Investigations of Quantum Effect Semiconductor Devices – The Tunnel Switch Diode and the Velocity Modulation Transistor

Thesis by

Erik S. Daniel

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

1998
(Submitted November 24, 1997)

© 1998
Erik S. Daniel
All Rights Reserved

To my family

List of Publications

Work related to this thesis has been or will be presented in the following papers:

Macroscopic and Microscopic Studies of Electrical Properties of Very Thin Silicon Dioxide Subject to Electrical Stress,
E. S. Daniel, J. T. Jones, O. J. Marsh, T. C. McGill, *J. Vac. Sci. Technol. B* **15**, pp. 1089-1096, 1997.

A Transistorless Current Mode Static RAM Architecture,
H. J. Levy, E. S. Daniel, T. C. McGill, accepted for publication in IEEE J. Sol. St. Circ.

Dependence of the *I-V* curve of a Metal Insulator Semiconductor Switch on Insulator Thickness – an Experimental and Theoretical Investigation,
P. O. Pettersson, A. Zur, E. S. Daniel, H. J. Levy, O. J. Marsh, T. C. McGill, submitted to IEEE Tr. Elec. Dev.

Dual Mode Behavior of Polysilicon Gate Tunnel Switch Diodes,
E. S. Daniel, T. C. McGill, in preparation.

Drift-Diffusion Simulations of the Tunnel Switch Diode,
E. S. Daniel, D. Z.-Y. Ting, T. C. McGill, in preparation.

Mobility Modulation and Novel Transport Phenomena in a GaSb/InAs/AlSb Double Quantum Well System,
E. S. Daniel, D. A. Collins, D. Z.-Y. Ting, T. C. McGill, in preparation.

Acknowledgements

It would be impossible to give due credit in the few pages allowed here to everyone who has supported me in the pursuit of my doctoral degree, so I hope to convey my gratitude to most in person. I would, however, like to formally thank at least a few of those who have helped me through the years.

Describing Thomas McGill in a few sentences is a bit like trying to replicate a Rembrandt with a paint roller – you could probably get the job done, but you could never reflect a decent likeness of the original. He has incredibly sharp instincts, an amazing awareness of and respect for his students' abilities, and a healthy laugh. While he is demanding of himself and of his students, he also understands the value of a life outside the lab. Few research advisors can provide the kind of environment that Tom does, in which students are tossed into a sea of equipment and encouraged to learn and pursue whatever they find interesting. While I've learned a fair amount of Physics from Tom, there are a number of other, perhaps more useful things I have learned from him over the years: how to give a talk on a subject you know nothing about, how to roll with the punches, how to sniff out good research directions, and how to sniff out a good Mexican restaurant in an unfamiliar town.

Perhaps Tom's greatest strength is the ability to attract and train an incredibly talented and incredibly diverse group of students and staff. I am indebted to many members of this group as they have contributed more to my education than I ever would have expected. Nearly every skill I have picked up during my term as a graduate student, I have learned from Harold Levy, David Ting, and Doug Collins. Harold is probably the most brilliant, creative person I have ever met. His broad interests, his infinite patience, and his creative interpretations of the English language continue to influence me. David's ability to convey his vast knowledge of solid-state physics to others is truly remarkable, and his leadership during his years as a postdoc in the group has helped a large number of students to develop and to succeed. I wish I had

inherited Doug's awesome ability to perform miracles in the lab, but I consider myself lucky just to have been able to watch him in action and to learn what I could before some lucky company stole him away from us.

I have had the pleasure of working with a number of other students in the research group: Mike Wang, Peo Petterssen, Chris Springfield, Xiao-Chang Cheng, Eric Piquette, Alicia Alonzo, Paul Bridger, Zvonimir Bandic, Joel Jones, and Robert Beach. Each of these talented people has had a distinct and notable effect on me, and though I could elucidate on my experiences with each one, I'll only say that I wish each of them the best, and I hope they have benefitted from our interactions a fraction as much as I have. I have also enjoyed getting to know some of the newer students: Cory Hill, Matt Barton, Xavier Cartoixa, and Robert Goettler.

The group is blessed with a group of three true scholars who constantly provide wisdom, encouragement, and direction to the rest of us. Ogden Marsh must spend every waking hour scouring every scientific journal known to man, looking for nuggets of information relevant to the group's activities. Gerry Picus has helped to keep us all appraised of the goings-on in the real world by arranging a marvelous series of video conference lectures from outside sources. Jim McCaldin seems to have a personal rapport with every element of the periodic table, almost as if they were all old drinking buddies. Each of these sages has steered me away from disaster at one time or another, and for that I will always be in their debt.

I could not have remained sane during my years as a graduate student without a number of terrific friends dragging me off to have fun once in a while. While I have enjoyed the company of a large number of interesting people during these years, three in particular, Steve Ouellette, Adam Leibovich, and Torrey Lyons, have been keeping me entertained and relatively stress-minimized since I began my graduate career. Thanks for all of the hours of entertainment, all of your advice, and all those pennies.

Most of all, I thank my family for supporting me throughout my student years. In particular, I thank my wife, Tanja, for being patient with me, and for giving me a good kick in the rear when I needed it.

Abstract

In this thesis we present the results of experimental and theoretical studies of two quantum effect devices – the Tunnel Switch Diode (TSD) and the Velocity Modulation Transistor (VMT).

We show that TSD devices can be fabricated such that they behave (semi-quantitatively) as predicted by simple analytical models and more advanced drift-diffusion simulations. These devices possess characteristics, such as on-state currents which range over nearly five orders of magnitude, and on/off current ratios which are even larger, which may allow for a practical implementation of a very dense transistorless SRAM architecture and possibly other novel circuit designs.

We demonstrate that many TSD properties can be explained by analogy to a thyristor. In particular, we show that the thin oxide layer in the TSD plays a critical role, and that this can be understood in terms of current injection through the oxide, analogous to transport through the "current limiting" layer in a thyristor. As this oxide layer can be subjected to extreme stress during device operation, we have studied the effect of this stress on device behavior. We demonstrate many significant stress-dependent effects, and identify structures and operation modes which minimize these effects.

We propose an InAs/GaSb/AlSb VMT which may allow for larger conductance modulation and higher temperature operation than has been demonstrated in similar GaAs/AlAs structures. Fundamental differences in device operation in the two materials systems and unusual transport mechanisms in the InAs/GaSb/AlSb system are identified as a result of the band lineups in the two systems. Boltzmann transport simulations are developed and presented, allowing a qualitative description of the transport in the InAs/GaSb/AlSb structure. Band structure calculations are carried out, allowing for device design.

While no working VMT devices were produced, this is believed to be due to

processing and crystal growth problems. We present methods used to overcome or circumvent a number of processing problems which resulted in shorting of the gate to the source/drain contacts. The results of electrical measurements are given, which may allow for identification of further obstacles to the production of working devices.

Contents

List of Publications iv

Acknowledgements v

Abstract vii

1 Introduction 1

1.1 Introduction to Thesis 1

1.2 Summary of Results 1

1.3 Motivation 2

1.3.1 The Tunnel Switch Diode 3

1.3.2 The Velocity Modulation Transistor 6

1.4 Chapter Outline 7

Bibliography 9

I Experimental and Theoretical Investigations of Si-based Tunnel Switch Diodes (TSD) 10

2 Background and Motivation for Study of the TSD 11

2.1 Introduction and Chapter Outline 11

2.2 Introduction to the Tunnel Switch Diode (TSD) 11

2.2.1 Background 11

2.2.2 Basic TSD Device Operation 13

2.3 Applications of the TSD 19

2.3.1 Transistorless SRAM Application 19

2.3.2 Other Applications 25

2.4 Properties and Analysis of Thin Oxides 27
2.4.1 Background 27
2.4.2 Electrical Properties and Characterization 27
2.4.3 Structural Properties and Characterization 35
2.4.4 Summary 37

Bibliography **38**

3 Device Design and Experimental Procedures **43**
3.1 Introduction and Chapter Outline 43
3.2 Design of Target Structures and Description of Commercial Wafers Used 43
3.3 Fabrication Procedures 45
3.3.1 Device Geometry 45
3.3.2 Process Steps Used at Caltech 47
3.3.3 Process Steps Used at Stanford Nanofabrication Facility . . 50
3.3.4 Process Steps Used at Intel Components Research 53
3.4 Measurements and Characterization 54
3.4.1 Simple I-V Measurements 55
3.4.2 C-V Measurements 57
3.4.3 Electrical Stressing 60
3.4.4 Multiprobe I-V Measurements 61
3.4.5 Ellipsometry 67

Bibliography **68**

4 Experimental Results **70**
4.1 Introduction and Outline 70
4.2 Effects of Stress on TSD Devices 70
4.2.1 Constant and Incremental Current Stressing 71
4.2.2 Voltage Limiting Behavior 74
4.2.3 Dual Mode Behavior of Poly Gate Devices 77

4.3 Scaling Behavior of TSD Devices 80
4.3.1 Introduction and Early Results 80
4.3.2 Stress-Induced Modification of Scaling Behavior 82
4.3.3 Comparison of Scaling of Field Oxide and Flat Devices . . . 84
4.3.4 Scaling of High and Low Current Modes of Poly Gate Devices 88
4.3.5 Scaling Behavior of Avalanche Mode Devices 94
4.4 Dependence of TSD *I-V* Curve on Design Parameters 97
4.4.1 Oxide Thickness 98
4.4.2 Epilayer Properties 102
4.5 Summary and Discussion 106

Bibliography **108**

5 Simulations **109**
5.1 Introduction and Outline 109
5.2 Drift-Diffusion Simulations 109
5.2.1 Basic Semiconductor Equations 109
5.2.2 Models of Physical Parameters 113
5.2.3 Discretization of Equations 115
5.2.4 Solution of Discretized Equations - Relaxation Method . . . 118
5.2.5 Matrix Methods 121
5.2.6 Tests and Results 123
5.3 Tunnel Current Simulations 124
5.3.1 General Formalism 124
5.3.2 Calculation of Transmission Probabilities 128
5.3.3 Tests and Results 131
5.4 Union of Drift-Diffusion and Tunnel Simulations 132
5.5 Results 133
5.5.1 Studies of *pnpn* Thyristors 133
5.5.2 Studies of Thin Oxide MOS Devices 139
5.5.3 Studies of the TSD 143

5.6 Summary and Discussion 149

Bibliography **153**

II Experimental and Theoretical Investigations of an InAs/GaSb/AlSb Double Quantum Well Velocity Modulation Transistor (VMT) **156**

6 Background **157**

6.1 Introduction and Outline 157

6.2 The Velocity Modulation Transistor (VMT) 157

6.2.1 Device Concept 157

6.2.2 Simple Performance Estimate 159

6.2.3 GaAs Implementations of the VMT 160

6.3 An InAs/GaSb/AlSb Implementation of the VMT 161

6.3.1 Basic Concept 161

6.3.2 Comparison of GaAs/AlGaAs and InAs/GaSb/AlSb Implementations 164

6.3.3 Transport Mechanisms 166

Bibliography **169**

7 Modeling **171**

7.1 Introduction and Outline 171

7.2 Band Structure Calculations 171

7.2.1 Description of Calculations 171

7.2.2 Discussion of Results and Design Considerations 173

7.3 Boltzmann Transport Studies 178

7.3.1 Simple Band Model 178

7.3.2 Boltzmann Transport Formulation 181

7.3.3 Results 184

Bibliography **188**

8 Experiment **189**

8.1 Introduction and Outline 189

8.2 Crystal Growth and Characterization 189

8.3 Processing Techniques 191

8.3.1 Early Attempts 191

8.3.2 SiO_2 Buffer Layer Scheme 194

8.3.3 Modified Mask Designs 196

8.3.4 Refined Fabrication Steps 196

8.4 Results and Discussion 199

8.4.1 Characterization of Samples 199

8.4.2 Discussion of Related Work at NRL 203

Bibliography **207**

A Resistive Network Analysis **209**

Bibliography **213**

B Matrix Norms and Condition Numbers **214**

Bibliography **218**

List of Figures

1.1 N-type and S-type Negative Differential Conductance 4

2.1 Tunnel Switch Diode 12
2.2 Basic TSD Device Operation 14
2.3 Effect of Light Injection on TSD *I-V* Curve 16
2.4 Calculated TSD Peak Voltages 18
2.5 Transistorless SRAM Static/Read Scheme 21
2.6 Transistorless SRAM Write Scheme 22
2.7 Density Maximizing TSD SRAM Fabrication Scheme 23
2.8 Comparison of Thick and Thin Oxide MOS Devices 29
2.9 *C-V* Curves for a Thick Oxide MOS Device 31
2.10 *C-V* Analysis of Thin Oxide MOS Devices 34

3.1 Flat and Field Oxide Device Geometries 47
3.2 LOCOS Procedure 52
3.3 Experimental Setup for TSD Electrical Measurements 56
3.4 Parasitic Parallel Impedances in Electrical Measurement Setup . . . 58
3.5 Multiple Probe Measurement Setup 62
3.6 Typical Measured Multiprobe *I-V* Curve 64
3.7 Resistive Network Model 64
3.8 Measured Gate Voltage Drop 66

4.1 Constant Current and Incremental Current Stressing of Poly Gate TSD Devices 72
4.2 Constant Current Stressing of Al Gate TSD Devices 73
4.3 Voltage Limiting Behavior During Incremental Current Stressing . . 75
4.4 Analysis of Limiting Voltage Versus Oxide Thickness 76

4.5 Dual Mode Behavior of Poly Gate TSD Devices 78
4.6 Early Scaling Results 81
4.7 Low Voltage Scaling Behavior of the TSD with Light Injection . . . 83
4.8 Scaling of Flat Al Gate Devices 86
4.9 Scaling of Field Oxide Al Gate Devices 87
4.10 Scaling of Low Current Mode of 4.5 nm Oxide Poly Gate Devices . 89
4.11 Survey of Poly Gate Devices in Low Current Mode 90
4.12 Survey of Poly Gate MOS Devices 92
4.13 Survey of Poly Gate Devices in High Current Mode 93
4.14 Survey of Poly Gate Avalanche Devices 95
4.15 Scaling of Stanford Flat and Field Oxide Avalanche Devices 96
4.16 Current Spreading in the TSD Epi Layer 98
4.17 Dependence of Low and High Current Mode Poly Gate Punch-through Devices on Oxide Thickness 99
4.18 Dependence of Poly Gate Avalanche Devices on Oxide Thickness . . 100
4.19 Survey of Stanford Wafer Punch-through and Avalanche Devices . . 101
4.20 Survey of Caltech Punch-through Devices 103
4.21 Comparison of 1.5 μm and 2.5 μm Epi Thickness Punch-Through Devices 104

5.1 Relaxation Method in One Dimension 119
5.2 Simulation of Short p-n Junction 125
5.3 Comparison of Tunnel Current Calculation to Measured Data 132
5.4 Comparison of TSD and Thyristor Devices 135
5.5 Calculated Thyristor I-V Characteristics 136
5.6 Calculated Thyristor Band Diagrams 138
5.7 Calculated Thin Oxide MOS Device Band Diagrams 142
5.8 Dependence of Calculated TSD J-V Curves on Recombination Lifetimes 145
5.9 Calculated Band Diagrams for a 2.5 nm Oxide TSD 147

5.10 Dependence of Calculated J-V Curves on Oxide Thickness – Fixed Lifetimes 148
5.11 Dependence of Calculated J-V Curves on Oxide Thickness – Variable Lifetimes 150

6.1 The Velocity Modulation Transistor 158
6.2 Measured Response of GaAs/AlGaAs VMT 161
6.3 Energy Band Properties for III-V Heterojunctions 162
6.4 Proposed InAs/GaSb/AlSb VMT Structure 163
6.5 Comparison of Homogeneous and Inhomogeneous Double Well VMT Systems 165
6.6 Simple Transport Model Applied to the VMT Systems 167

7.1 Calculated Sub-bands of InAs/GaSb/AlSb VMT Structures 174
7.2 Probability Densities at a Fixed Point in **k**-space 176
7.3 Probability Density Distribution in Sub-bands 177
7.4 Boltzmann Transport Model 180
7.5 Boltzmann Transport Simulation Results 186

8.1 Earliest Target Device Geometry 193
8.2 Buffer Layer Lift-Off Processing 195
8.3 Modified Mask Designs 197
8.4 TEM Images of Alloyed Au/Ge Contact 200
8.5 Un-Annealed Source-Drain I-V Curves 201
8.6 Results of Related NRL Work 204
8.7 Dual Gate Structure 206

A.1 Resistive Network Model – Definitions 210

List of Tables

2.1 Calculated and Reported Avalanche Mode TSD Peak Voltages . . . 17

2.2 TSD Memory Performance Estimates 24

2.3 DC Electrical Characteristics of Commercially Available SRAM and DRAM Chips 25

2.4 Comparison of the TSD and the RTD 26

2.5 Comparison of TSD and MOS Stress Parameters 37

3.1 Target TSD Parameters for SRAM Application 44

3.2 Commercial Wafers Used for TSD Devices 46

3.3 TSD and MOS Devices Produced at Caltech 50

3.4 TSD Wafers Produced at Stanford 53

3.5 TSD and MOS Wafers Produced at Intel 55

4.1 Summary of TSD Device Behavior 105

5.1 Silicon Mobility Parameters 114

5.2 Silicon Avalanche Generation Parameters 115

5.3 Tunnel Current Structure Parameters 131

7.1 Boltzmann Transport Calculation Parameters 185

8.1 MBE-Grown Samples Used for VMT Device Fabrication 190

8.2 Results of Hall Measurements 191

Chapter 1 Introduction

1.1 Introduction to Thesis

Each of the two parts of this thesis describes experimental and theoretical investigations of a semiconductor device in which quantum effects play a critical role. The first part describes the study of the Tunnel Switch Diode (TSD) – a highly nonlinear device which operates much like a thyristor, but whose behavior is controlled by tunnel current injection through a thin oxide layer. The second part describes an implementation of a Velocity Modulation Transistor (VMT) – a device in which the action of the gate is to effectively modulate the mobility of the carriers in a coupled quantum-well channel, rather than the concentration of these carriers, by adjusting the localization of the carrier wavefunctions. Experimental and theoretical studies of each device are presented and discussed.

1.2 Summary of Results

We have developed a much better understanding of the TSD device behavior, and the role of device structure, oxide properties, and stress effects in determining this behavior. Devices were fabricated at several different facilities and electrically characterized. It was found that the device properties, most notably the I-V characteristics, were heavily dependent on the electrical stress history of the device, as a result of stress modification of the thin oxide layer. Two "modes" of device behavior were discovered, depending on the fabrication details and the amount of electrical stress the device had withstood. A "low current mode" resulted in certain devices which had only been subjected to small amounts of stress, and a "high current mode" was observed in nearly all devices which had been subjected to large amounts of stress. The properties of these two "modes," such as the scaling of device current with active

area and the dependence on oxide thickness, were found to be profoundly different. Devices with extremely large ranges of on-state currents extending over nearly five orders of magnitude and with even larger on/off current ratios were produced in the "low current mode," suggesting that these devices might be useful in a number of practical applications, including a transistorless SRAM architecture.

A drift-diffusion simulation of the TSD device was developed. Many aspects of the TSD device behavior were found to be quite analogous to the behavior of a $pnpn$ thyristor. For example, the device behavior was found to be quite sensitive to the oxide tunneling properties (analogous to the transport properties of the current-limiting layer of a thyristor) and to the recombination lifetimes in the silicon parts of the device. The simulated results were found to describe the behavior or the "low current mode" observed in the experiments, and they also suggest some explanations for the unusual properties of the "high current mode."

An InAs/GaSb/AlSb VMT device was proposed. A simple theoretical description of the operation of this device was developed, and compared qualitatively to that of the analogous GaAs/AlAs device. The fundamental limitations and characteristics of the two devices were found to be quite different due to the opposite dispersion relations of the coupled carriers in the InAs/GaSb/AlSb system. A simple Boltzmann transport model was developed to describe the transport in the InAs/GaSb/AlSb system which is dictated by an unusual distribution of states in $\mathbf{k}$-space. The subband structure in the coupled well system was calculated and analyzed, allowing better understanding of the theoretical device behavior and providing information needed to engineer a device structure to be fabricated. A number of attempts were made to produce the proposed InAs/GaSb/AlSb VMT device. While none of these attempts were successful, a number of advances were made in developing the necessary processing techniques.

1.3 Motivation

The success of the vast Silicon VLSI logic industry has been guaranteed over the past decade by the ability to shrink the characteristic dimensions of the MOSFET devices

which comprise the underlying circuits. As the future success of this industry will depend critically on the continuation of this trend, much effort has been spent in the pursuit of answers to many relevant questions:

1. What technological advances will be necessary in order for these scaling trends to continue?

2. What will be the ultimate limit to which the devices can be scaled, and what physical, practical, or economic factors will dictate this limit?

3. Once this limit is reached, will other devices besides the MOSFET or other materials besides Si become more practical?

As discussions of these topics can easily fill volumes, and can often degenerate to more religious than scientific arguments, no such discussion will be presented here, but these questions do serve to define the research goal embodied by this work. Given that a minimum limit on MOSFET device size will eventually be reached by some mechanism, it is perhaps worthwhile to investigate alternative devices which (a) may be arranged into circuits which can provide superior performance to a MOSFET circuit of equivalent size, at least for some range of applications, or (b) can be fabricated and made to function at dimensions below that of the smallest feasible, functional MOSFET. The studies of the devices presented in this thesis are primarily aimed at objective (a), but it is possible that, in the end, (b) may hold true as well, as the quantum effects governing the behavior of both devices depend on the creation of very small physical structures, at least in one dimension. While the factors motivating the study of each of these devices will be described in detail in later chapters, a brief account is given here for each device.

1.3.1 The Tunnel Switch Diode

Much interest has been generated recently in devices which can exhibit highly nonlinear I-V characteristics. This is because a single such device could potentially provide the functionality of a large number of traditional transistors in a circuit, allowing for a

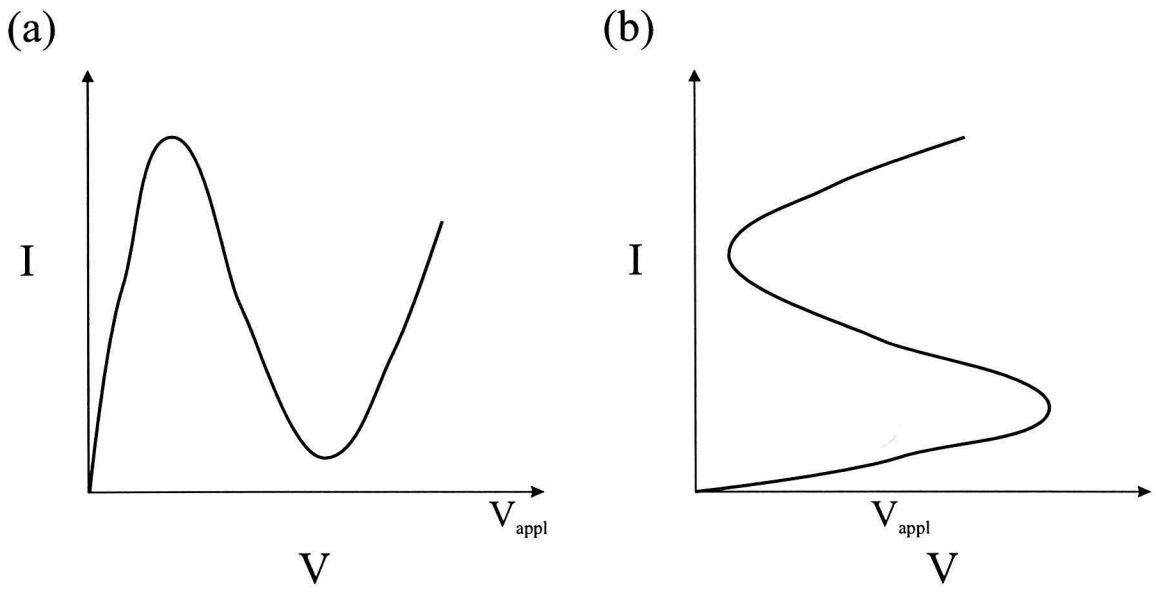

Figure 1.1: (a) N-type and (b) S-type negative differential conductance I-V curves.

faster and/or more compact implementation. While an unlimited number of nonlinear devices could be conceived, such a device is only useful if is can be used effectively in a circuit. Therefore, systems issues as well as device parameters must be examined simultaneously.

Of particular interest are devices which can support multiple stable states for a given range of boundary conditions, and hence may be useful in electrical memory systems. Such devices could allow for improvements over typical CMOS static RAM architectures, which require at least 4-6 transistors per bit [1]. One such class of devices are those which exhibit Negative Differential Resistance (NDR), *i.e.*, a region in the plot of voltage versus current for which the slope of the curve is negative. Two types of I-V characteristics which fall into this class are shown in Figure 1.1 – one which is "N-shaped," and one which is "S-shaped." The "N-type" devices can support multiple voltages for a range of currents, and the "S-type" devices can support multiple currents for a range of voltages.

We pause here to make an important semantic distinction. Note that as these curves are plotted as current versus voltage, rather than voltage versus current, it is more precise to describe this behavior as differential conductance rather than differen-

tial resistance. In fact, it is easy to see how this terminology could lead to confusion if not carefully used, as an "N-type" differential conductance I-V curve is also an "S-type" differential resistance I-V curve. Therefore, while we will refer to this behavior in general as Negative Differential Resistance, we will take care to use the proper notation when describing "N-type" or "S-type" curves.

Most recent work in this area has focused on devices with "N-type" differential conductance curves – most notably the Resonant Tunneling Diode (RTD) [2]. It has been shown, however, that an "S-type" differential conductance curve may prove more useful in memory circuits due to power dissipation, speed, and density considerations [3, 4]. The Tunnel Switch Diode, effectively consisting of a silicon p-n junction in series with a thin oxide MOS junction, exhibits "S-type" differential conductance, and may prove to be an ideal candidate for a very dense static memory architecture [4]. In addition, the TSD possesses other properties, such as light sensitivity and the ability to be used as a three terminal device, which could potentially be utilized in a number of circuit architectures. Perhaps most importantly, the TSD can be fabricated in silicon, unlike the RTD, and may therefore be compatible with conventional CMOS circuits as well.

Unfortunately, the TSD device has heretofore been rather poorly understood. While the basic operating principles are known, the details are obscured, largely due to difficulties in describing and predicting large current transport through the thin oxide layer. The work presented in this thesis focuses on furthering the understanding of the behavior of the TSD, particularly the transport through the oxide and its influence on the overall device behavior.

Another motivation for the study of the TSD involves the advancement of silicon CMOS technology. Over the past two decades, the characteristic dimensions of Si MOSFETs have decreased steadily, and the trends are likely to continue into the near future [5, 6]. Following these trends, the gate oxides will decrease in thickness from current values of 40-60 Å to less than 30 Å over the next several years. Therefore, an understanding of transport through such thin oxides will become crucial to the survival of this industry. While a tremendous amount of work has been done in this area, it is

possible that the study of another device which relies critically on the properties of a thin oxide layer, namely the TSD, may provide additional useful information.

1.3.2 The Velocity Modulation Transistor

Most traditional transistors employ the use of a gate contact to modulate the conductivity between two other contacts – the source and drain (although these three terminals are often known by different names). This modulation is brought about by a change in the carrier concentration somewhere between the source and drain. This type of process presents fundamental speed limitations. In order to bring about the change in carrier concentration, a capacitive coupling between the gate and the source-drain channel is necessary, leading to RC delay times. In addition, some transit time is necessary in order for the charges to move to their new positions.

Speed can be improved by reduction in the characteristic sizes of devices, as capacitances are thereby decreased, and the distances over which carriers must travel are reduced. Given that, at some point, the continued reduction in size will not be possible, this method cannot be used to further increase speed. It is also possible to use other choices of materials which have smaller effective masses, and hence, higher mobilities, leading to higher conductivities, lower RC delay times, and lower transit times, and hence, higher speed. There are a finite number of materials choices, however, and many may not be suitable for production of transistors for practical reasons. Therefore, it is interesting to consider the possibility of a transistor which may not be subject to these same fundamental limitations.

Given that the electrical conductivity in a semiconductor is given by the product of a charge density and a mobility (*i.e.*, the ratio of the carrier velocity to the applied electric field), it is evident that a change in charge density is not necessary in order to affect a change in conductivity – a change in mobility would work just as well. If it were possible to bring about a change in mobility without having to disturb much charge, it might be possible to create a device which could supersede the bounds described above. This concept was presented by H. Sakaki in 1982 [7], as well as a proposal for a

device which could produce such an effect, called the Velocity Modulation Transistor (VMT), or Mobility Modulation Transistor (MMT).

Sakaki's proposed device consisted of coupled quantum wells of differing mobilities and a gate electrode which could shift the relative energies of the quantized states in the wells. By shifting the wells in and out of resonance with the gate bias, the localization of the wavefunctions in the wells could be controlled, changing the effective mobility of the coupled well system. Such a device has been successfully demonstrated in the GaAs/AlAs material system [8]. Unfortunately, the modulation effect was not particularly large (a factor of three) and could only be observed at cryogenic temperatures (4.2 K). This is because the mobilities of the two GaAs quantum wells in the device could only be forced to be significantly different at low temperatures by introduction of impurities into one of the wells.

In this work, we investigate the operation and fabrication of a similar device in the InAs/GaSb/AlSb material system. The unusual band alignments of these materials allows the valence band states in the GaSb to overlap in energy with the InAs conduction band states, so quantum wells made out of these materials can couple, with GaSb hole states interacting with InAs electron states. As these materials exhibit very different mobilities, even at room temperature, it is possible that a VMT device could be produced that would operate at a higher temperature, and might exhibit stronger conductance modulation.

1.4 Chapter Outline

Chapter 2 provides a detailed description of the TSD device, its applications, and a discussion of thin oxides which are critical to its operation. In Chapter 3, the experimental procedures used to produce and characterize the TSD devices are described. The results of these experiments are presented in Chapter 4. The development of a device simulator capable of handling the TSD device and the results of these simulations follow in Chapter 5.

The VMT device is introduced, and its implementation in both GaAs/AlAs and

InAs/GaSb/AlSb systems is described and contrasted in Chapter 6. Chapter 7 contains theoretical studies of the proposed InAs/GaSb/AlSb device. Finally, the details of the attempts at fabrication of the device are given in Chapter 8.

Bibliography

[1] K. Itoh, *Proc. IEEE* **83**, 524 (1995).

[2] M. Tsuchiya, H. Sakaki, and J. Yoshino, *Jpn. J. Appl. Phys.* **24**, L466 (1985).

[3] H. J. Levy, *Application and Integration of Quantum-Effect Devices for Cellular VLSI*, PhD thesis, Caltech, 1994.

[4] H. J. Levy, E. S. Daniel, and T. C. McGill, (to be published).

[5] C. A. Mead, *Analog Integrated Circuits and Signal Processing* **6**, 9 (1994).

[6] Semiconductor Industry Association, The national technology roadmap, 1994.

[7] H. Sakaki, *Jpn. J. Appl. Phys.* **21**, L381 (1982).

[8] Y. Ohno, M. Tsuchiya, and H. Sakaki, *Appl. Phys. Lett.* **62**, 1952 (1993).

Part I

Experimental and Theoretical Investigations of Si-based Tunnel Switch Diodes (TSD)

Chapter 2 Background and Motivation for Study of the TSD

2.1 Introduction and Chapter Outline

This chapter will provide much of the framework necessary for discussion of this section of the thesis, which concentrates on the Tunnel Switch Diode. The first section (Section 2.2) will introduce the TSD, describe its behavior, and explain the basic device physics relevant to its operation. Section 2.3 describes a number of possible applications of the TSD. In particular, a detailed description of an SRAM architecture which relies on the TSD will be discussed, and the device parameters necessary for the SRAM to be useful will be defined. As we shall find that the TSD device behavior relies critically on the properties of its thin oxide layer, a discussion of many properties of thin oxides (Section 2.4), with a strong emphasis on SiO_2, will be presented.

2.2 Introduction to the Tunnel Switch Diode (TSD)

2.2.1 Background

The Tunnel Switch Diode (TSD), also known as the Metal Insulator Semiconductor Switch (MISS) or Inversion Controlled Switch (ICS) was first discovered by Yamamoto, *et al.*, in 1972 while conducting studies of Au-Si Schottky contact gate FET devices [1]. They found that if a thin layer of native oxide were allowed to form between the metal gate and the silicon, and if a p-n junction were present in the silicon, then the I-V curve of a two-terminal device consisting of the metal/oxide/n-Si/p-Si stack (Figure 2.1(a)) exhibited negative differential resistance (NDR), as seen in Figure 2.1(c), much like a $pnpn$ thyristor structure. Thus, if the TSD is placed in

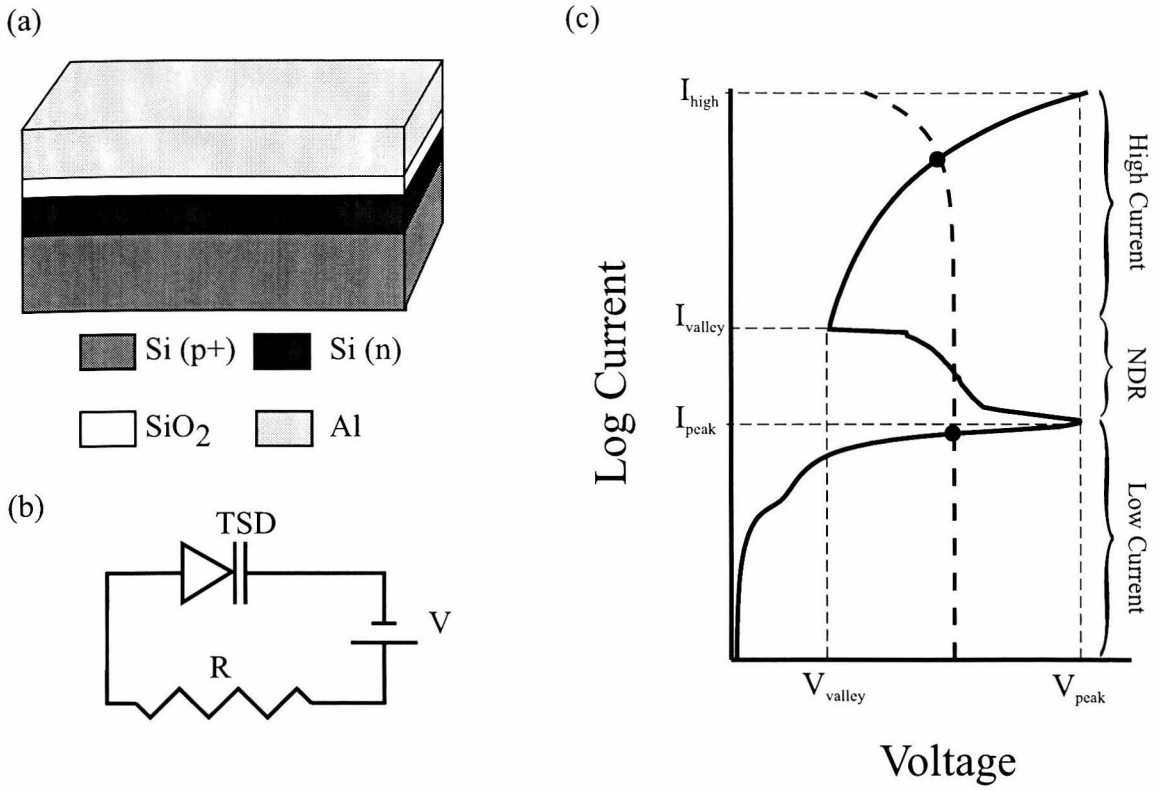

Figure 2.1: Tunnel switch diode (TSD). (a) The TSD device consists of a p^+ silicon substrate with an n silicon epilayer, typically a few microns thick, a tunnel oxide (10 - 40 Å), and a conductive (*e.g.* Al) gate on top. (b) Load-line analysis circuit. (c) A measured I-V curve for a 2.5×10^{-5} cm^2 device plotted along with a 2.5 V, 10 Ω load line, illustrating the two stable current states. For this particular device, I_{high}=100 mA, I_{valley}=100 μA, I_{peak}=9 μA, V_{valley}=1 V, and V_{peak}=4 V.

a circuit as in Figure 2.1(b), there is a range of voltages for which two stable current states are possible, as illustrated by the load line in Figure 2.1(c). Yamamoto further discovered that, unlike a thyristor, the TSD can be switched between high and low current states very rapidly, having observed switching speeds of $\lesssim$ 2 ns for large area devices [2].

This discovery led to the investigation of a wide variety of similar structures for use in a variety of applications, including high power, high speed switching and light-triggered switching. Devices with "insulating layers" composed of SiO_2 [2–4], Si_3N_4 [3, 5], SiO_xN_y [3], SiC [6], polysilicon [3, 7, 8], amorphous silicon [9, 10], and amorphous Ge [3] grown on either p^+-n or n^+-p silicon junctions have all been shown to exhibit

the same type of negative differential resistance.

While the behavior of the TSD is not completely understood, the basic physical processes which lead to the observed I-V characteristics are generally agreed upon [3, 4, 8, 11], and will be discussed in the next section. In particular, the discussion presented here will focus on the role of the thin insulating layer. For the purpose of this discussion, we will assume the insulating layer is SiO_2, as this is the barrier material used in the majority of the studies presented in this thesis.

2.2.2 Basic TSD Device Operation

The device is most easily described as a p-n junction and an MOS junction in series. If positive bias is applied to the gate electrode, the I-V curve is dominated by the reverse biased p-n junction and contains no unusual features. As negative bias is applied to the gate electrode, the MOS junction depletion layer begins to extend into the n-Si layer (Figure 2.2(a)). In this state, the current is low (Figure 2.2(b)), limited by generation/recombination processes in the reverse-biased depletion layer, and the bulk of the bias is dropped across this layer. As long as the oxide layer is not too thick, further bias will cause the depletion region to grow, rather than creating a surface inversion layer (see Section 2.4). Once the depletion layer reaches the p-n junction depletion region, further bias causes the p-n junction to be turned on, flooding the n-Si layer with holes. If the oxide layer is not too thin, the holes will accumulate in the n-region, creating an inversion layer. This shifts the bulk of the bias from the depletion layer to the oxide (Figure 2.2(c),(d)), causing a large increase in tunnel current. The injection of negative charge, in turn, causes the Fermi level in the n-Si layer to be pulled up, forward biasing the p-n junction even more. This positive feedback mechanism switches the device into the high current state, in which the n-Si layer is nearly charge neutral, highly populated by both holes and electrons, with a hole inversion layer at the surface providing the large field across the oxide. This state can only be supported as long as the oxide is thin enough to allow large tunnel currents, but thick enough to support the hole inversion layer.

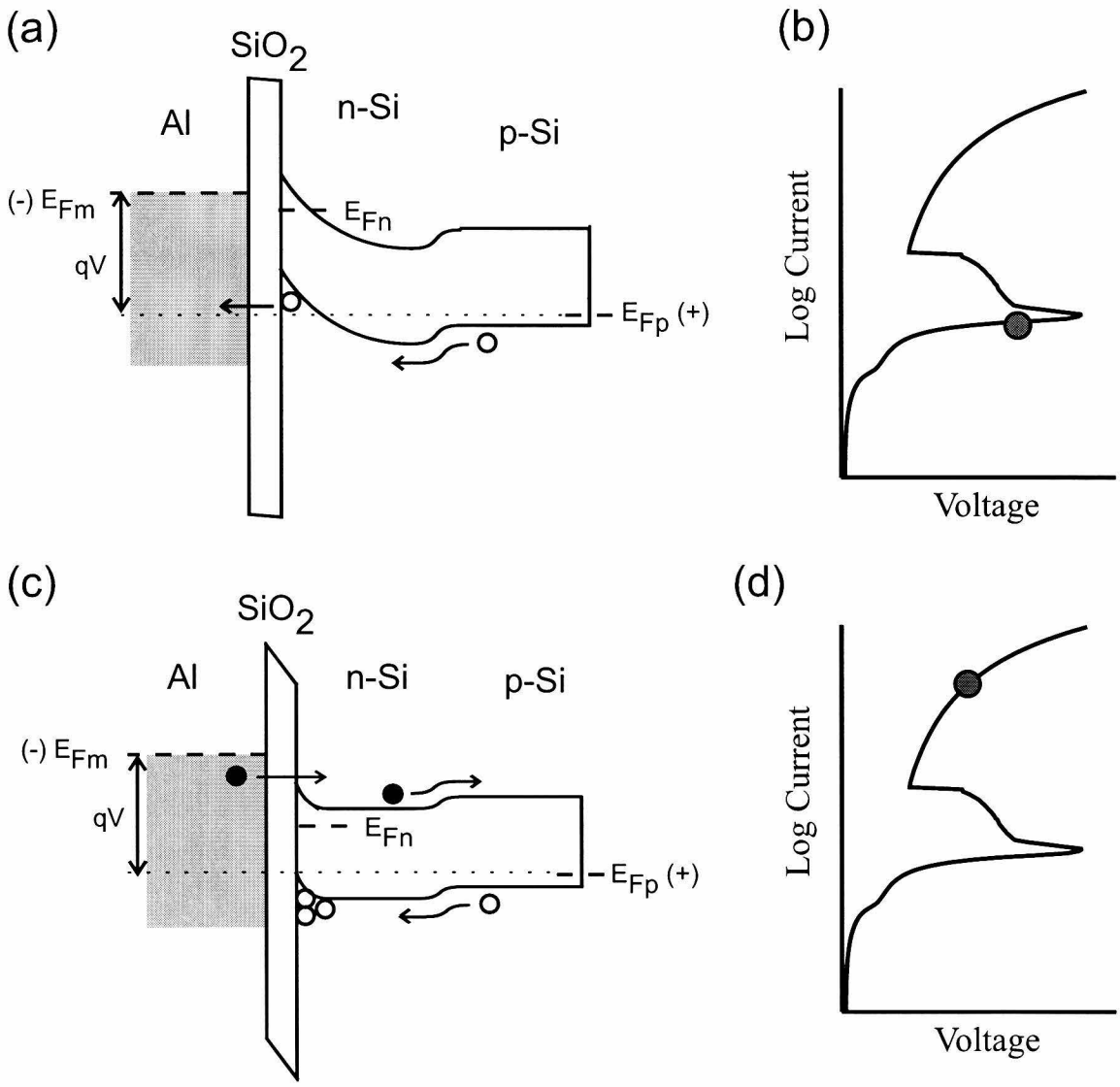

Figure 2.2: Basic TSD device operation. (a) A band diagram of the TSD in the low current state just below the peak voltage. E_{Fm}, E_{Fn}, E_{Fp} and V are the metal Fermi level, electron quasi-Fermi level at the SiO_2/n-Si interface, the Fermi level in the flat band region of the p-Si substrate, and the applied bias respectively. (c) A band diagram of the TSD in the high current state. The indicated circles on the *I-V* curves in (b), (d) correspond to the bias illustrated in (a), (c) respectively.

Note that just before switching, the bulk of the bias across the device is dropped across the depletion region, as it does not take much bias across the forward biased p-n junction or the oxide to provide the small current. Therefore, the peak voltage can be approximated by the voltage needed to create a depletion layer extending all the way to the edge of the p-n junction depletion region. Using the textbook formula for the bias across a depletion layer [12], we obtain

$$V_{peak} = \frac{qN_C w^2}{2\epsilon_s} \tag{2.1}$$

where N_C is the dopant concentration, w is the distance from the oxide-semiconductor interface to the edge of the p-n junction depletion region (not the junction itself), and ϵ_s is the semiconductor dielectric constant. This formula has been shown to predict peak voltages in TSD devices with a fair amount of accuracy for devices with top silicon layers relatively lightly doped (5×10^{14} /cm^3 - 5×10^{15} /cm^3) and several microns thick [3, 11]. As the device switching is triggered by the surface depletion region "punching through" the flat band region, it is said to operate in the "punch-through" (P-T) mode.

Other mechanisms can be employed to switch the device from the low to high current states. For example, light injection into the n-Si layer causes electron-hole pairs to be produced, setting up the necessary carrier distribution in the n-Si layer at a much lower bias [2]. In Figure 2.3, we see that increased light injection intensity results in a decrease in the peak voltage of the TSD, until eventually the peak disappears completely. Charge injection via a third terminal, connected to the n-Si layer, results in a similar modification of the I-V characteristic [4, 8].

If the top silicon layer is doped heavily enough, the electric field at the oxide interface will become great enough to initiate avalanche generation of carriers, which can switch the device from the low to high current state [3, 11]. If the bias at which this occurs is less that the bias needed to deplete the entire layer (the punch-through voltage), then this process will determine the peak voltage, rather than the punch-through pro-

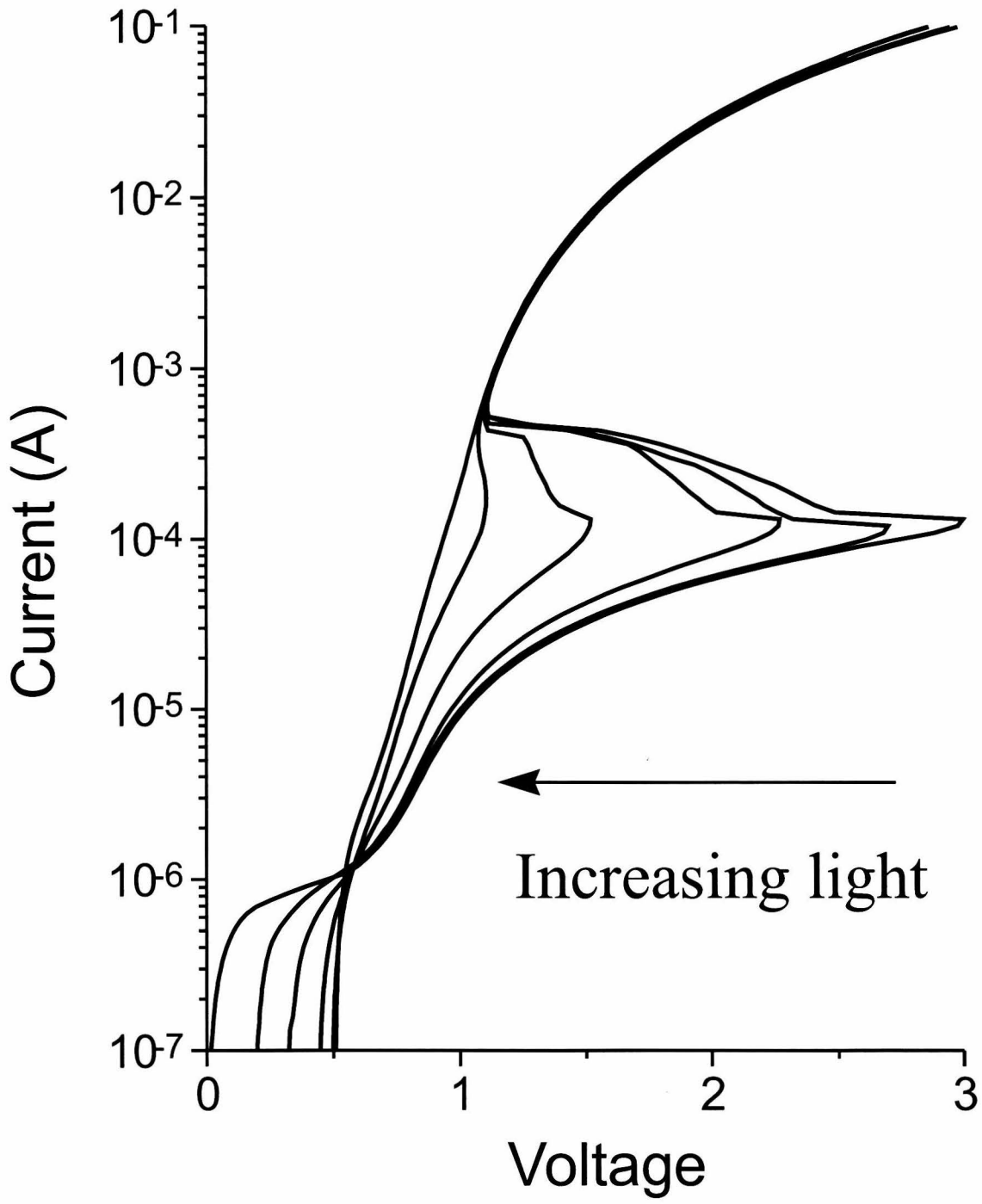

Figure 2.3: Measured I-V curves of a 150×150 μm^2 device for different intensities of injected light. Increasing light intensity decreases the peak voltage.

Table 2.1: Experimental and theoretical peak voltage values for avalanche mode TSD devices with various top layer doping densities, taken from [11].

N_d (cm^{-3})	V_{peak} - theory	V_{peak} - expt.
6.8×10^{17}	2.5	1.7
6.2×10^{17}	2.7	2.4
6.0×10^{17}	2.7	2.2
5.2×10^{17}	3.1	2.8
4.6×10^{17}	3.4	2.6
4.0×10^{17}	3.8	3.2
3.3×10^{17}	4.5	4.8
2.0×10^{17}	6.3	5.3

cess described above. The peak voltage for this process can be approximated by the avalanche breakdown voltage of an abrupt one-sided junction [12]:

$$V_{peak} = 60(E_g/1.1)^{3/2}(N_C/10^{16})^{-3/4} \tag{2.2}$$

where E_g is the bandgap in eV and N_C is the dopant concentration in cm^{-3}. A graphical comparison of avalanche and punch-through peak voltages for different top silicon layer doping levels and thicknesses is presented in Figure 2.4. Note that the thicknesses indicated in the plot refer to the total silicon layer thickness, not the punch-through depletion width, w, from equation 2.1. The width w is obtained by subtracting the thickness of the n-side of the p-n junction from the total layer thickness.

One reported set of experimental and theoretical avalanche mode device parameters is listed in Table 2.1. The experimental values are uniformly lower than the calculated values, presumably due to concentrated fields at the edges of the device or non-abrupt junction effects.

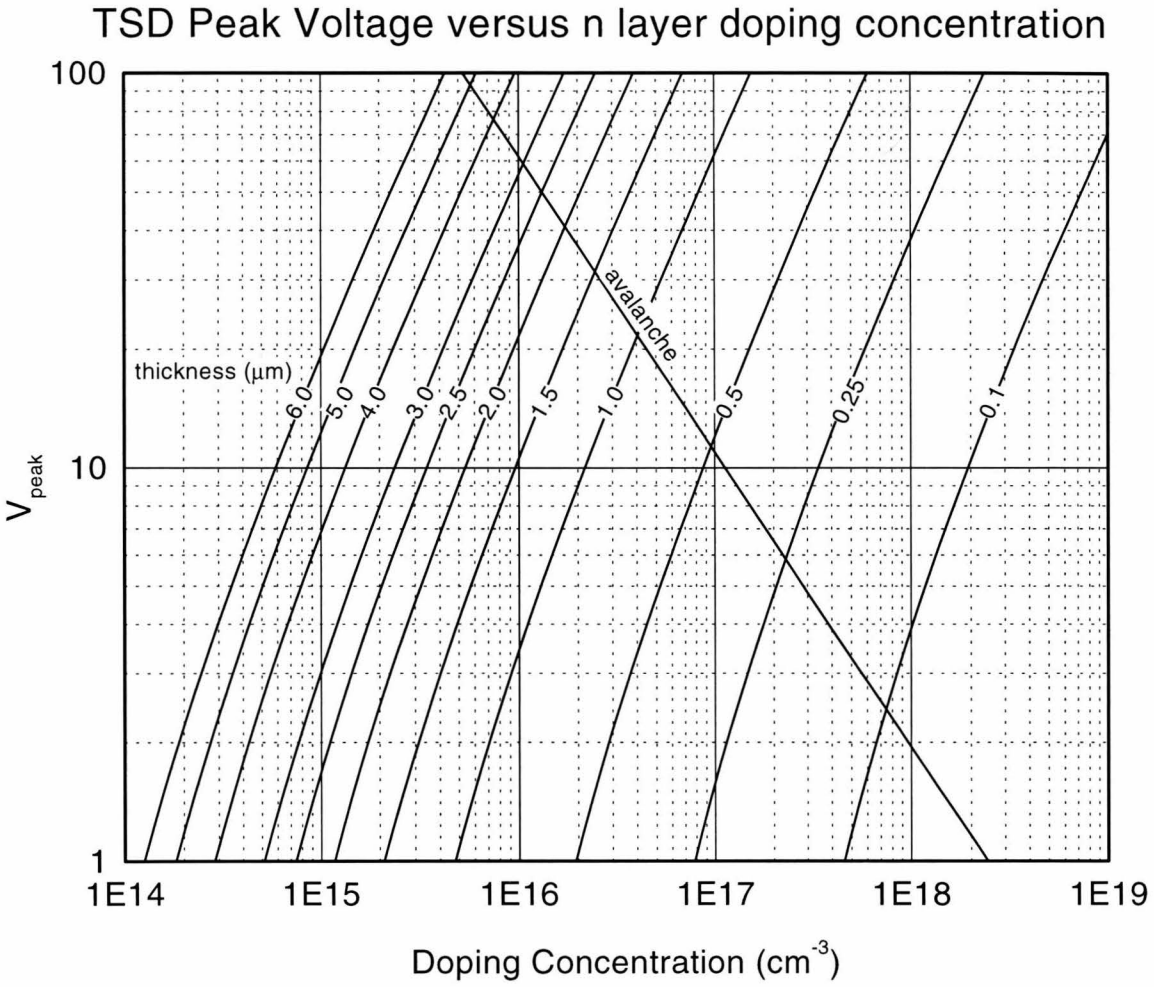

Figure 2.4: Calculated peak voltage for avalanche mode and punch-through mode TSD devices for a range of doping concentrations and thicknesses of the top silicon layer. A substrate doping concentration of 2×10^{18} was assumed.

2.3 Applications of the TSD

2.3.1 Transistorless SRAM Application

Background

While recent improvements in microprocessor technology have pushed clock rates higher and higher, DRAM access times have not improved at the same rate. This is because DRAM capacitor-discharge currents are too small – much less than SRAM transistor currents for comparable memory cell density – and it is these currents which are integrated over the sense time t_{sense} to provide the necessary sense charge Q_{sense} to a digital amplifier [13]. As a result, faster SRAM cache memory has become critically important in minimizing wasted processor cycles. While SRAM access speeds comparable to processor speeds can be achieved, a severe density penalty must be paid as a typical SRAM cell consists of at least five devices, and hence DRAM remains the solution of choice for mass storage [14, 15].

The need for both types of memory could be obviated by a memory architecture consisting of a very compact bit circuit capable of delivering large amounts of current for a "one" state read cycle, supporting fast write cycles, and maintaining many of the advantages common to both SRAM and DRAM designs, such as an exponentially lower "zero" state current, low quiescent power, and ease of large scale fabrication and integration. Levy, *et al.* [16–18], have demonstrated that the TSD alone can be used in a compact current-mode static RAM architecture, provided that certain constraints on the TSD I-V curve can be met. In the next section, the proposed SRAM architecture is presented in order to clearly define how the TSD must be tuned in order to produce a technologically and economically useful SRAM device.

Circuit Operation

The proposed memory architecture [16, 17] consists of an array of single TSD devices (Figure 2.5(a)), with the bit information stored in the current state of each device. Because the information is stored as a current rather than a voltage, there is no need

for isolation transistors between devices, allowing perhaps maximal density as each bit consists of a single, simple device. We shall see that, by adjusting the voltages on the bit and word lines, it is possible to read from and write to individual bits.

First, consider a static situation in which no read or write is being performed. In this situation, it is desirable to keep a low bias, V_s, across each bit such that power consumption is minimized, while the bit information is preserved (see Figure 2.5(b)). This is easily done by applying a voltage V_s^+ to the word lines and V_s^- to the bit lines such that $V_s^+ - V_s^- = V_s$.

When a bit is to be read, the "one" state current should be maximized in order to minimize t_{sense} while still preserving the bit state, so the bias across that bit should be increased to V_{read} (see Figure 2.5(b)). One method for performing this is illustrated in Figure 2.5(a) in which the voltage on a given word line is raised by an amount $\delta = V_{read} - V_s$, thereby increasing the currents of all bits along that row to the appropriate read currents. If the selected bit along a particular bit line is in the "one" state, the read current generated by that bit will dominate the total bit line current as it is orders of magnitude larger than any of the quiescent currents. If this bit is in the "zero" state, the current along the bit line will be exponentially smaller. Thus, in either case, the entire row of bits can be read out simultaneously by sense amplifiers connected to the bit lines.

A simple write scheme can be similarly executed (Figure 2.6). To write a "one" state, for example, the word line corresponding to that bit is raised by an amount Δ while the bit line corresponding to that bit is lowered by Δ, where Δ is chosen such that $V_s + \Delta < V_{peak}$ and $V_s + 2\Delta > V_{peak}$. This forces the selected bit into the high current state while preserving all of the other bit states. As the additional biases are removed, the newly written bit relaxes along the high current portion of the I-V curve, preserving its state. The method for writing a "zero" state is identical, where now Δ is negative, satisfying $V_s + \Delta > V_{valley}$ and $V_s + 2\Delta < V_{valley}$.

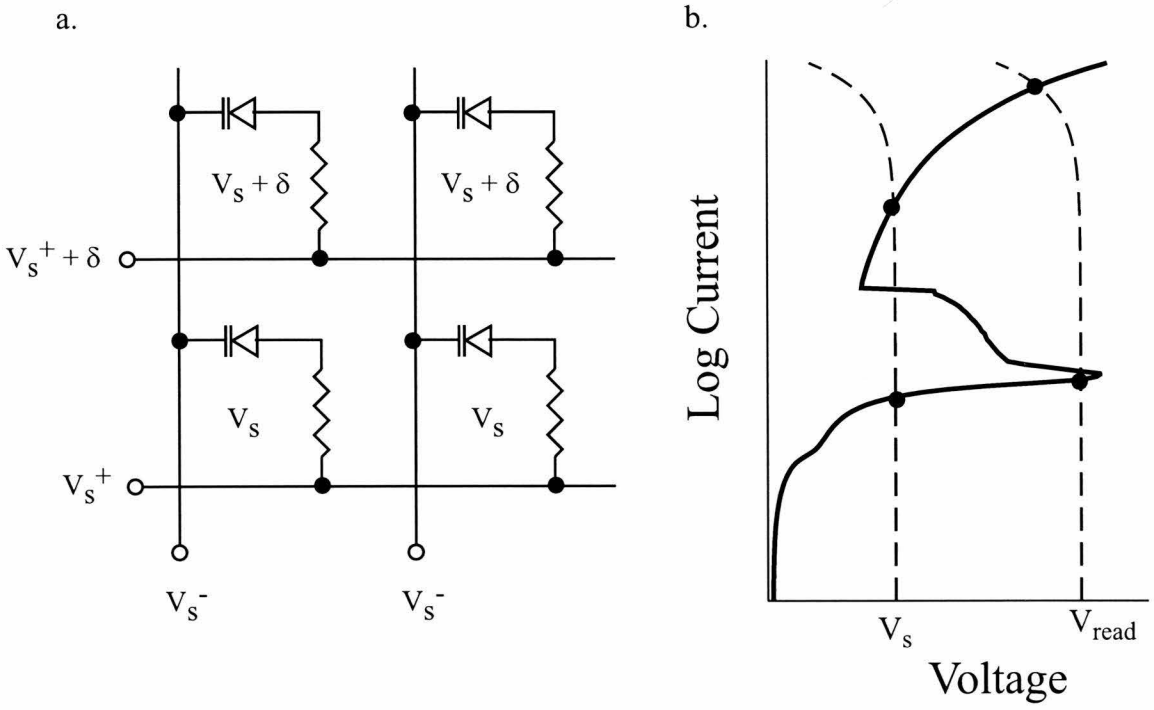

Figure 2.5: Transistorless SRAM Read Scheme. (a) A sample 2×2 bit array in which the top row is selected to be read out. (b) Typical TSD I-V curve with load lines corresponding to static bias and read bias.

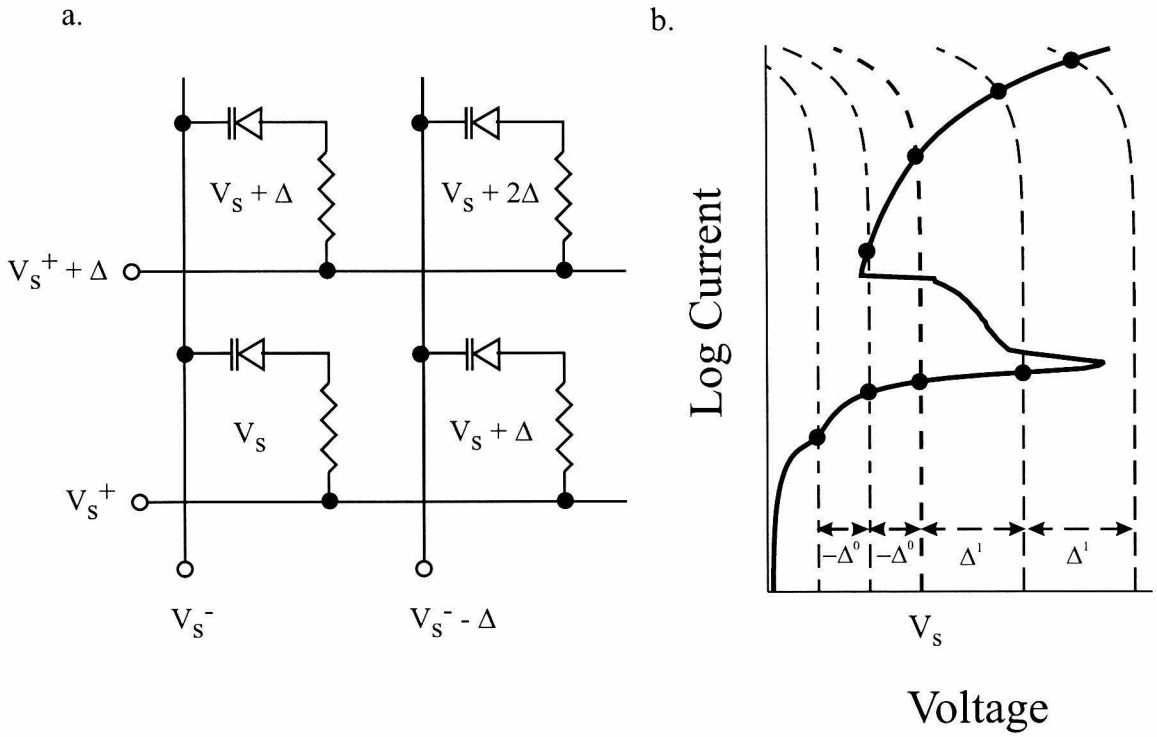

Figure 2.6: Transistorless SRAM Write Scheme. (a) A sample 2×2 bit array in which the bit in the upper left corner is to be written. (b) Typical TSD I-V curve depicting load lines employed for writing both the "one" and "zero" states.

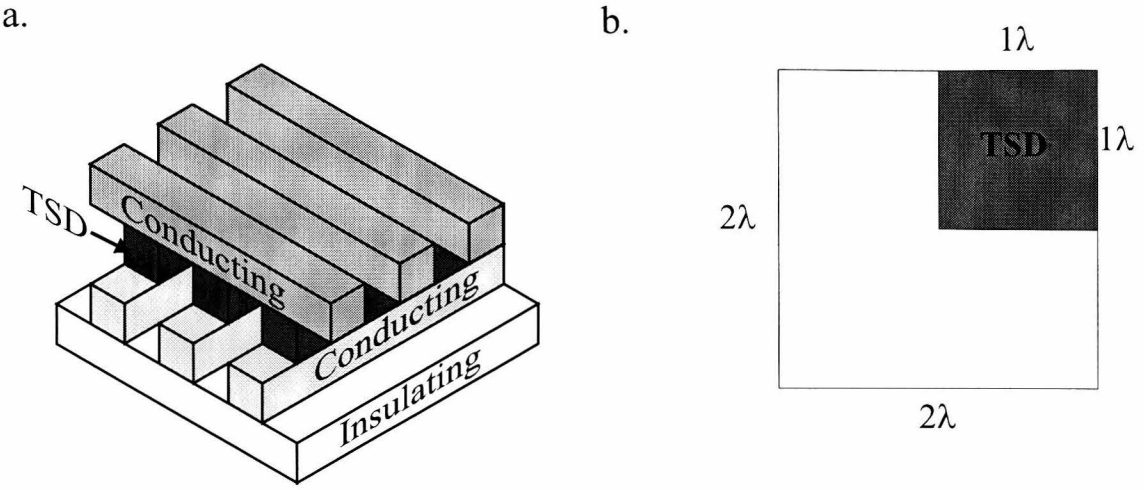

Figure 2.7: Density maximizing TSD SRAM fabrication scheme. (a) Self-aligned scheme for fabricating a dense addressable TSD array. (b) The area of the cell, 4 λ^2 is the maximum possible for a planar array of isolated devices.

