HOT CARRIER EFFECTS IN CMOS FIELD EFFECT TRANSISTORS AT CRYOGENIC TEMPERATURES

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING

AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

By

Albert Karl Henning

Stanford University

August 1987

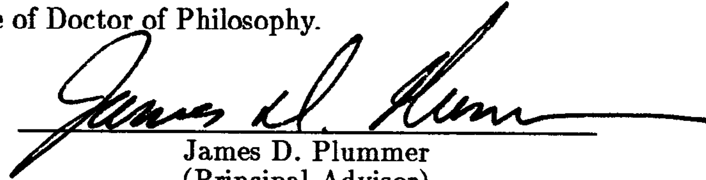
© Copyright 1987

by

Albert Karl Henning

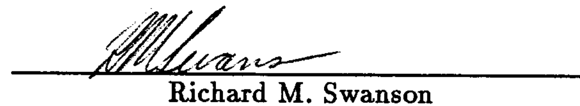
Stanford University

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



James D. Plummer
(Principal Advisor)

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

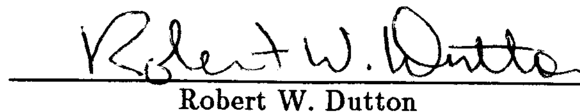


Richard M. Swanson

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

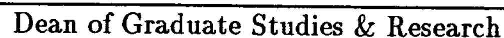
Malcolm R. Beasley

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



Robert W. Dutton

Approved for the University Committee on Graduate Studies:



Dean of Graduate Studies & Research

Abstract

Silicon is the material of choice for fabrication of high circuit density, low defect density and high speed integrated devices. CMOS technology provides the additional advantage of low power dissipation. These features make CMOS technology an attractive candidate to take advantage of the performance enhancements available through liquid nitrogen temperature operation. However, low temperature operation may increase the hot carrier generation of both substrate and gate currents - which can degrade device performance and reliability.

This dissertation begins briefly with an overview of the advantages and drawbacks of cryogenic device operation. The focus then shifts to hot carrier effects, since they prove detrimental to operation at both normal and cryogenic temperatures. In particular, characterization of the temperature, channel length, and voltage dependences of the weak avalanche substrate current between 77K and 300K will be presented. A microscopic, physical model based on Shockley's lucky electron approach will be described which explains this impact ionization behavior. The model incorporates a Maxwell-Boltzmann distribution of hot carrier energies beyond the band minima, and is implemented in the 2-D device simulators CADDET and PISCES. Specific tools have been developed in PISCES for analyzing hot carrier effects, using the results of this model.

Device gate current in short-channel NMOS FET's is also characterized at low temperatures and realistic biases. The measurements have implications for gate current modelling, device reliability, and reliability modelling. These implications are discussed in detail, and specific, quantitative suggestions are made on the necessary attributes for a 2-D gate current model.

Acknowledgments

As with every endeavor, this one was accomplished with the help and support of many people, and such assistance merits acknowledgement. First and foremost, then, I wish to thank my wife Carol Blue Muller, and our children Kaethe Blue Henning and Scott Anders Henning. The joy they have brought to my life and work I shall treasure and savor as long as I have a thinking mind. I look forward to our future with excitement, knowing that I am fortunate enough to experience it with them.

To my advisor, Jim Plummer, I offer my gratitude for taking me on as a graduate student in 1982. His decision has played a crucial and positive role in my career, in terms of both his technical and ideological leadership. In providing his students with superb tools to explore important technical areas, he achieves directly the dual purpose of challenging good minds, and adding to the critical store of knowledge in these areas. By doing so with good humor and a light but challenging touch, he achieves the indirect purpose of being a fine role model. His tolerance for extracurricular activity is especially noteworthy, as it humanized the graduate student experience. One could hardly ask for more in a person; I will look forward to working with him in the future, and passing along all I have learned.

Nelson Chan of Intel Corporation collaborated with me during the earlier efforts to formulate an impact ionization model for MOSFET simulation, using the simulator CADDET available to him at Intel. The experience he brought to bear on the problem saved much of the usual time and effort that might otherwise have been expended. I am also grateful for the personal example he set when his health was less than perfect. His inspiration will also be a guide for the future.

My graduate associates, Jason Woo and Jeff Watt, provided a challenging forum for discussion of ideas in the area of low temperature device physics. Jason's processing assistance

in our joint experiment helped assure its timely completion, and his questions and suggestions helped hone the idea of an energy distribution as applied to hot carrier effects. Jeff also brought great insight, particularly into the area of MOSFET mobility. His measurement software provided a solid basis for the gate current work. And, his forays into PISCES provided a spur to grapple with that large program on my own. I am very appreciative of their ideas, skills, comments and criticisms. And again, it would be my great pleasure to work with them in the future.

It is unfortunate that the assistance of the Integrated Circuits Laboratory staff has become so routine as to go almost without mention. Yet, without their superb skills and experience the work of the lab would amount to little, indeed. Dr. Terry Walker and Dr. Ernie Wood assisted with the computer details of the mask generation for my experiments. Dr. Dave Dameron and Paul Jerabek provided the masks themselves. Dr. Jim McVittie, Dr. John Shott, Nancy Latta, Margaret Prisbe, Robin King, Marnel King, Sherrie Bernard, Len Booth and Jibreel Mustafa provided much assistance with the actual processing, mostly in terms of advice on errors to avoid. Danny Roman, Thorwald Van Hooydonk, and Luke Meisenbach were essential in maintaining the equipment, all the while 'maintaining' a good sense of humor, as well. Joyce Pelzl provided invaluable office staff support and expertise which I shall miss, and for which I am most grateful.

Several professionals in the field, outside of Stanford, fulfilled the roles of mentor or peer at various times. Dr. Boaz Eitan gave encouragement for returning to the academic arena after a stint at Intel, and insight into experimental aspects of hot carrier effects. Dr. Fritz Gaensslen suggested the floating gate measurement technique for MOSFET gate current, and offered encouragement for the prospects of cryogenic device research. Mr. Paul Heremans and Dr. Nelson Saks spoke with me at length about their experiences with the floating gate technique, which were very helpful in applying the technique to cryogenic temperatures. Dr. Pat Dishaw offered his friendship as well as technical criticism and advice, for all ofwhich I shall be ever grateful.

Colleagues at Stanford gave material help along the way. Conor Rafferty, Mark Pinto, Hal Yeager, Mark Law, Doreen Cheng, and Dr. Chang-Gyu Hwang gave guidance on the intricacies of the two-dimensional device simulator PISCES, and on its strengths and foibles. They pointed out the right direction to go when the next step was obscure. Dr.

Gary Patton's temperature control software saved time and effort in the substrate current characterization. Discussions with Dr. Sergio Bampi on the gate current measurements, and expectations for them, were invaluable. Dr. Mike Reed's plotting software saved time with the gate current measurement analysis. His unusual sense of humor kept me out of mental ruts at various stages. Dr. Stan Swirhun collaborated on an interesting experimental diversion, looking at light emission from MOS transistors. Professors Dick Swanson, Bob Dutton, Jim Meindl, Jim Harris, Dave Bloom, Tony Siegman, Walter Harrison, Malcolm Beasley, and Malcolm McWhorter set the standards of excellence in the lab and in the classroom. Their examples will be difficult - but challenging nonetheless - to follow.

Contract support through the Joint Services Electronics Program is gratefully acknowledged.

Contents

Abstract v

Acknowledgments vii

1 Introduction 1

1.1 The case for silicon and CMOS 2

1.2 The case for low temperature 2

1.2.1 Advantages 3

1.2.2 Disadvantages 5

1.3 Objectives 7

1.4 Organization 7

1.5 Summary 8

2 Measurements of CMOS substrate current at cryogenic temperatures 9

2.1 Introduction 9

2.2 Motivation 9

2.2.1 Substrate current and reliability 10

2.2.2 Substrate current and gate current 10

2.3 General characteristics 12

2.3.1 Substrate current at low V_D 13

2.3.2 Temperature dependence of I_B : V_{xover} 15

2.3.3 Relation of V_{xover} and reliability 16

2.4 Fabrication and electrical parameters 17

2.4.1 Test pattern layout and mask-making details 17
2.4.2 Processing details and process characteristics 20
2.5 Characterization methodology 23
2.6 Measurement results 26
2.6.1 Drain current characteristics 26
2.6.2 Substrate current characteristics 28
2.6.3 Voltage crossover 29
2.6.4 Temperature effects 31
2.7 Summary 40

3 Impact ionization in silicon 41
3.1 Introduction 41
3.2 Definition of terms 41
3.3 Questions of physics 42
3.4 Historical review 43
3.5 The energy of a carrier 47
3.5.1 Derivation of the energy distribution 47
3.5.2 The average carrier temperature 51
3.6 Recapitulation 53
3.6.1 Assumptions in energy distribution 53
3.6.2 Assumptions in carrier temperature 55
3.6.3 Failings of previous models 56
3.7 Summary 56

4 Modelling impact ionization in CMOS FET's 57
4.1 Introduction 57
4.2 Historical review 58
4.2.1 Macroscopic models 59
4.2.2 Microscopic, local models 61
4.2.3 Microscopic, non-local models 62
4.3 A new, comprehensive, and efficient model 66
4.3.1 Description 66

4.3.2 Implementation 71
4.4 New model performance 73
4.4.1 CADDET results: I_B simulation 73
4.4.2 CADDET results: I_{BMAX} simulations 75
4.4.3 V_{xover} versus channel length 75
4.4.4 PISCES results: I_B simulation 80
4.4.5 Model parameter sensitivities 90
4.4.6 Ionization contours 98
4.4.7 Comparison of energy distribution functions 102
4.5 The effects of low temperature 102
4.6 λ, T_e, and the ionization threshold 106
4.6.1 Optical phonon mean free path 106
4.6.2 Impact ionization mean free path 107
4.6.3 λ temperature dependence 108
4.6.4 Inversion layer effect 109
4.6.5 The average carrier temperature, T_e 109
4.6.6 The ionization threshold 110
4.7 The new tools and device physics understanding 111
4.8 Summary 113

5 Gate current in NMOS FET's at low temperatures **116**
5.1 Introduction 116
5.2 Definition of terms 117
5.3 Questions of physics 120
5.4 Historical review and general characteristics 120
5.5 Gate current characterization methodology 127
5.6 Measurement results 131
5.6.1 Measurement specifics 133
5.6.2 Reproducibility: interface degradation 136
5.7 Modelling implications 138
5.7.1 Auger vs. thermionic emission 138

5.7.2 Drain avalanche hot carriers 139
5.8 Gate current models 140
5.8.1 Previous I_G models 140
5.8.2 Proposed I_G model 143
5.9 Summary 154

6 Conclusions and Recommendations **156**
6.1 Conclusions 156
6.2 Recommendations 158

7 Appendix: PISCES usage guide for impact ionization **161**
7.1 Introduction 161
7.2 Mesh generation and bias application 161
7.3 Known bugs 164
7.4 Example 165
7.5 New PISCES card and existing card changes 165

8 Appendix: Mobility models for 2-D impact ionization simulation **168**
8.1 CADDET mobility model 168
8.2 PISCES mobility model 170

9 Appendix: Thermionic emission probability **173**
9.1 Introduction 173
9.2 Derivation 174
9.3 Discussion 175

References **179**

List of Figures

1.1 Comparison of speed-power product for various IC technologies 3
1.2 Schematic of impact ionization processes in MOSFET's 7
2.1 Correlation of N-channel substrate current and reliability 11
2.2 Correlation of P-channel I_B and I_G with reliability 12
2.3 Standard substrate current characteristic. 13
2.4 2-D potential distribution: Region II 14
2.5 2-D potential distribution: Region III 14
2.6 Observation of I_B for V_{DS} less than the ionization threshold 15
2.7 Initial characterization of crossover voltage for MOSFET 16
2.8 Correlation of V_{xover} and device reliability 17
2.9 The layout of a typical PMOS transistor 19
2.10 The layout of a typical CMOS inverter 20
2.11 CMOS process cross-section 21
2.12 NMOS source-drain profile. 22
2.13 PMOS source-drain profile. 22
2.14 N-channel threshold voltage 23
2.15 P-channel threshold voltage 24
2.16 N-channel doping profile 24
2.17 P-channel doping profile 25
2.18 Example of the transconductance monitor 26
2.19 Example of the high-field I_D measurement 27
2.20 Generation efficiency of impact ionization process in MOSFET. 28
2.21 Normalized peak substrate current vs. temperature 29

2.22 Crossover voltage versus channel length for N-channel MOSFET's 30
2.23 I_D vs. temperature for N-channel devices with V_D=5V and V_G=5V 32
2.24 I_D vs. temperature for N-channel devices with V_D=50mV and V_G=5V . . . 33
2.25 I_D vs. temperature for P-channel devices with V_D=-5V and V_G=-5V 34
2.26 I_D vs. temperature for P-channel devices with V_D=-50mV and V_G=-5V . . 35
2.27 Threshold voltage for 25/1.15 N device, versus temperature. 36
2.28 Threshold voltage for 25/1.17 P device, versus temperature. 36
2.29 Ring oscillator gate delay 38
2.30 Ring oscillator gate delay, normalized to the 294K value 39
3.1 Baraff's ionization rate curves 45
3.2 Ionization rate temperature dependence 46
3.3 Single-particle impact ionization processes 48
4.1 Relation of current and field to I_B 58
4.2 Schematic of carrier energy distribution in a MOSFET channel 67
4.3 Schematic of model implementation in a PISCES mesh element 68
4.4 Flowchart for PISCES impact ionization model 72
4.5 I_B simulation vs. experiment: T=300K; W/L= 25/25N 74
4.6 I_B simulation vs. experiment: T=77K; W/L= 25/2.15N 74
4.7 I_B simulation vs. experiment: T=77K; W/L= 25/1.15N 75
4.8 I_B simulation vs. experiment: T=300K; W/L= 25/25P 76
4.9 I_B simulation vs. experiment: T=77K; W/L= 25/1.17P 76
4.10 Simulation of peak I_B : 1.15μm N 77
4.11 Simulation of peak I_B : 25μm N 77
4.12 Simulation of peak I_B : 2.15μm N 78
4.13 How a MB distribution explains V_{xover} 79
4.14 Simulation of V_{xover} versus L_e 80
4.15 Mesh used in PISCES simulations 81
4.16 PISCES I-V simulation: 25/25N, T=299K 82
4.17 PISCES I-V simulation: 25/25N, T=77K 83
4.18 PISCES I-V simulation: 25/0.8N, T=300K 84
4.19 PISCES I-V simulation: 25/0.8N, T=77K 85

4.20 PISCES I_B simulation: 25/25N, T=299K 86
4.21 PISCES I_B simulation: 25/25N, T=77K 87
4.22 PISCES I_B simulation: 25/0.8N, T=299K 88
4.23 PISCES I_B simulation: 25/0.8N, T=77K 89
4.24 Sensitivity of I_B simulation to E_{cut} : 25/0.8N, T=299K 90
4.25 Sensitivity to number of current contours: 25/0.8N, T=299K 92
4.26 Sensitivity to surface reduction factor: 25/25N, T=299K 93
4.27 Sensitivity to surface reduction factor: 25/25N, T=77K 94
4.28 Sensitivity to surface reduction factor: 25/0.8N, T=299K 95
4.29 Sensitivity to surface reduction factor: 25/0.8N, T=77K 96
4.30 Sensitivity to λ_0 at low T: 25/0.8N, T=77K 97
4.31 Schematic of α in PISCES element 99
4.32 α contour: V_G=5V, T=300K 100
4.33 α contour: V_G=2V, T=300K 101
4.34 α contour: V_G=5V, T=77K 101
4.35 α contour: V_G=2V, T=77K 102
4.36 2-D potential distribution along a current contour 103
4.37 Comparison of models: energy distributions 104
4.38 1-D ionization rate and electric field comparison: T=299K 106
4.39 1-D ionization rate and electric field comparison: T=77K 107
4.40 Current contours: T=300K, V_D=5V, V_G=3V 112
4.41 Electric field contours: T=300K, V_D=5V, V_G=3V 112
4.42 α contours: T=300K, V_D=5V, V_G=3V 113
4.43 Potential contours: T=300K, V_D=5V, V_G=3V 114
5.1 Schematic of the N-channel I_G vs. V_G characteristic 117
5.2 Schematic of the Auger emission mechanism in a MOSFET 118
5.3 CHC and DAHC schematic definitions 119
5.4 Gate current in N-channel MOSFET. 122
5.5 Lateral electric field vs. channel position 123
5.6 Gate current for very short, N-channel MOSFET. 124
5.7 Floating gate I_G measurement 124

5.8 I_G vs. Auger probability 125
5.9 P-channel gate current 126
5.10 Floating gate I_G at very low levels 126
5.11 Schematic of the external floating-gate measurement technique 128
5.12 Example of g_m measurement for I_G extraction 129
5.13 Typical I_D vs. time characteristic 130
5.14 Schematic of various I_G vs. time regimes 131
5.15 I_G for V_D=2.5V 132
5.16 I_G for V_D=4.5V 133
5.17 CHE gate current vs. temperature 134
5.18 Extraction of gate voltage and current from I_D vs. time 136
5.19 Order dependence of gate current measurement 137
5.20 Dependence of gate current on interface degradation 138
5.21 Gate current vs. Auger probability (process 1) 139
5.22 Schematic of previous I_G model 141
5.23 Two-dimensional schematic of CHE gate current 145
5.24 Band diagram under CHE bias conditions 145
5.25 Two-dimensional schematic of DAHC gate current 146
5.26 Band diagram under DAHC bias conditions. 146
5.27 Optical phonon scattering probability 151
7.1 PISCES input deck for mesh generation 163
7.2 Listing of PISCES job deck for I_B simulation 166
7.3 Manual page for new **IMPACT** card 167
8.1 Schematic of mobility extraction in PISCES 171
9.1 Band diagram for thermionic emission 176
9.2 Emission probability vs. large, normalized barrier height 177
9.3 Emission probability vs. small, normalized barrier height 178

Chapter 1

Introduction

Nearly half a century of work on semiconductor devices has produced a vigorous industry, at once pioneering and mature. Since the seminal papers on the unipolar, inversion layer transistor by Lilienfeld [Lili 30], the bipolar junction transistor and $p-n$ junction theory by Bardeen and Brattain [Bard 48] and Shockley [Shoc 49], and the junction field-effect transistor by Shockley [Shoc 52], researchers have pushed manufacturable silicon technology to submicron dimensions. Levels of transistor integration have reached greater than 16 million active components for memory devices [Mano 87]. More remarkable, these advances have been accomplished with chip sizes of roughly one square centimeter or less. Integration of gallium arsenide devices continues to drive the technology learning curve, too, with the report of 4Kb GaAs MESFET SRAM's [Taka 87]. In essence, the search for semiconductor devices and technologies, which give at once faster device speed and smaller device size - greater performance at cheaper cost - continues with increasing force and power.

