CMOS Inverse Doping Profile Extraction and Substrate Current Modeling

by

Eric Pop

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

Master of Engineering in Electrical Engineering and Computer Science

and

Bachelor of Science in Electrical Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 1999

Author .....
Department of Electrical Engineering and Computer Science
May 14, 1999

Certified by .....
Dimitri A. Antoniadis
Professor
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

CMOS Inverse Doping Profile Extraction and Substrate Current Modeling

by

Eric Pop

Submitted to the Department of Electrical Engineering and Computer Science
on May 14, 1999, in Partial Fulfillment of the Requirements for the Degrees of
Master of Engineering in Electrical Engineering and Computer Science
and
Bachelor of Science in Electrical Science and Engineering

Abstract

CMOS substrate current considerations play an important role in modern device design. Powerful, reliable and predictive simulation capabilities are essential to this effort. Such accurate substrate current simulations demand two requirements: knowledge of the E-field distribution, hence of the 2-D device doping profiles, and knowledge of the hot-carrier distribution both in momentum and position space. This thesis investigates the use of inverse doping profile extraction from device capacitance measurements with the help of a non-linear optimization program based on the Levenberg-Marquardt algorithm. It is shown that such a method leads to 2-D doping profiles that can be used for good device capacitance and current simulations. This thesis also implements a simple new impact ionization model based on a parameterized carrier distribution function with a high-energy tail. The new model is implemented in the device simulator FIELDAY and it is calibrated by comparisons of substrate current simulations and data. It is shown that the optimized doping profiles are essential for accurate simulations of the substrate current in MOSFETs.

Thesis Supervisor: Dimitri A. Antoniadis
Title: Professor

Acknowledgments

The work presented in this thesis was conducted between June 1998 and January 1999 while I was a co-op within the VI-A program at IBM Microelectronics in Essex Junction, VT. The thesis document itself was written upon my return to MIT in the spring of 1999. As such, there are many people, at both institutions, that have contributed in one way or another and have helped me along the way.

First and foremost I would like to thank Jim Slinkman, my main IBM mentor. A lot of this thesis would not have been possible without his help, patience and guidance. Among other things, he has taught me the intricacies of TCAD device simulations and has provided me with inspiration when I was stuck. I must also thank Prof. Dimitri Antoniadis, my MIT thesis advisor, for agreeing to supervise this project. He has provided me with precise help and guidance while I was away at IBM, and has continued to do so when my thesis work began to crystallize into the final document after my return to MIT.

I must thank Steve Furkay and Jeff Johnson of IBM for many helpful discussions related to device or process simulation, and Bill Clark of IBM for discussions related to device physics. I have learned a lot from them. My IBM manager Peter Cottrell has been a model of efficiency and resourcefulness. Also from IBM, I must thank Robert Gauthier and Don Cook who have helped me obtain all the hardware and the data I needed for my work.

Back at MIT I would like to thank my VI-A advisor, Prof. Steve Senturia who has also served as my academic mentor for the past three years. I must also thank my friend Ken Esler for providing me with the extra storage space on his own Linux workstation when the amount of data I brought back from IBM exceeded my Athena quota.

Finally, I would like to thank IBM for funding my research assistantship during the fall term, and Prof. Jesus del Alamo for taking me on-board as a teaching assistant for 6.012 during the spring semester.

Contents

1 Introduction **15**

1.1 Evolution 15

1.2 Today's Problems 15

1.2.1 Uneven Scaling 16

1.2.2 Hot Carrier Effects 16

1.3 TCAD Simulation 18

1.4 The Scope of This Work 19

1.5 Organization 20

2 Some Existing Doping Profiling Methods **21**

2.1 Destructive Methods 21

2.2 Non-Destructive Methods 23

3 Inverse Doping Profiling from C-V Measurements **27**

3.1 Junction Capacitance 27

3.1.1 Relationship to Substrate Current 28

3.1.2 The Inverse Problem 29

3.1.3 Solving for the Junction Capacitance 32

3.1.4 Experimental Measurements 34

3.1.5 Optimization Results 35

3.2 Gate Capacitance 37

3.3 Gate-to-Source Capacitance 39

3.3.1 Capacitance Components 39

3.3.2 Experimental Measurements 42
3.3.3 The 2-D Problem 43
3.3.4 The FITDRF Optimizer 43
3.3.5 Gate Voltage Dependence 44
3.3.6 Source Voltage Dependence 49
3.4 Drain Current Simulations 50
3.5 Summary 51

4 Substrate Current Modeling **53**
4.1 Motivation 53
4.2 Impact Ionization 54
4.3 Historical Background 56
4.4 Device Simulation 57
4.4.1 The Post-Processed Approach 58
4.4.2 The Self-Consistent Approach 60
4.5 Temperature-Dependent Impact Ionization Modeling 60
4.5.1 The Schöll-Quade Model 62
4.5.2 The Modified Distribution Function 64
4.5.3 The Modified Impact Ionization Rate 65
4.5.4 FIELDAY Implementation 66
4.6 Substrate Current Simulations 67
4.7 Discussion 69
4.8 Summary 72

5 Conclusions **75**
5.1 Summary 75
5.2 Discussion and Suggestions for Future Work 77

A The Levenberg-Marquardt Algorithm **81**

B The FITDRF Optimizer **85**
B.1 Purpose 85

B.2 Usage 85
B.3 The Input File 86
B.4 Other Input Files 87
B.5 Program Output 88
B.6 Timing and Speed Issues 88
B.7 Other Technical Issues 89

C Sample Input Files **91**
C.1 DOPING Input File 91
C.2 REGRID Input File 92
C.3 FIELDAY Input File 93
C.4 FITDRF Input File 93

Bibliography **95**

List of Figures

1-1 Schematic of impact ionization processes in n-MOSFETs. The circle represents the place where the impact ionization event took place and the new electron-hole pair was created. 17

3-1 The junction capacitance of a MOSFET 28

3-2 General schematic of the inverse method used to extract a structure's doping profile when its C-V characteristics are known. 31

3-3 Comparison of junction capacitance per unit area measured across 3 different chips (symbols) and simulation (lines) before and after the doping profile optimization. 36

3-4 Vertical junction net active doping $|N_d - N_a|$: initial and extracted profiles. The left side of the junction is the n+ source and the right side is the p-type substrate. 36

3-5 High frequency (HF) and quasi-static (QS) C-V data averaged over seven consecutive measurements in order to reduce experimental noise. The extracted interface trap distribution as a function of gate voltage is shown in the insert. 38

3-6 The various components that make up the gate-to-source capacitance, C_{gs}. Also marked on the figure are the gate-source overlap L_{ov}, the oxide thickness t_{ox}, and the gate thickness t_g. 40

3-7 Mesh used in FIELDAY for C_{gs} simulation. Compare with Figure 3-6 and note the presence of the gate side-wall spacer and the top passivation oxide. The mesh has been optimized for device simulation using the program REGRID. 44

3-8 Schematic of the inverse modeling method used to extract the 2-dimensional doping profiles based on C_{gs} measurements. The dotted line surrounds the three programs that make up the "forward" solver. 45

3-9 Plot of gate-to-source capacitance data across 3 chips (symbols) and several simulations (lines) as a function of gate voltage. The solid line represents the simulation with the optimized doping profile and the dotted lines represent simulations done using a lower channel doping and a lower gate length, respectively. 47

3-10 Plot of gate-to-source capacitance data across 3 chips (symbols) and several simulations (lines), as a function of gate voltage. The simulations were run with and without the top oxide (passivation) and with and without the gate side-wall spacer. 47

3-11 Comparison of gate-to-source capacitance per unit gate width measured across 3 different chips (symbols) and simulation (lines) before and after the doping profile optimization — as a function of source voltage. 48

3-12 Lateral junction net active doping: initial guess and extracted profile at 0.1 microns below the oxide/silicon interface. The left side of the junction is the n-type source and the right side is the p-doped substrate. 48

3-13 Measured and simulated drain currents per unit width for devices with 0.5, 0.6, 1.0 and 5.0 microns gate length. The width of the measured devices was 20 microns. The drain bias was set at 5 V whereas the source and the substrate were grounded. 51

3-14 Log scale comparison between simulated drain currents and data for the same devices as in Figure 3-13. 52

4-1 Schematic representation of the screened electron-electron interaction corresponding to impact ionization in an indirect band gap semiconductor (such as silicon). The top parabola represents the conduction band, while the bottom one is the valence band. 55

4-2 Block diagram of a device modeling scheme, such as the one in FIELDAY II. 61

4-3 Log scale comparison between the simple Maxwellian $f_{sq}(k)$ (solid line) used in Schöll-Quade's model and the new high energy tail distribution function $f_{het}(k)$ (dotted line). The comparison is done for $T = 300$ K and $r = 1.8$. . 66

4-4 Log scale comparison between the original Schöll-Quade impact ionization rate $G^{ii}_{sq}(u)$ (solid line) and the modified high energy tail $G^{ii}_{het}(u)$ model (dotted line). The comparison is done for $T = 300$ K, $r = 1.8$ and $E_{th} = 1.12$ eV (the silicon band gap). 67

4-5 Measured (symbols) and simulated (lines) substrate currents for the four devices under investigation — with gate lengths of 0.5, 0.6, 1.0 and 5.0 microns (from top to bottom). The impact ionization parameters used were ETHN=1.12, TAUN0=1.26E-14 and RHETN=1.32. The drain bias was $V_{ds} = 4$ V. 69

4-6 Comparison of measured (symbols) and simulated (lines) substrate currents for the same devices as in Figure 4-5. Only the high-energy tail parameter was changed to RHETN=1.0, forcing the impact ionization rates to be computed with the old (simple Maxwellian) energy distribution function. 70

4-7 Another comparison of measured (symbols) and simulated (lines) substrate currents. The simulations above were obtained from devices with the non-optimized doping profiles used as initial "guesses" in the inverse modeling procedure described in chapter 3. RHETN=1.32 was used. 71

Chapter 1

Introduction

1.1 Evolution

As the semiconductor industry has progressed over the past thirty years, integrated circuit densities have increased tremendously. In fact, the number of transistors on a chip has been doubling every 18 to 24 months, an observation that has come to be known as "Moore's Law" after Gordon Moore, the man who first noted it. Intel's first processor, the 4004, contained 2,300 transistors whereas today's complex microprocessors incorporate close to 10 million transistors and are up to a quarter million times faster. Unfortunately, such aggressive decrease in device size and increase in circuit density have not been possible without bringing along a variety of limitations.

1.2 Today's Problems

Despite the nearly exponential decrease in integrated circuit feature size over the years, it is apparent that this trend cannot continue going on forever. Both business and real physical limitations are sooner or later likely to slow it down. As chip densities rise, the cost of production goes up almost exponentially. As circuit complexity increases it has become virtually impossible to exhaustively test a computer chip. And as the minimum feature size drops below 0.1 microns — or a couple of hundred atoms across — the atomic and quantum mechanical nature of materials start creeping up and introducing new problems. It is now

generally believed that due to a combination of the limitations described above, "Moore's Law" will significantly slow down in the next 20 years.

1.2.1 Uneven Scaling

Although the size of the transistor has been aggressively scaled down in search for ever higher processor speeds and corporate profit margins, the power supply voltage has often escaped scaling for the sake of compatibility with existing systems and maintaining circuit speed margins. For example, the power supply voltage was kept at 5 V from the mid seventies, when transistors had typical channel lengths of about 5 microns and gate oxides around 1000 A, until the late eighties when the average transistor dimensions had shrunk by about a factor of 5 (see Table 1.1). Some relief came with the introduction of the 3.3 V, and more recently the 2.5 V supplies, but today's sub-micron transistors are still experiencing electric fields that are higher than ever, leading to numerous concerns regarding their reliability and further scaling.

	Year					
Parameters	1976	1980	1984	1988	1992	1996
Gate Length (μm)	5	3	2	1	0.5	0.35
Gate Oxide (nm)	100	60	40	20	12	8
Supply Voltage (V)	5	5	5	5	3.3	2.5

Table 1.1: Average industry-wide device scaling trends over the last quarter century.

1.2.2 Hot Carrier Effects

It is very common to find transistors with channel lengths under 0.5 microns and gate oxides below 100 A operated under 3.3 or even 5 V power supplies in modern integrated circuits. Moreover, some technologies that are designed to operate at 3.3 V need to be modified to accommodate 5 V devices on their chips (e.g. for I/O purposes). This combination of high voltage and small dimensions leads to very high electric fields that can reach more than 100 kV/cm during the normal operation of a transistor. The high electric

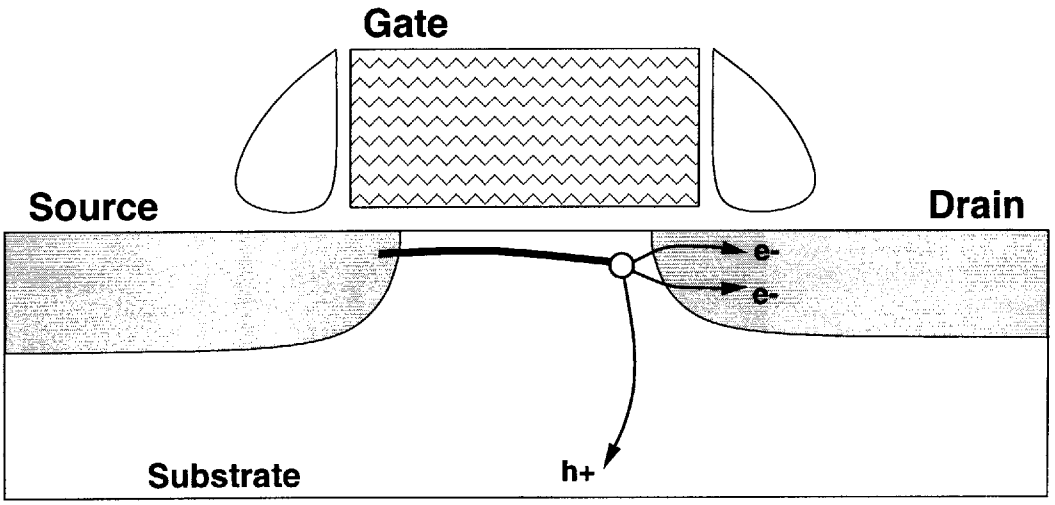

Figure 1-1: Schematic of impact ionization processes in n-MOSFETs. The circle represents the place where the impact ionization event took place and the new electron-hole pair was created.

fields in turn accelerate the mobile charge carriers to very high velocities, leading to what are known as "hot-carrier" effects. When these highly energetic carriers travel through a semiconductor there are two main phenomena that can occur. First, a carrier may acquire enough energy to break a lattice bond in the semiconductor. This phenomenon, also known as impact ionization, has been recognized and studied from the earliest days of the semiconductor industry [1, 2]. In the case of an n-MOSFET, the hole of the generated electron-hole pair travels towards the substrate contact, where it is collected in the form of the substrate current [3]. The impact ionizing electron and the generated electron are both usually collected by the device drain, as illustrated in Figure 1-1. Hot carrier phenomena are less of a concern in p-MOSFETs because the channel carriers' mobility and impact ionization rates are typically several times lower than in similar n-MOSFETs. The typical p-channel device substrate current is about three orders of magnitude lower than that of an n-channel device [4], and thus substrate current studies (including this one) generally focus on n-MOSFET issues.

Secondly, the channel carriers (or even the secondary generated carriers) may be scattered towards the silicon/insulator interface after a collision in which their momentum

changes direction by just the right amount. If their energy is large enough, these "lucky" carriers [5] can create interface states, or fill interface or bulk traps — either of which can lead to the accumulation of fixed charge at the silicon/insulator interface and ultimately to the degradation of the device. Moreover, some of these carriers may have enough energy to get injected directly into the conduction band of the insulator and then drift to the gate where they are collected as gate current. Therefore, due to their origin, both the substrate and the gate current have been used to monitor the hot-carrier, device-degrading effects taking place in modern MOSFETs.

High substrate currents by themselves can be damaging as well, as they can lead to overload of the circuit substrate-bias generator and can induce snap-back breakdown and CMOS latch-up [6, 7]. Hot carriers are also responsible for the photo-current that can degrade DRAM refresh times, whose origin is found in the bremsstrahlung radiation [8] emitted when an energetic carrier is decelerated by an impurity ion.

1.3 TCAD Simulation

Given the impact that hot carrier effects have on modern device reliability, it has become very important that they be modeled accurately and consistently. In fact, with the increase in our available computational power, Technology Computer-Aided Design (TCAD) simulations have become an essential part of the design of new generations of semiconductor devices [9]. Accurate, predictive simulations can save millions of dollars (and months of production time) that would have otherwise been spent testing wafers with countless process variations. New technologies and devices are much more easily tested in a "virtual fab", and essential design considerations can thus be made long before wafers need to be sent for testing in the real fab.

As devices are shrunk to sub-micrometer sizes, subtle details of the 2-dimensional (2-D) and 3-dimensional (3-D) redistribution of dopants, due to thermal diffusion during the fabrication process, strongly determine the device short-channel effects. It is these effects which ultimately limit device operation and performance. For the sake of accurate simulations it has thus become very important to understand what the exact doping profiles of a mod-

ern device are. Such full-scale simulation is a two-part problem: process simulation [10], which leads to the formation of the device and its physical properties (gate length, oxide thickness, 2-D doping profiles) and device simulation [11], the actual I-V or C-V electrical calculations. Device simulators need (and rely on) accurate process models for their good operation. Therefore accurate device doping profile information, if possible as a function of more than one space coordinate, is an important prerequisite for accurate device simulation.

Unfortunately, much is still unknown about the diffusion of impurities in silicon. Although complex computer models such as the process simulator SUPREM [10] exist, there are no measurements that can be performed to directly and accurately determine the 3-dimensional spread of impurities across a device's volume. Such experimental doping profile extraction methods, if available, could be used to:

- provide a check on the fabrication process
- serve as input for device simulators
- increase our understanding of dopant atom redistribution in semiconductors, thus also enabling the verification of process simulator models
- help minimize the amount of expensive test hardware used in technology development.

1.4 The Scope of This Work

The first goal of this thesis is to demonstrate the use of inverse 1- and 2-dimensional doping profile extraction both as a process simulator check and as a reliable input for device simulator calibration. Simulation results that are particularly sensitive to the doping profile distribution especially benefit from such an approach. In this thesis, doping profile extraction is treated as an inverse problem, whose outputs (e.g. electrical capacitance measurements) are known, but whose inputs (the device-specific doping profiles) are to be found.

The second goal of this thesis is to introduce a new parameterized impact ionization model and to calibrate it through substrate current simulations. The impact ionization model is based on a simple, yet physically-based high-energy-tail correction to the carrier distribution function and is shown to be easily implemented in an existing device simulator.

Moreover, it is demonstrated that the substrate current simulations are particularly sensitive to the device doping distribution. Therefore the previously inverse modeled doping profiles are shown to be essential for the accurate calibration of any such new transport models.

1.5 Organization

Chapter 2 of this thesis reviews some of existing doping profile extraction methods and discusses their individual strengths and weaknesses.

Chapter 3 formulates 1- and 2-dimensional doping profiling as an inverse problem whose starting points are the device C-V characteristics. The chosen C-V measurements are discussed and the implementation of the extraction technique is explained. The results are analyzed, and their reliability is assessed. The validity of the extracted MOSFET doping profiles is further supported by good agreement between both C-V *and* I-V simulations and data.

Chapter 4 begins by describing several approaches that have been taken to model impact ionization in semiconductors. The theory and assumptions behind a particular temperature dependent model [12] are then explored in more depth. A parameterized high energy tail is introduced in the carrier distribution function and the impact ionization rate is re-derived and implemented in an existing device simulator, FIELDAY [11]. The new impact ionization model is calibrated within the context of the previously determined device doping profiles.

Chapter 5 provides an overall conclusion of this thesis. The results are analyzed and several pointers are offered for future work.

The three appendices contain, in order, a summary of the least-squares Levenberg-Marquardt algorithm, a thorough description of the inverse modeling and parameter extraction program written for the purposes of this thesis, and several examples of simulator input files. The work described in the following four chapters and three appendices should provide enough detail to enable anyone with similar resources to duplicate the results of this thesis.

Chapter 2

Some Existing Doping Profiling Methods

This chapter reviews some existing doping profile extraction methods and discusses their individual strengths and weaknesses. All existing doping profile extraction methods can be classified in two broad categories: destructive, such as SIMS, RBS, spreading resistance and AFM, or non-destructive, such as 1- or 2-dimensional capacitance-voltage methods and sub-threshold current-voltage methods.

2.1 Destructive Methods

The main characteristic (and disadvantage) of destructive doping profiling methods, as their name suggests, is that the semiconductor wafer is at least partly destroyed in the process. Application of destructive methods to process monitoring is therefore undesirable. In destructive methods, usually thin layers of semiconductor material are removed from the surface of the device (or special test structure). Next, either the contents of the removed layer is analyzed or the behavior of the remaining device is measured. Layers may be removed by sputter etching with an ion beam, by beveling or by anodic oxidation followed by a selective wet etch to remove the oxide layer.

