lower temperature. During the subsequent higher temperature anneal

there is no metal film for the silicon to diffuse along, avoiding the formation of voids and stringers, but epitaxial $CoSi_2$ still forms.

The initial results of duplicating the standard silicide process showed exactly these void and stringer problems (Figure B-13). It was determined that the temperature in the RTA that the thermocouple was reading was both lower than the actual wafer temperature, as well as significantly lagging behind the wafer temperature during the temperature ramp (Figure B-14). The wafer was thus being exposed to much higher temperatures that what the thermocouple in the RTA was indicating. Comparing the temperature of the thermocouple to a full wafer thermocouple, two temperature points were correlated. A series of different temperature anneals were performed to find the RTA setting that gave the expected silicide behavior (Figure B-15). Note that the setting for 550 °C is high enough to have transformed the silicide into its lowest resistance phase ($CoSi_2$). Using these new settings with lower ramp rates, excellent silicide results with no voids or climbing were achieved (Figure B-16).

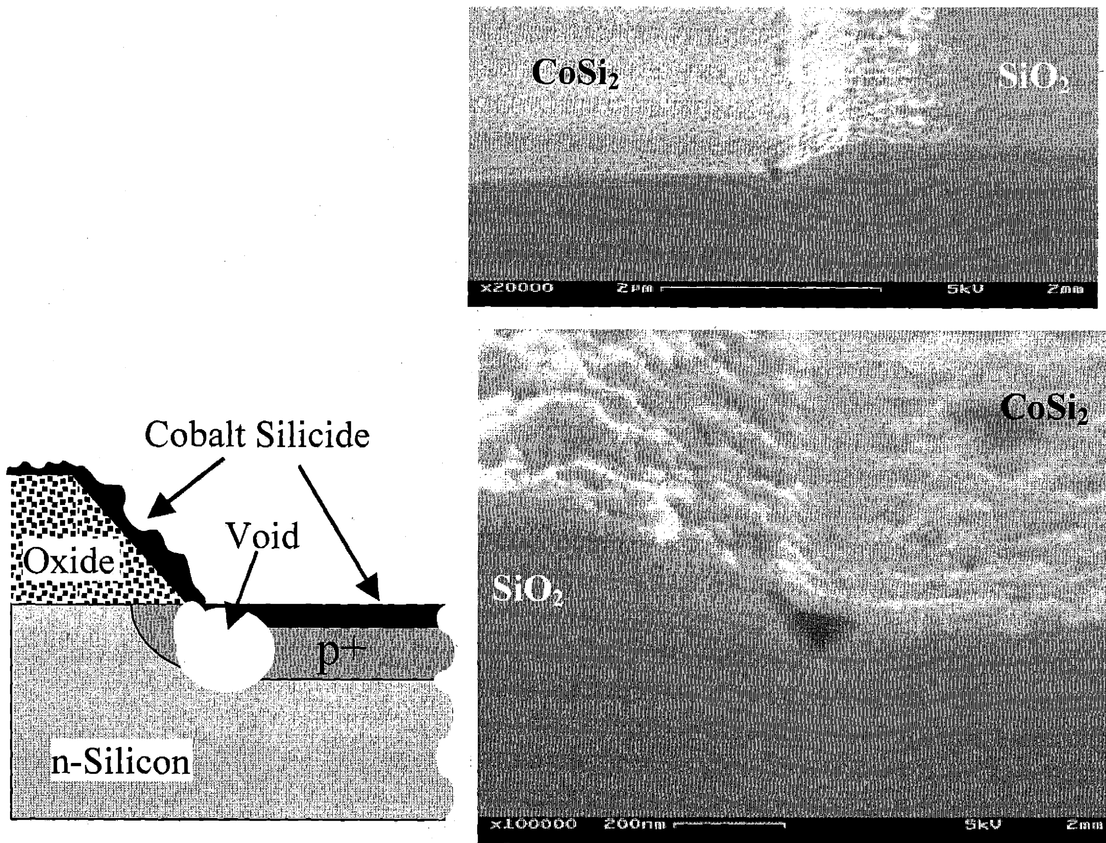


Figure B-13: Severe voids and climbing of silicide onto the isolation

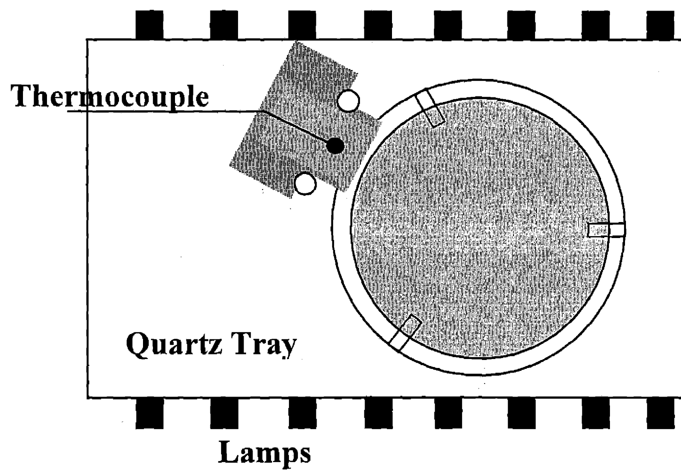


Figure B-14: Top down schematic of the RTA system.

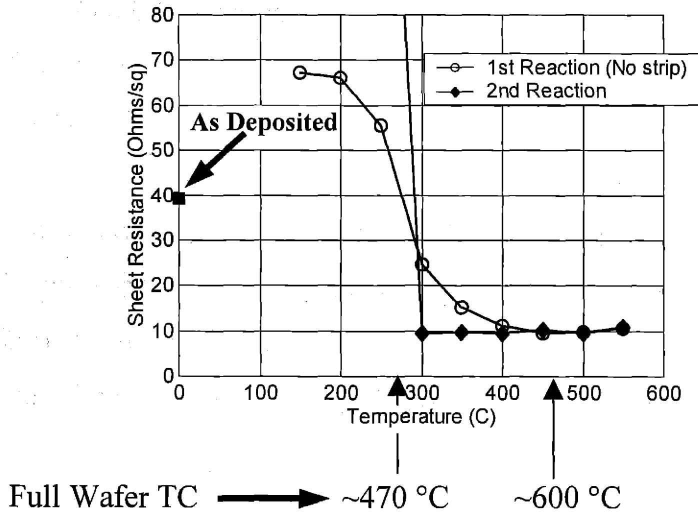


Figure B-15: Measured silicide sheet resistance versus temperature setting (equivalent to the thermocouple reading) on the RTA.

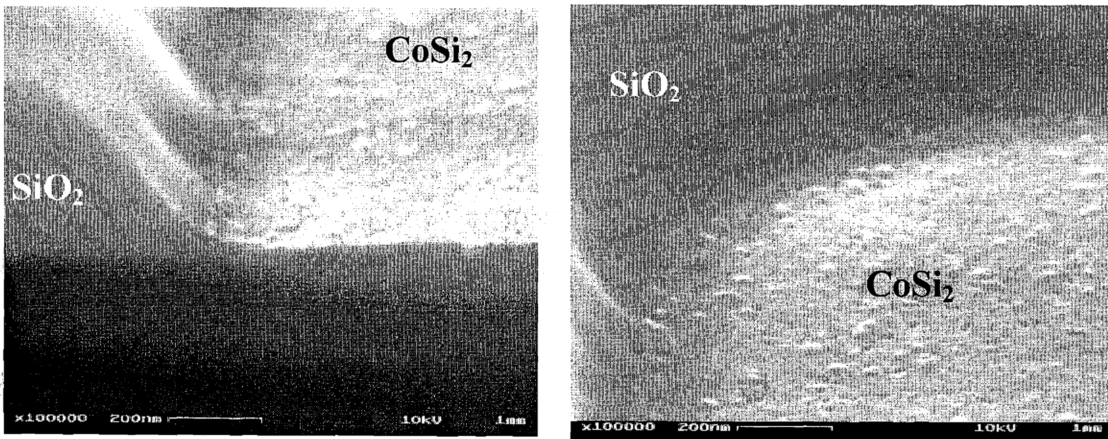


Figure B-16: Cobalt silicide results at the adjusted temperature settings. No climbing or voids are visible.

B.13 Contact Cuts

The etch process for the contact cuts was identical to the one used for the active area etch (Table B-2). The sloped sidewalls preserve the continuity of the metal and stopping with about 100 Å of oxide left again protects the source/drain surface. This oxide was removed in 50:1 HF during the pirhanna pre-metal clean right before loading into the e-beam for metal deposition.