But this search has also demonstrated fundamental limits to device performance and integration density [Mein 83]. Systems designed with fixed, five volt power supplies lead to increased device [Take 85] and interconnect [Gard 87] reliability problems as scaled dimensions [Denn 79,Sara 82] are implemented. Because of these limits, designers of devices, technologies, and systems are looking to alternative means to achieve both higher performance and lower cost.

1.1 The case for silicon and CMOS

Historically, silicon has led other semiconductor materials in this regard. An electrically and structurally stable oxide can be formed on Si for both device isolation and MOS gate formation [Nico 82], which is its principal distinction from other semiconducting compounds. With this advantage, Si technologies exceed their nearest competitor, GaAs, by almost three orders of magnitude in integration density.

Unlike GaAs, Si also offers two charge carriers with roughly the same mobility [Sze 81b], which have been used to create a complementary device technology [Whit 66] based on the unipolar MOS structure. These CMOS devices offer circuit advantages, such as more compact layout in dynamic random access memories, full-swing logic levels, reduced power consumption, and noise immunity.

A principal advantage for CMOS, however, is its power dissipation. Because the main component of digital circuitry is the inverter, and because the DC power drawn by a CMOS inverter for either a high or low input is zero, power consumption in static CMOS is eliminated when compared to the conventional NMOS technologies of the past. Dynamic operation of CMOS circuits of course requires power; but even so, the power consumption is still reduced for CMOS versus its NMOS cousin. Complementary HEMT technology [Ciri 85] may one day rival CMOS, but at this juncture lags CMOS in process stability, density, and cost.

1.2 The case for low temperature

Figure 1.1 shows a comparison between many technologies of interest, using inverter speed and power consumption as the quantities for comparison. Without optimizing the process for low temperature, cryogenic operation of Si CMOS nevertheless meets or exceeds Si bipolar (ECL) speeds, and approaches GaAs speeds. This performance is accomplished at greatly reduced power dissipation. When coupled with the greater thermal conductivity of Si at 77K, for instance [Sze 81a], one sees immediately that exceptionally high levels of integration can be achieved with a cryogenic CMOS process, vs. either Si ECL or GaAs technologies. When designed optimally for low temperature operation, both CMOS [Sun 87]

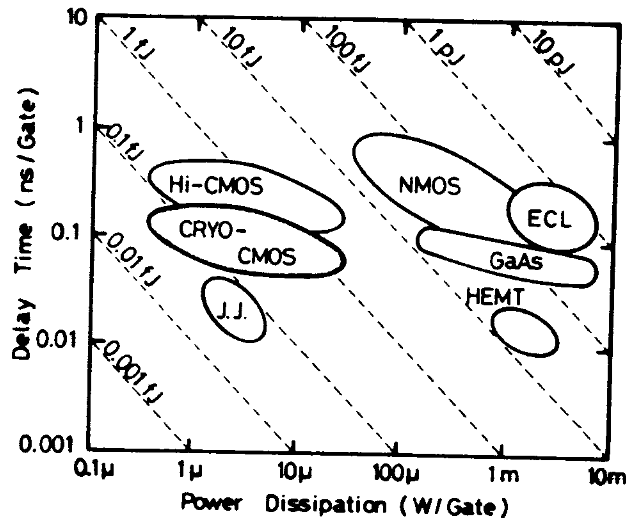


Figure 1.1: Comparison of speed-power product for various IC technologies. After [Hana 87].

and NMOS [Watt 87a] technologies offer significant improvements over the performance available from a conventional process, or from a room temperature process cooled to 77K.

1.2.1 Advantages

Gaensslen, *et al.* [Gaen 77], made the seminal suggestions for using VLSI MOSFET's at cryogenic temperatures. They measured and discussed such advantages as improved low-field mobility, saturated velocity, leakage, and metal conductivity. Tewksbury [Tewk 81] gave the first comprehensive characterization of N-channel device performance using the perspective of circuit modelling parameters like those used in SPICE [Quar 86]. His results showed traditional methods of circuit parameter extraction yielded meaningful results for temperatures down to 10K; the N-channel transistors behaved remarkably well, with no unaccountable behavior observed.

Drain-induced barrier lowering (DIBL) concerns device designers desiring to scale technologies [Trou 79]. DIBL has been shown to diminish at low temperatures [Woo 86], primarily because the barrier height at the device source, normalized to the lattice temperature, increases at lower temperatures. This makes the drain less likely to be a parasitic controller of the channel conduction current than at room temperature, for a given drain potential.

The subthreshold slope, characteristic of the exponential turn-on of the MOSFET, improves dramatically at low temperatures [Kamg 82b]. This feature implies the threshold voltage can be scaled downward without fear of encroaching on thermal noise margins. The lowered V_T then allows the power supply to be reduced with no loss in gate drive - while improving the device reliability through the reduced drain fields. A reduced subthreshold slope means channel leakages in DRAM transfer gates are reduced manyfold, to the extent that the circuits may be operated in a static mode at cryogenic temperatures [Ande 86]. For the same reason, clock circuit performance will improve with the reduced channel and junction leakage obtained at low temperatures. Finally, a reduced threshold implies reduced channel doping. This increases the carrier mobility in the channel at low temperature by reducing the charged impurity scattering component.

Improvements in metallic interconnect conductivity are expected [Watt 87c], which will reduce signal RC delays. Lowered metal resistance also means extra circuits - such as repeaters [Henn 83,Bako 85] - will be less necessary for the fastest transport of long-distance signals. Contact resistance is expected to improve as well [Swir 86].

CMOS latch-up is virtually eliminated by cryogenic operation [Dool 84]. The principal reason involves the large reduction in bipolar device gain for the dopings typically found in CMOSFET's, as the device temperature descends toward 77K. The shunt resistances also will decrease, improving the resistance to latch-up still further.

Because it is a temperature-activated process, electromigration is expected to decrease dramatically, allowing finer metal lines to be used for a given current density [Gard 87].

Circuit design and specific circuit advantages are evident. Designers of cryogenic system chips need not be concerned with ensuring their designs work in several temperature operating windows. In logic applications, low temperature NMOS technology can achieve the performance of room temperature GaAs MESFET technology for a roughly comparable channel length [Glor 86].

The recent, exciting news of full superconductivity at 94K in a yttrium-copper-barium oxide [Chu 87] gives rise to the immediate expectation that thin films of the new material can be fabricated [Robi 87]. Combined with the high speed and packing density of CMOS technology, one can imagine a *system* without the constraint of significant interconnection delays: very fast [Kwon 87], highly integrable, using a highly stable and reliable semiconductor process. However, the individual chips will still rely on traditional metallization, since the delay over long lines is device, not interconnect, limited [Watt 87c].

1.2.2 Disadvantages

As with any technology, there are drawbacks to cryogenic operation of high-performance circuits. Special system cooling is required [Long 87], which must be weighed against room temperature system costs. However, modern supercomputers such as the Cray 2 already require exotic cooling, which may make the cost of a liquid nitrogen-cooled system competitive [Iver 84].

The reduction of threshold voltage possible at low temperatures can lead to the problem of V_T control in submicron devices. Tolerances of plus/minus ten percent become more difficult to meet when V_T is less than or equal to 0.2V. Statistical variation of the dopant distribution in the submicron channel must also be considered a part of this tolerance, which will become more difficult to meet at reduced dimensions.

Besides V_T control, the ion implant techniques employed in modern device fabrication can lead to junction leakage currents higher than predicted by normal generation-recombination theory [Ande 86]. While not impacting normal device operation, this drawback may eliminate static operation of DRAM circuits at low temperatures.

Device reliability, however, is the main concern of those contemplating cryogenic devices. The underlying cause of device performance degradation - regardless of temperature - is the high energy carrier distribution developed through charge transport in modern, submicron transistors. The 'hot' carriers so developed give rise to the phenomena which this work seeks to examine.

Figure 1.2 shows a cross-section of an N-channel MOSFET, and provides a qualitative basis for understanding subsequent discussions of impact ionization processes in this dissertation. Substrate current in MOSFET's has been examined by Abbas [Abba 74]. Majority

carriers enter the high field region of the device channel. There, due to the high field and the probabilities for ballistic, or unscattered, motion over a certain distance, two phenomena may occur. First, the carriers may acquire enough energy to break a lattice bond in the semiconductor. In this case, the hole of the generated electron-hole pair typically travels toward the substrate contact, where it is collected and called substrate current, I_B. The electron is typically collected by the device drain. However, if the fields between the drain and gate are appropriate, the secondary carriers may be injected toward - and, if energetic enough, over - the semiconductor-insulator barrier. For this case, the charges may create interface states, or fill interface or bulk traps - either of which degrade the performance of the device.

Second, the channel carriers may be scattered toward the semiconductor-insulator interface themselves. If their energy is great enough, again they may surmount the interface barrier thermionically, and alter the interface or the insulator in the manner described above.

In both Si and GaAs unipolar FETs, source-drain breakdown (SDBD) [Kenn 73] is another possible outcome of the impact ionization outlined above, and a serious impediment to device design. Here, the secondary hole is collected, not at the substrate, but at the source contact. The potential change caused by this current flowing through the substrate resistance forward biases the source junction. This starts a regenerative feedback phenomenon by turning on the lateral, parasitic bipolar transistor inherent in any MOSFET. Additional holes collected at the source cause more forward-bias-injected electrons, which in turn create more holes through impact ionization - and the regenerative loop accelerates.

Low temperatures exacerbate all these phenomena. The scattering events decrease, the carriers travel farther between such events, and acquire more energy. The greater energy means increased probability of breaking a lattice bond and creating substrate current; or, of crossing the thermionic barrier at the interface, damaging it en route. An understanding, then, of the causes of these phenomena on a microscopic level can help in the reliable design of broad classes of devices, over wide bias and temperature ranges.

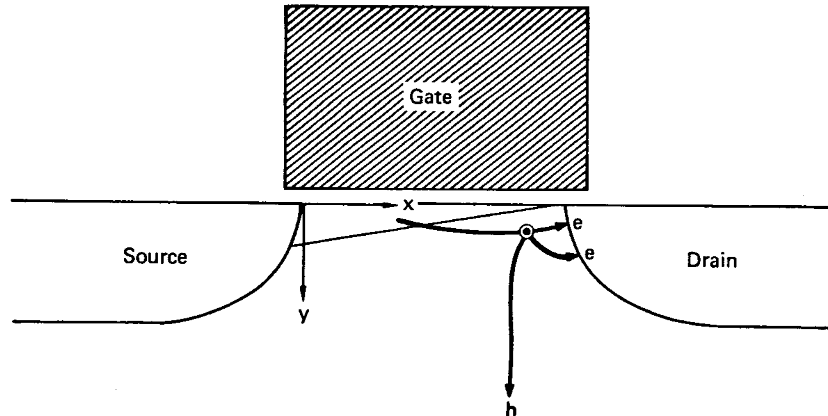


Figure 1.2: Schematic of impact ionization processes in MOSFET's

1.3 Objectives

This dissertation seeks to achieve several objectives. First, to explore the physical phenomena - impact ionization, substrate current, and gate current - which govern device reliability. Second, to characterize them over wide ranges of temperature and electrical bias. Third, to determine physical theories and models which can describe the observations. The clear intent is to develop an understanding on a microscopic scale, which may be incorporated in a two-dimensional (2-D) device simulator, and extended to devices of many designs. Such a model should be free of fitting parameters over any range of temperature or bias of interest, to insure its utility for all types of devices and technologies.

Temperature is used as a tool in this process, not only because of the device implications for cryogenic CMOS device operation, but because it can help isolate the most important physical causes for the observations.

1.4 Organization

Chapter Two presents the characterization of substrate current (I_B) in CMOS FETs, which includes a way of exploring device reliability expectations at low temperature by looking at peak I_B versus temperature. The important early works on impact ionization in bulk Si are detailed in Chapter Three, along with some modern treatments which explicate the

important physical processes on a microscopic scale. Using transport theory, the crucial concept of the carrier energy distribution is derived. Chapter Four provides the bulwark of this thesis, by exploring the history of I_B investigations in Si MOSFETs; the empirical and 2-D models which have been used to explain previous I_B observations; the reasons these models fail; a new, physical model which explains in microscopic detail all of the features of I_B; and a comparison of this model with measurements.

Gate current is the focus of Chapter Five, which explores previous I_G models and measurements, sets out the characterization of I_G , and discusses the implications of the measurements for a 2-D I_G simulator. Finally, Chapter Six reviews the important conclusions of the dissertation, and recommends further work in specific areas. Several Appendices discuss particular features of device measurement or simulation.

1.5 Summary

Cryogenic operation of CMOS circuits executed in silicon has been advocated as a candidate for high-speed computation. The advantages include increased speed due to improved device mobility and lower interconnect resistance; improved device packing density due to increased thermal conductivity and latch-up resistance; and, improved interconnection reliability due to increased electromigration resistance. The disadvantages include the potential for degraded device performance over time, due to the same phenomena which give rise to impact ionization, substrate current, and gate current in MOSFETs. This dissertation will explore these hot carrier phenomena in a systematic fashion, and demonstrate how they may be explained on a microscopic level. The results have implications for device physics and design, and are directly applicable to cryogenic technology development.

Chapter 2

Measurements of CMOS substrate current at cryogenic temperatures

2.1 Introduction

Motivating the interest in measuring and observing substrate current phenomena in CMOS FET's in general - and at low temperature in particular - is the initial task of this chapter. The general characteristics of substrate current are described qualitatively. The crossover voltage, V_{xover}, is defined, and proposed as a power supply limit for hot carrier reliability at cryogenic temperatures. Some of the fabrication details for the experimental devices - including layout and processing - are then given, though some discussion is reserved for the Appendices. Electrical parameters characteristic of the process are detailed. The measurement procedure is demonstrated, and the relevant results are presented. The summary enumerates the measurement characteristics which a successful model will need to explain quantitatively.

2.2 Motivation

The correlation between substrate current and MOSFET reliability is the outstanding reason for achieving a microscopic understanding of impact ionization. This understanding - combined with a microscopic picture of the insulator and interface physics in the MIS

system - can lead to full simulation of device reliability. Such a capability would reduce the cost and effort of developing reliable MIS processes, by allowing simulation to replace costly, time-consuming processing experiments.

As mentioned in Chapter One, low temperature device operation may exacerbate the generation of substrate current by impact ionization. This in turn correlates to possible increased device degradation [Take 83a,Hori 86a], under either static or dynamic stress. Reliable device design thus requires minimization of substrate current - making an accurate 2-D model of I_B extremely attractive. In particular, I_B is far easier to monitor than gate current, or the amount and location of trapped oxide charge - yet it provides a measure of device degradation caused by I_G, due in part to the correlation between I_B and I_G [Tam 82]. Once I_B is well-predicted, one can move toward the more complicated tasks of measuring and modelling I_G and insulator trapping and trap generation - toward a full, 2-D, DC device and reliability simulator.

2.2.1 Substrate current and reliability

Figure 2.1 shows a typical correlation between I_B and device relability. g_m is the device transconductance, and N_{ss} is the averaged surface state density. Changes in these parameters are used as measures of the reliability of a device - its resistance to the destructive mechanisms invoked by high field, DC bias stress. For this short-channel device, the clear implication is that I_B and reliability are correlated; however, the empirical nature of the analysis defeats understanding of the exact, microscopic mechanisms.

2.2.2 Substrate current and gate current

A more telling correlation is that between I_B and gate current, I_G. The transport of charge across the semiconductor-insulator interface is the direct cause of changes in the state of the interface. These changes are usually measured by either N_{it} or N_{ot}, the interface states or the oxide trapped charge, as defined in [Nico 82]. N_{it} is analogous to the N_{ss} of Figure 2.1. The development of interface charges is a direct result of the amount and energy of the charge transported to and across the interface, and the trapping cross-section of this charge with the interface states and traps. If the energy of the gate current carriers is large enough,

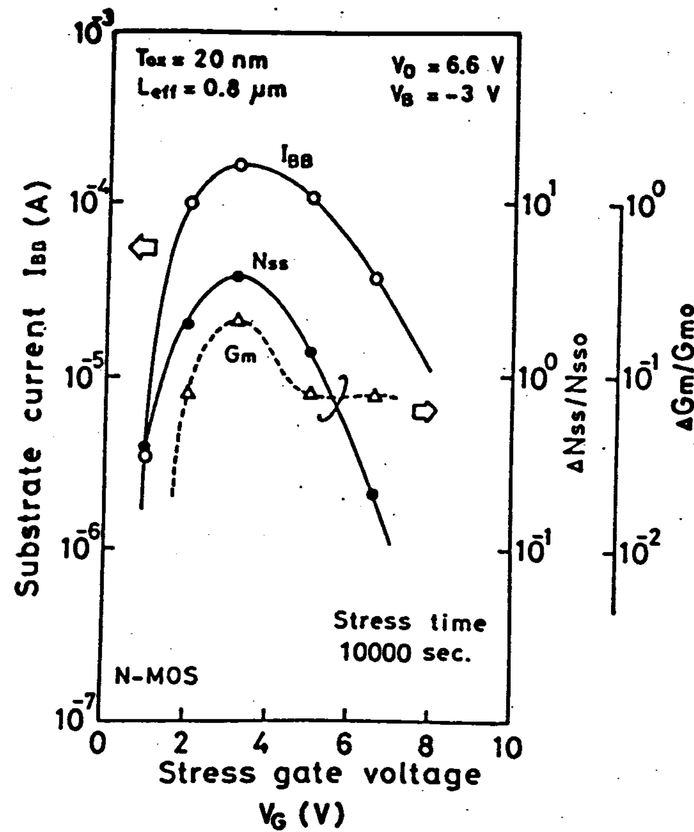


Figure 2.1: Correlation of impact ionization substrate current, transconductance degradation g_m and surface state density increase N_{ss}. I_B is the DC substrate current obtained before stress for the biases shown. The transconductance decrease and surface state density increase are measured after a 10^4 second DC stress at the gate bias shown. Such changes in device parameters degrade both DC and AC performance. After [Take 83b].