Secondary Ion Mass Spectroscopy (SIMS) uses an ion beam (e.g. Cs^+ at 10 keV) to continuously remove layers from the top of the semiconductor surface [13]. The ionized

particle stream eroded from the sample is analyzed by a mass spectrometer. If the erosion speed is known, the evolution of each species as a function of time can be used to find its distribution as a function of depth into the sample. SIMS analysis usually requires large test structures and the obtained data is not very reliable near an interface. Also, this technique only measures the total chemical concentration of dopants, not the concentration of ionized dopants — although it is the latter that is mainly responsible for a device's electrical properties. SIMS is by nature a 1-dimensional doping extraction technique, and perhaps the most commonly used one, despite its requirement for expensive equipment.

Rutherford Backscattering Spectroscopy (RBS) uses a 1 - 3 MeV $^4He^+$ ion beam to penetrate the semiconductor surface [13]. The incident ions are detected after they are backscattered at various energies by elastic collisions with the different atomic species present in the semiconductor sample. A depth profile can be obtained by monitoring the number of backscattered ions as a function of their energy. Unlike for SIMS, no calibration with standards is required to obtain accurate quantitative results, but this technique isn't as sensitive at lower doping levels.

In spreading resistance profiling (SRP) the sheet resistance ρ_s of a sufficiently large area of a semiconductor layer is measured. To obtain a depth profile, a large number of thin, uniform layers are removed [14]. For silicon, this is usually done by anodic oxidation of the surface, followed by a wet etch of the oxide layer. Alternatively, a depth profile can also be obtained by beveling the semiconductor surface at a small angle and probing down the bevel. Spreading resistance techniques are mostly of historical importance, but they still offer some perspective in the profiling of highly doped layers since they are less expensive than SIMS.

Atomic Force Microscopy (or AFM) is probably the newest among all destructive doping profile measurement methods. It also requires relatively expensive equipment and extensive sample preparation, but it can be used to directly explore cross-sections of actual MOS devices. The technique, also known as Scanning Capacitance Microscopy (or SCM), requires the use of an AFM machine to position a tiny conducting tip over the semiconductor surface. As the tip is scanned across the surface, the change in capacitance measured by the tip is held constant by varying the amplitude of the bias applied to the sample with

a feedback control. This leads to large bias voltages in heavily doped regions and small biases in lightly doped regions. The amplitude of the bias voltage can then be related to the dopant density through a conversion algorithm based on a quasi-3-D model of the tip sample capacitor. Using this method relatively good resolutions of vertical dopant profiles have been recently reported, in good agreement with SIMS measurements [15]. Although some advances towards the achievement of quantitative 2-dimensional doping profiles have also been recently made [16], the results are highly dependent on the quality of the probe tip and of the surface preparation. The AFM/SCM technique is still in its infancy, but it may hold great potential for the future. Nevertheless, the color map profiles that can be obtained today can still be used to at least qualitatively gauge the relative distribution of dopants across the 2-dimensional cross-section of a semiconductor device.

2.2 Non-Destructive Methods

Non-destructive doping profile extraction methods generally use radiation or electrical data to obtain the necessary information. Radiative methods are not too accurate, and can only be used to obtain approximate doping profiles. Their doping sensitivity is not very high, and they are also limited by the maximum penetration depth of the radiation type used. Electrical methods on the other hand are quite popular for a variety of reasons:

- they are non-destructive, and thus useful for on-line process monitoring with standard measurement equipment
- measurements can be directly obtained from the devices whose doping profile must be determined, or from test structures manufactured in the same fabrication process
- their experimental acquisition is straightforward
- they are most closely related to the final goal of the doping profile determination: the understanding of electrical device behavior.

The capacitance-voltage (C-V) method for 1-dimensional doping profiling was first mentioned by Schottky in the 1940's. The early application of the method to Ge diode profiling was reported in the 1960's [17] and many derived methods, too numerous to cite, have since

been described. The C-V method uses the small signal capacitance of the depletion layer as its starting point. As the reverse-bias voltage across the p-n or MOS structure is varied, the measured capacitance changes due to changes in the depletion layer width. The depletion layer width is also influenced by the spatial variation of the doping profile. In the simplest case, the doping profile can be calculated analytically if it is assumed flat on both sides of the p-n junction. A more realistic scenario however must assume that the doping profile is not flat. A computer program is then needed to determine the depth-dependent 1-D doping profile by searching for doping values whose simulated C-V characteristics match the experimental ones.

With the increase in available computational power, there have been several attempts to implement 2-dimensional inverse doping profile extraction techniques in recent years. The first comprehensive review of various such methods including their reliability and error analysis was first given by Ouwerling [18]. His investigations were however limited to specially designed test structures, more relevant to CCD cells than to transistor devices. More recently, Khalil and Faricelli have demonstrated the use of similar techniques by extracting doping profiles from regular transistor-related test structures, such as fingered overlap capacitors [19]. They used cubic splines to model the 2-dimensional doping profiles and they extracted the splines' parameters with the help of a nonlinear least-squares solver. Their extracted doping profiles were shown to yield good C-V agreement with data for a variety of bias voltages.

Another inverse modeling doping profile extraction method based on electrical measurements was recently described by Lee et al. [20, 21]. Their method extracts the 2-D doping profile of sub-micron MOS transistors by using I-V characteristics in the sub-threshold region. They rely on the fact that short-channel effects such as drain-induced barrier lowering (DIBL), sub-threshold slope and punch-through are strongly (exponentially) dependent on the 2-D device doping profiles, and only linearly dependent on other factors such as mobility or gate width. Therefore a relatively accurate doping profile extraction could be performed with the help of a nonlinear least-squares solver, while the uncertainties of the employed mobility model were shown to have only a marginal impact. It is currently believed that such sub-threshold I-V methods are typically more useful when extracting the channel (in-

cluding halo) doping profiles, while the C-V methods are more reliable when describing the source/drain regions of a device. It should also be noted that both the C-V and the I-V methods provide only indirect measures of the device doping. Unlike destructive methods such as SIMS, the C-V or I-V methods measure only the electrically active dopant concentration, without distinguishing between dopant species of the same type. For example, arsenic and phosphorus produce the same C-V and I-V "signatures" because they are both n-type dopants and they are virtually equivalent from an electrostatic point of view as far as their influence on the electrical device characteristics is concerned. This however is sufficient for accurate device simulation, for the exact same reason — because it is only the type and the active doping levels of a device that determine its electrical characteristics.

The next chapter describes the inverse modeling work done in this thesis. The current work is similar to [19], but it uses Gaussian functions (as opposed to cubic splines) to model the doping profiles. Two different gate-to-source capacitance measurements are used in conjunction to extract the 2-dimensional source-drain doping profiles. This work also combines the extraction of most necessary parameters from various experimental measurements and thus makes minimal use of a process simulator. In the end, the validity of the extracted doping profiles is further supported by good agreement between both C-V *and* I-V simulations and data. Like in the work of Lee [20], it is shown that the extracted doping profiles can be used to calibrate device I-V models. In this work the calibration is taken one step further and the extracted doping profiles are used to adjust a new substrate current model.

Chapter 3

Inverse Doping Profiling from C-V Measurements

This chapter is dedicated to formulating 1- and 2-dimensional doping profiling as an inverse problem whose starting points are the device C-V characteristics. The chosen C-V measurements are discussed and the implementation of the extraction technique is explained. The results are analyzed, and their reliability is assessed. The validity of the extracted doping profiles is further supported by good agreement between I-V simulations and data.

The doping profiling work done in this thesis specifically relied on depletion capacitance measurements. Several measurements were made, such as junction capacitance (C_j), gate capacitance (C_g) and gate-to-source capacitance (C_{gs}). The junction and gate capacitance measurements were used to determine 1-D aspects of the device doping profiles, such as the vertical junction 1-D profile, the junction depth, oxide thickness and oxide charges. The gate-to-source capacitance measurements were used to provide insight into the lateral and 2-dimensional distribution of dopants in the source (and drain) region of the device.

3.1 Junction Capacitance

The junction capacitance between the source (or drain) of a MOSFET and its substrate is an important device parameter. It holds clues to the operation speed of the MOSFET, its junction depth, and it can also be used to learn more about the nature of the vertical

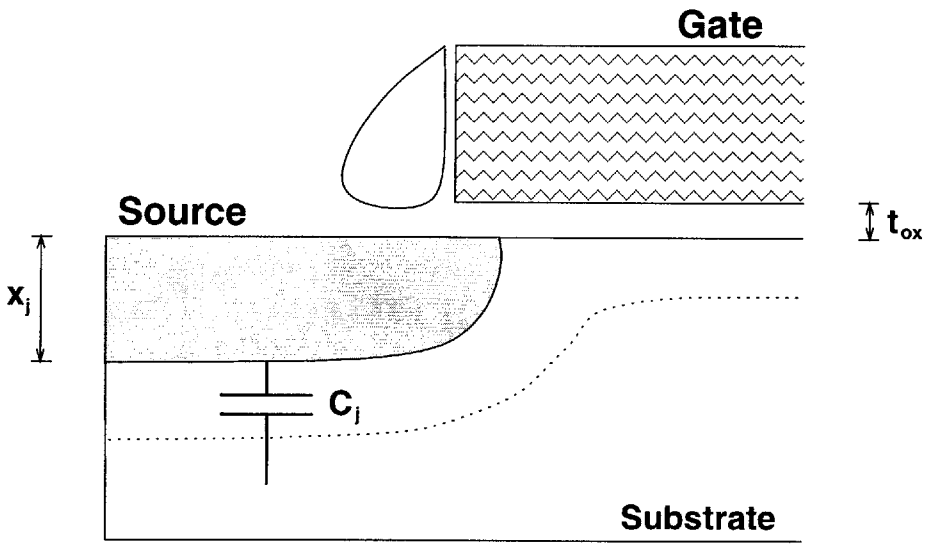

Figure 3-1: The junction capacitance (C_j) of a MOSFET. The oxide thickness (t_{ox}) and metallurgical junction depth (x_j) are marked on the figure as well. The dotted line represents the edge of the depletion region into the substrate. The drawing shows half of a typical MOSFET device, including the gate, source, spacer and substrate.

doping profile in this particular region (see Figure 3-1).

3.1.1 Relationship to Substrate Current

It is important to keep in mind that one of the final goals of this thesis is the calibration of a substrate current model. The MOSFET substrate current is usually represented as an exponential function of the maximum electric field (E_{max}) in most simple first order models [22]:

$$I_{sub} = \alpha I_d \exp\left(-\frac{\beta}{E_{max}}\right) \tag{3.1}$$

where I_d is the drain current and α and β are parameters. The maximum electric field occurs in the channel, near the drain, and can be expressed as:

$$E_{max} = \frac{V_{ds} - V_{dsat}}{l} \tag{3.2}$$

where l should be thought of as an effective ionization length and has been shown to be

directly dependent on the source/drain junction depth x_j [23]:

$$l = 0.22\ t_{ox}^{1/3} x_j^{1/2}. \tag{3.3}$$

The equation above was empirically determined for long channel and thick oxide devices, but a similar relationship exists for short channel devices [24]. Hence, even in a simple model the substrate current is shown to be directly dependent, to first order, on the MOSFET source (and drain) junction depth. More complex, 2-dimensional substrate current models must have good information not only about the junction depth, but also about the vertical variation of the source (and drain) doping profiles in order to correctly reproduce experimental results. The doping information provided by the inverse method presented in this chapter is therefore very valuable to the substrate current model calibration described in chapter 4.

3.1.2 The Inverse Problem

Small signal capacitance measurements with the junction in reverse bias show that C_j is a function of the applied voltage: as the reverse bias across the junction grows, the depletion region widens and the carriers on either side of the junction are pushed apart. This leads to a decrease in C_j as the voltage ($V > 0$) applied to the n-type source diffusion increases[1]. It has been shown [25] that even for an arbitrary doping profile, the measured junction capacitance per unit area is always inversely proportional to the depletion region width:

$$C_j = \frac{\epsilon_{si}}{W_{dep}} \tag{3.4}$$

where ϵ_{si} is the silicon dielectric constant. For the case of a simple 1-dimensional step junction with uniform donor (N_d) and acceptor (N_a) profiles, the capacitance has a simple

[1]The described junction capacitance measurements, as well as the rest of this thesis focus on n-channel MOSFETs. The reason for this focus is ultimately due to the fact that n-channel devices exhibit substrate currents that are about three orders of magnitude higher than those present in p-channel devices. Thus, any other devices described in this work should be implicitly considered to have an n-type channel, source, drain and gate and p-type substrate.

analytical dependence on the applied bias across it [22]:

$$C_j(V) = \sqrt{\frac{\epsilon_{si} q N_d N_a}{2(N_d + N_a)(\phi_{bi} + V)}} \tag{3.5}$$

where q is the magnitude of the electron charge and ϕ_{bi} is the junction built-in potential. Unfortunately in practice the dopants on either side of the junction are rarely uniform: rather they are strongly varying functions of at least one spatial coordinate (depth). In the general case, the capacitance also depends on this spatial variation of the doping, since the depletion region will be less likely to widen into the higher doped regions. By consequence, this property of the voltage- and dopant-dependent capacitance can be used to extract the spatial distribution of the doping across the p-n junction. The doping profile generally does not exhibit any discontinuities and therefore it can usually be modeled by a parameterized analytical function, like

$$f(p_1, p_2, ..., p_m;\ x) \tag{3.6}$$

where $(p_1, p_2, ..., p_m)$ are parameters and x is the spatial coordinate for the 1-dimensional junction capacitance problem. The parameters can then be extracted with the help of a computer program that will search for the set $\{p_i \mid i = 1..m\}$ whose doping profile yields simulated C-V curves that best match the experimental results.

In essence, the technique described above is the definition of inverse modeling. The problem is treated as a "black box" whose outputs (experimental C-V curves) are known but whose inputs (the doping distributions) must be found. In practice, the computer program most often enlisted for help in the search for appropriate doping coefficients is a Levenberg-Marquardt nonlinear least-squares solver [19, 26].

In order to solve the inverse problem an initial guess of the doping profile is first needed. The initial guess may be provided by SIMS analysis, by a process simulator run (e.g. SUPREM) or by using the known doping profiles previously determined for another (similar) technology. In this work, the latter two options were preferred, since using SIMS to provide the initial guess would undermine the non-destructive property of the C-V method!

The general flow of the inverse profiling method is depicted in Figure 3-2. Once the initial guess is provided, a program (the "forward solver") is needed to simulate the first

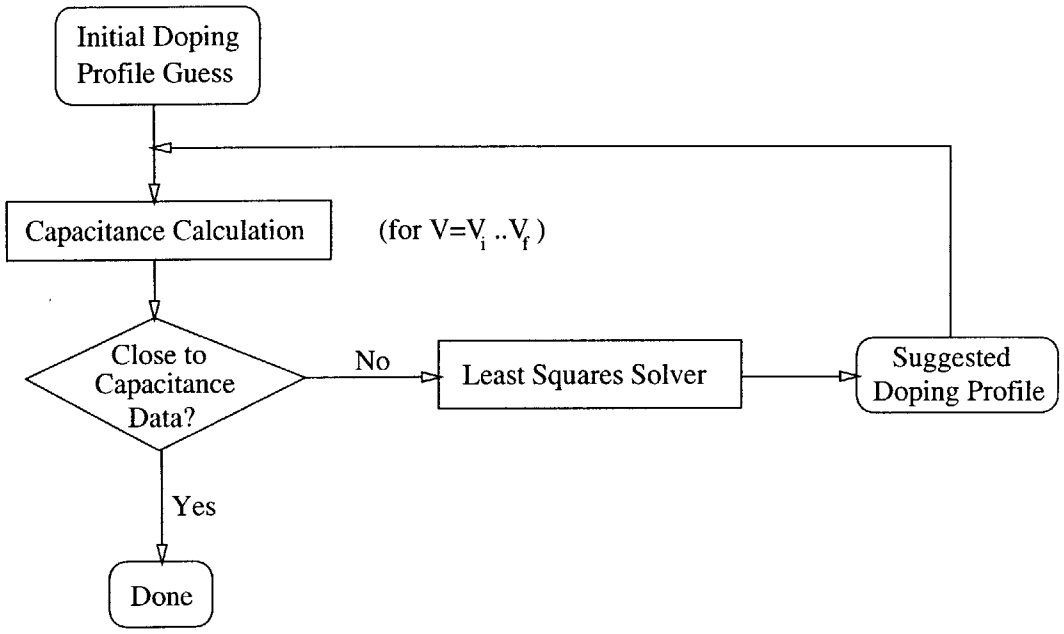

Figure 3-2: General schematic of the inverse method used to extract a structure's doping profile when its C-V characteristics are known.

expected C-V curve. The forward solver can be any program that takes in a set of doping profiles and other device parameters and outputs the computed C-V curve. In other words, it can be a full-blown device simulator (e.g. FIELDAY [11] or MEDICI [27]) or it can be a simple and fast Poisson solver specifically tailored to solve this particular problem. The inverse method treats the forward solver as a black box and does not require any knowledge about its internal workings. The only reason to choose one forward solver over another has to do with its sheer speed and convergence properties. Since the forward problem must be solved many times during the execution of the inverse extraction method, it is important to pick a fast and stable forward solver.

Once the simulated C-V curve is available, the result is compared with the experimental C-V data points. If the mean of the squared differences between the two data sets is deemed too large (by comparison with some user-specified value) the least-squares solver is invoked to find a better set of doping profile parameters. Although not shown on the schematic diagram in Figure 3-2, the least-squares solver in turn reruns the forward problem once

more for each parameter p_i, with slightly different values $p_i + \delta p_i$. This way, numerical derivatives of the capacitance with respect to each parameter ($\partial C/\partial p_i$) are calculated and the sensitivity of the problem to variations in each parameter is gauged.

When the least-squares optimizer finds a new set of parameters, the forward problem is again rerun, the new simulated C-V curve is compared with the data — and if the difference is still deemed too large the procedure briefly described above is repeated. The problem exits only when a suitable new set of parameters is converged upon. What "suitable" really means, as well as a more detailed discussion of the least-squares optimization procedure is provided in appendix A.

3.1.3 Solving for the Junction Capacitance

In the present work, a forward solver for the procedure described above was to be chosen between FIELDAY or a simpler, specially designed Poisson solver. After some experimentation it was decided that a simple program written specifically for the task of solving the 1-dimensional C-V problem was enough, and in fact faster than FIELDAY. After the least-squares-based optimizer was written in C, there was enough code infrastructure that building a fast 1-D Poisson solver and integrating it with the existing code did not present major problems. The entire forward problem is based on an iterative numerical solution of Poisson's equation using Newton's method on a fine enough grid:

$$\nabla^2\phi(x) = -\frac{\rho(x)}{\epsilon_{si}} = -\frac{q}{\epsilon_{si}}\left[p(x) - n(x) + N_d(x) - N_a(x)\right] \tag{3.7}$$

where ϕ is the potential and n and p are the electron and hole concentrations, respectively. The donor and acceptor doping profiles (N_d and N_a) can be expressed with the help of parameterized analytical functions as in equation 3.6. In this work, the analytical functions used to describe the doping profiles are sums of Gaussians, e.g. for N_d:

$$N_d(p_1, ..., p_m;\ x) = \sum_{\substack{i=1 \\ i=i+3}}^{m-2} p_i \exp\left[-\frac{(x - p_{i+1})^2}{2p_{i+2}^2}\right] \tag{3.8}$$

where p_i are parameters whose total number, m, must be a multiple of 3, and x is the spatial coordinate of the 1-dimensional junction. Gaussians were chosen because they are the doping profile shape predicted by the simplest theory of ion implantation in semiconductors [28]. However, due to various heat treatments of a wafer after ion implantation, a single Gaussian may not be enough to describe the final ion distribution. Since three Gaussians would bring too many parameters into the problem, it was decided to choose a reasonable compromise and represent most doping profiles in this thesis as sums of two Gaussians per implant dose.

The electron and hole concentrations in Poisson's equation (3.7) can be obtained from Maxwell-Boltzmann statistics written with respect to the carrier quasi-Fermi levels. Heavy doping effects are accounted for by using an effective intrinsic concentration (n_{ieff}) as first suggested by Slotboom [29]:

$$n(x) = n_{ieff} \exp \frac{q(\phi(x) - \phi_n)}{kT} \tag{3.9}$$

$$p(x) = n_{ieff} \exp \frac{q(\phi_p - \phi(x))}{kT} \tag{3.10}$$

where the quasi-Fermi levels ϕ_n and ϕ_p are determined by the voltages applied to the device's terminals, and n_{ieff} is Slotboom's empirically determined function of doping and temperature. Since the net charge density is given by

$$\rho(x) = q\left[p(x) - n(x) + N_d(x) - N_a(x)\right] \tag{3.11}$$

the total charge associated with the device terminals can be calculated by integrating

$$Q = \int \rho(x) dx. \tag{3.12}$$

Although most experimental measurement setups use small-signal (e.g. 50 mV) sinusoidal test voltages, for the purposes of this simulation the capacitance computed in the electrostatic approximation

$$C_j(V) = \frac{dQ}{dV} \simeq \frac{Q(V + \delta V) - Q(V)}{\delta V} \tag{3.13}$$

is valid and can be used.