Performance Estimates and Requirements of the TSD

Notice that the device consists of a simple stack of semiconductor and oxide layers, and nothing further is needed to construct a bit. Therefore, in principle, a dense memory array could consist of closely spaced rows of conducting word lines, separated vertically from perpendicular columns of bit lines by a TSD stack (Figure 2.7) [16, 17]. Assuming these features can all be made as small as the minimum feature size, λ, this yields a cell size of $4\lambda^2$ with an active device area of λ^2 – the minimum possible cell size for an isolated 2-D array structure. For the purpose of this discussion, we will assume this density maximizing fabrication scheme, keeping in mind that any scheme utilizing larger cell areas, through the use of lateral contacts for example, will require more stringent constraints on the TSD.

In order to evaluate the performance of such an architecture, we make the assumptions that we are constrained to a minimum Q_{sense} value (we assume 100 fC, a conservative CMOS value [14]), and that a maximum of 1 A/cm^2 of current may be dissipated in the quiescent state, corresponding to $\sim$ 1 W/cm^2 for a typical V_s $\sim$1V. Then, given the maximum read current, J_{read}, one can calculate the necessary ratio $\beta = J_{read}/J_s$ in order to meet the power dissipation requirement for an array of cells,

Table 2.2: TSD memory performance estimates. For a given maximum read current density, J_{read}, the necessary current ratio $\beta = J_{read}/J_s$ for a 1 A/cm^2 power dissipation, and the possible device size and density for several possible read times are listed, assuming a 100 fC sense charge. Adapted from [16, 17].

Max. J_{read} (A/cm^2)	Required β For $1A/cm^2$	TSD Size (μm) and Cell Density (cm^{-2}) For $Q_{sense} = 100fC$		
		$400\mu A \times .25ns$	$40\mu A \times 2.5ns$	$4\mu A \times 25ns$
4×10^5	10^5	$.32 \times .32$ 250Mb	$.10 \times .10$ 2.5Gb	$.03 \times .03$ 25Gb
4×10^4	10^4	1.0×1.0 25Mb	$.32 \times .32$ 250Mb	$.10 \times .10$ 2.5Gb
4×10^3	10^3	3.2×3.2 2.5Mb	1.0×1.0 25Mb	$.32 \times .32$ 250Mb
		Mb = 10^6 bits, Gb = 10^9 bits		

as well as the possible cell size for the quiescent current density J_s and a set of possible sense times. The results are presented in Table 2.2 [17].

Note that there is a tradeoff between cell density and sense time, as the device can simply be made larger in order to source more current in a given time. Also, notice that in order to achieve low power dissipation and a short sense time simultaneously, β should be as large as possible.

In order to evaluate a reasonable set of target parameters for the TSD device, consider the characteristics of a few typical commercially available SRAM and DRAM chips presented in Table 2.3. Comparing this data to that of Table 2.2, we see that a maximum J_{read} of $\sim 10^4$ and a *beta* $\sim 10^3 - 10^4$ are needed for the TSD to compete with current technologies in terms of speed, density, and power consumption. Note, however, that a simultaneous improvement in speed and density may warrant higher power consumption.

Other device characteristics will be important as well, such as switching speed and device lifetime. Bidirectional switching times in the 1-2 ns range have been reported for large area discrete devices [2, 4], and such devices have been operated in excess of 10^4 hours without failure. [2]

Table 2.3: DC Electrical Characteristics of Commercially Available SRAM and DRAM Chips. The reported refresh current for the DRAM chips is an average over the maximum possible refresh period (64 ms in both cases). The quiescent power is estimated as $I_{refresh} \times V_{supply}/\#bits$ for DRAM, and as $(I_{snooze}, I_{standby}) \times V_{supply}/\#bits$ for SRAM.

Chip Name	V_{supply} (V)	$I_{operating}$ (mA)	$I_{refresh}$ (mA)	$I_{standby}$ (mA)	I_{snooze} (mA)	$P_{quiescent}$ (nW/bit)
Micron 4Mb 7.5ns SRAM [19]	3.3	360	—	90	5	3.9-71
IBM 4Mb 6ns SRAM [20]	3.3	450	—	150	120	94-118
Micron 64Mb 50ns EDO DRAM [21]	3.3	170	170	—	—	8.4
IBM 64Mb 8ns SDRAM [22]	3.3	140	120	—	—	5.9

2.3.2 Other Applications

One interesting property of the TSD is the ability of the state of one device to influence the switching behavior of a neighboring device [16]. This can occur due to lateral flow of the high population of carriers present in the silicon layer nearest the oxide in a device in the "on" state. Flow of such carriers has the same effect as injection of current via a third terminal, discussed in Section 2.2.2 – the peak voltage of the device is reduced, allowing switching to the "on" state with lower than usual bias.

While this property can be undesirable (as with the SRAM application where each device must be isolated in order to prevent interactions), it can also prove useful. Many examples of cellular circuits and neural networks, for example, rely on the ability of a given cell to interact with its neighbors [23]. Several neural network concepts involving the TSD have been devised and demonstrated [16], but no single one has been aggressively pursued. In order for such circuits to be feasible, the effects of carrier injection must be well understood. While quantitative data and qualitative understanding exist [3, 7, 8], a detailed understanding of the device physics is lacking.

The light sensitivity of the TSD can also be exploited. A TSD device held at a given voltage in the "off" state will be switched into the "on" state by an amount of light injection determined by the holding voltage [9, 11]. In light of this property, it is easy to see how the SRAM architecture from the previous section could accommodate optical input into each bit. All bits could be written in to the "zero" state, and then held at some voltage between V_s and V_{read} (see Figure 2.5), and subsequent light

Table 2.4: Comparison of the TSD and the RTD.

RTD	TSD
Bistable voltages for a given range of current	Bistable currents for a given range of voltages
Modest difference between bistable states, even in III-Vs (a factor of 10-100 in voltage)	Huge difference between bistable states (many orders of magnitude in current)
Not capable of low power (low current) and high drive (high current) in a single device	Capable of low power (low current) and high drive (high current) in a single device
Not light sensitive	Light Sensitive
Not easily converted to a three terminal device	Can be easily used as a three terminal device
Seems to require (at minimum) epitaxial layers	Does not require epitaxial layers or any similar major constraint
So far, no working devices have been fabricated in Si with any of the many proposed barrier layers (SiO_2, $Si_xGe_yC_{1-x-y}$, ZnS, CaF_2,)	Working devices have been produced with many different insulating layers (SiO_2, p-Si, a-Si, Si_3N_4, Ge,)
Scaling to small lateral dimensions is likely to cause detuning of the resonant state	Scaling to small lateral dimensions seems to be possible

illumination of selected bits would force them into the "one" state. Again, quantitative understanding of carrier injection into the top silicon layer would be necessary for design of such a device.

Another popular negative differential resistance device, the Resonant Tunneling Diode (RTD), has been employed in a variety of high speed, low power logic circuit designs [24–26]. Although much effort has been expended in attempts to fabricate RTD's in silicon for compatibility with the vast existing silicon MOSFET logic technology, none of these efforts have yet produced a working device [25]. As the TSD functions in many ways as the dual to the RTD, possesses many similar features, and is easily produced in silicon with a wide variety of insulating layers as mentioned earlier, it may be possible to design similar circuits employing silicon-based TSD's. Properties of each of these devices are compared in Table 2.4.

Other applications such as chemical-sensitive switching [27, 28] and traditional high power switching [6, 10] have been explored, but fall outside of the scope of the discussion in this thesis.

2.4 Properties and Analysis of Thin Oxides

2.4.1 Background

As mentioned in Section 2.2.2, the TSD device operation depends heavily on the transport properties of the insulating layer. On one hand, this complicates the analysis of the TSD in that we must understand the insulator properties if we are to understand the device operation in detail. On the other hand, it also suggests that through study of the TSD, it may be possible to extract properties of the insulating layer. As Si MOSFET technology progresses, and thinner gate oxides are used, the study and understanding of their properties will become extremely important, so the use of the TSD as an oxide probe may prove as valuable an exercise as the study of the device in its own right. In any event, a discussion of the properties of thin oxides with an emphasis on SiO_2 is warranted.

2.4.2 Electrical Properties and Characterization

I-V and *C-V* Characteristics

The electrical properties of an oxide layer are most commonly and readily described in terms of a 1-D two terminal MOS device. Consider the band diagram for an MOS structure consisting of a metal (say, Al) layer, a SiO_2 layer, and an n-Si layer (Figure 2.8). In forward bias (Figure 2.8(a),(b)), majority carriers in the silicon accumulate at the oxide-silicon interface. Although the thickness of the oxide will determine how many of these carriers pile up at the interface and how many will be allowed to tunnel through the oxide, the basic behavior is the same regardless.

In reverse bias, however, two different behaviors are observed for thick and thin oxides (Figure 2.8(c)-(f)). If the oxide is thick enough to prevent appreciable tunnel currents from flowing for the range of bias considered, then very little current will flow in the silicon layer. Thus the carriers in the silicon will be nearly in electronic equilibrium, and can be described by a single flat Fermi level determined by the applied bias. Even if small tunnel currents are allowed, this picture remains valid. As the oxide

is thinned, however, eventually tunnel currents become larger than generation currents in the depleted silicon region. At this point, if bias is increased across the junction, the additional bias is dropped across the depletion region, rather than the oxide, since this region has become the current limiting part of the device. Therefore, the electron and hole quasi Fermi levels will split apart and bend in order to accommodate this bias. This turning point will be bias dependent, and will depend on many properties of the structure such as doping concentration and recombination lifetimes in the silicon, oxide transport properties, and gate-oxide barrier height. For an Al gate, SiO_2 barrier, and lightly doped n-Si semiconductor, this crossover occurs for oxide thicknesses of roughly 30-40 Å [29, 30].

We pause here to illustrate a distinction in tunneling in MOS structures that will become important in later discussion. In Figure 2.8(e), notice that the electric field across the oxide is strong enough that the electrons tunneling from the metal contact towards the silicon enter allowed SiO_2 conduction band states before entering the silicon. Note that a minimum bias, depending only on the relative band alignments of the electrodes and the oxide, must be dropped across the oxide in order for this situation to arise. This so called Fowler-Nordheim tunneling regime is qualitatively different from the direct tunneling regime depicted in Figure 2.8(f), in which electron must tunnel completely through the SiO_2 layer, only entering allowed states in the silicon layer. For example, that the effective thickness through which the electrons must tunnel is constant for direct tunneling, but is field dependent for Fowler-Nordheim tunneling. As a result, the I-V characteristics of the Fowler-Nordheim regime are much more weakly dependent on the details of the band structure of the oxide and the electrodes. This distinction is important in comparing thick and thin oxide behavior. For instance, for appreciable tunnel currents to flow through a thick oxide, enough bias must be applied that the Fowler-Nordheim regime is the only one of importance, but as the oxide is thinned, direct tunneling becomes important as well. In fact, for very thin oxides, the Fowler-Nordheim regime cannot be reached without breakdown of the oxide.

While the effects of thin oxides on I-V curves (increased currents and redistri-

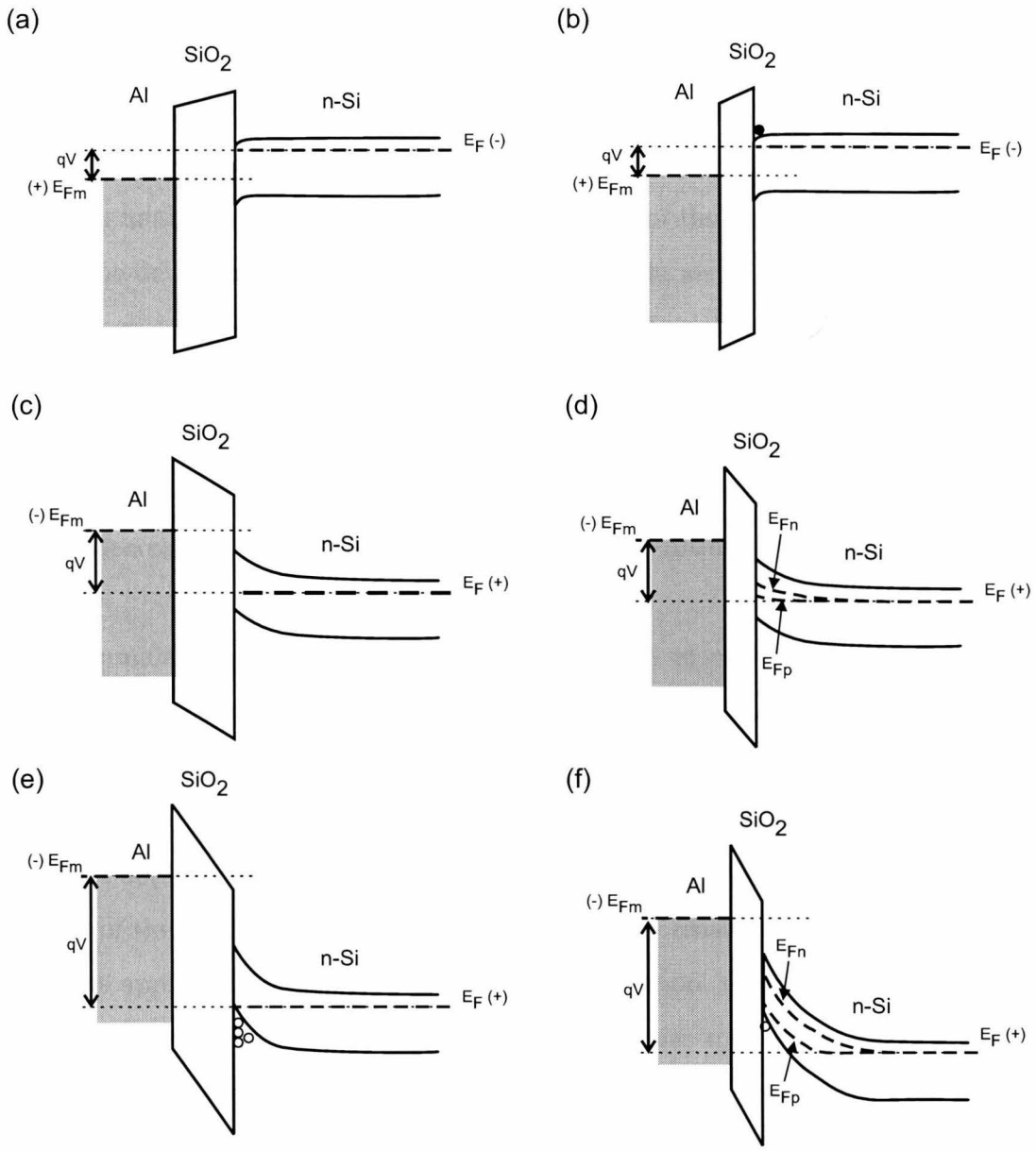

Figure 2.8: Comparison of Thick and Thin Oxide MOS Devices. Band diagrams for metal/SiO_2/n-Si MOS diodes in (a),(b) accumulation, (c),(d) depletion and (e),(f) inversion. Figures (a),(c),(e) assume the oxide layer is thick enough that the carriers in the silicon layer are all in electronic equilibrium. Figures (b),(d),(f) assume that the oxide layer is thin enough that this equilibrium is upset. In each figure, the applied voltage, V, the metal Fermi level, E_{Fm}, and the silicon contact Fermi level, E_f, are shown. In figures (d),(f), the electron and hole quasi Fermi levels, E_{Fn} and E_{Fp}, are also shown.

bution of potential drop in reverse bias) can be readily understood from Figure 2.8, an understanding of C-V behavior requires a bit more scrutiny. Before examining C-V characteristics of oxides which allow finite tunnel currents, however, we will first examine the textbook case of a completely insulating oxide layer [12].

Consider the same MOS structure described above and illustrated in Figure 2.8. In order to analyze the small signal AC capacitance of the device at a given DC bias, we treat the device as a simple capacitor with the gate metal as one electrode and the part of the silicon which experiences no band-bending as the other. The characteristic "thickness," d, of the capacitor used in the analysis can be determined from the band diagrams (Figure 2.8) by considering where additional charge piles up as an AC modulation bias is added to the fixed DC bias. The capacitance per unit area will then decrease as this thickness increases. The resulting C-V curve is depicted in Figure 2.9.

In accumulation, the AC bias causes modulation of carriers at the silicon-oxide interface, so the characteristic thickness of the capacitor is simply the oxide thickness. In depletion and inversion, however, the characteristic thickness depends on the frequency of the AC signal. First, consider low, near-static frequencies. As negative DC bias is first applied, the characteristic thickness is given by the oxide thickness plus the thickness of the depletion region which grows with increased DC bias. When enough DC bias is applied to force the device into accumulation, the characteristic thickness decreases once again to the oxide thickness as AC bias modulates minority carriers to the interface (see Figure 2.8(e)). Note that in order to reach inversion, electron-hole pairs must be generated in the depletion region, and the holes subsequently are forced to the surface while electrons are forced towards the semiconductor contact. As this process is relatively slow in silicon due to the long recombination lifetimes, at a high enough frequency, the process is disallowed. Therefore, the AC bias only can move carriers to the edge of the depletion layer, and hence the characteristic thickness remains fixed at the oxide thickness plus the static depletion width.

If the effective area of the MOS capacitor is known, then the thickness of the oxide can be extracted from the accumulation capacitance, and the substrate doping level

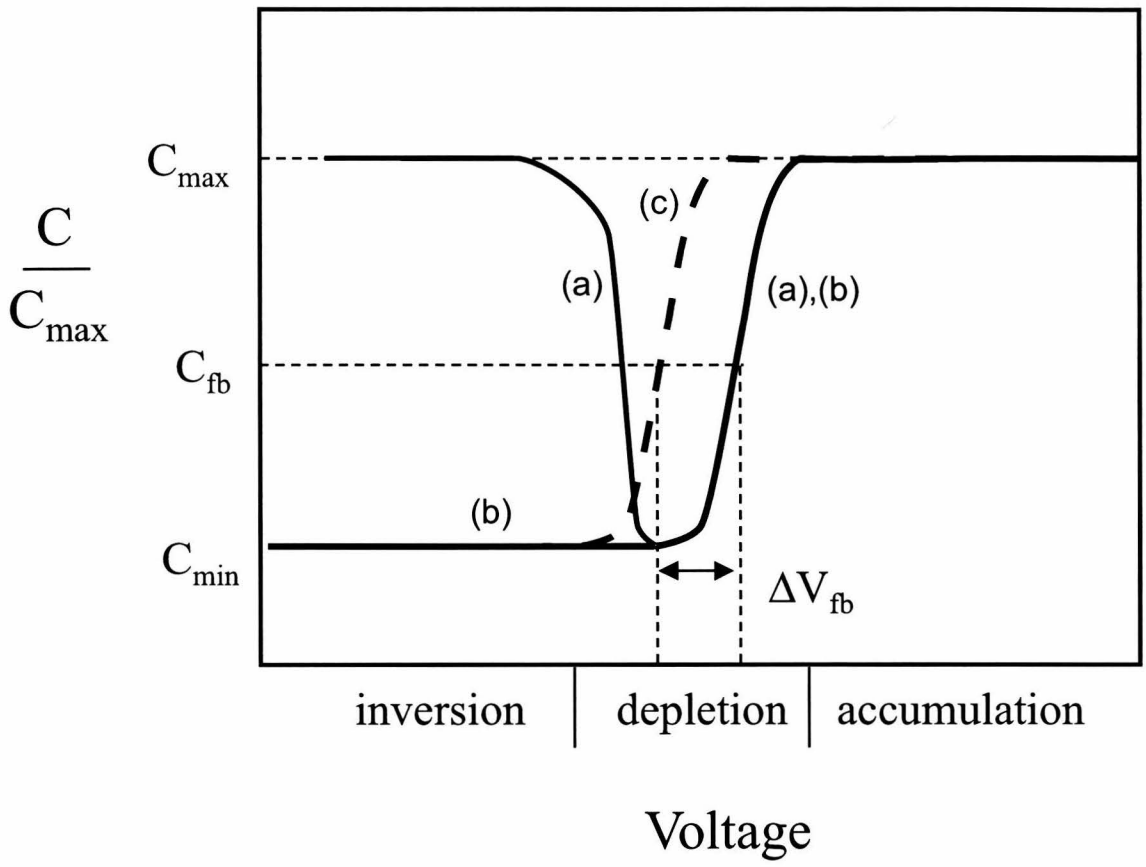

Figure 2.9: (a) Low frequency and (b) high frequency C-V curves for a thick oxide MOS device. (c) Represents the ideal high frequency C-V curve for an MOS device with no fixed charge at the oxide-silicon interface, shifted in voltage from the measured curve. The quantity of fixed charge can be determined by measuring the shift in voltage of the curve at the flat band capacitance, C_{fb}. Adapted from [12].

can be extracted from the depletion and inversion capacitance. Also, note that if a quantity of charge is permanently stored at the oxide-silicon interface, this will induce a shift in the voltage of the C-V curve, proportional to the amount of charge present and to the oxide thickness, as it induces an additional bias across the device (compare Figures 2.9(a),(c)). Such a population of charge is typically present in MOS devices in small amounts [12]. Additional charge found in trap states differs in amount with varying bias, and hence leads to a nonuniform shift in the C-V curve. Note that as the oxide gets thinner, however, this voltage shift decreases, becoming harder to detect.

The above analysis certainly is valid for very thick oxides, but also remains valid for oxides down to roughly 30 Å if modified slightly. In order to treat tunnel currents, for example, a parallel resistor is added to the effective circuit used in extracting the capacitance values, allowing the capacitor to be "leaky." Most LCR bridge measurement circuits can automatically make this adjustment. Figure 2.8(f) illustrates another necessary modification in the reverse bias regime. Instead of going into accumulation, the device enters "deep depletion" as bias is increased. Therefore, the C-V curve does not flatten out, but continues to drop with increasing bias. As the thickness of a depletion region is proportional to the square-root of the bias across it, the capacitance falls off as $1/\sqrt{V - V_{fb}}$, where V_{fb} is the flat band voltage. Another adjustment is necessary when the desired resolution of the oxide thickness measurement is comparable to the Debye length in the accumulation or inversion region of the semiconductor. The carriers accumulated at the interface are distributed throughout a thickness characterized by the Debye length, so this length must be subtracted from the measured accumulation thickness to obtain the true oxide thickness. Similar modifications must be made to account for the spread of carriers on the gate side of the oxide, particularly when the gate material is, say, doped polysilicon instead of metal [31, 32]. All of these changes are readily implemented, however, and still allow for extraction of useful information.

As the oxide thickness is decreased further, however, tunnel currents increase, and the behavior of the quasi Fermi levels becomes more complicated, as discussed above. At this point, the small signal operation of the device can no longer be described

simply as a capacitor in series with a resistor. Figure 2.10 depicts measured C-V curves and corresponding I-V curves for thin oxide Al/SiO_2/p-Si devices which will be described in Section 3.4.2. While the C-V curve corresponding to the thickest oxide device closely resembles the ideal curve in Figure 2.9, the thinner oxides exhibit gross departures. Some researchers have successfully devised more complicated effective circuits for analysis of such thin oxides [33–35] (see Figure 2.10(c)-(e)), but analysis of these circuits relies on separation of the measured signal into many different components, which requires additional measurements or explicit models of these components.

Electrical Stress

As far as the MOSFET community is concerned, one of the most serious complications that arises with thin oxides is damage caused by electrical stress. Over the past decade, a staggering amount of effort has been directed towards stress measurements of silicon MOS devices across a wide range of parameter space [34, 36–43], investigations of the mechanisms involved in the subsequent degradation of the oxide [37, 38, 43–46], attempts to bolster oxide resistance to such degradation [34, 47–51], and studies of the physical and chemical nature of the induced damage [39, 52–54]. Although it is not feasible to summarize all of this work in this thesis, we will touch on a few of the most important points.

Typical measured breakdown fields of thicker oxides (hundreds of Å) are roughly 5-10 MV/cm [55], which suggests that a few volts could be dropped across an oxide 20-30 Å thick, but the true situation is significantly more complicated. Even if a voltage below these breakdown voltages is applied to a thin oxide MOS device, the tunnel currents that flow damage the oxide, making it more conductive. More current is then allowed to flow, and the oxide is further degraded. As this process continues over time, the oxide is degraded to a "breakdown" point. The definition of this point is somewhat ill-defined for very thin oxides, but it is usually characterized by a sharp increase in conductivity [37, 43]. There is evidence that even after the oxide has been degraded to this point, it has not become an electrical short – it appears to still

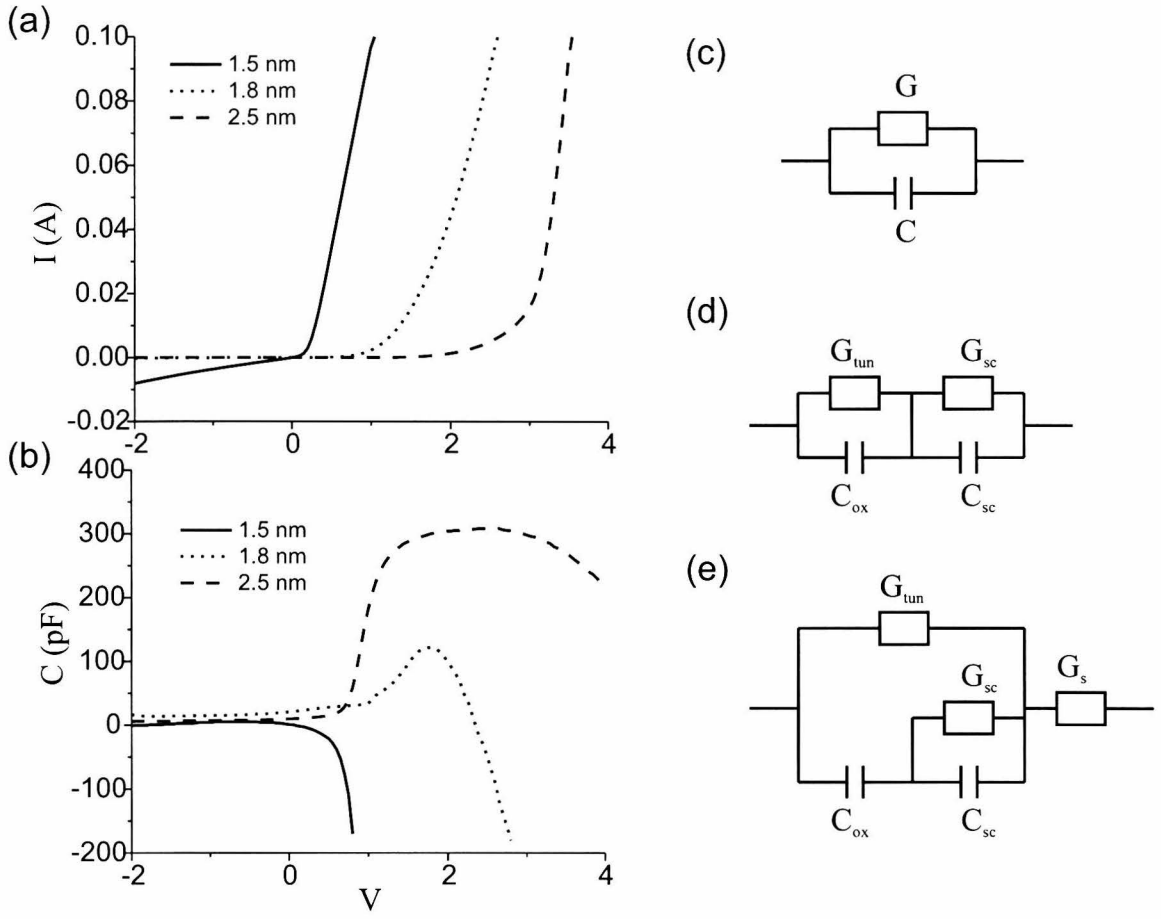

Figure 2.10: C-V analysis of thin oxide MOS devices. Measured (a) I-V curves and (b) C-V curves for Al/SiO_2/p-Si structures with indicated oxide thicknesses. (c) Simple effective circuit assumed in extracting the capacitance data in (b). (d),(e) More complicated effective circuits that have been used to describe thin oxide MOS C-V characteristics. C_{ox} represents the oxide capacitance, C_{sc} represents capacitance of the depletion (space charge) region, G_{sc} represents the conductance of this region, G_{tun} represents tunnel current conductance, and G_s accounts for series resistance.

act as a partial barrier, although transport may involve some kind of "trap-assisted" tunneling process [37, 39, 43]. The total integrated charge that has traversed the oxide up to this point is called the "charge to breakdown," Q_{bd}, and is quoted as a figure of merit in describing oxide integrity.

While the use of Q_{bd} as a figure of merit is widespread, it should be noted that this parameter is by no means an ideal benchmark [32]. It depends heavily on a number of parameters, increasing with decreased stress currents and decreased oxide thickness, for example [37, 56]. Local stressing using Ballistic Electron Emission Microscopy (BEEM) indicates much greater values of Q_{bd} than obtained from measurements on macroscopic devices when similar stressing parameters are employed [57, 58]. Therefore, it is useful primarily in comparing two similar structures in which a small change has been made in the processing of the structure.

The mechanism of oxide degradation through current injection is still not completely well understood, but studies of this process have contributed greatly to the understanding of the chemical and physical nature of the Si-SiO_2 interface. These structural properties of MOS devices, and their relationships to electrical stress, will be discussed in the following section.

2.4.3 Structural Properties and Characterization

In an attempt to probe thin oxides and the Si-SiO_2 interface, and to relate electrically induced changes to physical and chemical changes, many measurement techniques have been employed such as ellipsometry [59–61], X-ray Photoelectron Spectroscopy [39, 52–54], and local probe techniques such as Scanning Tunneling Microscopy (STM) and Ballistic Electron Emission Microscopy(BEEM) [57, 58, 62, 63]. Again, only a few select results of this body of work will be presented here.

As Si and SiO_2 have quite different crystal structures and lattice parameters, SiO_2 layers grown on Si are typically amorphous, although some evidence of partial crystallinity has been observed for very thin oxides [39]. In fact, as the SiO_2 structure is not particularly dense, many Si atoms on the Si surface cannot be bonded to O

atoms at the interface without generating tremendous stress energy [53]. These atoms, therefore, create dangling bonds which project into the oxide, and are believed to be terminated with Hydrogen atoms [58, 64]. In addition, experimental [52, 54] and theoretical [53] studies suggest that, in order to relieve strain, roughly 10-20 Å of the oxide adjacent to the interface is not stoichiometric SiO_2, but is oxygen deficient SiO_x.

One widely publicized mechanism of oxide degradation through electrical stressing involves current-induced desorption of the Hydrogen attached to the non-bonded silicon atoms, which leaves behind a trap in the form of a Si dangling bond. In order for the hydrogen atoms to be stripped from the silicon sites, they must gain roughly 2 eV of energy from the impinging electrons [58, 64]. This suggests that relaxation of hot electrons must be involved, such as those tunneling into the SiO2 conduction band in the Fowler-Nordheim regime, and some evidence has suggested that such an energy threshold to damage creation does exist [58]. This cannot be the only mechanism of degradation, however, as much lower voltage stresses can also cause damage. One suggestion is that energy can be integrated from a stream of electrons, stored and accumulated as vibration mode energy, allowing for lower voltage damage [50]. Other damage mechanisms involving structural degradation of the oxide at the Si-SiO_2 interface [39, 58] and charge trapping [40, 41, 65] have been proposed.

Note that for very thin oxides, the presence of a sub-oxide layer at the interface could mean that the majority of the oxide may have very different properties than those of bulk SiO_2 . For example, dielectric constants, electrical barrier heights, and optical constants may all be a function of position in the oxide. In fact, for such thin oxides, many of these properties are ill-defined. This will have a significant impact on standard measurements of such properties as oxide thickness. If the thickness is determined from ellipsometry, the choice of optical constants may be important [59, 61]. In addition to the problems described above with C-V determination of the thickness of a thin oxide, a gradient in dielectric constant will worsen the problem. In short, it may not be possible to use such measurements to obtain absolute measurements of such parameters as oxide thickness, but such measurements could still be used for differential measurements, such as determining whether one oxide is thicker than

Table 2.5: A comparison of maximum stress operating parameters of the TSD devices produced in our labs with those of current thin oxide MOS research structures [37, 45]. For the purposes of a rough comparison, we define "Q_{max}" as the areally-averaged maximum fluence of charge that results in a nonfunctioning device, and we compare this with Q_{bd} in MOS devices. As our time-tested TSD devices did not expire in the course of our testing, we can only give a lower bound on this number.

	t_{oxide} (Å)	Q_{max}, Q_{bd} (C/cm^2)	J_{oxide} (A/cm^2)	Electric Field (MV/cm)
TSD	<35	$>10^7$	$>10^4$	10-20
MOS	30 - 60	$<10^5$	<10	<15

another.

2.4.4 Summary

This terse discussion of thin oxide properties illustrates that as oxides are thinned down to roughly 30 Å, many of the traditional characterization methods and concepts of transport break down, stress-induced damage becomes important, and structural imperfections may comprise the majority of the material. In general, while much has been learned through years of research, there is still much to learn.

It should be noted that the bulk of the past research in this area has been aimed at describing and improving the characteristic of the Si MOSFET – a device in which oxide tunnel current is to be avoided. In this thesis, we consider the Tunnel Switch Diode, in which tunnel currents are desirable, and in fact critical to the device operation. A rough comparison of TSD operating parameters to those of MOS research (Table 2.5) indicates that we are likely to be exploring a different region of parameter space. Therefore, we must on one hand expect difficulties in fabricating and characterizing the oxides utilized in our studies of the TSD, but we must also acknowledge and embrace the possibility that new insight into the properties of thin oxides may be gained along the way.

Bibliography

[1] T. Yamamoto and M. Morimoto, *Appl. Phys. Lett.* **20**, 269 (1972).

[2] T. Yamamoto, K. Kawamura, and H. Shimizu, *Solid–State Electron.* **19**, 701 (1976).

[3] H. Kroger and H. A. R. Wegener, *Solid–State Electron.* **21**, 643 (1978).

[4] H. Kroger and H. A. R. Wegener, *Solid–State Electron.* **21**, 655 (1978).

[5] H. Kroger and H. A. R. Wegener, *Appl. Phys. Lett.* **23**, 397 (1973).

[6] J. D. Hwang, Y. K. Fang, K. H. Chen, and H. Y. Chiu, *IEEE Trans. Electron Devices* **42**, 2246 (1995).

[7] H. Kroger and H. A. R. Wegener, *Appl. Phys. Lett.* **27**, 303 (1975).

[8] N. T. Ali and R. J. Green, *IEEE Trans. Electron Devices* **42**, 1978 (1995).

[9] Y. K. Fang, K.-H. Lee, K.-H. Wu, and C.-Y. Tsao, *IEEE Trans. Electron Devices* **44**, 34 (1997).

[10] S. S. Ang, *IEEE Trans. Electron Devices* **35**, 1378 (1988).

[11] J. G. Simmons and A. A. El-Badry, *Radio and Elec. Eng.* **48**, 215 (1978).

[12] S. M. Sze, *Physics of Semiconductor Devices*, Wiley, New York, NY, 2nd edition, 1981.

[13] K. Itoh, *IEEE J. Solid-State Circuits* **25**, 778 (1990).

[14] K. Itoh, *Proc. IEEE* **83**, 524 (1995).

[15] H. Nambu et al., *IEEE J. Solid-State Circuits* **27**, 632 (1992).

[16] H. J. Levy, *Application and Integration of Quantum-Effect Devices for Cellular VLSI*, PhD thesis, Caltech, 1994.

[17] H. J. Levy, E. S. Daniel, and T. C. McGill, (to be published).

[18] H. J. Levy and T. C. McGill, Patent # 5,535,156, 1996.

[19] Micron Technology, Inc., Data sheet - MT58LC128K32/36D8 128K x 32/36 syncburst SRAM, 1997.

[20] IBM Corporation, Data sheet - IBM04184ARLAA 256K x 18 SRAM, 1997.

[21] Micron Technology, Inc., Data sheet - 16 Meg x 4 EDO DRAM, 1997.

[22] IBM Corporation, Data sheet - IBM0364804C 64Mb Synchronous DRAM, 1997.

[23] L. O. Chua and L. Yang, *IEEE Trans. Circuits Syst.* **35**, 1257 (1988).

[24] A. Seabaugh et al., *GaAs IC Symposium* (1997).

[25] P. van der Wagt, A. Seabaugh, and E. Beam III, *Technical Digest IEDM*, 425 (1996).

[26] W. Williamson III et al., *IEEE J. Solid- State Circuits* **32**, 222 (1997).

[27] G. B. Barbi, J. S. Blnco, M. Baroffio, J. Agapito, and F. J. Gutierrez, *Sensors and Actuators B* **18-19**, 93 (1994).

[28] S. Nakagomi and T. Yamamoto, *Jpn. J. Appl. Phys.* **33**, 6136 (1994).

[29] J. Shewchun, M. A. Green, and F. D. King, *Solid–State Electron.* **17**, 563 (1974).

[30] R. A. Clarke and J. Shewchun, *Solid–State Electron.* **14**, 957 (1971).

[31] M. Depas, B. Vermeire, P. W. Mertens, R. L. Van Meirhaeghe, and M. M. Heyns, *Solid–State Electron.* **38**, 1465 (1995).

[32] D. A. Buchanan and S.-H. Lo, *Proc. Electrochem. Soc.* **96-1**, 3 (1996).

[33] M. Depas, R. L. Van Meirhaeghe, W. H. Laflere, and F. Cardon, *Semicond. Sci. Technol.* **7**, 1476 (1992).

[34] M. Depas, R. L. Van Meirhaeghe, W. H. Laflere, and F. Cardon, *Solid–State Electron.* **37**, 433 (1994).

[35] S. Kar and W. E. Dahlke, *Solid–State Electron.* **15**, 221 (1972).

[36] W. Weber, M. Brox, R. Thewes, and N. S. Saks, *IEEE Trans. Electron Devices* **42**, 1473 (1995).

[37] K. F. Schuegraf and C. Hu, *J. Appl. Phys.* **76**, 3695 (1994).

[38] R. Rofan and C. Hu, *IEEE Electron Device Lett.* **12**, 632 (1991).

[39] M. Hirose, J. L. Alay, T. Yoshida, and S. Miyazaki, *Proc. Electrochem. Soc.* **96-1**, 485 (1996).

[40] M. O. Andersson, K. R. Farmer, and O. Engstrom, *J. Appl. Phys.* **71**, 1846 (1992).

[41] J. Maserjian and N. Zamani, *J. Vac. Sci. Technol.* **20**, 743 (1982).

[42] M. Depas, B. Vermeire, and M. M. Heyns, *J. Appl. Phys.* **80**, 382 (1996).

[43] M. Depas, T. Nigam, and M. M. Heyns, *IEEE Trans. Electron Devices* **43**, 1499 (1996).

[44] S. Nagano, *J. Appl. Phys.* **75**, 3530 (1994).

[45] T. Yoshida, D. Imafuku, J. L. Alay, and S. Miyazaki, *Jpn. J. Appl. Phys.* **34**, L903 (1995).

[46] D. J. DiMaria and J. H. Stathis, *Appl. Phys. Lett.* **70**, 2708 (1997).

[47] V. V. Afanas'ev, M. Depas, J. M. M. de Nijs, and P. Balk, *Microelectronic Engineering* **22**, 93 (1993).

[48] M. Bhat et al., *IEEE Electron Device Lett.* **15**, 421 (1994).

[49] H. Fukuda, M. Yasuda, T. Iwabuchi, and S. Ohno, *IEEE Electron Device Lett.* **12**, 587 (1991).

[50] J. W. Lyding and K. Hess, *Appl. Phys. Lett.* **68**, 2526 (1996).

[51] Y. Okada, P. J. Tobin, P. Rushbrook, and W. L. DeHart, *IEEE Trans. Electron Devices* **41**, 191 (1994).

[52] Z. H. Lu, M. J. Graham, D. T. Jiang, and K. H. Tan, *Appl. Phys. Lett.* **63**, 2941 (1993).

[53] A. Pasquerello, M. S. Hybertsen, and R. Car, *Phys. Rev. B* **53**, 10942 (1996).

[54] M. H. Hecht, R. P. Vasquez, F. J. Grunthaner, N. Zamani, and J. Maserjian, *J. Appl. Phys.* **57**, 5256 (1985).

[55] S. Wang, *Fundamentals of Semiconductor Theory and Device Physics*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[56] M. Depas, B. Vermeire, P. W. Mertens, M. Meuris, and M. M. Heyns, *Semicond. Sci. Technol.* **10**, 753 (1995).

[57] R. Ludeke, A. Bauer, and E. Cartier, *Appl. Phys. Lett.* **66**, 730 (1995).

[58] R. Ludeke, H. J. Wen, and E. Cartier, *J. Vac. Sci. Technol. B* **14**, 2855 (1996).

[59] E. A. Irene, *Thin Solid Films* **233**, 96 (1993).

[60] M. Weidner, A. Röseler, and M. Eichler, *Thin Solid Films* **234**, 337 (1993).

[61] D. Tonova, M. Depas, and J. Vanhellemont, *Thin Solid Films* **288**, 64 (1996).

[62] B. Kaczer, Z. Meng, and J. P. Pelz, *Phys. Rev. Lett.* **77**, 91 (1996).

[63] B. Kaczer and J. P. Pelz, *J. Vac. Sci. Technol. B* **14**, 2864 (1996).

[64] D. J. DiMaria, E. Cartier, and D. Arnold, *J. Appl. Phys.* **73**, 3367 (1993).

[65] K. R. Farmer, M. O. Andersson, and O. Engstrom, *Appl. Phys. Lett.* **60**, 730 (1992).

Chapter 3 Device Design and Experimental Procedures

3.1 Introduction and Chapter Outline

This chapter describes in detail the experimental procedures and analysis techniques used in our studies of the TSD and of the thin oxides used in these devices. Section 3.2 describes the choices of target parameters for the TSD device fabrication, as well as the logic and motivation behind these choices. Section 3.3 describes in detail the procedures used in preparation of all samples. Due to the limitations of the processing capabilities available at Caltech, devices were produced at two other facilities as well: The Stanford Nanofabrication Facility (SNF), and Intel Components Research. The procedures used at each of these facilities are discussed separately. Section 3.4 discusses the measurement and characterization techniques used and their limitations.

3.2 Design of Target Structures and Description of Commercial Wafers Used

The primary motivation for the production and testing of these structures was the exploration of parameter space of the TSD in order to optimize performance for the SRAM application. As outlined in Section 2.3, in order for this application to be realized, certain TSD device parameters must be brought into ranges defined by existing technologies. Therefore, memory characteristics and corresponding TSD parameters were defined as targets for this study as outlined in Table 3.1 (see Figure 2.1 for definitions of parameters).

Table 3.1: Desired properties of an SRAM memory, and the corresponding necessary TSD properties. See Section 2.3.1 for discussion of the origin of these relations, and see Figure 2.1 for definitions of TSD parameters.

SRAM Feature	TSD Constraint
Low quiescent state current	$J_{valley} \times V_{valley} \lesssim 1$ W/cm^2
Fast read times, small device areas	$I_{high} \times t_{read} \sim 100$ fC, $t_{read} \lesssim 10$ ns $\Rightarrow I_{high}/I_{low}$ huge (10^4 - 10^5)
Fast write times	Fast switching times ($\lesssim 10$ ns)
Reasonable operating voltage range	$V_{peak} \sim 5$ V, $V_{valley} \lesssim 1$ V
Temperature insensitive performance	Temperature insensitive parameters

Unfortunately, as the details of the device physics relevant to TSD operation are not completely well understood, and as the bulk of the previous work on TSD devices was focused on high power devices, there is little information available for use in the engineering of our TSD devices. Therefore, the majority of the desired parameters could not be predicted, but rather, required exploration of parameter space. The only parameter which had been studied enough to allow for design is the peak voltage, V_{peak}, although most reported target voltages have been substantially greater than ours, so we might not expect particularly good accuracy. As discussed in section 2.2.2, however, we can design a device with a given peak voltage in punch-through mode by choosing the top silicon layer doping concentration and thickness, or an avalanche mode by choosing just the doping concentration, and requiring the thickness to be greater than the punch-through thickness at the avalanche voltage.

All of the wafers used for production of TSD devices, consisting of p^+-Si substrates with n-Si CVD-grown epilayers, are listed in Table 3.2. The wafers with epilayers doped at 2×10^{15} were designed to operate in punch-through mode with peak voltages of 5-10 V, and the wafers with 2×10^{17} epilayers were designed to operate in avalanche mode with peak voltages of about 5 V (see Table 2.1 and Figure 2.4). As most reported devices using SiO_2 as an insulator material used p-type substrates with n-type epilayers, we chose the same polarity for our devices.

All of the samples produced at Caltech used wafer set S1. Unfortunately, sub-

sequent measurements[1] indicated an epilayer thickness of only 1.0-1.4 μm – substantially lower than the target. In addition, while the carrier concentration was 2-3$\times 10^{15}$ cm^{-3}, roughly what was expected, the doping level was not particularly uniform across the test wafer, as indicated by resistivity mapping which showed an annulus of increased resistivity near the outer edge of the wafer. As will be shown in the next section, however, it was still possible to obtain useful information from the devices prepared with these wafers.

Once this error was found, wafer set S2 was ordered from SQI. We hoped to be able to order a set of avalanche mode wafers as well, but SQI could not grow epi layers with as high a dopant concentration as was required (2×10^{17} Phosphorus/cm^3). With this set of wafers, SQI provided spreading resistance profile data, indicating an epilayer thickness of 3.3 μm, epilayer resistivity of 2.3 Ω-cm, and a substrate resistivity of 1.5×10^{-2} Ω-cm. These values correspond to doping concentrations of roughly 2×10^{15}/cm^3 for the epilayer and 2×10^{18}/cm^3 for the substrate if mobilities of 1400 cm^2/V-s and 50 cm^2/V-s are used respectively. Delivery of the wafers was delayed for more than a year, so no devices were produced at Caltech using these wafers. They were used, however, in fabrication of devices at Stanford, along with wafer set H3. Wafer sets I1-I3 were used in fabrication of devices at Intel.

3.3 Fabrication Procedures

3.3.1 Device Geometry

Although many different mask sets were used in the production of the devices, only two basic geometries were employed. One device type which we will label a "flat device" consists of the silicon layers, followed by the thin tunnel oxide layer, and a patterned gate pad on top (Figure 3.1(a)). With this type of device, the size of the pad defines the size of the device. For all devices investigated, the size of the pad was much larger than the thickness of all of the layers down to the substrate, so in the

[1]Spreading resistance profiling and resistivity mapping by Solecon Labs

Table 3.2: Commercial wafers used for TSD devices. All wafers are (100) p-type Boron doped substrates with n-type Phosphorus doped CVD-grown epilayers. All measured values were obtained from spreading resistance measurements performed by the manufacturer, except for wafer set S1, in which case the spreading resistance measurements were performed by Solecon Labs. Estimated peak values were obtained from Figure 2.4.

		Sub. Conc. (cm^{-3})		**Epi Conc. (cm^{-3})**		**Epi Thick. (μm)**		**Est.**
Index	**Source**	**Target**	**Measured**	**Target**	**Measured**	**Target**	**Measured**	**V_{peak}**
S1	SQI	2×10^{18}	2×10^{18}	2×10^{15}	$1\text{-}3\times10^{15}$	3.0	1.0-1.4	<1
S2	SQI	2×10^{18}	2×10^{18}	2×10^{15}	2×10^{15}	2.5-3.5	3.3	12
H1	Hughes	2×10^{18}	2×10^{18}	2×10^{15}	2×10^{15}	2.5	2.5	5
H2	Hughes	2×10^{18}	2×10^{18}	2×10^{15}	2×10^{15}	3.0	3.0	8
H3	Hughes	2×10^{18}	2×10^{18}	2×10^{17}	2×10^{17}	1.0	0.9	6
I1	Intel	2×10^{18}	–	2×10^{15}	–	1.5	–	1
I2	Intel	2×10^{18}	–	2×10^{15}	–	2.5	–	5
I3	Intel	2×10^{18}	–	2×10^{17}	–	1.0	–	6

absence of large charge concentrations in the layers, one would expect the device to act roughly as a one-dimensional stack with a size dictated by the size of the gate.

The second type which we will label a "field oxide device" consists of the silicon layers, followed by a thick "field" oxide containing a patterned window reaching down to the tunnel oxide layer, and a patterned gate pad covering the window (Figure 3.1(b)). In this case, the window defines the area of the device. Again, in the absence of large charge concentrations and for windows larger than the thickness of the silicon epi layer, the device should act roughly as a one-dimensional stack. This geometry is preferable to the flat geometry for several reasons. First of all, the gate pad can be made much larger than the window, so the active device area can be quite small, even if the gate must be large to accommodate contacting with a probe tip or a wirebond. Second, consider the process of making a contact to the gate pad. With a flat device, a metal probe or a wirebond probe tip must press down on the gate, possibly damaging the thin tunnel oxide underneath. With a field oxide device, however, the contact to the gate can be made on top of the field oxide instead, leaving the tunnel oxide undamaged. Third, and perhaps most important, a flat device will have larger fringing fields at the edges and corners, whereas the extended gate of a field oxide device tends to diminish these effects (hence the name "field" oxide).

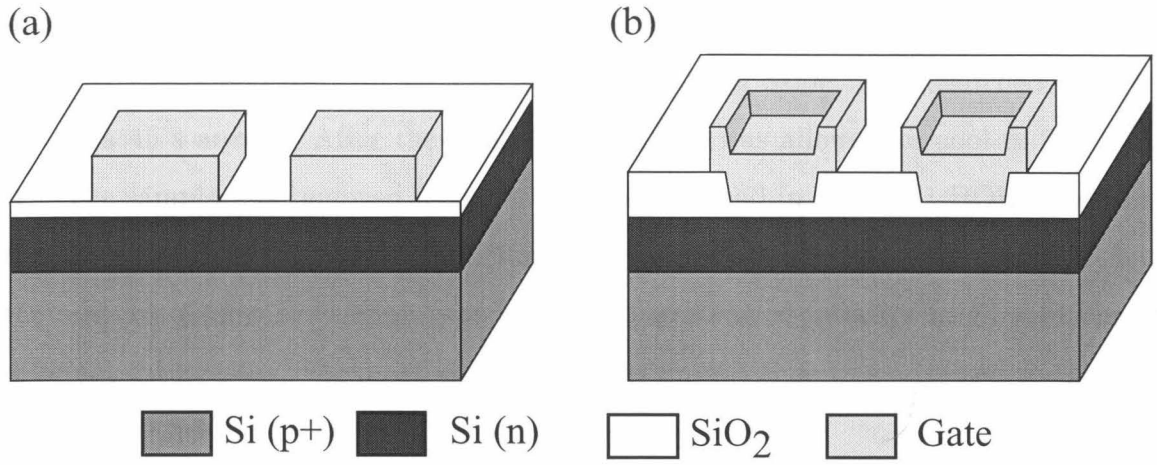

Figure 3.1: Flat and field oxide device geometries. (a) Flat device. Only the gate layer is patterned, and it lies directly on top of the tunnel oxide which covers the crystal layers. (b) Field oxide device. A thick field oxide covers the silicon, but has patterned windows which contain the thin tunnel oxide. The equipotential gate pad extends across the windows, diminishing fringing fields at the edges.

3.3.2 Process Steps Used at Caltech

Fabrication of flat devices consisted of three major steps: cleaning, oxidation, and gate deposition and definition. Each of these will be described in detail here. [1, 2]

A roughly 1 cm^2 piece of a wafer was cleaved off and degreased in Acetone, Isopropanol, and DI-H_2O for 30 s each in ultrasound. The sample was then cleaned using a modified RCA process [3]. This process consists of immersion in two solutions: SC-1 (NH_4OH:H_2O_2:H_2O 1:4:20 with 1 ppm of Ethylenediaminetetraacetic acid (EDTA) at 80°C) and SC-2 (HCL:H_2O_2:H_2O 1:1:5 at 80°C). The sample was placed in the SC-1 solution for 10 minutes followed by an overflow rinse in DI water for another 10 minutes. Then, the sample was spun dry and immersed in SC-2 for 5 minutes. To remove any native oxide and oxide chemically grown during the SC-2 immersion, the sample was dipped in 10% 10:1 Buffered Oxide Etchant(BOE) (the buffer was NH_4F) followed by a dip in DI water.

Immediately after the DI water dip, the sample was placed into a rapid thermal processing oven[2] (RTP). After a purge of the sample chamber with Ar, the sample temperature was ramped to 900°C at 200°C/s in Ar ambient. After the process

[2]Heatpulse Minipulse 310, AG Associates, Inc.

temperature of 900°C was reached and held for 15 s, the ambient gas was switched from Ar to O_2 and held for 60 s to grow a sacrificial oxide, then switched back to Ar for a 45 s anneal. After the anneal, the sample was allowed to cool to 250°C in Ar. The sample was removed from the oven and dipped in 10% 10:1 BOE to remove the sacrificial oxide. Then the tunnel oxide was grown using the same procedure, but the amount of time the sample was held in O_2 was varied in order to grow oxides of different thicknesses. For most samples, the Ar anneal time was adjusted so that the sample was held at 900°C for a total of 120 s.

Immediately after tunnel oxide growth, the sample was transferred to a magnetron plasma sputtering chamber with base pressure $\sim 5 \times 10^{-8}$ Torr. An Al layer, typically 75 nm but sometimes as thick as 200 nm, was deposited using DC magnetron sputtering in $\sim 4 \times 10^{-3}$ Torr Ar. Square photoresist pads of varying sizes were lithographically defined, and the Al layer was etched in the developer (AZ 422 MIF) used for developing the photoresist (AZ 4210). An Al back-side contact to the substrate was also deposited in the sputter chamber, following a 50W RF magnetron sputter of the sample back side for ~1 min. in order to remove any native oxide. The sample was then annealed at 400°C in order to improve the back side contact, and to form the Al-SiO_2 interface. The anneal was later determined to have little effect on device characteristics, however.

A field oxide process was attempted, using the following additional steps. After the cleaning process, a thick (100-150 nm) field SiO_2 layer was deposited on the sample using 150 W RF magnetron sputtering of a SiO_2 target in Ar ambient ($\sim 4 \times 10^{-3}$ Torr). Photoresist was spun on, windows were patterned, and the field oxide within the windows was removed with BOE etchant. Care had to be taken to bake the sample before spinning on the photoresist, or else it would delaminate during the etch, presumably due to moisture adhering to the oxide [4]. The hottest bake possible with the available Blue M ovens was 150°C so a long bake (several hours) was necessary. After the etch, the photoresist was removed in Acetone, and the sample cleaned in Isopropanol and D-H_2O. The processing then continued with the sacrificial oxide growth.

This field oxide process did not appear to produce reliable TSD devices. One

possibility that was considered was that the photoresist removal process after the field oxide etch contaminated the silicon in the windows so that subsequent tunnel oxide growth was flawed. We attempted to address this problem by etching only part of the way through the field oxide, removing the photoresist, and then etching the entire oxide layer a bit further so that all field oxide was etched away from the windows, but some was left elsewhere. The quality of the sputtered oxide was poor, non-uniform, and irreproducible, so the timing of etch rates necessary for this process was not possible. Therefore, only results on flat devices produced at Caltech will be reported. Future attempts at this process might be successful using process steps similar to those discussed in the following sections.

Some samples were produced using N_2O instead of O_2 as the process gas, which produces a SiO_xN_y oxide layer [5–7]. We did not seem to be able to reproduce this process reliably, however, as the device behavior (switching voltages, stress effects, etc.) was not at all reproducible. Therefore, this process was abandoned, and no further results of this process will be discussed.

A large array of samples was produced, but modifications in many of the process steps listed above were made from one sample to the next in order to establish a reliable fabrication procedure. Therefore, we present here results on a small subset of these samples which were fabricated with the same process steps, except for differences in tunnel oxidation time and maskset choice. These samples are listed in Table 3.3, indexed by their original identification numbers.

In order to estimate the thickness of the SiO_2 layers, the following method was used. Simple MOS devices were fabricated using the process described above, using a piece of a moderately doped p-type silicon (100) wafer in place of the p-n wafers used in TSD production. The oxide thickness was then measured using the C-V method which will be described in Section 3.4.2.

For the 60 s oxygen flow sample, a thickness of 2.5 nm was determined. This value is about 1 nm less than reported in published calibration curves (SiO_2 thickness, as measured by ellipsometry, versus oxygen flow time in RTP oxidation) [8], [9]. A similar offset of 0.9 nm has been reported by Duzelier *et al.* [10] when comparing C-V

Table 3.3: TSD and MOS devices produced at Caltech.

Index	Epi Wafer	Mode	O_2 Flow Time (s)	SiO_2 Thick. Estimate (nm)
C19	S1	P-T	60	2.5
C25b	p-type	MOS	15	1.5
C26b	p-type	MOS	30	1.8
C27b	p-type	MOS	60	2.5
C29	S1	P-T	60	2.5
C45	S1	P-T	15	1.5
C46	S1	P-T	30	1.8
C47	S1	P-T	45	2.1
C48	S1	P-T	60	2.5

thickness measurements with that of high resolution transmission electron microscopy and ellipsometry, suggesting that the constant offset could be attributed to the aluminum reacting with the SiO_2 layer, decreasing the effective thickness as measured by C-V techniques. We therefore find our measurement in good agreement with these results.

While the C-V technique yielded reasonable results for the 60 s sample (C27b), C-V analysis of the thinner oxides did not produce usable results for the reasons discussed in Section 2.4.2. The SiO_2 thicknesses of the thinner oxide samples were estimated by extrapolating from the measured 60 s thickness using a slope of 0.023 nm/s, the average value of the published curves mentioned above.

3.3.3 Process Steps Used at Stanford Nanofabrication Facility

The Stanford Nanofabrication Facility (SNF), residing on Stanford University campus, consists of a 10,000 square foot class 100 clean room facility and supporting equipment for wafer-level semiconductor processing. Much of the resident equipment (including oxidation furnaces, CVD reactors, Plasma etchers, wet benches, and lithography tools) is automated, a large fraction is dedicated to silicon processing, and all is maintained by teams of technicians. These enhanced capabilities allowed for more complex processing than was possible using the Caltech facilities.

Processing consisted of four major steps: cleaning, field oxide formation, tunnel

oxide formation, and gate patterning. Each will be discussed below.

As most of the equipment could not handle pieces of wafers, entire 4 inch wafers were processed. The processed wafers were cleaned using an RCA procedure similar to the one described above. Wafers were spun and blown dry.

The field oxide/tunnel oxide structure was grown using a LOCOS (LOCal Oxidation of Silicon) procedure (see Figure 3.2). A thermal "pad" oxide was grown on the wafer. Then, Si_3N_4 was CVD deposited over the oxide, and patterned. A thermal field oxide was then grown. The Si_3N_4 was removed using a hot phosphoric acid bath. A dilute HF bath was used to remove the pad oxide. Then, the thin tunnel oxide was grown in a furnace at 750°C for times ranging from 5 to 15 minutes.

The Al gate was deposited immediately following the oxidation. Standard lithography was used to pattern photoresist pads, and the Al gates were defined with an etch. Due to time constraints, the back contact could not be deposited at Stanford, and the 4 inch wafers could not easily fit into the deposition chamber at Caltech. Therefore, the back contacts to the wafers were made by scratching the surface to break through the native oxide, and melting In/Ag solder onto the wafer with a soldering iron.

A custom set of masks was designed for these samples, containing many different kinds of structures. For example, both field oxide and flat devices of varying sizes were present on every wafer. Flat devices could be produced using this field oxide process by placing large silicon nitride pads on some areas of the wafer, thus allowing no field oxide to grow there. Devices of different shapes were present on the mask, including comb-like structures with large edge-to-area ratios, but only the field oxide and flat square devices were studied in detail.

A total of six wafers were processed according to these steps, as listed in Table 3.4. Subsequent references to these samples will invoke the use of the wafer index from this table. Different tunnel oxide recipes were used as shown. The oxide thicknesses were obtained from ellipsometry on unpatterned substrate wafers which were oxidized at the same time as the epi wafers used to grow the devices. The "Stanford" measurements were performed using a single wavelength and angle, and the "Caltech"

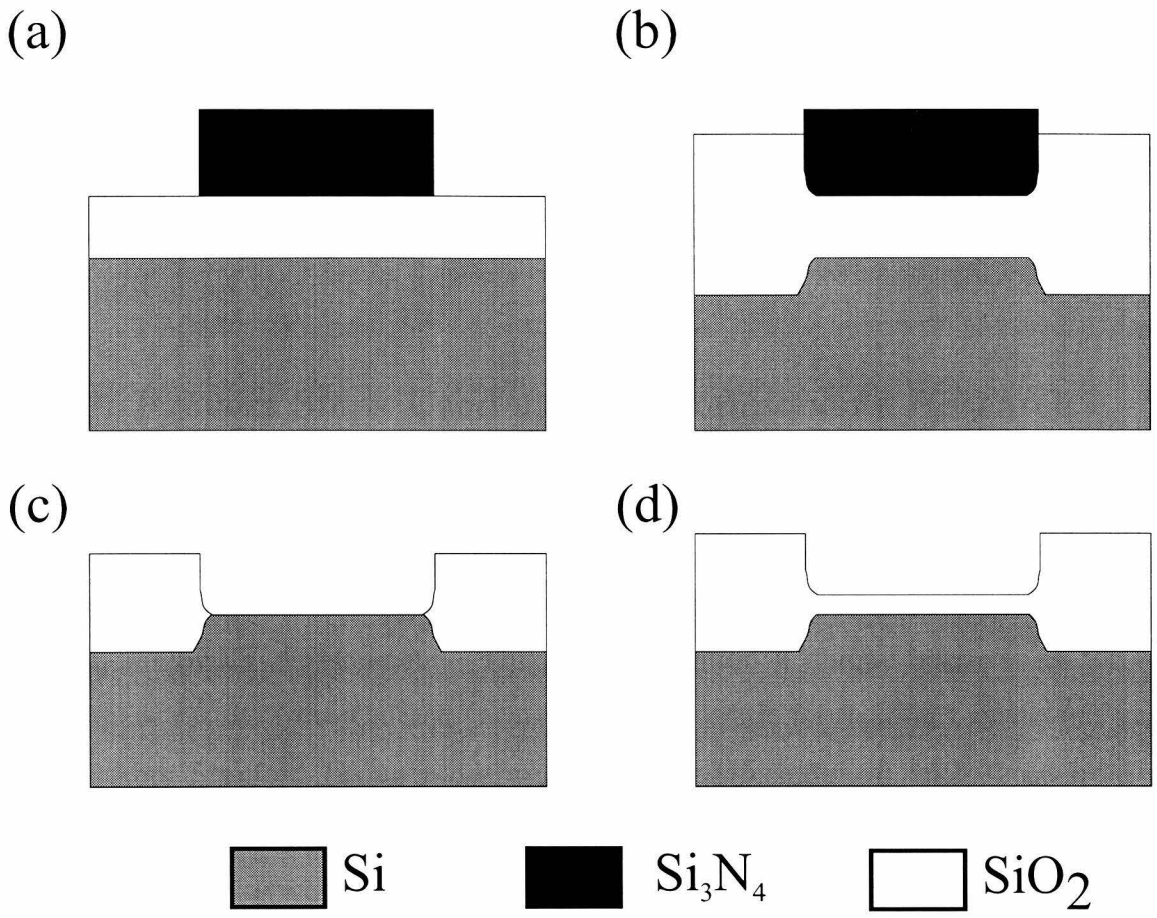

Figure 3.2: LOCOS procedure. (a) Silicon nitride is deposited and patterned on a "pad" oxide on silicon. (b) The silicon nitride acts as an oxygen barrier during further oxidation which produces the field oxide. (c) The silicon nitride is stripped off, and the oxide is etched until the pad oxide is completely removed in the windows. (d) A thin oxide layer is grown. The profile of the resulting oxide is commonly referred to as a "bird's beak."

Table 3.4: List of TSD wafers produced at Stanford.