The contact cuts play a crucial role in the use of the Ti barrier layer with Aluminum metalization. Without the oxide layer and contact cuts, it is possible for the edge of the Ti/Al metal stack to land on silicon. In this case the Al can easily diffuse over the edge of the Ti and spike into the silicon and short the source or drain junction.

B.14 Metal Deposition and patterning

The metal contacts for this lot used a 1000 Å Ti layer next to the silicon and 1 μm of Al on top, both deposited in an e-beam evaporator at pressures in the 10^{-7} Torr range. The titanium acts as a diffusion barrier to the aluminum and prevents it from spiking into the silicon and shorting the junctions of the source or drain.

The backside of the wafers were also metalized with 1μm of aluminum to create a low resistance backside contact (Al spiking is okay here since the entire substrate is of the one doping type). In order to be able to pattern the front-side aluminum using a wet etch, 1000 Å of titanium was deposited on top of the 1μm of backside aluminum. The Pan-Etch solution used to wet etch aluminum attacks the titanium very slowly, allowing the front-side aluminum to be patterned without etching the backside aluminum.

Finally, to wet etch the 1000 Å of titanium on the front-side of the wafer, a dilute 200:1 H_20:48% HF solution was used that gave a Ti etch rate around 50 Å/s, Al etch rate of < 30 Å/s, $CoSi_2$ etch rate of ~15 Å/min, and thermal oxide etch rate of ~30 Å/min. A delay of about 15 seconds occurred before bubbles appeared on the wafer indicating that the titanium was etching. Previous experience showed that the faster rates of more concentrated HF solutions gave very uneven etching as bubbles formed and protected the metal in spots across the wafer.

B.15 Device Results

Measurements of the devices show that they work well and do not suffer from the gate to source/drain leakage that the initial devices exhibited. The increase in the thickness of the re-oxidation and adjustment of the wafer cleans seem to have solved that problem. Figure B-17 shows the results for a L_{eff} =0.2 μm, W=22 μm PMOS device.

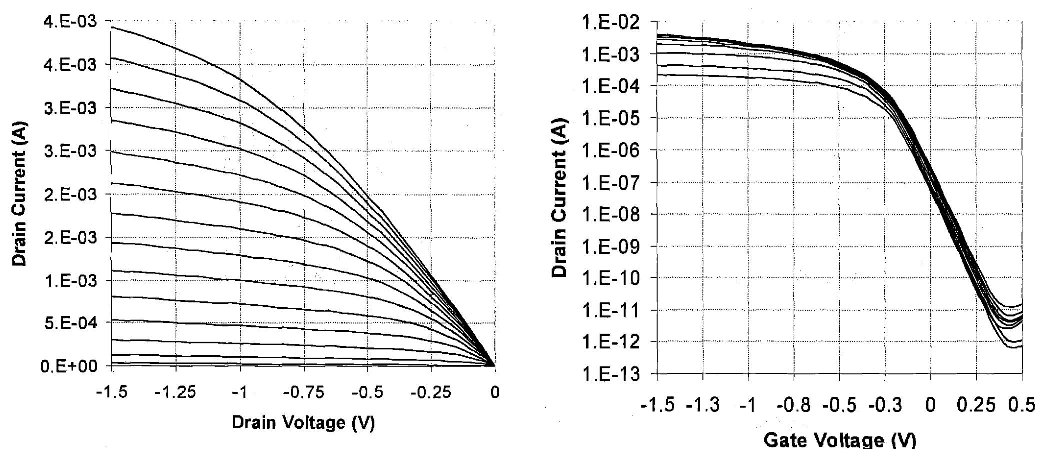


Figure B-17: Measured data from a L_{eff} = 0.2 μm, W=22 μm PMOS device.

Although the devices work well, a couple major setbacks occurred due to the gate etch profile. As mentioned in section B.6.3.1.1, the gates ended up having long sloping tails due to the initially poor shape of the hardmask (Figure B-18). These tails cause the channel length to be much longer that expected. The shortest devices are around 0.2 μm. In addition, the tails were thick enough to stop the shallow source/drain extension implants from reaching the silicon substrate, but not thick enough to stop the halo implants. This extra length of halo doping significantly raised the threshold voltage of the devices, with the lowest V_{th}s around 0.4 V.

With tuning of the hardmask process, this fabrication technology should produce far shorter working devices.