3.1.4 Experimental Measurements

The devices studied in this thesis were part of a 3.3 V technology that had been modified to run at 5 V. This was necessary because the core circuitry ran at 3.3 V, but the devices which communicated with the outside world needed to run at 5 V. Since both types of devices had to be built on the same wafer, a few extra process steps were taken to "convert" some of the devices to run out of 5 V power supplies. For example, the 5 V devices were given a thicker, dual oxide layer (roughly 120 A thick) as opposed to the single oxide layer used for the 3.3 V devices (roughly 70 A). The high-voltage devices also had larger minimum channel lengths (0.55 μm versus 0.35 μm) and a different channel implant dose to insure a higher threshold voltage.

Hot carrier effects were diminished by adding an extra source/drain extension implant, in the form of an LDD (Lightly Doped Drain) displaced from the regular high-dose implant by the presence of a spacer. Nevertheless, these devices' measured substrate currents were still relatively high, despite the less steeply graded source/drain junction profiles. The combination of high substrate currents (to be measured), graded doping profiles (to be determined) and a relatively thick oxide (rendering quantum mechanical surface effects somewhat negligible) made these devices a good choice for the study in this thesis.

The junction capacitance C-V data was taken on roughly rectangular, large area and minimum perimeter STI-bound diffusion capacitors. The use of large area capacitors (75,435 μm^2) is useful because the measured capacitances are proportionally larger and the error due to instrumental accuracy and line noise is minimized. Using large area and minimum perimeter structures also minimizes the error due to the side-wall component of the measured capacitance, and the emphasis is kept on the junction capacitance component. To minimize other parasitic capacitance effects, the setup was calibrated by taking readings with the probes lifted off the wafer and subtracting those values from the actual junction capacitance measurements. To get a sense of the general validity of the acquired data, three different chips were measured on the same wafer, and the C-V measurements were repeated several times and averaged for the diffusion capacitor on each chip.

3.1.5 Optimization Results

The measured devices' n-type source and drain were formed with a regular high-dose implant and an LDD implant (both being phosphorus), while their substrate was made up of two boron implants (a shallow and a deep one) and the relatively constant background doping (about 5×10^{15} cm^{-3}).

Using two Gaussian functions for each implant dose (with three parameters for each Gaussian) quickly leads to a total of twenty-four coefficients to be optimized for the entire problem. Such a problem is clearly something best left for a computer to solve. However due to large computation time demands and the possibility of doping coefficients' divergence beyond physically reasonable limits only two or three parameters were optimized at a time, the others being held constant.

The initial guess was provided by fitting sums of Gaussians to the n- and p-type doping profiles extracted from a SUPREM run. Both intuitively and after a few simulation runs it became apparent that the parameters determining the Gaussians' displacement and standard deviation (e.g. p_{i+1} and p_{i+2} respectively in equation 3.8) were most strongly responsible for the shape of the C-V curve. Those parameters were allowed to adjust first, other ones being held constant. Afterwards the parameters determining the peak donor and acceptor concentrations were allowed to adjust. In general however, it was found that degenerate peak concentrations (above 10^{19} cm^{-3}) had little to no effect on the outcome of the C-V curve.

The final results of these computations are displayed in Figures 3-3 and 3-4. The experimental data in Figure 3-3 came from three diffusion capacitors (across three chips) on the same wafer, and several measurements were performed and averaged on each chip. The doping profile shown in Figure 3-4 was optimized using the average of the C-V data over the three chips. The origin of the x axis in Figure 3-4 is at the surface of the wafer and the extracted junction depth is thus approximately 0.265 μm, in very good agreement with that extracted by SIMS analysis on the same devices at a later point in time.

Despite the good agreement between the final simulated C-V curve and the data, the limitations of the extracted doping profile must be understood. As mentioned before, the extraction method loses its sensitivity for degenerate doping levels (above 10^{19} cm^{-3}) be-

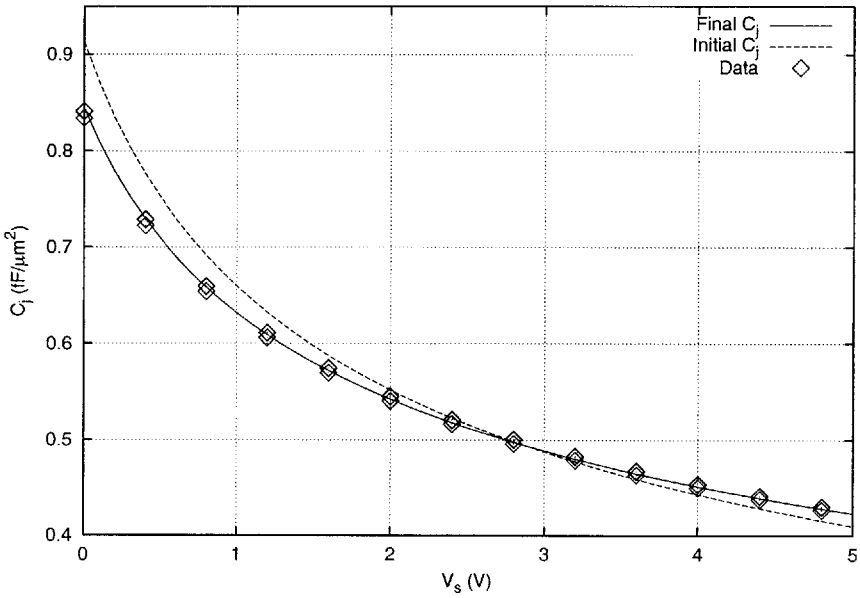

Figure 3-3: Comparison of junction capacitance per unit area measured across 3 different chips (symbols) and simulation (lines) before and after the doping profile optimization.

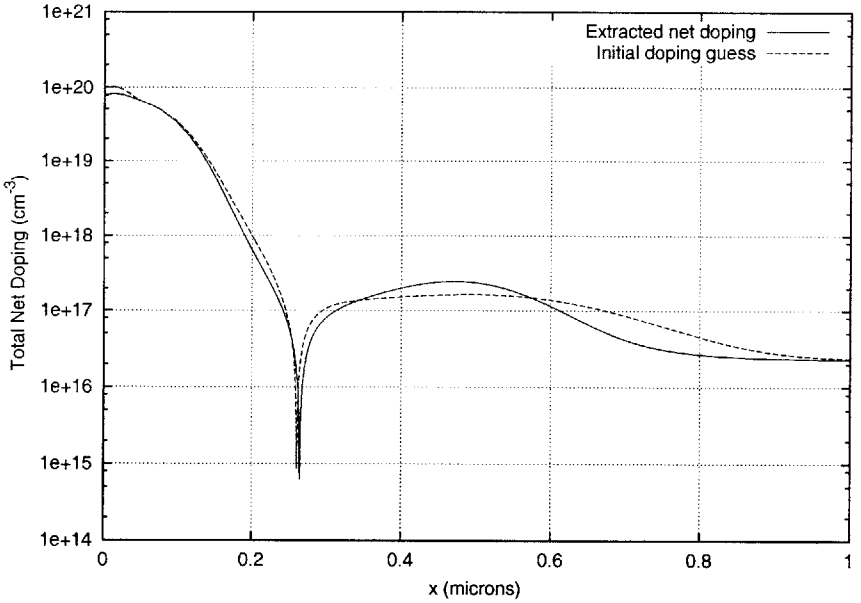

Figure 3-4: Vertical junction net active doping $|N_d - N_a|$: initial and extracted profiles. The left side of the junction is the n+ source and the right side is the p-type substrate.

cause the potential variation diminishes there. Also, since this sort of C-V inverse modeling relies on the voltage dependence of the depletion width, the extracted doping profiles are limited by the applied range of voltages. The applied range of voltages is in turn limited by the turn-on of the p-n junction at one end and by junction breakdown at the other. At the maximum applied voltage of 5 V, the maximum width of the depletion region is on the order of 0.3 μm, so the doping profile deeper into the substrate (e.g. for $x > 0.6$ μm) is susceptible to some error. However the doping profile at such depths has little influence on the general device characteristics. Also it should be noted that very fine details of the doping profile *within* the depletion region are likely to be averaged out, since the potential in Poisson's equation is a very smooth function in space (being a double integral of the space charge). This may not be a tremendous issue however, since in the processing of MOSFETs the diffusion of dopants also results in smooth doping profiles [20].

3.2 Gate Capacitance

Several gate capacitance measurements were performed in order to determine such device characteristics as the oxide thickness, the polysilicon gate doping and the density of interface traps (DIT) at the Si/SiO_2 interface. Like the junction capacitance, the gate capacitance was also measured on specially designed large area (65,457 μm^2) and minimum perimeter test structures in order to minimize 2-D fringing field effects and the contribution of the side-wall capacitance. The structures used to measure the gate capacitance were rectangular STI-bound capacitors — essentially just large MOS sandwiches with a 0.2 μm thick phosphorus doped polysilicon gate on top, the p-type silicon substrate underneath, and the dual, thicker oxide in between (corresponding to the 5 V devices). The gate capacitance structures were formed through the same process steps and on the same wafer as the junction capacitance structures previously described, and as the MOSFET devices to be later measured for their I-V characteristics.

In order to extract the DIT, the method described in [30] was followed: a high frequency (100 kHz) C-V measurement was first performed, followed by a quasi-static sweep with a slow voltage ramp (50 mV/sec). The interface traps can be easily filled or emptied during

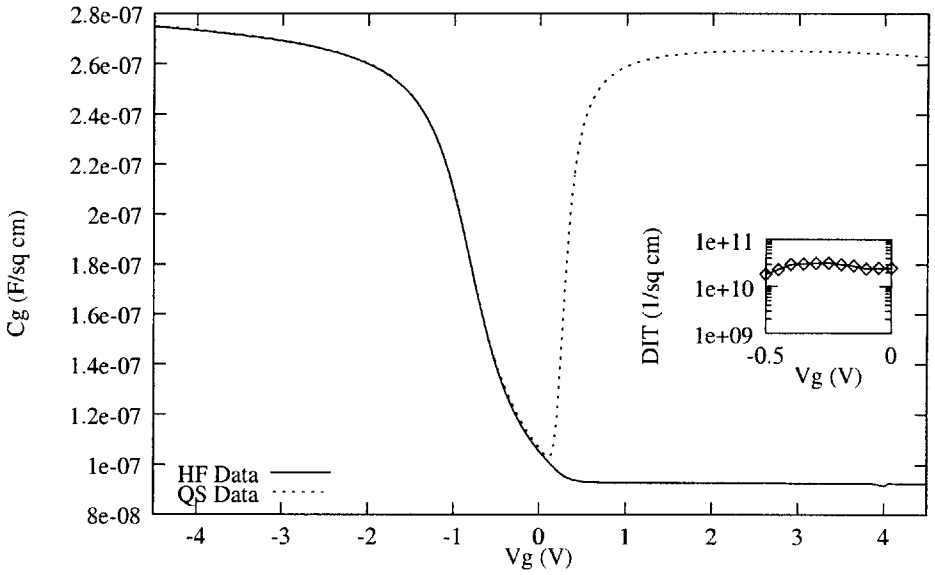

Figure 3-5: High frequency (HF) and quasi-static (QS) C-V data averaged over seven consecutive measurements in order to reduce experimental noise. The extracted interface trap distribution as a function of gate voltage is shown in the insert.

the quasi-static measurement, but they cannot keep up with the high frequency signal. Therefore a comparison of the two obtained C-V curves (see Figure 3-5), especially in the depletion region, can be used to gauge the DIT and its variation with the gate voltage. The oxide thickness can be computed by linearly extrapolating the high frequency C versus 1/V data in the strong accumulation region, yielding approximately 120 A; the oxide capacitance C_{ox} follows immediately. Quantum mechanical surface effects are nearly negligible in devices with such a thick oxide.

The interface trap contribution to the gate capacitance can be obtained directly from the measured C-V curves, as described in [30]:

$$D_{it} = \frac{C_{it}}{q} = \frac{(1/C_{QS} - 1/C_{ox})^{-1} - (1/C_{HF} - 1/C_{ox})^{-1}}{q}. \tag{3.14}$$

The results of the C-V measurements and the DIT extraction are shown in Figure 3-5. In addition, the polysilicon doping can be inferred by comparing C-V simulations with the acquired data at high gate voltages in the inversion region — where signs of polysilicon

depletion become apparent. Such comparisons indicated that the active polysilicon doping for the MOS structures being tested was around 6.7×10^{19} cm^{-3}. This value was later confirmed by a similar result obtained from 2-dimensional gate-to-source inversion capacitance measurements (see section 3.3.6).

3.3 Gate-to-Source Capacitance

The two capacitance measurements described thus far have only offered insight into the 1-dimensional (vertical) structure of the devices under study. On the other hand, the gate-to-source (or gate-to-drain) capacitance is essentially a 2-dimensional capacitance, and it can therefore offer insight into the 2-D structure of the MOSFET device.

The gate-to-source capacitance is a key parameter for device reliability and circuit speed. Moreover, its magnitude holds important clues about the lateral extent of the source diffusion under the gate. It has been shown [19, 31, 32] that the voltage dependence of the gate-to-source capacitance (C_{gs}) can be used to probe the extent of the source diffusion under the gate and to provide a good measure of the overlap length. Assuming a symmetric MOSFET structure, the gate-to-source and gate-to-drain capacitances are equivalent. Therefore any further discussion in this thesis referring to the structure and doping profiles of the source of a MOSFET is equally relevant about its drain: the two can be reversed by simply reversing the polarity of the applied bias. Hence any knowledge of the lateral MOSFET diffusion profiles obtained from gate-to-source C-V measurements is relevant for substrate current simulations described in chapter 4 — where the electric field, carrier temperature and impact ionization rate near the *drain* have a pronounced doping profile dependence.

3.3.1 Capacitance Components

The various capacitance components that make up the gate-to-source capacitance (C_{gs}) of a MOSFET are illustrated in Figure 3-6. C_{if} is the inner fringing capacitance associated with the electric field emerging from the inner side of the of the source and ending at the underside of the polysilicon gate. C_{ov} is the overlap capacitance associated with the

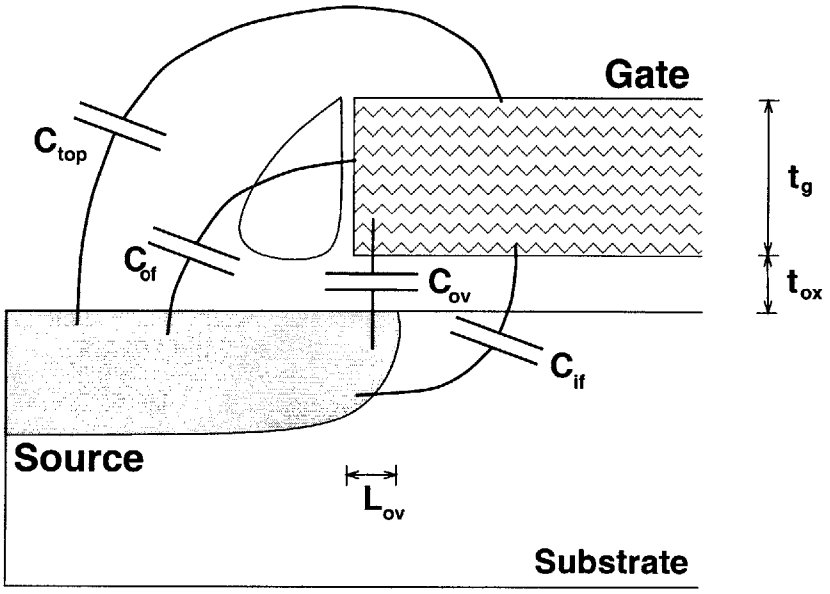

Figure 3-6: The various components that make up the gate-to-source capacitance, C_{gs}. Also marked on the figure are the gate-source overlap L_{ov}, the oxide thickness t_{ox}, and the gate thickness t_g.

voltage- and doping-dependent gate-to-source overlap region (L_{ov}). C_{of} is the outer fringing capacitance associated with the electric field emerging from the side of the gate, going through the side-wall spacer and ending at the top of the source region. Finally, C_{top} is the capacitance due to the electric field lines emerging from the top of the gate, going through the first passivation layer and ending at the top of the source.

C_{of} and C_{top} are virtually bias independent, being mainly determined by physical properties of the MOSFET such as the the gate and oxide thickness (t_g and t_{ox}), the gate length (L_g) and the choice of spacer and passivation layer dielectrics. C_{top} is perhaps the smallest component of the total gate-to-source capacitance because of the relatively large distance the electric field lines need to travel from the top of the gate to the top of source. C_{top} was in fact ignored in most simple calculations until recently, when an analytical formula for it was proposed [33]:

$$C_{top} = \epsilon_{ox} \ln\left(1 + \frac{L_g}{t_{ox} + t_g}\right) \tag{3.15}$$

where L_g is the polysilicon gate length and the use of ϵ_{ox} assumes an oxide passivation

layer. For the devices being studied $t_{ox} = 120\ A$, $t_g = 0.2\ \mu$m and the line width of the overlap structures is $L_g = 0.75\ \mu$m, so equation 3.15 predicts a top capacitance component of about $0.052\ fF/\mu$m.

The outer fringing capacitance is generally more significant and a simple analytical formula for it has been derived as well [34]:

$$C_{of} = \frac{2\epsilon_n}{\pi} \ln\left(1 + \frac{t_g}{t_{ox}}\right). \tag{3.16}$$

This formula assumes that most electric field lines pass through the side-wall spacer (in this case nitride, with dielectric constant ϵ_n) and that the side-wall of the gate forms a 90° ($\pi/2$) angle with the wafer surface. For the devices being studied in this thesis equation 3.16 predicts $C_{of} \simeq 0.116\ fF/\mu$m, more than twice the size of C_{top}.

Unlike the top and outer fringing capacitance components, both the overlap capacitance (C_{ov}) and the inner fringing capacitance (C_{if}) are bias-dependent. Although C_{ov} can be roughly approximated as

$$C_{ov} = \frac{\epsilon_{ox} L_{ov}}{t_{ox}} \tag{3.17}$$

this "parallel plate" approximation is very rudimentary due to edge fringing effects, and C_{ov} is usually best obtained from full 2-dimensional device simulations. However it is important to note that the overlap capacitance is still directly dependent on the overlap length (L_{ov}), and that they are both voltage-dependent.

The C_{ov} dependence on the source voltage (V_s) is due to the depletion of the source-to-substrate junction when V_s is varied, and it is also influenced by the local doping profiles of the source and substrate. When $V_s = 0\ V$, a small electron inversion layer exists near the channel-side of the junction, adding to the overlap capacitance. Because the net doping profile is lower there, the threshold voltage is lower as well. However as V_s is increased, the localized inversion layer diminishes, thus causing a drop in C_{ov} [32]. As V_s is increased further, the source side of the depletion region begins to meet the highly doped source diffusion, and the depletion width becomes almost a constant, flattening out the C_{ov} versus V_s curve (see the data Figure 3-11).

The overlap capacitance dependence on the gate voltage (V_{gs}) is even stronger, because

a high enough gate voltage will lead to the formation of an electron inversion layer connected to the source and stretching the entire length of the channel, thus suddenly increasing L_{ov} and C_{ov} [35]. At the low end of the gate voltage range, a low enough (negative) V_{gs} will induce a hole accumulation layer underneath the gate, thus shielding the inner fringing capacitance component (C_{if}) and decreasing the overall gate-to-source capacitance (see the data in Figures 3-9 or 3-10).

It is this voltage and doping dependence of the overlap, and consequently of the gate-to-source capacitance that enables the use of inverse modeling for the extraction of the lateral source and channel doping profiles.

3.3.2 Experimental Measurements

Gate-to-source capacitance measurements were performed on large perimeter (99,819 μm) and small area fingered, "comb"-like structures. Unlike for the junction and gate capacitance measurements described in the previous sections, structures with a large perimeter are necessary in gate-to-source capacitance measurements because C_{gs} is essentially a perimeter capacitance — typically measured in fF/micron. Having a small area to perimeter ratio also insures that other components (e.g. the gate to bulk capacitance) are minimized and the measurement results emphasize the perimeter capacitance. The fingered measurement structures used in this study had a polysilicon (gate) line width of $L_g = 0.75$ μm and a minimum line spacing of 2 μm — thus keeping the capacitive coupling between adjacent polysilicon fingers to a minimum. Otherwise, the measurement structures were formed on the same wafer and under the same process steps as all other devices described in this work.