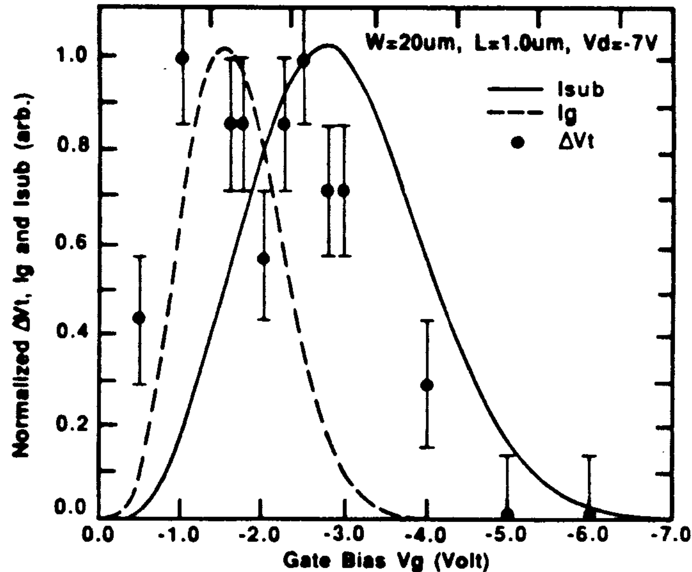


Figure 2.2: Correlation of substrate and gate currents with PMOS device reliability. The measured threshold shift after stress at the given gate and drain potentials can be seen to be due to trapped electrons in the insulator due to both gate and substrate currents. After [Bras 87].

states and traps can be created, as well.

Figure 2.2 shows the results from a recent study of I_B and I_G correlation. As can be seen, the correlation is not one-to-one. Nevertheless the similarities are marked enough to warrant the belief that understanding of substrate current will enhance that of gate current.

2.3 General characteristics

Figure 2.3 shows a standard I_B characteristic measured for an N-channel MOSFET with conventional, arsenic source-drain technology (see below for a full discussion of the process details). Here, W_e/L_e=25/0.85, in μm, and the temperature is 77K. The logarithm of I_B is plotted versus gate potential, with drain bias as the second variable. The dashed lines show where $V_D = V_G - V_T$, and thus provide demarcation for the linear and saturation regions of device operation.

Region I indicates the expected exponential behavior of I_B with increasing V_G. In this subthreshold region, I_D is an exponential function of V_G, and I_B is linearly related to I_D (see Chapter Four on the modelling of I_B for a fuller discussion of this characteristic). Region II

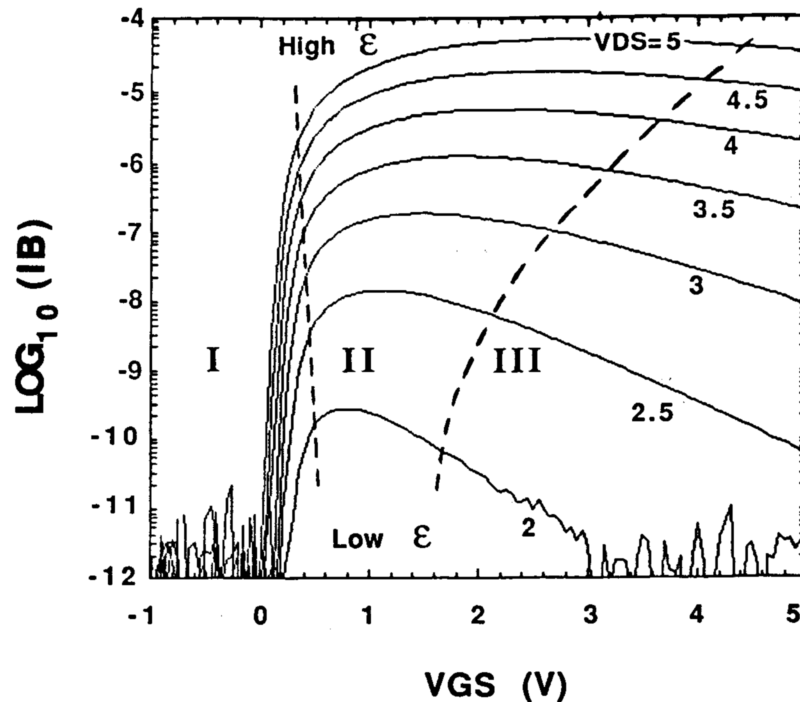


Figure 2.3: Substrate current vs. gate voltage for an N-channel MOSFET, at various drain voltages. $L_e = .85\mu$m; T=77K.

is the peak substrate current regime. V_D is greater than $V_G - V_T$, and the device is in saturation. The field in the pinch-off region - and, thus, the number of channel carriers with energy exceeding the ionization threshold - has reached a maximum, for any particular value of V_D. Region III shows the expected decrease in I_B, as the device goes out of saturation and the field in the pinch-off region diminishes.

Figures 2.4 and 2.5 show the potential contours near the device drain for biases in Regions II and III, respectively. These demonstrate the behavior of the pinch-off point in the device as a function of bias. Also, the relative separation of the potential contours shows the peak fields for Figure 2.5 are less than those for Figure 2.4.

2.3.1 Substrate current at low V_D

Figure 2.6 demonstrates another important characteristic of substrate current: the observation of I_B , even when the total available potential drain-to-source is less than the ionization

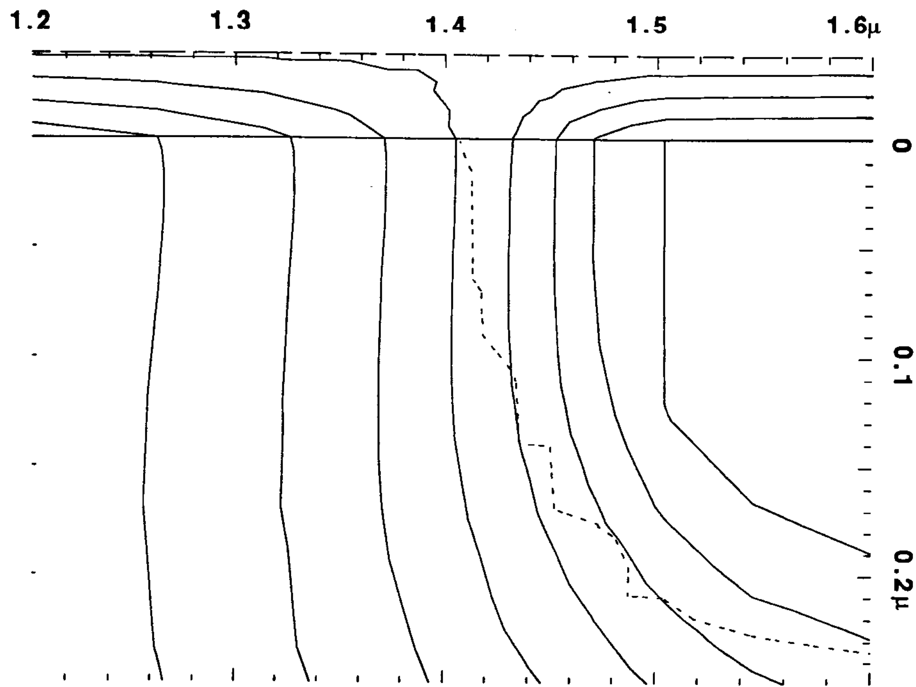


Figure 2.4: Two-dimensional potential distribution for the substrate characteristic of Figure 2.3. V_D=5V and V_G=3V. Potential contours are in 0.5V increments.

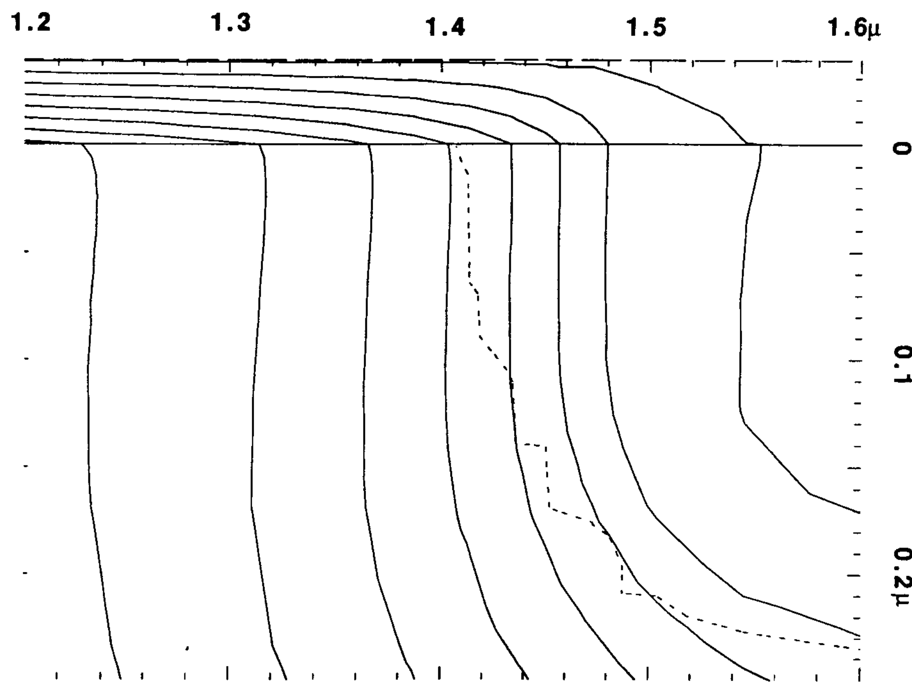


Figure 2.5: Two-dimensional potential distribution for the substrate characteristic of Figure 2.3. V_D=2V and V_G=5V. Potential contours are in 0.5V increments.

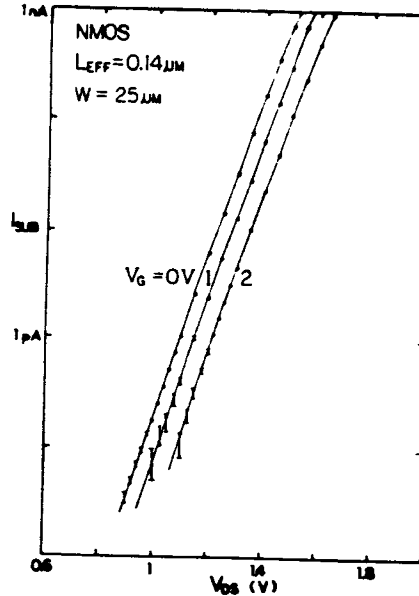


Figure 2.6: Observation of substrate current for V_{DS} less than the ionization threshold. After [Tam 83].

threshold energy. Despite the uncertainty in the literature of the exact value of the ionization threshold (see Chapter Three for more discussion and references on this issue), the value almost certainly ranges between E_g and $1.5E_g$, where E_g is the band gap energy of the semiconductor. One can see from the Figure that, for this silicon device, significant substrate current is still seen for values of V_D lower than this range. Since the total available potential energy due to external applied bias is less than this threshold, one infers immediately that some channel carriers may not lie at the minimum of the energy band, and are gaining energy from another source. As will be discussed later, ballistic energy gain from the electric field combined with subsequent optical phonon scattering gives rise to this required energy distribution.

2.3.2 Temperature dependence of I_B : V_{xover}

Any eventual model must also explain the effect first noted by Eitan, *et al.* [Eita 81a] and shown in Figure 2.7. Here, the crossover voltage V_{xover} is defined as that drain potential (1.75V) for which the substrate current is roughly a constant versus temperature.

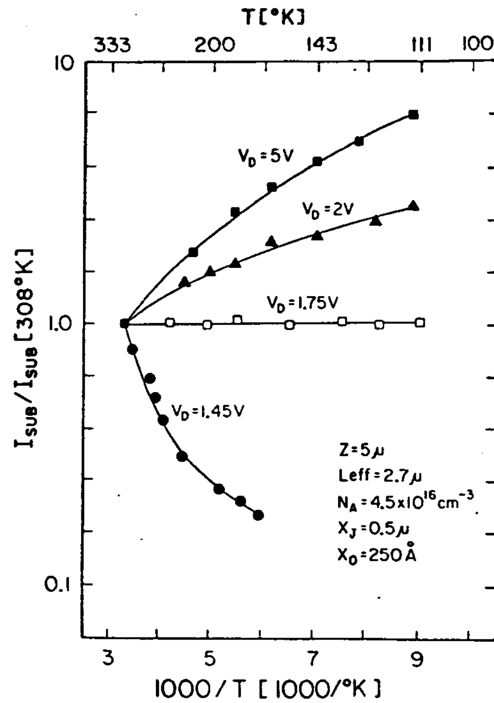


Figure 2.7: Initial characterization of crossover voltage for MOSFET substrate current. After [Eita 81a].

If substrate current is truly correlated to device degradation, then low temperature device operation at such potentials less than V_{xover} should be favorable for device reliability, *provided* one assumes a constant power supply voltage as the temperature is lowered. As shown in the next section, however, scaling temperature without scaling supply voltages simultaneously is an unrealistic comparison.

2.3.3 Relation of V_{xover} and reliability

Toriumi, *et al.* have performed the reliability correlation, to determine if device operation at the crossover voltage leads to improved reliability. Their results are shown in Figure 2.8, and demonstrate a close agreement between the crossover voltage, and the drain potential which leads to constant or improved reliability at low temperatures. The value of V_D at which 77K and 300K device reliability are equal is 1.8V, which is very nearly Eitan's value of 1.75V, and close to the value found for the devices measured in this work, to be presented in Figure 2.22. A more realistic comparison of relative reliability of technologies is seen by

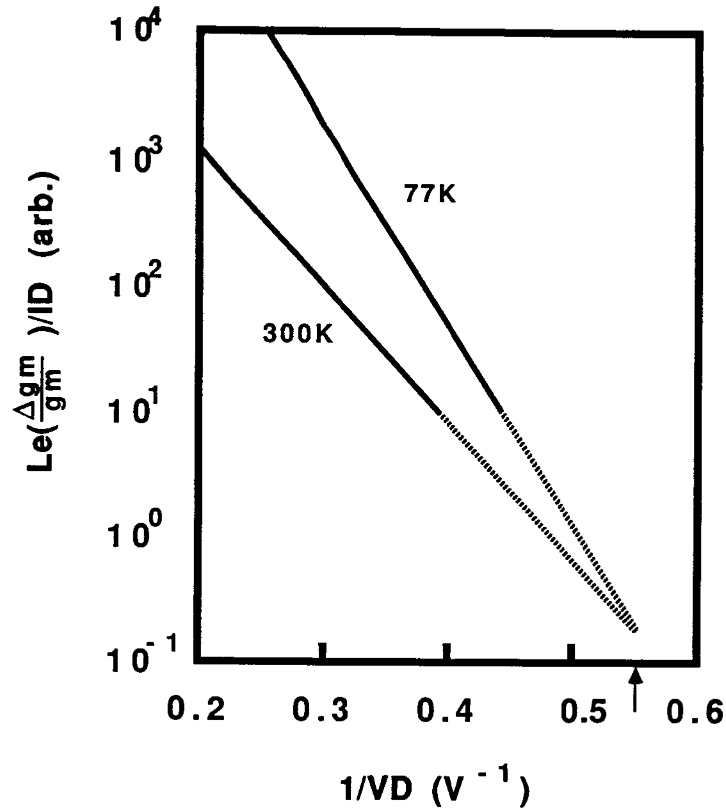


Figure 2.8: Normalized degradation vs. drain voltage and temperature, leading to comparison with V_{xover} . Increases in the parameter plotted along the y-axis correspond to decreased reliability. Solid lines are fits to their data, while dashed lines are extrapolations. After [Tori 86].

comparing the plotted value for the 300K curve at 5V, and the 77K curve at 3V. These potentials correspond closely to those at which actual technologies would operate for those temperatures. By means of this comparison, the 77K technology is still more reliable than its 300K counterpart, at a higher supply voltage.

2.4 Fabrication and electrical parameters

2.4.1 Test pattern layout and mask-making details

A test pattern to characterize substrate current and hot carrier effects in CMOS FET's was designed for eventual fabrication. The layout was generated on an Applicon AGS system.

MASK NUMBER	MASK DESCRIPTION
1	N-Well implant
2	Active area definition
3	NMOS field implant
4	NMOS threshold adjust
5	Polysilicon gate definition
6	NMOS source-drain contact
7	PMOS source-drain contact
8	Contact definition
9	Metal definition

Table 1: **Mask levels for CMOS test pattern layout.**

An N-well CMOS process requiring nine masking levels was employed (see Table 1). Transistors were designed without common pads, to avoid the testing problems associated with shared pads. This criterion was important, since exploration of gate current was a goal of the work. A variety of gate lengths and widths were included; very short gate lengths were drawn, in the hope these transistors could eventually be formed with e-beam lithography. Junction and MOS capacitors were added to find out the relevant parameters for later calculations of gate and field oxide thickness and capacitance, and junction capacitance. Van der Pauw structures and contact cross-bridges were drawn to find the electrical critical dimensions (CD's), resistivities, and areal contact resistivities for the various layers. Inverters and ring oscillators were drawn, intended to evaluate the CMOS process from a simple circuit point of view.

The layout of a typical PMOS transistor is shown in Figure 2.9. The dimensions are 25 microns by 25 microns. An intimate contact is made to the N-well; a similar situation holds for the N-channel device, even though the wafer backside is available as substrate contact. (This proved necessary at low temperatures, where the wafer backside contact became highly resistive due to oxide and $n+$ polysilicon layers atop the p-substrate.)

A typical inverter used in the 23-stage ring oscillators is shown in Figure 2.10. Both devices have W/L ratios of 25/2 in microns. Nominal electrical channel lengths are 1μ. Fan-in and fan-out were both unity.