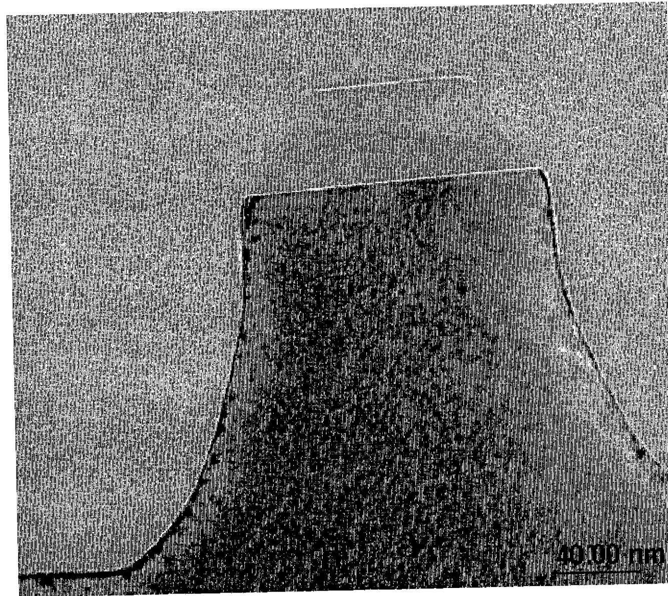


Figure B-18: TEM picture of a polysilicon gate directly after the gate etch. Note the similarly sloped shape of the nitride hardmask and polysilicon gate.

B.16 One-Level MOSFET Critique

One of the many concepts the final set of devices explored was the idea of creating one mask level MOSFETs (Figure B-19, Figure B-20). The concept was to eliminate the isolation step and backend contact cuts and metalization. By fabricating structures with ring-like geometries, only the core part of the MOSFET process would need to be performed to gain single measurable devices, and by eliminating excess process steps, significant time could be saved. In theory, such a quick turnaround structure could give at least measurements of the subthreshold characteristics and allow easy device prototyping.

From the vantage point of measuring the fabricated devices, it is clear that this concept does not make sense in a university research lab for such complicated devices. The largest fabrication time and effort correlates with the tightest tolerance part of the process which relates to the actual MOSFET fabrication from gate stack through silicide. Thus eliminating the isolation, contact cuts and metalization does not save that much time.

Even more importantly, the ring geometries with large polysilicon pads limits the measurements that are easily taken. The standard measurements of capacitance needed for device characterization and mobility extraction are made extremely difficult by the large parasitic capacitance of the surrounding gate and polysilicon pad which sit on the gate oxide.

The direct probing of silicide to measure the devices does work, but very large contact resistances (80-100 Ω total) were encountered which prevents accurate measurement of the devices' on-current. Applying a large current through the probe tips before measuring the devices reduced the contact resistance suggesting a thin oxide layer may have existed on top of the silicide, possibly leftover titanium oxide from the silicide reaction. A short HF dip as a final step would probably help make the probing easier.

It is worth noting that a one mask level approach has been found useful for very large devices to evaluate material properties of strained silicon [34].

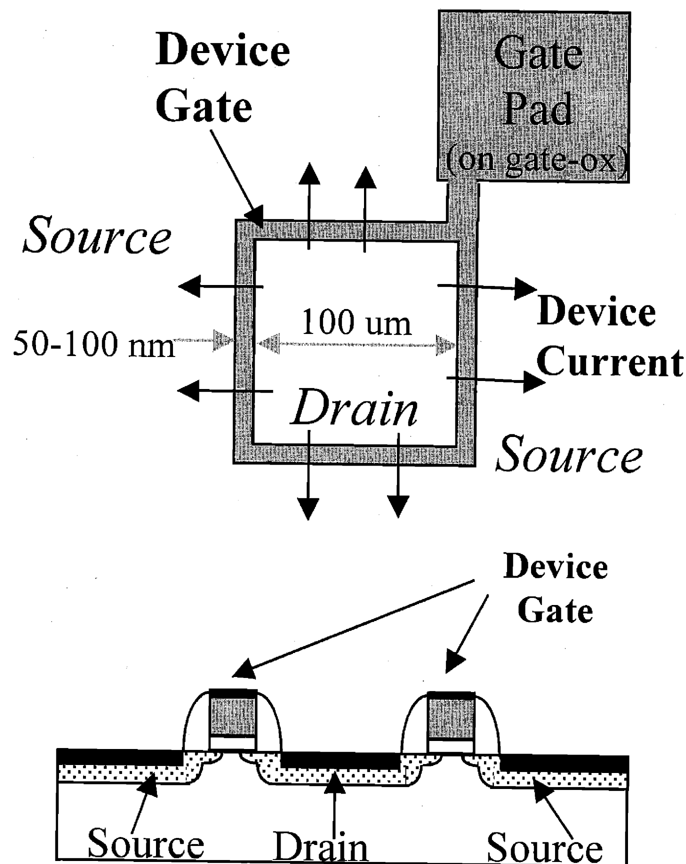


Figure B-19: Ring design one mask level MOSFET, top view and cross-section.