Two different measurements were performed on the 3-terminal fingered overlap structures. In both cases the substrate terminal was grounded ($V_{sub} = 0V$). In the first case the gate was grounded as well ($V_g = 0V$) and the voltage was varied on the source terminal, $-5V < V_s < 5V$. In the other case the source was grounded ($V_s = 0V$) and the gate voltage was ramped from -5 to 5 V. Because of the nature of the fingered test structures, the source diffusion surrounded the entire perimeter of the polysilicon gate fingers, therefore also serving as the drain diffusion. From an electrostatic point of view the fingered overlap structures were thus equivalent to MOSFETs with the source and drain shorted

together to form a single terminal. The data obtained from the measurements are shown in Figures 3-10 and 3-11. The physical interpretation of the data and a description of the 2-D inverse modeling technique are provided in subsequent sections of this chapter.

3.3.3 The 2-D Problem

Unlike for the junction capacitance calculations, it was decided that a full-blown device simulator (FIELDAY) would be better suited for inverse modeling the 2-dimensional gate-to-source capacitance. The suite of TCAD simulation software available at IBM was therefore used: the program DOPING [36] was used to place analytical Gaussian doping profiles on top of an otherwise "blank" mesh. Since 2-D computation is much more numerically demanding than 1-D computation, the mesh needed to be optimized with the program REGRID [37] before a FIELDAY device simulation could be run. All REGRID meshes used in this work were carefully chosen to be small enough to allow for quicker solutions, but dense enough not to introduce mesh-related errors in the outcome of the device simulation (see appendix C for sample input files used in the inverse modeling process). A typical half-device mesh used for C_{gs} modeling contained roughly 4500 nodes. An illustration of such a doped and optimized mesh is provided in Figure 3-7.

3.3.4 The FITDRF Optimizer

The general optimization procedure described in section 3.1.2 applied to the 2-D problem as well, with a few caveats: the analytical Gaussians became 2-dimensional Gaussians with both x and y coordinate parameters, and the Levenberg-Marquardt nonlinear optimizer had to be rewritten in order to communicate with all other programs that were utilized during one loop of the optimization process: DOPING, REGRID and FIELDAY. A schematic of the 2-D inverse modeling method is presented in Figure 3-8. The initial doping profile was provided again by a set of Gaussians fitted to the output of a SUPREM simulation. However the coefficients that determine the vertical doping profiles were fixed based on the results of the previous junction capacitance investigations. The only doping profile parameters that were allowed to vary were therefore the ones determining the lateral spread of the source diffusion under the gate.

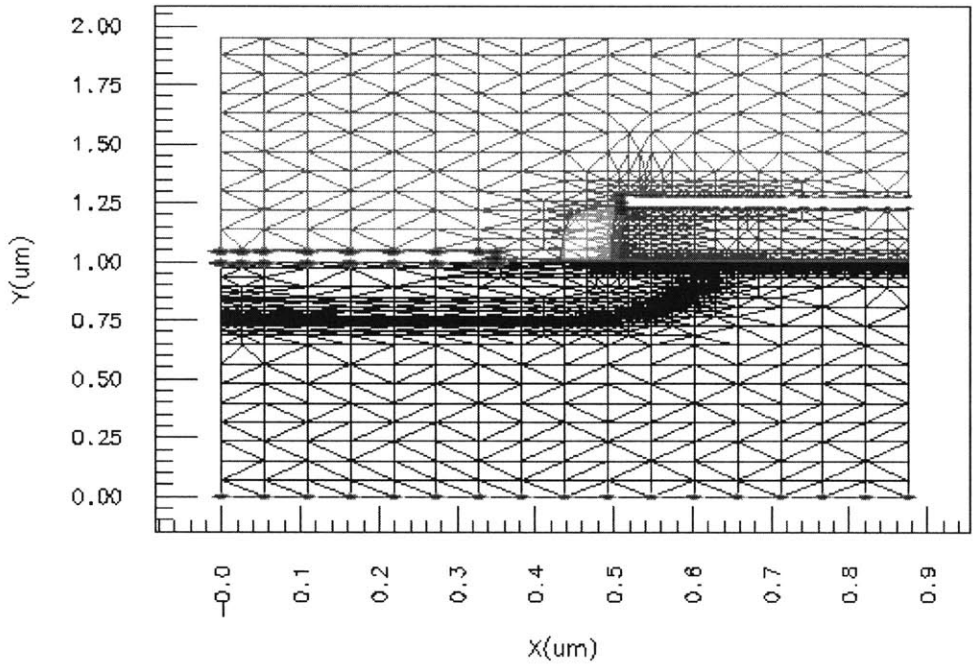

Figure 3-7: Mesh used in FIELDAY for C_{gs} simulation. Compare with Figure 3-6 and note the presence of the gate side-wall spacer and the top passivation oxide. The mesh has been optimized for device simulation using the program REGRID.

The program FITDRF was created as a general-purpose optimizer, not only for doping profile parameters (like in the scheme of Figure 3-8), but as an optimizer for particular REGRID or FIELDAY parameters as well. The program is invoked with its own input file and command-line arguments that fully define the user's intentions. FITDRF was built on top of the Levenberg-Marquardt infrastructure used in the simpler 1-dimensional junction capacitance case, but it represents a much more general optimizer, that can be used along with any of the three IBM TCAD programs mentioned above. A more in-depth overview of FITDRF, its functions and usage has been relegated to appendix B.

3.3.5 Gate Voltage Dependence

Following the method of Koldyaev [35], the gate-to-source capacitance dependence on the gate voltage (V_{gs}) was investigated first. The initial "blank" (devoid of doping) mesh was taken from a SUPREM simulation, including the gate and spacer dimensions and

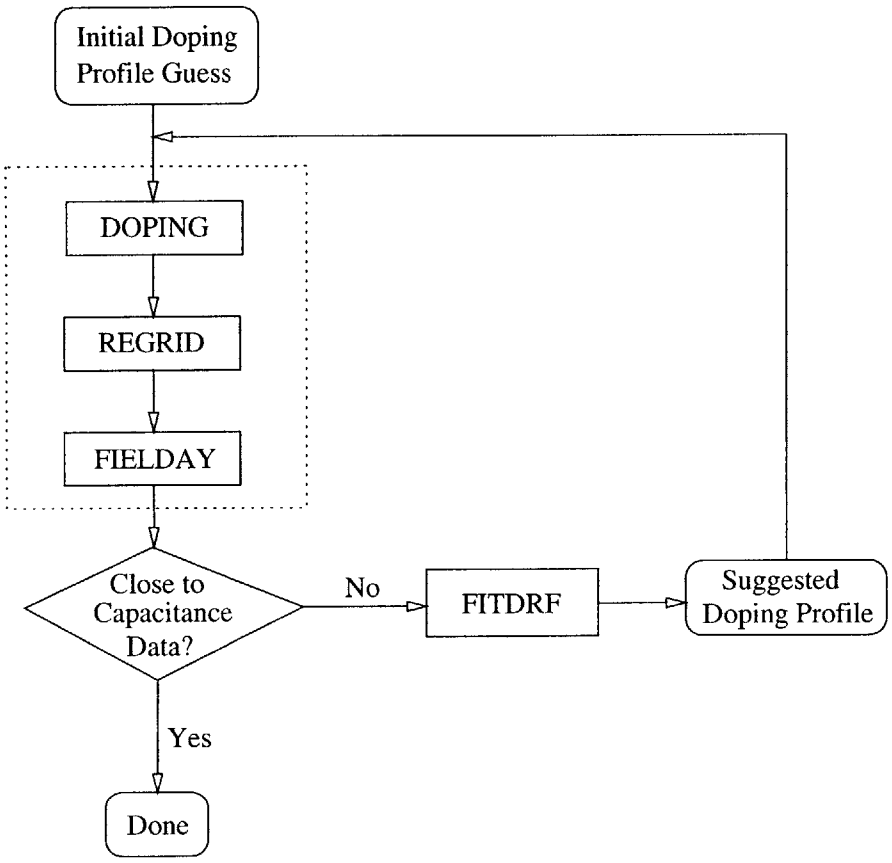

Figure 3-8: Schematic of the inverse modeling method used to extract the 2-dimensional doping profiles based on C_{gs} measurements. The dotted line surrounds the three programs that make up the "forward" solver.

the passivation and gate oxide thickness. The gate oxide thickness (t_{ox}), a strong factor in determining C_{gs}, was carefully set based on the previously extracted value from high frequency gate capacitance measurements (section 3.2). As discussed in section 3.3.1, the gate-to-source capacitance suddenly increases when the applied gate voltage exceeds the threshold voltage and an inversion layer connected to the source is formed. This transition point is very sensitive to the channel doping (N_{ch}), and therefore N_{ch} can be directly extracted via inverse modeling. Moreover, the gate-to-source capacitance in the inversion region is strongly dependent on the overlap structure's gate length, so L_g can then be extracted as well.

Figure 3-9 illustrates the effects of a different channel doping and gate length on the gate-to-source capacitance characteristic. The solid line represents the simulation with the optimized doping profile, whereas the dotted lines show simulations done with a lower channel doping (56 % of the extracted value) and a lower gate length (0.5 microns). The lower channel doping allows the channel to invert at a lower gate voltage, and thus the newly formed electron inversion region, which is electrically connected to the source, immediately increases C_{gs}. The shorter gate length induces a proportionally smaller value of C_{gs} in inversion, because the shorter L_g leads to a smaller inversion area — and thus a smaller inversion capacitance component. The values extracted by FITDRF for the channel doping and the gate length were $N_{ch} = 1.75 \times 10^{17}$ cm^{-3} and $L_g = 0.75$ μm. The extracted value for the gate length was in very good agreement with the designed line width of the fingered gate-to-source overlap structures: 0.75 μm.

Figure 3-10 shows comparisons of the extracted doping profile's capacitance simulation (solid line) with and without the top oxide passivation layer and the gate side-wall spacer. As theoretically expected, the passivation oxide and gate side-wall spacer have voltage-independent capacitance contributions, thus causing a constant down-shift of the C-V curve when removed from the simulation mesh. From the 2-dimensional computation it appears that the passivation oxide contribution to C_{gs} is about 0.05 fF/μm, in good agreement with the analytical result for C_{top} from equation 3.15: 0.052 fF/μm. On the other hand, the voltage-independent value of the gate side-wall spacer contribution to C_{gs} appears to be about 0.065 fF/μm when extracted from the 2-D simulation results in Figure 3-10.

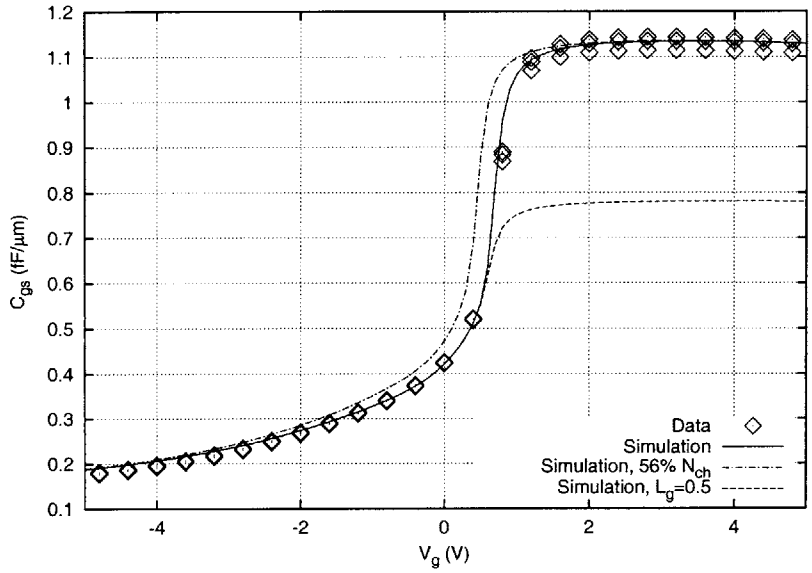

Figure 3-9: Plot of gate-to-source capacitance data across 3 chips (symbols) and several simulations (lines) as a function of gate voltage. The solid line represents the simulation with the optimized doping profile and the dotted lines represent simulations done using a lower channel doping and a lower gate length, respectively.

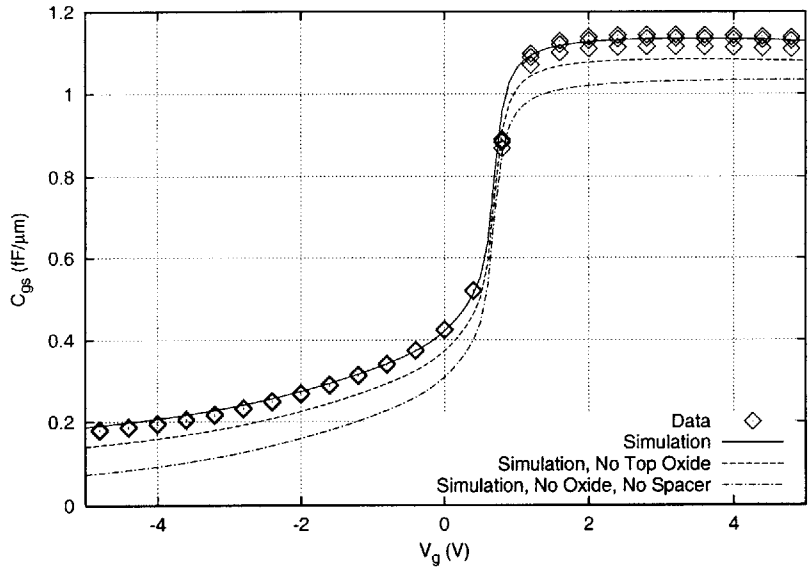

Figure 3-10: Plot of gate-to-source capacitance data across 3 chips (symbols) and several simulations (lines), as a function of gate voltage. The simulations were run with and without the top oxide (passivation) and with and without the gate side-wall spacer.

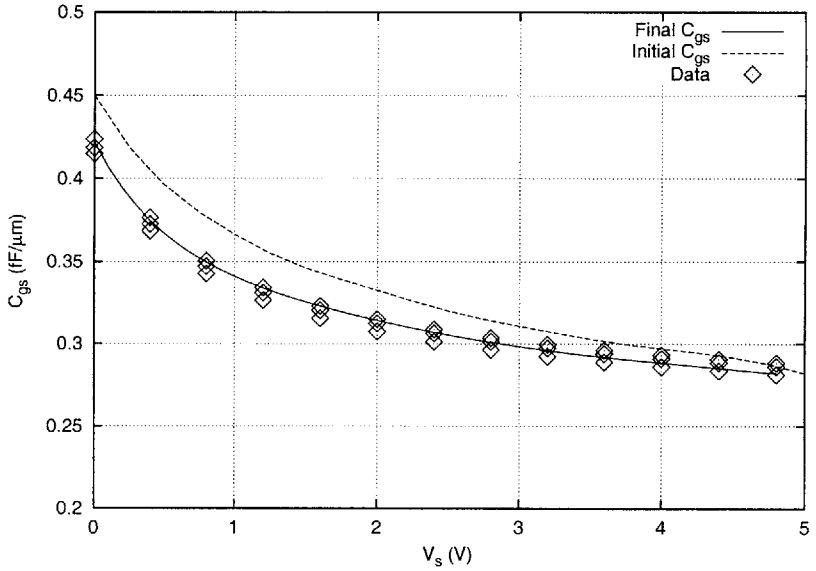

Figure 3-11: Comparison of gate-to-source capacitance per unit gate width measured across 3 different chips (symbols) and simulation (lines) before and after the doping profile optimization — as a function of source voltage.

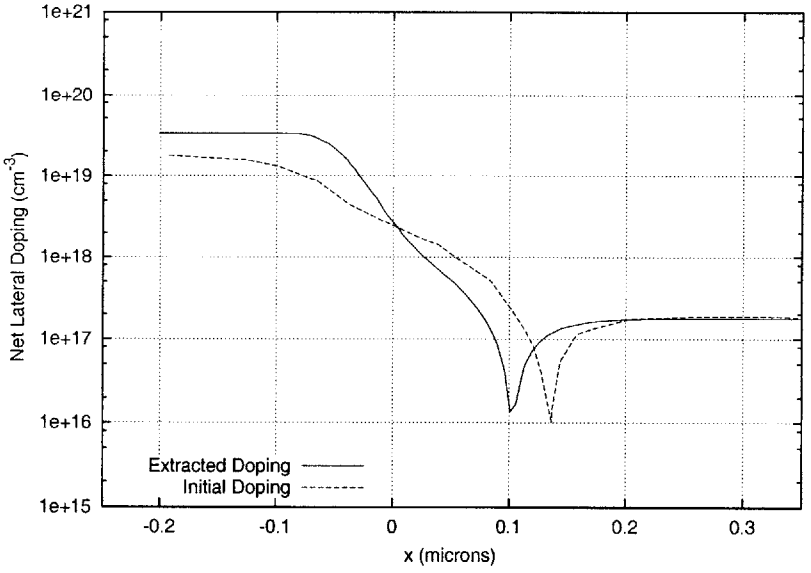

Figure 3-12: Lateral junction net active doping: initial guess and extracted profile at 0.1 microns below the oxide/silicon interface. The left side of the junction is the n-type source and the right side is the p-doped substrate.

This value is almost twice smaller than the value of $C_{of} = 0.116$ fF/μm predicted by equation 3.16. The discrepancy is most likely due to the approximate nature of the analytical equation and to the fact that the electric field lines from the gate side-wall to the source pass through a finite thickness (about 65 nm) of nitride, but continue the rest of their way through oxide. For simplicity, only the (higher) dielectric constant of nitride had been used in equation 3.16. Nevertheless, it is clear that both the passivation oxide and the gate side-wall spacer contributions to the gate-to-source capacitance are significant and must be included in the 2-dimensional simulations. The analytical equations' results should be used for quick order-of-magnitude estimates, but they are not accurate enough to be relied on for doping profile extraction calculations.

Finally, one more thing can be learned by carefully studying the strong inversion region ($V_{gs} > 2$ V) in Figure 3-10: the C_{gs} characteristic begins slightly "bending down" as the gate voltage is increased — the effect being due to slight polysilicon gate depletion. Since the n+ gate was modeled with a flat doping profile, the value of the active gate doping could be extracted through inverse modeling. A gate doping of about 7.2×10^{19} cm^{-3} led to a simulated capacitance that best matched the gate depletion curvature of the data. This value is in good agreement with that extracted from gate capacitance measurements in section 3.2.

3.3.6 Source Voltage Dependence

Following the suggestion of Oh et al. [32], the gate-to-source capacitance dependence on V_s was also investigated. For the inverse modeling procedure, the doping parameters previously extracted from the junction, gate, and gate-to-source capacitance analysis were kept constant — and only the lateral source doping parameters were allowed to vary. As explained in section 3.3.1, the gate-to-source capacitance is dependent on the applied source bias because the edge of the depletion region and therefore L_{ov} varies with changes in V_s (also see Figure 3-6). This dependence is also strongly influenced by the spatial variation of the source-to-channel doping profile in the lateral direction — thus enabling the inverse doping extraction technique.

Figure 3-11 shows the simulated gate-to-source capacitance before (dotted line) and

after the doping profile optimization (solid line) — as compared with data taken on fingered overlap structures across three different chips. The lateral coefficients for the initial doping guess were provided by fitting Gaussians to lateral cross sections obtained from a SUPREM simulation. Figure 3-12 shows these lateral profiles before and after the doping profile optimization. Everything else being constant, the lateral edge of the source region had to "retract" from underneath the gate, since the initial capacitance guess was too high when compared with the data. Clearly, even relatively small differences in the capacitance curves lead to significant extracted differences in the doping profiles, therefore making it even more important to include the side-wall spacer and the passivation oxide in the C-V inverse modeling procedure. As with the vertical doping extraction from junction capacitance measurements it was noted that degenerately high doping levels are less accurately extracted — though from an electrical point of view small variations in such high doping levels don't play a very important role in the device behavior, since the potential varies only very little.

3.4 Drain Current Simulations

The device doping profiles previously obtained via inverse modeling from C-V data were finally put together to simulate full 2-dimensional MOSFET drain current characteristics.

A 2-D MOSFET doping-free mesh was obtained from SUPREM and the extracted analytical doping profiles were added to a half-device with the program DOPING (see Figure 3-7). The doped mesh was then prepared for the 2-D FIELDAY current simulation by adding a few lines to the REGRID input file shown in appendix C:

```
&MIRROR MIRROX='R',
        CHOPX='R',
&END
```

These lines "reflected" the half device mesh and converted it into a fully symmetric MOSFET ready for FIELDAY simulation. The data was taken on devices with gate lengths of 0.5, 0.6, 1.0 and 5.0 microns — so the simulated devices' gate lengths were adjusted accordingly. All measured devices were 20 microns wide.