Pad sizes throughout the test pattern were 120 microns on a side. Masks were made

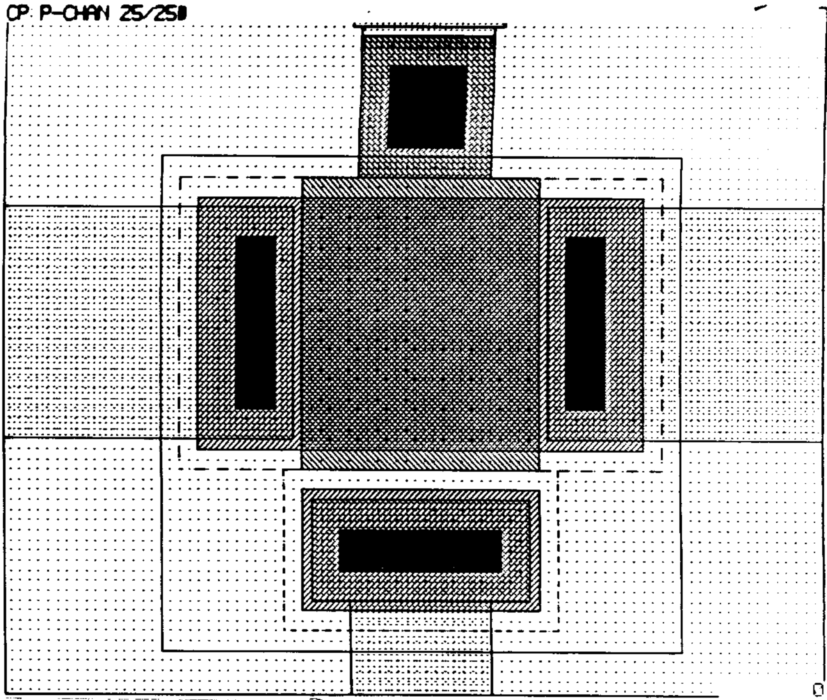


Figure 2.9: The layout of a typical PMOS transistor

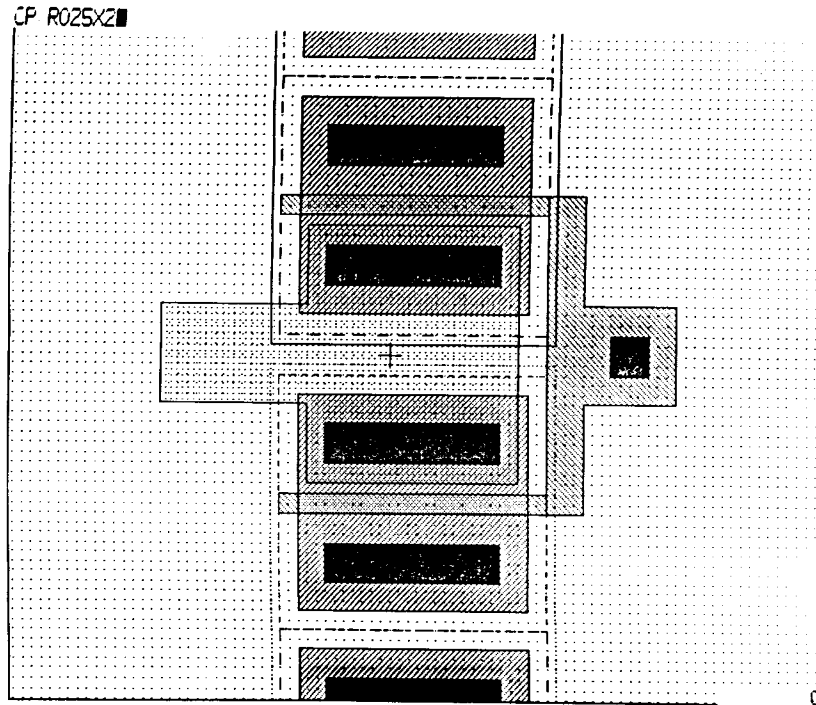


Figure 2.10: The layout of a typical CMOS inverter

using the MEBES mask-making facility at Stanford. Exposures on the three inch wafers were carried out using a Canon 4:1 nine-position stepper.

2.4.2 Processing details and process characteristics

A conventional N-well CMOS technology was chosen for processing the experimental devices [Pfie 84]. The N-channel devices use a standard arsenic source-drain, with no LDD structure. Previous work [Sun 84] indicated certain N-channel LDD structures can lead to serious problems for liquid nitrogen temperature operation, as trapped charge above the LDD region - combined with freezeout effects - turns off the device channel. The P-channel devices used conventional boron source-drain regions.

Gate oxide thickness, t_{ox}, was 385Å. Field oxide thickness varied between 6000 and 8000 Å, depending on the wafer chosen and the doping of the substrate beneath the measurement location. Both of these thicknesses were measured using both optical as well as capacitive

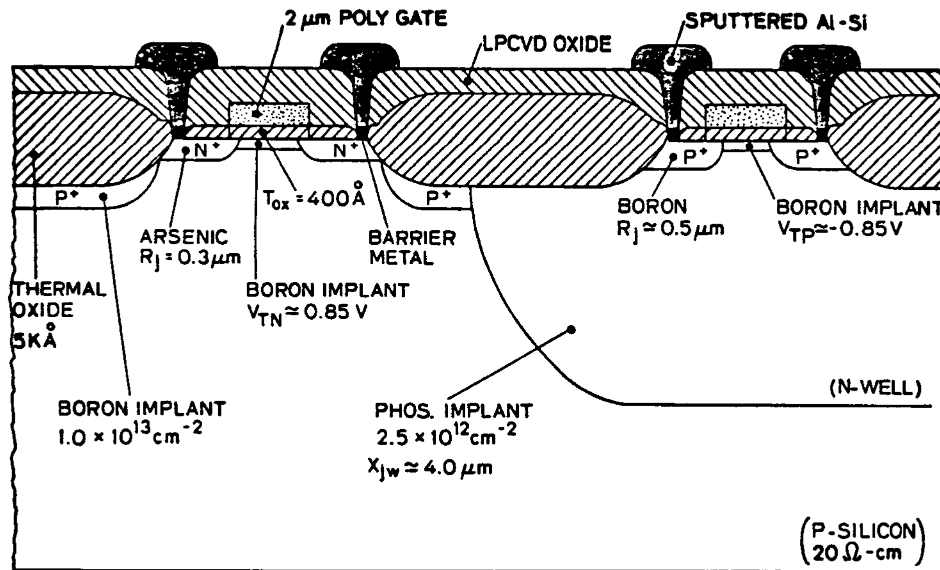


Figure 2.11: CMOS process cross-section. After [Pfie 84].

techniques. The n^+ junction depth was 0.24μm, while the p^+ junctions were 0.62μm deep. The N-well depth was 3μm. The junction doping profiles, from SUPREM III, are shown in Figures 2.12 and 2.13. Sheet resistivities for the n^+, p^+, and N-well layers are 73, 167, and 3100 Ω/□. These were measured using both Van der Pauw and resistor structures.

Threshold voltage versus substrate bias is shown in Figures 2.14 and 2.15 for the long-channel devices (25/25) at room temperature. V_T was measured using the extrapolation technique in the low-V_D linear regime (V_D=50mV) [Yau 74]. The measured characteristics are consistent with the predicted threshold voltages using both SUPREM III and CADDET [Mock 73]. Later, PISCES II-B [Pint 84] simulations with the same profiles correctly predicted the linear and subthreshold characteristics for both long and short channel devies. The channel profiles used are given in Figures 2.16 and 2.17. Channel surface low-field mobilities are 720 cm^2/V-s and 200 cm^2/V-s for electrons and holes, respectively. The short-channel effect caused a threshold voltage decrease of roughly 150mV between the long channel and 1μm devices.

It should be noted that an error in processing reduced the N-channel enhancement threshold voltage adjust implant to one tenth its planned value. This had the fortunate consequence of lowering the value of V_{TN} to almost the ideal value for cryogenic operation,

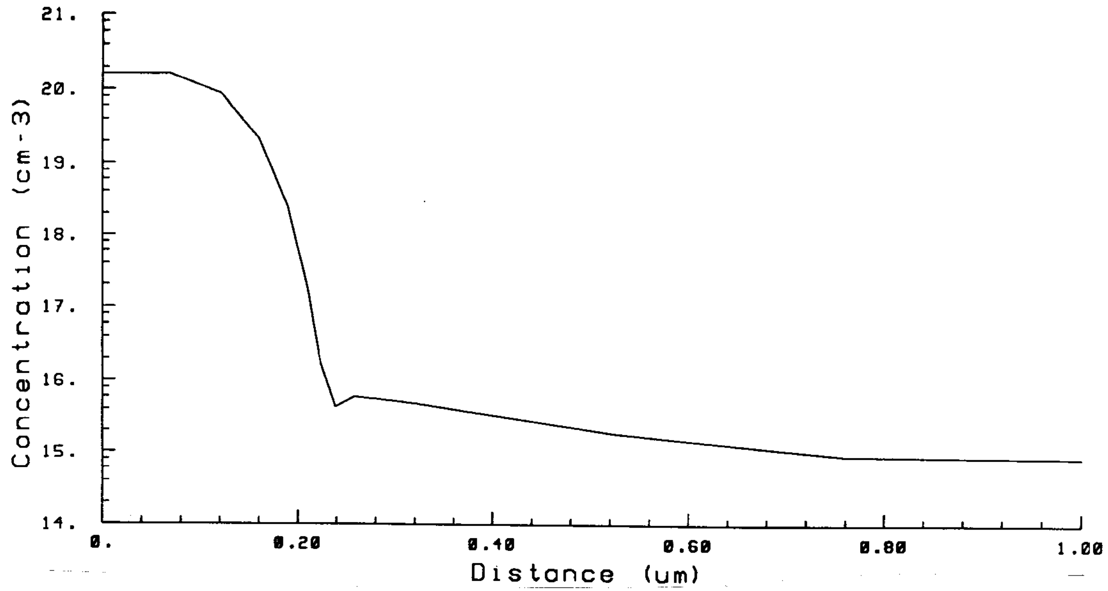


Figure 2.12: NMOS source-drain profile.

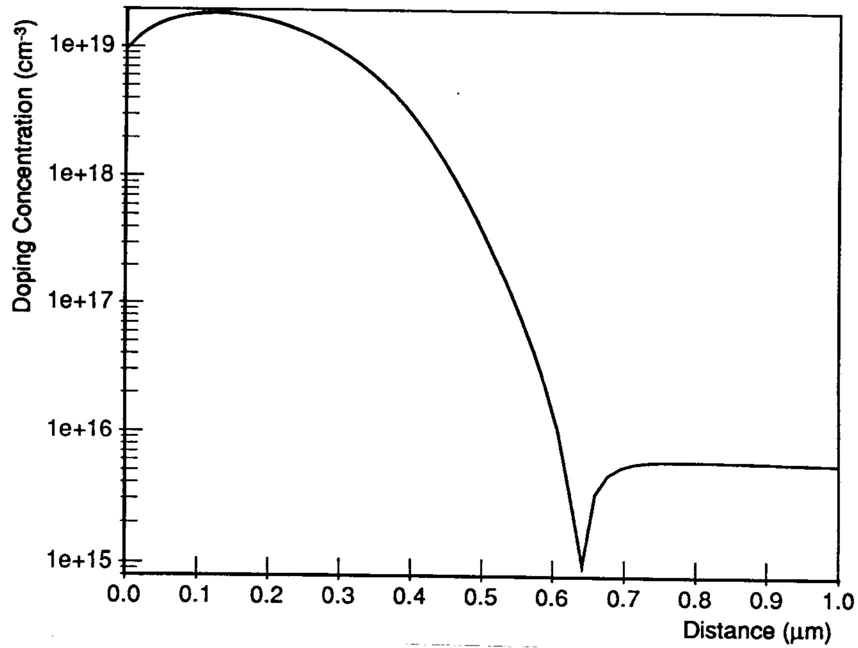


Figure 2.13: PMOS source-drain profile.

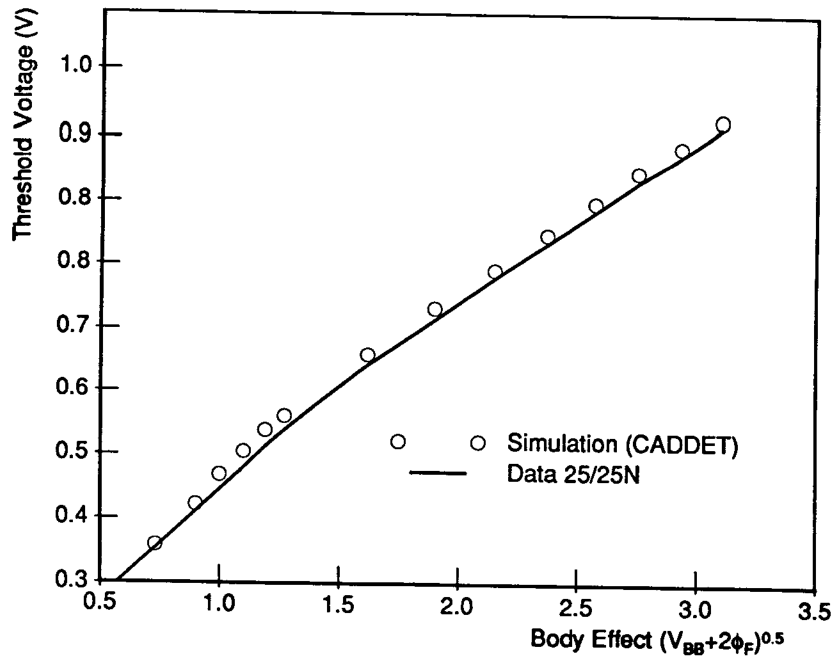


Figure 2.14: N-channel threshold voltage

assuming a reduced power supply voltage [Sun 87].

2.5 Characterization methodology

The following procedure was used to characterize CMOS substrate current. First, I_D was measured for $|V_G|$=-1 to 5V and $|V_D|$=50mV. This measurement was performed versus V_B as well, with $|V_B|$ ranging between 0V and -9V. This measurement served several purposes: monitor transconductance, g_m, for any changes due to measurement-induced degradation; channel length calculation; channel doping calculation; subthreshold slope calculation; mobility calculation.

Second, I_D was measured for the same range of V_G, but now varying $|V_D|$ between 0.5V and 5V. Substrate voltage was held at 0V. These measurements allowed later simulation comparison, to ensure that the I_B modelling results were based on knowledge of the device fields versus these biases.

Third, I_B was measured for the same voltage conditions as the second set of curves.

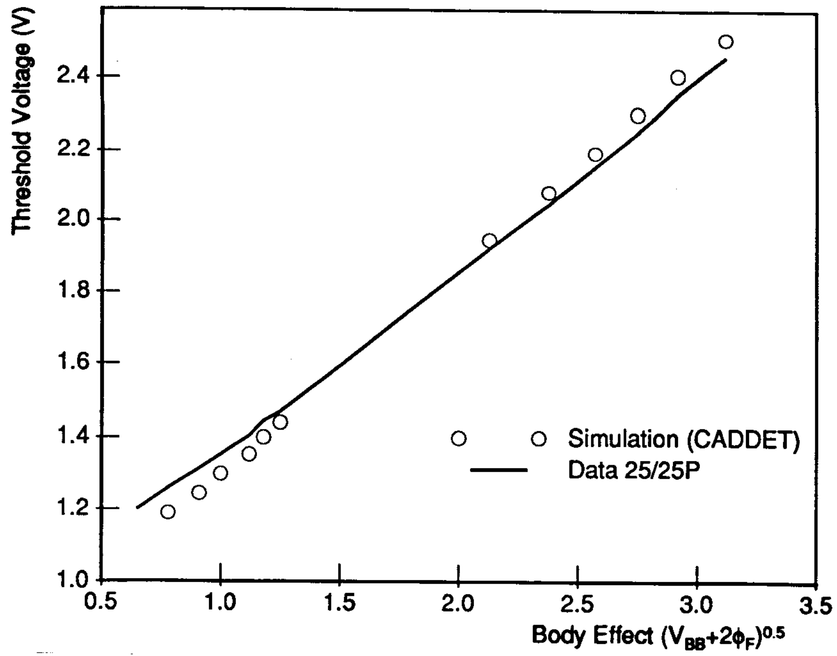


Figure 2.15: P-channel threshold voltage

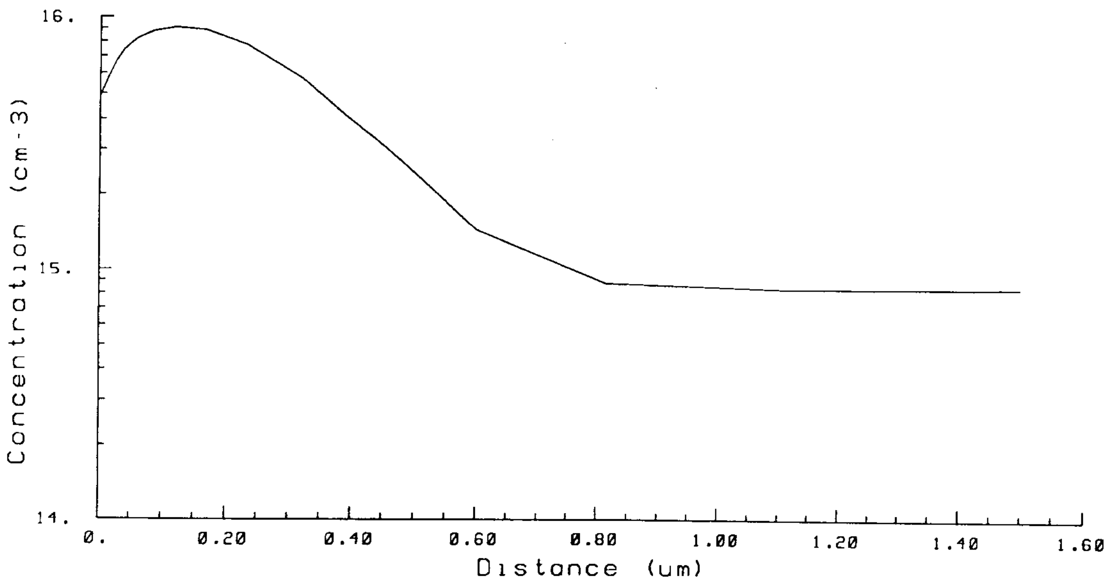


Figure 2.16: N-channel doping profile

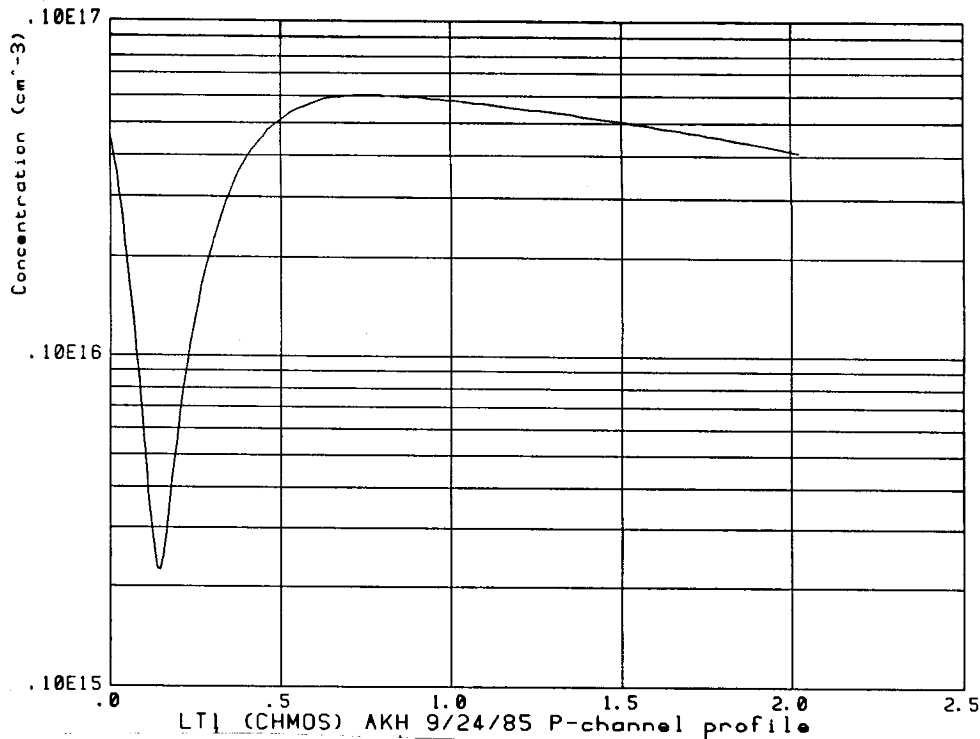


Figure 2.17: P-channel doping profile

These are the fundamental measurements of impact ionization in the MOSFET. As in Figure 2.21 below, the peak substrate current will be used extensively, as a presentation tool.