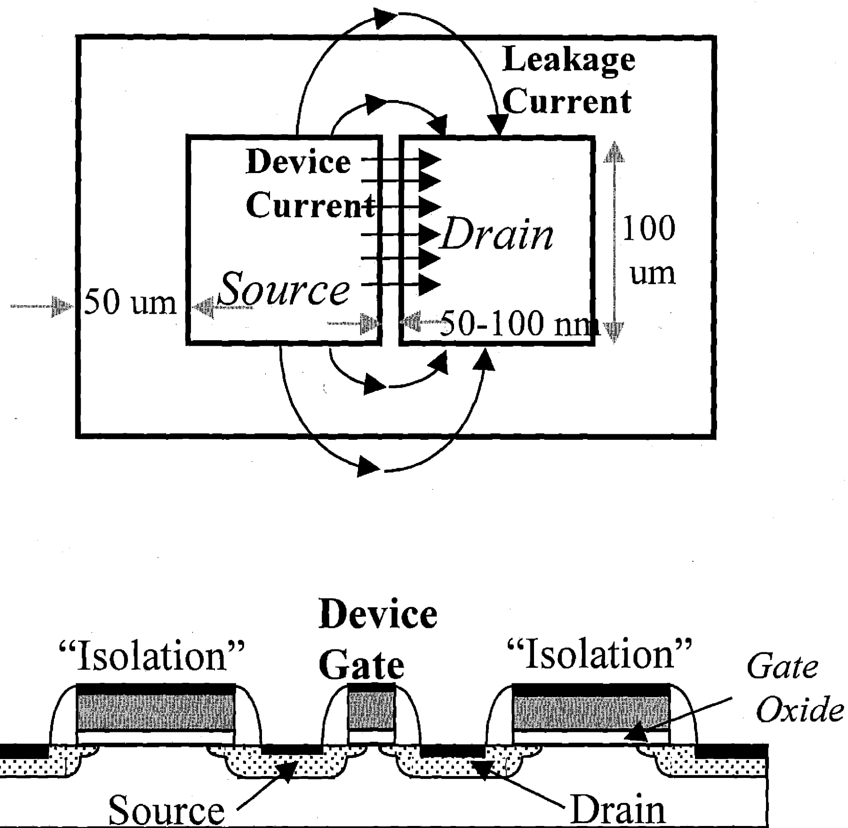


Figure B-20: Figure-8 design of the one mask level MOSFET, top view and cross-section.

References

[1] TOMCAT Manual, UT Austin.

[2] TSUPREM4, Avant! Corporation.

[3] MEDICI 1999.2, Avant! Corporation.

[4] H. Hu. *Experimental Study of Electron Velocity Overshoot in Silicon Inversion Layers*. MIT EECS Doctoral Thesis, 1994.

[5] Semiconductor Industry Association, *The International Technology Roadmap for Semiconductors*, 1997.

[6] A. Agarwal, et al. "Effect of Ramp Rates During Rapid Thermal Annealing of I_{on} Implanted Boron for Formation of Ultra-Shallow Junctions." *Journal of Electronic Materials*, Vol. 38, No, 12, 1999, pp. 1333-1339.

[7] P. H. Singer. "Dry Etching of SiO_2 and Si_3N_4." *Semiconductor International*, May, 1986, pp. 98-103.

[8] P. C. Karulkar and M. A. Wirzbicki. "Characterization of etching of silicon dioxide and photoresist in a fluorocarbon plasma." *Journal of Vacuum Science and Technology B*, vol. 6, 1988, pp. 1595-1599.

[9] A.G. Nagy. "Sidewall Tapering in Reactive Ion Etching." *Journal of the Electrochemical Society: Solid-State Science and Technology*, vol. 132, 1985, pp. 689-693.

[10] D. A. Buchan. "Scaling the gate dielectric: Materials, integration and Reliability." *IBM Journal of Research and Development*, Vol. 43, No. 3, May 1999, pp.245-264.