Index	Epi Wafer	Mode	Oxide Recipe	Oxide Thickness (nm)	
				Stanford	Caltech
W1	S2	P-T	750°C, 30 min.	2.8	3.82
W2	S2	P-T	750°C, 15 min.	2.3	3.33
W3	S2	P-T	750°C, 15 min.	2.3	3.33
W4	H3	Av.	750°C, 15 min.	2.3	3.33
W5	S2	P-T	750°C, 5 min.	1.9	3.16
W6	H3	Av.	750°C, 5 min.	1.9	3.16

measurements were performed using variable angles and wavelengths one or two days later, after the wafers were shipped to Caltech. See Section 3.4.5 for more details on these measurements.

3.3.4 Process Steps Used at Intel Components Research

A few members of the technical staff at Intel Components Research became interested in this project as a means for studying thin oxide properties, and offered to produce a set of TSD devices for study using their extremely well-tuned process techniques used in MOSFET development. This offer presented a unique opportunity for us to compare results from Caltech (using somewhat crude fabrication techniques) and results from Stanford (using more well established techniques) with those from a world-leading fab. Unfortunately, some flexibility had to be sacrificed as the process steps were relatively rigid and in some cases proprietary, and a custom mask set could not be produced, but this seemed a small price to pay.

The process steps used were similar to those used at Stanford, consisting of cleaning and other preliminary steps, field oxide growth, tunnel oxide growth, and gate patterning. More detail is presented here.

After a simple cleaning step, a 13 nm thermal "pad" oxide was grown, and the wafers were laser scribed with identification codes, followed by another cleaning step. A stoichiometric Si_3N_4 layer was grown by Low Pressure CVD (LPCVD) at 780°C to a thickness of 140 nm. The nitride layer was patterned lithographically and dry (plasma) etched. Then, 700 nm of thermal field oxide was grown at 900°C. The Si_3N_4

was removed with hot phosphoric acid, and the wafers were dipped in dilute HF (50:1) to remove the 13 nm pad oxide. A thermal sacrificial oxide was grown and etched in the same HF bath to remove any residual Si_3N_4.

The wafers were then subjected to an RCA clean consisting of an SC-1 solution of H_2O_2, NH_4F, and H_2O, and an SC-2 solution of H_2O_2, HCl, and H_2O. A tunnel oxide of varying thickness was then grown using proprietary recipes, and a 300 nm polysilicon CVD layer was grown immediately afterwards.

The poly gate layer was doped n-type with an As implant and subsequent RTP anneal. Discrete poly pads were defined using standard lithography and a dry etch. A back contact was formed by scratching the back surface to penetrate any residual oxide, and sputtering roughly a micron of Al metal.

The mask set used was not designed specifically for this project, and therefore contained a number of structures which required further processing steps and could not be used. Large square field oxide devices were the only structures studied in detail, but other useful structures such as "combs" and other large edge-to-area ratio devices, and structures for measuring the poly gate sheet resistance were present.

A total of 48 wafers, three identical wafers of each type, were produced by Intel as listed in Table 3.5. The p-type wafers used for samples W1-W12 were (100) boron doped p^{++} substrates with 6 μm boron doped epi (12 Ω/sq). Oxide thickness targets were as specified in the table, but no direct measurements of these thicknesses were performed. Subsequent references to these processed wafers will be made by wafer number.

3.4 Measurements and Characterization

Nearly all electrical measurements were performed with the set of equipment depicted in Figure 3.3. An Alessi probe station was used for making electrical contacts to devices on processed chips or wafers. The samples were placed on the conducting sample stage, which provided the back contact, and tungsten probes were used to contact the top gates. Pads greater than 20 μm on a side could be easily contacted in

Table 3.5: List of TSD and MOS wafers produced at Intel.

Index	Wafer	Mode	Oxide Target Thick. (nm)	Index	Wafer	Mode	Oxide Target Thick. (nm)
W1-W3	p-type	MOS	4.5	25-27	I1	P-T	4.5
W4-W6	p-type	MOS	3.5	28-30	I1	P-T	3.5
W7-W9	p-type	MOS	2.5	31-33	I1	P-T	2.5
W10-W12	p-type	MOS	2.0	34-36	I1	P-T	2.0
W13-W15	I2	P-T	4.5	37-39	I3	Av.	4.5
W16-W18	I2	P-T	3.5	40-42	I3	Av.	3.5
W19-W21	I2	P-T	2.5	43-45	I3	Av.	2.5
W22-W24	I2	P-T	2.0	46-48	I3	Av.	2.0

this way. In a few cases, a piece of a chip or a wafer was cleaved off after processing, and mounted on an 8 pin header. The conductive surface of the header provided contact to the substrate. Contacts between the pins and the device pads or the back surface of the header could be made by wirebonding. As the wirebond pad footprint was roughly 50 μm$\times$100 μm, pads of roughly 100-150 μm on a side could be easily contacted. A device mounted in a header could be mounted in an HP test box which provided external connections to the desired pins.

The leads from either the probe station or the HP test box could be connected to a Tektronix curve tracer for a quick look at I-V behavior or for stressing, an HP 4192 LCR Impedance Analyzer for C-V measurements, or an HP 4145 Semiconductor Parameter Analyzer for I-V measurements. The latter two of these could upload data to a computer via a GPIB bus for storage and further analysis.

3.4.1 Simple I-V Measurements

Most I-V measurements were taken using the HP 4145 and the probe station, although the curve tracer was often used for quick evaluation of device performance without computer acquisition. A single tungsten probe was used to contact the gate pad, and the conducting sample chuck contacting the semiconductor substrate provided the second terminal. In most cases, the gate pad was grounded while positive bias was applied to the substrate, as this is the "interesting" bias direction for the TSD. The room lights were turned off for all measurements, as the ambient light injection was

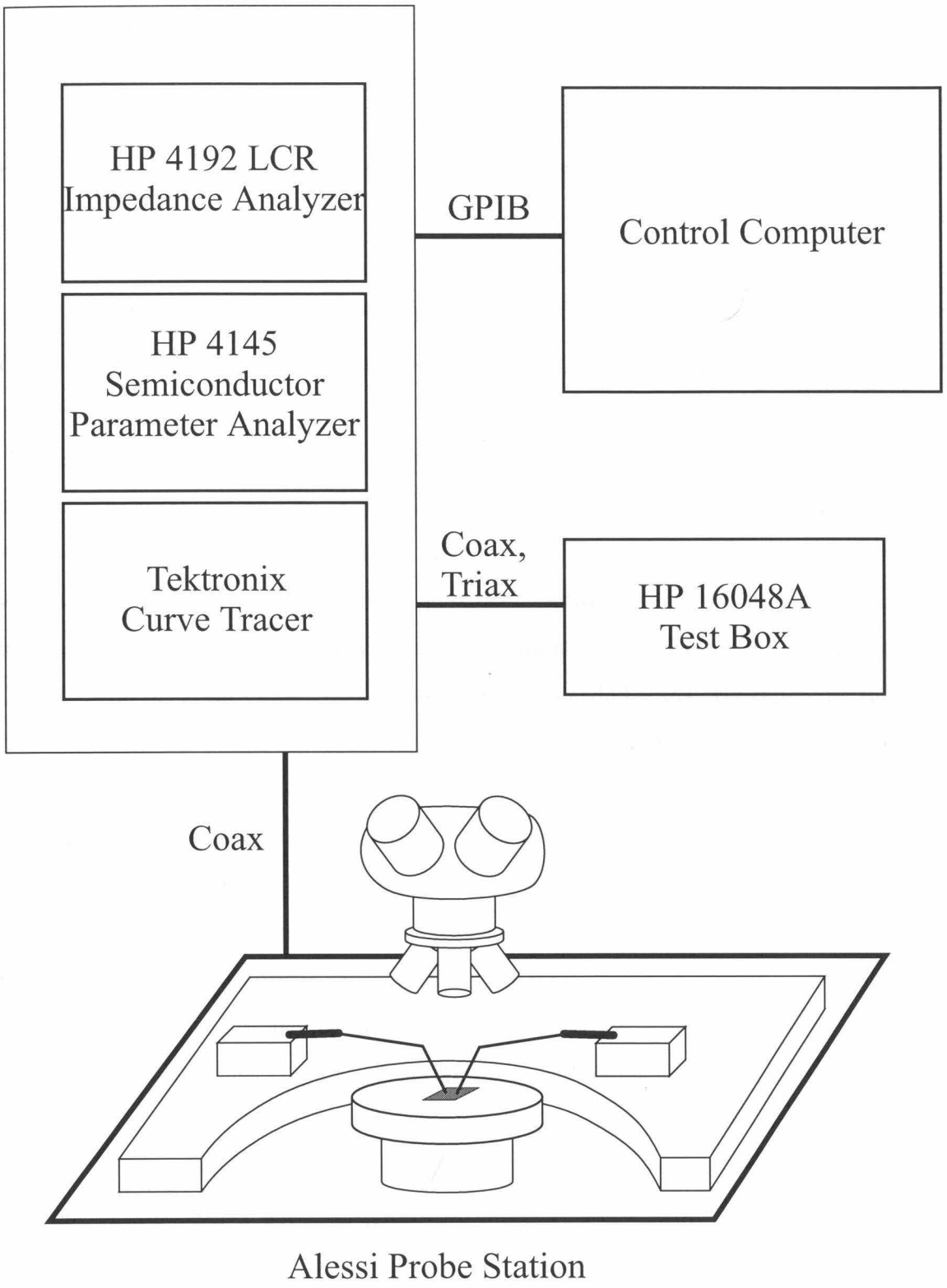

Figure 3.3: Experimental setup for TSD Electrical Measurements.

often enough to noticeably lower the TSD peak voltage (see Section 2.2.2).

The integration time of the HP 4145 was usually set to "long," corresponding to the most possible signal averaging during measurement. A setting of "medium" usually resulted in noisier TSD curves, but was often useful for MOS device measurements. For TSD I-V curves, the HP 4145 sourced current to the device and measured voltage so that no hysteresis needed to be detected. Given that the current source circuitry is active, attempting to feed back an appropriate voltage to match the setpoint current, the measured response in the NDR regions of the I-V curves is likely to be inaccurate. For some simple MOS devices, as no NDR behavior was expected, voltage was sourced and current measured.

Electrical contact to the probe tips is provided via banana plug connectors in the base of the probe station. The two banana plugs were combined into a single coax line which was connected to the HP 4145. The guarded triax connections on the HP 4145 were therefore not used. The lack of guard connections placed an upper limit on the measurable impedance of roughly 2×10^{10} Ω as illustrated in Figure 3.4, resulting from the parasitic resistance and capacitance of the unguarded leads. By replacing the coax line to the banana plug connectors with triax, this limit may have been improved, but the guard line would have to terminate at the probe station connection, leaving a large unguarded section of conductor leading to the probe and down the probe tip, possibly still providing a large parasitic capacitance. Unfortunately, this setup could not be tested as the necessary connectors have not been readily available since the time that this limitation was discovered. Most devices tested were of sufficiently low impedance that this limit was unimportant, but several measurements made on the Intel wafers would have benefited from an improved setup allowing higher impedance measurement.

3.4.2 C-V Measurements

All C-V measurements were performed using the HP 4192 connected to the HP device test box, containing a sample mounted on a header. For all measurements, the parallel

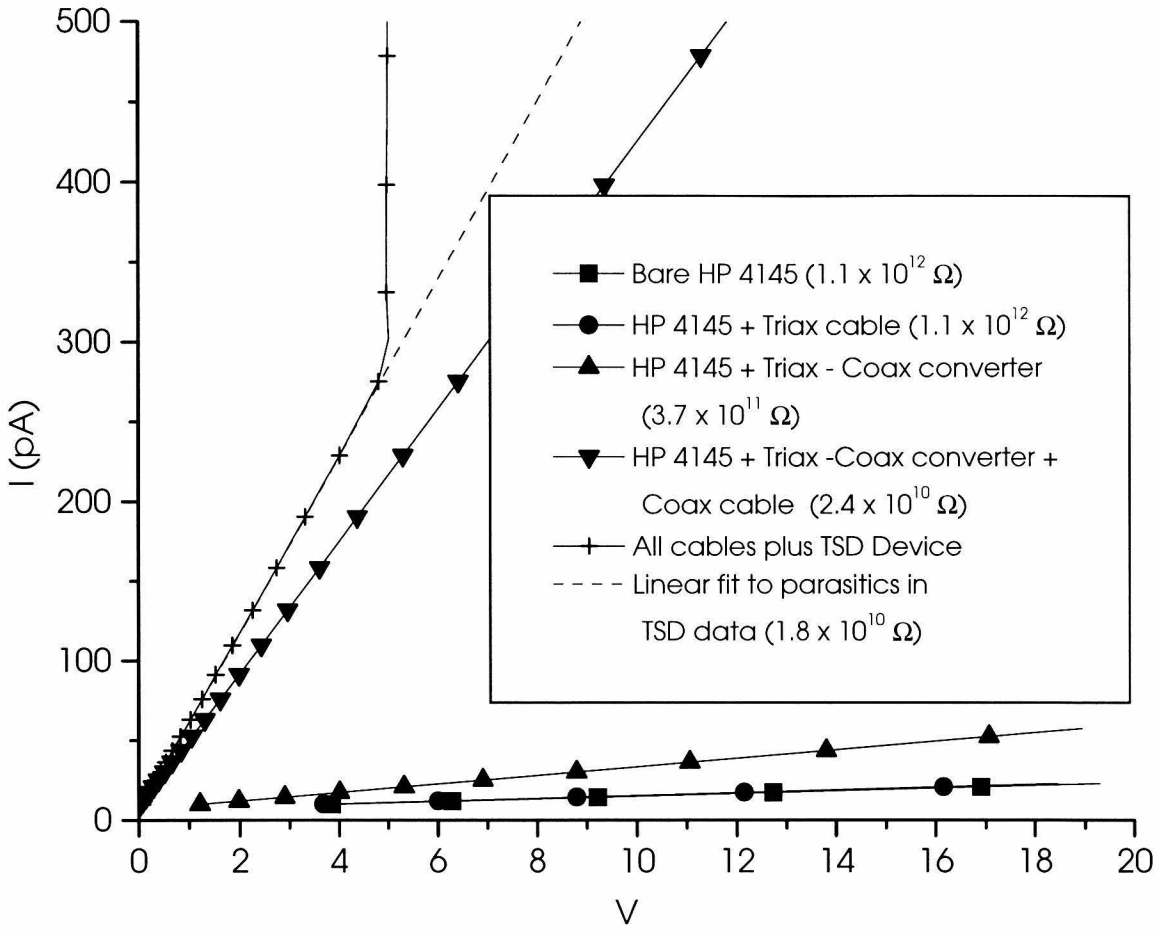

Figure 3.4: Parasitic parallel impedances in the electrical measurement setup. The *I-V* curves shown were taken with the indicated connection leads connected to the parameter analyzer. The leads were left open in all cases except for the TSD device curve. The triax cable connections do not degrade the 10^{12} Ω internal impedance of the HP 4145, but the triax-coax connector and the coax cable do appear to substantially limit the open circuit impedance.

effective circuit shown in Figure 2.10(c) was assumed, and the reported capacitance is that of the capacitor in this effective circuit. At each frequency used, prior to C-V measurement, the parasitic capacitance and conductance were subtracted out using the HP 4192 "short circuit zero" and "open circuit zero" functions. Attempts were made to perform C-V measurements with samples on the probe station, but the results were erratic, presumably due to the large parasitic capacitance relative to the test device capacitance.

This C-V technique was used primarily for determination of the oxide thicknesses of the Caltech MOS samples. Typical 1 MHz C-V curves of MOS samples with varying thicknesses (C25b, C26b, C27b) were illustrated in the previous chapter (Figure 2.10). The thickness estimate was calculated for the thickest oxide sample simply by taking the maximum capacitance of the accumulation part of the C-V curve, and calculating the thickness using the simple expression for the a parallel plate capacitor:

$$C = \frac{\epsilon_s A}{d} \tag{3.1}$$

where C is the capacitance, ϵ_s is the oxide dielectric constant (the bulk value of 3.9 ϵ_0 was used), A is the capacitor area, and d is the oxide thickness. For the C-V curve shown in the figure, $C_{max} = 310$ pF, $A = 150{\times}150$ μm^2, giving a calculated thickness of 2.5 nm. A variation of a few percent was observed across this sample.

While this method of choosing the maximum capacitance value is questionable, and has been shown to give underestimates of the oxide thickness in devices with polysilicon gates (resulting from a depletion layer in the poly) [11, 12], there are many other elements of this analysis that could produce errors of similar magnitudes (*i.e.*, the use of bulk SiO_2 dielectric constants and the assumption that the oxidation rate of the p-substrate used to make the MOS devices is the same as that of the p-n epi wafer). In light of the fact that the results of this calculation were in good agreement with other thin oxide work (see Section 3.3.2), the results of this simple analysis were determined to be an acceptable estimate.

3.4.3 Electrical Stressing

Electrical stressing of devices was carried out in three ways, two of which were unlike typical MOS stress procedures in which a device is held at a constant current and the voltage is monitored over time, or vice versa [13–16]. Therefore, it may be difficult or unreasonable to directly compare the stress results reported here with those found in MOS literature.

One stressing method involved connecting the device to the curve tracer, and sweeping the voltage at the line frequency (60 Hz) either from zero to some set voltage (unipolar), or from negative to positive voltage (bipolar) for some period of time. This method was motivated by TSD literature in which devices were switched from the low to high current states in order to determine whether the switching mechanism induced damage [17]. As it has been shown that the polarity of stressing is important in MOS device [16, 18], bipolar stressing may produce different results than unipolar stressing for MOS devices. For TSD devices, however, in reverse bias, the majority of the bias is dropped across the reverse biased p-$n$junction, so little current flows through the device, and little bias is dropped across the oxide. For these reasons, unipolar and bipolar stressing of TSD devices should yield similar results. A detailed study of this point was not performed, however. This stress method will be termed "AC constant current stressing," as the maximum current is held fixed (unless the device I-V characteristic changes significantly during stress).

Another stress method simply involved subjecting the device to the stress incurred during the course of an I-V measurement taken with the HP4145 as described above. A typical stress measurement would proceed as follows. A device which had never been subjected to electrical stress would be connected to the parameter analyzer, and an I-V curve would be scanned to some small maximum current. This process yielded the "virgin" scan. While this measurement proceeded, the stress induced by the measured currents may have modified the device. A subsequent scan to a higher maximum current would then reveal a modified I-V curve, while further stressing the device. Another scan to a still higher maximum current could then be taken, and so

on. We will call this method "incremental current stressing." This stress method was used extensively for several reasons. First of all, many devices changed appreciably through the course of a measurement, so this method provides a way to evaluate a global change in the *I-V* curve as a result of this stress. In addition, unlike the curve tracer method outlined above, data is accumulated during the stressing period, so any sudden changes are recorded. This method was also quite easy to implement. While this technique is also difficult to directly relate to constant current or constant voltage MOS stressing, there seems to be no reason that either stressing method should be preferred *a priori*.

Constant current/voltage stressing, similar to that used in MOS studies, was also used occasionally. This method involved connecting the device to the curve tracer, switching the curve tracer to DC mode, ramping the current and voltage to some constant value, and holding this value for some period of time. As the curve tracer attempts to hold a fixed voltage value across the device and an internal current-limiting resistor placed in series, any change in the device *I-V* response could cause both the voltage and current to change during the stressing time, so this method only approximates a true constant current/voltage stress measurement when little change is observed. This stress method will be referred to as "DC constant current stressing."

3.4.4 Multiprobe *I-V* Measurements

The single probe *I-V* measurement technique as described above works well as long as there is a negligible voltage drop across the probe-gate contact, and as long as there is a negligible voltage drop across the gate pad (*i.e.*, the gate is very nearly an equipotential). For the metal gate devices, both of these conditions hold except for the highest applied currents, and even then, the deviations are not great. The polysilicon gate devices produced by Intel, however, are far from meeting either of these conditions in the higher current ranges. The probe-gate contact resistance can be quite high due to the native oxide that grows on the polysilicon, leading to significant voltage drops across this junction. Also, the relatively high polysilicon sheet resistance (measured

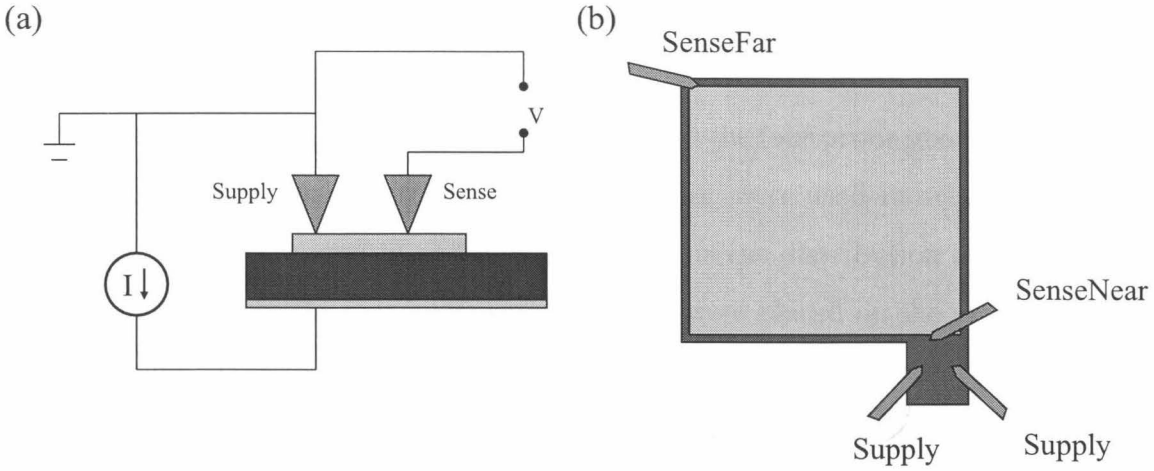

Figure 3.5: Multiple probe measurement setup. (a) Typical supply and sense measurement of a generic device. Both the supply and sense probes make contact with the device pad with the same contact resistance, but only a tiny amount of current flows through the sense probe due to the high voltmeter input impedance, so the voltage drop across the sense probe contact is negligible. (b) Four probe arrangement used in poly gate device measurement. The "Supply" probes supply the current, the "SenseNear" probe is used to subtract the effect of the contact resistance, and the "SenseFar" probe is used to determine the voltage drop across the sample.

at about 500 Ω/square), leads to a significant voltage drop across the gate for higher currents. To make matters worse, the Intel mask set only provided fairly large area devices ($\geq$85 μm on a side), accentuating this problem.

The effects of the probe-gate contact resistance can be completely removed by the use of a source probe and a sense probe as shown in Figure 3.5(a). The source probe supplies the current to the device as in the single probe measurement, so any contact resistance will lead to a voltage drop from the probe to the gate. The sense probe, however, is connected to a high impedance voltage measurement circuit, so very little current flows across the probe-gate contact. Therefore, very little voltage is dropped across the contact and an accurate measurement of the gate voltage can be made, as long as the contact resistance is significantly smaller than the voltmeter impedance.

Unfortunately, the effects of a voltage drop across the gate cannot be easily circumvented, but this voltage drop can at least be measured as shown in Figure 3.5(b). By placing sense probes at opposite sides of the device, with one sense probe close to the supply probe, it is possible to monitor the voltage drop across the device. A

large voltage drop across the gate pad will have important consequences as will be discussed shortly.

All measurements made on the Intel wafers used the four probe geometry shown in Figure 3.5(b). Two supply probes, shorted together, were used in order to minimize the contact resistance, and in order to make the current distribution at least slightly more uniform across the gate. The supply probes were placed on the area of the gate pad which lay on top of the field oxide. The sense probes were placed near the corners of the windows of the field oxide, lying at least partially on the part of the gate which extended over the tunnel oxide.

A typical I-V curve taken on an 85×85 μm^2 Intel TSD device using this four probe method is shown in Figure 3.6. Notice that the sense curves start to separate from the supply curve at roughly 0.1 mA due to the probe contact resistance. Then, the sense curves begin to split at a slightly higher current, indicating a voltage gradient across the sample.

In order to reduce the contact resistance of each probe contact, current was forced through the pair of supply probes and subsequently through the pair of sense probes in order to break down any oxide layer present on top of the polysilicon gates without stressing the underlying device. Typically, this required a stress of several volts across each pair of probes. Although no current was forced through the device during this process, the large current flowing through the gate in close proximity to the thin oxide layer during the sense probe stress may have caused some oxide damage.

As mentioned above, if the gate pad cannot be held at an equipotential, it is not easy to compensate for the change in I-V behavior. Here, we will examine in detail what influence a gate voltage drop can have, and what analysis can be used to explain this behavior and perhaps recover useful information.

In order to describe a device in which the gate cannot be considered to be at an equipotential, we can use a resistive network model, as depicted in Figure 3.7. The gate is modeled as a mesh of identical resistors, with the active device now split into a number of identical microscopic devices, each connected to a point on the gate and to a substrate equipotential. By applying Kirchoff's laws to each of the components, and

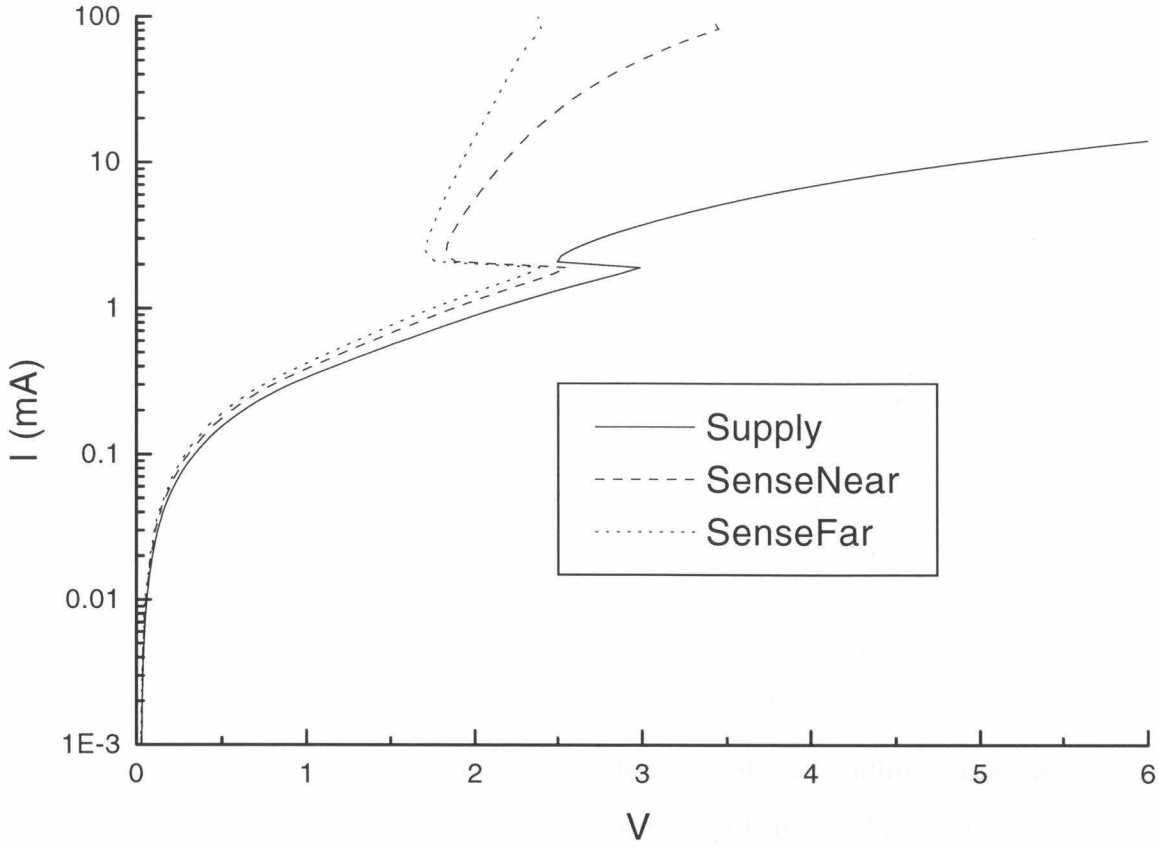

Figure 3.6: Typical measured multi-probe I-V curve for an 85×85 μm^2 device from Intel W47 (avalanche mode, 2.0 nm oxide). The three curves correspond to the three contacts described in Figure 3.5(b).

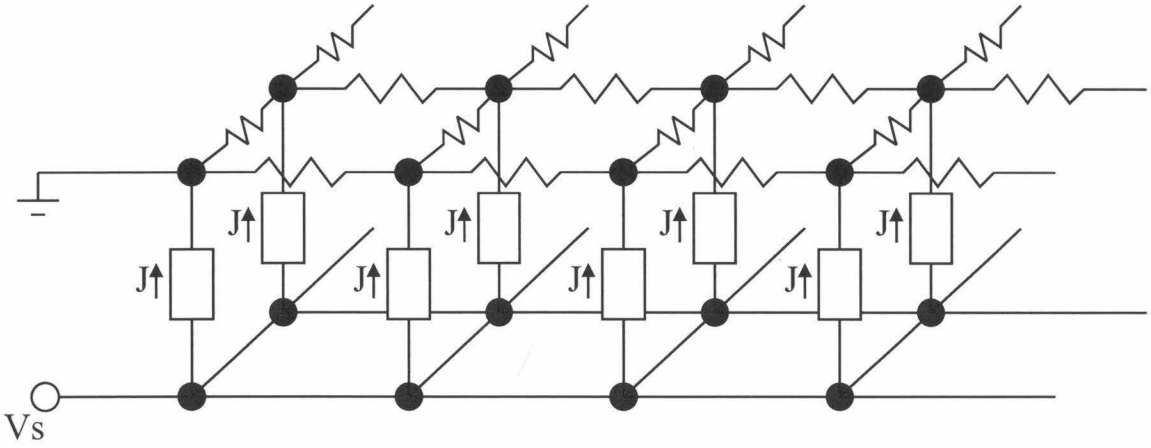

Figure 3.7: Resistive network effective circuit for analysis of gate voltage drop. The gate is modeled as a resistive mesh, and the device is split into a grid of microscopic devices connected to the mesh.

by taking the limit of an infinite number of infinitesimal mesh elements, the following equations result (see Appendix A):

$$J(x,y) = \frac{t_g}{\rho_g}\nabla^2 V(J(x,y)) \tag{3.2}$$

where $J(x,y)$ is the current density flowing through the device as a function of position across the gate, t_g, ρ_g are the thickness and resistivity of the gate (so t_g/ρ_g is the sheet resistance), and $V(J)$ is the voltage drop across each infinitesimal device element for a given current density, J.

Consider the special case in which the device elements are simply resistive. Then, $V(J)$ is proportional to J, and Equation 3.2 can be solved, giving

$$J(x,y) \sim \exp(-(x+y)/r), \qquad r = \sqrt{\frac{\rho_{dev} L_{dev}}{\rho_g/t_g}} \tag{3.3}$$

where $\rho_{dev} L_{dev}$ is the resistivity times the "length" of each infinitesimal device (or equivalently, the reciprocal of the conductance per unit area). Therefore, the current through the device drops off exponentially with a characteristic length, r. The gate voltage, $V_g(x,y)$, will be proportional to $J(x,y)$, so it too will drop off exponentially. As the gate gets more resistive, or the device more conductive, the characteristic length gets shorter and the voltage drops off more rapidly.

In order to compare this model to the true device behavior, the voltage drop across a $500\times500\mu\text{m}^2$ avalanche TSD device (Intel wafer W47) was measured by taking I-V curves with the "SenseFar" probe placed at various points across the pad. The I-V curves with the "SenseFar" probe placed at the far edge of the device as usual are plotted in Figure 3.8(a), and the gate voltage as a function of position is plotted for several different points on the TSD I-V curve. The bias voltages listed in the legend of Figure 3.8(b) are the measured "SenseNear" voltages. When the TSD is in the low current state, the gate voltage is roughly constant across the pad, but in the high current state, the gate voltage drops off sharply, as expected from the model. Consider, for example, the device biased at 4.5 V in the high current state. The gate voltage

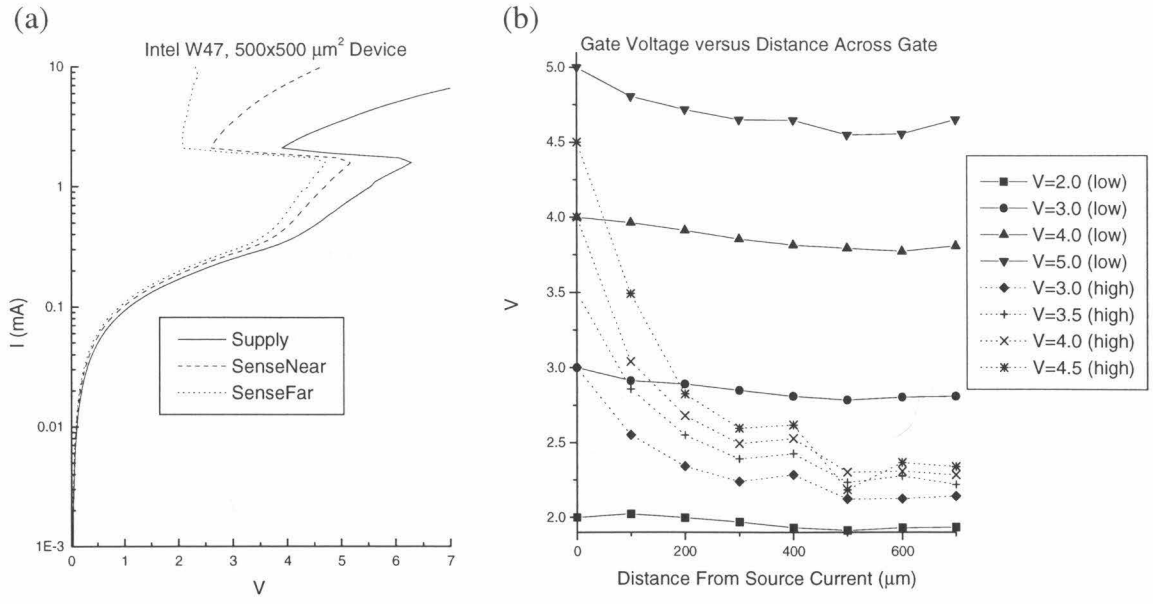

Figure 3.8: (a) The measured "Supply," "SenseNear," and "SenseFar" I-V curves for a $500\times500\mu m^2$ avalanche device (Intel W47). (b) The gate voltage for the device in (a) plotted versus diagonal distance from the SenseNear probe for a number of different applied (SenseNear) voltages, both in the low and high current states of the device.

drops rapidly across 300 μm or so, and then reaches the valley voltage. Further drop in the gate voltage forces the device back down into the low current mode, so the conductivity drops, and the gate voltage remains roughly constant across the rest of the gate.

The above analysis is too simple to be used to quantitatively fit the data in Figure 3.8(b) for several reasons. First of all, the device I-V response is far from resistive - the effective conductance changes with applied bias, causing the characteristic decay length to be a function of bias. In addition, any interaction between microscopic device elements is ignored, whereas in reality, charge can flow between different areas of the device in the top silicon layer, particularly when it is heavily populated in the high current state. Nonetheless, it is easy to see that if the currents are too large, the current transport through the poly gate devices will be far from uniform as the effective active area of the device is reduced.

Any I-V curves of poly gate devices shown subsequently in this thesis will reflect the voltage sensed by the "SenseNear" probe as this is perhaps the best indicator of

the "true" voltage across the device. For large gate poly devices, however, it must be considered that the effective area of the device may be substantially less than the size of the gate.

3.4.5 Ellipsometry

Ellipsometry was used to perform oxide thickness measurements on several samples. Two different instruments were used: A Gaertner single wavelength (He Ne), discrete angle ellipsometer at Stanford, and a J. A. Wollam Variable Angle Spectroscopic Ellipsometer (VASE) at Caltech. Thickness estimates of the oxide were obtained in both cases by fitting the measured ψ and Δ values to those calculated from a model consisting of a simple layer of Silicon and a simple layer of oxide. Bulk optical constants were used for both layers [19].

The VASE instrument was used for most of the measurements. Data was taken over a range of wavelengths (200-715 nm), and often over a range of angles (70°, 75°, 80°). This range of parameters was found to give quite repeatable and reproducible results. The fit of the data to the theory was characterized by the mean standard error (MSE), typically in the range from 7.0-9.0, corresponding to a standard deviation of about 1%. The Gaertner ellipsometer was used for immediate measurements of oxide thicknesses during processing at Stanford, but did not produce as repeatable or reproducible results as the VASE system. This does not necessarily mean that the VASE results are closer to the true thicknesses (better accuracy), but small differences in thickness seemed to be better represented by the VASE data (better precision).

Bibliography

[1] P. O. Pettersson, *Silicon Heterojunctions*, PhD thesis, Caltech, 1996.

[2] P. O. Pettersson et al., (submitted for publication to IEEE Tr. Elec. Dev.).

[3] Q.-Y. Tong et al., *J. Electrochecm. Soc.* **142**, L201 (1995).

[4] W. R. Runyan and K. E. Bean, *Semiconductor Integrated Circuit Processing Technology*, Addison-Wesley, Reading, MA, 1990.

[5] H. Fukuda, M. Yasuda, T. Iwabuchi, and S. Ohno, *IEEE Electron Device Lett.* **12**, 587 (1991).

[6] M. L. Green et al., *Appl. Phys. Lett.* **65**, 848 (1994).

[7] A. E. T. Kuiper, H. G. Pomp, and P. M. Asveld, *Appl. Phys. Lett.* **61**, 1031 (1992).

[8] A. M. Hodge, C. Pickering, A. J. Pidduck, and R. W. Hardeman, *Mat. Res. Soc. Symp. Proc.* **52**, 313 (1986).

[9] S. E. Lassig, T. J. Debolske, and J. L. Crowley, *Mat. Res. Soc. Symp. Proc.* **92**, 103 (1987).

[10] S. Duzelier, G. Sarrabayrouse, J. L. Prom, and G. Hollinger, *Electronics Letters* **27**, 1399 (1991).

[11] D. A. Buchanan and S.-H. Lo, *Proc. Electrochem. Soc.* **96-1**, 3 (1996).

[12] M. Depas, B. Vermeire, P. W. Mertens, R. L. Van Meirhaeghe, and M. M. Heyns, *Solid–State Electron.* **38**, 1465 (1995).

[13] R. Rofan and C. Hu, *IEEE Electron Device Lett.* **12**, 632 (1991).

[14] K. F. Schuegraf and C. Hu, *J. Appl. Phys.* **76**, 3695 (1994).

[15] M. Depas, R. L. Van Meirhaeghe, W. H. Laflere, and F. Cardon, *Solid–State Electron.* **37**, 433 (1994).

[16] M. Depas, T. Nigam, and M. M. Heyns, *IEEE Trans. Electron Devices* **43**, 1499 (1996).

[17] T. Yamamoto, K. Kawamura, and H. Shimizu, *Solid–State Electron.* **19**, 701 (1976).

[18] W. Weber, M. Brox, R. Thewes, and N. S. Saks, *IEEE Trans. Electron Devices* **42**, 1473 (1995).

[19] E. D. Palik, *Handbook of Optical Constants in Solids*, Academic Press, Orlando, FL, 1985.

Chapter 4 Experimental Results

4.1 Introduction and Outline

The procedures outlined in the previous chapter produced working TSD devices quite readily, and did not require (or in some cases allow) a great deal of fine tuning. Nearly all devices produced "worked" in the sense that they exhibited both high and low current states. Even the "failed" field oxide processes employed at Caltech produced working devices by this definition. However, the details of the device behavior seemed to depend on a large matrix of parameters involving the fabrication and measurement techniques. In this chapter, we will discuss the many factors that were found to influence the device behavior. As many of these were discovered through the course of performing measurements, the sets of experiments performed on earlier sets of samples (*i.e.*, the Caltech chips) do not coincide exactly with those performed on samples obtained later on, but attempts will be made to draw comparisons where applicable.

In Section 4.2, the effects of electrical stress on the various types of TSD devices will be discussed. Then a discussion of how device currents scale with device area will be presented in Section 4.3. Given the results presented in these two sections, the dependence of device behavior on design parameters, such as oxide thickness, and silicon epilayer doping and thickness, will be compared to the expected behavior in Section 4.4. Finally, a summary of results and discussion of several issues follows in Section 4.5.

4.2 Effects of Stress on TSD Devices

The *I-V* characteristics of all devices tested were heavily dependent on stress history. Even the act of taking an *I-V* measurement often stressed the devices enough to

make a noticeable change in subsequent measurements. Thus, before any comparison of devices with different fabrication parameters or different geometries can be made, it is necessary to explore the influence of stress in order to determine how each device should be stressed in order to make useful comparisons.

4.2.1 Constant and Incremental Current Stressing

For nearly all devices tested, it was found that significant immediate changes in the I-V characteristic occurred if the device were subjected to a larger current than it had ever encountered before. In contrast, devices which were subjected to a fixed maximum current did not appear to change substantially over stressing periods of many minutes, or perhaps an hour. Examples of these two different stressing methods, described in Section 3.4.3 as "incremental current stressing" and "constant current stressing" respectively, are shown in Figures 4.1, 4.2.

In Figure 4.1(a), we see that the virgin scan (taken before the device had been subjected to any stress) differs significantly from the subsequent scan, reflecting a relatively large, sudden stress-induced change in the device during this scan. Note, however, that the device I-V characteristic does not make as large a change through the course of an additional 12 minutes of constant current stress. Only a total of 900 minutes of stress, corresponding to a total integrated charge density fluence of 4×10^6 C/cm^2, produces a similar magnitude of change in the I-V curve.

The set of scans in Figure 4.1(b) represent a sequence of I-V measurements of a single device with no other stressing in between scans, *i.e.*, incremental current stress. Each scan in this series of stress events is much like the one imposed on the device in Figure 4.1(a) during the virgin scan. While the lower current stress events do not induce much change, once the maximum current reaches ~ 1 mA, each subsequent scan causes a significant change in the I-V curve until finally, the device no longer exhibits any NDR. Notice that the I-V curves are relatively smooth up to the maximum current of the previous scan, but then become jagged above this current. This suggests that the majority of the stress modifications are likely taking place during the portion of

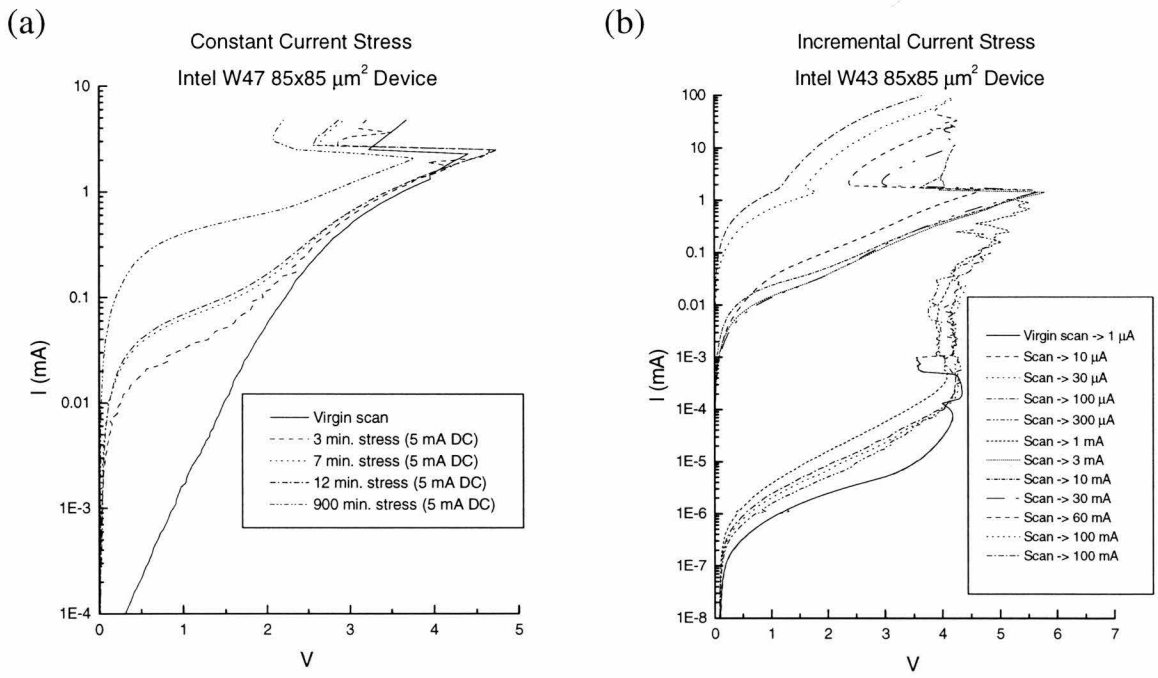

Figure 4.1: Constant current and incremental current stressing of poly gate TSD devices. (a) Avalanche device with 2.0 nm oxide subjected to a DC 5 mA stress current over time. (b) Avalanche device with 2.5 nm oxide subjected to a series of scans to an increasing maximum current.

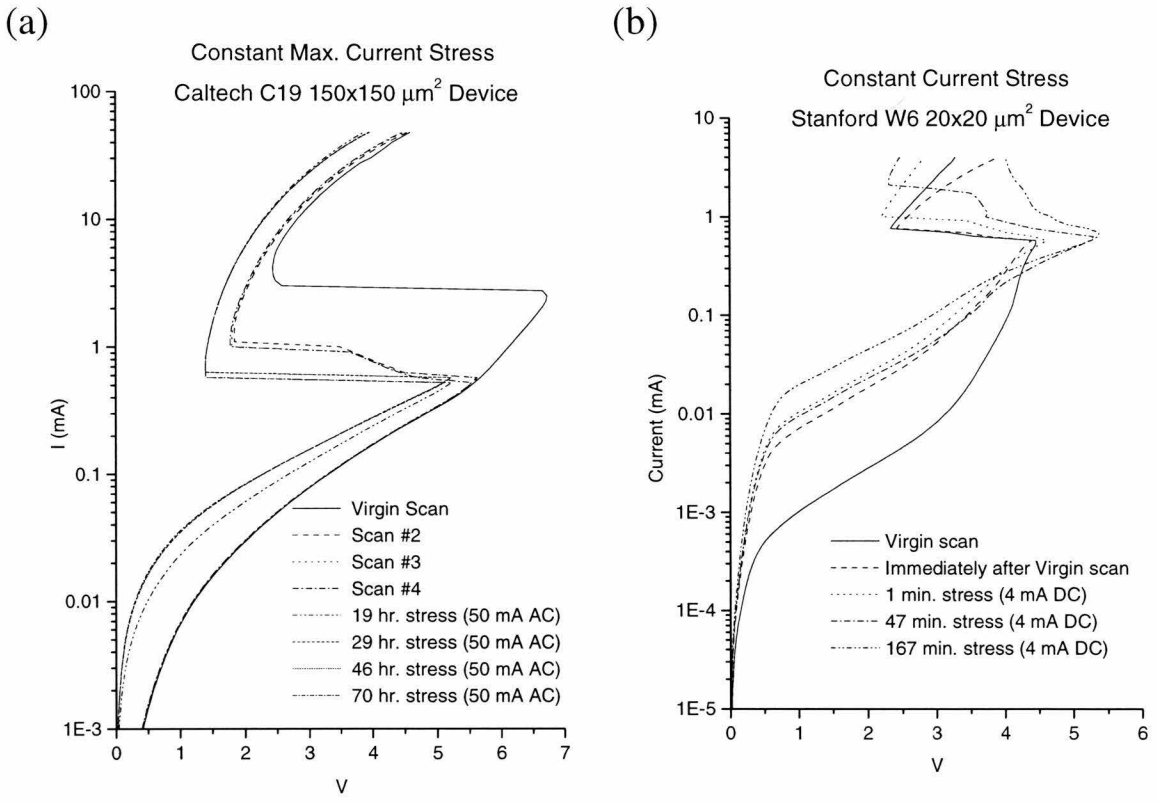

Figure 4.2: Constant current stressing of Al gate TSD devices. (a) Caltech punch-through device with 2.5 nm oxide subjected to a series of four scans with no stressing in between, followed by several scans separated by an AC (50 mA max) constant current stress. The total stress time preceding each scan is indicated. (b) Stanford avalanche device with 1.9 nm oxide subjected to a DC 4 mA constant current stress over time.

the scan above the previous maximum value. That is, given that the device has been stressed to some maximum current, if it is subjected to further current flow below this value, little device modification occurs, but if it is stressed above this maximum current, significant stress-induced changes occur immediately.

The Al gate devices produced both at Caltech and Stanford exhibit similar behavior. Figure 4.2 illustrates the same behavior under constant current stressing as for the poly gate device (Figure 4.1(a)) – small changes in device behavior over time after a large initial change during the virgin scan. While incremental stressing was not performed on the Al gate structures in the same way as for the poly gate structures, a similar effect could be observed by watching the I-V curve swept out by the curve tracer. As the maximum current was increased on the curve tracer, the entire I-V curve would move up in current in discrete jumps, stabilizing when the maximum current was again held steady.

4.2.2 Voltage Limiting Behavior

It is interesting to note that, in Figure 4.1(b), the voltage in the high current state of the device seems to be limited to roughly 4 V. As mentioned in Section 2.2.2, the bulk of the bias across the device in the high current state is dropped across the oxide, suggesting this as a limit on the total bias allowed across the oxide. Incremental stressing seems to make whatever modifications are necessary to the device (presumably in the oxide layer) in order to drop the voltage at the maximum stress current to this maximum value. The oxide appears to be sustaining some kind of "soft breakdown."

This behavior was observed in all poly gate devices on which incremental current stressing was performed, including the simple MOS devices. A comparison of this effect in devices with a range of oxide thicknesses (Figure 4.3) shows that the I-V curves in this voltage limiting stressing region are more jagged for the thicker oxides than for the thinner ones. This suggests that the "soft breakdown" process is "softer" for the thinner oxides: as the oxide thickness is increased, it begins to look much more like a series of sharp increases in conductivity, very much like one would expect from

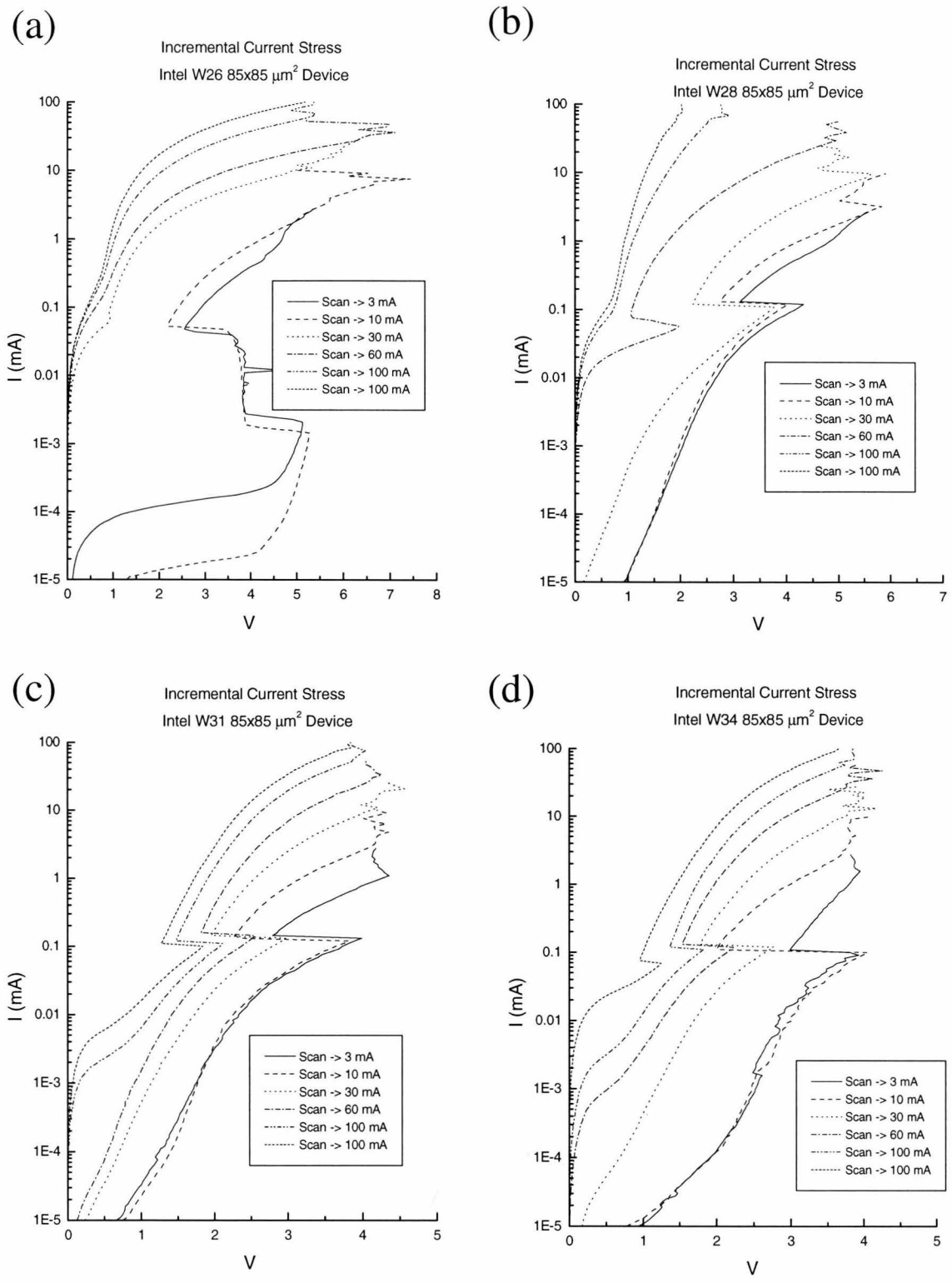

Figure 4.3: Voltage limiting behavior evident during incremental current stressing of (a) 4.5 nm, (b) 3.5 nm, (c) 2.5 nm, and (d) 2.0 nm punch-through mode devices.

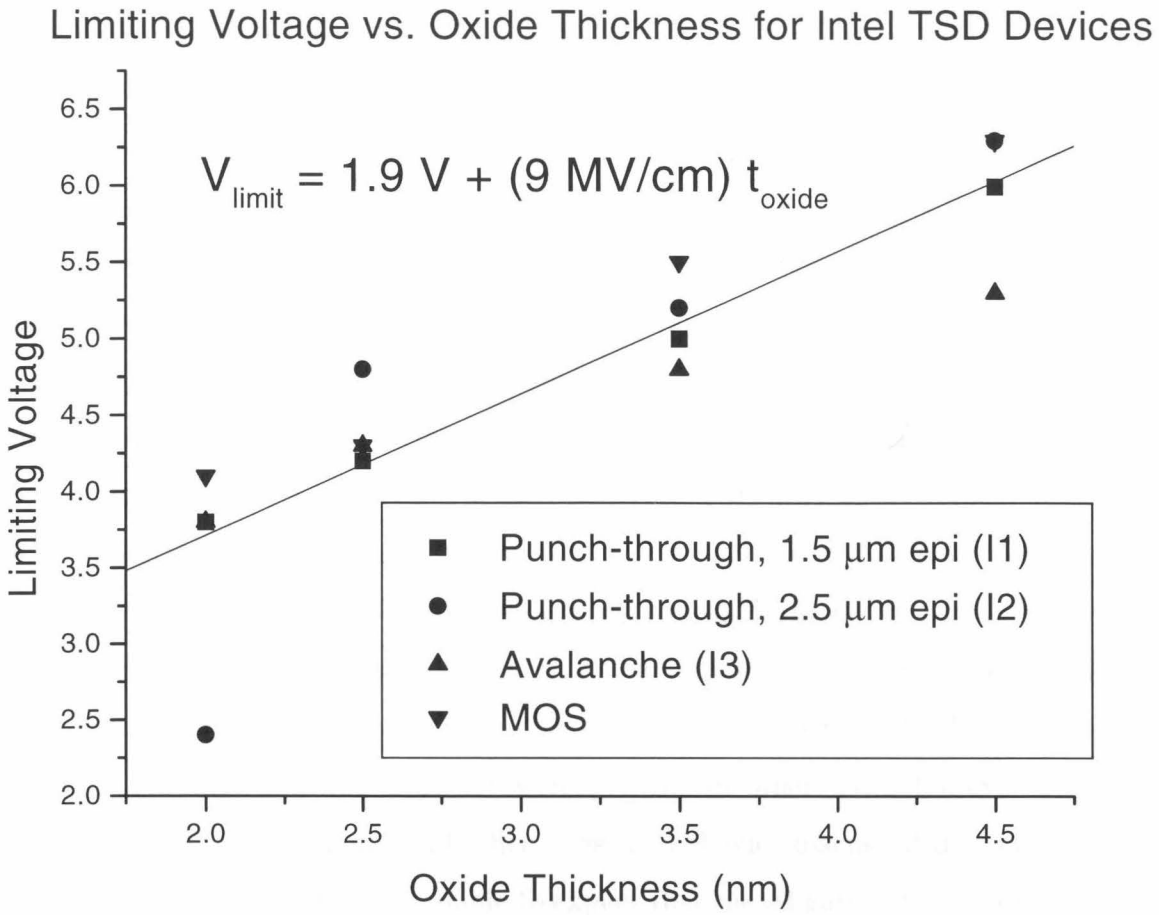

Figure 4.4: The limiting voltage for representative devices from the full array of poly gate wafers is plotted as a function of oxide thickness. A linear fit to all data is also shown.

traditional dielectric breakdown.

In addition, the limit voltage increases with increasing oxide thickness, suggesting a field-limiting process. In Figure 4.4 we plot the limiting voltage versus oxide thickness for poly gate devices of each oxide thickness and each type (MOS, avalanche TSD, and the two punch-through TSD). The limiting voltage was estimated by averaging the high and low voltage values in the high current, stress-inducing part of a set of incremental current stress curves (as in Figure 4.3) for an 85×85 μm^2 device on a wafer of each type. A linear fit to all of the data is also shown in the figure.

One possible interpretation of this data is obtained by splitting the voltage across the device into two parts: the voltage across the oxide (V_{ox}), and the voltage dropped across the other parts of the device (V_{other}). A linear fit to the points for each device

type would then have an intercept equal to V_{other}, and a slope equal to E_{limit} (the limiting electric field), assuming these parameters are thickness independent. This analysis yields roughly the same values of V_{other} ~2 V and E_{limit} ~9 MV/cm for each device type. Although an offset voltage of this magnitude might be possible for the TSD in which voltage could be dropped across the p-n junction or the n-layer, one would expect the offset voltage for the MOS device to be much smaller than this, as nearly all applied bias should fall directly across the oxide. Therefore, a more likely interpretation of this data might be that the critical field increases with decreasing oxide thickness, and that the bulk of the bias in the TSD devices is dropped across the oxide. Limited confidence can be placed in any conclusion drawn from such a small sample set of data obtained from such a subjective analysis, however.

Unfortunately, no incremental current stress measurements were taken on the Al gate devices, so it is unknown whether these devices also exhibit this behavior. Some hints of a voltage limiting behavior were present in other data, however (see Figures 4.20, 4.6). In addition, as all other stressing behavior discussed thus far has been similar for all devices, it is reasonable to expect that the Al gate devices would exhibit the voltage limiting behavior as well.

4.2.3 Dual Mode Behavior of Poly Gate Devices

Consider the incremental current stress curve shown in Figure 4.5. The three lowest stress scans sweep out a very low current bistable TSD I-V curve, and are very nearly identical except for a slight increase in the switching current with stress. After a maximum current stress of 0.1 mA (corresponding to a voltage of ~5.4 V), however, the I-V curve begins to change significantly. During the subsequent sweep, the voltage drops suddenly in what appears to be some kind of breakdown, but as the sweep continues, a higher current TSD I-V curve results, and remains (albeit modified) in subsequent scans. Thus, a "low current TSD mode" is observed when the device has been subjected to little or no stress, and a "high current TSD mode" abruptly takes over once some critical stress has been reached. Both "modes" exhibit bistable

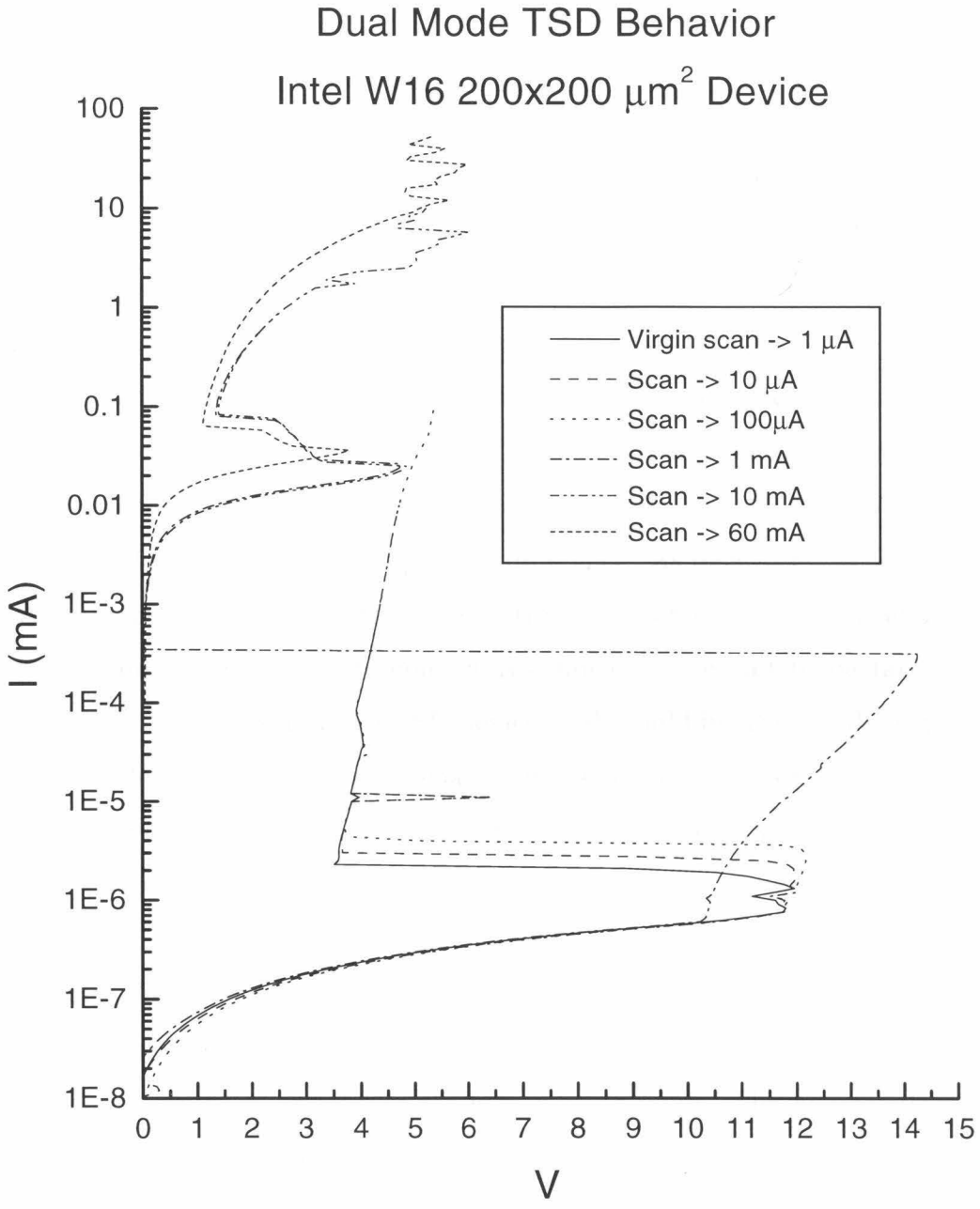

Figure 4.5: Incremental stress curves of a punch-through device with 3.5 nm oxide. The device exhibits two distinct "modes" of operation, each of which can support low and high current states characteristic of the TSD. A "low current mode" exists until the device has been stressed enough to force it into the "high current mode."

current states for a range of voltage. Note that the low current part of the "low current mode" I-V curve is limited by the parallel impedance of the measurement cables, so the impedance of the TSD device is likely much higher (see Section 3.4.1).

This dual mode behavior was observed in nearly all of the poly gate punch-through mode devices, and occasionally in the poly gate avalanche mode devices. It was not ever observed, however, in any of the Caltech or Stanford Al gate devices. As this behavior was discovered after all Al gate device measurements had already been performed, it is possible that it was not seen because we did not know to look for it, but many virgin scans of Al devices were taken (as in Figure 4.2), and none revealed anything like this dual mode operation. Similarly, all poly gate avalanche devices were also measured before this dual mode mechanism was recognized as such, and many incremental current stress scans of these devices were taken, so these same arguments could apply, with one important difference. As mentioned in Section 3.4.4, the supply and sense probe contacts were stressed prior to device measurements in order to diminish the probe-gate contact resistance. Care had to be taken not to overstress the contacts, or else the TSD device oxide could incur some damage during the process. Before the dual mode behavior was discovered, this precaution was not carefully observed. Therefore, it is quite possible that the dual mode behavior was usually not observed in the avalanche mode poly gate devices due to oxide damage induced during stressing of the probe contacts.