Fourth, the second and third measurements were repeated for $|V_B|$=-5V.

Finally, the g_m fingerprint as above was repeated, this time with back biases of 0V and -5V. Again, this monitored stresses experienced by the device during the high-field I_D and I_B measurements.

These measurements were then repeated at several temperatures, usually 77, 95, 100, 111, 125, 143, 167, 200, 250, and 300K.

The experimental set-up consisted of the following: a liquid nitrogen cryostat; an 18-pin DIP socket to hold the test devices; a Lake Shore Cryotronics controller, to control the heater coil and monitor the germanium temperature-sensing diode; an HP4145 semiconductor parameter analyzer connected to the test socket via coaxial BNC cables; and an HP9845 desk-top computer to monitor and control all the electronics, including the liquid nitrogen level sensor in the cryostat. Due to the limitations of the cryostat, direct immersion into

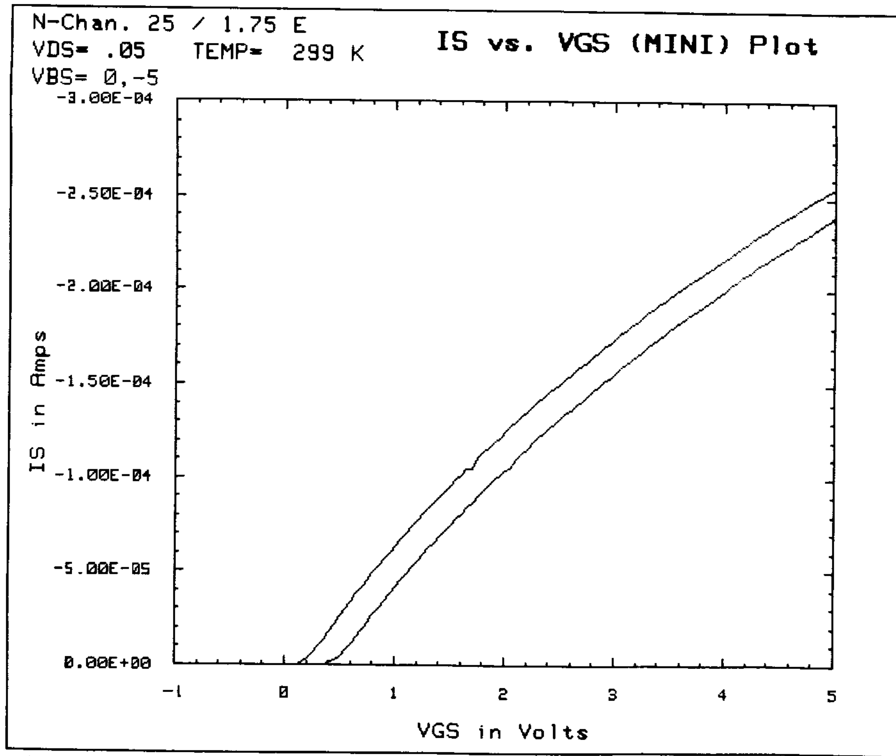


Figure 2.18: Example of the transconductance monitor of the measurement sequence. The device is drawn 25/2, with L_e =0.85, and T=299K.

liquid nitrogen was used for the 77K measurements. Both ultrasonic and gold-ball bonding were used to wire the devices into the DIP sockets.

2.6 Measurement results

2.6.1 Drain current characteristics

An example of the g_m monitor is shown in Figure 2.18. The high-field I_D measurement is shown in Figure 2.19. For channel lengths below 1μm, some surface punchthrough or drain-induced barrier lowering (DIBL) was noted, as evidenced by increased subthreshold slope. This punchthrough characteristic, however, could be eliminated by application of substrate bias, or operation at lower temperature.

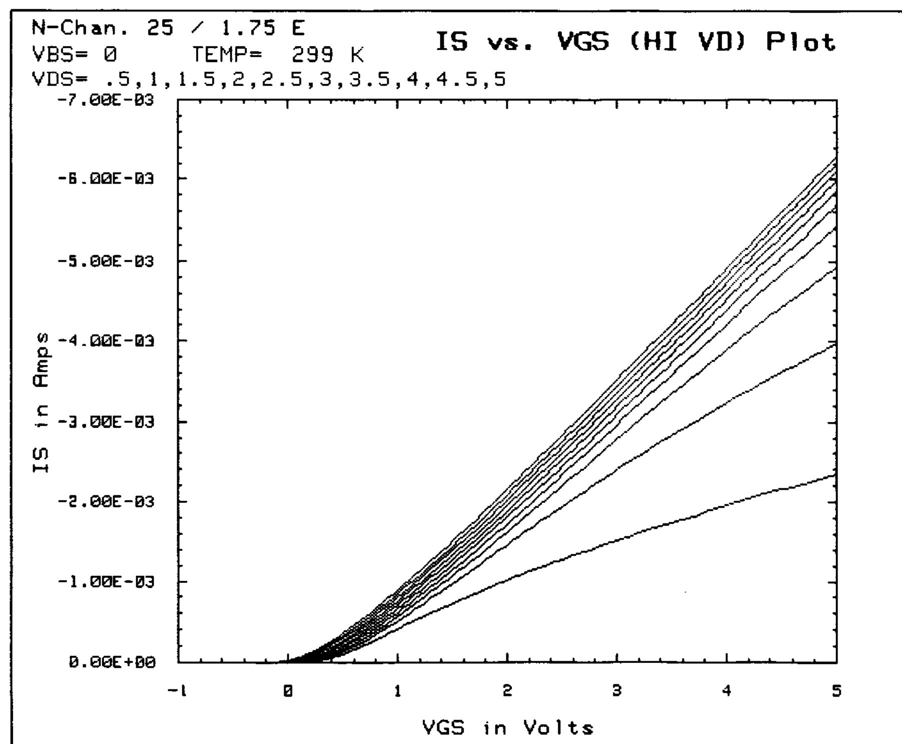


Figure 2.19: Example of the high-field I_D measurement, for the same device as in the previous Figure.

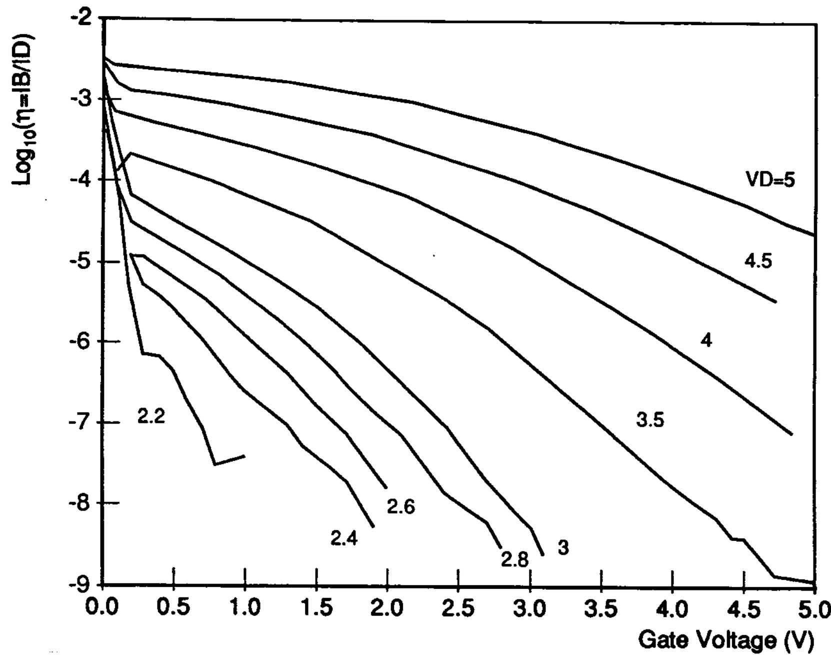


Figure 2.20: Generation efficiency of impact ionization process in MOSFET.

2.6.2 Substrate current characteristics

An example of the I_B measurement was previously shown in Figure 2.3. The substrate current characteristic can be plotted in a fashion which shows the impact ionization efficiency $\eta = I_B/I_D$ in the global device. Figure 2.20 shows this for an N-channel device with electrical dimensions of 25/2.15 in microns. Several features are worthy of note. First, the efficiency η is much less than unity. This means the substrate current is a small peturbation on the total current in the channel, and does not affect the solution of the device equations. Modelling of the substrate current will rely intimately on this fact. Second, the efficiency can be used as another monitor of device degradation, and has been so utilized elsewhere [Niss 86a].

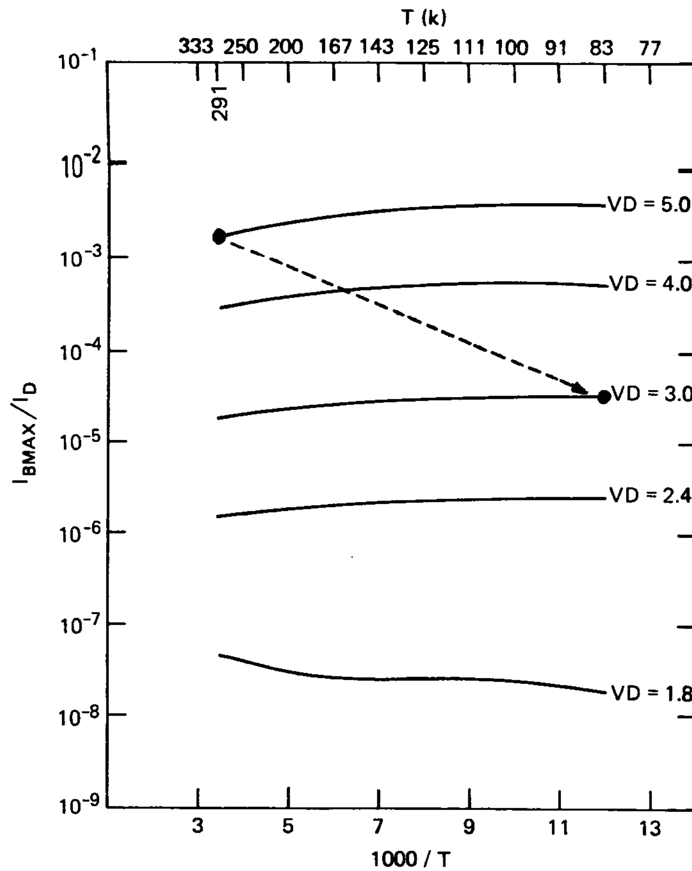


Figure 2.21: **Normalized peak substrate current vs. temperature, showing V_{xover}.** L_e=1.15μ. Also shown is a comparison between the peak I_B for 300K at 5V, and the peak for 77K at 3V. Clearly, the 77K value is much lower than that at 300K, leading to a prediction of improved reliability at 77K for realistic power supply reductions along with the temperature decrease. The mismatch between the decrease predicted here and that in Figure 2.8 indicates that other factors, such as a temperature-dependent trapping cross-section, must also be considered.

2.6.3 Voltage crossover

V_{xover} is demonstrated in Figure 2.21 for one of the N-channel devices used in this dissertation. At large drain voltages, the peak I_B (normalized to I_D as in [Tam 82], to remove the temperature dependence of the channel current itself) for this device with L_e =1.15μm, increases with decreasing temperature. However, at lower drain biases, the opposite is true. Again, this is precisely the effect one would want in a CMOS technology designed for cryogenic operation, to avoid hot carrier degradation effects. V_{xover} equals approximately 2V in Figure 2.21.

Using other channel length devices according to the process outlined below, Eitan's work

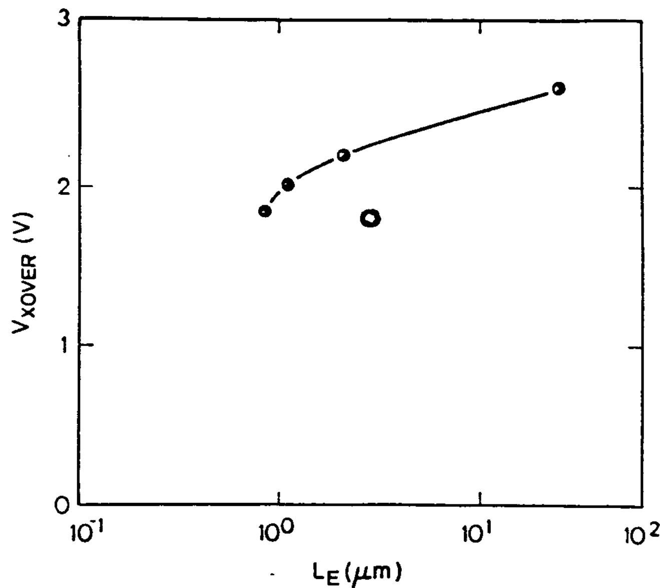


Figure 2.22: Crossover voltage versus channel length for N-channel MOSFET's. Filled circles: t_{ox}=38.5 nm. Open circle: t_{ox}=25.0 nm, after [Eita 81a].

has been extended. V_{xover} is plotted versus L_e for several N-channel devices in Figure 2.22, from which one may infer a power supply voltage of slightly less than 2V is required for a cryogenic CMOS technology designed so that hot carrier effects (as manifested by substrate current I_B) are no worse at low T than at room temperature. Other criteria may also be important when determining the reliability of a technology allowing scaled temperature and power supply. In particular, if one looks strictly at gate current and requires it to be constant as a reliability constraint, then the power supply may be considerably higher than 2V for a cryogenic technology [Kato 84].

The (V_{xover}, L_e) data from [Eita 81a] is also shown in Figure 2.22. Eitan's t_{ox} (250Å) is roughly sixty percent, and his peak channel doping (4.6 x 10^{16} cm^{-3}) is roughly five times, greater than that in the other N-channel devices. Both insulator thinning and increased doping cause field increases in a scaled device; one can infer, then, that V_{xover} is weakly dependent on scaling to smaller geometries. A stronger dependence might be expected, except that the channel current travels deeper in a scaled device, and so does not experience the same magnitude of peak field [Laux 84]. Thus, if reliability is a concern, and if V_{xover} is an indication of susceptibility to - or protection from - hot carrier effects, a performance enhancement trade-off will need to be made between temperature scaling and geometry scaling.

It should be noted that Lau, *et al.* [Lau 85] demonstrate a phenomenon similar to V_{xover} in their description of I_B/I_D vs. inverse pinch-off field: at lower pinch-off field, I_B/I_D decreased as T decreased. However, it was not possible to determine the peak I_B from their data; nor was any explanation of the phenomenon made.

It was not possible to observe V_{xover} directly in the P-channel devices used in this work, due to a minimum measurable I_B of roughly 100fA. However, extrapolation of the P-channel I_{BMAX} vs. T vs. V_D data is consistent with this 2V design requirement, deduced from the N-channel characteristics.

2.6.4 Temperature effects

Shown in Figures 2.23, 2.24, 2.25 and 2.26 are the absolute and normalized I_D data for various devices and bias conditions, versus temperature. The enhancements at low drain bias are comparable to expectations for low-field mobility increase, while those for high drain bias relate to expectations for saturated velocity increases in the devices. They compare quite well with high-field transconductance improvements reported by Aoki, *et al.* [Aoki 87] and Sun, *et al.* [Sun 87]. Note, however, that these data understate the actual velocity saturation improvement, since they are currents at absolute biases, not slopes for a given V_D and fixed gate drive, $V_G - V_T$.

The temperature dependence of threshold voltage is shown in Figures 2.27 and 2.28. The results are comparable to expectations based on the behavior of the built-in voltage versus temperature, $\phi_b \sim k_B T/q$.

g_m in mS/mm are given in Table 2. Again, these compare favorably with other work [Aoki 87,Sun 87].