The drain current simulations were run with FIELDAY II [38] using the post-processed impact ionization model described in chapter 4. The impact ionization parameters were

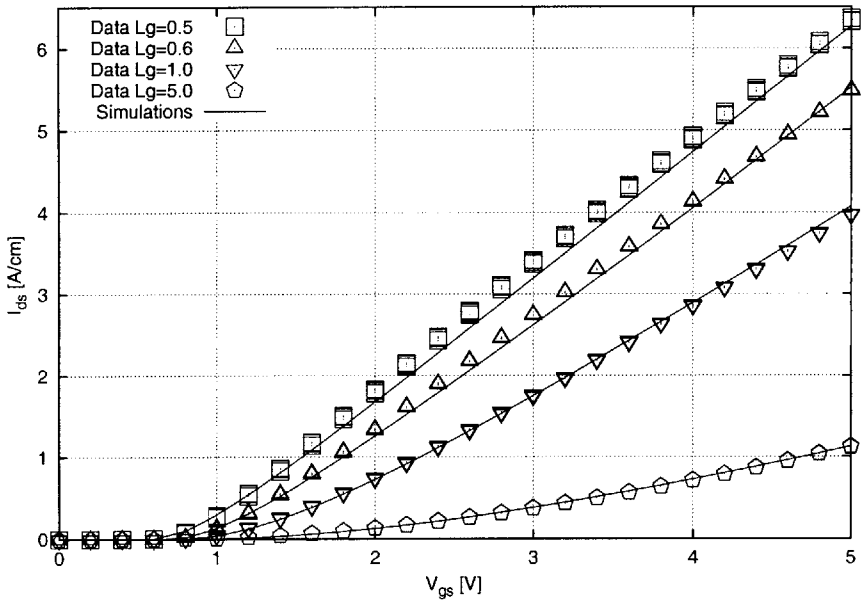

Figure 3-13: Measured and simulated drain currents per unit width for devices with 0.5, 0.6, 1.0 and 5.0 microns gate length. The width of the measured devices was 20 microns. The drain bias was set at 5 V whereas the source and the substrate were grounded.

manually set such that the simulated and measured substrate currents were relatively close. Otherwise most FIELDAY parameters were left at their default values. The MINIMOS mobility model [39] was used, but its parameters were also left at their default values.

The very first FIELDAY runs using the extracted doping profiles were in remarkable agreement with the data. Only one FIELDAY parameter, the source and drain contact resistance (RESISTOR=0.06), was optimized with FITDRF to obtain the plots displayed in Figures 3-13 and 3-14. The experimental data in these figures was taken across three different devices on the same wafer.

3.5 Summary

This chapter presented a 1- and 2-dimensional doping profile extraction procedure using C-V measurements. The procedure was treated as an inverse problem whose outputs (the device electrical characteristics) were known, but whose inputs (the device doping profiles) were to be found. The doping profiles were expressed as sums of Gaussians whose coefficients had

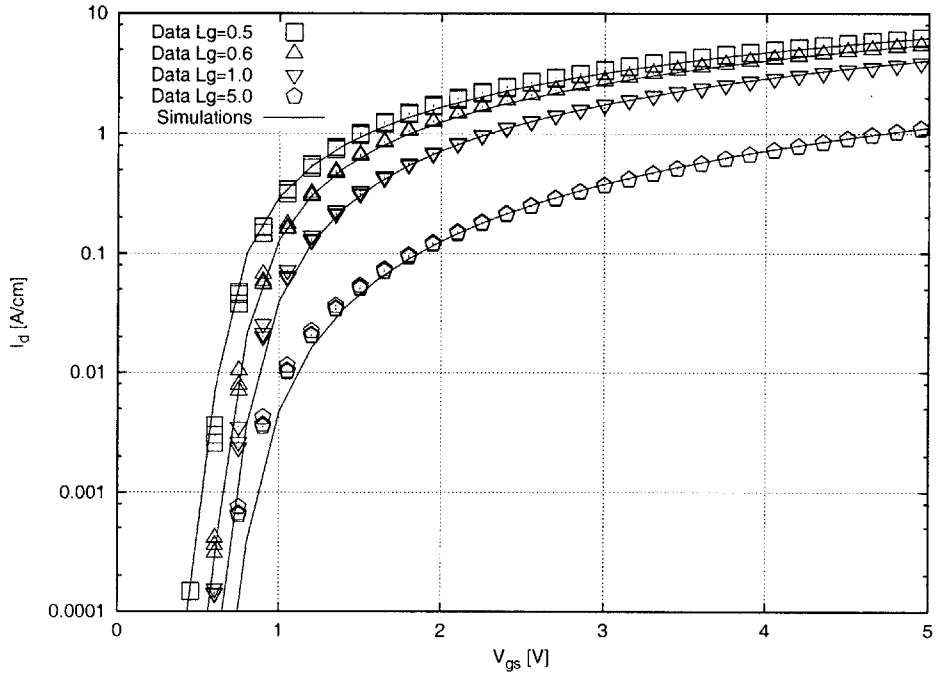

Figure 3-14: Log scale comparison between simulated drain currents and data for the same devices as in Figure 3-13.

to be solved for. FITDRF, a general-purpose optimizer based on the Levenberg-Marquardt least squares algorithm, was introduced and used to extract the Gaussian parameters. Because solving such a highly nonlinear problem with approximately 20 parameters at once would have led to immediate convergence problems, several measurements were made and some educated decisions were employed. In practice only two or three parameters were optimized at a time.

In the end, the extracted doping profiles presented C-V electrical properties that were remarkably close to the ones of measured devices. The doping profiles were also used for full device drain current simulations, and only one single FIELDAY parameter (the source and drain contact resistance) needed to be fine-tuned to produce very good agreement with data. All these results indicate that the extracted 2-dimensional doping profiles are reliable and physically accurate.

Chapter 4

Substrate Current Modeling

This chapter presents some of the issues behind accurate substrate current modeling in modern device simulators. The beginning of the chapter reviews a few of the approaches that have been taken to study impact ionization in semiconductors, and therefore substrate currents in MOSFETs. The rest of the chapter focuses on the theory and assumptions of an existing carrier-temperature-dependent impact ionization model [12], as implemented in the device simulator FIELDAY. A parameterized high energy tail is introduced in the carrier distribution function and the impact ionization rate is re-derived and implemented in FIELDAY. The new impact ionization model is calibrated and analyzed within the context of the previously determined device doping profiles. The accuracy of the doping profile information is shown to be essential for obtaining reliable substrate current simulations.

4.1 Motivation

In recent years MOSFET feature sizes have been continuously scaled down into the sub-micron range. This size reduction has caused an increase of the maximum field strength inside the device and thus, an increase of the substrate current. The amount of substrate current in turn is an important indicator of device aging and reliability [40]. Electric fields and therefore carrier heating effects are highly sensitive to the doping profile distribution inside the device. In practice, smaller substrate currents are generally obtained by carefully tailoring the device doping profiles in order to minimize a given device's electric fields.

Lower dose extension implants (e.g. Lightly Doped Drain or LDD) are usually added to a device's source and drain, offset with the help of a spacer. Such extra implants can help achieve less steep doping gradients, but they also introduce more unknowns in the doping profiles that are used for device modeling. Hence accurate inverse doping profile extraction methods (such as that presented in chapter 3) can and should be used to alleviate one of the main uncertainties related to substrate current modeling, such that attention can then be devoted to electric field or impact ionization rate calculations. This is the approach adopted in this thesis.

4.2 Impact Ionization

Impact ionization is the process of electron-hole pair creation through the breaking of a lattice bond by a charged carrier whose kinetic energy exceeds the threshold for bond breaking. This threshold is called the ionization threshold and is comparable to the band gap energy of the semiconductor. Impact ionization is essentially the inverse process of Auger recombination [41].

In the case of impact ionization by electrons, the impact ionizing electron loses most of its energy by interacting with a valence band electron. A lattice bond is broken and the valence band electron is then promoted to the conduction band, leaving behind a hole:

$$e_{c+} \rightarrow e_{c'} + e_c + h_v \tag{4.1}$$

Figure 4-1 shows this process schematically. The impact ionizing "hot" electron (e_{c+}) and the original valence band electron (e_c, initially present at the location of h_v) interact via their screened Coulomb potentials. The outcome of the process leaves the impact ionizing electron at a lower energy level in the conduction band (e'_c) and adds an electron-hole pair ($e_c - h_v$) to the total number of free carriers available in the semiconductor.

The ionizing carriers usually gain their energy from the electric field. The high field region near the drain of an n-channel MOSFET is one such cause, leading to the generation of electron-hole pairs by electron impact ionization. The biasing of a typical n-MOSFET in normal modes of operation causes the generated electrons to be drawn into the drain

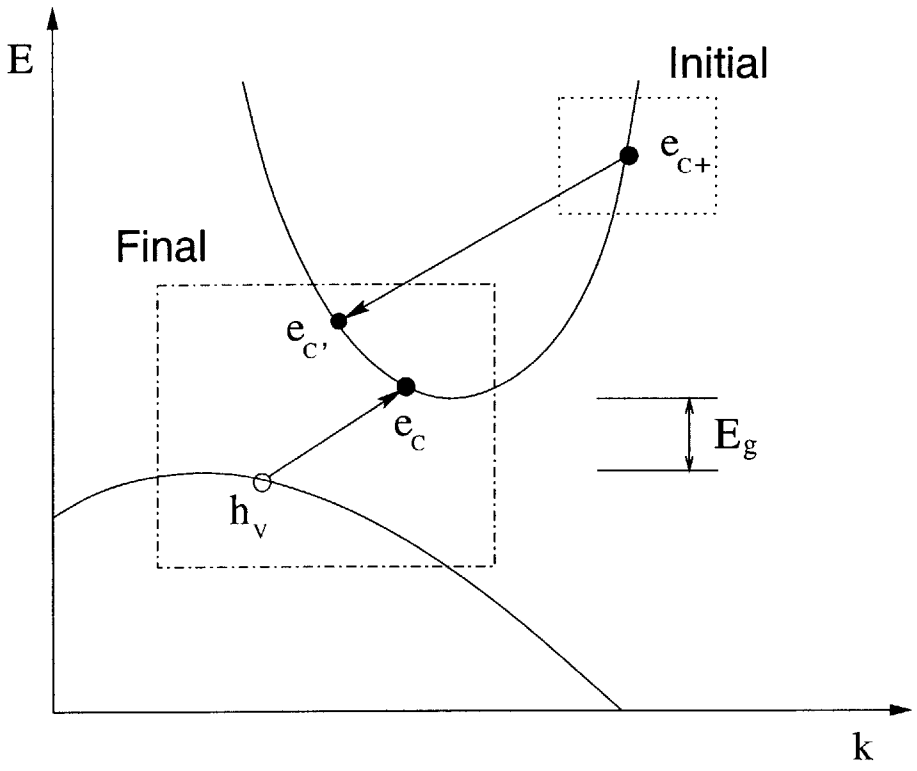

Figure 4-1: Schematic representation of the screened electron-electron interaction corresponding to impact ionization in an indirect band gap semiconductor (such as silicon). The top parabola represents the conduction band, while the bottom one is the valence band.

terminal, while the holes are collected at the substrate contact, in the form of the substrate current, as represented in Figure 1-1.

4.3 Historical Background

The first impact ionization investigations approached the phenomenon analytically, describing the probability of impact ionization as a function of the local electric field. In Shockley's lucky electron model [42] the impact ionization rate is proportional to the probability of an electron gaining the threshold energy E_{th} from the electric field $\mathcal{E}$ after traveling a distance $l = E_{th}/q\mathcal{E}$ without collision, i.e.

$$P_{ii} \sim \exp\left(-\frac{E_{th}}{q\mathcal{E}\lambda}\right) \tag{4.2}$$

where λ is the mean free path. An even earlier result belonged to Wolff [1], who investigated the high field regime above 200 kV/cm, assuming a constant electric field distribution in space (not constant in time, as is customary for device simulation). Wolff investigated impact ionization from the point of view of energy diffusion, and his approach yielded an ionization coefficient with an $\exp(-C/\mathcal{E}^2)$ dependence on the electric field $\mathcal{E}$, where C is a constant. Although physically sound, Wolff's model could not account for a lot of the data taken in small semiconductor devices. This situation was clarified by Chynoweth [43] and later by Baraff [44] who showed that for lower electric fields, when $q\mathcal{E}\lambda \leq \hbar\omega$ (the optical phonon energy), the ionization rate is better represented by a dependence like Shockley's, proportional to $\exp(-C/\mathcal{E})$.

Keldysh [45] extended the above treatments to finite temperatures. Although originally developed only for direct gap and parabolic band materials, his model has been extensively used in applications ever since. A more rigorous treatment of impact ionization should take into account the effects of band structure and Kane [2] attempted this. In his work he numerically determined the ionization rate using Fermi Golden Rule calculations, including a realistic band structure and a momentum-dependent dielectric function. More recently, Bude [46, 47] refined Kane's work by including collision broadening and intra-collisional field effects. Instead of Fermi's Golden Rule, Bude applied a quantum transport approach (e.g.

using density matrix formalism) combined with Monte Carlo methods to treat the impact ionization process. His work showed that the already "soft" threshold of the ionization process — as indicated in Kane's work and also suggested by other workers (e.g. Woods [48]) — is considerably broadened by electron-phonon collisions and the intra-collisional effect, so that a well-defined threshold does not really exist.

4.4 Device Simulation

Modeling impact ionization phenomena in semiconductors is not a trivial task. As previously mentioned, it is believed that an accurate description of the impact ionization process requires a full band structure, along with complex numerical calculations such as the Monte Carlo method [49, 50]. Unfortunately such approaches are very time-consuming and computationally intensive, thus being unsuitable for most routine device simulation work.

In order to calculate the substrate current within the context of a device simulator like FIELDAY or MEDICI it is necessary to use an accurate, physically motivated — yet computationally inexpensive impact ionization model. The standard Drift-Diffusion (DD) device simulation approach can only use a field-dependent impact ionization model because no carrier temperatures are available [51]. Therefore, the most commonly used impact ionization model for such device simulators includes only the simple exponential dependence of the ionization rate on the *local* electric field, as suggested by the work of Chynoweth, Baraff and Shockley:

$$G_{ii} \sim \exp\left(-C/\mathcal{E}\right). \tag{4.3}$$

Unfortunately, this kind of electric field dependence tends to overestimate substrate currents, especially in small devices [52]. In sub-micron MOSFET devices the impact ionization process occurs in the presence of rapidly varying electric fields and at very narrow peak field widths. Due to this large spatial variation of the electric field, carriers do not reach a steady state equilibrium with the local field. The impact ionization process is therefore considered *non-local* because it cannot be described as a function of the local field, doping and potential at a given point in the lattice alone. For the purpose of device simulation, such non-local effects must be taken into account when the typical thickness of space charge regions becomes

comparable with the carrier energy relaxation lengths [53].

To include non-locality, the impact ionization generation rate must be calculated using the local carrier temperatures (i.e. mean energies) instead of the local electric field. Such an approach is physically more plausible because the microscopic scattering mechanisms which control carrier transport depend mainly on the (microscopic) energy of the carriers [54]. For the purposes of device simulation the mean carrier energies can then be obtained through a moment expansion of the Boltzmann Transport Equation (BTE), as in the hydrodynamic approach [55, 56].

4.4.1 The Post-Processed Approach

Fully self-consistent hydrodynamic simulations are very time consuming and ill-conditioned, thus often exhibiting convergence problems in practice. Although favored over the Monte Carlo method in terms of speed, they are still somewhat unsuited for fast, practical substrate current modeling. A time-saving alternative is to post-process the temperature calculation within the device simulator. This approach has been shown to be quite successful — and it is the computation method of choice for relatively quick device simulations, such as the ones studied in this thesis. In the post-processed approach the Poisson and current continuity equations are first solved as a coupled system to produce solutions for the spatial distribution of the potential (ϕ), the carrier densities (n, p) and currents (J_n, J_p), without taking into consideration any carrier pair generation due to impact ionization:

$$\nabla^2\phi = -\frac{q}{\epsilon}(p - n + N_d - N_a) \tag{4.4}$$

$$\nabla \cdot \mathbf{J}_n = qU \tag{4.5}$$

$$\nabla \cdot \mathbf{J}_p = -qU \tag{4.6}$$

where N_d and N_a are the donor and acceptor concentrations and ϵ is the semiconductor's dielectric constant, as usual. The electron and hole currents can be written as:

$$\mathbf{J}_n = -qn\langle\mathbf{v}\rangle = q\mu_n n\nabla\left(\phi + \Delta\phi_c - \frac{k_B T_n}{q}\right) + qD_n\nabla n \tag{4.7}$$

$$\mathbf{J}_p = qp\langle\mathbf{v}\rangle = -q\mu_p p\nabla\left(\phi - \Delta\phi_v + \frac{k_B T_p}{q}\right) - qD_p\nabla p \tag{4.8}$$

where $\Delta\phi_c$ and $\Delta\phi_v$ contain any conduction or valence band energy corrections due to high doping levels or quantum effects. The carrier temperatures T_n and T_p are assumed constant and equal to the lattice temperature T_L. The carrier mobilities and diffusivities are represented by their usual symbols, $\mu_{n,p}$ and $D_{n,p}$ respectively. The net generation rate only includes Shockley-Read-Hall (SRH), surface and Auger recombination:

$$U = R_{srh} + R_{surf} + R_n^{Aug} + R_p^{Aug}. \tag{4.9}$$

Once convergence of the solution at a bias point is reached, the second moment of the Boltzmann Transport Equation (the energy balance equation) is then solved using the previously determined potential and carrier densities. The carrier temperatures are computed and therefore both electric field and temperature-dependent impact ionization rates can be included in the model. The electron-hole pairs created by impact ionization are then added to the respective terminals, as a first order correction to the main bias currents previously computed:

$$\nabla \cdot \mathbf{S}_n = \mathbf{E} \cdot \mathbf{J}_n - U w_n - n \frac{w_n - w_o}{\tau_{wn}} \tag{4.10}$$

$$\nabla \cdot \mathbf{S}_p = \mathbf{E} \cdot \mathbf{J}_p - U w_p - p \frac{w_p - w_o}{\tau_{wp}} \tag{4.11}$$

where $w_o = (3/2)k_B T_L$ is the average equilibrium energy, $w_n = (3/2)k_B T_n$ is the average electron energy and w_p is similarly defined. The electron and hole energy fluxes are

$$\mathbf{S}_n = n\langle E\mathbf{v}\rangle = -\kappa_n \nabla T_n - \left(w_n + \frac{k_B T_L}{q}\right)\mathbf{J}_n \tag{4.12}$$

$$\mathbf{S}_p = p\langle E\mathbf{v}\rangle = -\kappa_p \nabla T_p + \left(w_p + \frac{k_B T_L}{q}\right)\mathbf{J}_p \tag{4.13}$$

and the net generation rate includes carrier generation due to impact ionization:

$$U = R_{srh} + R_{surf} + R_n^{Aug} + R_p^{Aug} - G_n^{ii} - G_p^{ii}. \tag{4.14}$$

The energy transport equations are solved in the relaxation-time approximation [57] and the energy relaxation times (τ_{wn} and τ_{wp}) can themselves be modeled as energy- or doping-

dependent, or simply assumed constant. The general flow of the modeling scheme used in FIELDAY II and in this thesis is shown in Figure 4-2. The post-processed approach (option POST on the FIELDAY diagram) is generally considered to be a good compromise between speed and accuracy as long as impact ionization does not significantly affect the potential and carrier distributions — thus not representing a large portion of the drain current.

4.4.2 The Self-Consistent Approach

The fully-coupled device solution (option SELF with CTEMPN=HYDRO in FIELDAY, such that at least the electron energies are computed consistently) involves taking the temperature solution from the energy balance equations and using it to re-solve Poisson's and the continuity equations. The updated potentials, carrier densities and currents are then input again into the energy balance equations and another temperature solution is obtained. This procedure must be repeated until internal self-consistency is achieved and all output variables converge simultaneously — hence the time-consuming nature of the fully coupled approach.

It is generally considered a good approximation to assume the lattice temperature T_L constant (LTEMP=OFF) in most FIELDAY simulations — exception being ESD or other studies where device self-heating plays an important role. Similarly, for n-channel devices, when the impact ionizing carriers are electrons it is usually sufficient to compute only the electron energy equation consistently (equation 4.10), thus assuming constant hole temperatures (CTEMPN=HYDRO but CTEMPP=CONST).