[11] Personal Communication with Dr. Doug Buchanan (IBM), Andy Fan (MIT), and Isaac Lauer (MIT).

[12] I. Y. Yang. *Study of Sub-0.5 μm SOI-with-Active-Substrate (SOIAS) Technology for Ultra-Low Power Applications*. MIT EECS Doctoral Thesis, 1996.

[13] L. Y. Tsou. "High Rate and Selective Reactive Ion Etching of Polysilicon" *Dry Processing for Submicrometer Lithography*, SPIE Vol. 1185, 1989, pp. 110-114.

[14] Dr. Tony Lochtefeld - fabrication process.

[15] Dr. Andy Wei and Andy Ritenour personal communication

[16] T. J. Dalton. *Pattern Dependencies in the Plasma Etching of Polysilicon*. MIT Chemical Engineering Doctoral Thesis, 1994.

[17] A. M. El-Masry et al. "Magnetically Enhanced Reactive I_{on} Etching of Silicon in Bromine Plasmas." *Journal of Vacuum Science and Technology B*, Vol. 6, No. 1, Jan/Feb 1988, pp. 257-262.

[18] T. D. Bestwick and G. S. Oehrlein. "Reactive Ion Etching of Silicon Using Bromine Containing Plasmas." *Journal of Vacuum Science and Technology A*, Vol. 8, No. 3, May/June 1990, pp. 1696-1701.

[19] Takeshi Hattori, ed. *Ultraclean Surface Processing of Silicon Wafers: Secrets of VLSI Manufacturing.* (1998). Pp. 451 & 482.

[20] R. Mark Hall, et al. "Effect of SC-1 Process Parameters on Particle Removal and Surface Metalic Contamination." *Mat. Res. Soc. Symp. Proc.*; Vol 386, p. 127 (1995).

[21] K. K. Christenson, et al. "Effects of SC-1 Dilution and Temperature Variations on Etch Rate and Surface Haze." *Mat. Res. Soc. Symp. Proc.*; Vol 386, p135, (1995).

[22] Werner Kern. "The Evolution of Silicon Wafer Cleaning Technology." *Semiconductor Cleaning Technology*, p. 3 (1989).

[23] Sadao Adachi and Katsuyuki Utani. "Chemical Treatment Effect of Si Surfaces in $NH_4OH:H_2O_2:H_2O$ solutions studied by spectroscopic elipsometery." *Japanese Journal of Applied Physics*, L1189 (1993).

[24] Takahiro Suzuki and Sadao Adachi, "Chemical Treatment Effect of Si Surfaces in 1 NH_4OH : X H_2O_2 : X H_2O solutions studied by spectroscopic elipsometery." *Japanese Journal of Applied Physics*, p. 2689 (1994).

[25] D. Graf et al. "Influence of HF-H202 treatment on Si (100) and Si(111) surfaces." *Journal of Applied Physics*, p. p. 1679 (1993).

[26] Kobayashi et al, *Japanese Journal of Applied Physics,* Vol. 33 (1994) L15.

[27] Kazuyuki Kobayashi, et al. "Chemical Treatment effect of Si(111) surfaces in H2SO4:H202 solution." *Japanese Journal of Applied Physics*, p.5925 (1996).

[28] Chemistry of the Semiconductor Industry (book) pg 252.

[29] S.D. Houssain, *Extended Abstracts of the 1993 Spring meeting of the Electrochemical Society* (vol 93-1) pp. 787-788.

[30] Kirt R. Williams, et al. "Etch Rates for Micomachining Processing." *Journal of Microelectromechanical Systems*, Vol. 5, No.4, p. 256 (1996).

[31] Mitsushi Itano, et. al. "Particle Deposition and Removal in Wet Cleaning Processes for ULSI Manufacturing." *IEEE Transactions on Semiconductor Manufacturing*, Vol. 5, No. 2, p. 114 (1992).

[32] Mitsushi Itano, et. al. "Particle Removal from Silicon Wafer Surface in Wet Cleaning Process." IEEE Transactions on Semiconductor Manufacturing, Vol. 6, No. 5, p. 258 (1993).

[33] M. J. Sherony. *Design, Process, and Reliability Considerations in Silicon-on-Insulator (SOI) MOSFETs.* MIT EECS Doctoral Thesis 1998.

[34] M. A. Armstrong. *Technology for SiGe Heterostructure-Based CMOS Devices.* MIT EECS Doctoral Thesis 1999.