Note that there are several major differences between these two TSD modes. For example, the low current mode does not change appreciably with incremental current stress until the device is forced into the high current mode. The high current mode, however, changes substantially with each incremental scan. Perhaps more striking is the shape of the high current state in each of the two modes. The I-V curve in the high current state of the low current mode is exponential in shape (linear on the log-linear plot), and extremely steep, spanning nearly five orders of magnitude in current over a range of roughly two volts. The high current state of the high current mode, however, does not have an exponential shape, and increases only two orders of magnitude over a range of a few volts. It is possible the transport through the oxide in this part

of the curve is some sort of "trap-assisted" tunneling process, brought about due to stress in the oxide [1]. Still more differences between these modes will be discussed in subsequent sections. We believe that the discovery of these two modes of operation may be considered one of the most important results to come out of this body of work.

4.3 Scaling Behavior of TSD Devices

4.3.1 Introduction and Early Results

As the fabrication of TSD devices at Caltech was primarily inspired by the transistorless SRAM application, careful attention was given to the device parameters relevant to this application. In particular, increasing the ratio of the "read current" to the "quiescent current" (see Section 2.3.1) was considered one of the primary goals. Due to maskset limitations and problems with the field oxide process, the earliest TSD devices were fairly large area devices of a single size (150×150 μm^2). The high current state impedance could, therefore, be low enough that it could be overwhelmed by a relatively small parasitic impedance, and the measured current ratio would be lower than expected. By decreasing the device size, the device impedance would increase, eventually becoming greater than any parasitic impedance and thus allowing the current ratio intrinsic to the device to be measured. A new maskset with a range of smaller device sizes was produced in order to test this hypothesis.

The I-V curves for a $150\times150\mu$m^2 device from chip C19 and a $50\times50\mu$m^2 device from chip C29 are shown in Figure 4.6. The facts that the I-V curves are of roughly the same magnitude at the highest currents and scale roughly with area at the valley current are consistent with the idea of a parasitic impedance limitation. In fact, this trend allowed for an order of magnitude increase in the current ratio of interest, suggesting that further improvements might be made simply by reducing device size! However, note that the ratios of the peak currents and low state currents are not the same as the area ratio. The peak current, for example, is more than two orders of magnitude lower in the smaller device. As the two devices considered were

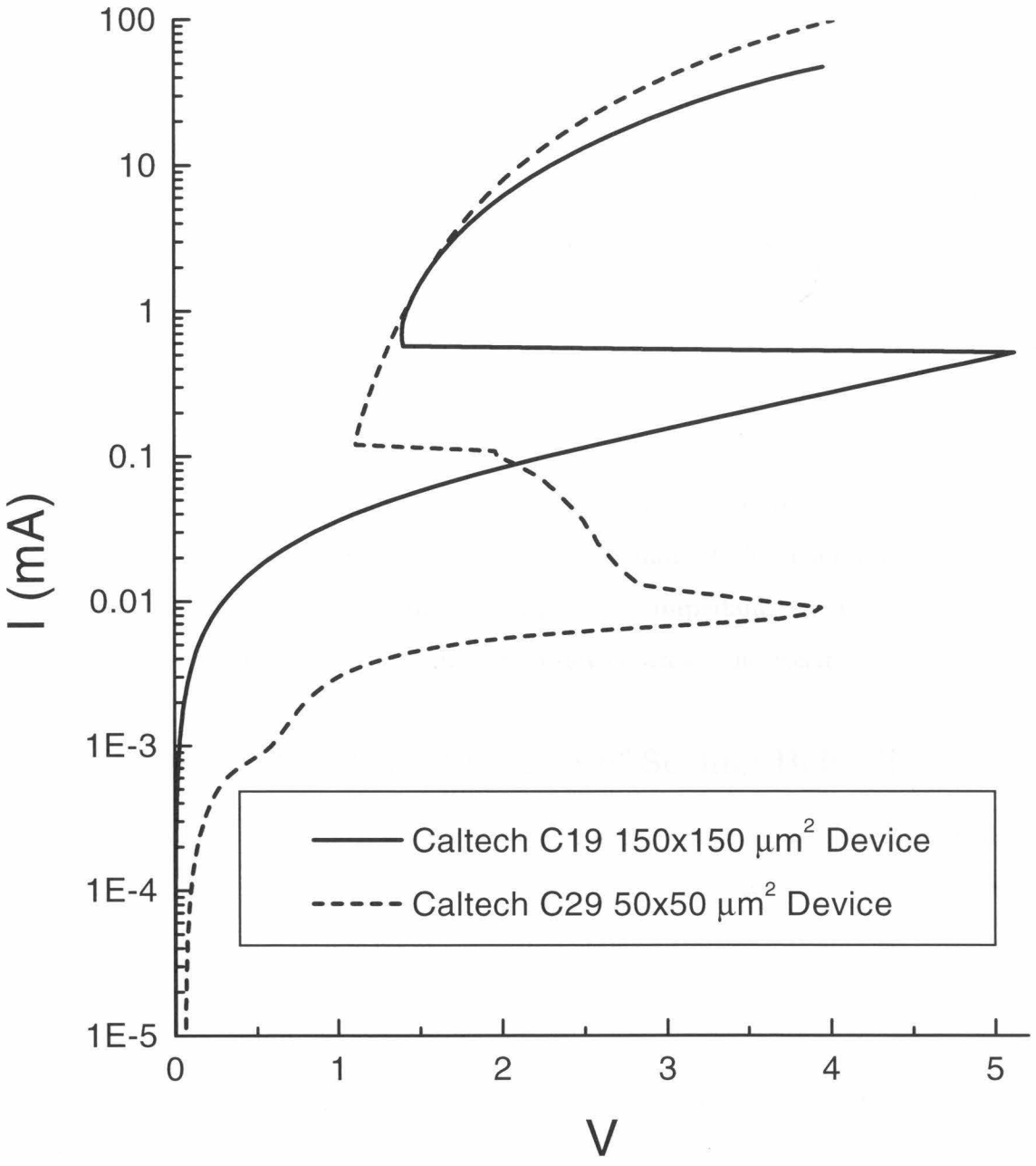

Figure 4.6: Early scaling results. A comparison of two different size devices on different chips, but with the same fabrication recipe, shows that a reduction of the device area by a factor of 9 results in a reduction of the valley current by roughly this factor, a reduction in the low state current of more than this factor, and no reduction (actually, an increase) in the highest currents.

fabricated separately, it is possible that these discrepancies are a result of slightly different processing details even though the recipes employed were the same. If such dramatic changes in the *I-V* shape are possible with minor, unintentional differences in fabrication, however, the comparison may be invalid.

One serious difficulty with the comparison of the two devices in Figure 4.6 is that both devices have been stressed significantly. As discussed in Section 4.2 the *I-V* response of any TSD device is strongly dependent on the maximum current which has been forced through the device. As is evident by examining the endpoints of the scans in Figure 4.6, the smaller area device had actually been stressed to a higher current than the larger area device. In fact, the voltage at the endpoints of both curves is about 4 V – roughly the same as the limit voltage for the 2.5 nm oxide devices described in Section 4.2.2. This provides an alternate explanation for the fact that the high current part of both curves are of the same order of magnitude. Therefore, before any conclusions can be drawn about the role of parasitic impedance, a closer look at role of stress in the scaling of the *I-V* curve with device area is necessary.

4.3.2 Stress-Induced Modification of Scaling Behavior

In studying scaling behavior, it would be ideal if it were possible to eliminate all stress effects. Unfortunately, as discussed in Section 4.2 and illustrated in Figure 4.2(a), for all Al gate TSD devices produced at Caltech, if enough voltage was applied to force the device into the high current state, the stress induced a significant change in the device *I-V* characteristic. Therefore, it appears that one cannot measure the high current part of the *I-V* curves of these devices without subjecting the device to significant stress.

One way to circumvent this, however, is by flooding the device with light. As discussed in Section 2.2.2, if enough light is injected into the device, the peak voltage can be reduced until all NDR disappears – the device is essentially forced to be in the high current state at any voltage. This allows measurement of at least the low current part of the high current state without ever applying a large voltage across the device.

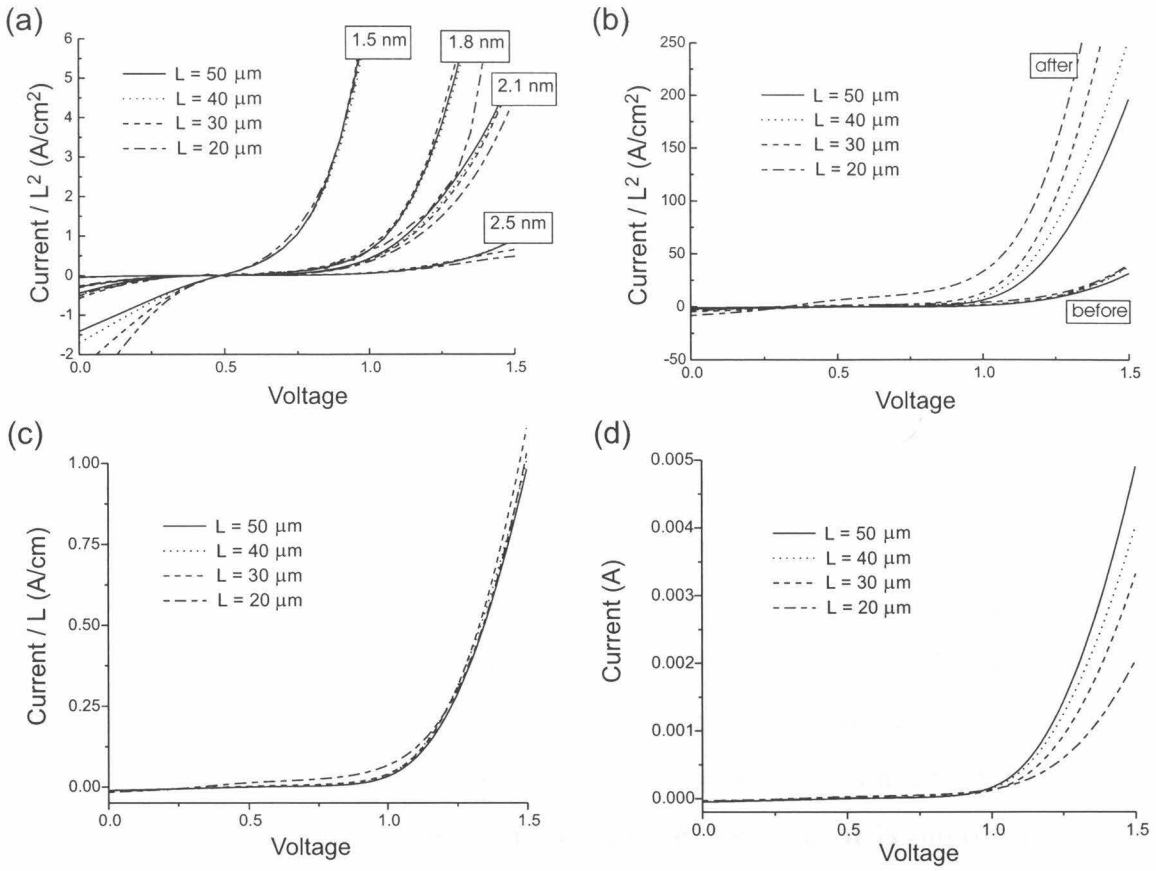

Figure 4.7: Low voltage scaling behavior of the TSD with light injection. (a) "Virgin" J-V curves taken before the devices had ever been subjected to high field stress. Devices of four different sizes are shown in groups corresponding to samples with different oxide thicknesses. (b) J-V curves for a group of different size devices on the 1.8 nm oxide sample taken both before and after high field stress. The "after" curves are also plotted as (c) Current/L^1 and (d) Current/L^0 in order to illustrate the modified scaling behavior.

Note that, as the bulk of the bias is dropped across the oxide in the high current state, this method effectively allows isolation of the behavior of the MOS part of the device.

In Figure 4.7(a), we plot virgin *J-V* curves for four different size devices on chips C45, C46, C47, and C48, possessing oxides of the four different thicknesses shown. We see that the curves very nearly collapse into four groups corresponding to the four different oxide thicknesses, suggesting nearly uniform current density through each of the oxides. While the curve corresponding to the 20×20 μm^2 device on the 1.8 nm oxide sample does not quite coincide with the other curves for this sample,

it is possible that the probe tip scraped off part of the Al pad, resulting in a smaller effective area.

In Figure 4.7(b), we plot similar J-V curves of a set of different size devices on chip C46 under illumination both before and after the devices had been subjected to 100 mA maximum AC stress for roughly 1 minute. We see, first of all, that the (average) current density through the devices increased by over an order of magnitude after stressing. We also note that the currents no longer scale with the area of the device, L^2. In fact, we see by comparing Figures 4.7(b), (c), and (d), that the current most closely scales linearly with the edge length, L. Similar behavior was observed for the devices on chips C45, C47, and C48.

As these devices used the simple "flat" geometry, it is likely that this modification of the scaling behavior is due to mild breakdown of the oxide at the edges of the device where the electric field is substantially higher, allowing greatly increased conduction around the perimeter. This increase in conduction at the edges must be considerable as it overwhelms the conduction of the rest of the device, so it is interesting to note that it does not completely short out the device – the TSD continues to operate with low and high current states. This suggests that, although the conduction through the oxide has been greatly increased, it seems to still act as a partial barrier.

4.3.3 Comparison of Scaling of Field Oxide and Flat Devices

As discussed in Section 3.3.1, a field oxide device geometry can be used to reduce edge effects present in flat gate devices. Therefore, such a device might not be as prone to the edge effect described above, and might continue to scale with area, even under stress. As both field oxide and flat devices were present on the masksets used in fabrication of devices at Stanford, these wafers allow for direct comparison of these different geometries with all other parameters equal.

Unfortunately, the behavior of all of the punch-through devices produced at Stanford was erratic at best, possibly because the wafers used had fairly thick epi layers, leading to large peak voltages (20-25 V) and therefore high stress when the device

switched from the low current state to the high current state. Therefore, it is difficult to make much of a comparison, but some trends were apparent. Although avalanche devices were also produced at Stanford, their scaling was a bit different, and will be discussed in Section 4.3.5.

As the *I-V* curve of a device depends heavily on the amount of stress that device has endured, some convention was needed for determining how to stress different size devices for a useful comparison. It was decided that the devices would be stressed to a constant (average) current density in order to encourage a uniform current transport. A value of 125 A/cm^2 was chosen as it seemed to produce reasonable results for the midrange device sizes. Each device was, therefore, stressed for a few seconds to a maximum current equal to 125 A/cm^2 times the area of the device. The resulting measured *I-V* curves for flat and field oxide devices from Stanford wafer W1 are shown in Figures 4.8 and 4.9. Similar results were obtained from the other punch-through wafers. Note that, while the devices were only stressed to 125 A/cm^2 before measurement, the *I-V* curves were scanned to a maximum current of 50 mA, so the highest current parts of the curves corresponding to any device smaller than $200\times200\mu$m^2 represents incremental current stress. The horizontal lines in Figures 4.8(c) and 4.9(c) denote the maximum stress applied before the *I-V* measurements were taken. Any data points lying above these lines represent additional stress, and are not considered in these scaling studies.

For all but the largest flat devices, it is difficult even to identify a definite peak point on the *I-V* curve, and in any event, the high current parts of the curves do not appear to scale with any power of the device edge length. The best evidence of any scaling behavior appears in the low current part of the flat devices, which appear to scale with area. The field oxide devices do not appear to scale as well, even in the low current regime, but note that the field oxide devices are smaller than the flat devices, in general, and the smallest devices have edge lengths of the order of the epilayer thickness, so one-dimensional behavior is not expected. The parts of the *I-V* curves just before where the peak should be (the nearly vertical parts) do seem to collapse more closely to a uniform current density in the field oxide devices than in the flat

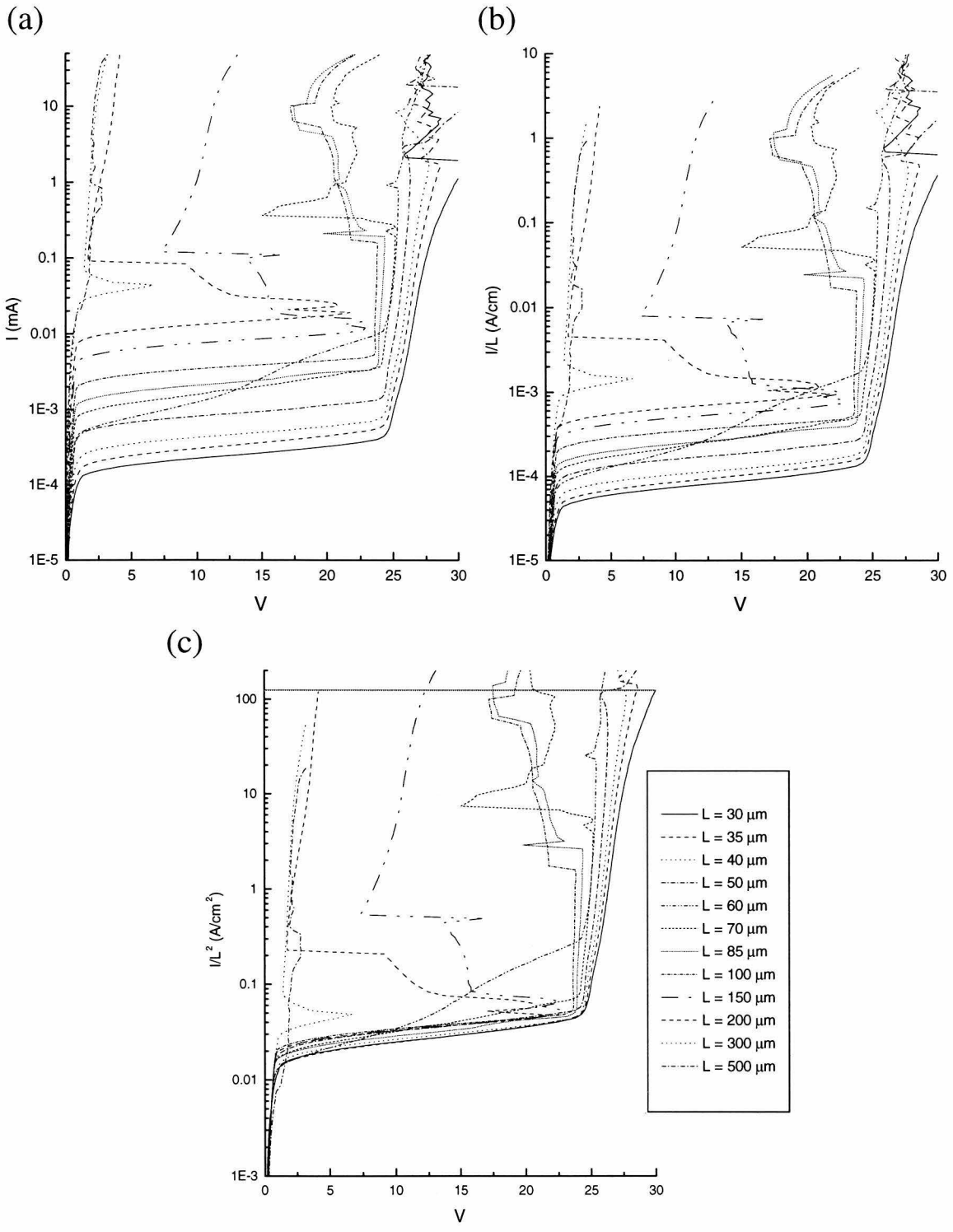

Figure 4.8: *I-V* curves of different size flat devices from Stanford wafer W1, stressed to 125 A/cm^2. The curves are plotted as (a) I vs. V, (b) I/L vs. V, and (c) I/L^2 vs. V, where L is the length of the edge of each device. The horizontal line in (c) indicates the maximum stress point.

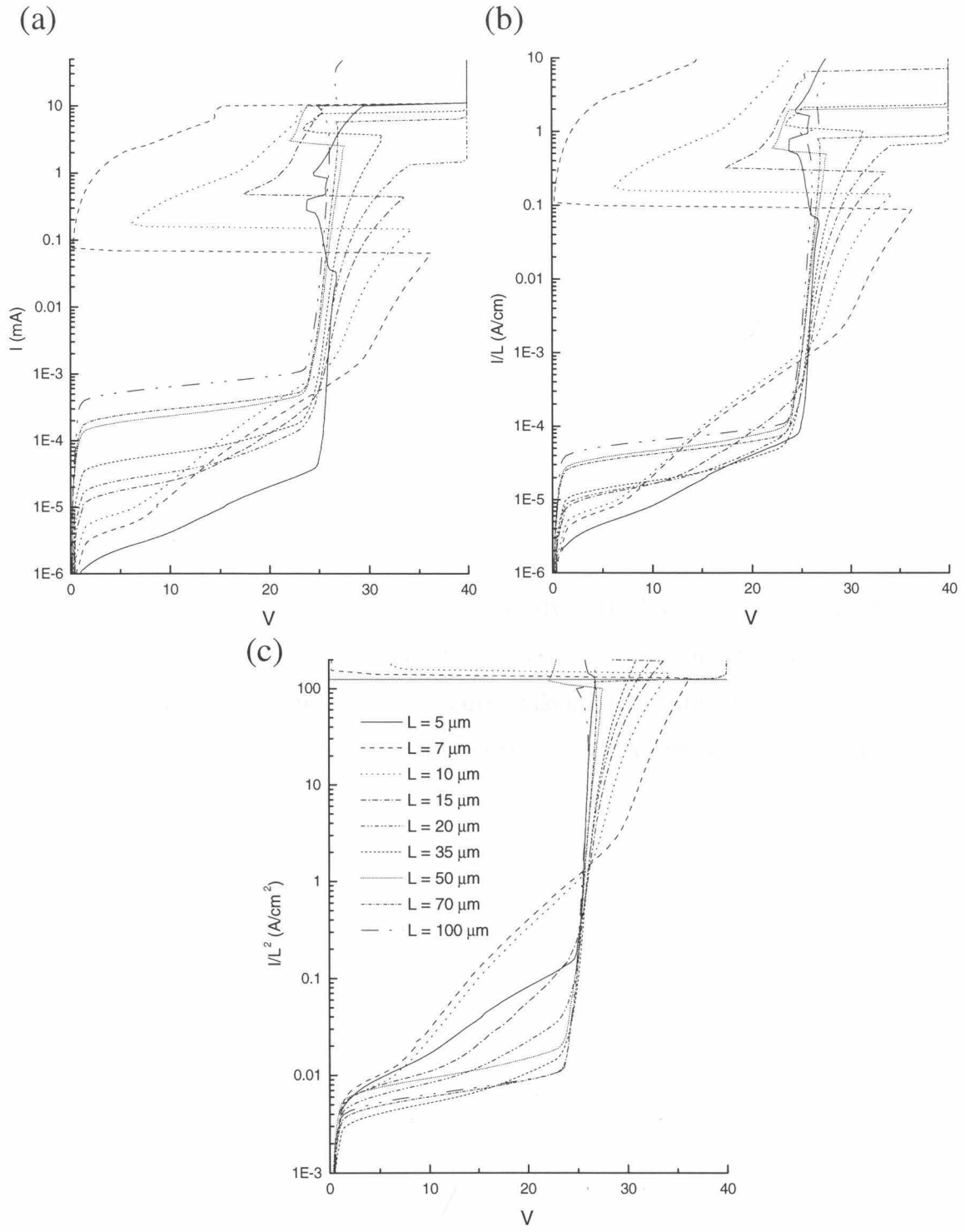

Figure 4.9: *I-V* curves of different size field oxide devices from Stanford wafer W1, stressed to 125 A/cm^2. The curves are plotted as (a) I vs. V, (b) I/L vs. V, and (c) I/L^2 vs. V, where L is the length of the edge of each device. The horizontal line in (c) indicates the maximum stress point.

devices, however. Note, also, that the smallest devices show the most deviation in this part of the curve.

It is interesting to note that the peak voltage of the largest flat devices decreases with increasing size until the peak is eliminated for the $500\times500\mu m^2$ device. As no devices of comparable size were fabricated in the field oxide geometry, it is unknown whether such devices would have exhibited the same behavior. The origin of this behavior is unknown, but similar results will be presented and discussed in Section 4.3.5.

4.3.4 Scaling of High and Low Current Modes of Poly Gate Devices

In order to investigate the scaling of the low and high current modes of the punch-through poly gate devices, *I-V* curves taken during incremental current stressing of different area devices can be compared. For study of the low current mode, the highest current *I-V* curve which did not reflect any major stress-induced change was chosen. For study of the high current mode, the curve taken immediately following a maximum of 10 mA stress was chosen. A constant current density stress, as was used in study of the field oxide and flat Stanford devices, was attempted, but as relatively large currents were required to force the smallest device into the high current mode (1-10 mA), the necessary current stress required for large area devices ($1320\times1320\mu m^2$) was infeasible (.25-2.5 A).

Low current mode *I-V* curves of four different area poly gate punch-through mode devices from Intel wafer W13 are plotted in Figure 4.10. The low current state is limited by parasitic impedance for all but the largest area device and thus cannot be expected to scale properly, but the high current state appears to scale with area. The very highest currents appear to begin to scale more closely with the edge, however, possibly due to edge effects that are too strong to be eliminated by the field oxide geometry. The currents are low enough that the measured voltage drop across the gate was negligible and could not account for this effect.

In order to compare this result with the scaling of devices with thinner oxides,

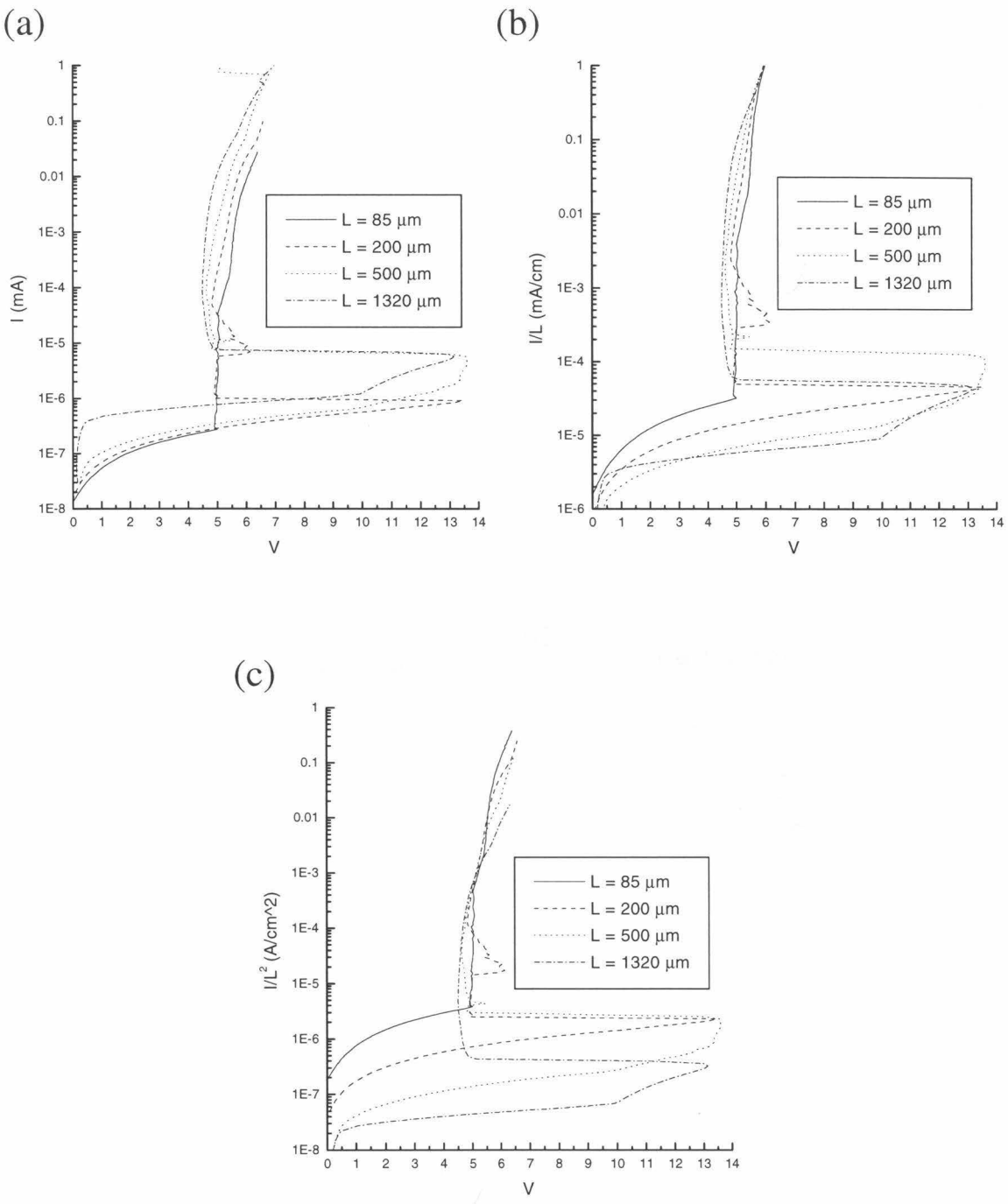

Figure 4.10: *I-V* curves of different size field oxide devices from Intel wafer W13. The curves are plotted as (a) I vs. V, (b) I/L vs. V, and (c) I/L^2 vs. V, where L is the length of the edge of each device.

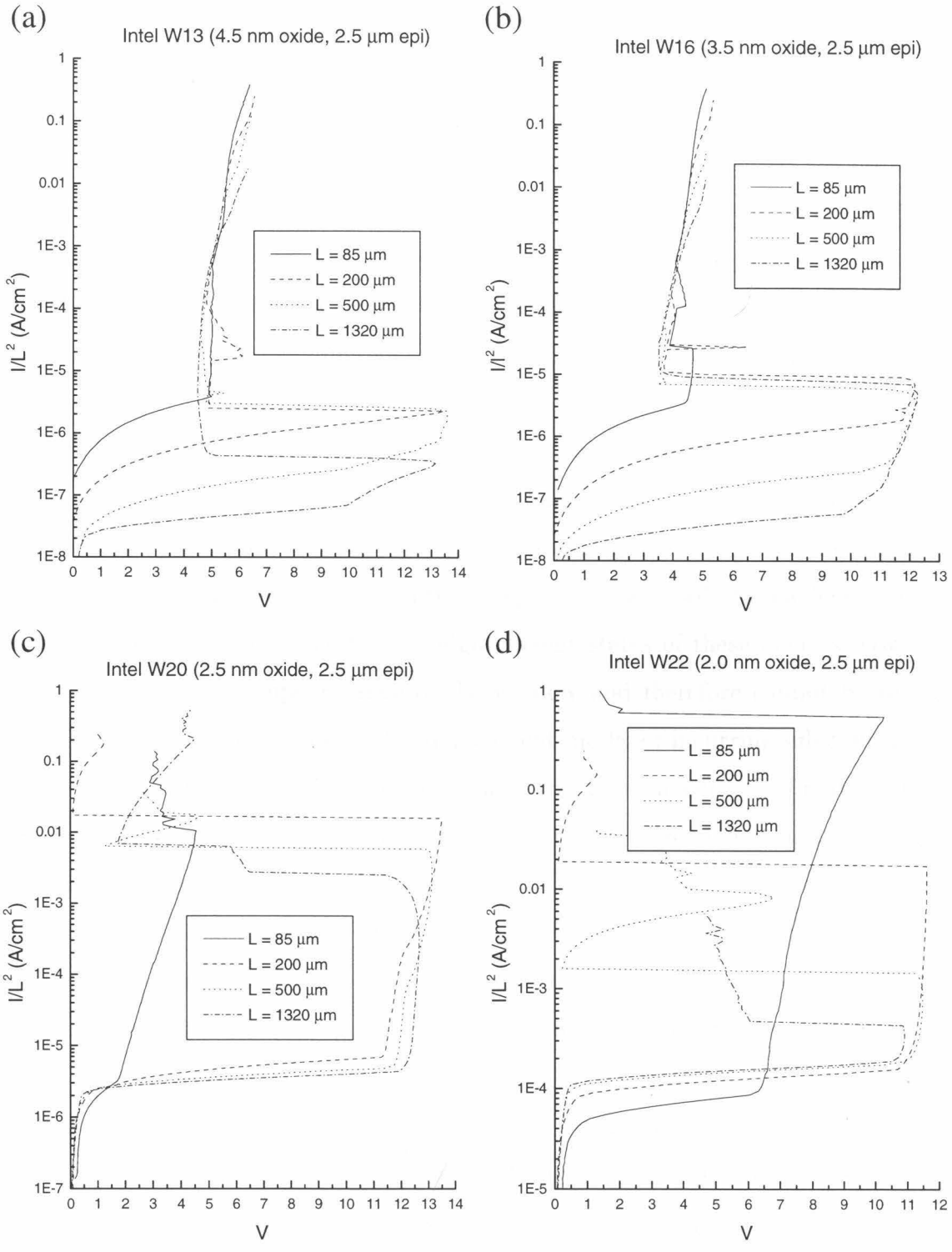

Figure 4.11: J-V curves of different size punch-through devices in low current mode with (a) 4.5 nm, (b) 3.5 nm, (c) 2.5 nm, and (d) 2.0 nm oxides.

the average current density versus voltage is plotted in Figure 4.11 for different size devices from 2.5 μm epi wafers with different oxide thicknesses (W13, W16, W20, W22). Recall from Section 3.4 that there is a window in current in which we can expect to obtain reasonable I-V data for scaling purposes, the low end defined by the parasitic impedance, and the high end defined by voltage drop across the poly gate reducing the effective area of the device, or by excessive stress which kicks the device into the high current mode. Also, note that as the oxide thickness decreases the entire TSD I-V curve shifts up in current. Therefore, by looking at devices with different oxide thicknesses, we move different parts of the TSD I-V curve through the current window.

In Figures 4.11(a), (b), only the high current state of the TSD device falls within the window, and this range is found to scale roughly with area. In Figures 4.11(c), (d), the low current state of the TSD has been brought into the window, and now it is found to scale roughly with area (with the exception of the smallest area devices which may have incurred some stress). The high current states of these devices, however, have moved out of the upper range of the window and therefore cannot be probed without stressing the devices into the high current mode, or incurring substantial gate voltage drop. Note that the drop in voltage near the high current end of many of these curves is the result of stress forcing the device into the high current mode, and does not represent the peak of the I-V curve, as can be verified by a subsequent scan. Therefore, while the entire low current mode TSD curve for a given device cannot be shown to scale with area, the fact that all parts of the curve scale with area when they are within the window suggests that the nature of this mode of operation is very nearly one dimensional.

Note the similarity of the scaling and shape of the low current TSD mode I-V curves to those of the unstressed simple poly gate MOS structures shown in Figure 4.12. This suggests that the high current state transport in the TSD is controlled by the MOS part of the device, as expected.

In contrast, transport in the high current TSD mode is decidedly non-uniform. Consider the curves shown for devices of different areas and oxide thicknesses in Fig-

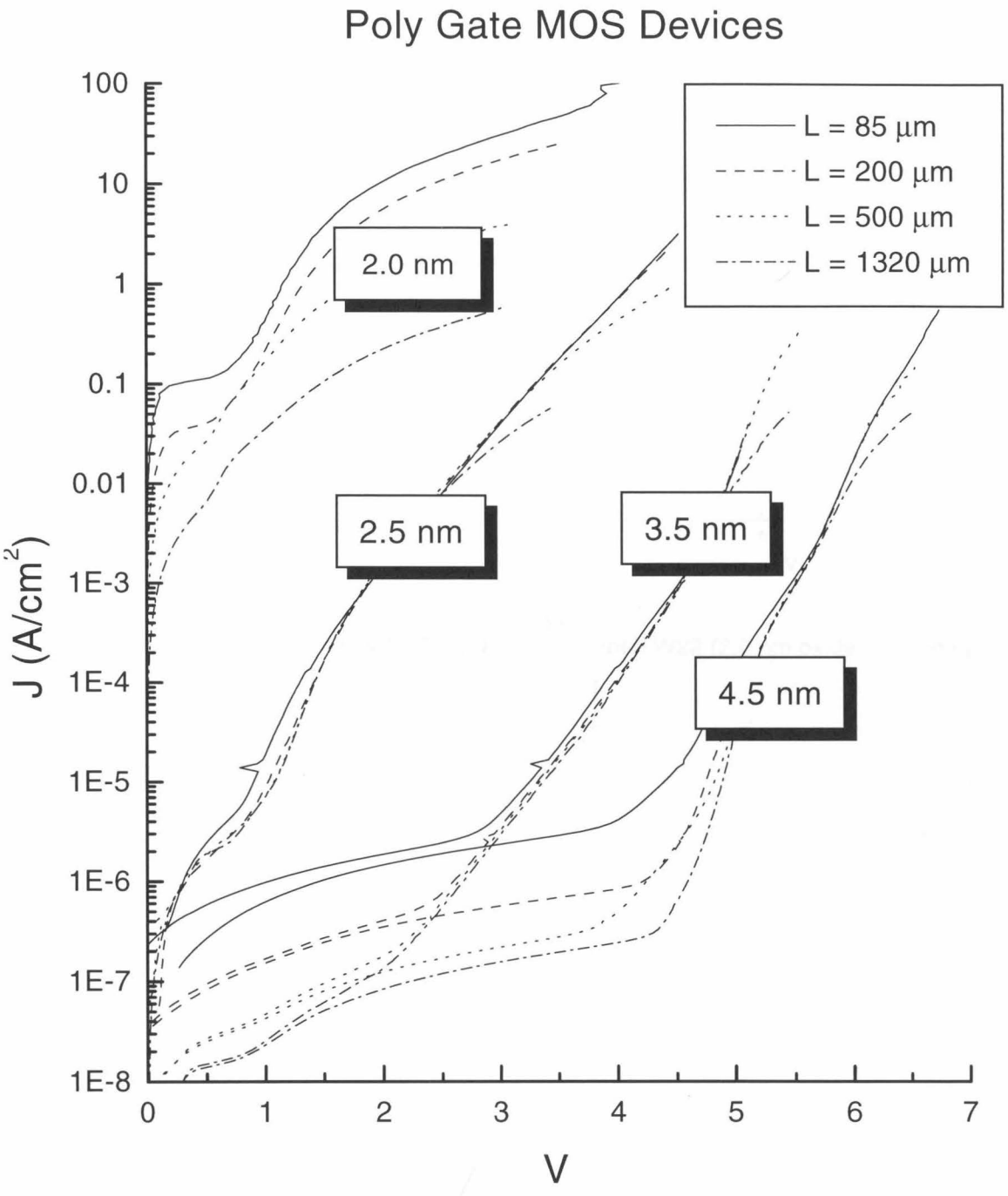

Figure 4.12: MOS device J-V curves for devices with different oxide thicknesses and areas. For all but the thinnest oxide, current transport appears to be roughly uniform.

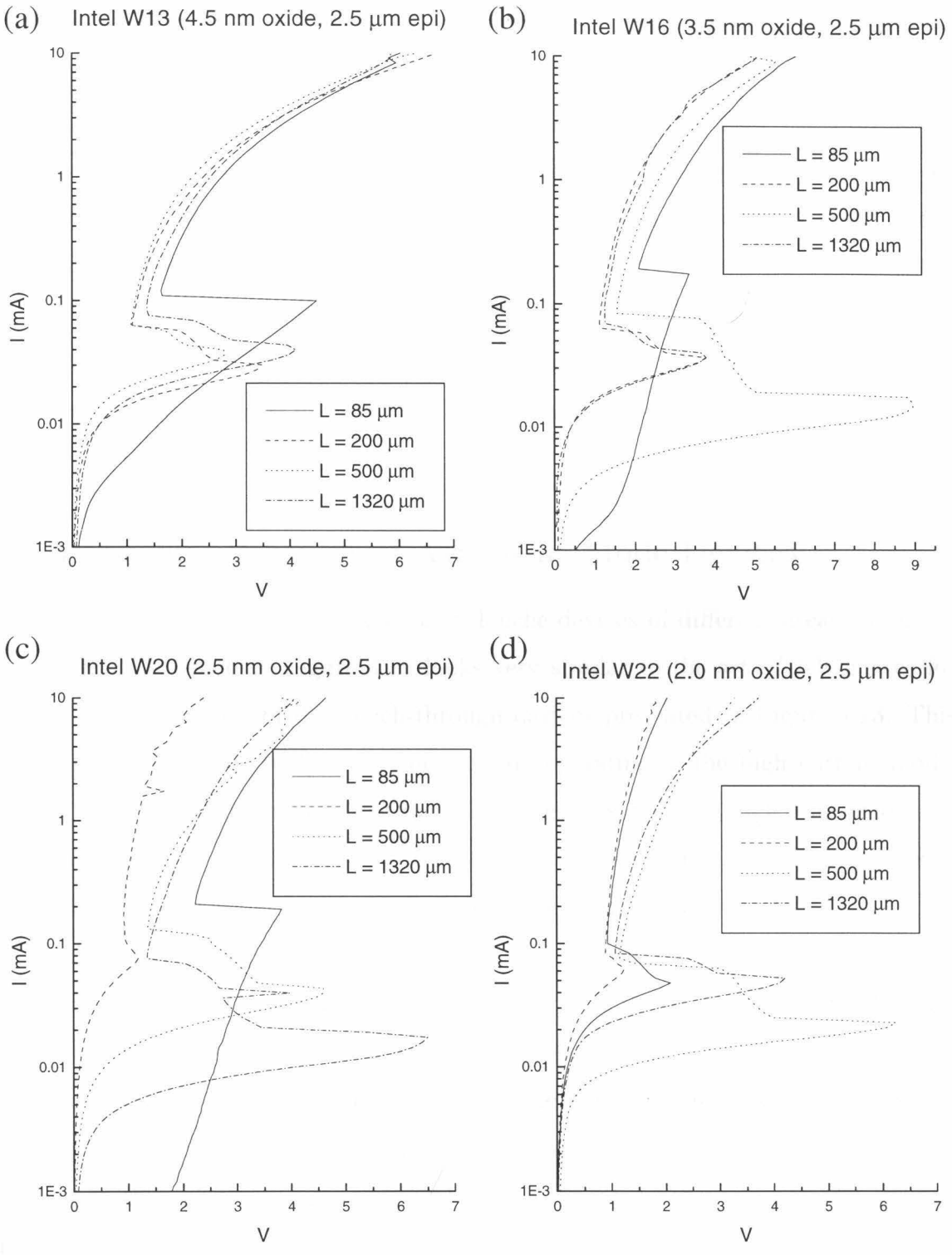

Figure 4.13: *I-V* curves of different size punch-through devices in high current mode with (a) 4.5 nm, (b) 3.5 nm, (c) 2.5 nm, and (d) 2.0 nm oxides.

ure 4.13. No definite trends are followed, but at best, the device I-V curve is roughly independent of device area and oxide thickness. In fact, valley currents in a fairly tight cluster of roughly .06-.3 mA were observed for high current TSD modes of all Intel punch-though mode devices, regardless of stress history, device area, oxide thickness, or epi thickness. Thus, is appears that the dominant current transport mechanisms are far from uniform in current. It might be suggested that this behavior may have been encouraged by the fact that devices were stressed to a constant maximum current (10 mA) rather than a constant current density as with the Stanford Al gate devices described above, but a range of stress histories were explored, producing roughly the same results. These same basic results for both modes were found for the 1.5 μm epi wafer set.

4.3.5 Scaling Behavior of Avalanche Mode Devices

The survey of I-V curves of poly gate avalanche devices of different areas and oxide thicknesses presented in Figure 4.14 looks very similar to the set of I-V curves for the high current mode of the punch-through devices presented in Figure 4.13. This is not surprising since both sets of devices are operating in the high current mode. Again, the shape and magnitude of the I-V curves was generally independent of area and stress history. Unfortunately, a low current mode could not be reliably produced with the avalanche wafers, so no scaling properties could be explored.

Compare these results to those from the set of Stanford wafers in Figure 4.15, however. Here, the current density curves in (b), (d), collapse into two bundles corresponding to the low and high current states of the TSD, suggesting uniform current transport, but the current at which the devices switch between these two states seems to be independent of area! Both field oxide and flat devices exhibit this behavior, although the pre-measurement stress current density was not high enough to force the smallest field oxide devices into the high current state, so they incurred additional stress instead of dropping down into the high current state during the high current part of the scan.

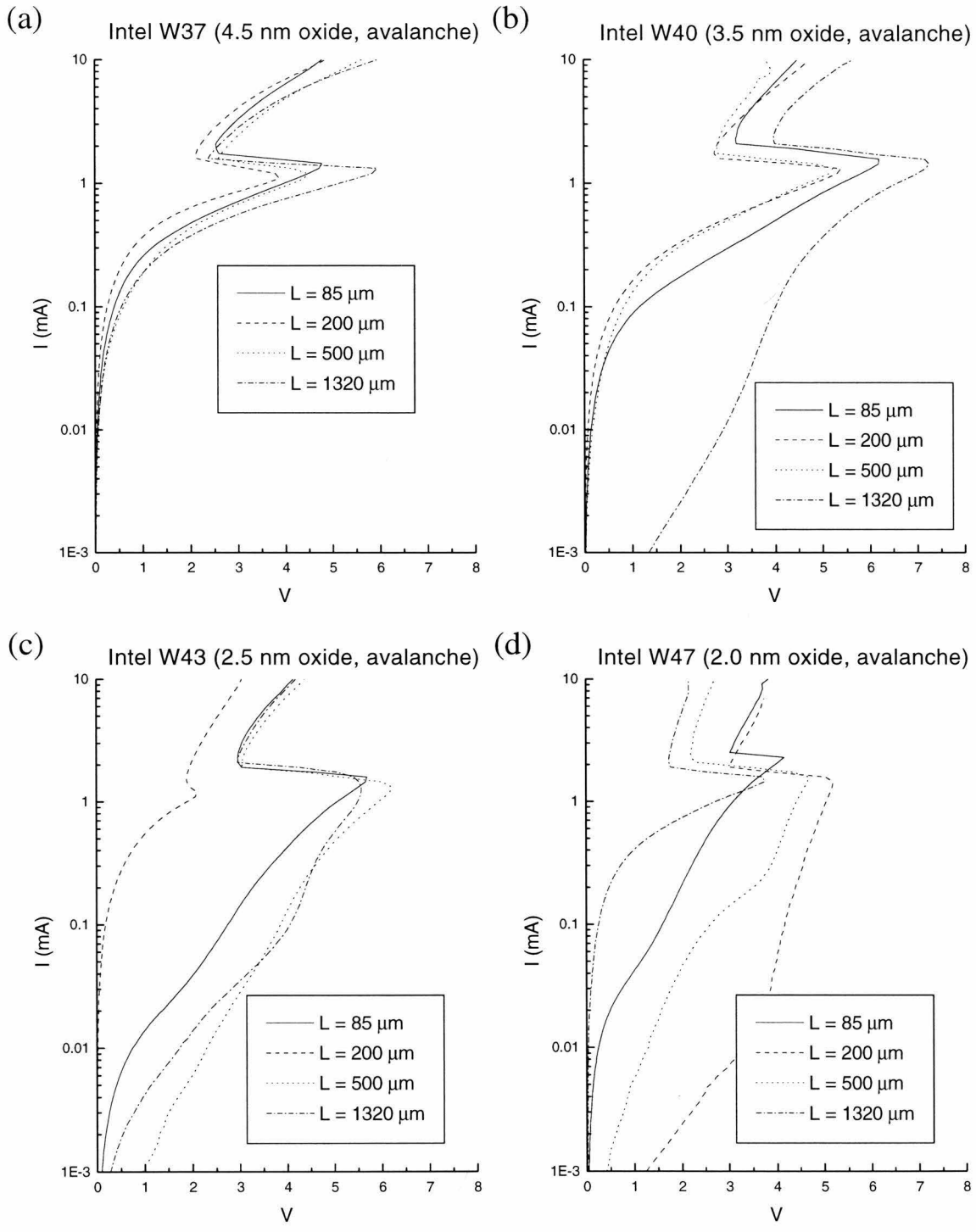

Figure 4.14: I-V curves of different area poly gate avalanche devices in high current mode with (a) 4.5 nm, (b) 3.5 nm, (c) 2.5 nm, and (d) 2.0 nm oxides.

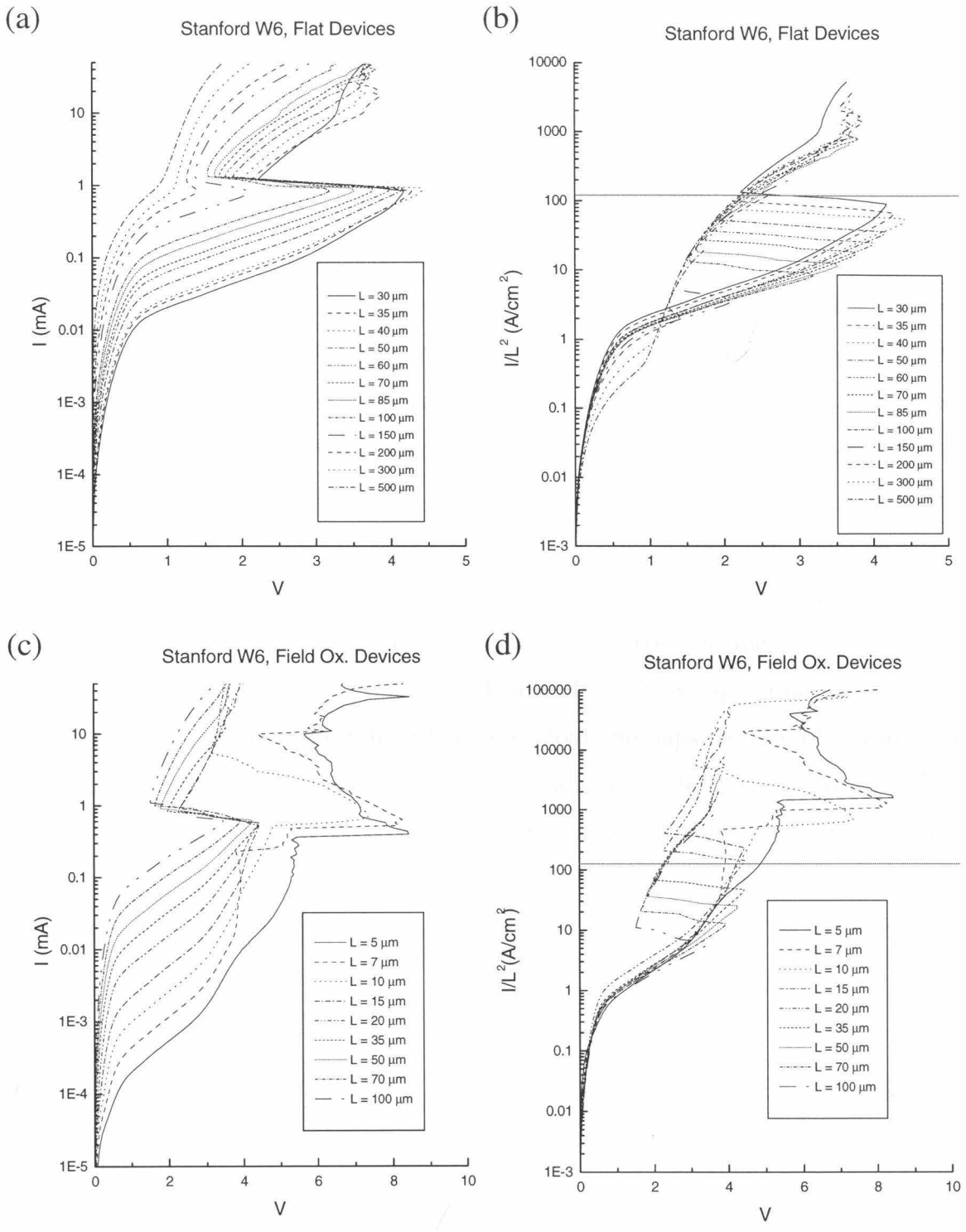

Figure 4.15: *I-V* curves of different area Al gate avalanche devices. Flat device curves are plotted as (a) I vs. V and (b) I/L^2 vs. V where L is the length of the edge of the device. Similar plots are made for field oxide devices in (c), (d). All devices have been stressed to a maximum of 125 A/cm^2 prior to measurement. This value is indicated by the horizontal lines in (b), (d).

This result is particularly puzzling as it suggests both uniform and nonuniform mechanisms are at play simultaneously. Similar behavior has been evident in results already presented, such as the variation of peak voltage with device size for the largest flat Al gate devices shown in Figure 4.8 and the distinct difference in scaling behavior of the low and high current modes of the poly gate devices. One possible explanation for all of these results may be tied with the fact that the device area is defined only by the size and geometry of the gate and field oxide layers. In particular, as the avalanche wafers are fairly heavily doped, it is quite possible that the silicon epi layer is conductive enough that it remains an equipotential far outside the area defined by the gate pad, forcing the p-n junction part of the device to have a much larger area (see Figure 4.16). Therefore, as the p-n junction is turned on when the device switches, this mechanism may be independent of the gate area. A similar argument could be made for the punch-through devices, given that recombination lifetimes are typically long in silicon and could allow for significant diffusion of the carriers in the epi layer outside of the bounds of the gate area. While the details of how these spreading mechanisms might come into play are uncertain at best, they could perhaps be tested by producing mesa etched devices or devices with isolation implants which force the active area of all layers of the device to be the same. Of course, these methods introduce other possible sources of nonuniform current transport, such as edge leakage in a mesa structure, or lateral depletion in an isolation implanted structure, but such studies might still be able to address this issue.

4.4 Dependence of TSD I-V Curve on Design Parameters

In the previous sections, we have seen that it is not a trivial matter to compare TSD devices fabricated with different sets of design parameters, such as oxide thickness, epilayer thickness, and epilayer doping level, as the device behavior is also heavily dependent on stress history, and in some cases, device area and fabrication details.

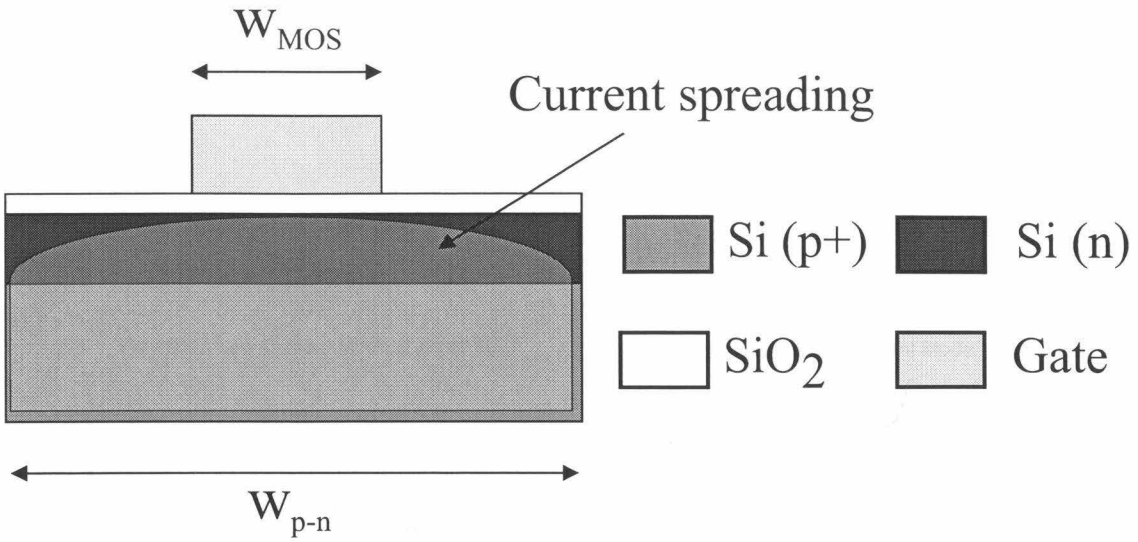

Figure 4.16: Either due to relatively high conductivity or long recombination times, and therefore long diffusion lengths, it is possible that the equipotential part of the top silicon layer in the TSD could spread out far beyond the edges of the gate. This would mean that the gate part of the device and the p-n junction part of the device would have different characteristic sizes.

In light of the results described in the previous sections, however, we will discuss the influence of different design parameters on TSD device behavior in a broad sense.

4.4.1 Oxide Thickness

It has already been shown in Section 4.3.4 that the two modes of operation of the punch-through TSD devices behave very differently with respect to changes in oxide thickness. This result is reiterated in Figure 4.17. The low current mode curves increase in current rapidly with decreasing oxide thickness, much like simple MOS devices (Figure 4.12). The high current mode curves, however, are relatively independent of oxide thickness. As significant stress is needed in order to reach the high current mode, it is possible that this stress "forms" the oxide in a self-limiting way such that, after the forming process, any oxide thickness is equally conductive. Some hint of such behavior was illustrated in the discussion of voltage limiting behavior in Section 4.2.2.

Poly gate avalanche devices also seemed to yield I-V curves relatively independent

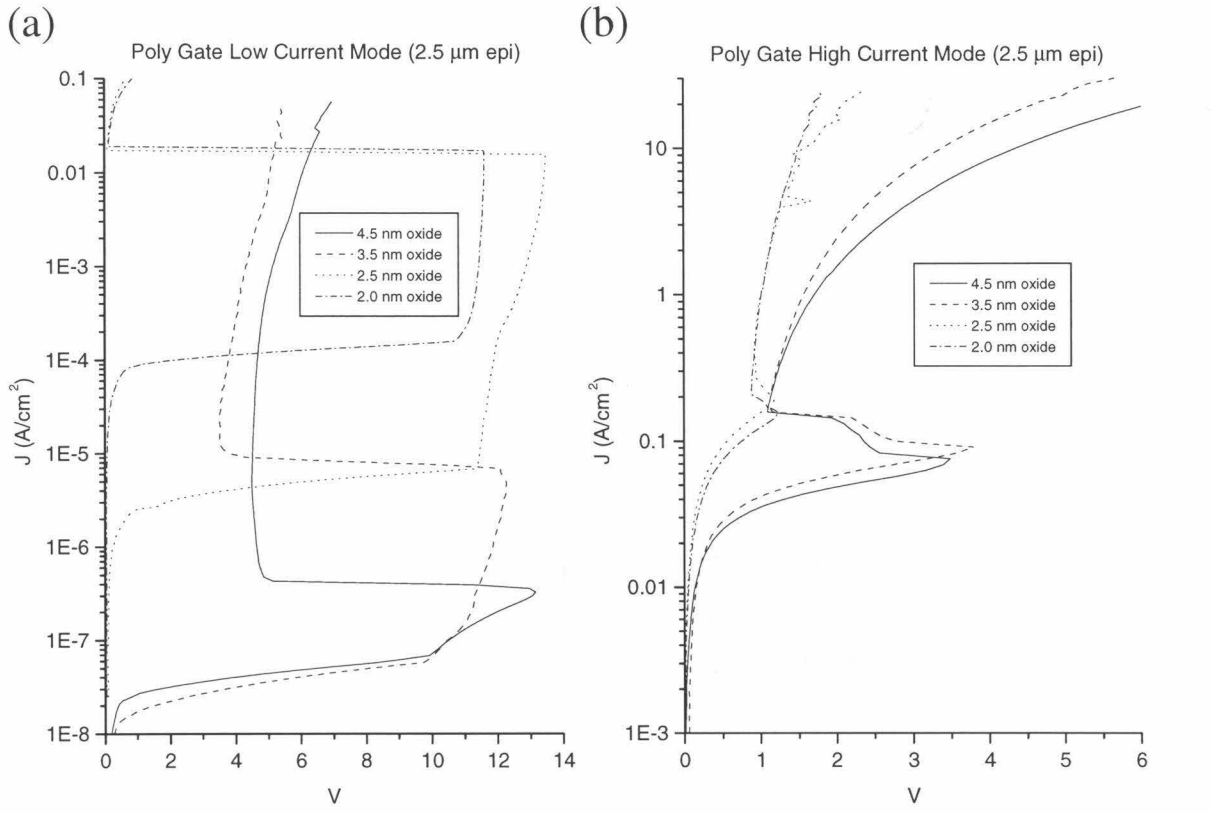

Figure 4.17: J-V curves for select devices of different oxide thicknesses are plotted for punch-through devices (2.5 μm epi) in (a) low current and (b) high current modes. In (a), 1320×1320 μm^2 devices were chosen for the 4.5, 3.5 nm oxides so that the low end of the curve was not limited by parallel impedance. 200×200 μm^2 devices were chosen for the 2.5, 2.0 nm oxides in order to minimize distortion of the high current end of the curve due to voltage drop across the gate. In (b), all devices chosen were 200×200 μm^2.

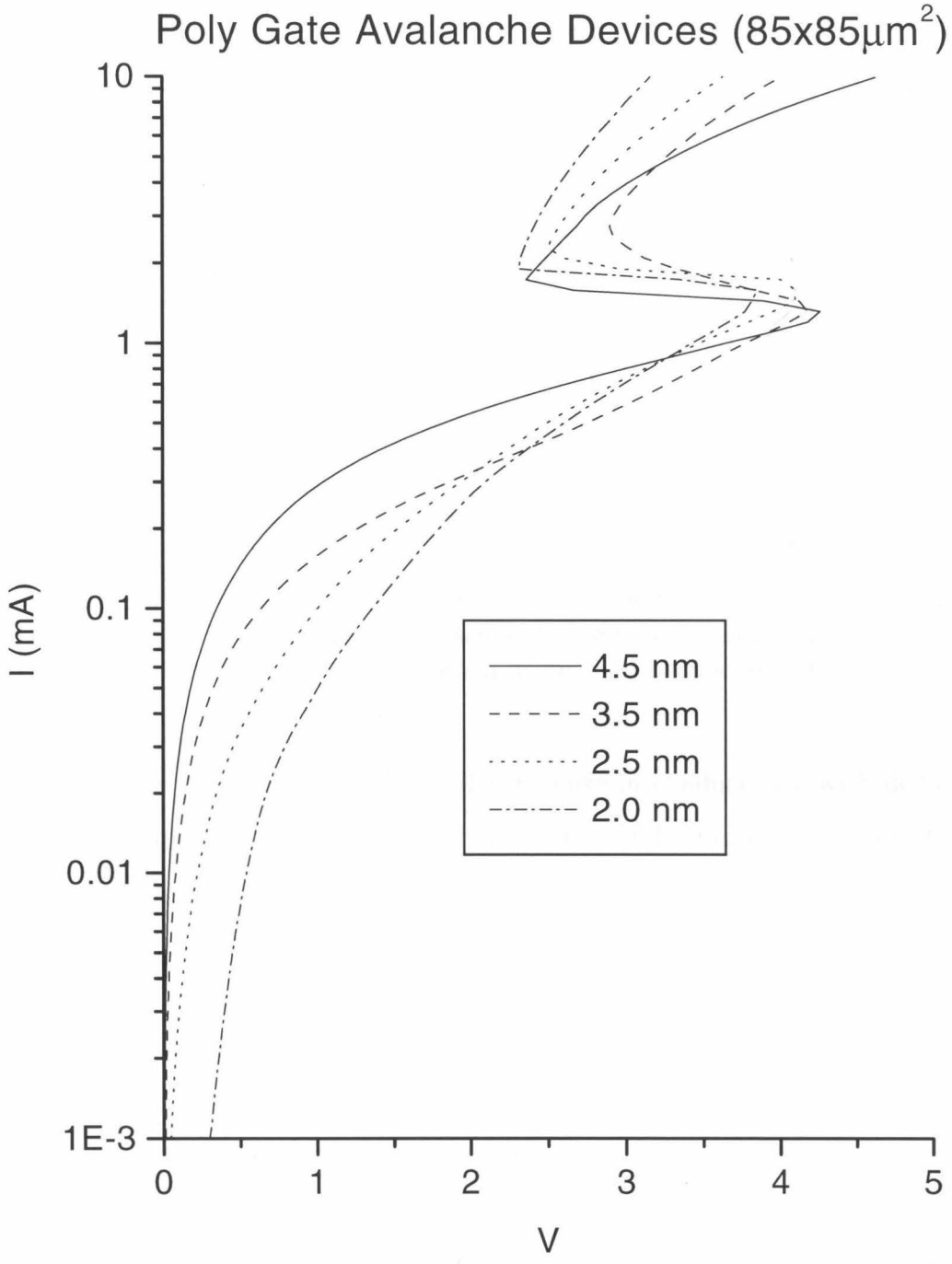

Figure 4.18: *I-V* curves for poly gate avalanche devices of different thicknesses (Intel wafers W37, W40, W43, W46). Each curve represents the average of eight curves taken from $85\times85\mu m^2$ devices across the wafer. Each device had been stressed to 10 mA prior to measurement.