Ring oscillators were measured for the CMOS process. The absolute delays for the oscillator layout of Figure 2.10 with 23 stages and a three-stage output buffer are shown in Figure 2.29, versus power supply voltage and temperature. The same data is plotted normalized to the room temperature number in Figure 2.30. The CMOS process employed was decidedly not optimized for low temperature operation - especially considering the imbalance between N-channel and P-channel threshold voltages. Nevertheless, a nearly 40% improvement in the propagation delay could be seen for all four power supplies measured,

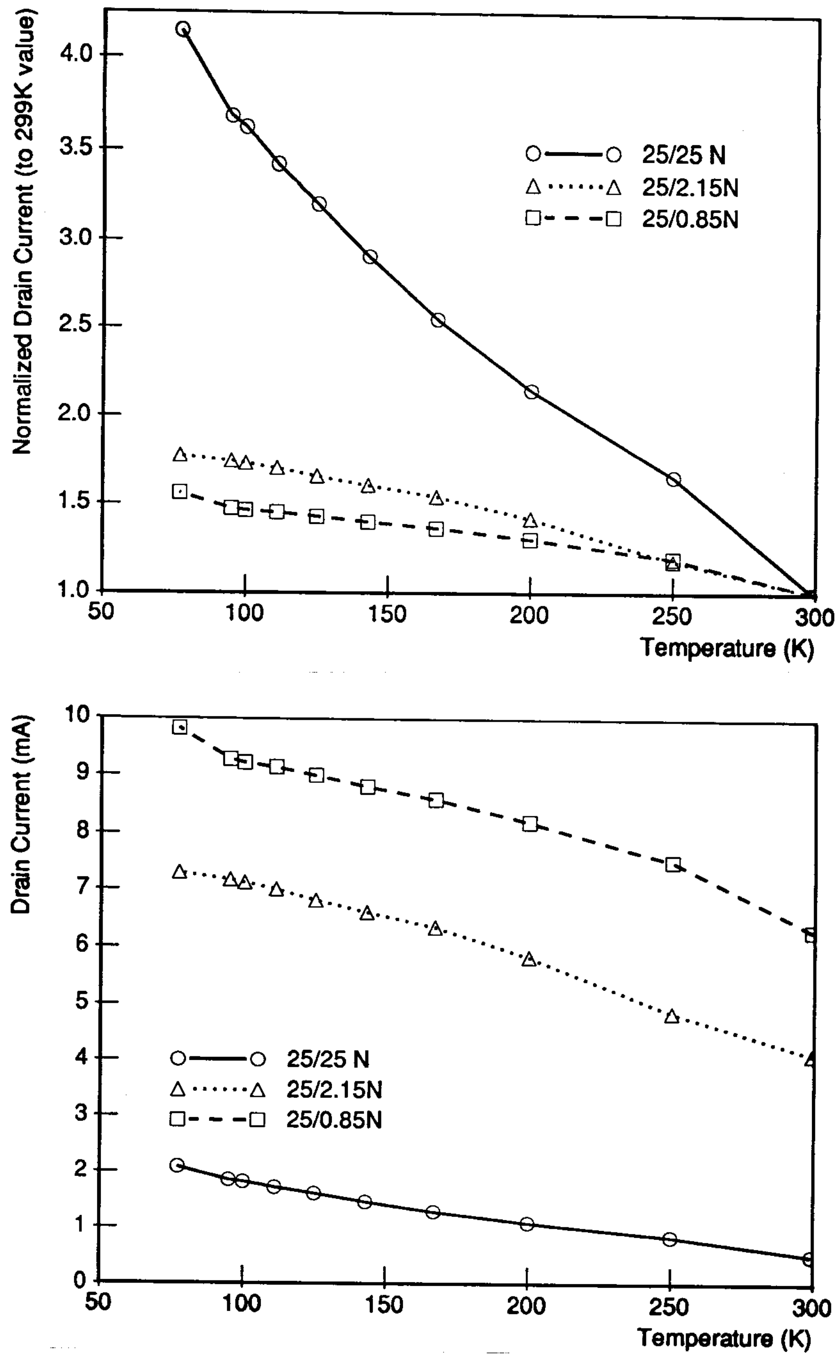


Figure 2.23: **Drain current characteristics versus temperature for N-channel devices**, with 5V on the **gate** and drain. The upper curves are normalized to the respective 299K values, while the **lower** curves are absolute numbers.

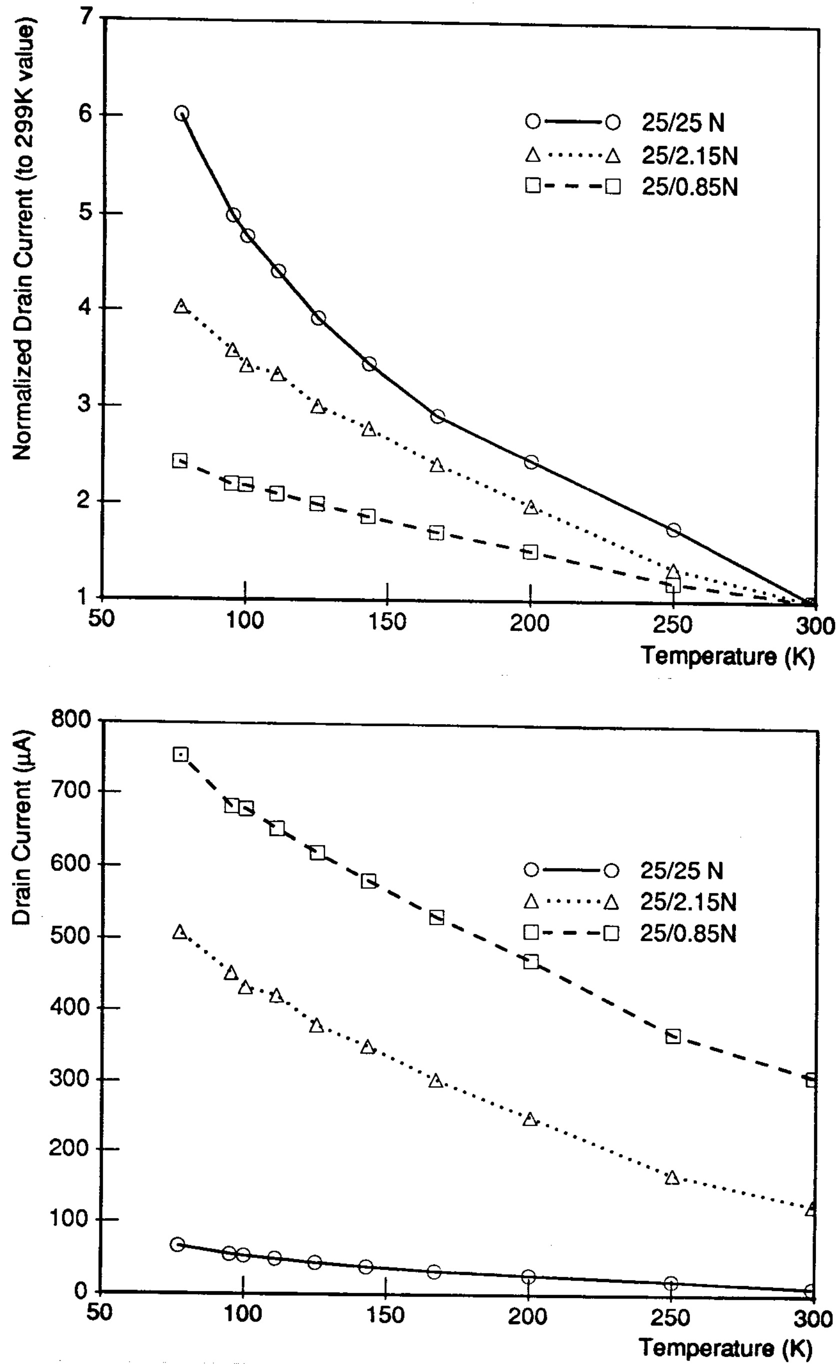


Figure 2.24: Drain current characteristics versus temperature for N-channel devices, with V_D=50mV and V_G=5V. The upper curves are normalized to the respective 299K values, while the lower curves are absolute numbers.

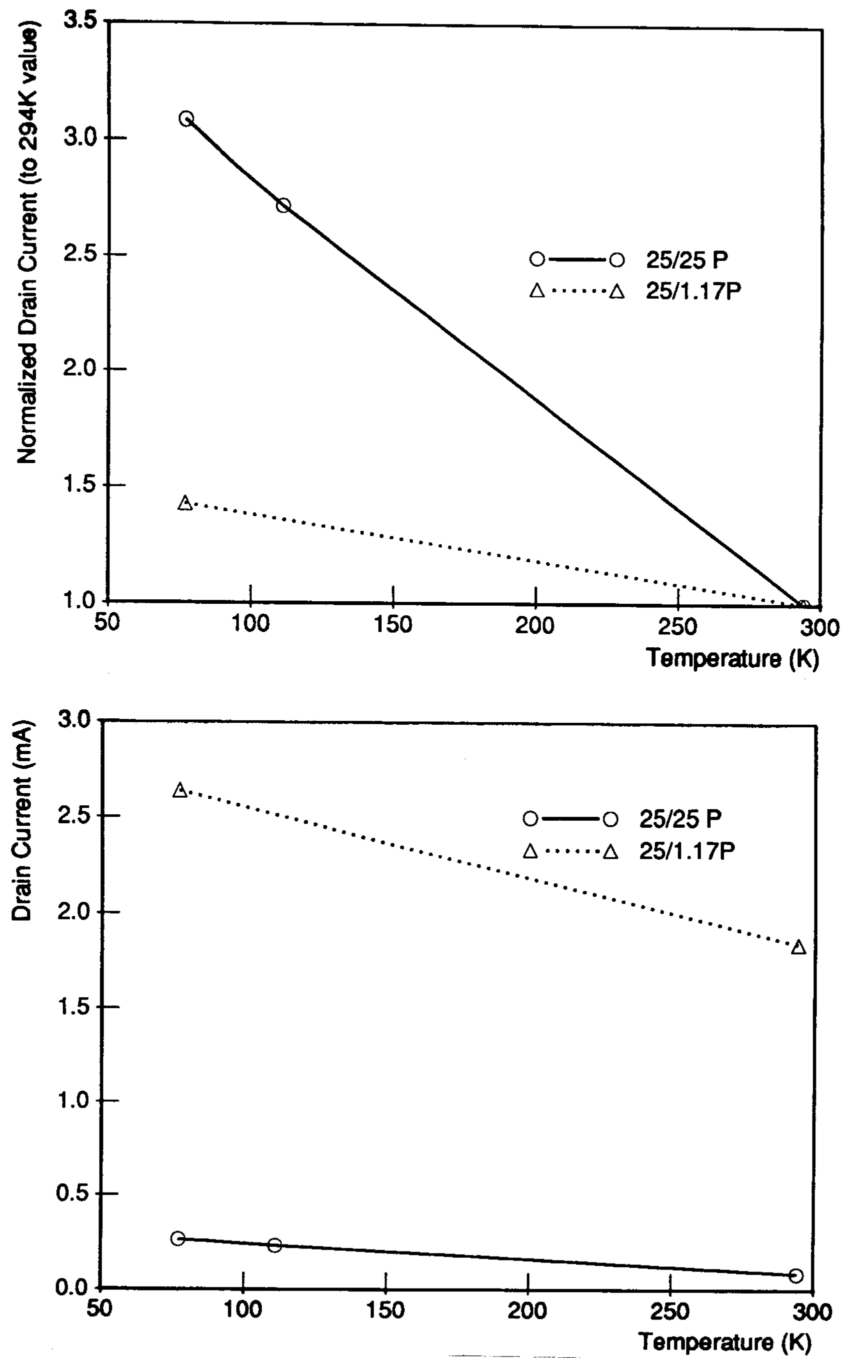


Figure 2.25: Drain current characteristics versus temperature for N-channel devices, with -5V on the gate and drain. The upper curves are normalized to the respective 299K values, while the lower curves are absolute numbers.

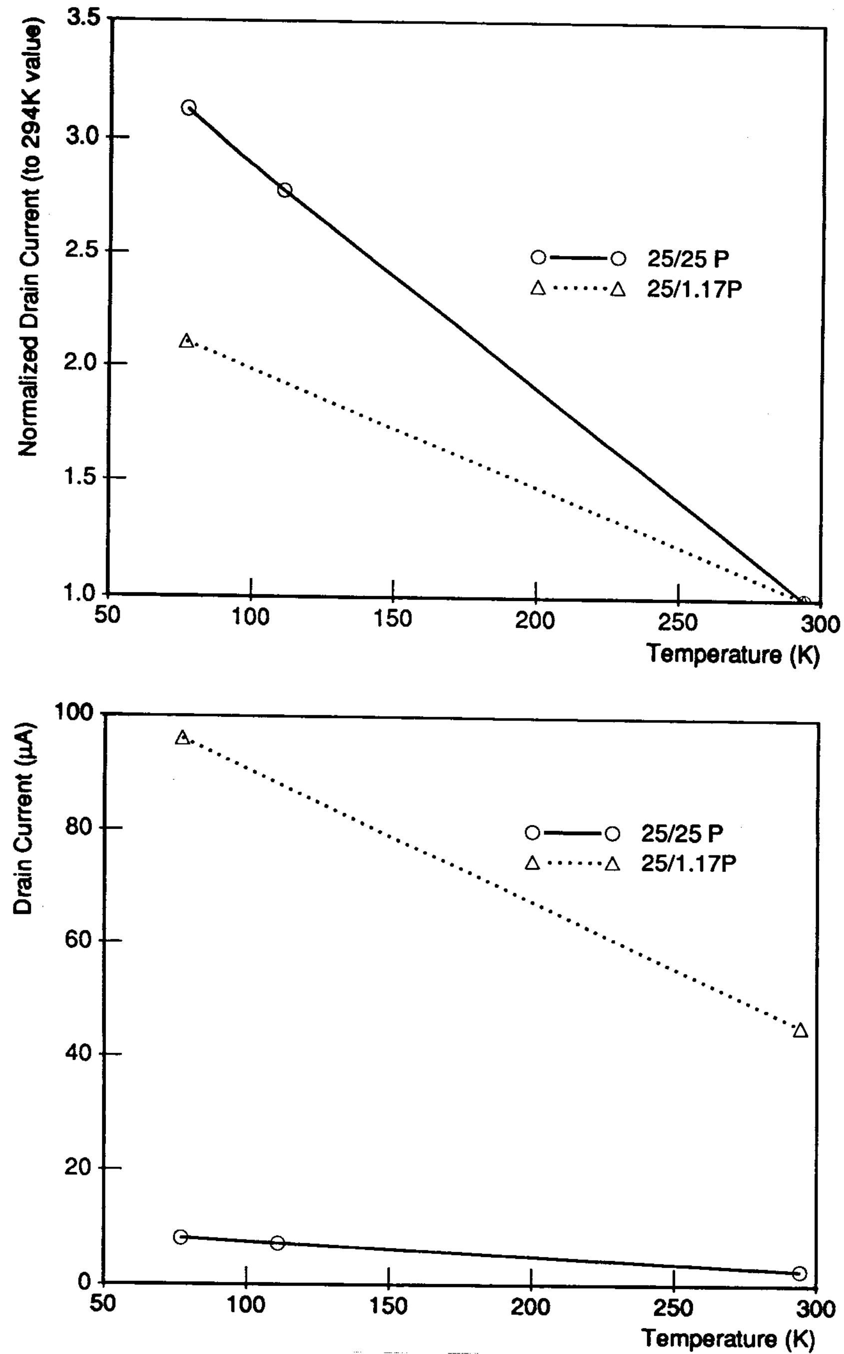


Figure 2.26: Drain current characteristics versus temperature for N-channel devices, with V_D=-50mV and V_G=-5V. The upper curves are normalized to the respective 299K values, while the lower curves are absolute numbers.

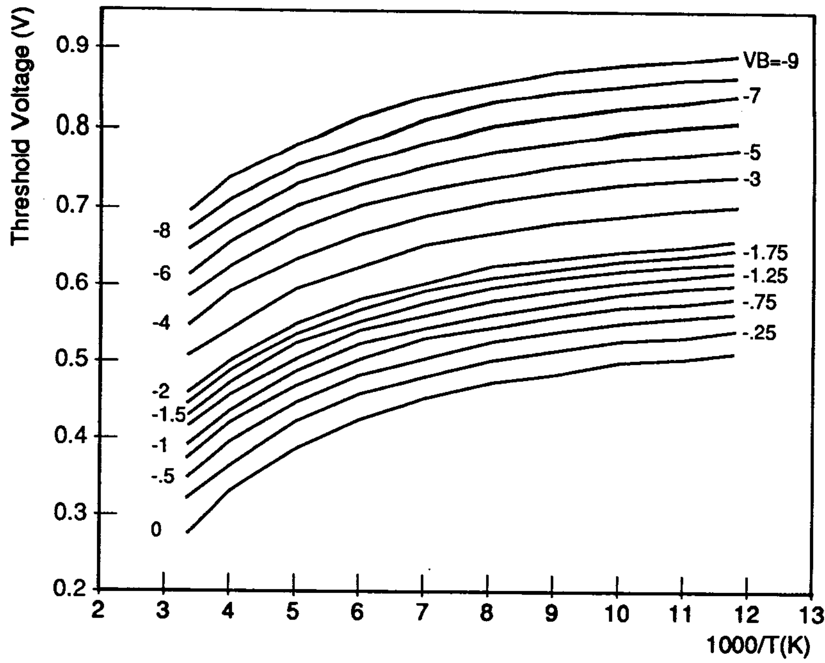


Figure 2.27: Threshold voltage for 25/1.15 N device, versus temperature.

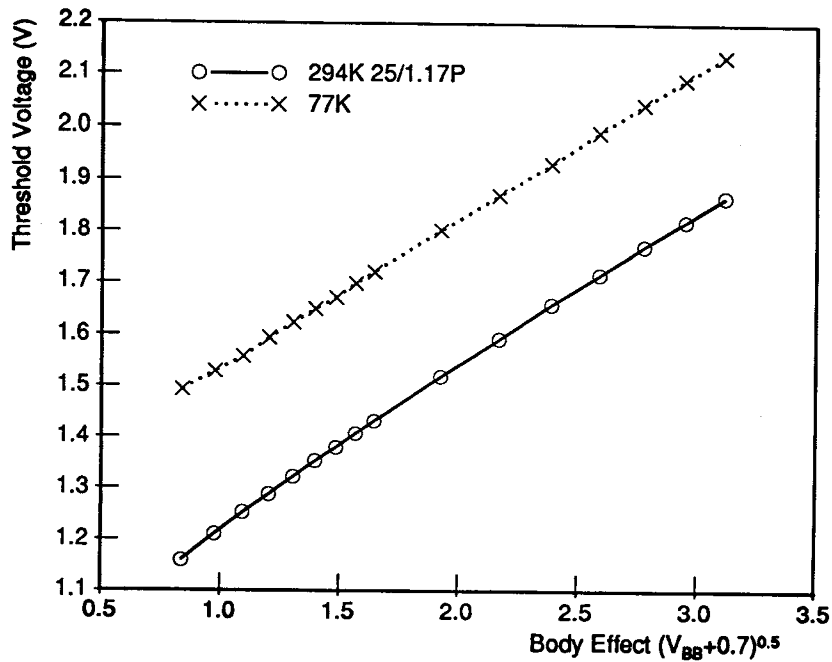


Figure 2.28: Threshold voltage for 25/1.17 P device, versus temperature.

g_m (mS/mm)		
Device \ T	300K	77K
25/25 N	7.33	29.8
25/2.15N	42.6	71.4
25/0.8N	55.2	87.3
25/25P	1.56	4.66
25/1.17P	24.1	32.5

Table 2: Device performance - transconductance

by reducing the temperature from 300K to 77K.

A comparison may be made to the ring oscillator results of other workers. Table 3 makes this comparison, but some notes should be made. First, the gate delay for 3V from the data of Aoki, *et al.* [Aoki 87] is inferred from their Figure 9. Second, the delay for Sun, *et al.* [Sun 87] is derived by dividing their given value of 450ps for a 1μm channel length by the fan-out. Also, they used a NAND gate, rather than the NOR gate of this work and that of Aoki, *et al.* The effect of a NAND gate is to increase the delay somewhat relative to the NOR configuration, since the parasitic gate capacitances are increased, and the charge transferred from one power rail to the other has a longer distance to travel. These are both small effects, however.