4.5 Temperature-Dependent Impact Ionization Modeling

Several approaches have been used in order to include an energy-dependent impact ionization model within the framework of 3-D device simulators like FIELDAY. Recent models have described impact ionization rates that depend on the high energy tail of the electron distribution or simply on the average carrier temperature. Among these, Scrobohaci and Tang [58] have focused on the hot electron subpopulation (HES) and have shown (by comparison to Monte Carlo simulations) that the average energy of the HES is a good vari-

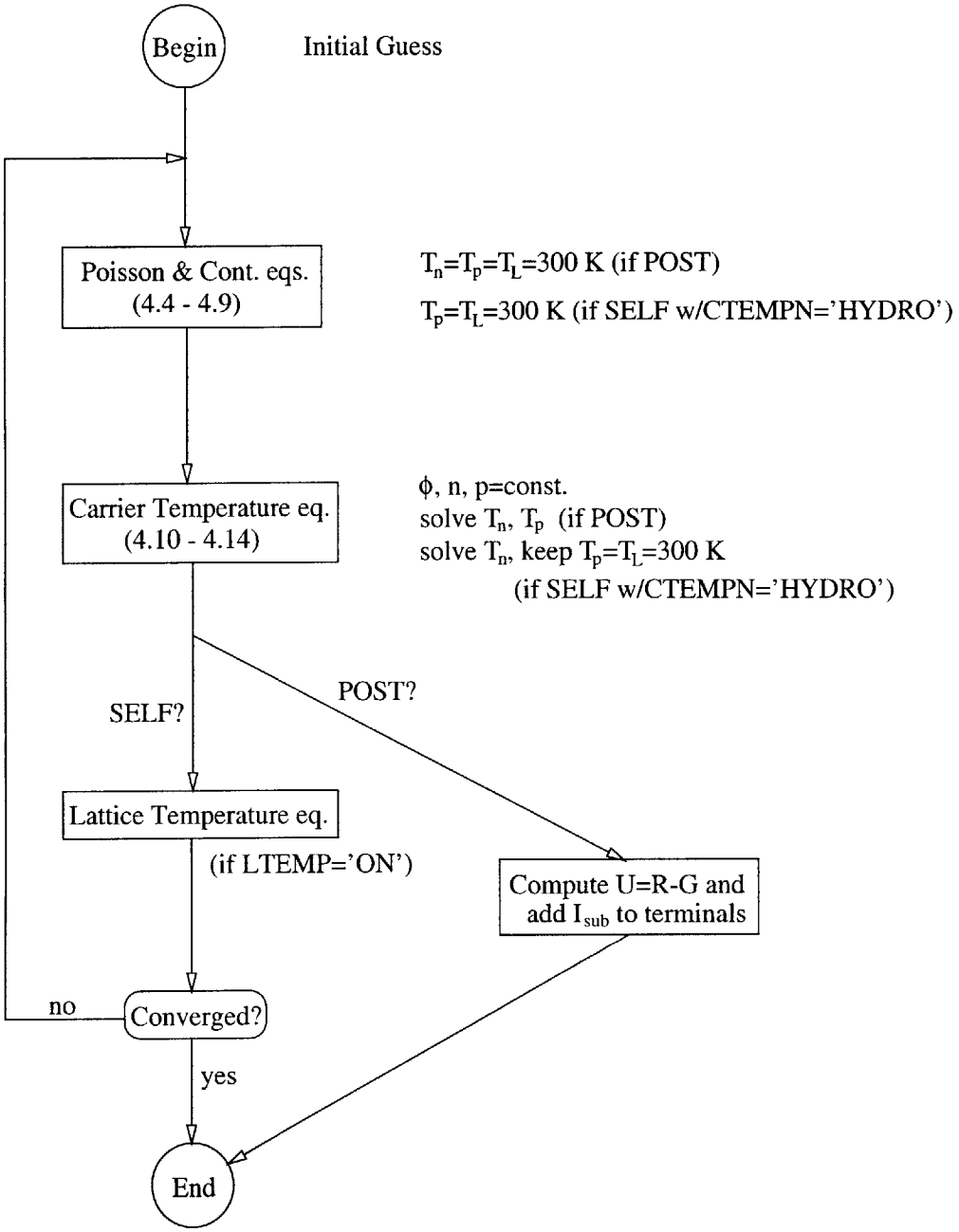

Figure 4-2: Block diagram of a device modeling scheme, such as the one in FIELDAY II.

able for the macroscopic quantification of the impact ionization rate. Unfortunately their approach is relatively complicated, because it requires a new set of modified transport equations to be derived from the Boltzmann transport equation and numerically implemented in a device simulator.

Ahn [59] and Yao [60, 61] have done similar work on a tail-electron hydrodynamic model (TEHD). For the treatment of tail electrons they have used the first four moments of the Boltzmann equation in a manner similar to that of the conventional hydrodynamic model — except they have performed integration only in the energy region $E > E_{th}$. The threshold energy for impact ionization E_{th} in silicon has been chosen somewhat arbitrarily at 1.5 eV. Their approach also requires a new set of equations to be derived and numerically implemented within a device simulator.

4.5.1 The Schöll-Quade Model

This thesis focuses on a simpler approach, attempting to modify only the impact ionization generation rate based on the average carrier temperature already available within the post-processed device solution (see section 4.4.1). This work is based on an extension to one of the most widely used temperature-dependent impact ionization models: the Schöll-Quade model [12]. Their model relates the impact ionization rate to the mean energy per carrier as computed from the second moment of the Boltzmann transport equation. Schöll and Quade use Fermi's Golden Rule to compute the transition probability per unit time for an elementary impact ionization process:

$$P_{ii} = \frac{24\pi}{\hbar}\left(\frac{4\pi q^2}{\epsilon_o}\right)^2 \frac{\delta(\mathbf{k_{c+}} + \mathbf{k_v} - \mathbf{k_c} - \mathbf{k_{c'}})\delta(E_{c+} + E_v - E_c - E_{c'})}{(|\mathbf{k_v} - \mathbf{k_c}|^2 + \kappa^2)^2} \tag{4.15}$$

where $c+, c'$ and c are conduction band states and v is the original valence band state, following the same notation as in equation 4.1 and Figure 4-1. The Kronecker δ functions represent momentum and energy conservation, respectively. Schöll and Quade's derivation assumes a screened Coulomb interaction with screening length κ^{-1} between the two involved carriers:

$$U(\mathbf{r_v}, \mathbf{r_{c+}}) = \frac{q^2 \exp(-\kappa|\mathbf{r_v} - \mathbf{r_{c+}}|)}{\epsilon_o|\mathbf{r_v} - \mathbf{r_{c+}}|} \tag{4.16}$$

in the limit of a large screening length and overlap integrals approximated by unity. Their original paper [12] also assumed a semiconductor structure with a direct band gap, parabolic conduction and flat valence bands, hence the limit of a large hole effective mass ($m_h^* \gg m_e^*$). Some of these constraints were relaxed in a more recent publication [62] where a more universal formula was derived for an indirect band gap semiconductor, starting with a relatively general band structure.

The original carrier distribution function of the Schöll and Quade model was assumed to have a spherically symmetric component f_o and a component f_1 which was odd in k-space. However it was found that only the symmetric part contributed to the impact ionization calculations, and this was chosen as a heated Maxwellian:

$$f(k) = f_o(k) = \frac{nh^3}{2(2\pi m^* k_B T)^{3/2}} \exp\left(-\frac{\hbar^2 k^2}{2m^* k_B T}\right) \tag{4.17}$$

where n is the total number of carriers, h is Planck's constant and $\hbar = h/(2\pi)$. Attention should also be paid not to confuse the wave vector k with k_B, Boltzmann's constant. Finally, m^* and T represent the carrier (i.e. electron or hole) effective mass and temperature, respectively. The carrier distribution function is normalized such that the total carrier density can be obtained by simply integrating

$$n = \frac{2}{(2\pi)^3} \int f(k) dk^3. \tag{4.18}$$

over all of k-space, where the prefactor $2/(2\pi)^3$ represents the density of states including spin. Following Schöll and Quade, the impact ionization rate per unit time and unit volume ($cm^{-3}s^{-1}$) can be calculated by integrating from the impact ionization threshold k_{th} to infinity:

$$G^{ii} = \frac{1}{(2\pi)^3} \int_{k_{th}}^{\infty} \frac{f(k)}{\tau_{ii}(k)} d^3k \tag{4.19}$$

where

$$\tau_{ii}(k) = \frac{\tau_o}{\frac{1}{2}\left(\frac{k}{k_{th}} + \frac{k_{th}}{k}\right) - 1} \tag{4.20}$$

is the isotropic impact ionization scattering time in state $|\mathbf{k}| \geq |\mathbf{k}_{th}|$, which decreases monotonically with $|\mathbf{k}|$. Evaluating the integral in equation (4.19) using the simple heated Maxwellian, Schöll and Quade's result (subscript "sq") is obtained:

$$G^{ii}_{sq}(n,T) = \frac{n}{\tau_o}\left[\sqrt{\frac{u}{\pi}}\exp(-1/u) - \mathrm{erfc}(1/\sqrt{u})\right] \tag{4.21}$$

where $u = k_B T/E_{th}$, and the complementary error function is defined as

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty \exp(-t^2)dt. \tag{4.22}$$

The carrier temperature T is related to the mean energy $\langle E\rangle$ and mean momentum $\langle p\rangle$ by

$$\langle E\rangle = \frac{\langle p\rangle^2}{2m^*} + \frac{3}{2}k_B T \tag{4.23}$$

which implies $u = 2\langle E\rangle/3E_{th}$ because the convective energy term $\langle p\rangle^2/2m^*$ can usually be neglected. The scattering time constant τ_o is on the order of femtoseconds and the ionization energy threshold E_{th} is on the order of the semiconductor band gap E_g (about 1.12 eV for silicon). They are both entered as user-definable parameters in most device simulator implementations of this model.

4.5.2 The Modified Distribution Function

Recent Monte Carlo simulations have shown that the carrier distribution cannot always be described by a simple Maxwellian, and this is especially the case in the limit of sub-micron devices and very high electric fields [63]. Specifically, when hot carriers are injected into the drain of a MOSFET they do not immediately achieve thermal equilibrium with the cold bath of majority carriers available there; rather, the carrier distribution function exhibits what has been termed a "high energy tail" [50, 64]. What is happening is that the total carrier energy distribution function is essentially a superposition of two carrier sub-populations: a "cool" one formed by the majority carriers available in the drain, and a "hot" one consisting of the high energy carriers that have just been injected from the channel and have not yet achieved thermal equilibrium with their surroundings.

The drain region of a MOSFET is also the place where most impact ionization events take place — the overwhelming majority of which are due to the high energy tail carriers. The work in this thesis seeks to account for the excess hot carrier sub-population through a simple adjustment of the Maxwellian distribution in equation 4.17. It is hoped that being able to include a high energy tail correction in the carrier impact ionization model would help achieve more accurate substrate current simulations, while keeping the model relatively simple. Specifically, the use of a single parameter $r \geq 1$ is proposed, in order to simply relate the average high energy carrier temperature with the low-energy one, such that $T_H = rT_L$. In order to obtain a distribution function that represents this mixture of hot and cool carriers, a mixing ratio has to be chosen — for example 1/2. Introducing any other parameters would undermine the simplicity of this approach. With these assumptions, the new distribution function can be written as

$$f(k,r) = \frac{nh^3}{2(2\pi m^* k_B T)^{3/2}} \left[\frac{1}{2} \exp\left(-\frac{\hbar^2 k^2}{2m^* k_B T} \right) + \frac{1}{2r^{3/2}} \exp\left(-\frac{\hbar^2 k^2}{2m^* k_B T r} \right) \right] \tag{4.24}$$

which is properly normalized, satisfying equation (4.18). The astute reader should note that the new distribution reduces to the usual Schöll-Quade Maxwellian in the limit $r = 1$, as expected. A comparison between the new distribution (with $r = 1.8$) and a simple Maxwellian at $T = 300$ K is made in Figure 4-3.

4.5.3 The Modified Impact Ionization Rate

With the modified carrier distribution function, a new impact ionization formula can be easily derived. Following the integration procedure outlined in section 4.5.1 and using the new $f(k)$ from equation (4.24), the new impact ionization coefficient with r as a parameter is arrived at:

$$\begin{aligned} G_{het}^{ii}(n,T,r) &= \frac{n}{2\tau_o} \left[\sqrt{\frac{u}{\pi}} \exp(-1/u) - \operatorname{erfc}(1/\sqrt{u}) \right] \\ &\quad + \frac{n}{2r^{3/2}\tau_o} \left[\sqrt{\frac{ru}{\pi}} \exp(-1/ru) - \operatorname{erfc}(1/\sqrt{ru}) \right] \end{aligned} \tag{4.25}$$

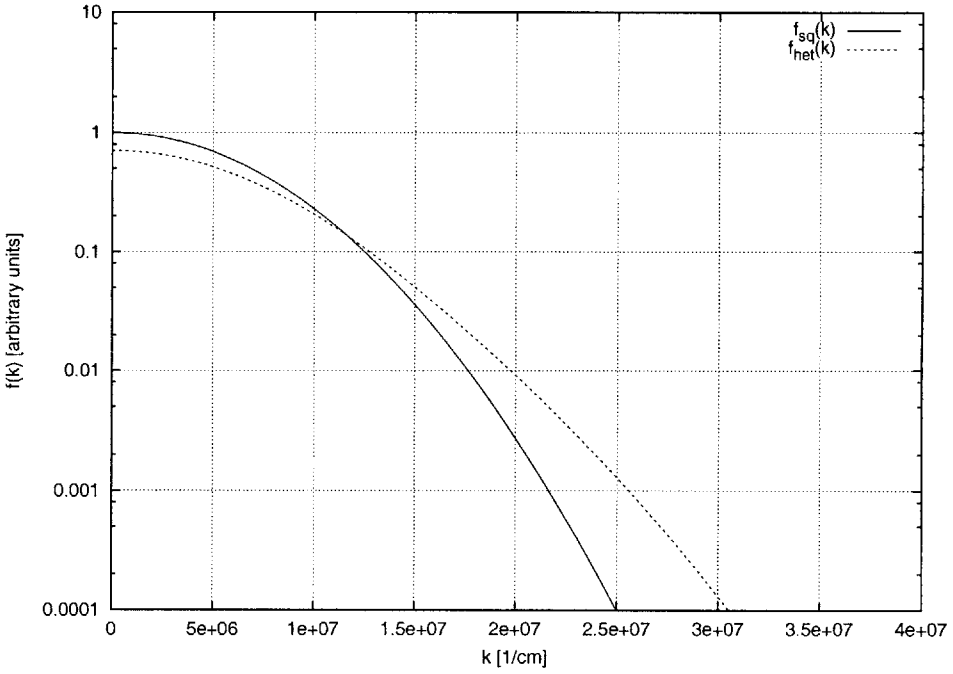

Figure 4-3: Log scale comparison between the simple Maxwellian $f_{sq}(k)$ (solid line) used in Schöll-Quade's model and the new high energy tail distribution function $f_{het}(k)$ (dotted line). The comparison is done for $T = 300$ K and $r = 1.8$.

where $u = k_B T / E_{th}$ and the subscript "het" stands for "high energy tail". Note that this new impact ionization rate simply reduces to Schöll-Quade's result from equation 4.21 in the limit of $r = 1$, as expected. A comparison between the new band to band impact ionization rate (with $r = 1.8$) and the original model is made in Figure 4-4. The presence of a high-energy tail in the electron distribution makes impact ionization more likely at average carrier energies at or around the impact ionization threshold. As a result, a "softer" impact ionization threshold is observed, as expected — in agreement with the previously mentioned Monte Carlo results, yet at a much lesser simulation time expense.

4.5.4 FIELDAY Implementation

The modified impact ionization generation rate described above was introduced within FIELDAY's post-processed substrate current simulation. The implementation was relatively easy, because only a few files needed to be modified in order to introduce the new

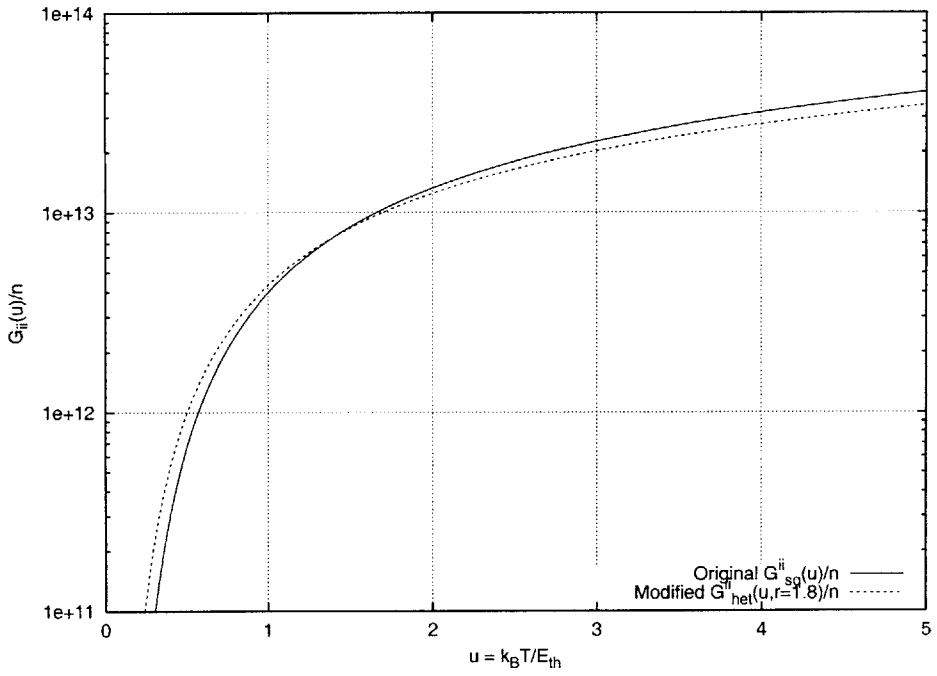

Figure 4-4: Log scale comparison between the original Schöll-Quade impact ionization rate $G^{ii}_{sq}(u)$ (solid line) and the modified high energy tail $G^{ii}_{het}(u)$ model (dotted line). The comparison is done for $T = 300$ K, $r = 1.8$ and $E_{th} = 1.12$ eV (the silicon band gap).

formulation. The parameter r was extracted and made available for the user under the name RHETN (for electrons) and RHETP (for holes) within FIELDAY's input files. The default values of both parameters are RHETN=RHETP=1, so simulations are equivalent to the old Schöll-Quade formulation if the parameters are omitted. A sample FIELDAY input file showing the inclusion of the new parameters is included in appendix C.

4.6 Substrate Current Simulations

Substrate current simulations were carried out using the newly derived impact ionization model, as implemented in FIELDAY II. The simulations were run on the inverse modeled devices described in chapter 3. The post-processed approach described in section 4.4.1 was used for all substrate current calculations. As for the drain current simulations, all FIELDAY parameters were either left at their default values or set at the values derived in

chapter 3 (e.g. RESISTOR=0.06 for the source and drain contact resistance). The MINIMOS mobility model was chosen because it adequately describes the variation of mobility as a function of depth from the surface and as a function of the electric fields.

The threshold energy for electron impact ionization was set equal to the band gap energy of silicon: ETHN=1.12 eV in FIELDAY. The model parameters for hole impact ionization were left unchanged because virtually all impact ionization events are caused by electrons in n-channel devices like the ones being investigated in this thesis. The electron impact ionization scattering time τ_o and the high energy tail parameter r introduced in the previous sections (TAUN0 and RHETN respectively in FIELDAY) were fine-tuned using the general-purpose optimization program FITDRF described in appendix B. While investigating the simulated substrate current sensitivities to both parameters as a function of gate voltage (V_{gs}) it was found that they both affected the magnitude and V_{gs} dependence of the *peak* substrate current, to a certain extent. Hence, the two parameter values were extracted simultaneously by requiring the peak simulated substrate current to fit the peak measured current for the device with the 0.6 micron gate length at a drain bias of 4 V. The 0.6 micron device was chosen for the purpose of this fit because it was the closest in length to the special test structures the inverse C-V doping extraction had been done on — whose polysilicon gate line width was 0.75 microns. The extracted values were TAUN0=12.6 femtoseconds and RHETN=1.32.

Substrate current simulations for all four investigated devices (with gate lengths of 0.5, 0.6, 1 and 5 microns) were carried out using the extracted impact ionization parameters. The results of the simulations compared with experimental data are displayed in Figure 4-5. The substrate current data was taken on the same devices that were used for the drain current measurements described in chapter 3. The data displayed in Figure 4-5 was taken with the gate bias being ramped from 0 to 5 V. The drain bias was set at 4 V while the source and substrate contacts were both grounded. All devices were 20 microns wide. For low gate voltages the substrate current increases at first, because the drain current goes up, therefore increasing the number of electrons available for impact ionization. On the other hand, as the gate voltage keeps increasing the drain saturation voltage (V_{dsat}) also goes up — and therefore the peak electric field decreases for a fixed drain bias (see equation 3.2).