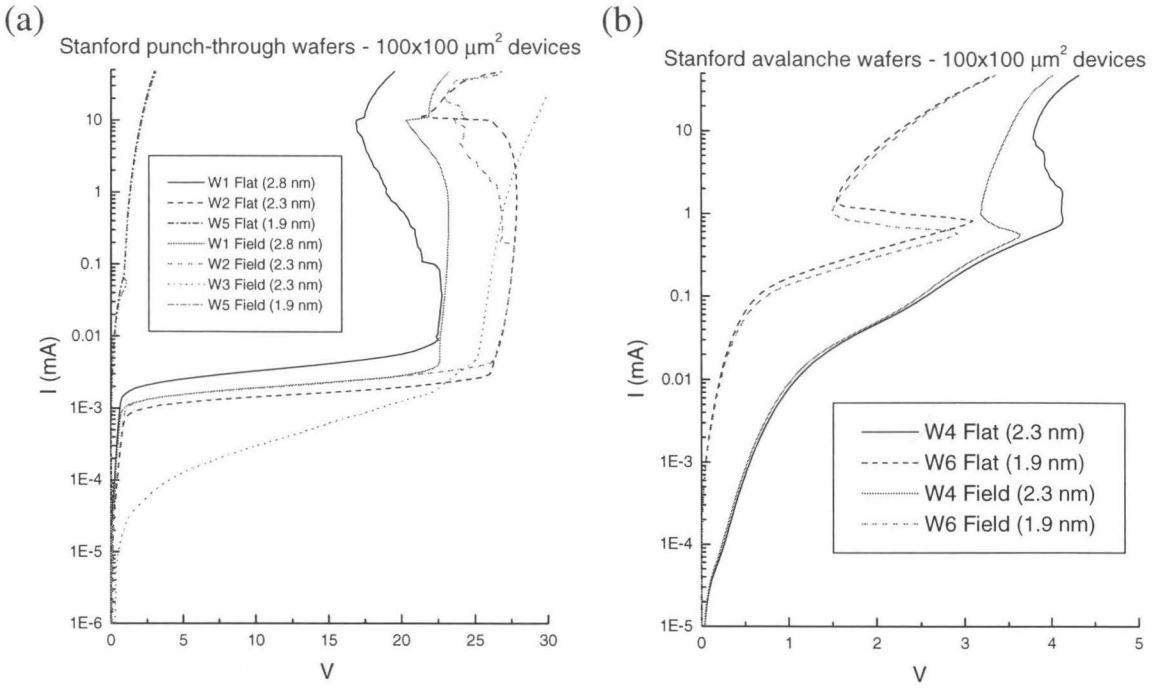

Figure 4.19: *I-V* curves for Al gate (a) punch-through and (b) avalanche wafers produced at Stanford. Each curve represents the average of eight curves taken from $100\times100\mu m^2$ devices across the wafer. Each device had been stressed to 50 mA prior to measurement.

of oxide thickness (Figure 4.18). While the increase in conductivity with decreasing oxide thickness apparent in the high current state would be easy to explain, the opposite behavior exhibited in the low current state poses more of a problem. One possibility is that, as thicker oxides would require more stress damage to bring their conductivity up to the roughly constant level observed, they may become more conductive at lower voltages than thinner oxides. It should be noted, however, that these variations in the *I-V* curves from one oxide thickness to another are of the same order of magnitude as the variations from one device to the next across a given wafer, so the trends may be coincidental. The fact that the illustrated curves were taken as averages of eight devices across each wafer adds credibility to the trends, but more work would be necessary to make a more definitive statement.

Figure 4.19 summarizes the behavior of punch-through and avalanche devices with both flat and field oxide geometries produced at Stanford. The punch-through devices were quite difficult to characterize, so again, the differences between the curves cor-

responding to different oxide thicknesses were of the same order of magnitude as the variations from one device to the next. Although the thinnest oxide devices exhibited no peak, recall that the peak size was found to depend on device area for these devices (see Figure 4.8). In fact, smaller devices on wafer W5 did have peak voltages of the same size as those on W1, W2, and W3 ($\sim$25 V). It appears that the "critical size" at which the peak begins to appear may decrease with decreasing thickness, however. Stanford Al gate avalanche devices were only produced with two different oxide thicknesses, so a trend cannot really be established, but the thinner oxide devices seem to allow higher currents than the thicker oxide devices, as might be expected.

Caltech punch-through devices of different oxide thicknesses are compared in Figure 4.20. The peak voltage increases with increasing oxide thickness, but the high current state currents seem relatively independent of oxide thickness. In fact, given that the only curve that deviates from the rest is the one that represents lower maximum current stress prior to the measurement, this data supports the idea that the "forming" process that takes place during incremental current stressing forces the different thickness oxides to all be equally conductive at the maximum current. The question then arises, if the oxides are all equally conductive, why do the peaks differ? Given that stressing appears to induce conduction primarily around the edges of these devices (see Figure 4.7, Section 4.3.2), it is possible that edge conduction provides the current in the high current state, but that the oxide covering the rest of the device plays a role in the lower current parts of the I-V curve, such as the peak and valley.

4.4.2 Epilayer Properties

Recall that the epi wafers were designed such that the TSD devices produced with them would operate either in the avalanche or punch-through mode with specific peak voltages, as listed in Table 3.2. It is difficult to compare these values directly with measured curves, however, as the peak voltage has been found to vary with device area and stress history. As stress appears to predominantly lower the peak voltage (see Figure 4.1), the larger values obtained before significant stress modification will

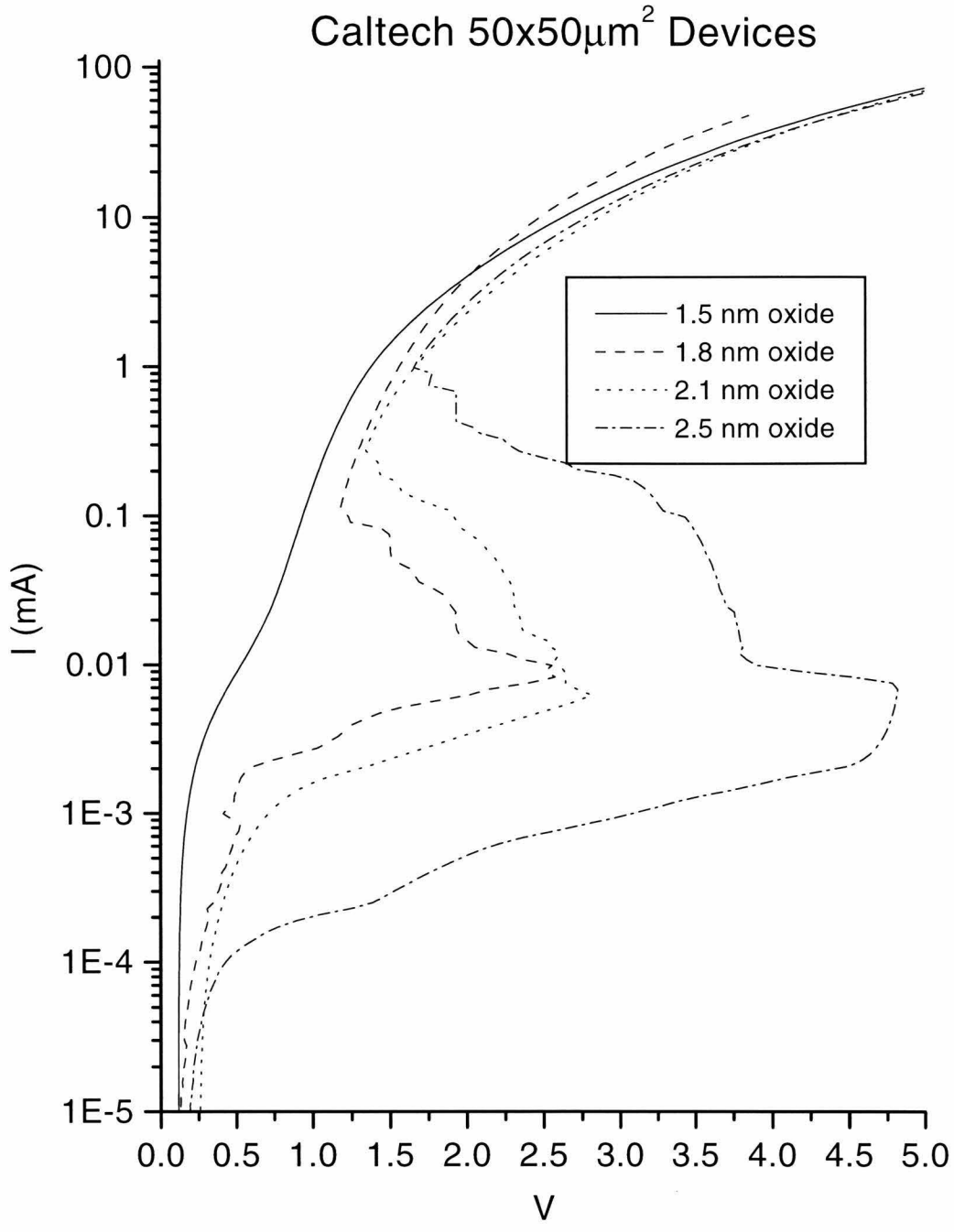

Figure 4.20: *I-V* curves for Al gate punch-through devices of different oxide thicknesses produced at Caltech. Each curve represents the average of eight curves taken from $50\times50\mu m^2$ devices across the chip. Each device had been stressed to 100 mA prior to measurement, except for the 1.8 nm oxide devices which were accidentally only stressed to 50 mA.

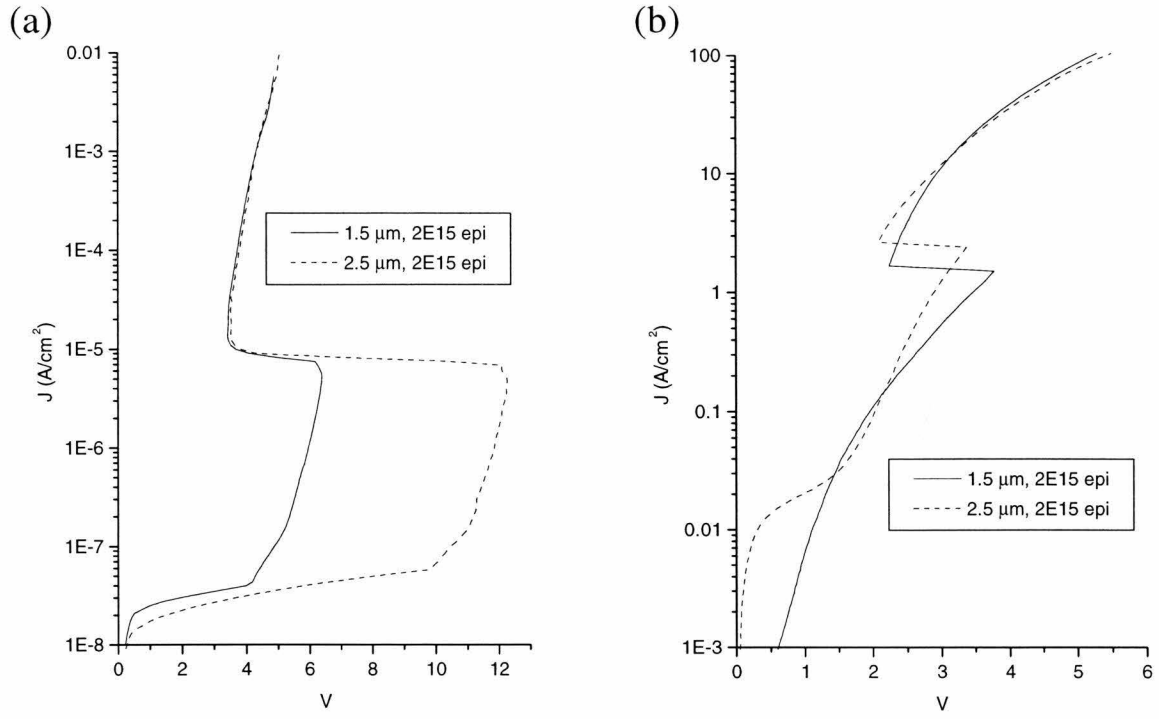

Figure 4.21: J-V curves for select (a) low current mode and (b) high current mode punch-through devices with different epi layer thicknesses. All devices had a 3.5 nm oxide (wafers W16, W28). The high current mode devices were stressed to 10 mA prior to measurement.

be considered. Areas in the range of (50-100 μm)^2 will be considered as they are neither too small for the geometry to be manifestly two-dimensional, or too large for gate voltage drop (in the case of poly gate devices) to cause multidimensional effects.

The overall range of all avalanche device peak voltages was in the range of 5-10 V, in fair agreement with the estimated value of 6 V (see Figures 4.14, 4.19). The punch-through devices, however, typically exhibited peak voltages much greater than those predicted. This result was quite unexpected, given that a fair amount of data has been gathered, supporting the models used in the calculations, albeit mostly for devices with fairly large peak voltages [2–4]. Although this result would be most easily explained by multidimensional effects which would lead to larger voltages dropped across the depletion layer, many of the devices exhibited some signs of uniform current transport, particularly in the low current state in which the depletion occurs. Perhaps most notably, the poly gate punch-through devices which seem to scale with area

Table 4.1: Summary of TSD device behavior. For each class of device studied (first four columns), the scaling behavior, peak voltage (measured and calculated), observation of voltage limiting behavior, and dependence on oxide thickness are listed.

Fab	**Type**	**Geom.**	**Curr. Mode**	**Scaling**	**V_{peak} Meas.**	**V_{peak} Calc.**	**V limit Obs.?**	**t_{oxide} Dep.**
Caltech	P-T	Flat	—	L^1, L^2	3-6	<1	Yes?	Little
Stanf.	P-T	Flat	—	L^2 ??	20-25	12	?	?
Stanf.	P-T	Field	—	L^2 ??	20-25	12	?	?
Stanf,	Av.	Flat	—	L^2 ?	3-7	6	?	None?
Stanf,	Av.	Flat	—	L^2 ?	3-7	6	?	None?
Intel	1.5 μm P-T	Field	Low	L^2	5-7	1	—	Large
Intel	1.5 μm P-T	Field	High	$\sim L^0$	3-5	1	Yes	None
Intel	2.5 μm P-T	Field	Low	L^2	12-14	5	—	Large
Intel	2.5 μm P-T	Field	High	$\sim L^0$	5-10	5	Yes	None
Intel	Av.	Field	Low	insufficient data			—	?
Intel	Av.	Field	High	$\sim L^0$	5-10	6	Yes	None

completely, follow the expected trend of increasing peak voltage with epi thickness, but have much higher voltages than estimated (see Figure 4.21). It is possible that this discrepancy depends on the fact that the device area is defined only by the gate, leading to some lateral depletion which would require more voltage. As the depletion voltage increases with the square of the depletion region thickness, it is possible that a small underestimate of the epilayer thickness could be responsible, but it seems unlikely that all of the spreading resistance measurements would have underestimated the thicknesses.

Note that as the epilayer thickness is increased in the poly gate punch-through devices in low current mode, the peak region grows in voltage, but the rest of the *I-V* curve remains predominantly invariant (Figure 4.21). In fact, the high current state is almost identical, suggesting that essentially no additional voltage is dropped across the thicker epilayer region in the high current state. This supports the belief that the voltage across the device in the high current state drops predominantly across the oxide.

4.5 Summary and Discussion

The results presented in this chapter have been summarized in Table 4.1. For each class of device, the table indicates how the devices were observed to scale with the length (L) of the edge of the square devices, what the expected and measured ranges of peak voltages were, and whether the *I-V* curves were found to depend on oxide thickness. While nearly every entry requires some qualifying statements, this table allows for a quick comparison of the behavior of the various samples.

Perhaps the most interesting result obtained from this work was the discovery and characterization of the low current mode of operation in the poly gate punch-through devices. These devices possess a number of useful qualities: they scale with area, they change little during incremental current stressing until the high current mode takes over, the high current state part of the *I-V* curve is extremely steep, and the absolute magnitude of the current can be adjusted by changing the oxide thickness. These qualities all would be desirable for the transistorless SRAM application, for example. In fact, if the current range of the device could be increased by thinning the oxide further, the device parameters would enter the very interesting range represented by the upper right entries in Table 2.2. Unfortunately, the less desirable high current mode takes over before this occurs. It is possible that replacing the SiO_2 layer with an insulating layer (such as Si_3N_4) with a lower barrier height and hence larger tunnel probability, would allow for thicker insulating layers to be used. Therefore, a large current could be supported without the need for a prohibitively large electric field across the insulating layer.

As the high current mode of the poly gate devices seems to be relatively independent of a large class of parameters, it seems likely that the current transport mechanism is far from one-dimensional due to the lack of confinement of carriers in the silicon epi layer (see Figure 4.16). If this is the case, this mode of operation may take over when the current reaches some critical value, even if stress has not forced the change from the low current mode to the high current mode. This might explain why some of the thinner oxide devices switch from the low to high modes even before the low current

mode can reach its high current state to stress the oxide (see Figure 4.13(c),(d)).

It is not clear why the Al gate devices exhibited somewhat different behavior from the poly gate devices. Note, however, that the Al gate device behavior seems to be much like the high current mode of the poly gate devices in many respects. The shape of the curves is similar, the effects of stress are similar, and while the scaling properties are not quite the same, there are similarities. As some evidence has shown that Al can react with the oxide [5], it is possible that the effective thicknesses of these oxides is reduced by this reaction so that the current range of a would-be low current mode is pushed up to where the high current mode must take over.

Notice that the transition from the low current mode to the high current mode in the poly gate TSD devices is reminiscent of the transition from "leakage current" mode to "soft breakdown" observed in MOS devices, as described in Section 2.4.2. In fact, the small changes in the low current mode TSD curve with increasing stress may be a result of leakage current allowed by stress of the oxide.

While the high current mode of the poly devices, and the only mode of the Al gate devices are in many ways inferior to the low current mode, they do possess many interesting properties. In order to explore and understand these further, however, it may be necessary to fabricate devices with some sort of lateral isolation as mentioned in Section 4.3.5.

Bibliography

[1] M. Depas, T. Nigam, and M. M. Heyns, *IEEE Trans. Electron Devices* **43**, 1499 (1996).

[2] H. Kroger and H. A. R. Wegener, *Solid–State Electron.* **21**, 643 (1978).

[3] J. G. Simmons and A. A. El-Badry, *Radio and Elec. Eng.* **48**, 215 (1978).

[4] A. C. F. D. Mattos and G. Sarrabayrouse, *Phys. Stat. Solid. (A)* **87**, 699 (1985).

[5] S. Duzelier, G. Sarrabayrouse, J. L. Prom, and G. Hollinger, *Electronics Letters* **27**, 1399 (1991).

Chapter 5 Simulations

5.1 Introduction and Outline

During the early phases of data collection, it became clear that simple descriptions of the device operation presented in the literature were ill-equipped to provide any explanations of many of the unusual results presented in the previous chapter. Therefore, in an attempt to better understand the theoretical operation of the TSD device, and in the hopes of being able to design device parameters, a TSD device simulation was developed. This chapter describes the development of a model capable of simulating the physical processes relevant to the device behavior and the preliminary results obtained with it.

Given the success and flexibility of the drift-diffusion model in the description of semiconductor devices, it was decided that this would serve as the basis for the TSD device simulations. This model and the details of its implementation are described in Section 5.2. In order to account for the tunnel current in the device, the standard drift-diffusion formalism had to be extended to allow for tunnel currents at a boundary. Section 5.3 describes the tunnel current model used, and Section 5.4 describes the union of the two models. The results of simulations of TSD structures and some simpler, related structures are then presented in Section 5.5.

5.2 Drift-Diffusion Simulations

5.2.1 Basic Semiconductor Equations

Solutions of the drift-diffusion semiconductor equations have been extensively studied and utilized for simulation of semiconductor devices over the past few decades [1–3]. While this topic has reached a level of maturity, and many well-developed techniques

are widely used, it is not usually possible to simply plug any semiconductor device parameters into a generic black box simulator. Each new device may present certain pitfalls which cannot be dealt with using generic algorithms. Therefore, while this section primarily outlines a simulation procedure which is well rooted in theory and well studied in practice, certain adjustments will be described which reflect device-specific problems.

All of the established methods are aimed at solving the following set of equations [1, 3], or some variation thereof:

$$\nabla^2\psi - \frac{q}{\epsilon}(n - p - C) = 0 \quad \text{(Poisson Equation)} \tag{5.1}$$

$$\nabla \cdot \vec{J_n} - qR(\psi, n, p) = 0 \quad \text{(Electron Current Continuity)} \tag{5.2}$$

$$\nabla \cdot \vec{J_p} + qR(\psi, n, p) = 0 \quad \text{(Hole Current Continuity)} \tag{5.3}$$

with

$$\vec{J_n} = qD_n\nabla n - q\mu_n n\nabla\psi \tag{5.4}$$

$$\vec{J_p} = -qD_p\nabla p - q\mu_p p\nabla\psi \tag{5.5}$$

where ψ is the electric potential, n and p are the electron and hole carrier concentrations, q is the electron charge, ϵ is the semiconductor dielectric constant, C is the net concentration of ionized dopants ($= N_d^+ - N_a^-$), $\vec{J_n}$ and $\vec{J_p}$ are the electron and hole current densities, R is the net recombination rate, D_n and D_p are the electron and hole diffusion constants, and μ_n and μ_p are the electron and hole mobilities. For simulations of some devices, such as devices in which only majority carriers are important, a subset of these equations can be used. For example, the "Thomas Fermi" model results if the electron and hole quasi Fermi levels are assumed piecewise constant, and hence only Equation 5.1 is needed.

This set of equations can be extended to allow for heterojunction structures [2, 4]. While the simulations of the TSD did not require these extensions, these modified

equations were employed to allow for more general use, so we present them here:

$$\nabla^2\psi - \frac{q}{\epsilon}(n - p - C) + \frac{1}{\epsilon}\nabla\epsilon \cdot \nabla\psi = 0 \tag{5.6}$$

$$\vec{J_n} = qD_n\nabla n - q\mu_n n\nabla\psi - \nabla\chi - \frac{qD_n}{N_c}\nabla N_c \tag{5.7}$$

$$\vec{J_p} = -qD_p\nabla p - q\mu_p p\nabla\psi - \nabla\chi - \nabla E_g + \frac{qD_p}{N_v}\nabla N_v \tag{5.8}$$

where χ is the electron affinity (*i.e.*, energy difference between the vacuum level and the semiconductor conduction band), E_g is the bandgap, and N_c, N_v are the conduction and valence band effective density of states. These replace Equations 5.1, 5.4, and 5.5, while Equations 5.2 and 5.3 are maintained.

In addition to these equations, boundary conditions must be applied. Although many such conditions are possible, two will be considered here. An ideal ohmic contact held at a fixed potential is described by the following equations [3]:

$$\psi = \psi_0 + V_{appl} \tag{5.9}$$

$$n = n_0 \tag{5.10}$$

$$p = p_0 \tag{5.11}$$

where V_{appl} is the potential applied at that contact, and ψ_0, n_0, and p_0 are the values of the corresponding variables in bulk material in equilibrium (no space-charge).

An equipotential contact on top of a perfectly insulating material (*i.e.*, a metal gate on top of an oxide layer) can be described by the following set of equations, applied at the edge of the semiconductor layer [3]:

$$\vec{e_n} \cdot \vec{J_n} = -qR_s \tag{5.12}$$

$$\vec{e_n} \cdot \vec{J_p} = qR_s \tag{5.13}$$

$$\vec{e_n} \cdot (\epsilon_i\nabla\psi_i) - \vec{e_n} \cdot (\epsilon\nabla\psi) = q\sigma_i \tag{5.14}$$

where $\vec{e_n}$ is the normal direction to the surface, R_s is the surface recombination velocity,

ϵ_i is the insulator dielectric constant, and σ_i is the concentration of sheet charge at the semiconductor-insulator interface (useful for simulating oxide trapped charge).

Given the semiconductor structure (physical structure, doping profiles) and materials properties (bandgaps, mobilities, recombination parameters, etc.), these equations can be solved for (ψ, n, p) throughout the region of interest. From these variables, other quantities of interest can be computed, such as currents, electric fields, and band diagrams. The equations can be recast in terms of other variable sets, such as (ψ, E_n, E_p) where E_n and E_p are the electron and hole quasi Fermi levels, but the (ψ, n, p) variable set seems to give good computational results for a broader class of devices and bias situations [1, 5].

Several other relations are needed to carry out the simulations. The ionized dopant concentrations can be computed as follows [4]:

$$N_a^- = \frac{1}{1 + 4\exp[(E_a - E_p)/k_bT]} \tag{5.15}$$

$$N_d^+ = \frac{1}{1 + 2\exp[(E_n - E_d)/k_bT]} \tag{5.16}$$

where E_a, E_d are the acceptor and donor energy levels. The quasi Fermi levels can be computed from the carrier concentrations using [1]

$$E_n - E_c = k_bT\ \mathcal{F}_{1/2}^{(-1)}\left(\frac{n}{N_c}\right) \tag{5.17}$$

$$E_v - E_p = k_bT\ \mathcal{F}_{1/2}^{(-1)}\left(\frac{p}{N_v}\right) \tag{5.18}$$

where $\mathcal{F}_j$ is the Fermi integral of order j, defined as [6]

$$\mathcal{F}_j = \frac{1}{\Gamma(j+1)}\int_0^\infty \frac{\epsilon^j d\epsilon}{1 + \exp(\epsilon - \eta)} \tag{5.19}$$

Also, the modified Einstein's relation was used to compute diffusion constants from mobilities [1, 4, 7]:

$$\frac{D}{\mu} = \frac{k_bT}{q}\frac{\mathcal{F}_{1/2}(\eta)}{\mathcal{F}_{-1/2}(\eta)} \tag{5.20}$$

where η is $(E_n - E_c)/k_bT$ for electrons and $(E_v - E_p)/k_bT$ for holes.

While more general, these relations employing Fermi-Dirac statistics can present computational difficulties. For example, not only must numerical approximations [6, 8] for $\mathcal{F}_{1/2}$ and $\mathcal{F}_{-1/2}$ be used, but the approximations must obey the relation $\mathcal{F}'_{1/2} = \mathcal{F}_{-1/2}$ essentially exactly, or else the resulting equations are not self-consistent and may fail to converge. Perhaps more seriously, typical expressions for recombination (see below) do not employ Fermi-Dirac statistics, breaking down under degenerate conditions as $np = {n_i}^2$ fails to hold. Therefore, if Fermi-Dirac statistics are used elsewhere, the equation set is again not self-consistent. In fact, as no good recombination models employing Fermi-Dirac statistics were found, most of the results presented here used a less precise, but self-consistent set of expressions employing Boltzmann statistics. Even in degenerate conditions, Boltzmann statistics give correct trends even if the values are not completely precise, so simulations entering this regime can still be useful. In the future, it is hoped that a self-consistent set of equations employing Fermi-Dirac statistics can be developed for more general use.

5.2.2 Models of Physical Parameters

As a great deal of effort has been expended in the development of Drift-Diffusion modeling of silicon devices, quite accurate models of the physical parameters of silicon have been developed. Therefore, the following models and parameters were simply adapted from the literature [1, 3].

Mobilities were represented using the following equations:

$$\mu = \mu(C, \bar{c}, \mathcal{E}) = \mu(C, \bar{c})[1 + (\mu(C, \bar{c})\,|\mathcal{E}|\,v_{max})^{\beta}]^{-1/\beta} \tag{5.21}$$

where

$$\mu(C, \bar{c}) = \mu_{min} + \frac{\mu_{max} - \mu min}{1 + (N_T/N_{ref})^{\alpha}}$$

$$N_T = 0.34(N_d + N_a) + 0.66(\bar{c}), \quad \bar{c} = n + p$$

Table 5.1: Silicon mobility parameters (adapted from [3]).

	μ_{max} cm^2/V-s	μ_{min} cm^2/V-s	N_{ref} cm^{-3}	v_{max} cm/s	α –	β –
Electrons	1360	92	1.3×10^{17}	1.1×10^{7}	.91	2
Holes	520	65	2.4×10^{17}	9.5×10^{6}	.61	1

where $\mathcal{E}$ is the electric field. The parameter values for electron and hole mobilities are given in Table 5.1.

The recombination rate, R, was split into three parts: Shockley-Reed-Hall recombination (R_{SRH}), Auger recombination (R_{Aug}), and avalanche generation (G_{Av}). Direct band recombination was ignored, as it is negligible compared to these other processes in silicon.

The SRH recombination was calculated as follows:

$$R_{SRH} = \frac{np - {n_i}^2}{\tau_p(n + n_t) + \tau_n(p + p_t)} \tag{5.22}$$

with

$$n_t = N_c \exp[(E_t - E_c)/k_bT],$$
$$p_t = N_v \exp[(E_v - E_t)/k_bT].$$

where τ_n, τ_p are the electron and hole recombination lifetimes and E_t is the intermediate state trap level. For these calculations, we chose $n_t = p_t = n_i$, corresponding to a trap energy roughly midgap – the most effective trap level. The recombination lifetimes were chosen to be adjustable parameters, constant throughout the region of interest. Formulas for computing the recombination lifetimes as a function of dopant and carrier concentration exist, but were not used.

Auger recombination was described by

$$R_{Aug} = (C_n n + C_p p)(np - {n_i}^2) \tag{5.23}$$

Table 5.2: Silicon avalanche generation parameters (adapted from [3]).

	α_∞ (cm^{-1})	b (V/cm)
Electrons	7.03×10^5	1.231×10^6
Holes	1.582×10^6	2.036×10^6

with $C_n = 2.8\times10^{-31}$ cm^6/s, $C_p = 9.9\times10^{-32}$ cm^6/s. This term was typically much smaller than the SRH term.

Avalanche generation was considered using the following:

$$G_{Av} = \alpha_n \|\vec{J}_n\| + \alpha_p \|\vec{J}_p\| \tag{5.24}$$

with

$$\alpha = \alpha_\infty \exp(-b/\,|\mathcal{E}|)$$

where the parameters used are listed in Table 5.2.

5.2.3 Discretization of Equations

In order to solve the above set of equations numerically, the region of interest is divided up into a finite number of mesh points, and the differential equations are discretized, typically using a modified finite difference method. Then, the variables of interest are calculated at each mesh point by solving the resulting algebraic equations. Care must be taken in applying the discretization, as a simple finite difference scheme will produce nonideal results. To see this, note that a simple finite difference applied to a variable effectively assumes a linear dependence of this variable between mesh points. A linear change in ψ produces exponential changes in n, p, however, so the discretized equations would be far from self-consistent if this difference scheme were applied to all variables, unless the mesh spacing is made quite small [1, 3]. The widely used Scharfetter-Gummel discretization scheme [1–3] alleviates this problem by assuming a linear dependence of ψ between mesh points, and calculating the enforced

dependence of the carrier concentrations, given the analytical differential equations. Details of this method are not outlined here as the methods presented in [1] were followed closely. Slight modifications were made to allow for Fermi-Dirac statistics to be used, if desired, and for the heterojunction terms to be included. Only one-dimensional behavior was considered, so the mesh consisted of a set of non-equally spaced points along a given axis. The resulting one-dimensional discretized equations are presented here.

A finite difference scheme applied to Equation 5.6, assuming a constant electric displacement ($D = \epsilon\mathcal{E}$) between mesh points, produces the following result:

(5.25)

$$\begin{aligned} F_\psi\Big|_i \equiv & \frac{2}{(x_{i+1} - x_{x-1})}\left[\left(\frac{\psi_{i+1} - \psi_i}{x_{i+1} - x_i}\right)\left(\frac{2\epsilon_i\epsilon_{i+1}}{\epsilon_i + \epsilon_{i+1}}\right) - \left(\frac{\psi_i - \psi_{i-1}}{x_i - x_{i-1}}\right)\left(\frac{2\epsilon_i\epsilon_{i-1}}{\epsilon_i + \epsilon_{i-1}}\right)\right] \\ & + q(p_i - n_i + C_i) \\ & = 0 \end{aligned}$$

where the x_i are the spatial coordinates of the mesh points, and all variables dependent on these coordinates are notated as $f(x_i) \equiv f_i$. Note that in the case that all the ϵ_i are the same, this expression reduces to the result of a simple finite difference applied to Equation 5.1.

The Scharfetter-Gummel scheme applied to the current continuity equations (Equations 5.2, 5.3, 5.7, 5.8) produces the following:

(5.26)

$$F_{J_n}\Big|_i \equiv \frac{J_n|_{i+\frac{1}{2}} - J_n|_{i-\frac{1}{2}}}{(x_{i+1} - x_{i-1})/2} - qR\Big|_i = 0,$$

(5.27)

$$F_{J_p}\Big|_i \equiv \frac{J_p|_{i+\frac{1}{2}} - J_p|_{i-\frac{1}{2}}}{(x_{i+1} - x_{i-1})/2} + qR\Big|_i = 0,$$

with

(5.28)

$$J_n\Big|_{i+\frac{1}{2}} = q\frac{D_n|_{i+1} + D_n|_i}{2}\left(\frac{B(u_{i+\frac{1}{2}})n_{i+1} - B(-u_{i+\frac{1}{2}})n_i}{x_{i+1} - x_i}\right),$$

$$u_{i+\frac{1}{2}} = \frac{\mu_n}{D_n}[(\psi_{i+1} - \psi_i) + (\chi_{i+1} - \chi_i)] + (\log N_c|_{i+1} - \log N_c|_i),$$

(5.29)

$$J_p\Big|_{i+\frac{1}{2}} = q\frac{D_p|_{i+1} + D_p|_i}{2}\left(\frac{B(v_{i+\frac{1}{2}})p_i - B(-v_{i+\frac{1}{2}})p_{i+1}}{x_{i+1} - x_i}\right),$$

$$v_{i+\frac{1}{2}} = \frac{\mu_n}{D_n}[(\psi_{i+1} - \psi_i) + (\chi_{i+1} - \chi_i) + (E_g|_{i+1} - E_g|_i)] - (\log N_v|_{i+1} - \log N_v|_i),$$

where $B(x)$ is the Bernoulli function

(5.30)

$$B(x) = \frac{x}{e^x - 1}$$

and the quantity (μ_n/D_n) is evaluated using the Einstein relation (Equation 5.20). The quantities $J_n|_{i-\frac{1}{2}}$ and $J_p|_{i-\frac{1}{2}}$ can be obtained by replacing i with $i-1$ in Equations 5.28 and 5.29. While this set of equations is formidable, a few comments may clarify their origin.

The attempt to keep the set of equations self-consistent causes the appearance of the Bernoulli functions. They allow the carrier concentrations to vary exponentially between mesh points. In the limit that the argument is small, the Bernoulli function reduces to $B(x) \approx 1 - x/2$, and in this limit, the above equations reduce to those obtained from a simple finite difference scheme. Ignoring the heterostructure terms for the moment, and using the nondegenerate form of the Einstein relation, the arguments of these functions are of the form $q(\psi_{i+1} - \psi_i)/k_bT$. Thus, when the potential does not change much compared to k_bT/q between mesh points, the equations reduce to those obtained from a simple treatment. Conversely, a simple discretization would require mesh points for every k_bT/q change in the potential in order to remain accurate

– a fairly stringent requirement, particularly for multidimensional grids, or for low temperature simulations. The result is that the more advanced treatment allows for many fewer mesh points to be used, particularly in regions of a device which truly do have a nearly constant electric displacement, as in the depletion region of a p-n junction or a heterojunction.

The mesh points were manually chosen for each device. Automatic mesh generation [1, 5], and mesh adaption [5, 9] were considered, but not implemented as the necessary work seemed to outweigh the benefits, given the simple one-dimensional geometry.

5.2.4 Solution of Discretized Equations - Relaxation Method

A pictorial description of a single-variable relaxation method is presented in Figure 5.1. The goal is to find a root, z^*, of the function $f(z)$ shown. If some initial approximation, z_j, is given, a "better" approximation, z_{j+1}, can be obtained by calculating the derivative of the function at the approximate solution, and extrapolating along the tangent to the curve to the axis. The solution to the following equation carries out this procedure:

$$-f(z_j) = f'(z_j)(z_{j+1} - z_j) \tag{5.31}$$

Iteration of this procedure produces a series of approximate solutions which will converge to the true solution, provided the initial estimate is close enough [10].

The relaxation method used to solve the drift-diffusion equations above is essen-

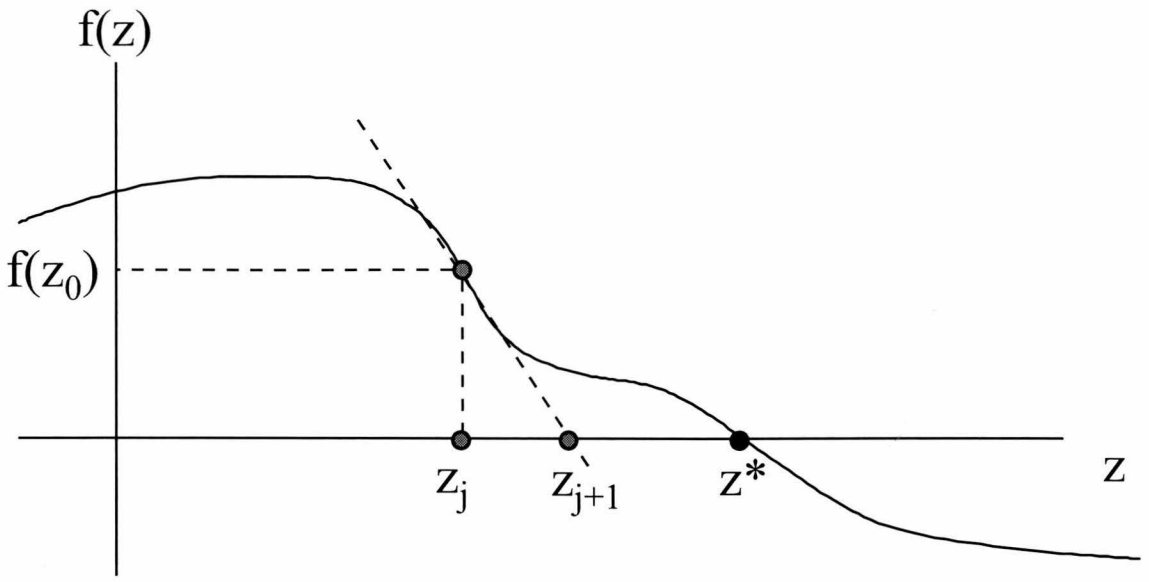

Figure 5.1: Graphical representation of the relaxation method for a function of a single variable, where z^* is the true solution of $f(z) = 0$, z_j is the first approximation to the solution, and z_{j+1} is a refined approximation.

tially identical, with the following transformations:

$$z \rightarrow \vec{z} \equiv [\psi, n, p]^{\mathrm{T}}, \psi = \{\psi_i\}, n = \{n_i\}, p = \{p_i\} \tag{5.32}$$

$$f(z) \rightarrow \mathbf{F}(\vec{z}) \equiv \begin{bmatrix} F_\psi(\vec{z}) \\ F_{Jn}(\vec{z}) \\ F_{Jp}(\vec{z}) \end{bmatrix} \tag{5.33}$$

$$f'(z) \rightarrow \mathbf{F}'(\vec{z}) \equiv \begin{bmatrix} \frac{\partial F_\psi(\vec{z})}{\partial \psi} & \frac{\partial F_\psi(\vec{z})}{\partial n} & \frac{\partial F_\psi(\vec{z})}{\partial p} \\ \frac{\partial F_{J_n}(\vec{z})}{\partial \psi} & \frac{\partial F_{J_n}(\vec{z})}{\partial n} & \frac{\partial F_{J_n}(\vec{z})}{\partial p} \\ \frac{\partial F_{J_p}(\vec{z})}{\partial \psi} & \frac{\partial F_{J_p}(\vec{z})}{\partial n} & \frac{\partial F_{J_p}(\vec{z})}{\partial p} \end{bmatrix} \tag{5.34}$$

$$\Rightarrow \qquad -\mathbf{F}(\vec{z_j}) = \mathbf{F}'(\vec{z_j})(z_{j+1} - z_j) \tag{5.35}$$

If there are N mesh points, the problem now requires iterated solutions of 3N non-linear algebraic equations, represented by the matrix equation in 5.35. Again, this iteration will converge, provided that the initial estimate of the solution variables is close enough [1]. Each step in this iteration is called a "Newton Step" as this method is essentially the Newton-Raphson root finding algorithm [10].

It turns out that the iteration can still converge, sometimes more quickly, if a modified version of the matrix $\mathbf{F}'(\vec{z})$ is used. The simplest modification is to multiply this matrix by a constant factor (amounting to taking a fraction of the step along the tangent line in Figure 5.1). Several iterative methods use a large factor early on when the approximate solution is further away, reducing this factor to one once the approximations improve [1, 10–12]. More complicated modifications can be used, such as the zeroing of parts of the matrix. One such method, the Gummel method [13], has been widely used. It is of particular interest in that it effectively splits the 3N×3N matrix problem into three successive N×N ones, reducing the memory and processing requirements necessary to solve the problem. For this reason, it was extensively used years ago when computational speed was orders of magnitude slower, and is still used in some form or another in multidimensional simulations in which many mesh points are used [14, 15], but it often does not converge well in high bias conditions. The Newton method, modified by the multiplier technique mentioned above, has been shown to converge much more universally [5]. Because only a single spatial dimension was considered in this work, computational power was not an issue, so a Newton method with a primitive multiplier algorithm was used [10].

Note that, since the matrix used in Equation 5.35 does not need to be the exact derivative of $\mathbf{F}(\vec{z})$, the functional dependence of many parameters can be ignored. For example, the model used for the mobilities (Equation 5.21) depends on ψ, n, p, but the dependence is relatively weak except when large electric fields are present. Therefore, this dependence can often be ignored in computing the derivatives, allowing for simpler equations and quicker computation. Therefore, in our calculations, many such simplifications were used, primarily in the treatment of mobilities and recombination rates.

The initial values needed for the relaxation technique were provided as follows. For zero bias conditions, the carrier concentrations and potential were set to equilibrium bulk values – the same as those used for ohmic contact boundary conditions (Equations 5.9, 5.10, 5.11). The results from the zero bias calculation were then used for a subsequent calculation in which a small bias was applied. By iterating this procedure,

the device bias could be stepped up to any desired value.

5.2.5 Matrix Methods

Through the use of the discretization methods above, the simulation has been reduced to iterative solving of Equation 5.35, which is of the form $\mathbf{A}\vec{x} = \vec{b}$, where $\mathbf{A}$ and $\vec{b}$ are known and $\vec{x}$ is to be determined. As many package routines are available for solving such equations, it appears that this should be a fairly straightforward task. Attempts to use routines from one such package [1] failed for nearly all test structures, however. In order to account for and correct this problem, the concepts of matrix conditioning and of the use of pivoting in solution methods had to be understood.

For any matrix $\mathbf{A}$ and vectors $\vec{x}, \vec{b}, \delta\vec{x}, \delta\vec{b}$ which satisfy $\mathbf{A}\vec{x} = \vec{b}$ and $\mathbf{A}(\vec{x} + \delta\vec{x}) = (\vec{b} + \delta\vec{b})$, it can be shown that

$$\frac{\delta\vec{x}}{\vec{x}} \leq \|\mathbf{A}\|\|\mathbf{A}^{-1}\|\frac{\delta\vec{b}}{\vec{b}} \tag{5.36}$$

for suitable definitions of the vector ratios and the matrix norm (see Appendix B for more details). The quantity $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ is called the "condition number" of the matrix, and is a measure of how close to singular the matrix is – a large condition number indicates a near-singular matrix.

Now, consider solving the matrix equation for $\vec{x}$. If the condition number is large, any small error in $\vec{b}$ could lead to a much larger error in $\vec{x}$, even if the solution algorithm is perfect. As the semiconductor equations involve variables of widely different orders of magnitude, it turns out that the condition number is often very high – values of 10^{44}-10^{50} were not uncommon. This means that, even if $\vec{b}$ is known to machine precision, $\sim 10^{-17}$ (the best possible case for an 8 byte double precision number), the solution calculated by a perfect solver could have a relative error as great as 10^{33}! This clearly leads to unacceptable solutions.

This difficulty can be reduced by "preconditioning" the matrix, *i.e.*, scaling the matrix in an attempt to reduce the condition number, usually by trying to scale all

[1] LAPACK linear algebra package routines DGBSV, DGBSVX

matrix entries and vector components to roughly the same order of magnitude. This is accomplished by choosing appropriate matrices $\mathbf{R}$, $\mathbf{C}$ and transforming the equation to

$$\mathbf{A}^*\vec{x}^* = \vec{b}^*, \tag{5.37}$$

where

$$\mathbf{A}^* = \mathbf{RAC}^{-1}, \tag{5.38}$$

$$\vec{x}^* = \mathbf{C}\vec{x}, \tag{5.39}$$

$$\vec{b}^* = \mathbf{R}\vec{b}. \tag{5.40}$$

If $\mathbf{R}, \mathbf{C}$ are diagonal matrices, this amounts to simple scaling of the rows and columns of the algebraic equations. Several successive conditioning steps can be used if necessary. While many possible choices of $\mathbf{R}, \mathbf{C}$ are possible [16, 17], the simple conditioning method used for these simulations is presented here [18].

First the columns of the matrix and the solution variables, (ψ, n, p), are scaled. All potentials are divided by k_bT/q, and the ith electron and hole concentrations divided by the previous solutions, n_i, p_i respectively. Then, the row scaling is performed. The ith electron and hole current continuity equations are divided by the previous solutions, n_i, p_i respectively. Finally, one last row scaling is applied as follows. For each of the three blocks along the diagonal of the matrix (*i.e.*, $\frac{\partial F_\psi(\vec{z})}{\partial \psi}$ in Equation 5.35), the trace of that block is calculated, and the rows of the full matrix which contain the block are scaled by the result of the trace. This method reduced condition numbers from roughly 10^{44}-10^{50} down to 10^6-10^{10}, producing much better results.

Although conditioning of the matrix improved convergence of the simulations dramatically, a change in the solution algorithm was perhaps even more effective. In order to understand this, first note that many of the entries in the matrix in Equation 5.35 are zero, as the equations only couple variables on a given mesh point with those corresponding to its two nearest neighbors. Thus, with a proper arrangement of rows

and columns, the matrix can be written in block tridiagonal form:

$$\begin{bmatrix} \mathbf{B}_1 & \mathbf{C}_1 & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{A}_2 & \mathbf{B}_2 & \mathbf{C}_2 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_3 & \mathbf{B}_3 & \mathbf{C}_3 & \dots & \mathbf{0} \\ \vdots & & & & \ddots & \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{A}_{N-1} & \mathbf{B}_{N-1} & \mathbf{C}_{N-1} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{A}_N & \mathbf{B}_N \end{bmatrix} \tag{5.41}$$

where each of the elements is a 3×3 matrix.

The LAPACK package routine used initially assumed a band diagonal form, and solved the matrix equation using an LU factorization technique [19, 20]. While a band tridiagonal matrix can also be considered band diagonal, this is in a sense too rigid a classification, as many of the band diagonal entries will be zeroes. As the band diagonal solution routine attempts to preserve the band diagonal geometry, the restrictions imposed are therefore more stringent than necessary. In particular, the use of pivoting (*i.e.*, swapping of rows or columns to put large elements on the matrix diagonal) is extremely limited in band diagonal routines. As pivoting is extremely important to the stability of many matrix algorithms [10], this restriction can reduce the effectiveness of the algorithm. Therefore, a routine employing Gaussian elimination on each row of blocks, using both row and column pivoting, was created and implemented. The new routine was found to be sufficiently robust that, in many cases, the preconditioning mentioned above was not needed as the pivoting process intrinsically conditioned the matrix.

5.2.6 Tests and Results

The drift-diffusion simulations were applied to a number of simple structures, such as bulk silicon "resistors" and *p*-*n* junctions in order to confirm the validity of the simulations. The results for a test *p*-*n* junction structure are shown in Figure 5.2. The calculated *I*-*V* curve is compared with the Shockley diode equation, modified for

short semiconductor regions [7]:

$$j = q{n_i}^2\left(\frac{D_p}{t_p {N_d}^+} + \frac{D_n}{t_n {N_a}^-}\right)\left(\exp\left(\frac{qV}{k_b T}\right) - 1\right) \tag{5.42}$$

where t_p and t_n are the thicknesses of the p and n flat band regions, replacing the electron and hole diffusion lengths used in the original equation. This substitution must be made when the thickness of the p and n regions becomes less than the diffusion length. The calculated I-V curve quantitatively matches this result with no adjusted parameters in medium range forward bias, as expected. The calculated reverse bias and small forward bias currents are much larger than those obtained from the Shockley equation due to generation/recombination currents in the depletion layer. The calculated large forward bias current drops below the Shockley result due to high-injection effects and the impedance of the silicon layers. These deviations correlate well with those described in the literature [21]. Band diagrams are also shown in the figure for a few bias conditions. Again, these display the expected behavior. The sharp drop in the quasi Fermi levels near the contacts under forward bias are a result of the linear drop in minority carrier concentration in the short semiconductor regions [7].

5.3 Tunnel Current Simulations

5.3.1 General Formalism

While a vast number of publications have been dedicated to the description of tunnel currents in solids, many of these specifically aimed at MOS structures, the number of different variations on the treatment of this subject seems almost as large. Certainly, the same physics underlies them all, but differences in implementation seem to be non-negligible, particularly in obtaining tunnel current densities from transmission coefficients. Therefore, the treatment used for these calculations, drawn from an apparently self-consistent subset of the available literature [22–25], will be presented here. This treatment assumes independent particle tunneling (*i.e.*, no collective effects, as in superconductors), no inelastic processes, and a band model of electronic states

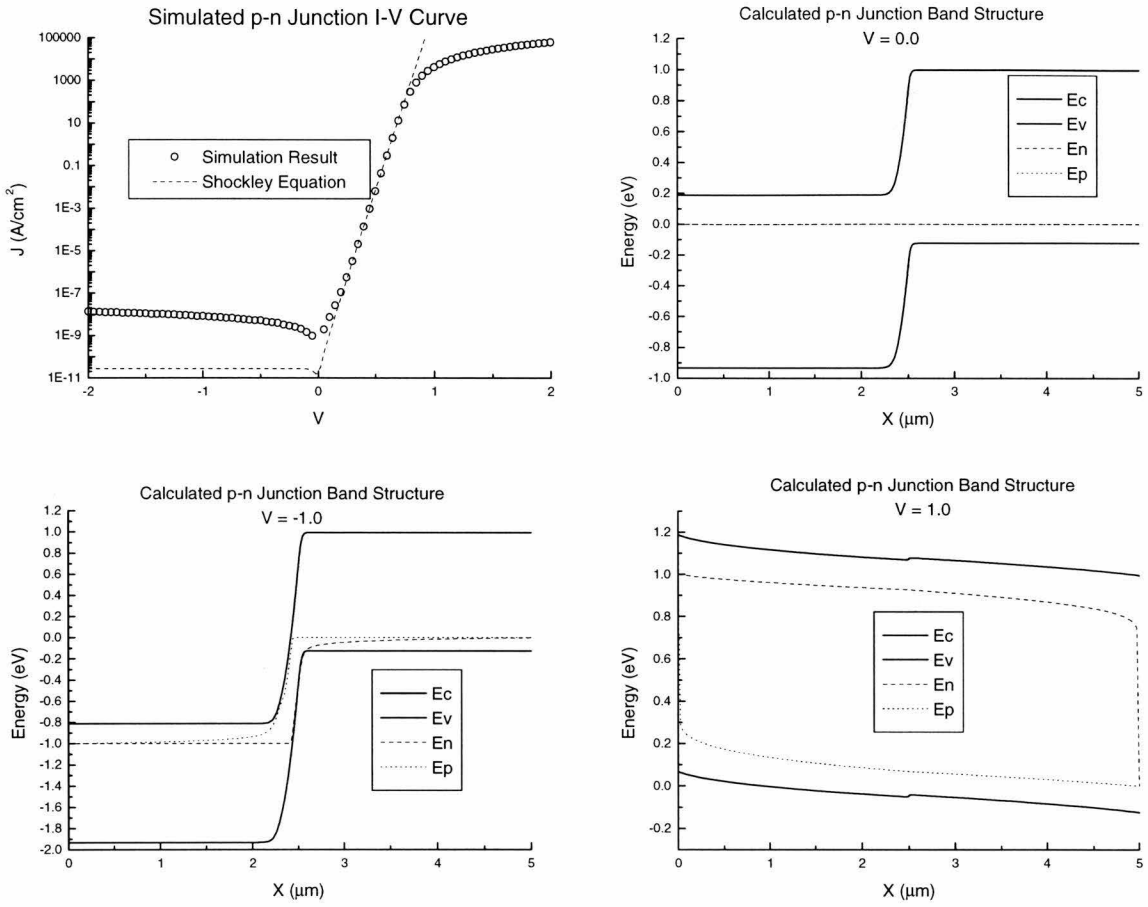

Figure 5.2: Simulated I-V curve and band diagrams of a short p-n junction with N_d = $2{\times}10^{16}$, N_a = $1{\times}10^{17}$, lifetimes = 1 μs. (a) The simulated I-V curve is compared to the Shockley diode equation for this structure, modified for short diffusion regions. Corresponding band diagrams for (a) 0.0 V, (b) -1.0 V, and (c) 1.0 V bias.

in each layer of material.

Consider tunneling from one electrode on the left side of an insulating layer to another electrode on the right side. For each available electronic band in a solid, the density of states in **k**-space is given simply by

$$dn = \frac{d^3\vec{k}}{(2\pi)^3}. \tag{5.43}$$

Therefore, the number of electrons in the left electrode with velocity along a given (x) axis between v_x and $v_x + dv_x$, and with perpendicular wavevector components between $\vec{k}_{\|}$ and $\vec{k}_{\|} + d\vec{k}_{\|}$, is given by

$$dn(v_x) = \frac{d^2\vec{k}_{\|}dk_x}{(2\pi)^3} f(E - E_L) \tag{5.44}$$

where $f(E - E_L)$ is the Fermi-Dirac distribution function for an energy of E and a chemical potential of E_L, and the dependence of the right-hand side on v_x, dv_x is contained implicitly in E, k_x.

The contribution of electrons within the same volume of phase space to the electric current moving from the left electrode to the right can then be obtained by multiplying the total charge of these electrons by their velocity, v_x, and by the probability that such an electron could make the trip to the right electrode. This probability is given by the transmission probability , $T(\vec{k})$, times the probability of an open state at the appropriate energy in the right electrode. Therefore, this current contribution is given by

$$dj_{L\to R} = qv_x dn(v_x)[1 - f(E - E_R)]T(\vec{k}). \tag{5.45}$$

Integrating over all states throughout the Brillouin zone to get the total current, we obtain

$$j_{L\to R} = q\int_0^{k_{xmax}} \frac{dk_x}{2\pi} \int_{BZ} \frac{d\vec{k}_{\|}}{(2\pi)^2} v_x f(E - E_L)[1 - f(E - E_R)]T(k_x, \vec{k}_{\|}) \tag{5.46}$$

where BZ represents the $\vec{k}_{\|}$ range of the Brillouin zone. Using the fact that $v_x = \frac{1}{\hbar}\frac{dE}{dk_x}$, and changing variables from $\{k_x, \vec{k}_{\|}\}$ to $\{E, \vec{k}_{\|}\}$, yields

$$j_{L\to R} = \frac{q}{h}\int_0^\infty dE f(E-E_L)[1-f(E-E_R)]\int_{BZ}\frac{d\vec{k}_{\|}}{(2\pi)^2}T(E,\vec{k}_{\|}). \tag{5.47}$$

An analogous argument produces a similar expression for the current flowing from the right to the left electrode. Summing these two expressions, we obtain

$$j_{TOT} = \frac{q}{h}\int_0^\infty dE[f(E-E_L)-f(E-E_R)]\int_{BZ}\frac{d\vec{k}_{\|}}{(2\pi)^2}T(E,\vec{k}_{\|}). \tag{5.48}$$

For all simulations, isotropic parallel band dispersion was assumed, so the Brillouin zone integration can be taken to be over all space, and angular part of the $\vec{k}_{\|}$ integral can be performed, yielding

$$j_{TOT} = \frac{q}{h}\int_0^\infty dE[f(E-E_L)-f(E-E_R)]\int_0^\infty \frac{dk_{\|}}{2\pi}k_{\|}T(E,k_{\|}). \tag{5.49}$$

Note that this result only accounts for a single band of states, and a sum of such expressions must be used for multiple bands. In particular, a factor of two is typically included to account for spin-degenerate bands, and another factor of two is included to account for the two X valleys along the x-axis in the case of tunneling into a silicon (100) surface. It is important to realize that the dependence of the tunnel current on densities of states in the electrodes and any **k**-space selection rules are included implicitly within Equation 5.49, primarily in the second integral containing the transmission probabilities.

The integrations were performed using successive Romberg integration routines [10] over energies extending at least 10 k_bT into the bands. The details of calculations of the transmission probabilities follow.

5.3.2 Calculation of Transmission Probabilities

Most reported treatments of tunnel currents in MOS structures use some form of a WKB approximation to evaluate the transmission probabilities, using the conduction band mass in the oxide and the metal-oxide barrier height as fitting parameters [26–29]. Using this method, the calculated I-V curves can be made to match most of the basic features of measured I-V curves. However, as this treatment typically only involves oxide properties, any information about the electrodes is ignored. This causes discrepancies, primarily in the low voltage part of the I-V curve where the tunnel currents are most sensitive to electrode properties. In addition, when large electric fields are present, the validity of the WKB approximation becomes questionable. For example, one feature of Fowler-Nordheim tunneling which cannot be reproduced by a simple model is a small oscillation in the I-V curve [30, 31]. While none of these discrepancies is particularly important for many applications, such as estimation of gate leakage currents in MOSFETs, or design of EEPROM tunnel currents, a more careful treatment may be necessary to accurately describe the TSD device as the tunnel currents important to the switching mechanism seems to depend on the interplay of the carrier distribution in the silicon and the tunneling properties of the oxide.

The Quantum Transmitting Boundary Method (QTBM) [32] was chosen for evaluation of the transmission probabilities. This method involves discretization of an effective mass Hamiltonian for the electron envelope function, and solution of the discretized equations subject to plane wave boundary conditions. It has proven very useful in evaluating transmission probabilities of complex structures, such as double barrier resonant tunneling diodes, which cannot be accurately treated using simpler methods. While more computationally expensive than a WKB method, for instance, it is well within the means of modern workstations. The formalism is also much more extensible than the WKB method, allowing detailed treatment of band structure and inclusion of multiple bands, if desired [33]. A one-band version of this method is described here, taken from [33–35], with minor modifications to allow for valence bands (negative effective masses) to be used.

The goal is the solution of the one-band effective mass equation:

$$-\frac{\hbar^2}{2}\frac{d}{dx}\left(\frac{1}{m^*(x)}\frac{dF}{dx}\right)+V(x)F(x)=EF(x) \tag{5.50}$$

subject to the boundary conditions

$$F_L(x)=Ie^{i\,\mathrm{sgn}(m_L^*)k_Lx}+re^{-i\,\mathrm{sgn}(m_L^*)k_Lx} \tag{5.51}$$

$$F_R(x)=te^{i\,\mathrm{sgn}(m_R^*)k_Rx} \tag{5.52}$$

where m^* is the effective mass, $F(x)$ is the envelope function, $V(x)$ is the potential profile, E is the electron energy, k is the electron wavevector, and the subscripts L, R refer to the left and right electrodes. The $\mathrm{sgn}(m^*)$ terms in the boundary conditions are needed if negative effective masses are used.

Care must be taken in the discretization of these equations, as with the drift-diffusion equations above, so that the resulting equations provide an appropriate physical description. The details of this procedure will not be discussed here, however. The resulting equations are

$$H_{\sigma,\sigma-1}F_{\sigma-1}+\bar{H}_{\sigma,\sigma}F_\sigma+H_{\sigma,\sigma+1}F_{\sigma-1}=0 \tag{5.53}$$

where $\bar{H}\equiv H-E$ and

$$H_{\sigma,\sigma\pm1}=-\frac{\hbar^2}{d^2}\left(\frac{1}{m_\sigma^*+m_{\sigma\pm1}^*}\right), \tag{5.54}$$

$$H_{\sigma,\sigma}=V_\sigma-H_{\sigma,\sigma-1}-H_{\sigma,\sigma+1}, \tag{5.55}$$

with plane wave boundary conditions

$$F_1 = I + r, \tag{5.56}$$

$$F_2 = Ie^{i\,\mathrm{sgn}(m_L^*)k_L d}, \tag{5.57}$$

$$F_{N-1} = t, \tag{5.58}$$

$$F_N = te^{i\,\mathrm{sgn}(m_R^*)k_R d} \tag{5.59}$$

where d is the (fixed) mesh point spacing, and N is the total number of mesh points. Note that the potential must be flat across the first two and last two mesh points in order for the plane wave boundary conditions to make sense.

The boundary conditions can then be rearranged, eliminating t, r, yielding

$$F_1 - e^{i\,\mathrm{sgn}(m_L^*)k_L d}F_2 = I(1 - e^{i\,\mathrm{sgn}(m_L^*)2k_L d}), \tag{5.60}$$

$$F_{N-1}e^{i\,\mathrm{sgn}(m_R^*)k_R d} - F_N = 0 \tag{5.61}$$

The set of equations 5.53, 5.60, 5.61 constitute a tridiagonal linear system which can be solved with a standard routine for complex matrices [10]. The transmission probability can then be determined using Equation 5.58 and

$$T(E, k_\parallel) = \|t(E, k_\parallel)\|^2 \frac{v_R}{v_L} \tag{5.62}$$

where v is the group velocity given by $v = \frac{dE}{dk}$. The dependence of t on E is apparent in Equation 5.53, but the dependence on $k_\parallel$ is hidden in the definition of $V(x)$. Assuming the parallel band dispersion is parabolic, this dependence can be explicitly written as

$$V(x, k_\parallel) = V_0(x) + \frac{\hbar^2 k_\parallel^2}{2m_t^*(x)} \tag{5.63}$$

where V_0 is the potential profile of the band edge, and the transverse effective mass, $m_t^*(x)$, characterizes the parallel band dispersion in each layer.

Note that the discretized Hamiltonian is analogous to a tight-binding formulation.

Table 5.3: Parameters used in the calculation of tunnel currents in Al/SiO_2 /Si structures. The conduction band edges are normalized to that of silicon. The Si effective masses and bandgap were obtained from [7], the SiO_2 parameters and band lineups from [36], and the Al band structure parameters from [37].

	Elec.		Hole					
	m^*	m_t^*	m^*	m_t^*	$\epsilon(\times\epsilon_0)$	E_c	E_g	E_F
Si	0.9	0.19	0.5	0.5	11.7	0.0	1.12	—
SiO_2	0.30	0.30	0.30	0.30	3.9	3.22	***	—
Al	1.0	1.0	—	—	100	-5.6	—	-.03

One consequence of this is that the energy dispersion in the x-direction is given by

$$\cos(kd) = 1 - \frac{m^* E d^2}{\hbar^2}. \tag{5.64}$$

This relation must be used in computing the group velocities above. Note that the band spans an energy range determined by the mesh spacing. While a typical tight-binding treatment uses a mesh spacing equal to the atomic spacing, the resulting energy band limits the bias range for which transmission probabilities, and therefore tunnel currents, can be calculated. For this reason, a somewhat smaller mesh spacing (typically 1 Å) was used to provide a wider energy band, rather than adding contributions from multiple bands. While a multiple band treatment might be more accurate, the details of the band structure of the layers are largely unknown, particularly in the thin oxide layer, so it is not likely that much would be gained.

5.3.3 Tests and Results

Transmission probability calculations were verified on a large number of test structures, but most full current calculations were performed on $Al/SiO_2/Si$ structures with the parameters given in Table 5.3. The SiO_2 band gap was adjusted during the TSD simulations in order to obtain reasonable hole tunnel currents, but was otherwise unused. The Al dielectric constant, used in calculating the band profile, was set to an arbitrary large value to force the electric field in the Al next to the Al-SiO_2 interface to be small.