With these in mind, the optimized technology of Sun, *et al.*, is clearly the best. Yet, because of the use of a NAND gate and lower threshold voltages, it is not that much different than that of Aoki, *et al.* In terms of an optimized, low temperature technology, the Stanford process suffered most from the high P-channel threshold, and the longer channel length.

Given these differences, then, the results of Table 3 are not mysterious. Surprising, however, is that Sun, *et al.* do not obtain even more improvement in delay vis-á-vis the other two processes. One infers that velocity saturation is a dominant component to signal transit through a MOSFET channel, and that true circuits will be hard-pressed to improve speed gains at low temperature beyond that allowed by this limiting factor.

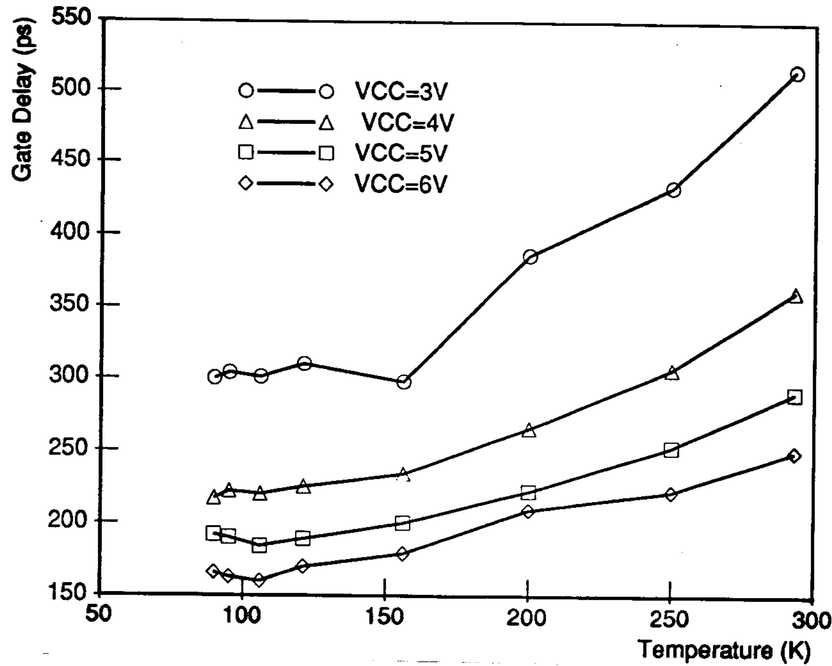


Figure 2.29: Ring oscillator gate delay

77K CHARACTERISTICS

	STANFORD	SUN	AOKI
TOX(Å)	385	125	200
VTN(V)	.4	.42	.63
VTP(V)	-1.4	-.43	-.85
LE(µm)	1.2	1.0	.80
td(ps)/ fanout (VCC=3V)	300	150	200

Table 3: Comparison of ring oscillator delays for three CMOS technologies

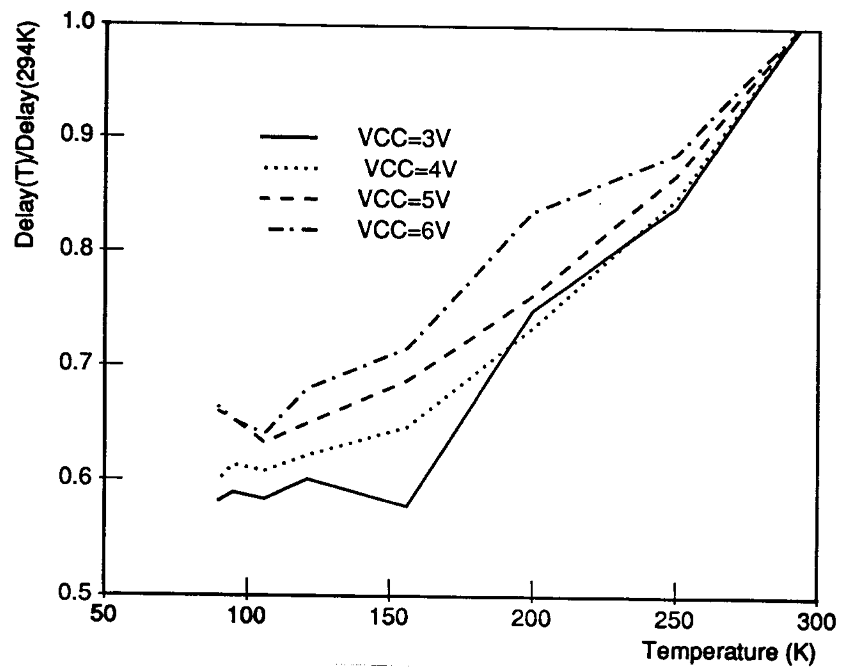


Figure 2.30: Ring oscillator gate delay, normalized to the 294K value

2.7 Summary

This chapter began by outlining previous characterizations of substrate current in MOSFET's, at nominal and low drain biases, and at room and low temperatures. The device design, process, and fabrication procedures used to make the characterization devices in this work were then outlined. Some of the pertinent, general electrical characteristics were given. The characterization procedure for substrate current was outlined, and the important results presented.

The concept of voltage crossover was defined, and extended to the regime of submicron channel lengths. Most important, the crossover voltage was promoted as a measure of device reliability over any temperature of operation: the voltage below which reliability will improve as the operating temperature decreases. Other work [Tori 86] has subsequently confirmed this concept.

Two important results of this and previous characterizations present themselves. First, I_B can be observed even when the drain-to-source bias is less than the ionization threshold. The physical implication of this observation is that channel carriers may not lie at the bottom of their energy band, but must have a probability for transport at elevated kinetic energies - that is, for travelling 'hot'.

Second, if the drain-to-source bias is low enough, I_B can be observed to decrease as temperature decreases. This result runs counter-intuitively to one's expectation based on simple considerations. That is, if temperature decreases, one expects the mean free path of a carrier to increase. If the mean free path increases, then the carrier's kinetic energy should, on average, increase - and so, one would guess, lead to a greater probability of obtaining the right amount of energy to break a Si-Si bond.

The next chapter will derive the form of the carrier energy distribution, and argue qualitatively that this distribution can explain both of these experimental observations. It will be left to Chapter Four, however, to implement the derived energy distribution in a 2-D device simulator, and show that it does, in fact, predict the observation of V_{xover}.

Chapter 3

Impact ionization in silicon

3.1 Introduction

This chapter begins with an historical review of the phenomenon of impact ionization in bulk silicon. The critical concept of the energy distribution of a charge carrier is then presented, by a derivation which begins with the Boltzmann Equation. The concept of the carrier temperature is introduced. The chapter closes by reiterating the assumptions of the various models, pointing out their failings, and making plausibility arguments that the energy distribution derived can help explain all of the important observations detailed in Chapter Two - which sets the stage for the new model presented in Chapter Four.

Questions of physics important to the understanding of impact ionization are pointed out as appropriate. The chapter also explains the relevant terminology: impact ionization; mean free path; ionization coefficient; local versus non-local models; macroscopic versus microscopic modelling; steady-state; equilibrium, or non-equilibrium, field; and the assumptions pertinent to treatment of the Boltzmann Transport Equation (BTE).

3.2 Definition of terms

Impact ionization refers to the breaking of a solid state lattice bond by a charge carrier, whose kinetic energy exceeds the threshold for bond breaking. This threshold is called the ionization threshold, and is comparable to the band gap energy in a semiconductor.

The mean free path is the *average* distance a charge carrier will travel, without scattering. Over this distance, the carrier motion is said to be ballistic, with the electric field being the motive force.

The ionization rate refers to the amount of current generated at a point in the crystal lattice due to the impact ionization process involving the highly energetic, principal charge carriers. Often, the local ionization rate is referenced to the steady-state current density in that vicinity.

A local ionization process depends only on the field, doping, temperature, and potential at a particular point in the lattice. A non-local ionization process depends on factors outside of the spatially local sphere where the charge is generated. In particular, non-local processes include consideration of the current paths in a material or device, and how parameters upstream of the generation point can affect the ionization rate at that point.

Macroscopic models seek to explain impact ionization processes on a scale large with respect to the lattice. For instance, empirical models of substrate current in MOS devices (see Chapter Four) often rely simply on the peak electric field in the device channel, and not on the local specifics or variations of the field. Microscopic models, on the other hand, seek to explain the measurement of gross terminal currents caused by impact ionization - substrate current in the case of an MOS device - by considering the physical processes of impact ionization on a scale comparable to that of the crystal lattice.

3.3 Questions of physics

Several questions of physics must be addressed to understand impact ionizations processes completely, and the observations from Chapter Two. The concept of the ionization threshold [see [Eita 81a] citations, for instance] relates to the energy a carrier must have before it can break a lattice bond, and will influence any theory or model investigating impact processes. As an example, the value of the threshold is $1.5E_g$ if one assumes spherical, nondegenerate bands with identical curvature in k-space (identical effective masses) [Wolf 54]. However, the use of such a high value in Si leads to inconsistencies with measured mean free paths for optical phonon scattering, as will be discussed in Chapter Four. A more appropriate value is the band gap E_g itself - which is, after all, the bonding energy between the lattice

atoms [Harr 80].

The mean free path itself is important in the derivation of the energy distribution of charge carriers. It also affects the evaluation of a carrier's probability to break a lattice bond. As will be shown, again in Chapter Four, this probability folded together with the energy distribution can explain the phenomenon of V_{xover}.

3.4 Historical review

Wolff [Wolf 54] was the first to explore the phenomenon of impact ionization in silicon. Concentrating on the high-field regime, in excess of 2 X 10^5V/cm, he arrived at a form for α, the ionization rate:

$$\alpha \sim \exp(-\frac{\text{const}}{\mathcal{E}^2}) \tag{3.1}$$

For future reference, note that MOSFET substrate current is related to I_D by $I_B \sim \alpha\ I_D$.

$\mathcal{E}$ is the local electric field. Wolff's treatment assumes 'equilibrium' field. By equilibrium, he and most other researchers mean constant in real space, not necessarily in time as is customary. In addition, it was a global model for the bulk semiconductor, and so did not address issues pertinent to MOSFET operation, where knowledge of local fields, mobility, and ionization are essential for effective simulation. Finally, the range of fields explored occurs only for sub-micron devices following constant-voltage scaling (e.g., L_e=.7μm at V_D=5V). Ideally, regions of a device containing such fields should be quite small for a reliable, logic technology.

Chynoweth [Chyn 58] followed similar assumptions as Wolff, but pursued lower fields more common in MOSFETs. His expression for the ionization coefficient is:

$$\alpha \sim \exp(-\frac{\text{const}}{\mathcal{E}}) \tag{3.2}$$

This formulation is more appropriate than Wolff's: the typical ionization efficiency in a MOSFET is less than 0.01; which is the region where Chynoweth's analysis is most accurate.

Keldysh [Keld 60] extended the above treatments to finite temperature, and transcended the region of electric field between them; this led to a form for the ionization coefficient of:

$$\alpha \sim \exp(-\frac{E_i S(T,\mathcal{E})}{q\lambda\mathcal{E}}) \tag{3.3}$$

E_i is the ionization threshold, and λ is the mean free path. The S function is solved using a transcendental equation, which unfortunately is not useful for device analysis.

Shockley [Shoc 61] arrived at a result similar to [Chyn 58], using physical arguments related to the probability that a carrier could be lucky enough to travel ballistically for many mean free paths without scattering - and have, at the end of travel, enough energy to break a Si-Si bond. With λ as the mean free path, and x the distance travelled, the ionization rate then looks like:

$$\alpha \sim \exp(-\frac{x}{\lambda}) \tag{3.4}$$

Baraff [Bara 62] solved the Boltzmann equation numerically, and was able to demonstrate the connection between the Wolff and Chynoweth/Shockley limits for the ionization rate, over the full range of field. His assumptions also included equilibrium field. His treatment was numerical, and so obscures the mechanisms of the impact ionization process. However, the result yielded a set of universal curves, shown in Figure 3.1, which can be useful guides to α values as a function of electric field.

Crowell and Sze [Crow 66] plotted some of Baraff's results at low temperature, using an analytic expression for the optical phonon scattering mean free path (see Equation 4.5). This is shown in Figure 3.2. While quite good, the fit begins to be low by in excess of 5-10% at 100K near the electric field strengths of interest in MOSFET's. This discrepancy will be important in Chapter Four, when the modelling of substrate current is detailed.

None of the above treatments allows for the possibility of measurable I_B at low-V_D. Each assumes a carrier begins its travel through the equilibrium field at the minimum of the band energy. So, if the total potential drop in a MOSFET from drain to source is less than that needed to break a Si-Si bond, each would predict zero substrate current. Clearly, based on the results of Chapter Two, this prediction is not supported by measurement, which leads to the development of an energy distribution at any point in the semiconductor, as follows in the next major section.

Ridley [Ridl 83a] was the first to show the dominant impact ionization processes in an

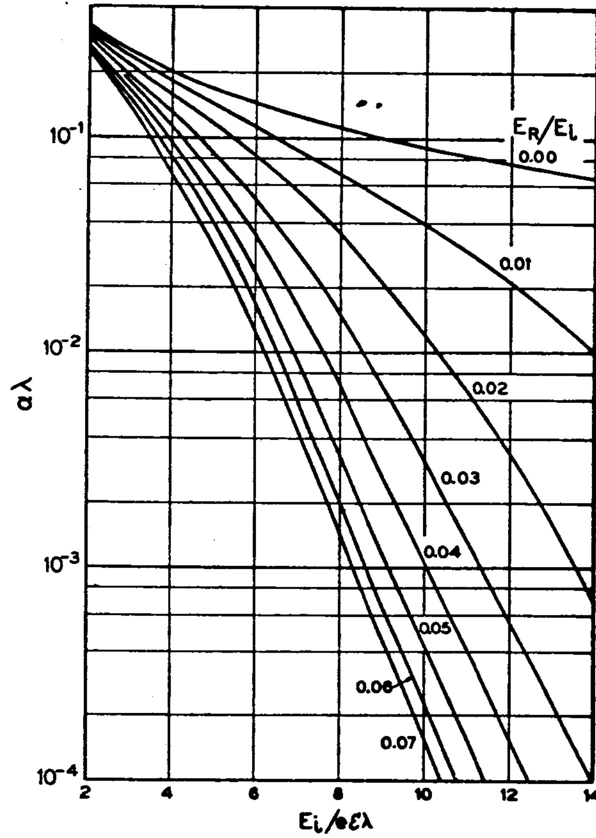


Figure 3.1: Baraff's ionization rate curves. E_i is the ionization threshold. E_R is the mean optical phonon energy.

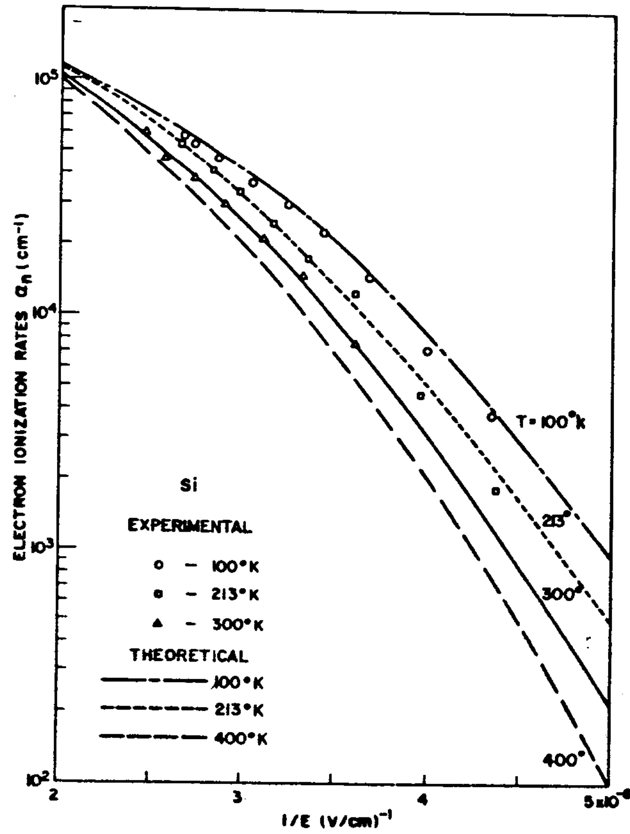


Figure 3.2: Ionization rate temperature dependence. The solid line is Baraff's universal curve. After [Crow 66].

analytical fashion, without obscuring the physics. The four important physical mechanisms are depicted in Figure 3.3. Comparison of his analytical result with Baraff's numerical one is quite favorable. Again, an assumption of equilibrium field is implicit to his treatment. It should be noted that Ridley ascribes the dominant contribution to α over the entire range of field to the processes involving lucky-drift, Figures 3.3b and 3.3d.

The intent of Ridley and others following this approach is to discern some universal traits of homogeneous and compound semiconductors with respect to impact ionization [Ridl 83b]. For instance, a universal relationship was determined relating the mean free path to the ionization threshold in a semiconductor, $\lambda \sim E_i^{-1/2}$. The simple theory, however, overpredicted the ionization coefficient in the silicon system for fields commonly found in submicron MOSFET's. Attempts were then made to modify the lucky-drift model to include the possibility for 'soft' ionization thresholds: thresholds which themselves varied with the energy of an incident carrier [Mars 87,Ridl 87,Wood 87]. However, the fitting procedures then followed shed little additional light on the subject.

Aside from the usual carrier-phonon scattering mechanisms, ionization rates have also been investigated theoretically from the standpoint of carrier-carrier scattering [Sing 85]. While this mechanism appears to be important in explaining observations in GaAs at fields in excess of 1MV/cm, it has not been investigated in silicon.