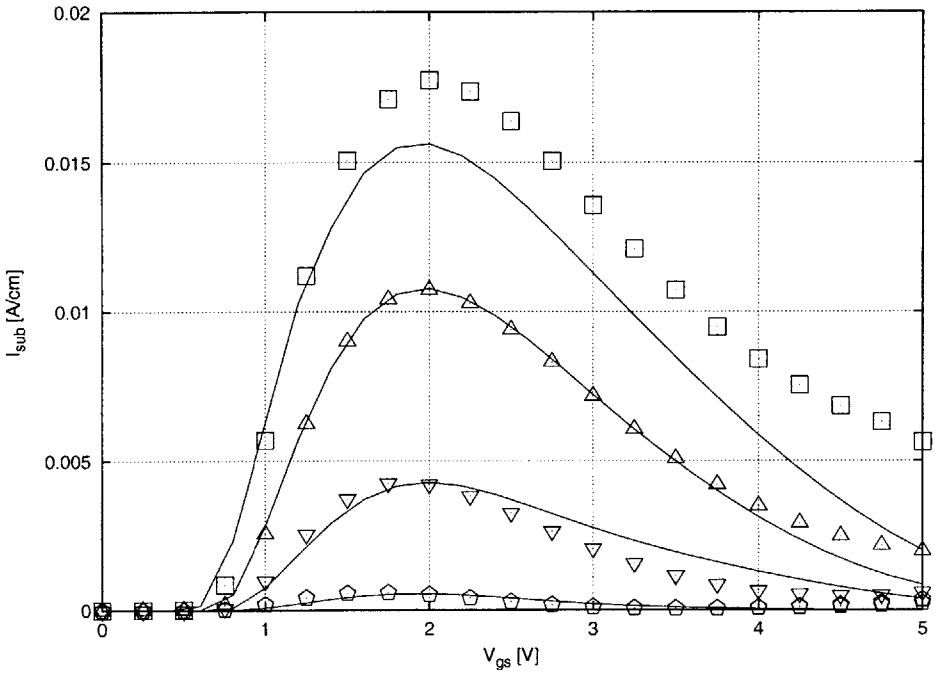

Figure 4-5: Measured (symbols) and simulated (lines) substrate currents for the four devices under investigation — with gate lengths of 0.5, 0.6, 1.0 and 5.0 microns (from top to bottom). The impact ionization parameters used were ETHN=1.12, TAUN0=1.26E-14 and RHETN=1.32. The drain bias was $V_{ds} = 4$ V.

As the peak field decreases, the channel electrons effectively begin "cooling off" near the drain, and the impact ionization rate and the substrate current decrease as well. The two competing effects (the increase in drain current versus the decrease in peak electric field with increasing V_{gs}) are approximately equal in magnitude for $V_{gs} \simeq V_{ds}/2$ and that is roughly where the substrate current reaches its maximum value.

4.7 Discussion

Figure 4-5 shows relatively good agreement between simulated and measured substrate currents across several device lengths. The roughly 10 percent discrepancy between the simulated and measured peak substrate current for the 0.5 micron device (the top curve) could be attributed to a similar discrepancy that can be seen in the simulated drain current in Figure 3-13. Because the simulated substrate currents for the other (longer) devices look

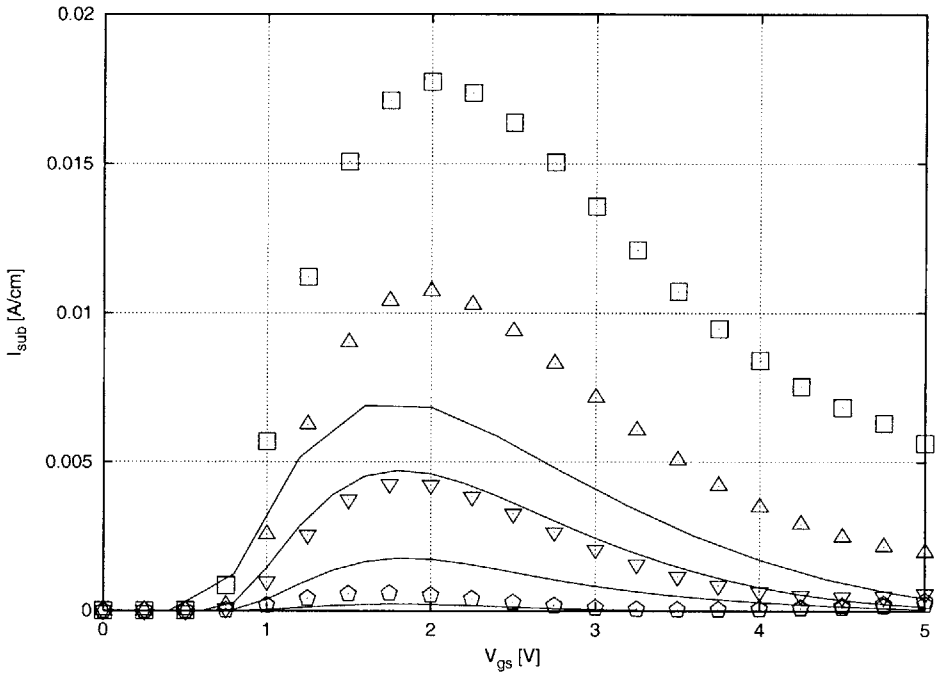

Figure 4-6: Comparison of measured (symbols) and simulated (lines) substrate currents for the same devices as in Figure 4-5. Only the high-energy tail parameter was changed to RHETN=1.0, forcing the impact ionization rates to be computed with the old (simple Maxwellian) energy distribution function.

relatively good, the discrepancy for the 0.5 micron device could be attributed more to the onset of some short-channel effects rather than to poor impact ionization modeling. For example, the channel doping was assumed to be laterally uniform in the inverse modeling procedure described in chapter 3. In real devices there could be a small increase in the channel side of the doping due to boron diffusion away from the channel/drain junction, despite the absence of a halo implant. Such a second-order doping effect would become more apparent at shorter channel lengths also because the C-V doping profiles were extracted from special structures of longer, 0.75 micron gate length. The C-V doping extraction procedure is generally believed to be more effective at inverse modeling the source or drain doping profiles, while the sub-threshold I-V technique [21] is perhaps better suited for capturing the details of the channel doping.

On the other hand, the discrepancy in the 0.5 micron device substrate current simu-

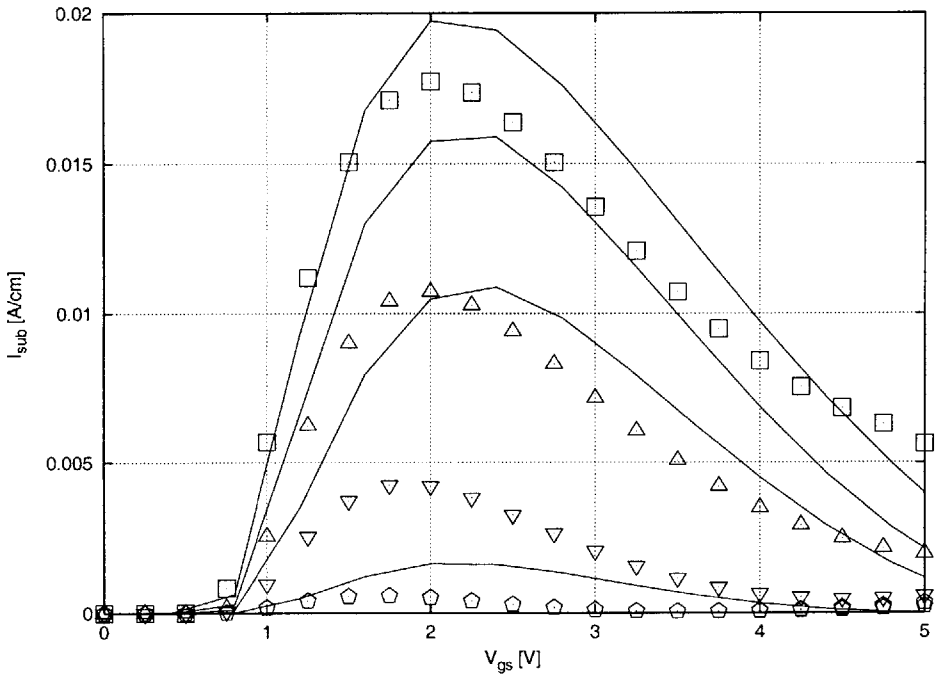

Figure 4-7: Another comparison of measured (symbols) and simulated (lines) substrate currents. The simulations above were obtained from devices with the non-optimized doping profiles used as initial "guesses" in the inverse modeling procedure described in chapter 3. RHETN=1.32 was used.

lation could also be due to approximations introduced by the post-processed method. It is well-known that the post-processed approach (see section 4.4.1) generally works as a good approximation when the substrate current represents only a small percent of the total drain current. The 0.5 micron device is the shortest among the ones being studied, and its substrate current is therefore the highest — due to higher electric fields and higher average carrier temperatures. Hence the error introduced by the post-processed approximation would be highest for the 0.5 micron device, and this may explain some of the discrepancy between the simulation and the data.

It is also relevant to explore the dependence of the simulated substrate currents on the high-energy tail parameter RHETN. Figure 4-6 shows a comparison between the same experimental data as the one presented in Figure 4-5, but the substrate current simulations were obtained with RHETN=1.0. Setting the high-energy tail parameter equal to

unity effectively forces the impact ionization rates to be computed using the old, simple Maxwellian energy distribution function. The simulated results are very different, about half the ones from Figure 4-5 where the optimized RHETN had been set equal to 1.32. This clearly highlights the importance of the high-energy tail parameter as introduced in the new model.

Finally, Figure 4-7 illustrates a comparison between the experimental data and simulations done on devices with non-optimized doping profiles. The doping profiles used as initial "guesses" in the inverse modeling procedure of chapter 3 were used. From top to bottom, the data sets and simulation lines correspond to the 0.5, 0.6, 1.0 and 5.0 micron devices. Thus the second line from the top corresponds to the data set of triangles (0.6 micron device) and the third line from the top is the simulation corresponding to the data set of inverted triangles (the 1.0 micron device). Although the parameter RHETN was set to its optimized value of 1.32, the discrepancy between the simulated and experimental substrate currents is large. By comparison with Figure 4-5, this clearly demonstrates the necessity of optimized doping profiles for substrate current simulations and the high degree of sensitivity these types of simulations have to the accuracy of the doping profiles used.

4.8 Summary

This chapter presented some of the issues confronting accurate substrate current modeling in modern device simulators. The chapter began by reviewing a few of the approaches that have been taken to model impact ionization in semiconductors, including full-band Monte Carlo methods, high-energy tail studies, carrier temperature-dependent models and even the simpler field-dependent approach. The rest of the chapter focused on temperature-dependent impact ionization calculations as implemented in the post-processed FIELDAY model.

The theory and assumptions behind an existing impact ionization model for device simulation were presented. Although the carrier energy distribution is typically represented by a heated Maxwellian, a few recent Monte Carlo results suggest that this approximation no longer holds in the limit of short devices and high biases. Instead, the carrier distribution

seems to exhibit a high energy tail component, especially in the drain of MOSFETs where most impact ionization events take place. A simple high energy tail correction was introduced into the carrier energy distribution function and the resulting impact ionization rate was implemented within the device simulator FIELDAY. Post-processed substrate current simulations were carried out, and after the calibration of a few parameters it was decided that the new impact ionization model was generally good. Further explorations showed that the substrate current simulations are highly sensitive both to the value of the high-energy tail parameter and to the accuracy of the doping profiles being used. The most reliable substrate current results were obtained with the inverse modeled device doping profiles described in chapter 3 and an optimized value of the high-energy tail parameter.

Chapter 5

Conclusions

This chapter presents a summary of this work, followed by a discussion and some suggestions for future research.

5.1 Summary

This thesis has demonstrated the implementation of a doping profiling technique based on electrical measurements and its applicability to device simulation and model calibration. Electrical measurements can be used as a non-destructive way of inversely determining a device's doping profiles, and they have become especially attractive with the increase in available computational resources in recent years. Such accurate doping profile knowledge is necessary to account for device transport effects that are highly sensitive to the electric field and doping distribution, such as impact ionization. The doping extraction technique described in this thesis was applied to the calibration of a new MOSFET device substrate current model.

The 1- and 2-dimensional doping profile extraction procedure presented in chapter 3 was done using capacitance-voltage (C-V) measurements. The procedure was treated as an inverse problem whose outputs (the device electrical characteristics) were known, but whose inputs (the device doping profiles) were to be found. The doping profiles were expressed as sums of Gaussians whose coefficients had to be solved for. FITDRF, a general optimizer based on the Levenberg-Marquardt least squares algorithm, was written for the purposes

of this thesis and used to extract the Gaussian parameters. The parameter optimization method started with an initial guess of the doping profile (provided by a process simulator such as SUPREM) and relied on the minimization of the least squares error sum of the difference between the measured and simulated electrical characteristics of the device. Besides being useful for extracting doping parameters, the program FITDRF could also be used to extract other parameters related to device simulation, such as mesh coefficients for IBM's mesh optimizer REGRID or carrier mobilities in FIELDAY, IBM's TCAD device simulator.

It was shown that the extracted doping profiles presented C-V electrical properties that were remarkably close to the ones measured on real devices. The doping profiles were also used for drain current simulations with FIELDAY. Only a single FIELDAY parameter was fine-tuned in order to achieve very good agreement with drain current data over a large range of biases and device sizes. All this served as proof that the inverse doping extraction method is reliable and could be used in a variety of ways: for predictive device simulations, for calibrating a wide range of transport model parameters, or to provide a check on the fabrication process.

Another goal of this thesis was to use the extracted device doping profiles in order to develop and calibrate a new impact ionization model for substrate current simulations. Besides being dependent on the electric field distribution inside a device, the impact ionization rate is also strongly dependent on the "hot carrier" energy distribution. Until recently, the impact ionizing carrier energy distribution was usually modeled as a heated Maxwellian. However, a few recent Monte Carlo results have suggested that in the limit of small devices and high biases, when carriers do not achieve steady state with the local electric field, the electron distribution function exhibits a high energy "tail" — especially in the drain of MOSFETs where most impact ionization events take place. Although accurate impact ionization calculations require time intensive full-band Monte Carlo methods, this thesis has sought to introduce a simple high energy tail correction at the level of a device simulator like FIELDAY. The carrier energy distribution was modeled as a superposition of two Maxwellians, one for the dominant "cool" electrons, the other as a correction to account for the high-energy ("hot") electrons. The impact ionization rate was re-derived using Fermi's Golden Rule and the new formula was parameterized and implemented in FIELDAY. Post-

processed substrate current simulations were carried out, and the few model parameters (such as the optical phonon scattering time) were calibrated by comparison with substrate current data. Despite the simplistic approach, the new model was deemed relatively good — although more investigations would be necessary to determine its applicability to a wider range of devices and biases (see the discussion below). Nevertheless, the strong dependence of the substrate current model on the high-energy tail parameter and on the accuracy of the device doping profiles used for simulation was conclusively proven.

5.2 Discussion and Suggestions for Future Work

A number of comments have to be made about the validity of the inverse doping extraction method presented in this thesis. First, it is important to keep in mind that the conventional macroscopic model for semiconductor devices was used. Specifically, the device doping profiles were assumed to be smooth and continuous functions of their spatial coordinates — such that they could be modeled by sums of Gaussians. This assumption breaks down in the limit of very small devices. For example, assume a uniform doping level of 10^{17} atoms per cubic centimeter. At this doping level, an imaginary cubic device with a side of 0.05 μm would contain only about a dozen doping atoms. Such a situation can hardly be described by a continuous doping profile. Therefore, in the future, more statistically based models (such as the ones used for Monte Carlo simulations) may be required.

Another assumption was made about the C-V measurement data. The C-V data was taken on large perimeter or large area specially designed test structures, in order to improve the signal-to-noise ratio of the measurements and to isolate a particular capacitance component, as explained in sections 3.1.4 and 3.3.2. When the extracted doping profiles were applied to model a MOSFET device, it was implicitly assumed that the MOSFET doping profiles would be the same as those of the special test structures built very near it on the same wafer cell, using the same process steps. Such assumptions are routinely made when testing devices for process characterization in the semiconductor industry. However when accurate measurements are needed for the inverse modeling of doping profiles, special care must be taken. Unfortunately it is hard to gauge what uncertainty this kind of

assumption would introduce. The data for the C-V measurements described in chapter 3 was taken across three different cells on the same wafer, in order to get an idea of such process variations. However that data only described the process variations between the C-V measurement structures across *different* cells, and still did not yield any direct information about the doping variation between test structures and MOSFETs on the *same* cell. Fortunately, it can be said that any process variations between C-V test structures across *different* cells did not produce variations larger than 0.8 % in the measured junction capacitance or larger than 2 % for the gate-to-source capacitance. Nevertheless, a study of the uncertainty introduced by using special test structures (as opposed to actual MOSFETs) for C-V measurements should be carried out in the future.

Although only the C-V inverse modeling technique was explored in this thesis, it may be useful to combine it with the sub-threshold I-V technique [20]. One advantage of the I-V technique is that it can be used *in situ*, by taking direct measurements on MOSFET devices. Moreover, it is believed that the I-V technique has better sensitivity in the channel region (including halo) while the C-V method has better sensitivity in the source/drain regions. Matching sub-threshold I-V data and C-V data simultaneously may improve the accuracy of the results and should be explored in the future. Moreover, as direct techniques for measuring 2-dimensional doping profiles may become available, these inverse modeling techniques should be checked against them.

From a computing perspective, the C-V inverse modeling simulations are quite time-intensive. In this work care was taken to avoid numerical instabilities by making "good" initial guesses and choosing to optimize at most two or three doping profile parameters at a time. In the future, more work could be done to optimize both the FIELDAY simulations, as well as the FITDRF parameter extraction program. Of course, the simulation time could be also cut down as faster workstations are becoming available.

Finally, care must also be taken in evaluating the impact ionization model introduced in chapter 4. It is well-known that accurate impact ionization calculations require the help of a full-band Monte Carlo simulator. The work in this thesis is a simple attempt to make a physically-based adjustment to the analytical carrier temperature-dependent approach taken in the context of a TCAD simulator like FIELDAY. As such, it does not have the

pretense of being a universal impact ionization model. Although relatively good agreement between simulated substrate currents and data was seen as a function of gate voltage and device size (for a given drain bias) — problems were noted in terms of scaling with the drain bias (for a given gate bias). Specifically, simulated substrate currents matching the data at a 4 V drain bias were under-estimating the data by close to 50 % for a 5 V drain bias. Similarly, the data was over-estimated by nearly 50 % when going to a 3 V drain bias. This may suggest that instead of being a constant, the high-tail parameter r (RHETN for electrons as introduced in FIELDAY) may be a function of the drain voltage V_{ds}. Some recent experimental work [65] based on optical spectrum measurements of hot carriers seems to suggest that there is a roughly linear relationship between the average channel hot carrier temperature and the applied V_{ds}. Future work should investigate these claims, perhaps via Monte Carlo calculations of the hot carrier energy distribution as a function of V_{ds} at the point of maximum impact ionization in the drain of a MOSFET.

Appendix A

The Levenberg-Marquardt Algorithm

All inverse modeling work done in this thesis relies on fitting simulated results to experimental data by modifying model parameters. This is accomplished with the Levenberg-Marquardt nonlinear optimization algorithm, which finds values of model parameters such that the mean square error between simulated and experimental data is minimized:

$$\xi(\mathbf{p}) = \frac{1}{N}\sum_{i=1}^{N} w_i^2 \left[y_i^{data} - y_i^{sim}(\mathbf{p})\right]^2 = \sum_{i=1}^{N} h_i^2(\mathbf{p}) = ||\mathbf{h}(\mathbf{p})||^2 \tag{A.1}$$

where N is the total number of points, $\mathbf{p}$ is the vector of model parameters, y_i^{data} is the value of the i-th experimental data point, y_i^{sim} is the i-th simulated data point, the w_i's represent weights that can be assigned to each term and $\mathbf{h(p)}$ is the error vector. In general, mean square error minimization is achieved by starting with an initial estimate of the parameter vector $\mathbf{p}^o$ and iteratively updating it by taking a sequence of steps in error space

$$\mathbf{p}^{k+1} = \mathbf{p}^k + \delta\mathbf{p}^k \tag{A.2}$$

such that the new mean square error $\xi(\mathbf{p}^{k+1})$ is minimal along the chosen search direction. The vector $\delta\mathbf{p}^k$ represents the incremental update vector at each iteration k.