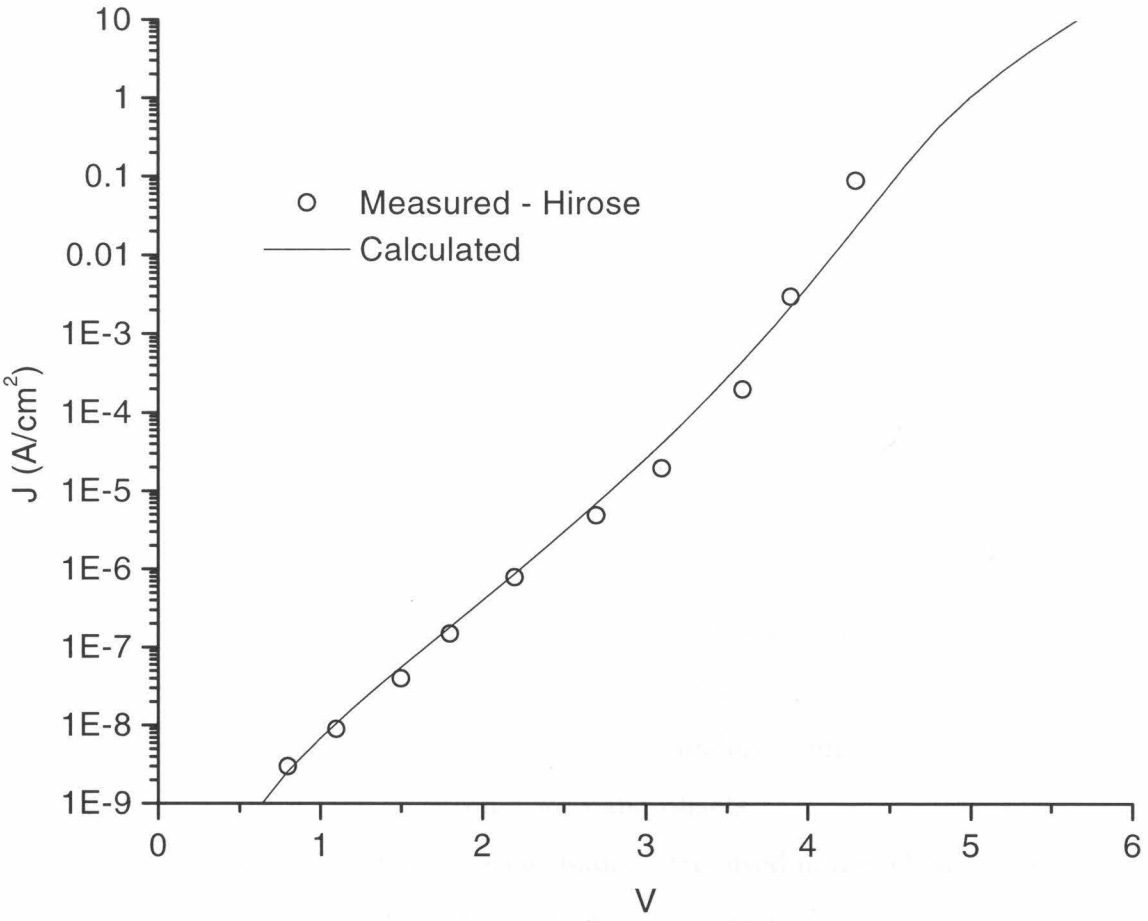

Figure 5.3: Comparison of a calculated MOS I-V curve to measured data [38].

A calculated I-V curve for a single conduction band in the Al/SiO_2/Si structure is plotted in Figure 5.3, along with experimental data from a similar n^+-poly/SiO_2/Si structure [38]. The oxide effective mass and the silicon Fermi level were adjusted to match the calculated curves roughly to the data, but a careful fit was not performed.

5.4 Union of Drift-Diffusion and Tunnel Simulations

In order to treat structures with a semiconductor section in series with a quantum transmission section, the drift-diffusion model was used to describe the semiconductor layers, and the electron and hole currents calculated by the QTBM method were imposed as boundary conditions at the edge of the drift-diffusion region. The potential profile and quasi Fermi levels used in the current calculations were obtained from the

drift-diffusion results at the interface. In order to accomplish this, modified versions of the boundary conditions found in Equations 5.12, 5.13, 5.14 were developed.

The modified equations are as follows:

$$J_n = -qR_s + J_{n\,calc} \tag{5.65}$$

$$J_p = qR_s + J_{p\,calc} \tag{5.66}$$

and

$$\psi = V_{appl} + \chi - \phi_m + \frac{t_i}{\epsilon_i}(\epsilon\nabla\psi + q\sigma_i) \tag{5.67}$$

The condition on ψ essentially sets the potential in the semiconductor to be the potential at the contact minus the change in potential across the oxide, assuming that the electric displacement is the same in the semiconductor and the insulator. The χ and ϕ_m terms are necessary to account for band offsets.

Note that in order to calculate the Jacobian matrix used in the relaxation scheme, it is necessary to compute the derivative of these equations with respect to the unknown variable set, (ψ, n, p). As the tunnel currents are calculated numerically, no analytical formula is available for computation of the necessary derivative. Therefore, tunnel currents were calculated for different electric fields and quasi Fermi levels, derivatives with respect to these variables were computed with a simple difference equation, and the desired derivatives were then obtained using the chain rule.

5.5 Results

5.5.1 Studies of *pnpn* Thyristors

One problem with the drift-diffusion formalism solved by the relaxation method is that the initial "guess" for the unknown variable set must be "close enough" to the true solution or else the iterations will not converge. This problem is compounded in highly nonlinear devices, such as the TSD, which can contain multiple solutions for

a given set of boundary conditions. It is unclear whether the iterations can converge to one solution or another in such a situation. Even if it can converge, is is unclear whether it is possible to "steer" the simulation to more than one solution for a given set of boundary conditions. While some reports of simulations of latching power devices indicate that these methods can be used with some success [9, 15], it is evident that particular attention must be paid to these complications in order to obtain reasonable results.

In the earliest attempts at simulating the TSD device, these complications, combined with the difficulties in treating the tunnel currents (which will be described in the next section), made execution of the TSD simulations intractable. Therefore, we sought a similar device to the TSD which exhibited a similar nonlinear, multistable *I-V* characteristic, and could be described by the one-dimensional drift-diffusion model without the added complications of tunnel currents. Exploration of simulations of such a device could provide some insight into simulation of multistable devices in general. One clear choice was the *pnpn* thyristor.

As the TSD *I-V* curve is similar to that of a thyristor, it is not surprising that the relevant physical mechanisms are similar in both devices. To illustrate this qualitatively, a TSD device and a *pnpn* thyristor are shown side by side in Figure 5.4. The four layers in each device serve essentially the same purpose – an injecting contact supplies electron flow which is controlled by the current limiting region, a stand-off region can be depleted in the low current state, supporting a large voltage drop, and a collector accepts the electron current and provides a supply of holes in the high current state. Therefore, taking a look at the detailed operation of the thyristor may aid in understanding the behavior of the TSD, as well as allowing study of the simulation of a highly nonlinear, multistable device.

It is perhaps easiest to present the simulation results and then to explain how they were obtained. Several simulated *I-V* curves for a fixed *pnpn* structure are plotted in Figure 5.5 for a range of Shockley-Reed-Hall recombination lifetimes, τ. For each lifetime, two curves are plotted. The black curves represent a sequence of solutions obtained by solving for the zero bias solution (as described in Section 5.2.4), and

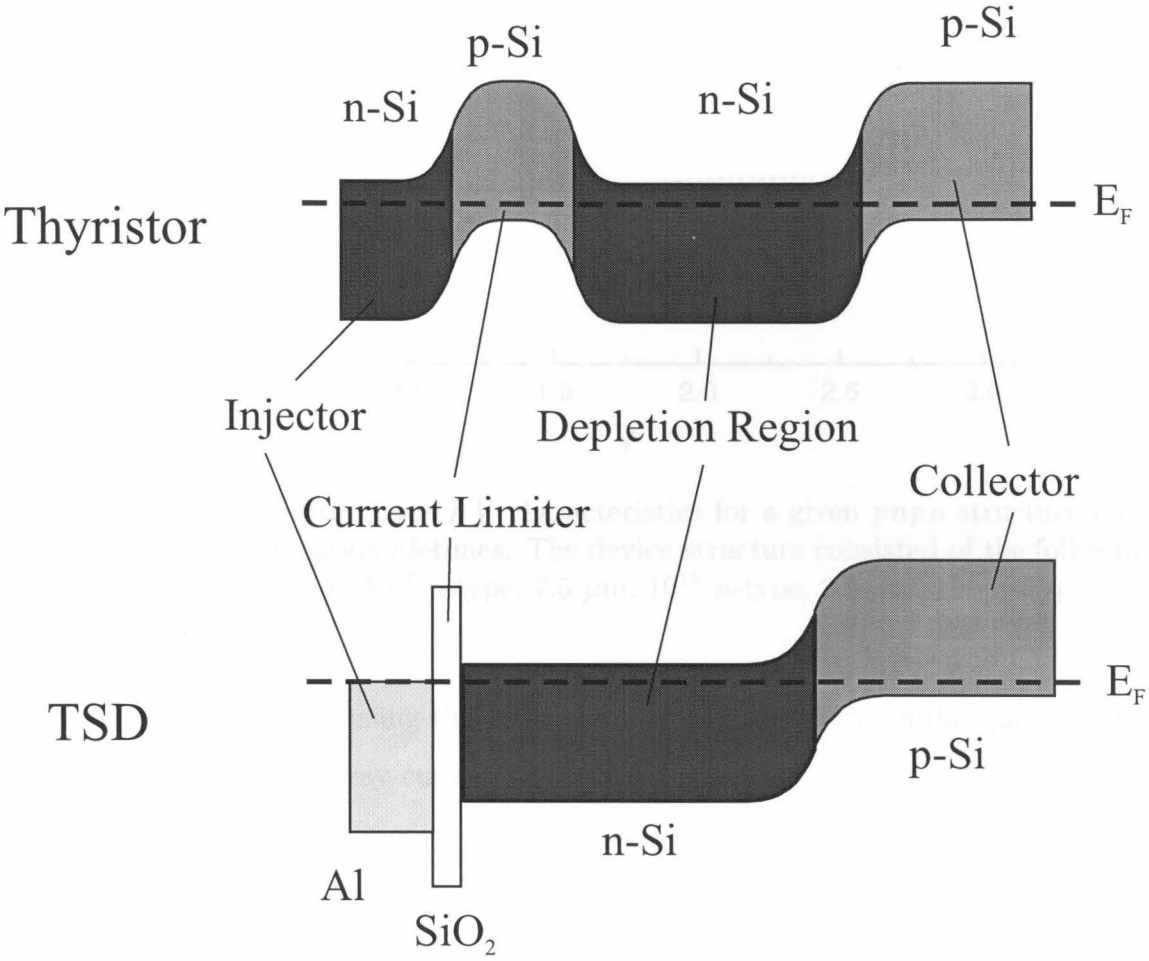

Figure 5.4: Comparison of TSD and thyristor devices.

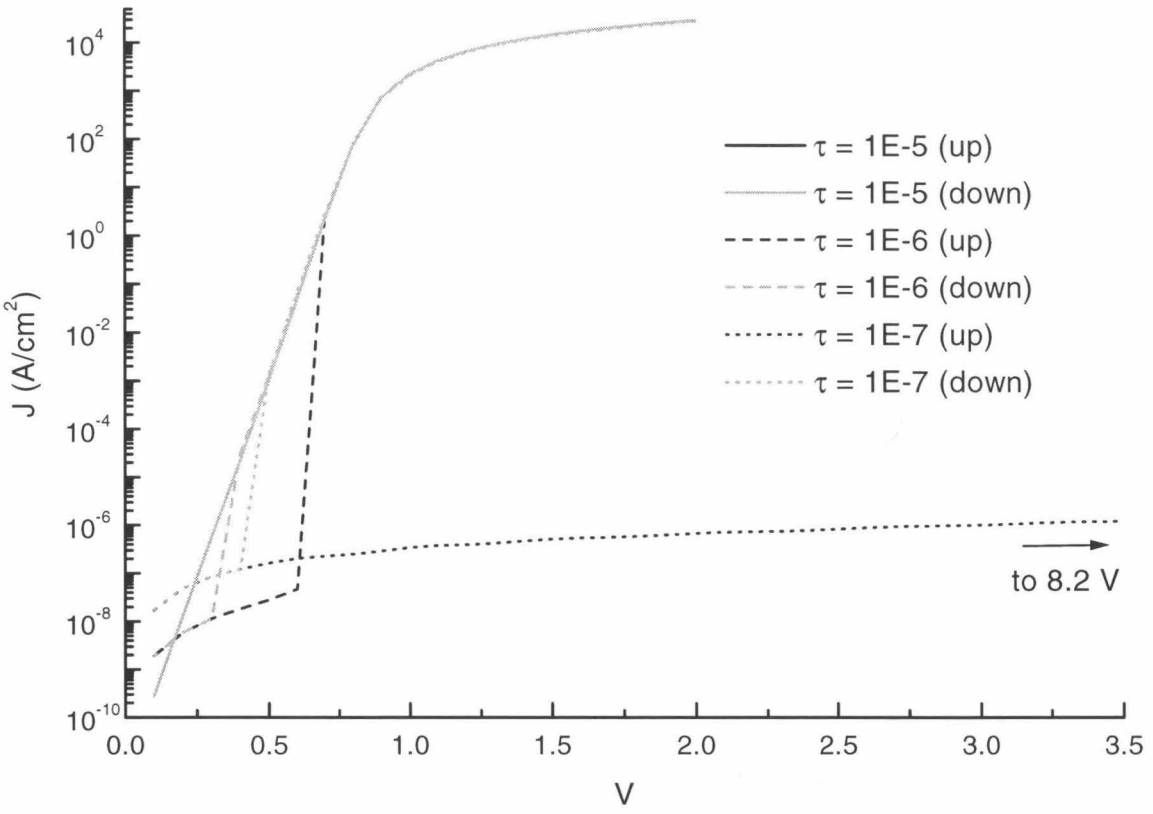

Figure 5.5: Calculated thyristor I-V characteristics for a given $pnpn$ structure for a range of SRH recombination lifetimes. The device structure consisted of the following stack of Si layers: 2.5 μm, 10^{17} p-type, 7.5 μm, 10^{15} n-type, 2.5 μm, 10^{17} p-type, and 2.5 μm, 10^{18} n-type.

stepping upward in bias, using the previous bias solution as the initial guess for the following iteration. The gray curves, on the other hand, represent a similar sequence of solutions obtained by stepping down in bias from a nonzero bias solution. In the case of $\tau = 10^{-5}$ s, the two curves are essentially identical. The $\tau = 10^{-6}$ s curves, however, exhibit some hysteresis. The increasing bias curve traces a low current path until a bias of 0.7 V is reached. Then all subsequent bias points fall on a high current path. The decreasing path traces the same high current curve down to 0.4 V, and then drops down to the low current path. Therefore, two stable current solutions were obtained for 0.4 to 0.7 V applied bias. No solutions in the negative resistance part of the I-V curve were obtained using this method, however.

The $\tau = 10^{-7}$ s curves exhibit similar behavior, but the low current part of the curve continues out to 8.2 V, leading to a complication. In this case, a further increase

in the bias condition does not result in a solution along the high current part of the curve – in fact the iterations do not converge at all. In general, no solutions in the high current part of the curve could be obtained for voltages larger than roughly 2.0 to 2.5 V. Therefore, another method had to be found to obtain a high current state solution for this lifetime. This was accomplished by using a high current state solution for $\tau = 10^{-6}$ as an initial guess for the $\tau = 10^{-7}$ iterations at the same bias condition. Once a $\tau = 10^{-7}$ solution was obtained using this method, the bias could be stepped either up or down, and the solutions would follow the high current curve.

In order to understand physically why the device I-V curve should be so sensitive to the recombination lifetime, consider the band diagrams in Figure 5.6 corresponding to different points on the $\tau = 10^{-7}$ I-V curve. Electrons are injected from the n-type layer on the far right of the structure, through the adjacent p-type layer into the rest of the structure. In the low current state plots ((b), (c), (d)), notice that the electron quasi Fermi level begins to drop just before reaching the reverse biased p-n junction, indicating some recombination. It is largely the control of this recombination in the p-type layer that determines the behavior of the rest of the device. If there is too little recombination, most of the electrons are injected into the thick n-type layer, so a depletion region cannot be supported. As a result, no low current state can exist. This effect also explains why the peak voltage (*i.e.*, the maximum voltage at which a low current state can be supported) is much lower than the punch-through voltage (87 V) for each τ considered. On the other hand, if there is too much recombination, the double-injection situation in the high current state in which carriers of both types drift across the entire structure (Figures 5.6(e), (f)) cannot be supported. As a result, no high current state can exist. This, incidentally, explains why the valley voltage (*i.e.*, the lowest voltage at which a high current state can be supported) increases with decreasing lifetime (see Figure 5.5). Therefore, the recombination lifetimes in the current limiting p-type layer must be such that the diffusion length of the injected carriers is comparable to the width of the layer in order for both high current and low current states to be supported.

As we have seen that the thyristor device behavior depends critically on the prop-

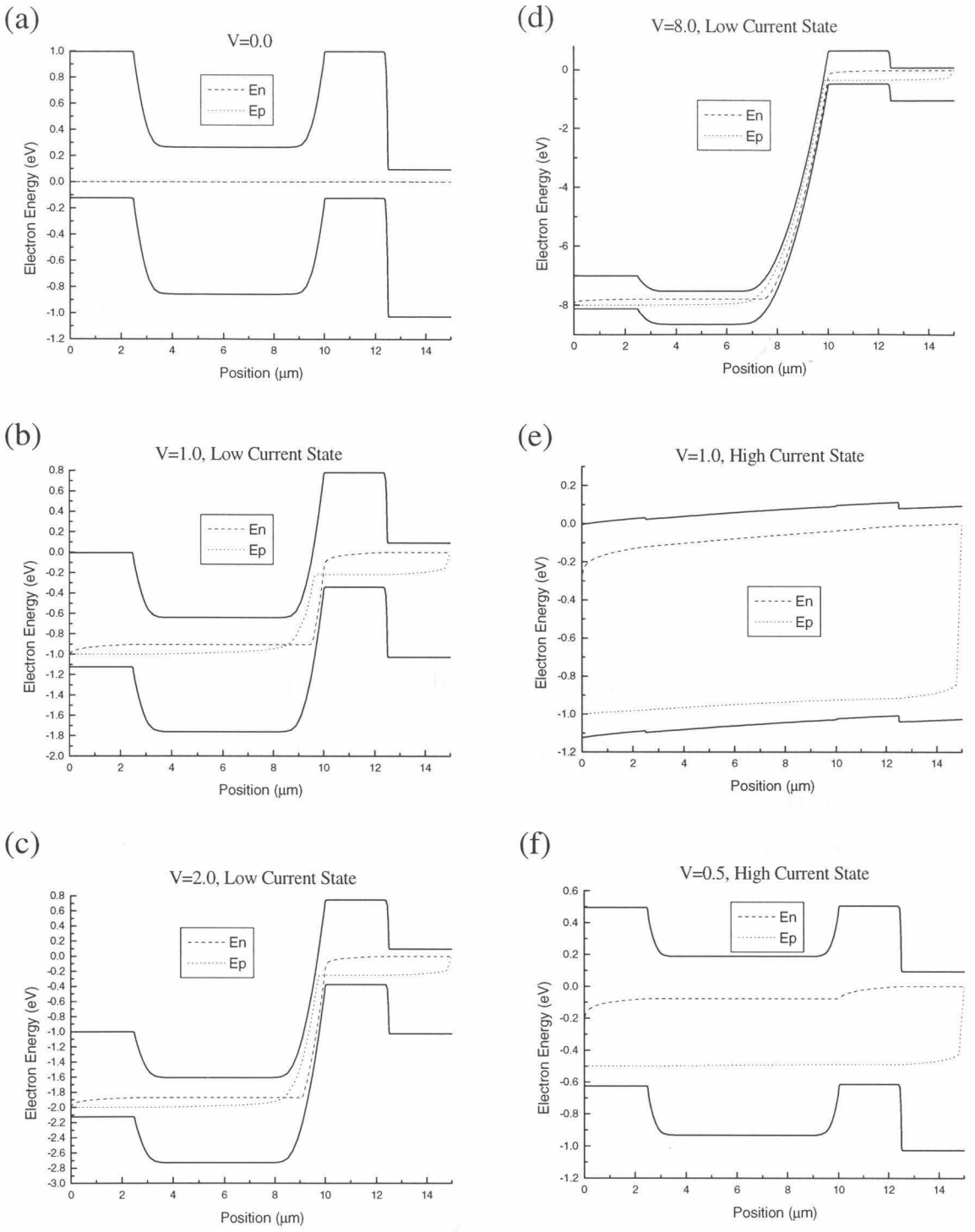

Figure 5.6: Calculated thyristor band diagrams for the $pnpn$ structure described in Figure 5.5 with $\tau = 10^{-7}$. The different diagrams represent different points on the I-V curve: (a) 0.0 V, (b) 1.0 V in the low current state, (c) 2.0 V in the low current state, (d) 8.0 V in the low current state, (e) 1.0 V in the high current state, and (f) 0.5 V in the high current state.

erties of the current limiting layer, we can expect that the behavior of the TSD will depend critically on the properties of the thin oxide layer, as already discussed in Section 2.2.2. Notice that, due to the symmetry of the $pnpn$ structure, many of the arguments applied to the transport in the p-type layer could equally well have been applied to the thick n-layer in the above discussion. For example, if the diffusion length in the thick n-layer were large compared to its thickness, carriers could be injected from the p-type contact on the left, diffusing all the way to the middle p-type layer, again preventing the support of a low current state. Therefore, the recombination properties in the silicon epilayer in the TSD will also play a role in determining device behavior. In particular, it is likely that we can use the same trick of using a long recombination lifetime in order to force the device into the high current state, then relaxing the lifetime back to a smaller value in order to obtain a high current state solution for a bistable device.

5.5.2 Studies of Thin Oxide MOS Devices

As boundary conditions for the drift-diffusion model at the silicon-oxide interface were being developed, an intermediate step was considered in which the boundary condition on the potential was maintained, but no tunnel currents were allowed – only recombination currents could flow at the interface. This allowed testing of the boundary conditions without the need for computing tunneling currents, which added significantly to the necessary computation time. Physically, this model would be accurate for thick oxides through which very little tunnel current could flow, although it could also be applied to a thin "ideal" oxide which would allow no current to flow.

A somewhat disturbing realization was made during simulation of MOS structures with these boundary conditions. While the simulations converged to the expected solutions if the MOS device was biased into accumulation (positive bias on the gate for an n-type silicon layer) or mild depletion (small negative bias on the gate for an n-type silicon layer), if the device was biased towards accumulation (larger negative bias on the gate), the iterations failed to converge. At first, it seemed that this may

have been the result of an error in the drift-diffusion equations or in the boundary conditions, but it was later brought to our attention that there are intrinsic numerical problems with simulation of a device structure such as this [39].

In order to understand the numerical problems, it is useful to consider the physical mechanisms taking place in an MOS device. Consider, for the sake of argument, an MOS device consisting of an n-type silicon layer, a thick (nonconducting) oxide, and a metal gate. Assume that the device is in steady state at the flat band bias. If a positive bias is suddenly applied to the gate electrode, electrons supplied from the n-Si layer will rush towards the oxide interface, rapidly creating an accumulation layer. On the other hand, if a negative bias is suddenly applied, an inversion layer cannot be rapidly created, as there is no nearby supply of holes. Therefore, initially, the device enters "deep depletion" – the bias drops across a depletion layer in the silicon as the negative carriers are swept out. Eventually, generation currents in this depletion layer create the holes needed to form the inversion layer at the oxide interface, but this process takes place on a relatively long time scale, dictated (in the case of silicon) by the Shockley-Reed-Hall recombination lifetime.

Numerically, this means that the part of the matrix equation describing the carriers near the oxide interface is very weakly coupled to the rest of the matrix when the MOS device is biased towards inversion. Therefore the carrier generation process gets "swamped" by the tight coupling in the rest of the matrix which is characterized by the dielectric relaxation time ($\tau = \rho\epsilon \sim 10^{-12}$ s sin silicon) [40], and the iterations fail to converge. In order to restore any hope of convergence, the coupling within the matrix must somehow be strengthened.

Perhaps the simplest way to increase the coupling in the matrix is to simply decrease the recombination lifetime in the silicon. If very short lifetimes were used ($\tau \leq 10^{-10}$ s), convergence was found to be very rapid in MOS devices over all bias ranges. For simplicity, electron and hole lifetimes were taken to be equal in all cases. While it can be argued that changing the lifetime can dramatically change the device behavior, this is not the case in an MOS device in which no current is allowed to flow. This is because, in the silicon layer, the electron and hole quasi Fermi levels must be

flat and equal if no current is allowed, so no recombination takes place.

If tunnel currents are allowed to pass through the oxide, the above argument no longer holds, so it is no longer clear whether decreasing the lifetime will have a significant impact on the solution. Luckily, the addition of tunnel currents can improve the situation somewhat. In order to understand this, consider the calculated band diagrams for MOS devices in reverse bias shown in Figure 5.7. Figures 5.7(a), (b) illustrate band diagrams for n-silicon MOS devices with 2.5 nm and 5.0 nm oxides respectively, in which no tunnel currents were allowed to flow. In order to get these simulations to converge, the recombination lifetimes were forced to be very small (10^{-19} s). A much longer lifetime (say, 10^{-10} s) could have been used, but the shorter lifetimes produce faster convergence, and the end result is the same as long as no currents are flowing. Compare these to the results for the same structures in which tunnel currents were allowed (Figures 5.7(c), (d)). In the case of the 5.0 nm oxide (Figure 5.7(d)), the tunnel currents are small enough that the band diagram is nearly identical to that in Figure 5.7(b). While a relatively small recombination time (10^{-11}) was needed to allow the simulation to converge, the band diagram was found to be essentially independent of the lifetime, as is expected, given that the quasi Fermi levels are nearly equal and flat. In the case of the 2.5 nm oxide, however, the tunnel currents are large enough that the device enters deep depletion, rather than inversion, as discussed in Section 2.4.2. While this means that a change in the lifetime will definitely affect the band diagram, as the quasi Fermi levels are no longer flat or equal, it also means that it is not necessary to generate holes as there is no inversion layer, so the problem of coupling in the matrix goes away. In fact, a much longer lifetime (10^{-4} s) was used to generate Figure 5.7(c). In addition, if hole tunneling is allowed, another channel is available for holes to enter the silicon region near the oxide if they are needed. Therefore, it appears that the matrix coupling is a problem primarily in the range of oxide thicknesses for which the lifetime can be reduced without influencing the solution, and for thinner oxides, the coupling is somewhat restored, so more reasonable lifetimes can be used. It is true, however, that the maximum usable recombination lifetimes will be dictated by the oxide thickness, so experimentation

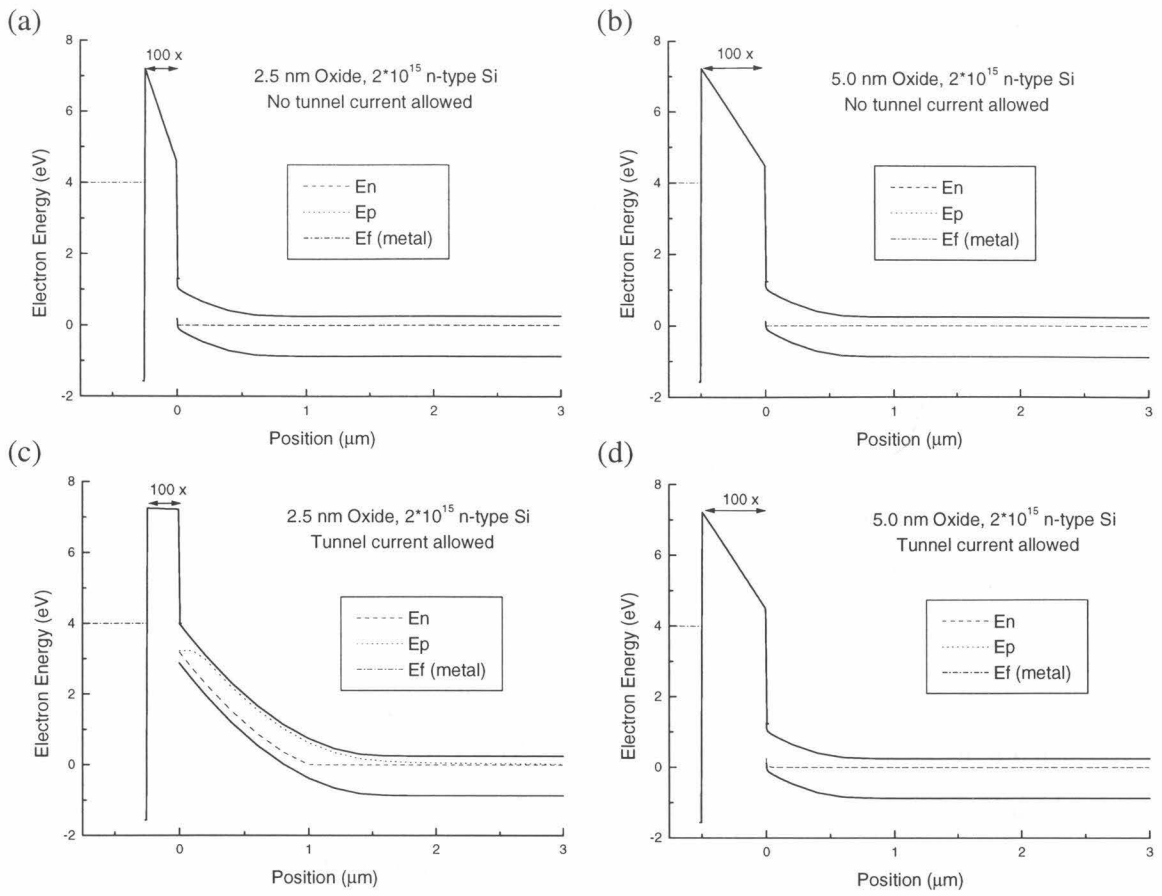

Figure 5.7: Calculated band diagrams for MOS devices consisting of (a),(c) 2.5 nm and (b),(d) 5.0 nm oxides on a 2×10^{15} cm^{-3} n-type silicon layer with Al gate electrodes. The total current was forced to be zero in (a), (b), while tunnel currents were allowed in (c), (d). The oxide region is expanded by a factor of 100 for clarity.

with a number of lifetimes may be necessary to obtain useful information.

It is perhaps possible that some conditioning of the matrix (*i.e.*, scaling of the rows and columns as discussed in Section 5.2.5), may be used to improve the matrix coupling, but such methods were not considered. Other solutions to the problem could involve modification of the physical structure modeled. For example in multidimensional MOSFET simulations, the presence of source/drain p-type implant layers next to the n-layer could supply the holes necessary to allow the inversion layer to be created more rapidly [39]. While it seems possible that the p-layer in the TSD might play a similar role, it is not in close proximity to the oxide, so holes can only reach the oxide if they drift or diffuse across the n-layer. This suggests that it may be more

useful in the high current state than in the low current state.

5.5.3 Studies of the TSD

Due to the difficulties resulting from the bistable nature of the TSD and the numerical problems with the reverse biased MOS part of the device as discussed above, simulation of a TSD device was found to be a nontrivial exercise. In order to obtain meaningful results, it was necessary to adjust recombination lifetimes, oxide thicknesses, tunneling parameters, matrix conditioning parameters, and a number of other parameters related to the convergence of the iterations. At present, there is still much to be learned about the effects of these various parameters, so the results presented here must be considered preliminary. In addition, only a few areas of parameter space have been explored. For example, only punch-through devices have been simulated thus far, even though there are no obvious obstacles to simulation of an avalanche mode device. Nonetheless, the results of the simulations performed thus far may provide some insight into the device operation.

Perhaps the most striking result of these studies is that it was necessary to include hole tunneling through the oxide as well as electron tunneling in order to achieve bistable device behavior. The first simulation attempts only allowed tunnel currents from the Al band through the SiO_2 conduction band, into the silicon conduction band. Hole current was only allowable through recombination in the bulk silicon, or at the silicon-oxide interface. No resulting simulated J-V curves could be forced into multiple current states for any range of tested voltages or recombination lifetimes. It was necessary to include tunneling from the Al band through the SiO_2 valence band into the silicon valence band for the device to be able to support both low and high current states. This is primarily due to the fact that the electron and hole current must be "decoupled" from one another in order for multiple states to be allowed. That is, for a given bias, a low current state will result if the majority of the current is electron tunnel current, and a high current state will result if both electron and hole tunnel current is present, but the simulation can only support both of these conditions if

the hole tunnel current can be independent of the electron tunnel current. For all simulations presented here, the SiO_2 bandgap was taken to be 7.0 eV – somewhat low compared to published bulk values [7, 36], but not unreasonably low given that the oxide layer is so thin. This value was chosen somewhat arbitrarily in an attempt to boost hole currents, so a study of the effect of this parameter on results should be investigated, but this has not been done to date.

The calculated J-V curves were found to be quite sensitive to the recombination lifetimes as illustrated in Figure 5.8. A multistable J-V curve resulted only if the lifetimes were taken to be 10^{-8} s. Shorter lifetimes resulted in J-V curves which traced a low current curve only, and longer lifetimes resulted in J-V curves which traced only a high current curve. These results are qualitatively consistent with the discussion of the thyristor in Section 5.5.1, but it seems unlikely that the window of lifetimes that would allow bistable states should be so narrow or in such a low range, particularly given that switching was observed in nearly every measured device (as discussed in Chapter 4), and no attempt was made to reduce or control the lifetimes during fabrication. It is quite possible that the simulations could not converge to multistable states for certain ranges of parameters due to the difficulties described in the previous sections.

Notice that the J-V curve for $\tau = 10^{-8}$ s in Figure 5.8 is presented differently than the thyristor J-V curves in Figure 5.5. To explain this, a discussion of the method in which the curves were obtained is required. The thyristor J-V curves were plotted as two curves, one set of points obtained by stepping the bias up from the zero bias solution, and another set of points obtained by stepping the bias down from a high current state solution. The TSD J-V curves were obtained in much the same way, with one interesting difference. As the bias across the TSD was stepped up past 3.5 V, the current increased abruptly, but no higher bias solutions were obtained. As the bias was stepped back down, it was found that the current increased with decreasing bias, indicating that the solutions represented the NDR region of the J-V curve, not the high current state. Given that an NDR state of a device in a true circuit is unstable in the presence of small signal fluctuations in a load line circuit, one might

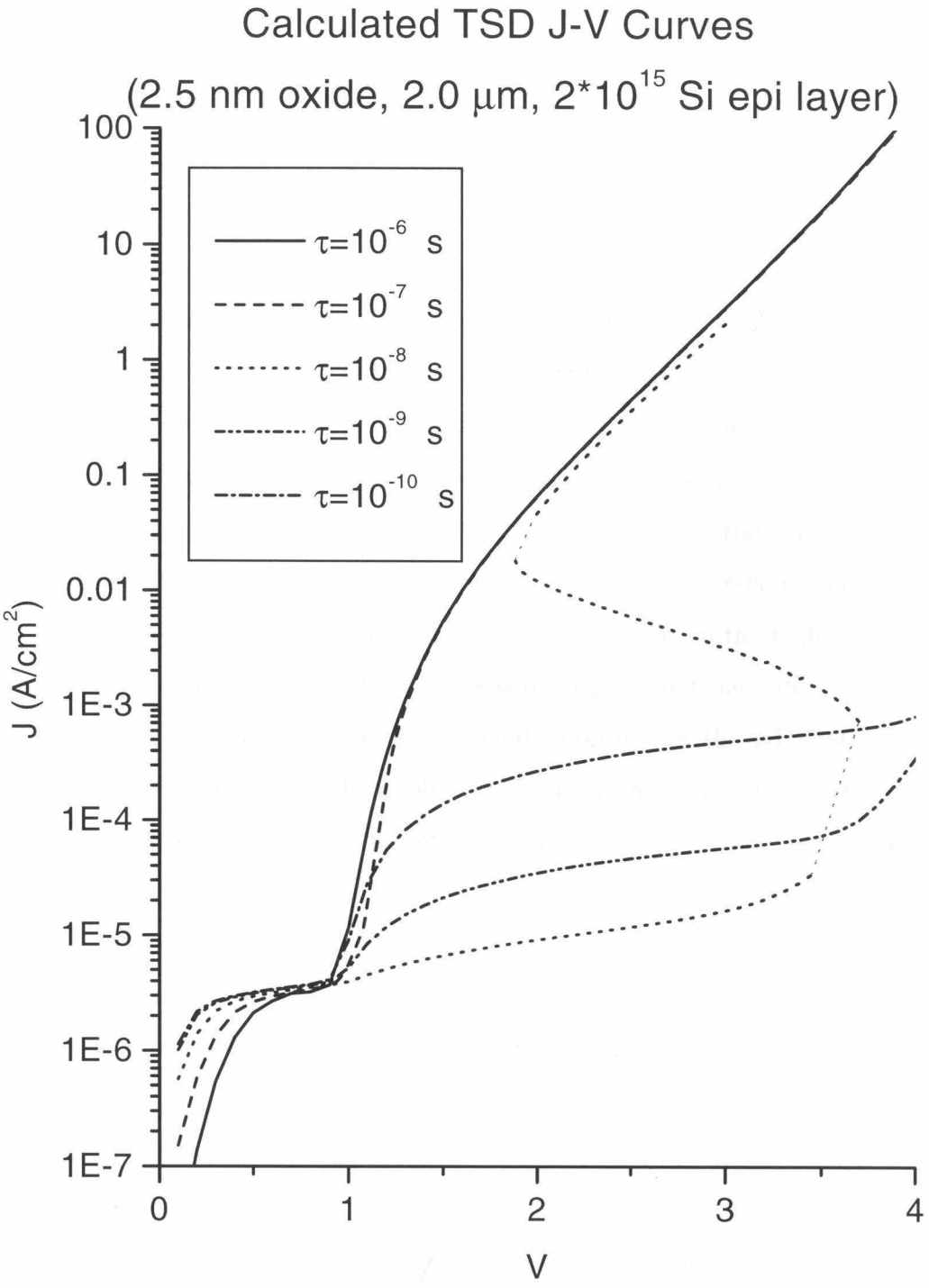

Figure 5.8: Dependence of calculated TSD J-V curves on recombination lifetimes. The simulated structure consists of an Al gate, 2.5 nm oxide, 2.0 μm, 2×10^{15} cm^{-3} n-silicon, and 1.5 μm, 10^{18} cm^{-3} p-silicon.

expect a numerical solution of such a state to be unstable in the presence of numerical fluctuations, so the realization that the drift-diffusion model could converge to NDR states came as somewhat of a surprise. As a result, the J-V curve shown in Figure 5.8 for $\tau = 10^{-8}$ s consists of three parts (low current, NDR, and high current) which were traced out independently, and "stitched" together by connecting the endpoints with lighter lines. Note that in order to enter the set of high current states, a longer lifetime solution was used as an initial guess, as described in Section 5.5.1.

The calculated band diagrams corresponding to several points on this J-V curve are shown in Figure 5.9 in order of increasing current. Notice the similarities between these plots and those for the thyristor in Figure 5.6. For example, in the low current state of both the TSD and the thyristor, the relative positions of the electron and hole quasi Fermi levels indicate a region of carrier generation (where the hole Fermi level is above the electron Fermi level), bracketed by regions of carrier recombination (where the electron Fermi level is above the Hole Fermi level). Also, in the high current state of both devices, carriers traverse the entire silicon region, and the bands are flattened out. It is also interesting to examine the band diagrams in the NDR state as they illustrate the shift of bias from the depletion region in the silicon to the oxide by the formation of an inversion layer at the interface. This is brought about as the electron and hole quasi Fermi levels swap relative positions in the n-silicon layer, changing from a state of generation to recombination.

In order to examine the effect of oxide thickness on TSD behavior, J-V curves were calculated for the same structure studied thus far, with $\tau = 10^{-8}$ s, and with several different oxide thicknesses. The results are presented in Figure 5.10. Again, we find that bistable devices result only for a narrow range of parameters (*i.e.*, oxide thicknesses between 1.5 and 3.0 nm). This can again be understood qualitatively in terms of the thyristor results. As the oxide in the TSD device acts much like the current-limiting p-layer in the thyristor (see Figure 5.4), if this layer is to transmissive, there is no barrier to electron injection, and a depletion layer cannot be supported. If the layer is too highly insulating, the double-injection condition necessary for the high current state cannot be met.

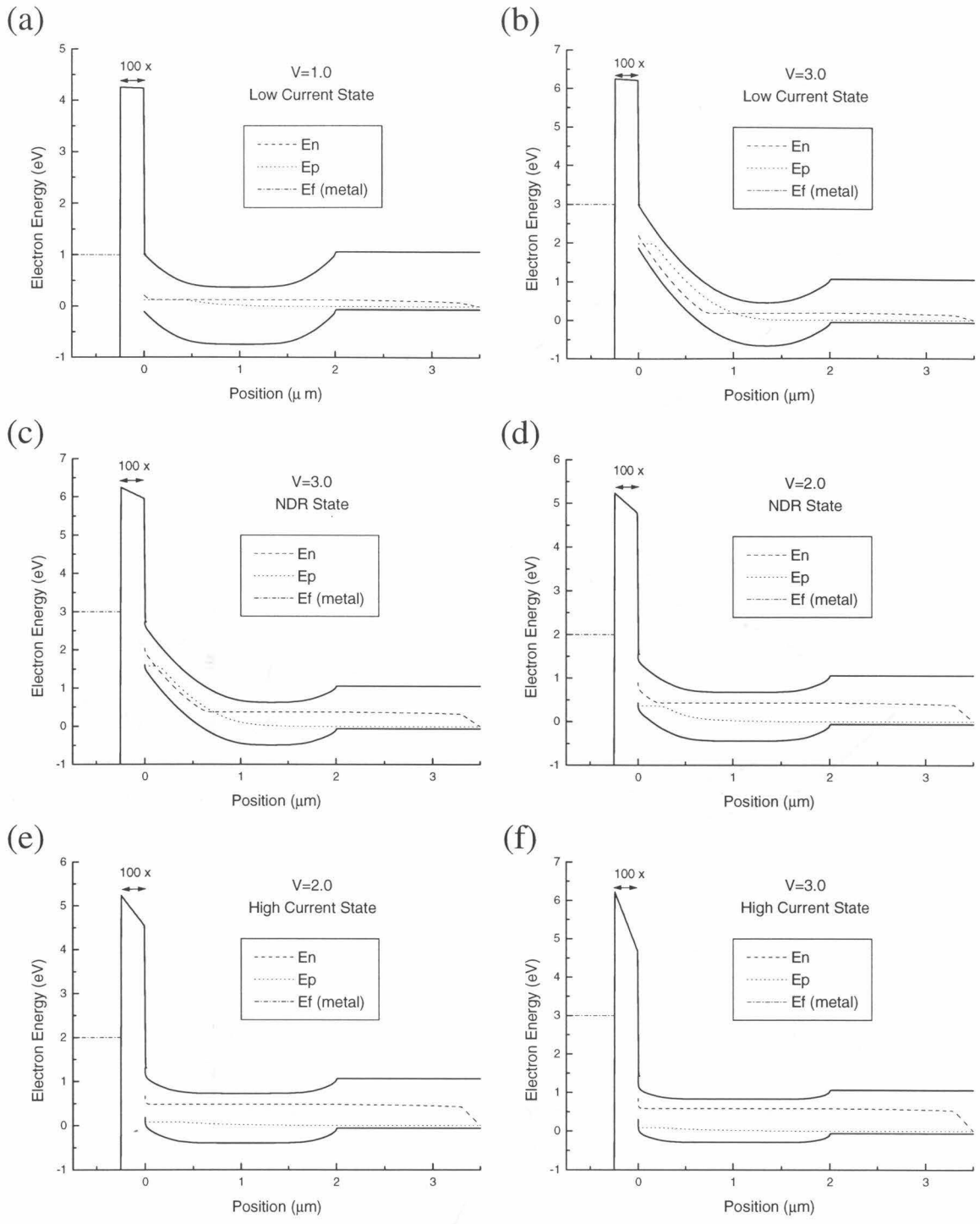

Figure 5.9: Band diagrams corresponding to several points on the TSD J-V curve shown in Figure 5.8 with $\tau = 10^{-8}$ s. The diagrams are arranged in order from (a) the lowest current to (f) the highest current. The oxide regions are magnified by a factor of 100 for clarity.

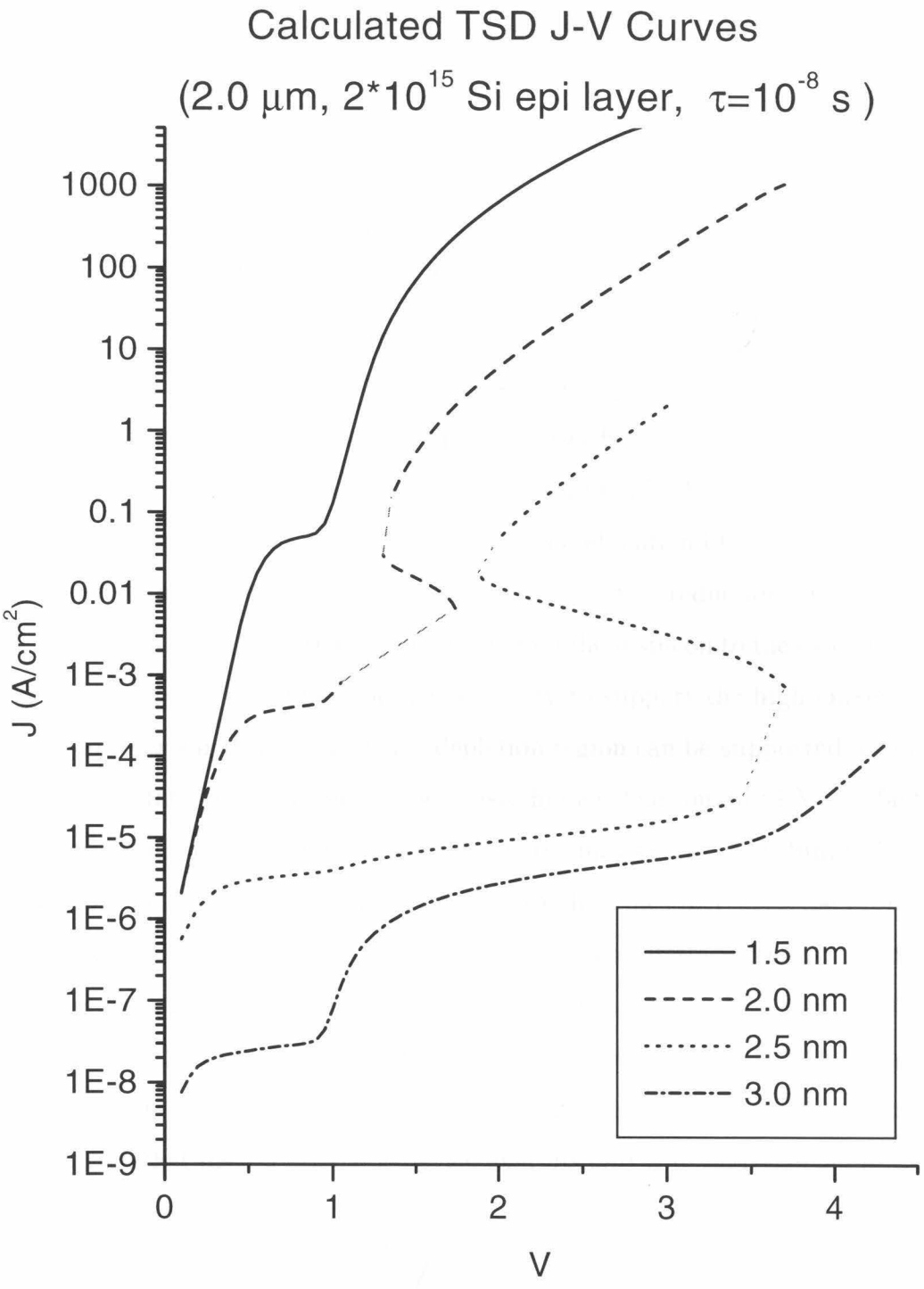

Figure 5.10: Calculated J-V curves are shown for TSD devices with varying oxide thicknesses. The simulated structure consists of an Al gate, 2.5 nm oxide, 2.0 μm, 2×10^{15} cm^{-3} n-silicon, and 1.5 μm, 10^{18} cm^{-3} p-silicon. For each structure, the lifetimes were taken to be 10^{-8} s.

Also, note that the peak voltage increases with increasing oxide thickness. This can also be understood in terms of the similarity of the thyristor current-limiting p-silicon region and the TSD oxide region. In the case of the thyristor, a decrease in the injected current flux through the p-type layer, brought about by a decrease in the recombination lifetime, allows a larger peak voltage, as more bias can be dropped across the reverse-biased p-n junction without increasing the injected current to the point the depletion layer can no longer be supported. An analogous argument holds for the TSD, where the decrease in injected current is brought about by increasing the oxide thickness. In all cases the peak voltages are comparable to the predicted value (2.5 V) based on the punch-through arguments described in Section 2.2.2.

The multistable behavior can be somewhat restored in both the 1.5 nm and 3.0 nm oxide devices from Figure 5.10 by adjusting the recombination lifetimes, as illustrated in Figure 5.11. In the case of the 1.5 nm oxide device, a reduction in the lifetimes decreases the number of holes that can make it from the p-silicon to the oxide interface, thereby increasing the minimum voltage necessary to support the high current state. While it is unclear whether a reasonable depletion region can be supported, given that the low current state failed to converge at biases higher than roughly 1 V, the fact that an NDR region could be reproduced suggests that it may be possible. Similarly, in the case of the 3.0 nm oxide device, an increase in the lifetimes increases the probability that injected holes will diffuse all the way to the oxide interface, thereby increasing the hole tunnel current, allowing support of the double injection, high current state. It is clear that the modifications of the lifetimes cannot completely offset the changes in carrier injection probability brought about by changes in oxide thickness, however. For example, notice that the 1.5 nm oxide device exhibits only a small range of multistable behavior.

5.6 Summary and Discussion

Given the difficulties in encouraging the simulations of the TSD device to converge, and given that only a limited range of parameter space has been explored thus far,

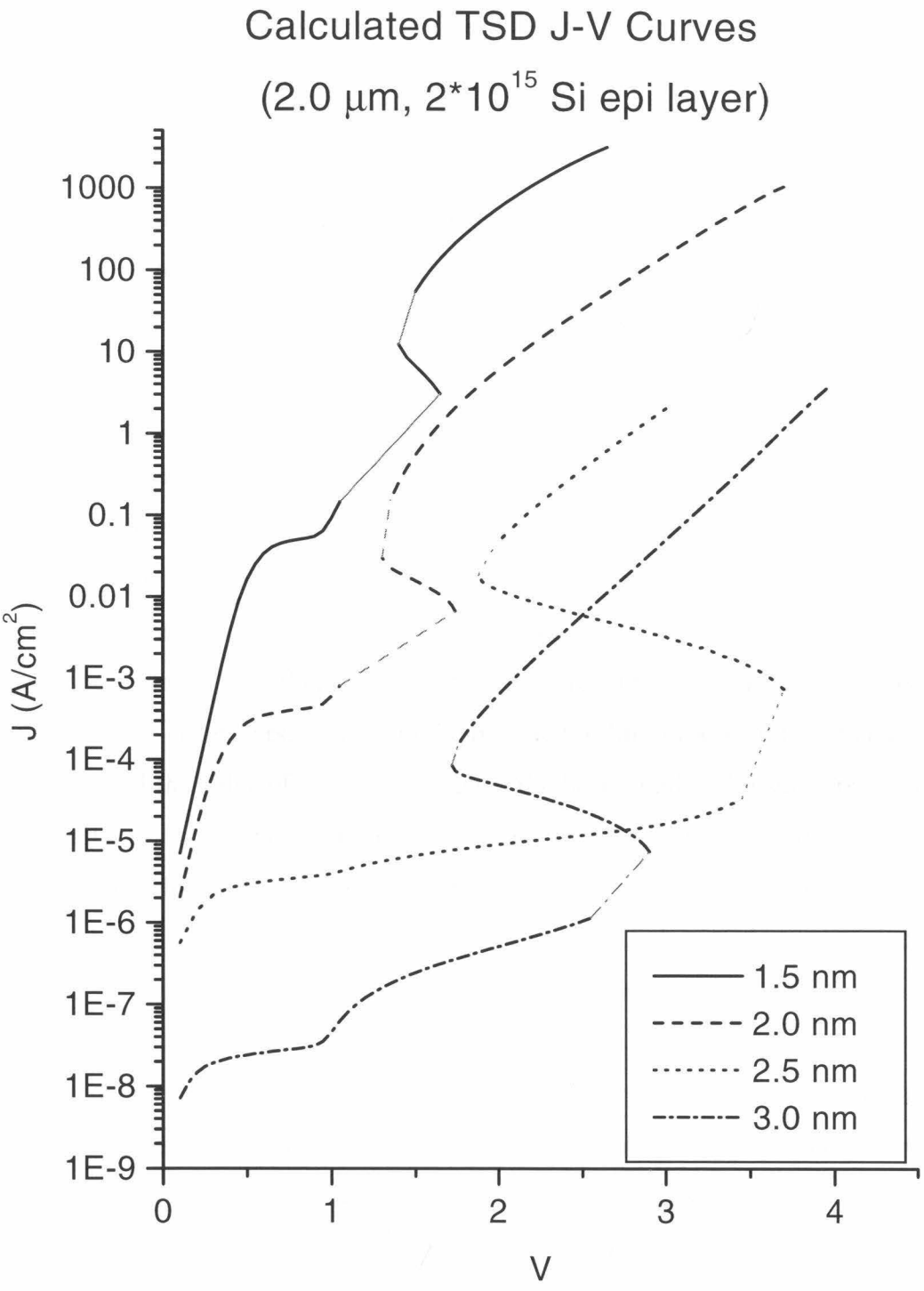

Figure 5.11: Same as Figure 5.10, but the lifetime of the 1.5 nm oxide device has been reduced to 3×10^{-9} s, and the lifetime of the 3.0 nm device has been increased to 10^{-7} s.

it would perhaps be unwise to put too much faith in the results of these simulations, but several general comments can be made with some degree of certainty.

Nearly all effects due to changes in oxide thickness and recombination lifetimes in the calculated J-V curves could be explained in terms of modification of carrier injection, in close analogy to thyristor operation. Therefore, we might expect that the effects of other parameters, such as layer thicknesses and oxide tunneling parameters, could be predicted from similar analogies to the thyristor. This could prove quite useful, as the thyristor is generally much better understood, and somewhat easier to simulate.

The fact that the range of oxide thicknesses and recombination lifetimes for which multistable J-V characteristics could be obtained was quite narrow may or may not have any real physical significance. On one hand, a much wider range of allowable parameters was observed for the measured devices presented in the previous chapter. Although no attempts were made to measure lifetimes, no deliberate attempt was made to modify or control them. In addition, several different vendors were used in production of the epi wafers, so it is unlikely that the lifetimes would be found to be clustered around the value of 10^{-8} s which gave the best simulated results, particularly as typical silicon recombination lifetimes are much longer (10^{-6}-10^{-3} s).

On the other hand, it is possible that this apparent discrepancy could explain, in part, why the measured peak voltages were so much larger than expected. To understand this, recall the idea presented in the previous chapter that, in a device with no lateral confinement in the silicon layers, it is possible that carriers could spread out over a large area in the silicon (see Figure 4.16), and the creation of a depletion layer in this extended region would presumably require a larger voltage drop. It is possible that such a two-dimensional arrangement could be favorable if the carriers "preferred" to diffuse laterally rather than through the layers as a result of the electric fields present across the device layers. In such an arrangement, the injection through the p-n junction could be reduced due to the lateral flow of charge, effectively looking like a reduction in the recombination lifetimes for carriers traveling through the layers. Unfortunately, a multidimensional model would be require to simulate such

a structure, but as mentioned in the previous chapter, fabrication of laterally confined devices might provide better understanding of some of these problems.

As mentioned previously, the fact that it was necessary to allow for hole tunneling currents in order to obtain multistable J-V curves most likely is a result of the requirement that the hole and electron currents be somewhat independent. It is possible that this result is of greater physical significance, however. As the tunnel currents in most MOS device structures are dominated by electron tunnel current, little is known about hole tunneling through SiO_2. It is possible that further study of TSD devices compared with simulations with a number of different tunneling parameters may provide some insight.

Bibliography

[1] S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer-Verlag, Wien, Austria, 1984.

[2] C. M. Snowden, *Introduction to Semiconductor Device Modeling*, World Scientific, Singapore, 1986.

[3] W. L. Engl, H. K. Dirks, and B. Meinerzhagen, *Proc. IEEE* **71**, 10 (1983).

[4] M. W. Wang et al., *J. Appl. Phys.* **73**, 4660 (1993).

[5] C. S. Rafferty, M. R. Pinto, and R. W. Dutton, *IEEE Trans. Computer-Aided Des.***CAD-4**, 462 (1985).

[6] J. S. Blakemore, *Solid–State Electron.* **25**, 1067 (1982).

[7] M. Shur, *Physics of Semiconductor Devices*, Prentice-Hall, New Jersey, 1990.

[8] X. Aymerich-Humet, F. Serra-Mestres, and J. Millan, *Solid–State Electron.* **24**, 981 (1981).

[9] W. M. Coughran, Jr., M. R. Pinto, and R. K. Smith, *J. Comp. Appl. Math.* **26**, 47 (1989).

[10] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, Academic Press, New York, 2nd edition, 1992.

[11] R. E. Bank and D. J. Rose, *SIAM J. Numer. Anal.* **17**, 806 (1980).

[12] R. E. Bank and D. J. Rose, *Numer. Math* **37**, 279 (1981).

[13] H. K. Gummel, *IEEE Trans. Electron Devices* **ED-11**, 455 (1964).

[14] S. Odanaka, A. Hiroki, K. Ohe, K. Moriyama, and H. Umimoto, *IEEE Trans. Computer-Aided Des.***10**, 619 (1991).

[15] K. H. Back, H. K. Dirks, B. Meinerzhagen, and W. L. Engl, *IEEE Trans. Computer-Aided Des.***10**, 1175 (1991).

[16] R. E. Bank, *Comp. Phys. Comm.* **53**, 201 (1989).

[17] R. E. Bank, T. F. Chan, W. M. Coughran, Jr., and R. K. Smith, *BIT* **29**, 938 (1989).

[18] R. K. Smith, private communication.

[19] E. Anderson et al., *Lapack Users Guide*, SIAM, University Science Center, Philadelphia, PA, 2nd edition, 1994.

[20] Clapack linear algebra routines version 2.0, Obtained from Netlib Database: http://netlib.org, 1994.

[21] S. M. Sze, *Physics of Semiconductor Devices*, Wiley, New York, NY, 2nd edition, 1981.

[22] C. B. Duke, *Tunneling in Solids*, Academic Press, New York, 1969.

[23] M. A. Green, F. D. King, and J. Shewchun, *Solid–State Electron.* **17**, 551 (1974).

[24] J. G. Simmons, *J. Appl. Phys.* **34**, 1793 (1963).

[25] W. A. Harrison, *Phys. Rev.* **123**, 85 (1961).

[26] M. Depas, B. Vermeire, P. W. Mertens, R. L. Van Meirhaeghe, and M. M. Heyns, *Solid–State Electron.* **38**, 1465 (1995).

[27] B. Brar, G. Wilk, and A. C. Seabaugh, *Appl. Phys. Lett.* **69**, 2728 (1996).

[28] D. A. Buchanan and S.-H. Lo, *Proc. Electrochem. Soc.* **96-1**, 3 (1996).

[29] A. Faigon and F. Campabadal, *Solid–State Electron.* **39**, 251 (1996).

[30] S. Zafar, K. A. Conrad, Q. Liu, and E. A. Irene, *Appl. Phys. Lett.* **67**, 1031 (1995).

[31] S. Zafar, Q. Liu, and E. A. Irene, *J. Vac. Sci. Technol. A* **13**, 47 (1995).

[32] C. S. Lent and D. J. Kirkner, *J. Appl. Phys.* **67**, 6353 (1990).

[33] D. Z.-Y. Ting, E. T. Yu, and T. C. McGill, *Phys. Rev. B* **45**, 3583 (1992).

[34] D. Z.-Y. Ting and T. C. McGill, (1995).

[35] D. Z.-Y. Ting, private communication.

[36] S. Nagano, *J. Appl. Phys.* **75**, 3530 (1994).

[37] K.-H. Hellwege and J. L. Olsen, *Landolt-Bornstein Numerical Data and Functional Relationships in Science and Technology*, volume 13 – Metals:Phonon States, Electron States and Fermi Surfaces, Springer-Verlag, Berlin, 1984.

[38] M. Hirose, J. L. Alay, T. Yoshida, and S. Miyazaki, *Proc. Electrochem. Soc.* **96-1**, 485 (1996).

[39] W. R. Frensley, private communication.

[40] S. Wang, *Fundamentals of Semiconductor Theory and Device Physics*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

Part II

Experimental and Theoretical Investigations of an InAs/GaSb/AlSb Double Quantum Well Velocity Modulation Transistor (VMT)

Chapter 6 Background

6.1 Introduction and Outline

In this section, we will introduce the concept of the velocity modulation transistor, and explore some of its properties. In Section 6.2, the original proposal of such a device will be discussed, along with details of progress that has been made in development of the device in the GaAs/AlGaAs system. Then, in Section 6.3, our proposed InAs/GaSb/AlSb device will be presented. The details of device behavior in both systems will be compared and several fundamental differences between the two systems will be identified.

6.2 The Velocity Modulation Transistor (VMT)

6.2.1 Device Concept

The concept of the velocity modulation transistor (VMT), a transistor in which the carrier velocity is modulated instead of the carrier concentration, was introduced by Sakaki in 1982 [1]. Such a device might not be limited by the same speed constraints as traditional field effect transistors, as charges need not be shuffled in and out of the conductive channel in order to effect a conductivity change.

One proposed implementation of a VMT device [1, 2] is shown in Figure 6.1(a). The device consists of two coupled quantum wells of differing mobilities, separated by an insulating barrier, and a gate contact on top. An applied gate voltage can shift the relative energies of the quantum wells, bringing them in and out of resonance. When the quantum wells are decoupled (Figure 6.1(b)), the dual quantum well system acts as two separate channels in parallel, so the conductivity is dominated by the channel with the highest mobility. If the wells are brought into resonance, however (Figure 6.1(c)),

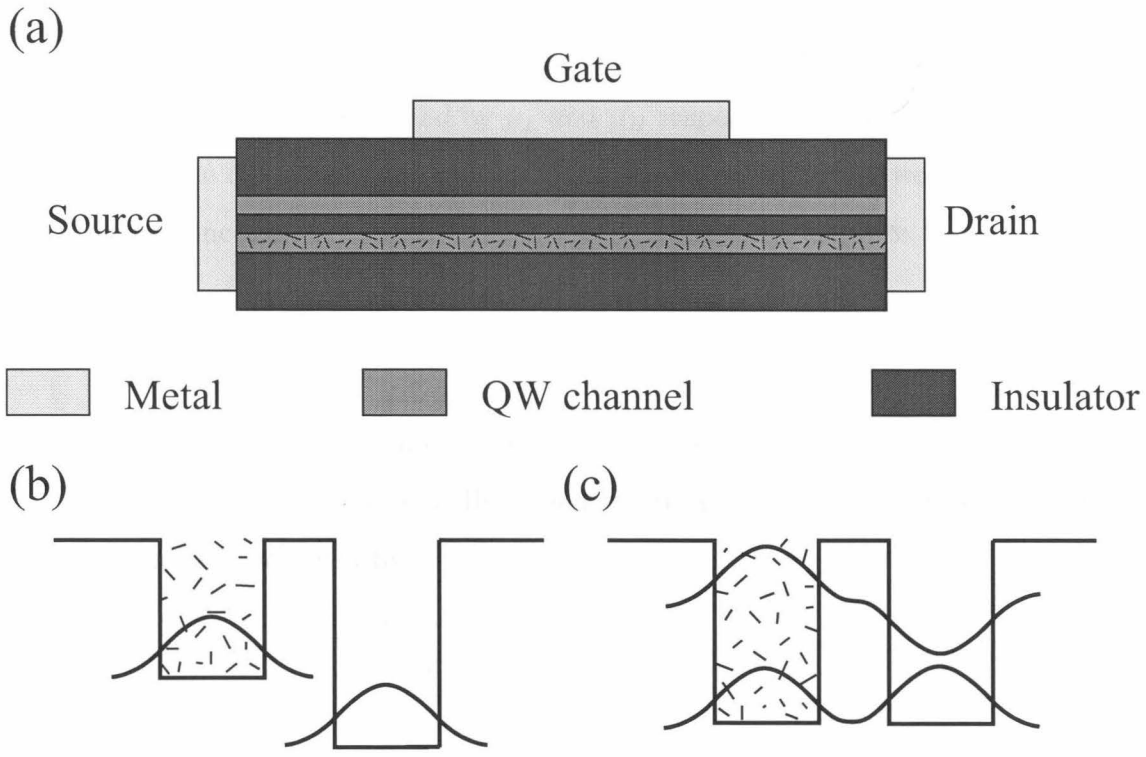

Figure 6.1: (a) One implementation of a velocity modulation transistor. Two coupled quantum wells, one high mobility and one low mobility (indicated by the stippled regions), are coupled together through a thin insulating layer, shorted together at both ends by source and drain contact. A gate contact interacts with the coupled wells through a thick insulating layer. (b) The electron wavefunctions are localized in each well when the wells are decoupled. (c) Symmetric and antisymmetric coupled states spread across both quantum wells when their energies are brought into resonance.

the wavefunctions become delocalized, spreading across both quantum wells. In this configuration, all electrons have some probability of scattering within the low mobility channel, so the overall mobility of the coupled system can be lowered.