3.5 The energy of a carrier

3.5.1 Derivation of the energy distribution

Beginning with the Boltzmann Transport Equation (BTE) [Bube 74,Swan 84], one may immediately use its steady-state form, so that the time-derivative of the distribution function is zero. Then, assuming zero magnetic field:

$$\frac{\partial f}{\partial t}|_{scat} = \frac{e}{\hbar}\mathcal{E} \cdot \nabla_k f + \mathbf{v} \cdot \nabla_r f \tag{3.5}$$

f is the distribution function, a function of both k-space (momentum) and r-space (position) coordinates. The gradients have the usual interpretation. The LHS includes scattering events in the thermodynamic volume under examination. The RHS includes

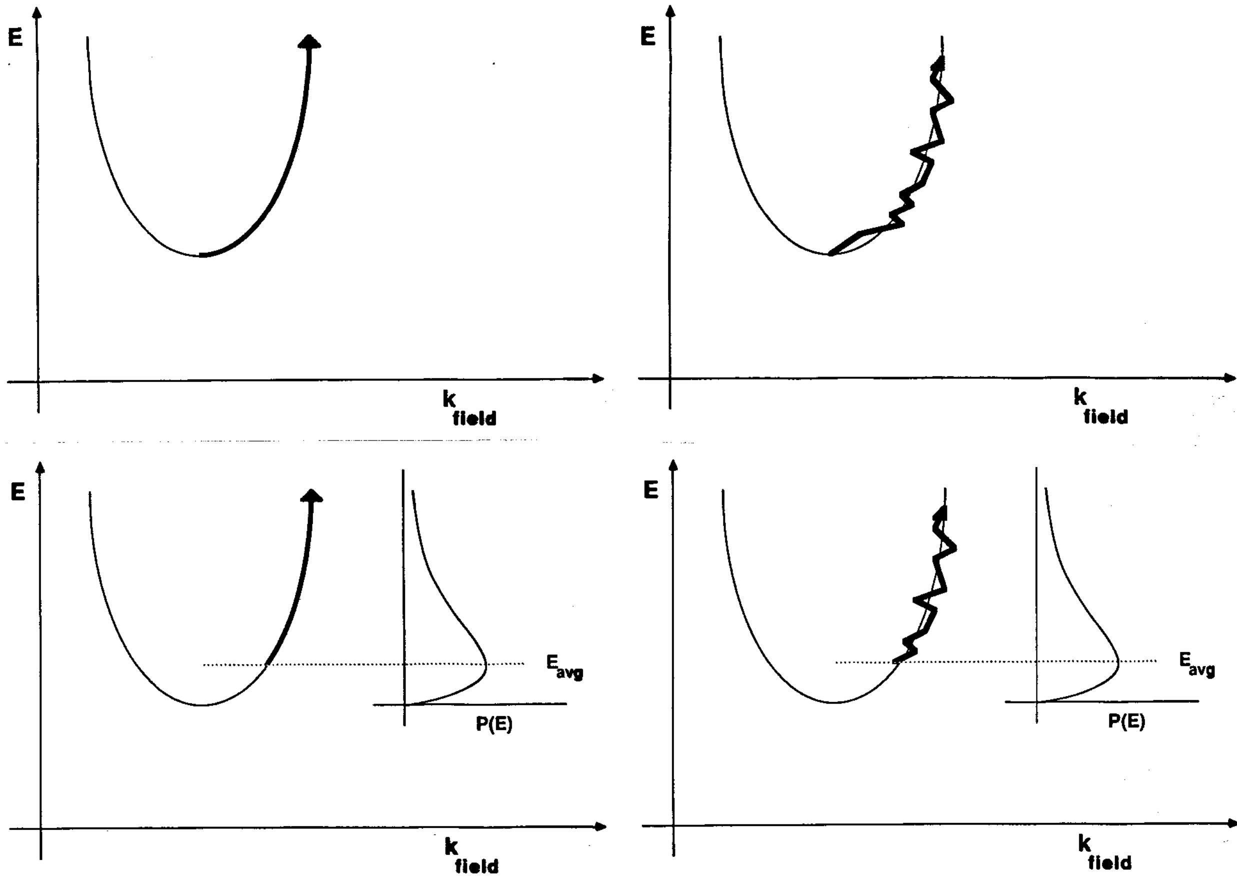


Figure 3.3: Single-particle processes leading to impact ionization. After [Ridl 82]. Lucky-drift is distinguished from ballistic, or lucky-electron, flight in that mostly-elastic scattering is allowed which re-directs the carrier momentum without reducing markedly the carrier energy. Upper left: pure lucky-ballistic. Carriers suffer no scattering in attaining the ionization threshold. Upper right: pure lucky-drift. The jagged path represents scattering to other states in E-k space, but off the ballistic or unscattered path parallel to the field. Thus, momentum may relax, but not energy. The momentum relaxation time is much less than the energy relaxation time. Lower left: lucky-ballistic motion from $2k_BT_e$, the average energy after the carrier is "thermalized": after the carrier reaches a steady energy state between field gain and phonon loss. Lower right: lucky-drift from the thermalized average.

forcing or driving terms due to, first, drift, and second, diffusion. Note the assumptions inherent in this equation. First, the effective mass approximation is inherent to the use of ∇_k, especially when momentum coordinates are transformed into energy through:

$$E(k) = \frac{\hbar^2 k^2}{2m^*} \tag{3.6}$$

The wave-particle duality is inherent in the definition of velocity:

$$\mathbf{v} = \frac{1}{\hbar}\nabla_k E(\mathbf{k}) = \nabla_k \omega \tag{3.7}$$

This is, however, just the group velocity of a quantum-mechanical wave packet. The final important assumption concerns the expectation value of the force on the wave packet, which must give an identical result as if the packet were treated as a classical particle. With these assumptions, the BTE (itself a classical equation) may be employed.

Assume that the electric field and thermal-diffusive forces are in the direction of the current flow; that is, one-dimensional. This yields:

$$\frac{\partial f}{\partial t}|_{scat} = \frac{e\mathcal{E}}{\hbar}\frac{df}{dk} + v\frac{df}{dx} \tag{3.8}$$

Write the distribution function as an equilibrium part, and a departure from equilibrium:

$$f = f_0(E) + f_1(k, x) \tag{3.9}$$

The relaxation time approximation is employed for the scattering term:

$$\frac{\partial f}{\partial t}|_{scat} \approx -\frac{f_1(k,x)}{\tau} \equiv \frac{\phi(k,x)}{\tau}\frac{df_0}{dE} \tag{3.10}$$

τ is the scattering time in this approximation. This leaves the BTE as:

$$\frac{\phi}{\tau}\frac{df_0}{dE} = \frac{e\mathcal{E}}{\hbar}\frac{df}{dk} + v\frac{df}{dx} \tag{3.11}$$

Next, the momentum and spatial derivatives of f need to be found. First, assume the spatial derivatives of f_1 are small compared to those of f_0, and may be neglected. Then the chain rule says:

$$\frac{df_0}{dx} = \frac{df}{dT}\frac{dT}{dx} \tag{3.12}$$

Next, the temperature derivative must be determined. From thermodynamics, it is known the distribution function depends on $(E - E_F)/T$, where E and T are independent state variables, and E_F is the Fermi level. So:

$$f_0 = f_0(\frac{E - E_F}{T}) \tag{3.13}$$

Again, applying the chain rule:

$$\frac{df_0}{dT} = -\frac{df_0}{dE}[\frac{E - E_F}{T} + \frac{dE_F}{dT}] \tag{3.14}$$

So:

$$\frac{df_0}{dx} = -\frac{df_0}{dE}[\frac{E - E_F}{T} + \frac{dE_F}{dT}]\frac{dT}{dx} \tag{3.15}$$

For df/dk, again using the chain rule and neglecting the f_1 derivative:

$$\frac{df}{dk} = \frac{df_0}{dE}\frac{dE}{dk} \tag{3.16}$$

Pooling terms:

$$\frac{\phi}{\tau}\frac{df_0}{dE} = \frac{e\mathcal{E}}{\hbar}\frac{dE}{dk} - v\frac{df_0}{dE}[\frac{E - E_F}{T} + \frac{dE_F}{dT}]\frac{dT}{dx} \tag{3.17}$$

This immediately defines ϕ. Neglecting dE_F/dT and using Equation 3.7:

$$\phi = \tau v[e\mathcal{E} - \frac{E - E_F}{T}\frac{dT}{dx}] \tag{3.18}$$

Neglecting the diffusion term due to the thermal gradient, the total distribution function is:

$$f = f_0 - \tau v e\mathcal{E}\frac{df_0}{dE} \tag{3.19}$$

With f, any measurable parameters of interest may be calculated. Of particular concern for our later modelling of substrate current is the electron current density. Noting that the

integral over k-space of the f_0 portion of f goes to zero because the integrand is odd, and changing variables in the standard way from crystal momentum to energy, one may define:

$$J_e = -e \int_{E_C}^{\infty} f_1 v N dE \tag{3.20}$$

$N(E)$ is the energy density of states over the Brillouin zone. For a semiconductor like silicon or germanium, the density of states can be written as $N(E) = C(E - E_C)^{1/2}$, where C is a constant of energy and E_C is the conduction band minimum. For a non-degenerate semiconductor, we can use the Boltzmann distribution for f_0:

$$f_0 = \exp[-\frac{E - E_F}{k_B T}] \tag{3.21}$$

Then, substituting into Equation 3.20:

$$J_e = -\frac{e^2 \mathcal{E}}{k_B T} \exp[-\frac{E_C - E_F}{k_B T}] C \int_0^{\infty} \tau(E) v^2 E^{1/2} \exp(-E/k_B T) dE \tag{3.22}$$

For longitudinal acoustic scattering, $\tau(E) = A\,E^{-1/2}$. Then, using Equation 3.7 once more:

$$J_e = -\frac{2e^2 \mathcal{E} k_B T}{m^*} \exp[-\frac{E_C - E_F}{k_B T}] C A \int_0^{\infty} \frac{E}{k_B T} \exp(-E/k_B T) \frac{dE}{k_B T} \tag{3.23}$$

If the pre-factor to the integral is written as J_0, then the amount of current between E and $E + dE$ is:

$$J(E) dE = J_0 \frac{E}{(k_B T)^2} \exp(-E/k_B T) dE = J_0 P(E) dE \tag{3.24}$$

Equation 3.24 is principle to the modelling work presented in Chapter Four. The energy distribution will allow a physical explanation of both V_{xover}, and the observation of substrate current for V_D less than the ionization threshold. Furthermore, it is required for explanation of the gate current phenomena presented in Chapter Five.

3.5.2 The average carrier temperature

The function $P(E)$ in Equation 3.24 at this point includes only the lattice temperature. As detailed in the review of assumptions below, this is primarily due to the inclusion of

longitudinal acoustic scattering only. However, experiments and previous theory [Hess 78] have demonstrated that optical phonon scattering is the principle mechanism for high-energy carriers. The following Ansatz is therefore made: to include the effects of saturated carrier velocity and optical phonon scattering, the energy distribution is described by a carrier temperature. This temperature, denoted T_e, may be different than the lattice temperature, especially at high fields.

The Ansatz is employed to keep the simplicity of the BTE energy distribution derived above, which lends itself very easily to inclusion in 2-D, numerical device simulators. In particular, the integral of $P(E)$ can be solved analytically for any integral limits.

Despite this Ansatz, an expression for T_e is still needed. Again, the method of [Bube 74] is followed, with one significant departure. The derivation is lengthy; thus, only the principle features will be given.

The essence of finding the carrier temperature is to balance the rate of energy gain from the electric field by a carrier with the rate of energy loss due to phonon emission. The rate of energy gain is:

$$\frac{dE}{dT} = e\mathcal{E}\cdot\mathrm{v} = e\mu|\mathcal{E}|^2 \tag{3.25}$$

The average energy change due to scattering for all carriers is:

$$\frac{dE}{dT} = \frac{\int_0^\infty(\frac{v}{\lambda}\langle\delta E\rangle_{avg})N(v)dv}{\int_0^\infty N(v)dv} \tag{3.26}$$

v/λ is the scattering rate. $N(v)$ is the number density of carriers with respect to velocity, similar to the energy distribution of carriers derived in Equation 3.24:

$$N(v)dv = \mathrm{C}\cdot v^2\exp[-\frac{E}{k_BT_e}]dv \tag{3.27}$$

In steady-state, the energy gain equals the loss. So:

$$e\mu\mathcal{E}^2 = -\frac{8\bar{v}^2(2\pi m^*k_BT_e)^{1/2}}{\pi\lambda T}(T-T_e) \tag{3.28}$$

$\bar{v}$ is the velocity of longitudinal sound in the semiconductor.

For longitudinal acoustic phonon scattering, the low- and high-field mobilities may be written:

$$\mu_0 = \frac{4e\lambda}{3(2\pi m^* k_B T)^{1/2}} \tag{3.29}$$

$$\mu = \frac{4e\lambda}{3(2\pi m^* k_B T_e)^{1/2}} \tag{3.30}$$

Combining these two equations with Equation 3.28 gives the final result:

$$\frac{T_e}{T} = \frac{1}{2}\{1 + [1 + \frac{3\pi}{8}(\frac{\mu\mathcal{E}}{\bar{v}})^2]^{1/2}\} \tag{3.31}$$

The full, not the low-field, mobility has been included in this equation, to account for the approach of T_e to a constant value for very high electric field. This must occur due to the saturation of the carrier velocity when optical phonon emission is included.

Thus, both the energy distribution $P(E)$ and the carrier temperature T_e have been defined. These will be crucial to the modelling of impact ionization substrate current demonstrated in Chapter Four.

3.6 Recapitulation

3.6.1 Assumptions in energy distribution

The derivation of the energy distribution of the channel carriers in a device has several limiting assumptions which should be noted. First, using the BTE at all assumes quantum mechancial packets of charge may be treated semi-classically, substituting the effective mass for the normal charge mass to account for the band structure of the semiconductor. The wave packet width estimates from the momenta found in MOSFET's are quite small relative to device channel lengths, making this a reasonable assumption. However, in MOS inversion layers the carriers are confined in one dimension. This affects the charge distribution near the semiconductor-insulator interface; in particular, the quantum mechanical charge must go to zero at the interface - while it reaches a maximum if treated classically. As discussed further in Chapter Five, such changes in the charge density will affect the channel hot carrier (CHC) injection into the MOS gate. In addition, the inversion mobility will be alterred if the confining field breaks the symmetry of the crystal, and alters the effective mass.

Use of the relaxation time approximation assumes the energy change on scattering is small related to k_BT. For longitudinal acoustic scattering, this is a valid assumption. However, for optical phonon scattering, the change in energy on scattering may exceed the energy derived from the lattice. The Ansatz employed in using T_e , however, ensures that the scattering energy change will be small relative to k_BT_e, even if optical phonon and impact ionization scatterings are included. Furthermore, for optical phonon scattering in silicon the phonon energy is only about $3k_BT$ at room temperature, which is still a small energy change.

The one-dimensional approximation for the BTE is justified, in that MOSFET substrate current will be modelled by looking at charge transport along the current contours. Ignoring diffusion for the moment (see below), the impact ionization rate will be calculated by following a charge packet along each current contour. In essence, the Cartesian space current contours are transformed into one-dimensional paths along an axis parallel to the electric field.

Only electric field drift was used in the derivation, which thus neglects diffusion components of the current. Replacing the electric field by the gradient of the quasi-Fermi level removes this difficulty, while maintaining the generality of the derivation.

The effective mass approximation was employed extensively. However, it is not clear that hot carriers lie on spherical energy bands, or what the degeneracy factors for these high-energy states in the conduction or valence bands are.

Non-degenerate placement of the Fermi level relative to the appropriate channel carrier band was assumed; this allowed the use of the Boltzmann distribution. This assumption is probably invalid in MOS inversion layers, which would then require full Fermi-Dirac statistics. Channel hot carrier gate current would be most likely affected. For such gate current, however, only the tail of the distribution has enough energy to surmount the large barriers seen in semiconductor-insulator systems. This tail can be described by a Boltzmann distribution as well, making this assumption seem non-critical. Long-channel substrate current also seems to be affected.

f_1 was assumed to be a small departure from equilibrium. This assumption allowed neglect of terms like df_1/dx, compared to df_0/dx. However, in a short-channel MOSFET with small gate and high drain biases, the lateral electric field is changing rapidly, and

departures from equilibrium may be much larger [Maha 85]. The effect is to push the average energy of the carrier distribution to higher energies. However, the analytic treatment used above becomes invalid, so that modelling is no longer a simple and computationally-compact effort (see Chapter Four).

Thermal gradients have been ignored in the derivation. Again, for a short-channel MOSFET T_e changes rapidly near the drain. Meinerzhagen and Engl [Mein 86] have indicated the thermal gradient may be important in treating hot carriers from the standpoint of impact ionization. Their treatment seeks to find the total ionization rate at a point from consideration of first principles only. The treatment here follows [Thur 85], in that non-local parameters affect the impact ionization at a local point. These distinctions will be discussed further in Chapter Four.

Longitudinal acoustic scattering was assumed in using the energy-dependent scattering time of Equation 3.22. Combined with the functional dependence of the density of states, this led to the simple expression for the carrier energy distribution in Equation 3.24. Inclusion of other scattering effects such as optical phonon scattering would then lead to a different form for the energy distribution: except that the full mobility is used in the T_e equation, to account for the important role of optical phonons in hot carrier processes.

3.6.2 Assumptions in carrier temperature

λ is independent of the carrier kinetic velocity. This allows a simple, analytic solution to the integral in Equation 3.26.

The use of the full mobility in Equation 3.31, and not the low-field mobility, is self-consistent with the Ansatz employed in using T_e instead of the lattice temperature to describe the carrier energy distribution. This is because optical phonon scattering is the dominant scattering mechanism for hot carriers [Chwa 79,Shic 81]. For T_e and the mobility to be consistent, then, the full mobility including velocity saturation - also due to optical phonon scattering - must be used.

3.6.3 Failings of previous models

The previous models of impact ionization have several shortcomings. First, they assume the carriers are by-and-large at the energy band minimum. With such a picture, no carrier could ionize a lattice bond if the driving potential were less than the ionization threshold energy. Second, the models assumed constant electric field in calculating the ionization α. This has the problem of neglecting the effect on local generation of electron-hole pairs of high-energy charge packets flowing into the region from elsewhere in the device. Furthermore, it neglects the effects of a fast-changing electric field near the drain of such a device. Finally, when these models are applied to simulation of impact ionization in MOSFET's, fitting parameters are required to explain the observations as functions of technology [Wern 84] and temperature [Lau 85].

3.7 Summary

This chapter has given an overview of the works preceding ours, giving rise to the need for further contributions. The notion of carrier energies beyond the band minimum has been promoted. An appropriate distribution of carrier energies has been derived using Boltzmann's Equation. The carrier temperature, or average energy, characteristic of this distribution has also been derived.

The proposed energy distribution has the proper characteristics to explain the observations of substrate current in MOSFET's as a function of bias and temperature, detailed in Chapter Two. The use of an energy distribution is needed to explain the observation of substrate current in MOSFET's at drain-to-source biases less than the ionization threshold. The distribution derived in this Chapter at least qualitatively explains such an observation. The energy distribution can explain the decrease of MOSFET I_B at low V_D with decreasing temperature, as well. As temperature decreases, the number of carriers in the tail of the distribution also decreases. At low V_D, these are the only carriers which have the right additional energy to break a lattice bond. This effect counterbalances the increased ballistic energy a carrier obtains, due to its enhanced mean free path at lower temperature. Both these effects will be discussed in detail and demonstrated quantitatively in Chapter Four.