6.2.2 Simple Performance Estimate

In order to estimate the possible modulation of the conductance in such a devices, consider the following simple treatment. Let the low and high mobilities corresponding to the two channels be designated by μ_L and μ_H respectively. Similarly, let n_L, n_H be the sheet charge concentrations in each channel. Then, when the wells are decoupled and the wavefunctions are localized, the sheet conductance is given by

$$g_{decoupled} = q(\mu_L n_L + \mu_H n_H). \tag{6.1}$$

When the wells are brought into resonance, we shall assume that both wavefunctions are spread equally across both wells. Then the total scattering cross section seen by each electron will be given by

$$\sigma_{coupled} = \frac{1}{2}(\sigma_L + \sigma_H), \tag{6.2}$$

where σ_L and σ_H are the total cross sections of the scattering mechanisms in the two wells. Assuming that the difference in mobility is completely due to the difference in cross sections, and using the fact that

$$\mu = \frac{q\tau}{m^*} \propto (1/\sigma), \tag{6.3}$$

where τ is the mean free time between relaxation collisions, and m^* is the effective mass, the effective mobility of all electrons is given by

$$\mu_{coupled} = 2(\frac{1}{\mu_L} + \frac{1}{\mu_H})^{-1}, \tag{6.4}$$

so, assuming the charge concentrations in the channels have not changed appreciably, the sheet conductance in the coupled state is given by

$$g_{coupled} = 2q(n_H + n_L)(\frac{1}{\mu_L} + \frac{1}{\mu_H})^{-1}. \tag{6.5}$$

If we assume that n_L and n_H are equal, then the sheet conductance ratio is given by

$$\frac{g_{decoupled}}{g_{coupled}} = \frac{(\mu_H + \mu_L)^2}{4\mu_L\mu_H} \approx \frac{1}{2} + \frac{1}{4}\frac{\mu_H}{\mu_L} \tag{6.6}$$

where the approximate form assumes $\mu_H \gg \mu_L$.

6.2.3 GaAs Implementations of the VMT

Several implementations of the VMT have been realized in the GaAs/AlGaAs system [3–5]. One device [5], consisting of coupled GaAs quantum wells (QWs) separated by an AlGaAs barrier, displays a strong conductance modulation of roughly a factor of 3 at 4.2 K (see Figure 6.2). Given that the mobility ratio was calculated to be roughly 10-20 [5], the simple result of Equation 6.6 (ratio of 3-5.5) is in good agreement with this result. As the difference in mobility between the two wells is brought about by introducing Si scattering centers within one of the wells, the mobility ratio approaches unity at higher temperatures, and even at 77 K, the effect is nearly washed out.

Note that the curves in Figure 6.2 represent the superposition of two effects. One is that of the velocity modulation transistor as described, which produces a dip in conductance when the wells are coupled. The other effect is the increase in concentration of channel carriers with increasing (positive) gate voltage, resulting in a monotonic increase in conductance.

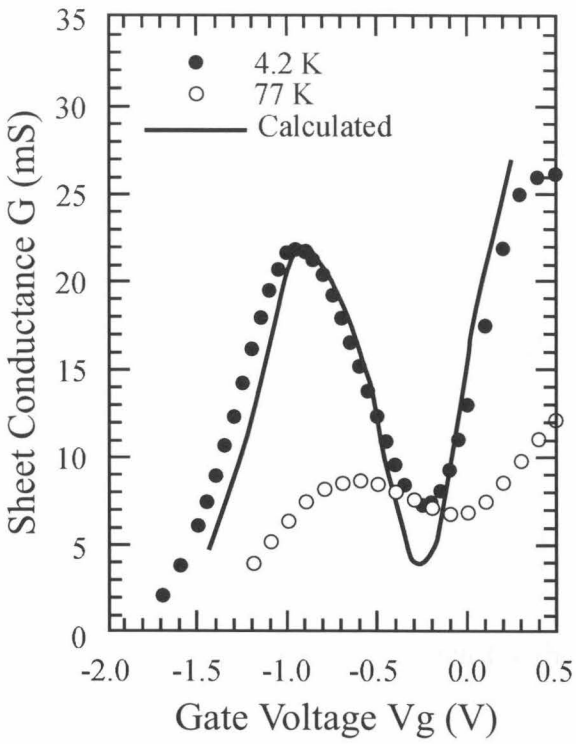

Figure 6.2: Measured response of GaAs/AlGaAs velocity modulation transistor (adapted from [5]).

6.3 An InAs/GaSb/AlSb Implementation of the VMT

6.3.1 Basic Concept

The near-lattice matched InAs/GaSb/AlSb materials system, shown in Figure 6.3, has been widely used in a variety of applications as a consequence of its versatility. The fact that the InAs conduction mobility can be made very high (>30,000 cm^2/V-s at room temperature in a quantum well geometry [10]) has generated interest in producing HFET structures [11, 12] and other high speed devices [7, 13]. Infrared devices can be implemented in this system, owing to the small bandgap of the InAs and GaSb, and the resulting properties of superlattice structures [14, 15]. Any type of heterojunction (type I, type II staggered, and type II broken) can be fabricated in this system, allowing for fabrication of a large class of quantum-effect devices [6].

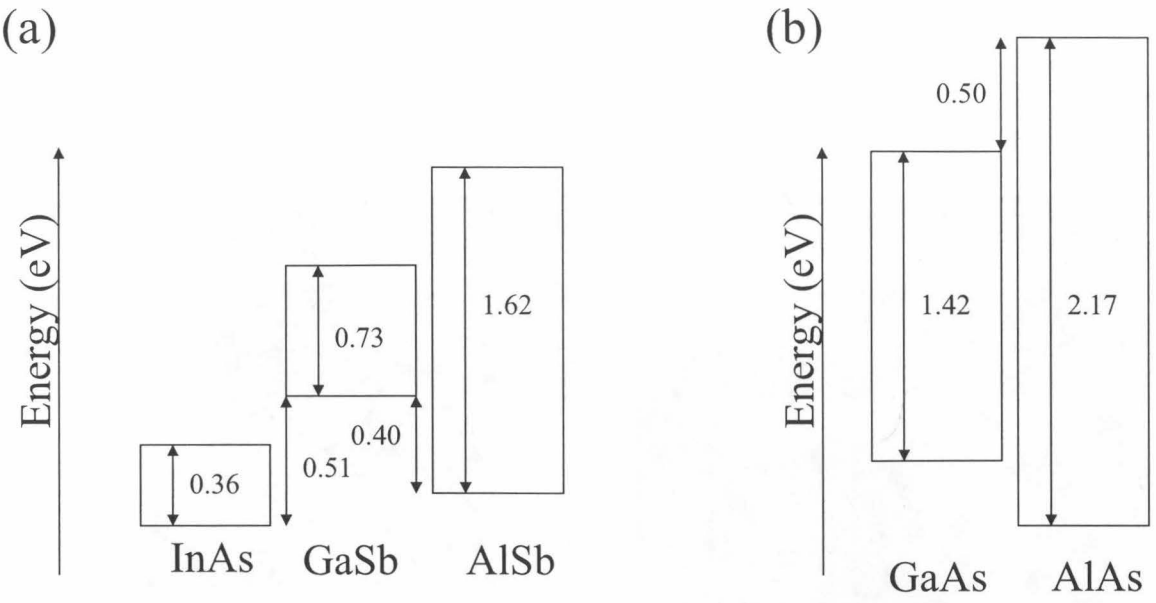

Figure 6.3: Energy bandgaps and alignments are shown for the (a) InAs/GaSb/AlSb system [6–8], and the (b) AlGaAs system [9].

A proposed InAs/GaSb/AlSb VMT device structure is shown in Figure 6.4. Note that the relative band alignments of GaSb and InAs allow for GaSb holes to be equal in energy to InAs electrons. These states are intrinsically very different, unlike the wells in the GaAs device implementations which must be forced to have different mobilities. As a result, the ratio of mobilities of the two states can be made quite high even at room temperature: given reported upper-limit values of 30,000 cm^2/V-s for InAs electrons [10] and 500 cm^2/V-s for GaSb holes [16], a ratio of 60 is possible, leading to a conductivity ratio of 15 using Equation 6.6. Assuming the GaSb mobility could be reduced to 100 cm^2/V-s by introducing interface roughness [10, 17] or some other degradation mechanism, the mobility ratio would become 300, and the conductivity ratio 75. While this estimate looks promising, the simple model fails to describe many subtleties of the VMT device operation in this material system, so a more detailed description will be presented here.

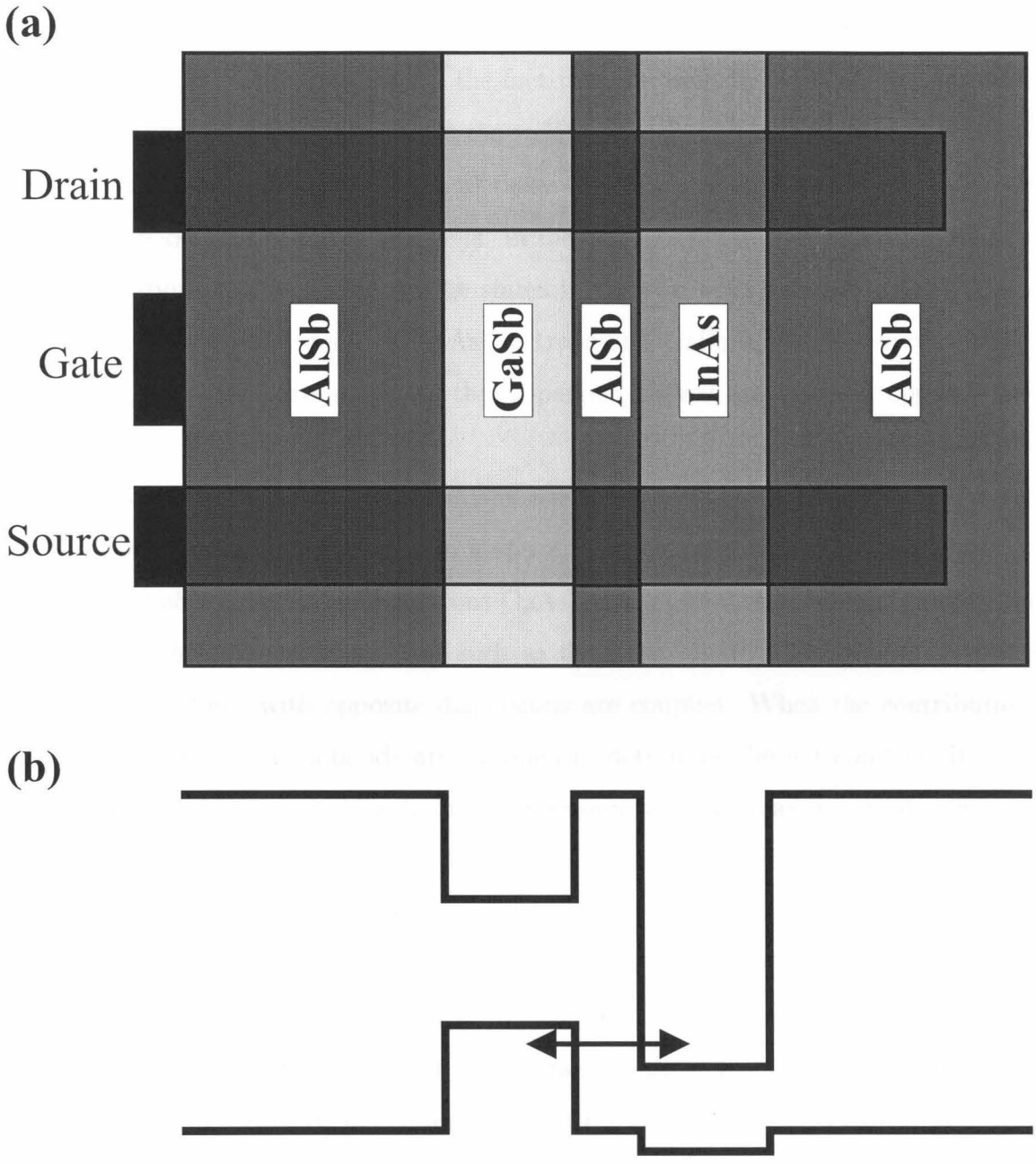

Figure 6.4: Proposed InAs/GaSb/AlSb VMT structure. (a) Device layer structure with gate and ohmic alloyed source and drain contacts. (b) Flat band diagram illustrating coupling of GaSb holes and InAs electrons.

6.3.2 Comparison of GaAs/AlGaAs and InAs/GaSb/AlSb Implementations

Several differences between the GaAs implementations of the VMT and the proposed InAs/GaSb/AlSb device result from the fact that the wells in the GaAs implementation are intrinsically the same, while the wells in the InAs/GaSb/AlSb implementation are intrinsically different. Most of these effects are related to the energy versus **k**-space dispersion relations in the wells. In the GaAs/AlAs implementations, the dispersion relations corresponding to the states in the two wells are the same, while in the InAs/GaSb/AlSb system, the InAs electrons and GaSb holes have very different dispersion behavior, particularly as the dispersion curves are directed opposite one another.

The discussion in the previous sections was simplified by considering a single pair of resonant states at a given point in **k**-space. The behavior of such a pair of states, as described above for the homogeneous GaAs/AlAs system, will be qualitatively the same as in a heterogeneous system, such as the InAs/GaSb/AlSb system, in which quantum well states with opposite dispersions are coupled. When the contributions of all **k**-states from all subbands are summed to determine the net conductivity at a given gate voltage, however, the **k**-space dispersion relations play a crucial role.

In analyzing this phenomenon, we shall assume that a given state in one well can only couple with a state of the same energy and **k**-vector as in the other well. That is, no interface scattering or multiple body phenomena will be considered. Then, in the case of a homogeneous double well system, as a changing gate voltage shifts the energies of the subbands in one well with respect to those in the other well, all states of a given pair of subbands move in and out of resonance simultaneously as shown in Figure 6.5(a)(i) due to the line-up of the homogeneous dispersion curves in **k**-space. Only states within $\sim k_b T$ of the Fermi level contribute strongly to the conductivity due to thermal smearing of the Fermi surface, but since all states have been brought into resonance simultaneously, the particular position of the Fermi level relative to the subband energies is irrelevant as long as any other subbands which may not be

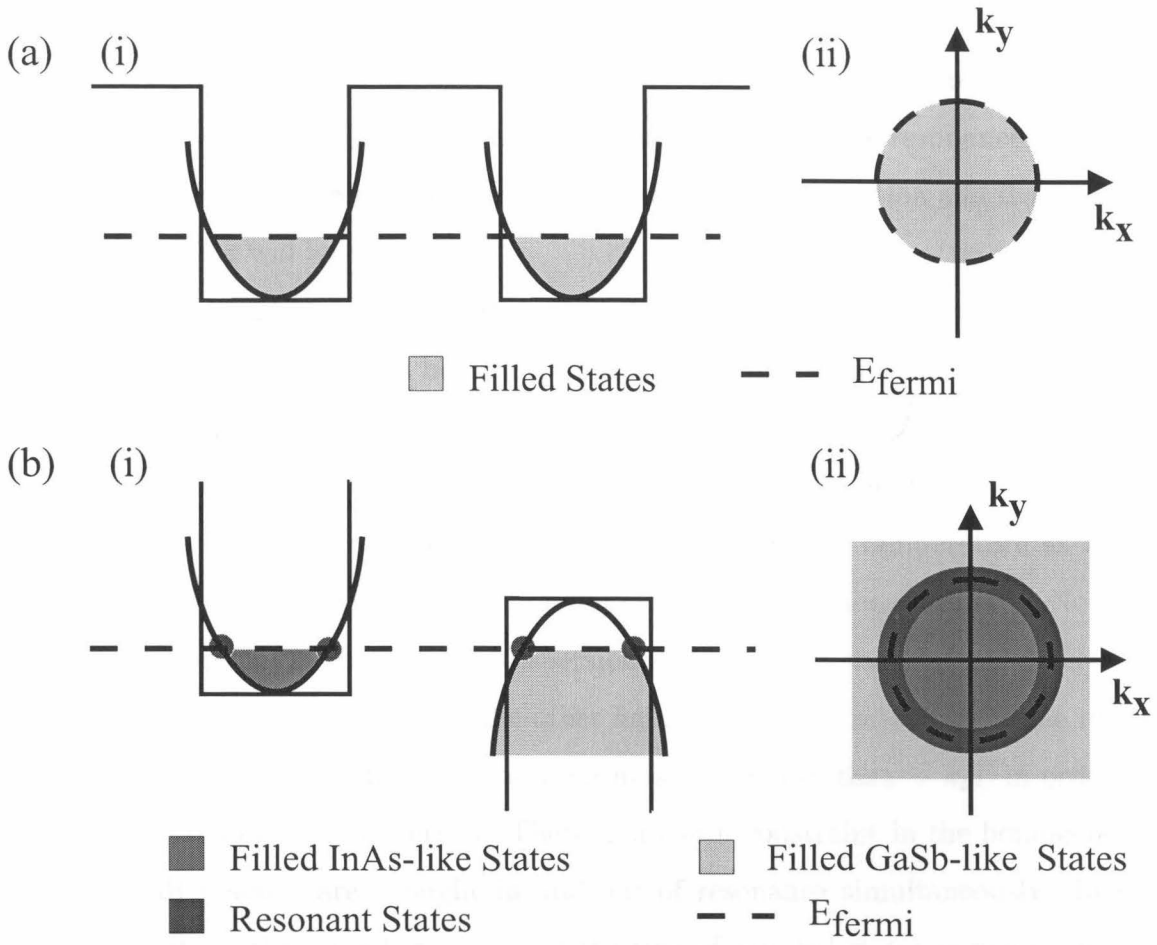

Figure 6.5: Comparison of (a) a homogeneous double well system and (b) an inhomogeneous system with opposite band dispersions in the wells. For each system, we show (i) the transverse dispersion curves (E vs. $k_{||}$) for both wells superimposed on the flat band diagram, and (ii) the distribution of states in transverse k-space in an (idealized) isotropic situation. The shaded regions indicate filled states of different mobility and localization as labeled.

resonant are more than $\sim k_bT$ away in energy.

In the case of the proposed heterogeneous system, for any given gate voltage, the majority of states are not resonant, but there always exists some set of states of some energy and **k** which are in resonance, as shown in Figure 6.5(b)(i), as long as the bottom of the lowest InAs conduction subband is not raised above the top of the highest GaSb valence subband. These resonant states form a ring in transverse **k**-space as shown in Figure 6.5(b)(ii) with a width determined by the coupling strength. As the gate voltage is changed, these resonant states are shifted in energy relative to

the Fermi level. If the Fermi level is further than $\sim k_bT$ away from these resonant states, they do not contribute strongly to conduction, but at the point at which the Fermi level matches the resonant state energy, if the width of the resonance is greater than $\sim k_bT$ in energy, these resonant states will dominate conduction and the mobility modulation effect will be observed.

The fundamental physical limitations on the maximum operating temperature at which velocity modulation can be observed are quite different in each of the described systems. The difference in mobility between the wells is brought about solely by introduction of impurities in the homogeneous case, but the well mobility is only dominated by this scattering mechanism below some fundamental temperature at which other scattering mechanisms, such as phonon scattering, do not dominate. No such restriction is present in the opposite dispersion system due to the intrinsic differences of mobility between the wells. On the other hand, the spread in energy of the resonant states in the opposite dispersion system must be greater than $\sim k_bT$ in order for velocity modulation to be observed. There is no such constraint in the homogeneous system as all **k**-states are brought in and out of resonance simultaneously. In our proposed system, the spread in energy of the ring of coupled states is maximized by reducing the width of the AlSb barrier, maximizing the coupling strength. In fact, the AlSb barrier is not necessarily needed to isolate the wells – the quantum wells remain distinct even in the absence of a barrier, as has been shown in discussion of other barrierless tunnel devices [8, 18, 19]. In both homogeneous and heterogeneous cases, the active QW subbands must be removed from other subbands by $\sim k_bT$, but this is not considered a fundamental limitation because the subbands may be spread out simply by narrowing the quantum wells until one of the above effects dominates.

6.3.3 Transport Mechanisms

Another distinction between a homogeneous double well system and an inhomogeneous double well system is discovered in considering the transport mechanisms in each system. One familiar simple picture of conduction in a degenerate semiconductor or

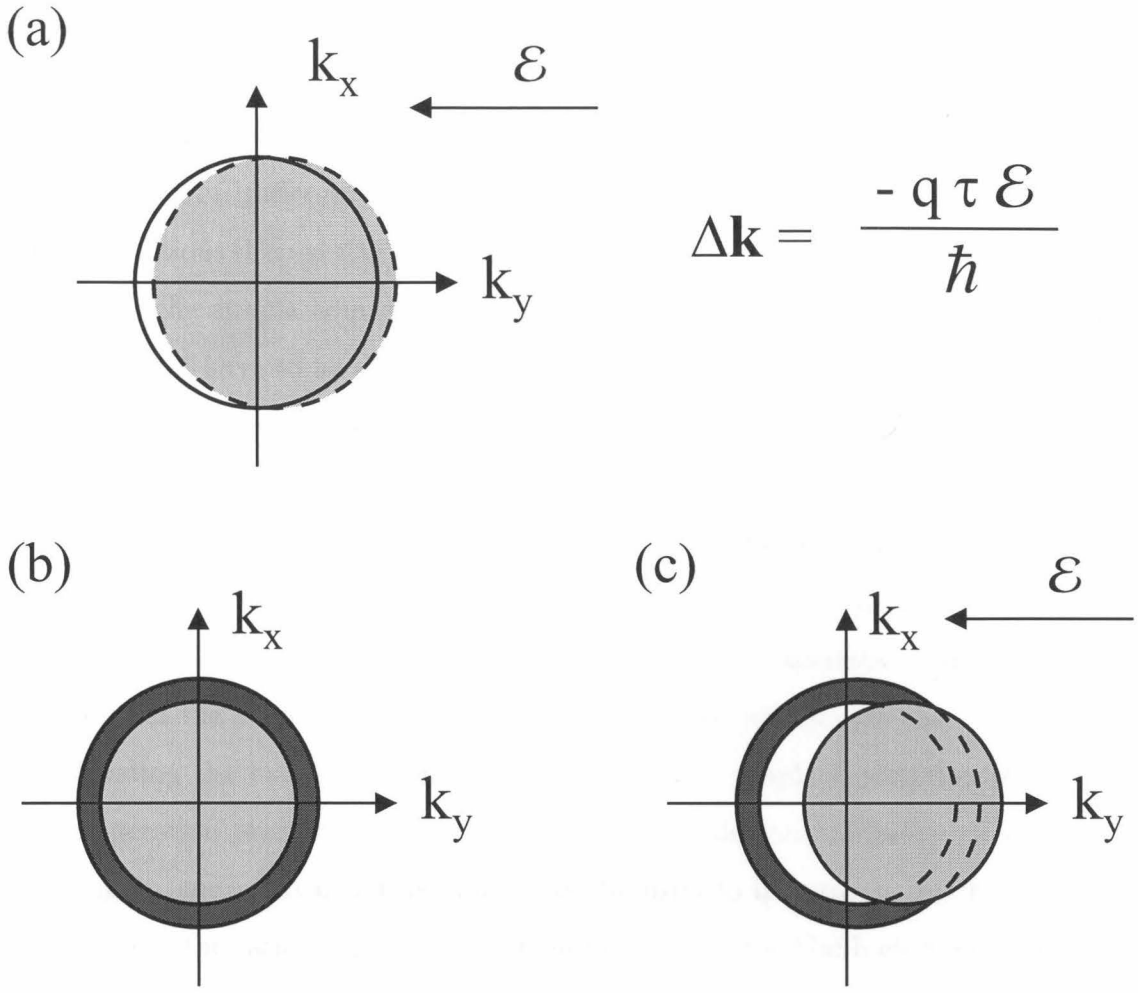

Figure 6.6: (a) Simple transport model for a metal or degenerate semiconductor. An applied electric field shifts the Fermi sea by an amount proportional to the mean free time. (b) Distribution of **k**-space states in a heterogeneous VMT system under no applied electric field. (c) Application of simple model in (a) to the structure in (b).

in a metal is illustrated in Figure 6.6(a). The effect of an applied electric field, $\mathcal{E}$ is to shift the Fermi sea of electrons in **k**-space by an amount proportional to the mean free time, τ. The addition of electrons along one edge of the sphere and the corresponding subtraction of states from the opposite edge lead to a net movement of charge. Consider what happens when this model is applied to both the homogeneous and heterogeneous VMT systems described above.

In the homogeneous well system, the states are uniform in character throughout **k**-space, composed either of a combination of low mobility and high mobility localized

states, or solely of delocalized states, depending on the gate voltage. Therefore, the simple model still makes sense, shifting the entire Fermi sea in each band by the same amount, though that amount is dependent on gate voltage. In the opposite dispersion system, however, different areas of **k**-space are characterized by different mobilities and localizations (Figure 6.6(b)). If the mobility difference is a result of different mean free times, the simple transport picture no longer makes sense as different areas of **k**-space would have to be shifted by different amounts (Figure 6.6(c)). This suggests that unusual transport behavior may be exhibited by such a system, which must be described by a more sophisticated model.

A difference in mobility can be brought about by a mass difference as well as by a difference in mean free time, as can be seen from Equation 6.3. In the InAs/GaSb/AlSb system, or in any system in which the dispersion relations in the two wells are not the same, the coupling between the wells will modify the subband structure, bending the bands in regions of coupling. The simple description of the VMT device operation presented in the previous section does not address this situation at all, so again, more advanced methods must be used to investigate this behavior.

Note that the ratio of the InAs electron mobility to the GaSb electron mobility for the highest mobility cases is roughly 60, while the ratio of the GaSb heavy hole mass to the InAs conduction band mass is (.23/.023)=10 [7, 16]. This suggests that both a difference in effective mass and in the mean free time are necessary to account for the observed mobility difference in the materials. Therefore, both of the above effects are likely to influence the transport mechanisms.

Bibliography

[1] H. Sakaki, *Jpn. J. Appl. Phys.* **21**, L381 (1982).

[2] B. Vinter and A. Tardella, *Appl. Phys. Lett.* **50**, 410 (1987).

[3] K. Hirakawa, H. Sakaki, and J. Yoshino, *Phys. Rev. Lett.* **50**, 1279 (1985).

[4] A. Palevski, F. Beltram, F. Capasso, L. Pfeiffer, and K. W. West, *Phys. Rev. Lett.* **65**, 1929 (1990).

[5] Y. Ohno, M. Tsuchiya, and H. Sakaki, *Appl. Phys. Lett.* **62**, 1952 (1993).

[6] T. C. McGill and D. A. Collins, *Semicond. Sci. Technol.* **8**, S1 (1993).

[7] A. G. Milnes and A. Y. Polyakov, *Mat. Sci. and Eng.* **B18**, 237 (1993).

[8] D. Z.-Y. Ting, E. T. Yu, and T. C. McGill, *J. Vac. Sci. Technol. B* **9**, 2405 (1991).

[9] M. Shur, *Physics of Semiconductor Devices*, Prentice-Hall, New Jersey, 1990.

[10] C. R. Bolognesi, H. Kroemer, and J. H. English, *Appl. Phys. Lett.* **61**, 213 (1992).

[11] C. R. Bolognesi, J. E. Bryce, and D. H. Chow, *Appl. Phys. Lett.* **69**, 3531 (1996).

[12] K. Yoh, K. Kiyomi, M. Yano, and M. Inoue, *J. Cryst. Growth* **127**, 29 (1993).

[13] D. H. Chow et al., *IEEE Electron Device Lett.* **17**, 69 (1996).

[14] D. H. Chow et al., *Appl. Phys. Lett.* **25**, 3700 (1995).

[15] E. R. Youngdale et al., *Appl. Phys. Lett.* **23**, 3160 (1994).

[16] A. Milnes and A. Y. Polyakov, *Solid–State Electron.* **36**, 803 (1993).

[17] H. Sakaki, T. Noda, K. Hirakawa, M. Tanaka, and T. Matsusue, *Appl. Phys. Lett.* **51**, 1934 (1987).

[18] D. Z.-Y. Ting, E. T. Yu, and T. C. McGill, *Appl. Phys. Lett.* **58**, 292 (1990).

[19] D. Z.-Y. Ting, E. T. Yu, D. A. Collins, D. H. Chow, and T. McGill, *J. Vac. Sci. Technol. B* **8**, 810 (1990).

Chapter 7 Modeling

7.1 Introduction and Outline

As discussed in the previous chapter, the details pertaining to the transport mechanisms in the proposed InAs/GaSb/AlSb VMT device cannot be explained by the simple models used to describe the behavior of the GaAs/AlGaAs device. In particular, the heterogeneous dispersion relations of the two quantum wells and the intrinsic differences in the well materials can lead to unusual behavior of the states distributed in **k**-space, and a nontrivial sub-band structure. These effects will be explored in this chapter.

In Section 7.2 we present calculations of the subband structure in the InAs/GaSb/AlSb layers used in the device. These calculations were used in part in the design of the structures fabricated and characterized, as described in the next chapter. In Section 7.3, we will explore the behavior of a simple system in which the mean free time is strongly dependent on position in **k**-space, analogous to the InAs/GaSb/AlSb system.

7.2 Band Structure Calculations

7.2.1 Description of Calculations

All calculations presented here are based on an "Effective Bond Orbital" formalism developed for superlattice band structure calculations [1]. The method uses a modified 8-band $\mathbf{k} \cdot \mathbf{p}$ band structure calculation with periodic boundary conditions to allow modeling of a finite number of crystal layers, replicated throughout space in a given "growth" direction. Given a one-dimensional layer structure and a number of bulk band parameters for each of the composite materials, the resulting band structure

can be calculated. This structure consists of 8N subbands (where N is the number of layers), the energy of which can be evaluated at any point in the Brillouin zone. These bands are filled with 6N valence electrons per lateral unit cell, so if the bands are placed in order of increasing energy, those roughly 6N up from the lowest energy will appear near the bandgap. In addition, the envelope function for a state in a given band and at a given point in **k**-space may be calculated.

The proposed InAs/GaSb/AlSb VMT structure does not, of course, consist of repeated superlattice layers, but as the electron wavefunctions decay to zero in the surrounding AlSb barriers for all bands of interest in this device, the wavefunctions for a periodic system should be the same as those for a single instance of the layer structure, provided that a thick enough AlSb barrier layer exists between the repeated quantum well layers. It was found that 15 layers of AlSb cells (92 Å total) on each side of the layer structure were enough to eliminate coupling of wavefunctions in neighboring instances of the wells.

It should be noted that these calculations neglect any band bending due to charge distributions throughout the structure – flat bands are assumed. As charge will be transferred from the GaSb well to the InAs well at zero bias in the true InAs/GaSb/AlSb structure as a result of the broken gap alignment, the results of these calculations will more closely resemble the characteristics of the structure under enough bias to bring it towards flat band. In order to address this problem, a self-consistent model, in which this calculation is performed iteratively with a Poisson model which adds in a potential resulting from the charge distributed throughout the structure [2], was pursued, but never completed. As the calculations depend on a number of fitting parameters, it is questionable whether the results could be believed to be quantitatively precise in any event, but an investigation of the basic trends is certainly of interest. We also note that all layers were assumed to be strained to the AlSb lattice constant (6.136 Å), and that the induced strain energy was included in the band calculations.

7.2.2 Discussion of Results and Design Considerations

Several calculated sub-bands in the (100) direction in **k**-space (perpendicular to the "growth" direction) for two InAs/GaSb/AlSb VMT structures are plotted in Figure 7.1 . The structure represented in Figure 7.1(a) consists of AlSb layers surrounding a 123 Å GaSb well and a 184 Å InAs well coupled through a 31 Å AlSb barrier. The structure represented in Figure 7.1(b) is identical, minus the middle AlSb barrier. Note that the bands appear to repel one another. The stronger the coupling, the greater the effective repulsion, leading to a wider separation in energy. Note that the structure without the AlSb barrier between the wells exhibits much stronger band coupling. In addition, the spin degeneracy is broken down somewhat is this structure due to the strong coupling between states of different angular momentum character (*i.e.*, valence and conduction band states).

For **k** vectors somewhat away from zero (the right side of the figure) bands of different dispersion character can be clearly distinguished for both structures. The upper five pairs of bands, increasing in energy with increasing **k**, arise from the InAs and (for the highest energy) GaSb conduction bands, and the lower five pairs, decreasing in energy with increasing **k**, arise from the GaSb valence band. For lower **k** values, however, the character of the bands is mixed as the bands couple to one another (where indicated by arrows in the figure) when they are near one another in energy and **k**-space. The character of a given band can be seen to "transfer" to another band in the coupling regions, however. Consider, for example, the lowest energy InAs conduction subbands (13,14) at large **k** in Figure 7.1(a). If these bands are traced back to $\mathbf{k} = 0$, several coupling regions are encountered, and the conduction band dispersion character is transfered first to bands (11,12), and subsequently, (9,10).

The character of the bands is not only apparent in the direction of dispersion, but also in the spatial distribution of the corresponding envelope functions. Figure 7.2 illustrates the probability densities derived from the envelope functions for fixed **k**-space (.026 a/π, 0, 0) states in each band from Figure 7.1(a). Most of the states are localized in one well or the other, clearly arising from the conduction band of the InAs,

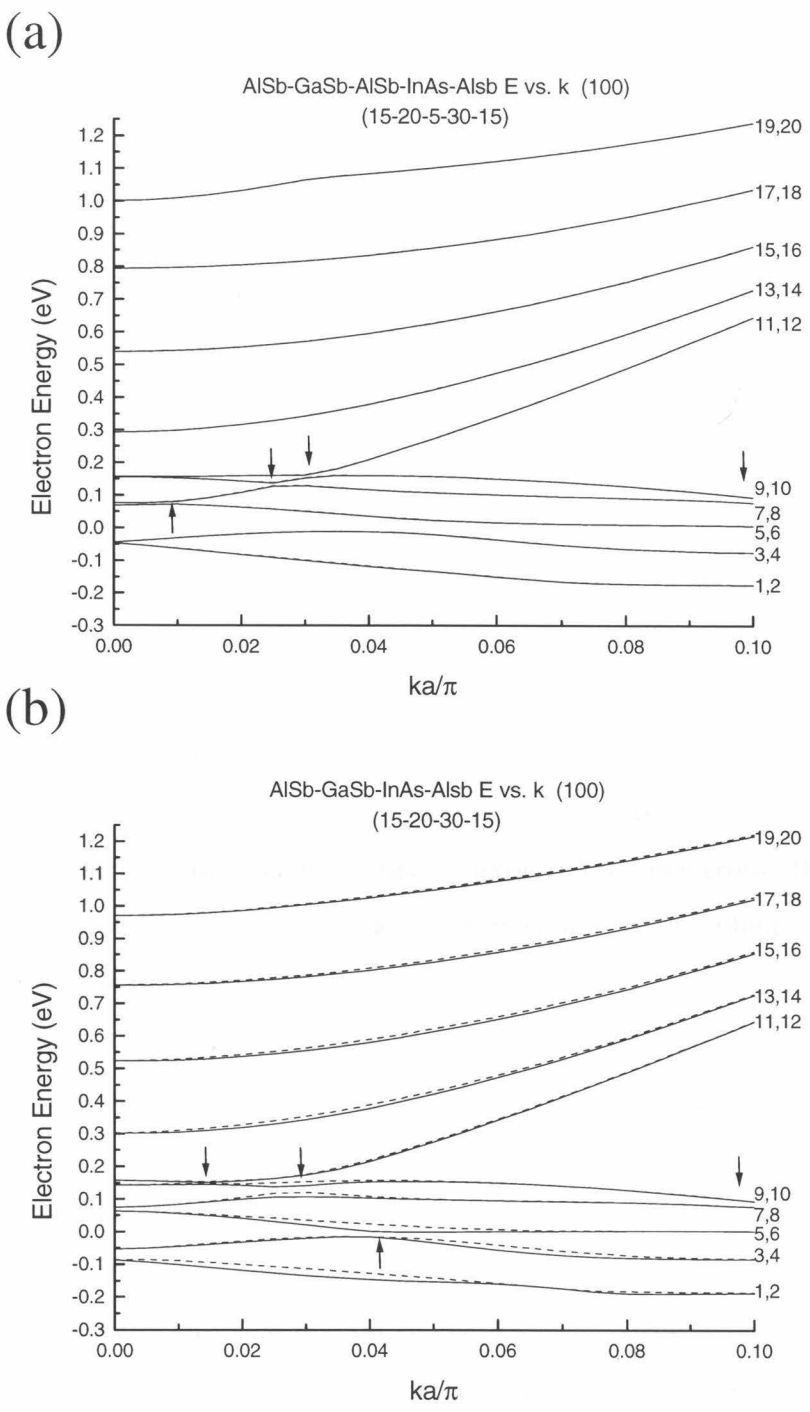

Figure 7.1: Several calculated sub-bands of (a) a VMT structure consisting of 15 layers of AlSb, 20 GaSb, 5 AlSb, 30 InAs, and 15 more AlSb, and (b) the same structure without the the intermediate AlSb layers. All layers have the thickness of the unit cell size (6.136 Å). The bands are plotted in pairs, indicated by solid and dashed lines, representing opposite spin polarities. Several regions of strong coupling are indicated by arrows. All bands are indexed in order of increasing energy for reference.

or the valence band of the GaSb. It is even possible to pick out the set of quantized states arising from each bulk band (*i.e.*, InAs conduction, GaSb conduction, GaSb heavy hole, and GaSb light hole) by examining the number of nodes in these envelope functions. The states in bands which are coupling at the chosen point in **k**-space, however, are spread across both wells, as expected.

The "transfer" of band character from one band to another can be seen by examining envelope function for a given band over a range of **k**-space, as shown in Figure 7.3 for coupling bands 8 and 10 from Figure 7.1(a). As **k** is increased, the envelope function of band 8 shifts from the InAs well to the GaSb well. The opposite is true for band 10, preserving the total probability density in each well at each point in **k**-space. Note that this kind of plot allows determination of the width of the coupling region in **k**-space, which reflects the coupling strength.

In all discussions thus far, and in most calculations performed, only the (100) direction in **k**-space has been considered. As the crystalline materials considered are not isotropic, the dispersion curves will vary somewhat as different directions are considered. While some effort was expended to explore the anisotropy, the same basic features were found to be present in all **k**-space directions in the plane parallel to the grown layers, so no extensive studies were performed.

As mentioned in the previous chapter, a larger effective mass can contribute to a lower mobility as well as a shorter mean free time. The band calculations presented here can be used to address this issue. Notice that the when two states of opposite dispersion couple, the slope of the bands must pass through zero in the interaction region (see Figure 7.1). Therefore, the effective mass of all such coupled states should be quite large, while the effective masses of the uncoupled states can still be similar to bulk values. This effect should enhance the decrease in mobility of the coupled system, thereby enhancing the mobility modulation effect.

In order to explore a range of design parameters for subsequent experiments, sub-band dispersion curves and envelope functions were calculated for a number of InAs/GaSb/AlSb VMT structures with varying layer thicknesses. The objective of this process was to design a structure that would be likely to exhibit the desired ve-

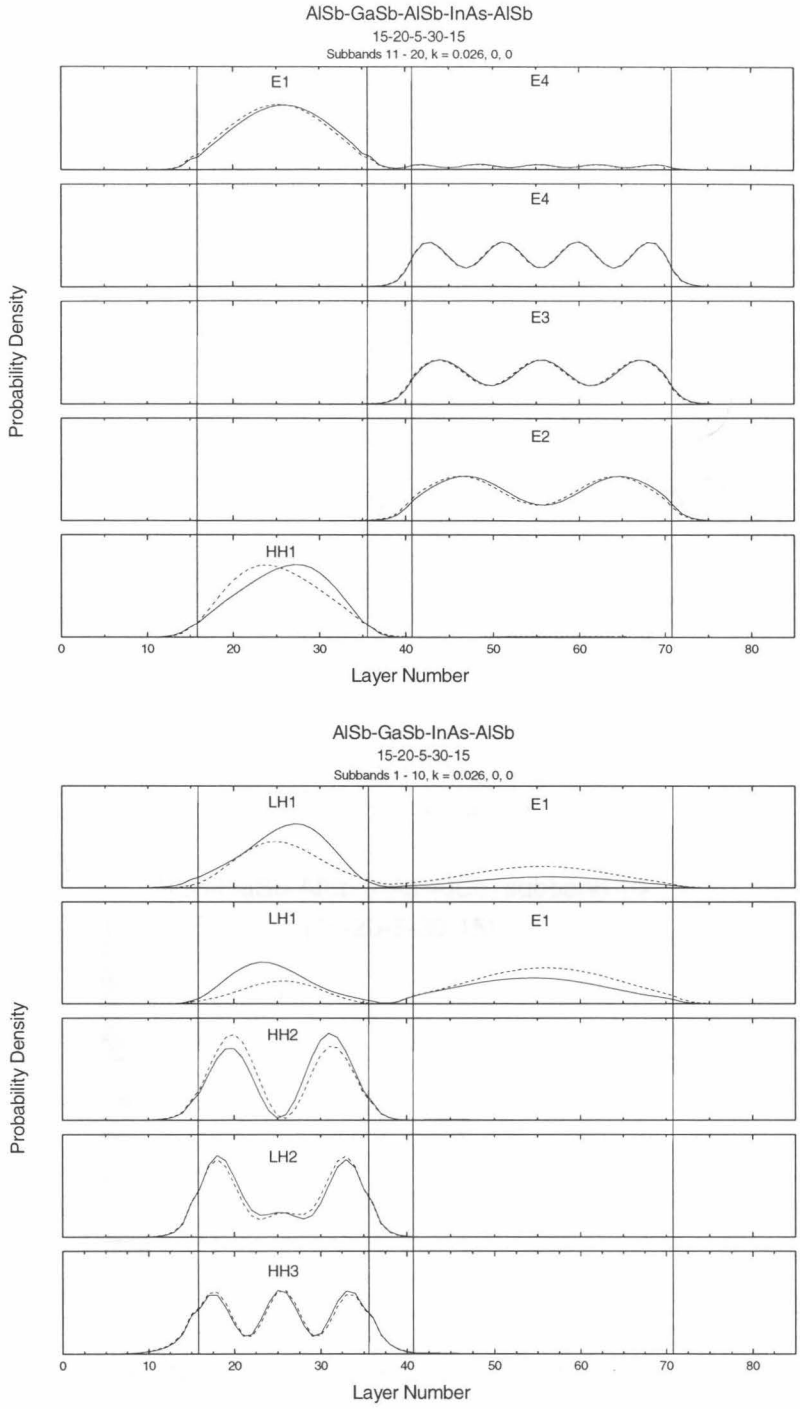

Figure 7.2: Calculated probability densities associated with each of the bands in Figure 7.1(a) at $\mathbf{k}$ = (.026 a/π, 0, 0), plotted as a function of layer number (*i.e.*, position in the layer structure). The plots are split into two groups for clarity, but are presented in order from the highest energy bands (top) to the lowest energy bands (bottom). The vertical lines indicate the layer interfaces. The character of the state in each well (E = electron, LH = light hole, HH = heavy hole) is indicated for each band, as well as an index of the quantized state of each character.

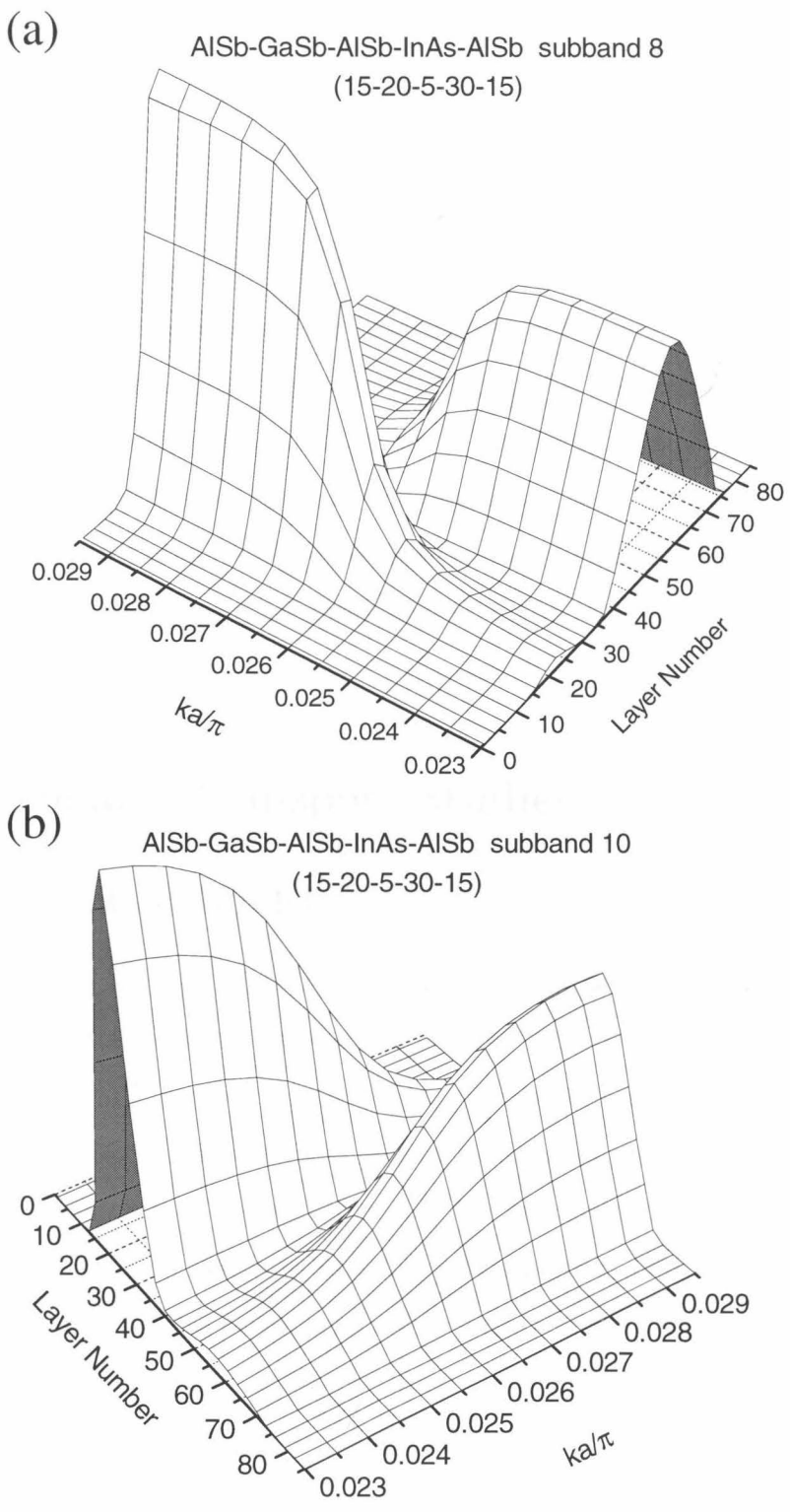

Figure 7.3: The probability density of states in bands (a) 8, and (b) 10 from Figure 7.1(a) is plotted as a function of layer number and **k**-space.

locity modulation effect under some range of possible operating parameters, even if those operating parameters were somewhat nonideal for practical applications. Given the fact that no reports of a similar device in this material system were known, and given the limitations of the calculations, the decision was made to design a system which possessed a reasonable number of strong resonances which could give rise to velocity modulation at some bias. Such a system, however, would likely have a number of bands relatively closely spaced in energy, so the system would have to brought to a low enough temperature to isolate the behavior of a given resonance. Therefore, the designed systems may require much lower temperature operation than limited by intrinsic physical properties, such as the maximum coupling strength of the InAs conduction band states to the GaSb hole states. After consideration of a number of structures, those already discussed (with 120 Å GaSb wells and 180 Å InAs wells) were chosen.

7.3 Boltzmann Transport Studies

7.3.1 Simple Band Model

The distribution in **k**-space of states of different character, as illustrated in Figure 6.6, suggests that simple transport concepts will not apply to the proposed InAs/GaSb/AlSb VMT device. In particular, it is not immediately apparent what kind of transport to expect in a system in which the mean free time is a strong function of **k**-space. In order to accurately model the operation of the VMT device, one would have to calculate the sub-band structure for a number of different gate voltages, determine the position of the Fermi level relative to the bands, compute the transport properties (*i.e.*, effective mass, mean free time, ...) for each of the bands based on calculated wavefunctions at different points in **k**-space, and sum up the contributions from each band. Such a calculation would require development of a fairly complex model, and a tremendous amount of computation time. As the various parts of this calculation would depend on a number of parameters that are relatively poorly known

in this system (*i.e.*, detailed band structure, details of scattering mechanisms, ...), it would likely only give a qualitative description of the system. For these reasons, a simple two-dimensional system with many of the same transport properties as the InAs/GaSb/AlSb coupled quantum well channels was developed in an attempt to understand qualitative transport behavior with relatively little computational effort.

As discussed in Section 7.2, the band structure of the InAs/GaSb/AlSb layers contains a large number of sub-bands, some of which lie in the energy range in which the GaSb states and the InAs states can interact. Two such bands are sketched in Figure 7.4(a): one containing localized InAs states at small $k_{||}$ values and GaSb states at high $k_{||}$ values, and the other with the opposite behavior. The wavefunctions of electrons in these bands are spread across both wells for the range of $k_{||}$ in which the bands are close in energy. The mean free time, τ, is sketched as a function of $k_{||}$ for each of these bands, assuming the intrinsic mean free time is lower in the GaSb well than in the InAs well (consistent with the discussion in Section 6.3.3).

A simple one-band model was developed to approximate the composite behavior of the more realistic two band system. A single parabolic energy band with the InAs conduction band mass was used (see Figure 7.4(c)), as the conduction will be dominated by the InAs well away from the resonance when the electrons are localized. The mean free time in this band was taken to be constant, with a notch at the resonance (Figure 7.4(d)), effectively a weighted sum of the contributions from the bands in the two band system.

In the true InAs/GaSb/AlSb system, an applied gate voltage would modify the sub-band structure, shifting the InAs and GaSb sub-bands with respect to one another, and the position of the Fermi level relative to the sub-bands would shift. In order to model an applied gate voltage in the simple one band system, however, a simpler method is used: the band structure is held fixed, while the Fermi level is shifted with respect to it. While this treatment neglects any changes in band structure, such as the shifting of the resonance states in **k**-space and energy, it allows the Fermi level to be swept past a resonance, which is the same qualitative effect expected in the true system. The increase of the Fermi level with respect to the band minimum in

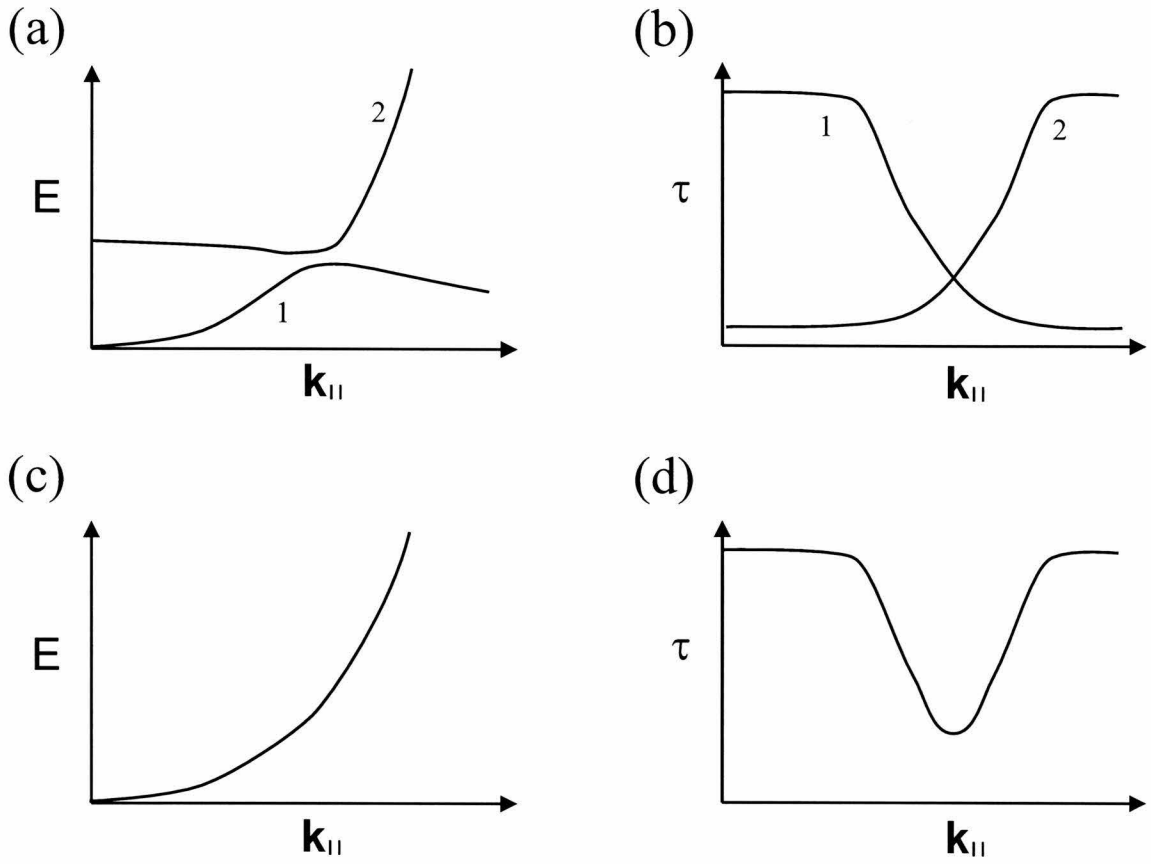

Figure 7.4: Models of parameters used for Boltzmann transport studies. (a) Energy versus $k_{||}$ dispersion relations for two interacting bands as expected in the InAs/GaSb/AlSb VMT structure. (b) Distribution of the mean free time versus $k_{||}$, corresponding to the bands in (a). (c),(d) One band approximations to the system described in (a),(b), used in the Boltzmann transport calculations.

the simple model will also cause an increase in the number of carriers in the system, analogous to traditional FET behavior, but the effect will likely be quantitatively inaccurate, as the relationship between an applied gate voltage and the Fermi level position is ignored in this simple treatment.

7.3.2 Boltzmann Transport Formulation

In thermal equilibrium, the properties of a system (of Fermions) can be described by the Fermi-Dirac distribution function, but when this equilibrium is disturbed by an external force, the distribution function is modified. The concepts of Boltzmann transport use the framework of statistical mechanics to describe semi-classical evolution of the distribution function, f, given a description of collision mechanisms and a model of the phase-space trajectory of a single particle in the absence of collisions. The path integral formulation of the evolution of f presented here follows that developed in [3].

Given a set of phase space coordinates, $\mathbf{r}, \mathbf{k}$, at a given time t, it is possible to compute coordinates, $\mathbf{r}_0, \mathbf{k}_0$ at some earlier time, t_0, from the equations of motion of a system, given that no collisions have occurred. Because f is invariant under time evolution in the absence of collisions (Liouville's Theorem), we can write

$$f(\mathbf{r}, \mathbf{k}, t) = f(\mathbf{r}_0, \mathbf{k}_0, t_0). \tag{7.1}$$

Now, we assume that particle collisions can be characterized by a mean free time, τ, and that a particle experiencing a collision is forced back to the equilibrium Fermi-Dirac distribution, f_0. Then, the distribution function for a set of particles which have suffered collisions at a time t_0 can be described by Equation 7.1, with the distribution function on the right side replaced by f_0. Contributions from all such collections of particles which have suffered collisions at all possible past times can be summed, weighted by the probability that the requisite amount of time had elapsed since the last collision without any subsequent collisions. This is the physical interpretation of

the following result, in the case of a constant τ:

$$f(\mathbf{r},\mathbf{k},t) = \int_0^\infty f_0(\mathbf{r}_0,\mathbf{k}_0,t-t')e^{-t'/\tau}\frac{dt'}{\tau} \tag{7.2}$$

where $\mathbf{r}_0, \mathbf{k}_0$ are phase space coordinates at time t' which evolve to coordinates $\mathbf{r}, \mathbf{k}$ at time t in the absence of collisions. When τ is allowed to be a function of phase space, a modified version of this equation results from averaging the weighting function over the traversed path:

$$f(\mathbf{r},\mathbf{k},t) = \int_0^\infty f_0(\mathbf{r}_0,\mathbf{k}_0,t-t')\exp\left(-\int_{t-t'}^{t}\frac{ds}{\tau(s)}\right)\frac{dt'}{\tau(t-t')} \tag{7.3}$$

where $t' \equiv t - t_0$, and now τ is a function of time as it depends on the phase space coordinates of a particle following a trajectory over time, determined by the equations of motion for the system.

Once the distribution function for a system is determined, other quantities of interest can be ascertained. In this case, we are interested in calculating the conductivity of the system: the ratio of the current density in a system to the electric field that brought about the current flow. The current density along, say, the x-direction in a two-dimensional system is given by

$$j_x = q\int_{\mathbb{R}^2}\frac{dk_x dk_y}{(2\pi)^2}f(\mathbf{r},\mathbf{k},t)v_x \tag{7.4}$$

where $v_x = \frac{1}{\hbar}\frac{\partial E}{\partial k_x}$. In a crystal, the integration should really be performed over the Brillouin zone volume, but if the Fermi level is much lower than the edge of the band, an integration over all space is a very good approximation.

Before Equation 7.3 is substituted into Equation 7.4, some useful transformations can be made given the specifics of the system under consideration. First, if we assume that the electric field is applied along the x-axis, then the relations between k_x, k_y and

k_{x0}, k_{y0} (as defined above) are as follows:

$$k_{x0} = k_x + t' q\mathcal{E}, \tag{7.5}$$

$$k_{y0} = k_y. \tag{7.6}$$

Then, realizing that in the steady state of this system, f will not depend on t or spatial variables, Equation 7.3 can be written as

$$\begin{aligned} f(k_x, k_y) = \int_0^\infty & f_0(k_x + \frac{q}{\hbar}\mathcal{E}t', k_y) \\ & \times \exp\left(- \int_{t-t'}^{t} \frac{ds}{\tau(k_x + \frac{q}{\hbar}\mathcal{E}(t-s), k_y)} \right) \\ & \times \frac{dt'}{\tau(k_x + \frac{q}{\hbar}\mathcal{E}t', k_y)}. \end{aligned} \tag{7.7}$$

With a change in variables ($s \to t - s$), this becomes

$$\begin{aligned} f(k_x, k_y) = \int_0^\infty & f_0(k_x + \frac{q}{\hbar}\mathcal{E}t', k_y) \\ & \times \exp\left(- \int_0^{t'} \frac{ds}{\tau(k_x + \frac{q}{\hbar}\mathcal{E}s, k_y)} \right) \frac{dt'}{\tau(k_x + \frac{q}{\hbar}\mathcal{E}t', k_y)}. \end{aligned} \tag{7.8}$$

An integration by parts then yields

$$\begin{aligned} f(k_x, k_y) = & f_0(k_x, k_y) \\ & + \frac{q}{\hbar}\mathcal{E} \int_0^\infty \frac{\partial f_0}{\partial [k_x]}(k_x + \frac{q}{\hbar}\mathcal{E}t', k_y) \\ & \times \exp\left(- \int_0^{t'} \frac{ds}{\tau(k_x + \frac{q}{\hbar}\mathcal{E}s, k_y)} \right) dt' \end{aligned} \tag{7.9}$$

where $\frac{\partial f_0}{\partial [k_x]}$ is the derivative of f_0 with respect to its first argument:

$$\frac{\partial f_0}{\partial [k_x]}(k_x + \frac{q}{\hbar}\mathcal{E}t', k_y) \equiv \frac{\partial f_0}{\partial u}(u, v) \Big|_{(u,v)\to(k_x + \frac{q}{\hbar}\mathcal{E}t', k_y)}. \tag{7.10}$$

The second term in Equation 7.9 represents the departure from the equilibrium distribution induced by the applied electric field. This is the only term that will contribute

to the current density – when the first term is substituted into Equation 7.4, the integrand is odd in k_x, so the integral reduces to zero. When this second term is substituted into Equation 7.4 and divided by $\mathcal{E}$ to obtain the conductivity, the result is

$$\sigma = \frac{q^2}{m^*}\int_{\mathbb{R}^2}\frac{dk_x dk_y}{(2\pi)^2}\int_0^\infty dt' k_x \frac{\partial f_0}{\partial[k_x]}(k_x + \frac{q}{\hbar}\mathcal{E}t', k_y) \times \exp\left(-\int_0^{t'}\frac{ds}{\tau(k_x + \frac{q}{\hbar}\mathcal{E}s, k_y)}\right) \tag{7.11}$$

where the parabolic band dispersion relation

$$E = \frac{\hbar^2\mathbf{k}^2}{2m^*} \tag{7.12}$$

was used. Note that it is possible for σ to depend on the electric field, so it is necessary to evaluate the expression for several electric fields to determine whether the conduction mechanism is ohmic. Also note that the dependence of the expression on the Fermi level is hidden in the definition of f_0.

Equation 7.11 was evaluated numerically using successive Romberg integration [4]. In order to factor out the effects of increasing charge density in the system, the result could be normalized to the result obtained from a constant mean free time,

$$\sigma_0 = \frac{q^2 n\tau}{m^*} \tag{7.13}$$

where

$$n = \int_{\mathbb{R}^2}\frac{dk_x dk_y}{(2\pi)^2} f_0(k_x, k_y). \tag{7.14}$$

7.3.3 Results

The mean free time, τ, was taken to have the following functional form:

$$\tau = \tau_0\left(1 - \frac{\Delta}{1+u^2}\right) \tag{7.15}$$

Table 7.1: Parameters used in the Boltzmann transport calculations.

τ_0 (s)	Δ —	k_0 (cm^{-1})	Δk (cm^{-1})	$\frac{\hbar^2 k_0^2}{2m^*}$ (eV)	$\frac{\hbar^2 k_0 \Delta k}{m^*}$ (eV)	m^* (m_e)
10^{-13}	0.9	2.5×10^6	2.0×10^5	0.1	0.016	0.023

with

$$u = \frac{\sqrt{{k_x}^2 + {k_y}^2} - k_0}{\Delta k} \tag{7.16}$$

where the various parameters were given the values listed in Table 7.1. This function is plotted versus energy in Figure 7.5(a). The parameters control the position, width, and depth of the "resonance." As indicated in the table, the width of the "resonance" is 0.016 eV in energy, or about 190 K when converted to a temperature.

The conductivity of the system was calculated for a range of electric fields, Fermi level energies, and temperatures. The results of these calculations are summarized in Figure 7.5. The plot of calculated conductivity (σ) and the normalization conductivity (σ_0) versus the Fermi level energy (Figure 7.5(b)) is analogous to a plot of conductivity versus gate voltage, as in Figure 6.2. Note that when the system temperature is comparable to or below that of the width of the "resonance," a strong modulation is observed, but as the temperature is increased above this limit, the effect is washed out, as predicted in the previous chapter.

The plots of σ/σ_0 in Figure 7.5(c) illustrate the velocity modulation effect separated from the effect of increasing channel charge. From these plots, it is apparent that at higher temperature, the mean free time notch contributes to a drop in conductivity averaged over a large range of Fermi level energies. The downward shift in the energy at the conductivity minimum is likely due to the anisotropy of the density of states, as the band edge exists at $E = 0$.

As the conductivity versus Fermi level data were taken at a fixed electric field (200 V/cm), any dependence of conductivity on electric field is neglected. Plots of conductivity versus electric field at 300 K are shown in Figure 7.5(d) for a few

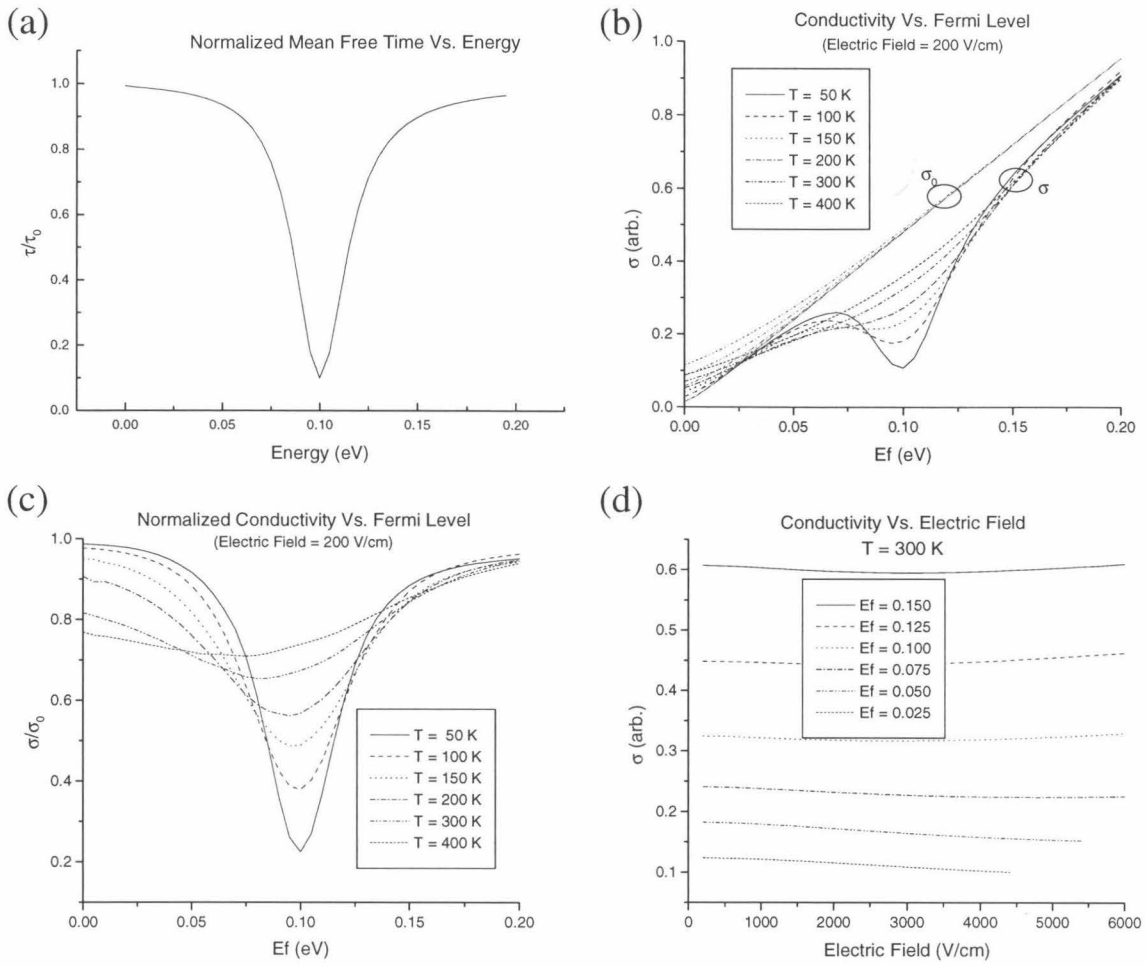

Figure 7.5: Results of Boltzmann transport calculations. (a) Functional form of the mean free time versus energy. (b) Calculated conductivity, σ, and reference conductivity, σ_0, as a function of Fermi level for a range of temperatures. (c) Same data as in (b), plotted as σ/σ_0. (d) Calculated conductivity versus electric field for a range of Fermi level energies. The curves for the lowest Fermi levels end at the point at which the simulation failed to converge.

different quasi Fermi level energies. The conductivity is relatively insensitive to electric field, although a relatively small (<20%) drop in conductivity with increasing field is noticeable at the lowest Fermi levels. This drop is likely due to the fact that a large electric field appears to disrupt the equilibrium distribution more if there are fewer carriers in the band, distributed over a smaller range of **k**-space.

In summary, these calculations confirm that a system with a mean free time which is consistent with the proposed InAs/GaSb/AlSb system can exhibit velocity modulation as long as the temperature of the system is below the equivalent temperature corresponding to the width of the reduced mobility region in energy. A more realistic model could be designed to give a more accurate description of the true InAs/GaSb/AlSb system, perhaps by introducing a more realistic band structure involving at least two interacting bands, but these results at least suggest that the proposed system could behave as expected, at a low enough temperature.

Bibliography

[1] Y.-C. Chang, *Phys. Rev. B* **37**, 8215 (1988).

[2] Y. Rajakarunanayake and T. C. McGill, *J. Vac. Sci. Technol. B* **4**, 1288 (1987).

[3] F. Reif, *Fundamentals of Statistical and Thermal Physics*, McGraw-Hill, New York, 1965.

[4] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, Academic Press, New York, 2nd edition, 1992.

Chapter 8 Experiment

8.1 Introduction and Outline

This chapter describes our attempts to fabricate and characterize the proposed InAs/GaSb/AlSb VMT device. Unfortunately, a working device was never produced, but much progress was made along the way, and in the latest experimental attempts, many of the early problems were overcome. Therefore, the results presented in this chapter may prove useful as a guide for future work.

The growth of the layered structures used in these experiments and the characterization of these samples is presented in Section 8.2. A large number of problems were encountered during processing of these samples. While most were eventually overcome or avoided, in the interest of completeness, the various failed attempts will be documented in Section 8.3. Finally, an analysis of the experimental problems and a discussion of closely related work in the literature follow in Section 8.4.

8.2 Crystal Growth and Characterization

All crystal samples were grown by Molecular Beam Epitaxy (MBE) in a UHV Perkin-Elmer system. Five samples were grown, the first three by Doug Collins, and the last two by Xiao-Chang Cheng (see Table 8.1 for details), but only two were actually used in fabrication of VMT device structures.

A semi-insulating GaSb substrate was used for all samples. A GaSb buffer layer was grown on top of the substrate, followed by a thick AlSb layer, the InAs quantum well, the AlSb barrier layer (where used), the GaSb quantum well, another thick AlSb layer, and a cap layer of AlGaSb or GaSb to prevent the AlSb from oxidizing. The growth of the GaSb well was interrupted after 70% of the layer had been grown, and the sample was subjected to an As soak in order to decrease the GaSb well mobility.

Table 8.1: MBE-grown samples used for VMT device fabrication. The cap layer grown on samples III-691, III-700, III-701 was $AlGa_{0.7}Al_{0.3}$, while GaSb was used for the remaining two. Sample III-727 contained additional grown layers beneath the quantum well layers to allow fabrication of a back-gated structure.

Sample Label	InAs Well Thick. (nm)	AlSb Barrier Thick. (nm)	GaSb Well Thick. (nm)	Top AlSb Layer Thick. (nm)	Cap Layer Thick. (nm)
III-691	18	3	12	25	3
III-700	18	1.8	12	50	3
III-701	18	3	14.5	50	3
III-726	18	—	12	50	10
III-727*	18	—	12	50	10

Several angstroms of each layer at each interface were grown by Atomic Layer Epitaxy (ALE) to ensure high crystal quality. The InAs quantum well interfaces were grown to be InSb-like. During growth, the monitored reflection high energy electron diffraction (RHEED) patterns suggested a high quality crystalline structure.

Hall measurements were performed on the GaSb substrates used for all sample growths, and on sample III-691. Square pieces of each sample were supplied with ohmic Au/Ge alloyed contacts (see Section 8.3), and measured using the Van-der-Pauw method at room temperature and at 77 K. The results are presented in Table 8.2. As the grown sample consists of a number of layers, the Hall data require some interpretation. The sheet resistance of the grown sample was found to be roughly the same as that of the GaSb substrate at room temperature. This suggests that the alloyed contacts reach down to the substrate, and that the rest of the device structure is shorted out by the substrate at this temperature. At 77 K, however, the substrate carriers are shorted out, while the wells remain heavily populated with carriers as a result of the broken gap alignment. Note that the lack of an insulating substrate and the use of alloyed contacts imposes a temperature constraint on observation of velocity modulation, perhaps more stringent than that of the coupled well system, as the substrate carriers in parallel with the wells must be frozen out before the well conduction can dominate.

The 77 K hall data from the grown crystal was interpreted by assuming that only the InAs quantum well contributed to conduction, as the GaSb well mobility

Table 8.2: Results of Hall measurements on the GaSb substrate, and on sample III-691. Bulk values were obtained from measured "sheet" values by assuming a thickness of 300 μm for the GaSb substrate, and 180 Å (the InAs well thickness) for the low temperature measurements on the layered structure.

	GaSb	GaSb	III-691	III-691
Temp.	300 K	77 K	300 K	77 K
N_{sheet} (cm^{-2})	+ 2.6×10^{15}	+ 2.4×10^{12}	- 2.6×10^{15}	- 3.1×10^{13}
N (cm^{-3})	+ 8.7×10^{16}	+ 8.1×10^{13}	—	- 1.7×10^{19}
ρ_{sheet} (Ω)	5.6	6000	3.4	45
ρ (Ω-cm)	0.19	181	—	8.2×10^{-5}
μ (cm^2/V-s)	370	430	—	4500

is much lower. While the dominant carriers were found to be negatively charged, supporting this assumption, the measured mobility of the system was quite a bit lower than expected for an InAs well, so it is possible that this assumption is inaccurate. Note that the low measured mobility value results in a rather high calculated carrier concentration – another indication that the InAs well might not be the dominant conduction path. One possible explanation for the lower mobility is that at zero bias, the states contributing to conduction might happen to be coupled GaSb-InAs states. This would not necessarily explain the large carrier concentration, however. Another possibility is that there are carriers present in the AlSb layers. While this seems unlikely at lower temperatures, there is some evidence that AlSb tends to be a poor electrical barrier, suggesting the possibility of the presence of mobile charge [1–3].

8.3 Processing Techniques

8.3.1 Early Attempts

The earliest fabrication attempts were aimed at producing the VMT device structure shown in Figure 8.1. The source/drain contacts to the quantum well channels are Au/Ge pads deposited on top of the crystal layers, alloyed down into the sample. The gate consists of a large contact pad and a finger that lies between the source and drain pads. The entire device sits on a mesa which eliminates the possibility of conduction

in the channel around the gate.

The following fabrication procedures were developed in order to produce the target structure. For all early attempts, samples III-691, III-700, and III-701 were used. Small chips (~10 mm^2) were cleaved from the grown crystal sample and cleaned in acetone, isopropanol, and deionized water. After baking dry at 85°C, AZ 4210 photoresist was spun on, exposed, and developed in MIF 422 developer to form windows for the source/drain contacts. Prior to developing, the photoresist was soaked in C_6H_5Cl to harden the surface, creating a favorable cross-sectional profile for lift-off. Several hundred nm of Au/Ge was thermally evaporated onto the sample, and the photoresist removed with acetone, leaving Au/Ge contact pads. In order to incorporate the Au/Ge into the sample, making contact to the buried channels, the sample was annealed at 405°C for 30 s in an AG 310 Minipulse rapid thermal processing (RTP) oven in an Ar ambient.

Photoresist was spun on and developed again to pattern windows for the gate contact. Again, the metal (Al) was deposited by thermal evaporation to a thickness of several hundred nm, and the photoresist and excess metal removed. Once again, photoresist was applied, exposed, and developed in order to define a photoresist mask for an etch step which would create the necessary mesa structure. At this point, it was discovered that the MIF 422 developed etched away the part of the Al gate that was not covered by the photoresist pad, however. In order to avoid the problem of etching of the gate, Au was substituted as the gate material in subsequent processing attempts. Al had been chosen as the gate material because of its large Schottky barrier to GaSb (0.65 eV) [4], suggesting a large Schottky barrier to AlSb as well, but it was hoped that the lower barrier Au gate would allow for enough range of applied voltage to see the desired modulation.

With this substitution of gate material, processing could be taken further. After definition of the photoresist etch mask, the mesa structure was produced by etching the surrounding layers in Br:HBr:H_2O (1:100:100) to a depth of 750-800 nm as measured by a Tencor Alpha Step 500 profiler. After removal of the photoresist, the chips were mounted on 8 pin headers using a conductive silver paste, and wirebond contacts were

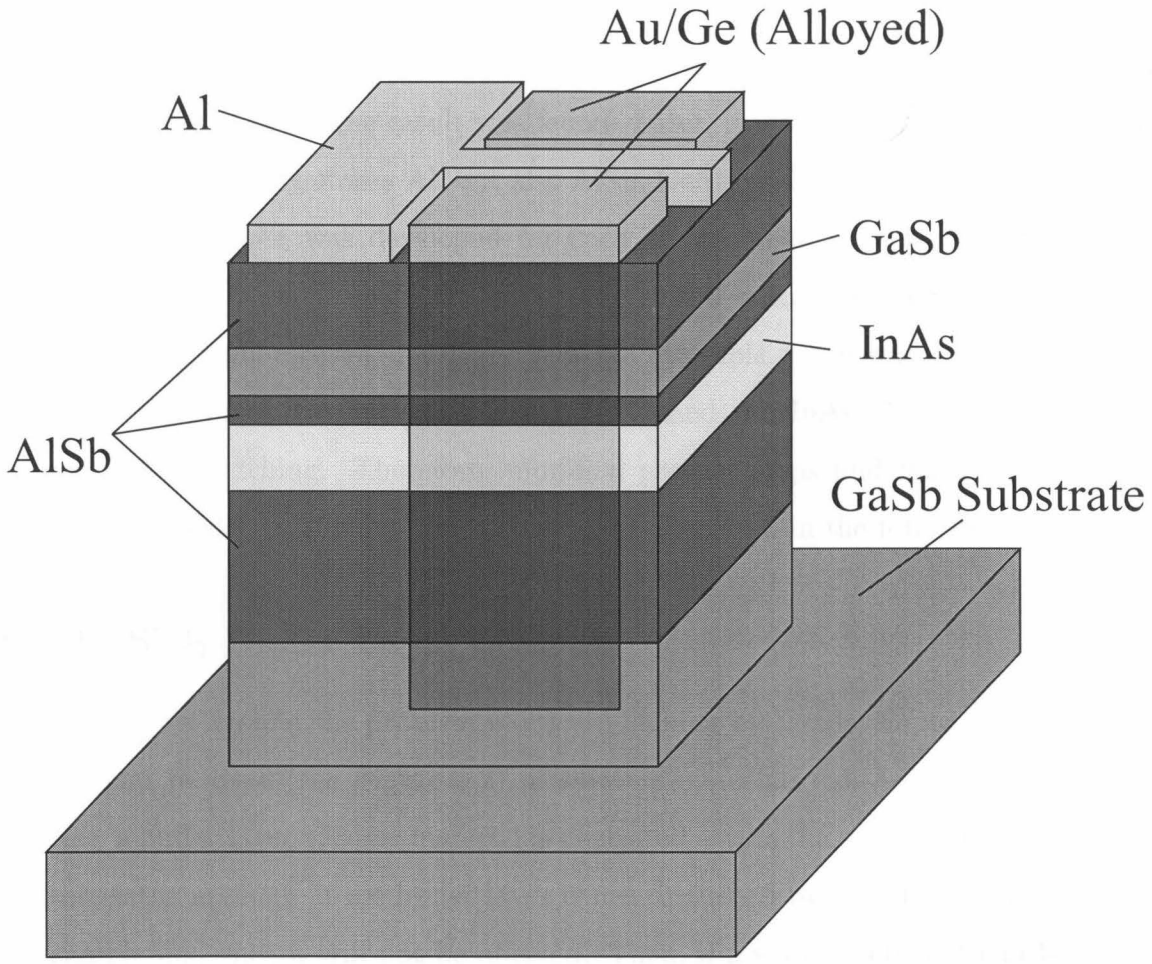

Figure 8.1: Target geometry for earliest fabricated devices. With the maskset used, the source and drain pads were 60×135 μm^2, the spacing between the source and drain was 12 μm, and the gate finger width was 6 μm.

made from the header pins to the device contacts. Header mounting was required in order to connect the devices to a probe suspended in a Janis dewar, allowing temperature control from 4.2 K to room temperature.

At room temperature, the gate-source/drain resistance was found to be comparable to the source-drain resistance, suggesting that the gate was shorted to the quantum wells. It was hoped that this could have been caused by carriers present in the AlSb layers which might freeze out at lower temperatures. When the samples were cooled to 4.2 K, however, the same result was found. Later, it was determined that the MIF 422 developer not only etches Al, but also AlSb, GaSb, and possibly InAs. Therefore, when the photoresist was developed for each of the metal lift-off steps, the active crystal layers, at least down to the InAs, were etched away, destroying the device.

While other photoresist/developer sets are available, none was known with certainty to be compatible with processing of Al and the InAs/GaSb/AlSb material system without etching. Therefore, modified process steps and mask designs were developed in order to alleviate this problem, as discussed in the following sections.

8.3.2 SiO_2 Buffer Layer Scheme

In order to get around the problem of crystal etching by the developer, a buffer layer scheme can be used (see Figure 8.2), if suitable materials can be found. With this process, a buffer layer of some material is deposited across the entire sample before the photoresist is applied. If the buffer layer is not disturbed by the lithography process, the sample underneath will not be affected. Then, if a suitable etchant can be found which etches the buffer layer, but not the underlying sample, then metal deposition and lift-off can be continued as usual.

It was found that SiO_2 was effective as a buffer material, using dilute HF:H_2O (1:10 - 1:100) as the selective etchant. The HF was found to etch both the GaSb substrates, and the grown crystal structures very slowly – at most 50 nm or so over 10 minutes in a 1:10 solution – and the MIF 422 developer was found to have a negligible effect on the SiO_2. The SiO_2 layer could be deposited on the crystal samples

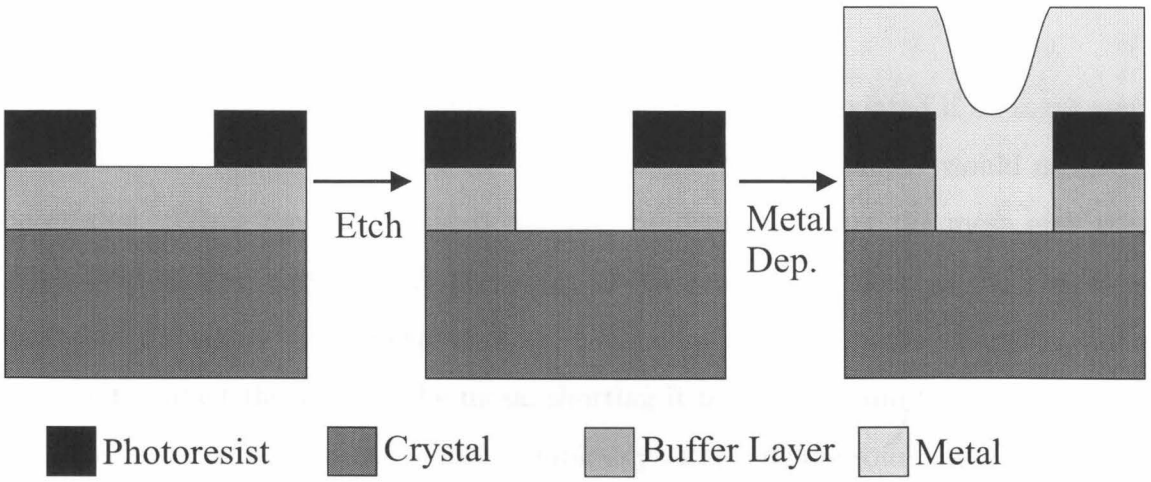

Figure 8.2: Graphical description of the use of a buffer layer in lithographic definition of metal pads by lift off.

by RF power magnetron sputtering of a Si target in an Ar:O_2 ambient. The etch rate of this sputtered SiO_2 layer in the HF solution was found to be much greater than that of thermal SiO_2 on Si, most likely due to a porous structure. It should be noted that before photoresist was spun onto the SiO_2 layer, the sample had to be baked in order to remove any moisture from the surface, or else adhesion would be poor, and the photoresist would delaminate during developing.

In addition to eliminating the problem with the developer etching the crystal materials, this buffer layer process can be modified slightly to improve the electrical isolation between the gate and the quantum wells. If the SiO_2 layer is not etched away after the lithography process, it remains underneath the gate metal, providing additional electrical insulation. As mentioned in Section 8.2, the AlSb layer has been shown to be a poor electrical barrier, so this process step was adopted.

While the buffer layer process solves the problem of etching of the crystal layers by the developer, the etching of the Al gate is not solved by this process. In order to address this issue, a modification to the mask design was made, as discussed in the following section.

8.3.3 Modified Mask Designs

The problem of the developer etching the Al gate could be alleviated if no mask steps were necessary following the gate metal deposition, as the developer would no longer be needed. While the gate could potentially be deposited after the mesa etch step, this would require very careful alignment of the gate to the mesa as the gate must extend to the edges of the mesa in order to cut off conduction paths around it, but it must not contact the sides of the mesa, shorting it to the quantum well channels. An easier solution is to use a gate which completely surrounds the source and drain pads, eliminating the need for a mesa structure.

A maskset allowing this geometry was produced, as illustrated in Figure 8.3(a). The gate consists of a mesh which extends across the entire sample surrounding the source and drain squares. Certain section of the mesh contain large pads to allow contact to the gate. Samples were produced with this mask set, using the buffer layer process described above, but the gate was again found to short to the source and drain contacts. This source of this problem was the fact that during gate metal deposition, some metal would coat the sides of the chip. As the gate mesh was a single contiguous piece of metal, this shorting at the edge caused the entire gate mesh to be shorted to the channels. This problem was solved by designing a modified mask set with the gate mesh broken up into a number of discrete pieces (Figure 8.3(b)).

The use of a process that does not require an etch step has other advantages. The Br:HBr:H_2O solution used to etch the InAs/GaSb/AlSb materials does not mix particularly well, and as a result, the etching can often be quite nonuniform. This effect is occasionally so detrimental as to render the etched sample unusable. In addition, mesa structures can often exhibit leakage along the etched edges as a result of surface states, so a planar device is often preferable.

8.3.4 Refined Fabrication Steps

In addition to the modifications of the process steps discussed in the previous sections, a few other minor changes were made from the earliest set of process steps. Therefore,

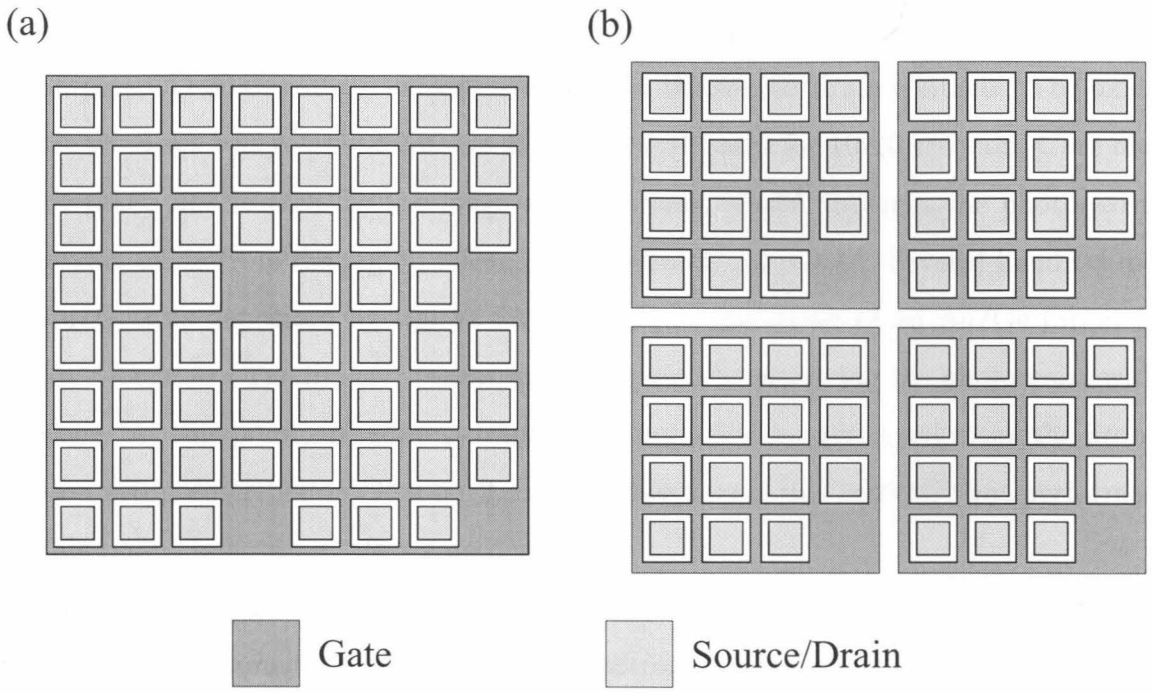

Figure 8.3: Modified mask designs. (a) The gate is a mesh extending across the entire sample, surrounding source and drain pads. (b) Same as (a), with the gate mesh broken up into a number of discrete sections. In each case, the source/drain squares were 145 μm on a side, the spacing between squares was 20 μm, and the width of the gate mesh was 8 μm.

the entire modified fabrication process will be described here. All samples produced using this process used MBE sample III-726.

Small pieces of the grown crystals were cleaved, cleaned, and baked dry as discussed previously. Then a SiO_2 buffer layer, roughly 100 nm thick, was deposited by magnetron sputtering of a Si target in a 100 W RF plasma, consisting of Ar:O_2 (160:3) at a pressure of 4×10^{-3} Torr. The base pressure in this sputtering system was typically 3×10^{-8} Torr. The sample was then baked at 120-140°C for several hours to drive off any moisture. Shipley 1813 photoresist was spun on, exposed, and developed in Shipley CD-30 developer to form windows for the source/drain contacts. The oxide in the windows was removed with an HF:H_2O (1:10) etch for 10-20 seconds. Care had to be taken not to etch for too long, or the etchant would undercut the photoresist, removing oxide from the region where the gate was to be placed. Several hundred nm of Au/Ge was deposited by 100 W DC magnetron sputtering of an Au/Ge target in 4×10^{-3} Torr Ar. The lift-off process was completed by removing the photoresist with acetone. At this point, the source-drain resistance could be measured using the probe station setup described in Chapter 3. The Au/Ge anneal was typically not performed at this point, but rather, after gate deposition.

A second lithography step defined the gate pattern. An additional 50 nm of SiO_2 was deposited immediately before the magnetron sputtering of several hundred nm of Al in order to further insure against gate shorting. The photoresist was removed, leaving the Al gate mesh structure on top of 150 nm of oxide. All contacts could be tested again on the probe station to verify that the gate was not shorted to the source and drain contacts. The Au/Ge was then alloyed into the crystal in a small bell jar system in Ar ambient, heated on a tungsten filament to 405°C for 30-60 seconds. Care had to be taken not to over-anneal the sample, as this resulted in very high contact resistances. One possible explanation for this is that the Au from the Au/Ge and the Al from the AlSb layer may form an alloy which is known to have low conductivity [5]. After annealing, the completed chip was then mounted on a header and wirebonded as discussed above.

8.4 Results and Discussion

8.4.1 Characterization of Samples

The modified procedure described above allowed successful production of devices in the sense that the gate was not shorted to the source and drain. Up to 10-20 V could be applied to the gate with no noticeable current transport to the source or drain pads. At higher voltages, however, the barrier would break down, allowing some current to flow. This breakdown voltage corresponds to a breakdown field of 1-2 MV/cm – a fair amount lower than typical values of 10 MV/cm for thermal oxide (see Chapter 2). This might be expected as the sputtered oxide is undoubtedly of poorer quality, as evidenced by the higher etch rate mentioned above. It is also possible that some etching of the oxide under the gate occurred during the HF etch of the oxide in the source/drain windows, as the etch typically undercut the photoresist barrier somewhat. This could have led to weak spots in the oxide barrier, causing lower breakdown fields.

Unfortunately, no velocity modulation effects were observed in any samples produced at any measured temperature (from 4.2 K to room temperature). In fact, the applied gate voltage had little effect on the source/drain properties at all. If the resistance of the source to drain of a device was monitored with a digital multimeter while the gate voltage was varied, small changes in resistance could be measured. This effect was determined to be due to injection of additional current from a slightly leaky gate contact into the channel, however, rather than any field effect. There are a number of possible explanations for the fact that no velocity modulation was seen. In order to explore these possibilities, first consider the following additional information.

Cross-sectional TEM was performed on early samples to verify that the Au/Ge anneal step was alloying the source/drain contacts into the crystal as expected. TEM images indicated features in some isolated spots underneath the pads, spiking down roughly 1 μm into the GaSb substrate (see Figure 8.4(a)). Chemical analysis by Electron Dispersion X-ray (EDX) confirmed that these spikes contained gold. Gold incorporation was found to be greater in the GaSb substrate than in the thick AlSb layers (see Figure 8.4(b)).

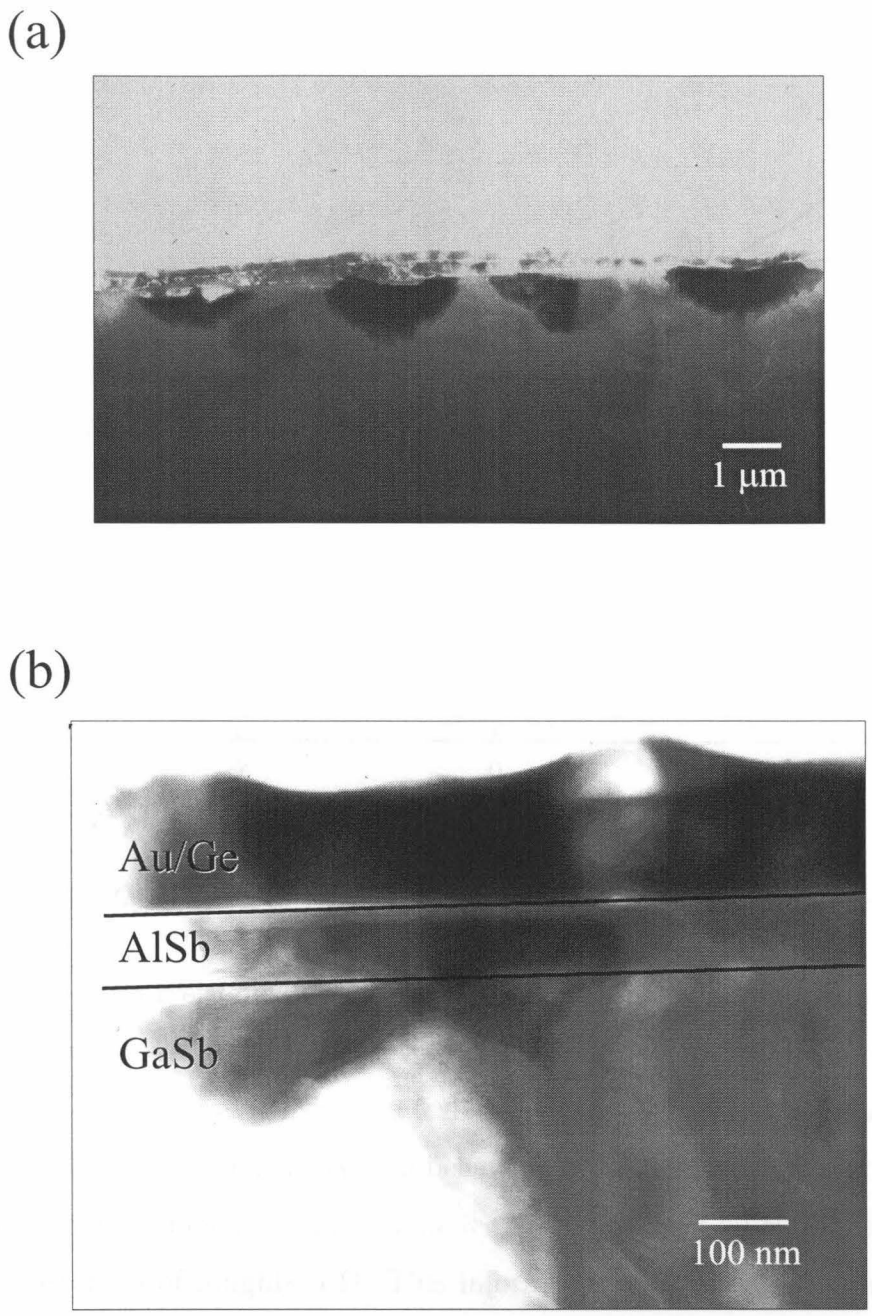

Figure 8.4: TEM images of a region of an alloyed Au/Ge contact on a processed piece of sample III-691. (a) An extended view shows spiking of the metal into the semiconductor. (b) A closer view allows resolution of the GaSb substrate, the thick AlSb layer, and the Au/Ge contact on top. Note that this sample does not contain a SiO_2 buffer layer, so all of the active layers, at least down to the InAs layer and perhaps all the way to the thick AlSb layer, were etched by the developer before the Au/Ge pad was deposited.

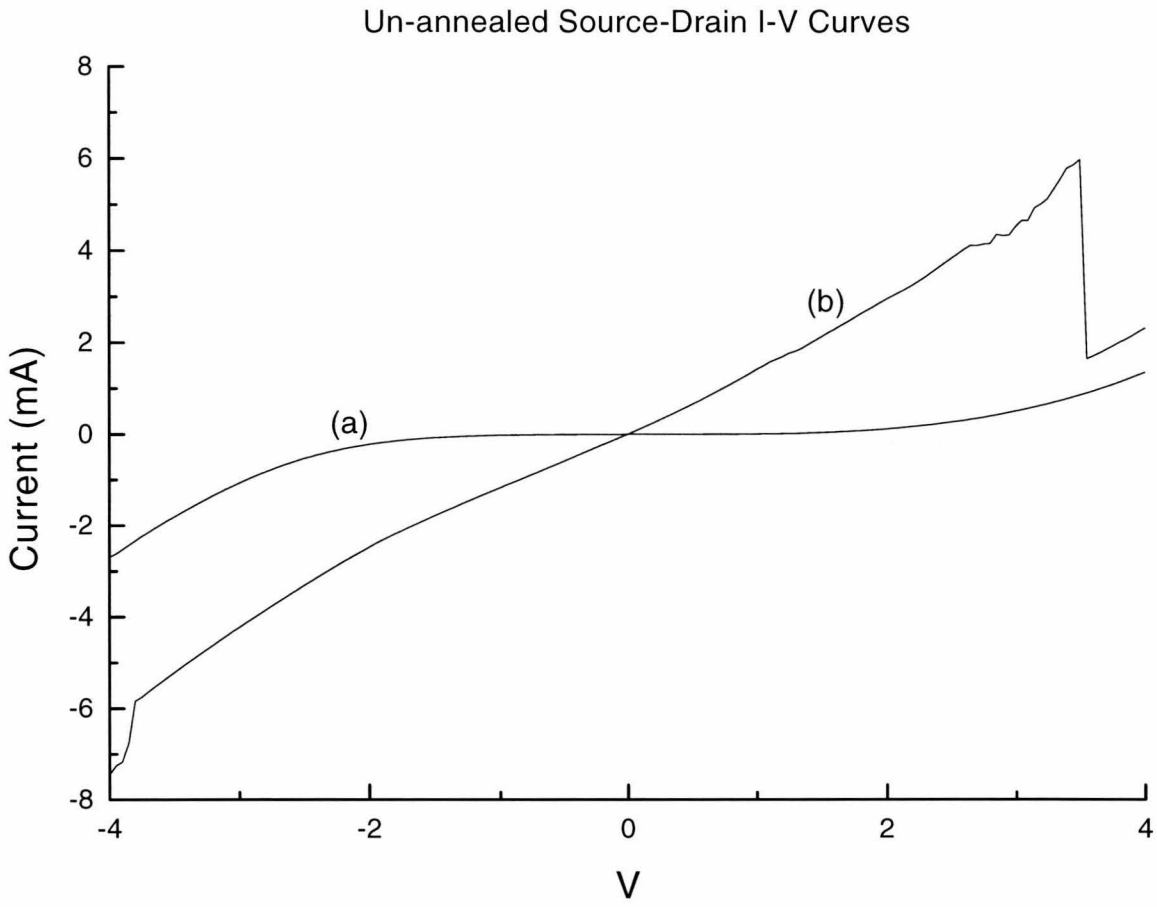

Figure 8.5: I-V curves of two pairs of un-annealed source-drain contacts.

The source-drain resistance was found to be relatively low, even before the anneal step was performed. A couple I-V curves of un-annealed source-drain contacts are shown in Figure 8.5. Curve (a) shows a (relatively rare) Schottky-like behavior, while curve (b) (more representative of typical behavior) appears more ohmic. Under stress of a few volts, most contacts became ohmic with resistances of 50-500 Ω, compared to annealed resistances of roughly 5 Ω. This information suggests that either the AlSb layer was etched somewhat during the removal of the SiO_2 buffer layer, or that the AlSb layer is not a particularly good electrical barrier.

In order to determine whether the source-drain resistance was limited by contact resistance or by the resistance of the channels between the contacts, the resistances between sets of source-drain contacts separated by varying distances across the chip were measured. If the source-drain resistance were limited by the resistance of the

quantum well channels, the resistance would increase with increasing pad separation as a longer distance along the channel would have to be traversed. If the contact resistance were the limiting factor, however, the measured resistances would be found to be nearly constant with increasing pad separation.

Measurements of both annealed and un-annealed source-drain resistances at room and cryogenic temperatures between pads of varying separations across the chip resulted in roughly constant values. This was true both before and after the chips had been mounted on the headers. Therefore, this suggests that the resistance was limited primarily by the alloyed contacts. Given the Hall data reported above for sample III-691 in Section 8.2 and the device geometry, the resistance between neighboring pads should be roughly 0.5 Ω at room temperature and 6 Ω at 77 K, while the resistance between pads 1 mm apart should be roughly ten times larger. This apparent contradiction suggests that some additional conduction channel was present. One possibility is that the AlSb layers contain mobile charge. Another is that the metal underneath the chip (the probe station chuck in the case of the un-mounted chips, or the header and conducting paint in the case of the mounted chips) shorts out the channel and substrate.

Even at 4.2 K, the resistance between the source/drain contacts and the header body was found to be very low (10-30 Ω). In order for the GaSb layer to provide this conductive path, a carrier concentration of roughly $10^8/\text{cm}^3$ would be needed, assuming a (generous) low temperature mobility of 50,000 $\text{cm}^2/\text{V-s}$ – many orders of magnitude more than would be allowed at 4.2 K, given that the carrier concentration drops off as $\exp(-Ea/k_bT)$ for low temperatures, where Ea is the acceptor energy of roughly 33 meV [4]. This suggests that some other conduction path must be present. One likely possibility is that the conducting paste used to bond the chip to the header made contact to the wells along the side of the chip.

The gate-source/drain capacitance was measured (see Chapter 3 for details) to be roughly 60 pF at room temperature. This is somewhat higher than the expected value of 23 pF, given the gate area of 10^5 μm^2, an oxide thickness of 150 nm, and an AlSb barrier thickness of 50 nm. One explanation for the discrepancy could be that the

majority of the initially deposited SiO_2 was etched via undercutting of the photoresist during the HF etch in the source/drain windows, leaving only the 50 nm deposited immediately before gate deposition. Note that with this capacitance, each volt applied to the gate should induce a change of 4×10^{11} charges/cm^2 in the channel, according to a simple parallel plate capacitor model. If the channel charge density is truly given by the result of the Hall effect calculations in Section 8.2 (3×10^{13}/cm^2), then 10-100 V would need to be applied to the gate to make a 10-100% change in charge density. A modulation of this magnitude is likely to be necessary to bring about a noticeable change in conductivity, either due to velocity modulation effects, or due to traditional FET channel charge modulation.

If all of these results are considered, the following list of possible reasons for failure can be compiled:

1. The AlSb layers contain mobile charge. These layers can then short out the channel and possibly screen the gate field.

2. The source-drain contact resistance is much larger than the channel resistance. Therefore, any change in the channel conductivity with gate bias is unobservable.

3. The channel conduction path was shorted out by the conducting paste and header body, as the conducting paste accidentally made contact to the sides of the chip.

4. The coupled quantum well channel and the gate are too weakly capacitatively coupled, and the channel charge density too high for the gate to cause enough shift in the Fermi level and/or bands in the channel to bring about the velocity modulation.

Each of these possibilities deserves consideration in future studies.

8.4.2 Discussion of Related Work at NRL

In the past few years, results have been reported by a research group from the Naval Research Laboratory (NRL) on a device essentially identical to the one we have proposed in this section [6, 7]. They have observed effects similar to those expected from

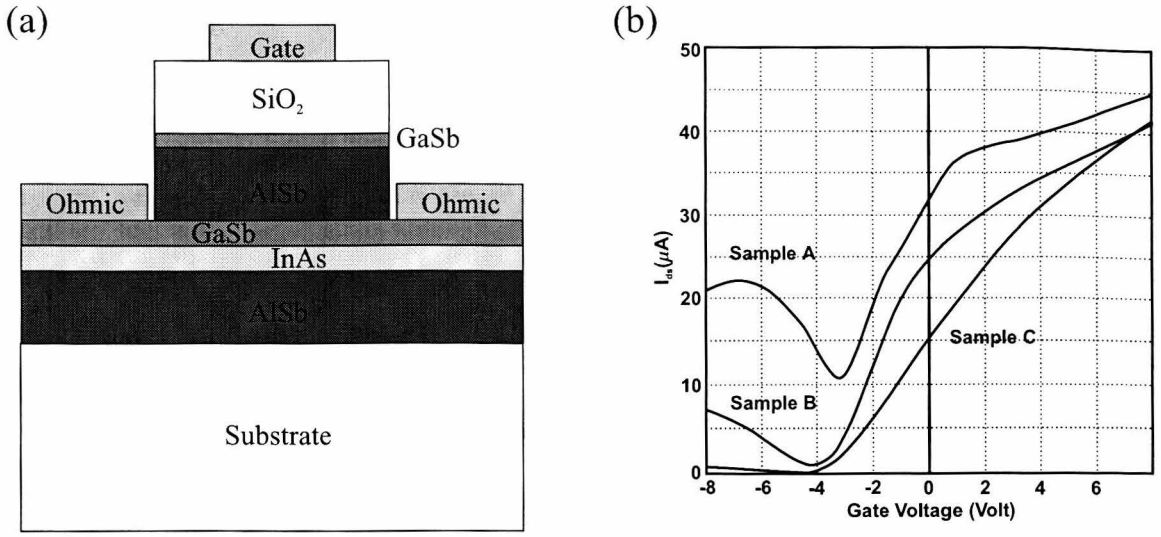

Figure 8.6: (a) Device structure and (b) source-drain current versus gate voltage plots at 4.2 K for samples described in [6]. Samples A, B, and C possessed InAs well widths of 14 nm, 8 nm, and 6 nm, and GaSb well widths of 4 nm, 5 nm, and 6 nm respectively. As the source-drain voltage was held fixed for these measurements, the source-drain current plotted here will be proportional to the channel conductance.

velocity modulation, but explained them simply in terms of modulation of charge density in the coupled well system. Here we shall examine their results and analysis.

A schematic of the devices studied in the NRL experiments is shown in Figure 8.6(a). It clearly is very similar to our proposed device structure, although selective etches were used instead of an alloyed metal to make contact to the quantum well layers, and a semi-insulating GaAs substrate with strained buffer layers was used instead of the GaSb substrates used in our studies. Devices with different well widths were fabricated, and the source-drain conductance was measured at different gate voltages for each device, as shown in Figure 8.6(b). They explained the increase in channel conductance as the gate voltage was moved away from the minimum at roughly -4 V as resulting from an increase in carriers in the coupled well system: negative carriers in the InAs well for a more positive gate voltage, and positive carriers in the GaSb well for a more negative gate voltage. This analysis was supported by Hall measurements of Sample A, illustrating a roughly linear modulation of channel sheet charge density from $+1.2\times10^{12}$ cm^{-2} at a gate voltage of -8 V to -2.2×10^{12} cm^{-2} at +8 V.

In the absence of any coupling between the quantum wells, we believe that this explanation would be accurate. Because the conductivity is proportional to the mobility and the carrier concentration, the increase in charge concentration would cause the conductivity to increase with a slope proportional to the carrier mobility. Given that the two carrier types have very different mobilities (10^4-10^5 for InAs, and 1000-5000 for GaSb as 4.2 K [3, 4, 8, 9]), the rise of the conductivity curve on either side of the minimum should be highly asymmetric.

This behavior is exactly what is observed for Sample C in Figure 8.6. As this sample consists of the thinnest InAs well, the sub-bands in this well are pushed far away from the bulk band edge energy, so no GaSb bands overlap with InAs bands in energy at flat band conditions. Given enough gate bias, the bands could be brought into resonance, but this would require considerable accumulation of charge in the channel. Thus, the Fermi level would likely need to be much higher than the position of the resonance, so the coupled states would likely not contribute to conduction.

Samples A and B, however, display conductance curves that are nearly symmetric about the minimum. In addition, the curve corresponding to Sample A appears to turn downward again at the far left of the plot. As these samples contained thicker InAs layers, and hence lower sub-band energies, it is possible that this symmetric feature is brought about by coupled InAs and GaSb sub-bands near the Fermi level, resulting in a velocity modulation effect.

Although we believe these arguments strongly suggest that the NRL samples A and B exhibit velocity modulation, a more convincing case could be made if the velocity modulation effect could be completely separated from the increase of carriers in the channel. One way to achieve this would be to use a dual-gated structure, as shown in Figure 8.7(a), as this would allow an electric field to be supplied across the channel without introduction of any additional net charge. To see this, consider the simple circuit model of the dual gate structure presented in Figure 8.7(b). If gate voltages are chosen such that $Q = C_{top}V_{top} = C_{bottom}V_{bottom}$, then the net charge induced in the channel will be zero, but a charge of $+Q$ will exist on one gate, and $-Q$ on the other gate, giving rise to an electric field across the channel.

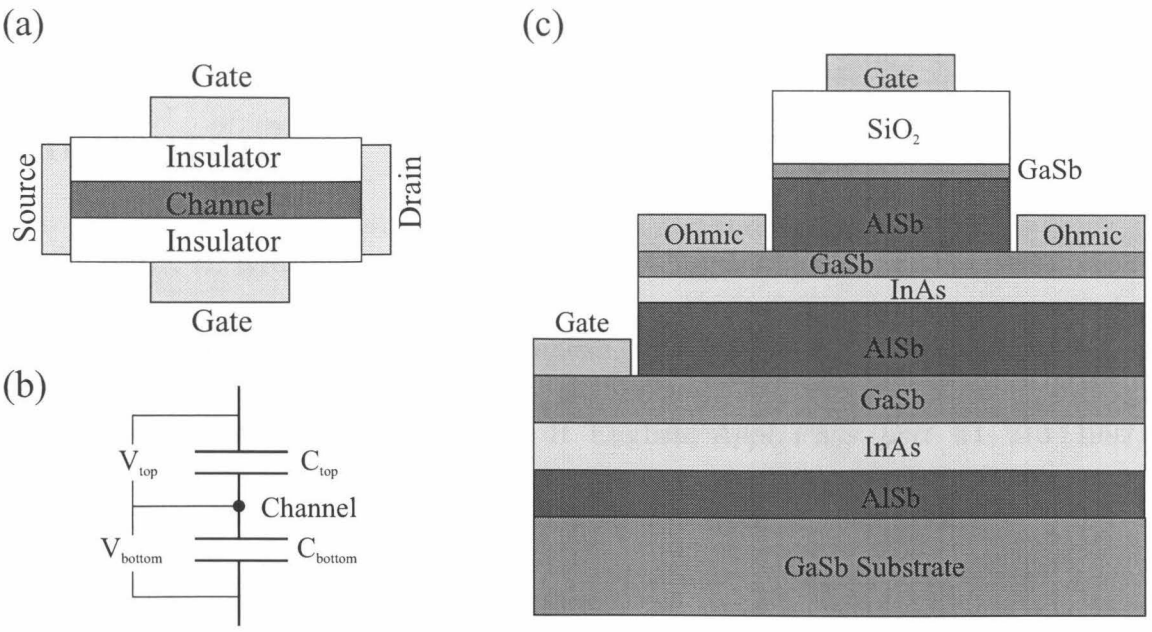

Figure 8.7: (a) Idealized dual gate FET. (b) Simple circuit describing the dual gate structure. (c) Proposed dual gate InAs/GaSb/AlSb VMT structure.

The back gate could be provided in the InAs/GaSb/AlSb material system by thick adjacent InAs and GaSb layers, acting as a pseudo-metallic layer (see Figure 8.7(c)). A sample with this structure was grown (III-727 in Table 8.1), but fabrication techniques and mask designs which could be used to produce the back-gated structure have not yet been developed.

Notice that the structure of NRL Sample A is very similar to the target device structure produced in our labs. The well thicknesses used (4 nm GaSb, 14 nm InAs) are of the same magnitude as ours (12 nm GaSb, 18 nm InAs), and even the oxide thickness (100 nm) and the AlSb barrier thicknesses (50 nm) are very close to those used in our experiments. It is therefore interesting that our measured channel carrier concentration ($-3.1\times10^{13}/cm^2$) is more than an order of magnitude larger than the NRL reported values ($-5\text{-}10\times10^{11}/cm^2$). This suggests that failure mechanisms 1 and 3 described above may be the most likely.

Bibliography

[1] G. Tuttle, H. Kroemer, and J. H. English, *J. Appl. Phys.* **65**, 5239 (1989).

[2] G. Tuttle, H. Kroemer, and J. H. English, *J. Appl. Phys.* **67**, 3032 (1990).

[3] C. R. Bolognesi, H. Kroemer, and J. H. English, *Appl. Phys. Lett.* **61**, 213 (1992).

[4] A. Milnes and A. Y. Polyakov, *Solid–State Electron.* **36**, 803 (1993).

[5] J. O. McCaldin, private communication.

[6] M. J. Yang, F.-C. Wang, C. H. Yang, B. R. Bennett, and T. Q. Do, *IEDM* (1995).

[7] M. J. Yang, F.-C. Wang, C. H. Yang, B. R. Bennett, and T. Q. Do, *Appl. Phys. Lett.* **69**, 85 (1996).

[8] C. R. Bolognesi, J. E. Bryce, and D. H. Chow, *Appl. Phys. Lett.* **69**, 3531 (1996).

[9] K. Yoh, K. Kiyomi, M. Yano, and M. Inoue, *J. Cryst. Growth* **127**, 29 (1993).

Appendices

Appendix A Resistive Network Analysis

Before I came across these concepts elsewhere [1], I developed the following set of equations to describe a macroscopic device with a resistive contact as a mesh of identical microscopic devices with a resistive network as the contact. Undoubtedly, the same concepts and (hopefully) the same equations have been derived before, but as I did not adopt the equations used in Chapter 3 from another source, I decided to include my analysis here.

Consider the effective circuit in Figure A.1. Applying Kirchoff's laws, the following equations result:

$$V_s = V_{j,k} = f(I_{zj,k}) \tag{A.1}$$

$$I_{zj,k} + I_{yj+1,k} + I_{xj,k+1} + I_{yj,k} + I_{xj,k} = 0 \tag{A.2}$$

$$V_{j,k+1} - V_{j,k} = RI_{xj,k} \tag{A.3}$$

$$V_{j+1,k} - V_{j,k} = RI_{yj,k} \tag{A.4}$$

where R is the resistance of each element of the network, and $\Delta V = f(I)$ is the voltage drop across a given microscopic device, drawing a current, I. These discrete equations can be transformed to continuous (differential) equations with the following

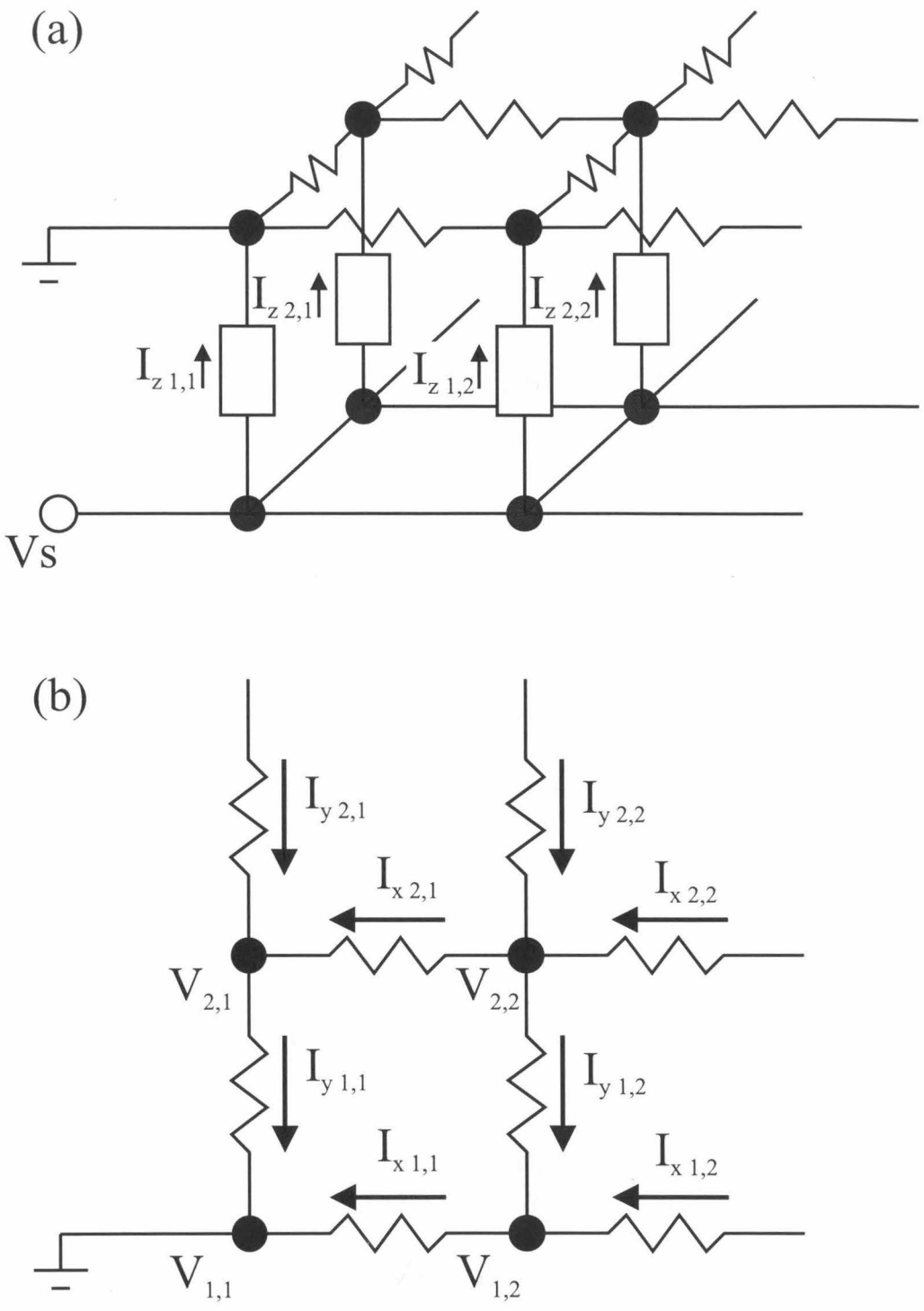

Figure A.1: (a) Three-dimensional view, and (b) top view of the resistive network effective circuit model for a macroscopic device with a resistive gate contact. Current and voltage variables used in the derivation of the equations are defined in the figures.

transformations:

$$V_{j,k} \rightarrow V(x,y) \tag{A.5}$$

$$I_{zj,k} \rightarrow J_z(x,y)\delta x \delta y \tag{A.6}$$

$$I_{yj,k} \rightarrow J_y(x,y)\delta x t_G \tag{A.7}$$

$$I_{xj,k} \rightarrow J_x(x,y)\delta y t_G \tag{A.8}$$

$$f(I_{zj,k}) \rightarrow F(J_z(x,y)) \tag{A.9}$$

$$R \rightarrow \rho L/A = \begin{cases} \rho_g \frac{\delta x}{\delta y t_g} & \text{for} J_x \\ \rho_g \frac{\delta y}{\delta x t_g} & \text{for} J_y \end{cases} \tag{A.10}$$

where t_g, ρ_g are the thickness and resistivity of the gate. Under these transformations, and taking the limits that all differential variables approach zero, Equations A.1-A.4 become

$$V_s - V(x,y) = F(J_z(x,y)) \tag{A.11}$$

$$-J_z(x,y) = t_g\left[\frac{\partial J_y(x,y)}{\partial y} + \frac{\partial J_x(x,y)}{\partial x}\right] \tag{A.12}$$

$$\frac{\partial V(x,y)}{\partial x} = \rho_g J_x(x,y) \tag{A.13}$$

$$\frac{\partial V(x,y)}{\partial y} = \rho_g J_y(x,y) \tag{A.14}$$

These equations can be condensed by substituting Equations A.13 and A.14 into Equation A.12:

$$-J_z(x,y) = \frac{t_g}{\rho_g}\left[\frac{\partial^2 V(x,y)}{\partial x^2} + \frac{\partial^2 V(x,y)}{\partial y^2}\right] = \frac{t_g}{\rho_g}\nabla^2 V(x,y) \tag{A.15}$$

and further condensed by eliminating $V(x,y)$, using Equation A.11:

$$J_z(x,y) = \frac{t_g}{\rho_g}\nabla^2 F(J_z(x,y)) \tag{A.16}$$

In the case in which the device elements are simple resistors, we can use a simple

expression for $F(J_z)$:

$$F(J_z) = J_z(x,y)\rho_{dev}L_{dev} \tag{A.17}$$

where ρ_{dev}, L_{dev} are the resistivity and effective "length" of each device. Equation A.16 then becomes

$$J_z(x,y) = r^2\nabla^2 J_z(x,y) \tag{A.18}$$

where the parameter

$$r = \sqrt{\frac{\rho_{dev}L_{dev}}{\rho_g/t_g}} \tag{A.19}$$

is an effective decay length.

Bibliography

[1] C. A. Mead, *Analog VLSI and Neural Systems*, Addison-Wesley, Reading, MA, 1989.

Appendix B Matrix Norms and Condition Numbers

While encountering problems with the drift-diffusion simulations described in Section 5.2, I conducted a brief search for a simple discussion of matrix norms, their properties, and condition numbers, as applicable to the numerical solution of an $\mathbf{A} \cdot \vec{x} = \vec{b}$ matrix problem. This search turned up very little useful information, however, so the following simple relations were developed in an attempt to better understand these concepts. The fact that the concepts presented here were scrawled on a paper napkin during a flight from L.A. to Portland hopefully does not compromise their validity.

All studies presented here assume the following definition of the matrix norm [1]:

$$\|\mathbf{A}\| \equiv \max_j \left(\sum_i \mathbf{A}_{ij} \right) \tag{B.1}$$

Apparently, other norms can also be defined, having similar properties [1]. Consider vectors which satisfy the following equations:

$$\mathbf{A} \cdot \vec{x} = \vec{b} \tag{B.2}$$

$$\mathbf{A} \cdot \vec{\delta x} = \vec{\delta b} \tag{B.3}$$

Expanding Equation B.2 into a sum yields

$$b_i = \sum_j \mathbf{A}_{ij} x_j \tag{B.4}$$

If we now sum this equation over i, we obtain

$$\sum_i b_i = \sum_i \sum_j \mathbf{A}_{ij} x_j \tag{B.5}$$
$$= \sum_j x_j \sum_i \mathbf{A}_{ij}$$

We now can introduce the matrix norm into the equation by using the following relation:

$$\sum_i \mathbf{A}_{ij} \leq \max_j \left(\sum_i \mathbf{A}_{ij} \right) \equiv \|\mathbf{A}\| \tag{B.6}$$

which can be substituted into Equation B.5, yielding

$$\sum_i b_i \leq \|\mathbf{A}\| \sum_i x_i \tag{B.7}$$

Applying the same set of manipulations to the inverse of Equation B.2 yields

$$\sum_i x_i \leq \|\mathbf{A}^{-1}\| \sum_i b_i \tag{B.8}$$

Similar relations can be developed for $\vec{\delta b}, \vec{\delta x}$.

Applying Equation B.7 to $\vec{b}, \vec{x}$, and Equation B.8 to $\vec{\delta b}, \vec{\delta x}$, and taking the ratio, we obtain

$$\frac{\sum_i \delta x_i}{\sum_i x_i} \leq \|\mathbf{A}\| \|\mathbf{A}^{-1}\| \frac{\sum_i \delta b_i}{\sum_i b_i} \tag{B.9}$$

where $\|\mathbf{A}\| \|\mathbf{A}^{-1}\|$ is defined to be the condition number, R, and the ratios of the vector sums are simple representations of fractional error, which we shall label $\vec{x}_{err}$ and $\vec{b}_{err}$. Note that if the role of x and b are switched, we obtain

$$\frac{\sum_i \delta x_i}{\sum_i x_i} \geq \frac{1}{\|\mathbf{A}\| \|\mathbf{A}^{-1}\|} \frac{\sum_i \delta b_i}{\sum_i b_i} \tag{B.10}$$

Consider what these relations imply about errors in the numerical solution of the matrix equation. The fraction error in the solution, $\vec{x}_{err}$, is bounded below by $\frac{1}{R}\vec{b}_{err}$, and above by $R\vec{b}_{err}$. When the condition number is large, therefore, the error in $\vec{x}$ is unpredictable, and may be very large, whereas if the condition number is reduced, the error in the solution can be kept within reasonable bounds.

Given this interpretation, it is useful to be able to estimate the condition number for a given matrix in order to evaluate whether solution of linear equations with that matrix is likely to result in useful results. While the matrix norm of a matrix, $\mathbf{A}$, can be readily computed from the definition (Equation B.1), it would be convenient to be able to estimate the norm of the inverse matrix without explicitly computing the inverse. One simple method is given here.

Using the definition of the inverse of a matrix, we can write

$$\sum_j \mathbf{A}^{-1}_{ij}\mathbf{A}_{jk} = \delta_{ik} \tag{B.11}$$

where δ is the Kronecker-δ symbol. Summing over i, we obtain

$$\sum_i \sum_j \mathbf{A}^{-1}_{ij}\mathbf{A}_{jk} = \sum_i \delta_{ik} = 1 \qquad \forall k \tag{B.12}$$

Rearranging, and using Equation B.6 for $\mathbf{A}^{-1}$, we can write

$$\left(\sum_j \mathbf{A}_{jk}\right)\left(\max_j \sum_i \mathbf{A}^{-1}_{ij}\right) \geq 1 \qquad \forall k \tag{B.13}$$

$$\Rightarrow \max_j \sum_i \mathbf{A}^{-1}_{ij} \equiv \|\mathbf{A}^{-1}\| \geq \frac{1}{\sum_j \mathbf{A}_{jk}} \qquad \forall k \tag{B.14}$$

As this equation holds for all k, the largest lower bound for the left-hand side can be determined by choosing the k corresponding to the smallest sum on the right-hand side:

$$\|\mathbf{A}^{-1}\| \geq \frac{1}{\min_k\left(\sum_j \mathbf{A}_{jk}\right)} \tag{B.15}$$

This allows a lower bound estimation of the condition number as follows:

$$R \equiv \|\mathbf{A}\|\|\mathbf{A}^{-1}\| \geq \frac{\max_j\left(\sum_i \mathbf{A}_{ij}\right)}{\min_j\left(\sum_i \mathbf{A}_{ij}\right)} \tag{B.16}$$

I really do not know how good an estimate this formula gives for the condition number – Lapack routines [2] and Mathematica were used to estimate the condition number in practice, rather than this formula. However, this relation does serve to illustrate that the condition number is related to the range of values within the matrix. As the unscaled semiconductor equations can give rise to a matrix with values varying over many orders of magnitude, particularly for wide bandgap materials in which the ratio of majority to minority carrier concentrations is enormous, it is apparent that the resulting condition number will be quite large.

Bibliography

[1] E. Anderson et al., *Lapack Users Guide*, SIAM, University Science Center, Philadelphia, PA, 2nd edition, 1994.

[2] Clapack linear algebra routines version 2.0, Obtained from Netlib Database: http://netlib.org, 1994.