In the vicinity of a minimum in parameter space, $\xi(\mathbf{p})$ can be approximated by a second

order multi-dimensional Taylor expansion around $\mathbf{p} = \mathbf{p}^k$:

$$\xi(\mathbf{p}) - \xi(\mathbf{p}^k) = \mathbf{g}(\mathbf{p}) \cdot (\mathbf{p} - \mathbf{p}^k) + \frac{1}{2}(\mathbf{p} - \mathbf{p}^k)^T H(\mathbf{p})(\mathbf{p} - \mathbf{p}^k) \tag{A.3}$$

where $\mathbf{g} = \nabla\xi$ is the gradient vector according to $g_j = \partial\xi/\partial p_j$ and H is the symmetric Hessian matrix of second derivatives given by

$$H_{ij} = \frac{\partial^2 \xi}{\partial p_i \partial p_j}. \tag{A.4}$$

The gradient can also be written as $\mathbf{g} = 2J^T\mathbf{h}$ where J is the Jacobian matrix given by $J_{ij} = \partial h_i(\mathbf{p})/\partial p_j$. Because the Hessian matrix H involves second derivatives which are computationally expensive, it is usually approximated using only first derivatives in the Gauss-Newton approach:

$$H \simeq 2J^T J. \tag{A.5}$$

The parameter vector $\mathbf{p}$ must be found in order minimize the error sum $\xi(\mathbf{p})$. For each iteration step the incremental parameter update vector can be obtained by solving the linear system of equations:

$$H(\mathbf{p}^k)\,\delta\mathbf{p}^k = -2J^T\mathbf{h}(\mathbf{p}). \tag{A.6}$$

The Levenberg-Marquardt method was introduced to regularize Newton's method because in practice the Hessian matrix often tends to be near-singular. The scalar parameter λ is added:

$$\left[H(\mathbf{p}^k) + \lambda D^k\right]\delta\mathbf{p}^k = -2J^T\mathbf{h}(\mathbf{p}) \tag{A.7}$$

where D is a diagonal matrix with $D_{ii} = H_{ii}$ and λ must be chosen such that $H + \lambda D$ is no longer near-singular. The Levenberg-Marquardt method reduces to the Gauss-Newton method for $\lambda \to 0$ and to the method of steepest descent for $\lambda \to \infty$. The Levenberg-Marquardt algorithm is summarized below, with user-definable parameters in boldface:

1. start with λ=**lambda_init** and an initial estimate for $\mathbf{p}^o$

2. solve the system in equation A.7 and let $\mathbf{p}^{k+1} = \mathbf{p}^k + \textbf{damping\_factor} * \delta\mathbf{p}^k$

3. if $\xi(\mathbf{p}^{k+1}) < \xi(\mathbf{p}^k)$ then

   $\lambda = \lambda/$**lambda_scale**

   else

   $\lambda = \lambda *$ **lambda_fail**

4. if not converged go to step 2.

The values of the parameters used in FITDRF (see appendix B) at compile-time are shown in Table A.1. Finally, to determine whether the iterative procedure has found a minimum,

Parameters	Value
lambda_init	0.01
damping_factor	1.0
lambda_scale	8.0
lambda_fail	10.0

Table A.1: A few default parameters built into the FITDRF optimizer.

or if the iteration should be stopped for other reasons, several convergence criteria can be applied in practice:

- if the value of the error sum $\xi(\mathbf{p}^k)$ is small, the algorithm may become limited by machine accuracy and should be stopped. The iteration may be stopped even sooner if the desired accuracy is reached.
- if the relative *change* in the value of $\xi(\mathbf{p}^k)$ is small, the iteration should be stopped.
- if the value of the parameter λ becomes greater than some λ_{max}, then the algorithm is likely not to be able to find further improvement and the method fails.
- if λ has decreased below some minimal λ_{min} then the algorithm is likely to be wandering at the bottom of some valley in parameter space and it must be stopped.

It should also be noted that in practice the iteration should generally *not* be stopped on a step where $\xi(\mathbf{p}^k)$ increases: that only shows that λ has not yet adjusted itself optimally [26].

A more thorough review of optimization techniques for inverse doping modeling has been done by Ouwerling [18]. Many other optimization algorithms are introduced and analyzed in D. Bertsekas' nonlinear programming textbook [66].

Appendix B

The FITDRF Optimizer

This appendix represents a modified and formatted version of the README file associated with FITDRF.

B.1 Purpose

FITDRF is a general-purpose optimizer that attempts to extract DOPING, REGRID or FIELDAY (hence the suffix "D-R-F") input parameters such that the mean square error between the simulated FIELDAY output and a user-provided data file is minimized. For this purpose FITDRF uses the least-squares Levenberg-Marquardt algorithm described in appendix A. FITDRF is written in C, but it also relies on a few GAWK calls. The user-provided data file can contain either capacitance-voltage (C-V) or current-voltage (I-V) measurements.

B.2 Usage

FITDRF can be invoked from the UNIX (AIX) command-line by typing:

```
> fitdrf fit.in .td -d/r/f
```

where `fit.in` is the program's input file and `.td` is the IBM Tdatabase that contains the simulation mesh all the runs will be performed on. The `.td` database must contain at least its original `raw` record (see appendix C). One of the `-d`, `-r` or `-f` switches must be supplied as

well, to indicate whether the parameter adjustments will be done in the DOPING, REGRID or FIELDAY input files, respectively.

B.3 The Input File

The program's input file (e.g. `fit.in`) must have a certain format. An example is provided in appendix C. As with the user-supplied data file, any lines preceded by either a `#`, `*`, `$` or `%` are considered comments and ignored. Any white spaces preceding an input line are also ignored. Each input statement must occupy an input line by itself. There must be no white spaces around the equal (=) sign. The statements

```
tdelete.path=...
doping.path=...
regrid.path=...
fielday.path=...
```

must contain the AFS paths for the four respective programs. If they are not supplied by the user, they will default to the system-dependent (local) program paths.

The user must also provide paths to the input files that need to be used during runtime with DOPING, REGRID or FIELDAY. If no such input files are supplied, FITDRF will look for files named `doping.in`, `regrid.in` and `fielday.in` in the local directory. An error will occur if any of these files cannot be found when needed by their respective programs.

The device and main record name inside the Tdatabase must be supplied as well. For example `devname=ngate` or `devname=nfet1@L=0.6`.

The experimental data file which the FIELDAY output will be compared to must also be given, for example `datafile=cap.dat`. This file must be in tab or space-separated vertical two-column format, with the first column being the independent variable (e.g. voltage in Volts) and the second one being either the experimental current or capacitance. The user should make sure that the units of the data are the same as the expected FIELDAY output units: for example fF/μm for overlap capacitance data, fF/μm^2 for junction capacitance data, A/cm for current data if FIELDAY III is used and mA/μm for current data if FIELDAY II is used.

The user must also specify which simulation file the FIELDAY output will be found in, generally either `acss2.data` for capacitance simulations or `results.summary` for current simulations. For example

```
fielday.out=acss2.data
```

which is also the default. Once it knows the name of the FIELDAY output file, FITDRF still needs to know which column to read the simulation results from. This is generally column 7 for gate-to-source capacitance and substrate current simulations in the `acss2.data` and `results.summary` files respectively, and column 9 for drain current simulation results in the file `results.summary`. The default is `ycolumn=7`.

Finally, FITDRF must be given initial values for the parameters it will attempt to extract, one value per input line. For example:

```
param1=3.1E-5
param2=-9.9
foobar=0.00012
```

These parameter names must be present inside the DOPING, REGRID or FIELDAY input files, wherever their numerical values are to be inserted during the runs (by a GAWK call). The parameter names are case-sensitive and can be called anything, with the exception of variable names reserved for use by DOPING, REGRID or FIELDAY. In other words, parameters should *not* be named anything like NAMES, TARGET, DPG, CON, SIGX, OUTPUT, END, REFINE, X0, CHARGE, MOBLTY or ACCAP.

B.4 Other Input Files

The DOPING, REGRID and FIELDAY input files (whichever are necessary to complete a particular run) must be provided by the user and must have their &NAMES SOURCE and TARGET set to point to the correct Tdatabase. The parameter names (`param1`, `param2`, etc.) must be placed *instead* of numerical values for the coefficients that are to be extracted inside the respective input files. For example, a statement like `SIGX=3.1E-05` inside a DOPING input file should be replaced by `SIGX=param1` if `param1` is to be extracted. Moreover

note that FITDRF cannot fit the parameters of two different programs at the same time: for instance trying to extract both DOPING and REGRID input parameters during the same run will not work.

Finally, for any parameter extraction, the FIELDAY input file must always have the word `vramp` substituted at the contact where the voltage will be varied, such as:

```
&CONTAC NUMBER=3, VO=vramp, &END
```

Note that FITDRF always considers voltage to be the independent variable.

B.5 Program Output

FITDRF sends most of its output to `stdout` and `stderr`. In addition, the most recent DOPING, REGRID and FIELDAY runs each store their outputs in their own directories, named `dopdir#`, `regdir#` and `fdirtemp#`, where the '`#`' sign represents the run number.

The output results of the Levenberg-Marquardt algorithm from FITDRF are stored in a file called `results`. The user can inspect how this file is periodically (but very slowly) updated by FITDRF by issuing a command like

```
> tail -f results
```

at the UNIX (AIX) prompt.

B.6 Timing and Speed Issues

Extracting parameters with FITDRF is not a quick operation, even if only one-dimensional junction capacitance calculations are needed. As a general rule of thumb, fitting FIELDAY parameters (e.g. fitting with the switch `-f`) should be faster because no REGRID or DOPING runs are made. Unfortunately this isn't always the case, especially when time-intensive substrate current FIELDAY parameter extractions are performed.

Each DOPING run takes on average less than a minute, while REGRID runs may take as much as 5 minutes, depending on the size of the mesh and the speed of the machine they are being done on. FIELDAY runs can average anywhere between a couple of minutes

(for a few capacitance points) to a couple of hours (for a few substrate current points). Moreover, FITDRF invokes two FIELDAY runs for each parameter that is extracted, for every iteration of the Levenberg-Marquardt algorithm. This is because the Levenberg-Marquardt algorithm needs to perform numerical derivatives for *all* simulated output points, as *each* input parameter is slightly perturbed.

In order to save time and avoid convergence problems, it is highly recommended that no more than three independent parameters be extracted at once. It is also recommended that for each N parameters to be extracted, at least 3N experimental data points be used — but not many more, due to the increased FIELDAY run-times.

B.7 Other Technical Issues

A number of parameters used internally by FITDRF's Levenberg-Marquardt algorithm or for the numerical derivatives are introduced with `#define` statements inside the C source code. A few of these parameters were shown in Table A.1 of appendix A. If necessary, these parameters can be modified and everything should then be re-compiled with:

```
> gcc -lm fitdrf.c
```

The FITDRF source code can be obtained by sending electronic mail to the author at `epop@alum.mit.edu`.

Appendix C

Sample Input Files

This appendix contains some sample input files used with the programs DOPING, REGRID, FIELDAY and FITDRF for the work done in this thesis. With the exception of the FITDRF input file, all other files are FORTRAN name-list type decks, with each card (e.g. &NAMES) starting at the second column of the file. Hence all lines that *do not* start with at least one space are considered comments and ignored.

C.1 DOPING Input File

A DOPING input file similar to the one used to generate the 2-dimensional Gaussian doping profiles for the inverse modeled 0.6 micron device is shown first:

```
 &NAMES SOURCE='.td::nfetl@L=0.6/raw',
        TARGET='.td::nfetl@L=0.6/dop', &END
$ source/drain part I:
 &DPG NSHAPE=1, CON=6.0E19, XLOC=0.0, YLOC=0.997E-4, ZLOC=0.0, NDIR=0,
                       XLEN=8.8776E-5, YLEN=0.0E-4, ZLEN=0.0E-4,
                       SIGX=3.4253e-06, SIGY=0.032E-4, SIGZ=0.0E-4, &END
$ source/drain part II:
 &DPG NSHAPE=1, CON=3.9E19, XLOC=0.0, YLOC=0.937E-4, ZLOC=0.0, NDIR=0,
                       XLEN=0.850E-4, YLEN=0.0E-4, ZLEN=0.0E-4,
                       SIGX=0.029E-4, SIGY=0.041E-4, SIGZ=0.0E-4, &END
$ LDD doping profile I:
 &DPG NSHAPE=1, CON=5.0E18, XLOC=0.0, YLOC=0.96E-4, ZLOC=0.0, NDIR=0,
                       XLEN=8.908E-05, YLEN=0.0E-4, ZLEN=0.0E-4,
                       SIGX=3.166E-06, SIGY=0.06E-4, SIGZ=0.0E-4, &END
```

```
$ LDD doping profile part II -- under gate, into channel:
 &DPG NSHAPE=1, CON=3.3E18, XLOC=0.0, YLOC=1.0E-4, ZLOC=0.0, NDIR=0,
                        XLEN=8.913e-05, YLEN=0.0E-4, ZLEN=0.0E-4,
                        SIGX=7.3404e-06, SIGY=0.10E-4, SIGZ=0.0E-4, &END
$ constant n-type background:
 &DPG NSHAPE=3, CON=4.0E6, XLOC=0, YLOC=1.0E-4, ZLOC=0.0,
                        XLEN=2.0E-4, SIGY=1.0E-4, ZLEN=0.0E-4, NDIR=-1, &END
$ channel doping implants:
 &DPG NSHAPE=1, CON=-1.3985e+17, XLOC=0, YLOC=1.0E-4, ZLOC=0.0, NDIR=0,
                        XLEN=2.0E-4, YLEN=0.0E-4, ZLEN=0.0E-4,
                        SIGX=0.0E-4, SIGY=0.20E-4, SIGZ=0.0E-4, &END
 &DPG NSHAPE=1, CON=-8.095E16, XLOC=0, YLOC=0.65E-4, ZLOC=0.0, NDIR=0,
                        XLEN=2.0E-4, YLEN=0.0E-4, ZLEN=0.0E-4,
                        SIGX=0.0E-4, SIGY=0.18E-4, SIGZ=0.0E-4, &END
 &DPG NSHAPE=1, CON=-1.616E17, XLOC=0, YLOC=0.5186E-4, ZLOC=0.0, NDIR=0,
                        XLEN=2.0E-4, YLEN=0.0E-4, ZLEN=0.0E-4,
                        SIGX=0.0E-4, SIGY=0.0847E-4, SIGZ=0.0E-4, &END
$ constant p-type background:
 &DPG NSHAPE=3, CON=-2.3E16, XLOC=0, YLOC=1.0E-4, ZLOC=0.0, NDIR=-1,
                        XLEN=2.0E-4, SIGY=1.0E-4, ZLEN=0.0E-4, &END
```

C.2 REGRID Input File

A sample REGRID input file used to convert the doped half-device mesh into a symmetrical full-device ready for 2-dimensional current simulations is shown below. An optimized half-device mesh can be obtained if the &MIRROR card is ommited and the TARGET record is instead named `regrid` (as opposed to `reflect`). Such an optimized half-device mesh is displayed in Figure 3-7 and can be used for gate-to-source capacitance simulations.

```
 &OUTPUT MSHWRT=1, DOPWRT=1, CNTWRT=1, GRMWRT=0, STRWRT=0,
                   POLYDOP=7.2E19, &END
 &NAMES SOURCE='.td::nfetl@L=0.6/dop',
        TARGET='.td::nfetl@L=0.6/reflect', &END
 &OPTIONS METAL_CNT='ON',
          SUBS_CNT='ON', &END
 &MIRROR MIRROX='R',
         CHOPX='R', &END
 &REFINE BLKGRD = 0.1000, INTTHK = 0.0150,
         INTFSP = 0.0015, THICKOX= 0.0130,
         JNCTSP= 0.0040, MAXGRD=1.5, &END
```

C.3 FIELDAY Input File

A FIELDAY II input file used for current simulations is displayed below. The file includes the parameter r (RHETN for electrons, in input card &GENR) as described in Chapter 5. Note that the card-ending statements &END and the forward slash (/) are equivalent.

```
&NAMES L_MASK=0.6,
       SOURCE='.td::nfetl@L=0.6/reflect',
       TARGET='.td::nfetl@L=0.6/reflect//IxVgr', &END
&ALLOC NCONEC=6, MAXRAM=80000000, /
&PHYSIX CSTATS='FDIRAC', DCBGAP='DA85', DVBGAP='DA85', QMCORR='ON',
        QMDEVICE='NFET', /
&GENR IMPACT='QADE', METHOD='POST', TAUNO=1.26E-14, ETHN=1.12, RHETN=1.32,
        PCONTACT=2, NCONTACT=4, /
&MOBLTY MOBN=16, MOBP=16, E_P_CALC='QFL', M_SD='MIN', /
&CHARGE TYPE='SURFACE', QNODE=3.0D10,
        POINT=0, 0.99D-4, 0,
              0, 1.00D-4, 0,
        NORMAL=0,1,0, 0,-1,0, /
&PROPS TEMP=296, /
&GEOMTR NGCCSO='ON', CCSSML=1E-8, /
&RECOM SRH='ON', AUGER='ON', SURFAC='ON', HURKX='ON', /
&SOLVER SOLTYP='DRCT', ORDR='MD', NTNIT=100, DTEPS=1.0E-3,
        NNWTIT=100, NGUMIT=100, DVEPS=1.0E-3, /
&CONTAC NUMBER=1, VO=0.0, DV=0.2,0.2,0.2,0.2,0.2,0.2,0.2,
                           0.2,0.2,0.2,0.2,0.2,0.2,0.2,0.2,0.2,0.2,
                           0.2,0.2,0.2,0.2,0.2,0.2,0.2,0.2, /
&CONTAC NUMBER=2, VO=0.0, /
&CONTAC NUMBER=3, VO=0.0, RESISTOR=0.06, /
&CONTAC NUMBER=4, VO=4.0, RESISTOR=0.06, /
```

C.4 FITDRF Input File

Finally, a typical FITDRF input file (`fit.in`) is included below:

```
# program paths
regrid.path=/afs/btv/data/vats/ef/bin/regrid212
fielday.path=/afs/btv/u/epop/bin/fielday
# fielday.path=/afs/btv.ibm.com/data/vats/ef/bin/fday304
doping.path=/afs/btv/data/vats/ef/bin/doping

# other programs' input files
```

```
doping.in=doping.in
regrid.in=regrid.in
fielday.in=fielday.in

# device name
devname=nfet1@L=0.75

# experimental data file
datafile=cov.dat

# which fielday file to find the output in
fielday.out=acss2.data

# which output column to read (i.e. what kind of data to fit):
#    Cov -- "acss2.data" column 7
#    Isx -- "results.summary" column 7
#    Ids -- "results.summary" column 9
ycolumn=7

# parameters to extract
param1=3.427E-06
param2=7.52E-6
```

More details about the sample input files included in this appendix can be found in the DOPING, REGRID and FIELDAY manuals [36, 37, 38] or in appendix B for FITDRF.

Bibliography

[1] P.A. Wolff. Theory of Electron Multiplication in Silicon and Germanium. *Physical Review*, 95(6):1415–1420, 1954.

[2] E.O. Kane. Electron Scattering by Pair Production in Silicon. *Physical Review*, 159(3):624–631, 1967.

[3] T. Kamata, K. Tanabashi, and K. Kobayashi. Substrate Current due to Impact Ionization in MOSFET. *Japanese Journal of Applied Physics*, 15(6):1127–1133, 1976.

[4] P.K. Chatterjee, W.R. Hunter, T.C. Holloway, and Y.T. Lin. The Impact of Scaling Laws on the Choice of n-Channel or p-Channel for MOS VLSI. *IEEE Electron Device Letters*, EDL-1(10):220–223, October 1980.

[5] S. Tam, P.-K. Ko, and C. Hu. Lucky-Electron Model of Channel Hot-Electron Injection in MOSFET's. *IEEE Transactions on Electron Devices*, ED-31(9):1116–1125, 1984.

[6] Y.W. Sing and B. Sudlow. Modeling and VLSI Design Constraints of Substrate Current. *IEDM Tech. Dig.*, page 732, 1980.

[7] J. Matsunaga et al. Characterization of Two Step Impact Ionization and Its Influence on NMOS and PMOS VLSI's. *IEDM Tech. Dig.*, page 736, 1980.

[8] S. Tam and C. Hu. Hot Electron Induced Photon and Photo-Carrier Generation in Silicon MOSFET's. *IEEE Transactions on Electron Devices*, pages 1264–1273, 1984.

[9] D. Cole, E. Buturla, S. Furkay, K. Varahramyan, J. Slinkman, J. Mandelman, D.P. Foty, O. Bula, A. Strong, J. Park, T. Linton, J. Johnson, M. Fischetti, S. Laux, P. Cot-

trell, H. Lustic, F. Pileggi, and D. Katcoff. The Use of Simulation in Semiconductor Technology Development. *Solid-State Electronics*, 33(6):591–623, 1990.

[10] D.A. Antoniadis, S.E. Hansen, and R.W. Dutton. SUPREM II – a Program for IC Process Modeling and Simulation. Technical Report 5019-2, Stanford University, 1978.

[11] E.M. Buturla, J.B. Johnson, S. Furkay, and P.E. Cottrell. A New 3D Device Simulation Formulation. *Proc. NASECODE VI, Dublin, Ireland*, pages 291–295, 1989.

[12] E. Schöll and W. Quade. Effect of Impact Ionization on Hot-Carrier Energy and Momentum Relaxation in Semiconductors. *Journal of Physics C, Vol. 20*, 1987.

[13] R.B. Marcus. *VLSI Technology*, chapter 12, Diagnostic Techniques. McGraw-Hill, 1983.

[14] W. Vandervorst and T. Clarysse. Recent Developments in the Interpretation of Spreading Resistance Profiles for VLSI Technology. *J. Electrochem. Soc.*, 137:679–683, 1990.

[15] J. Kim, J.S. McMurray, C.C. Williams, and J. Slinkman. Two-Step Dopand Diffusion Study Performed in Two Dimensions by Scanning Capacitance Microscopy and TSUPREM IV. *Journal of Applied Physics*, 84(3):1305–1309, 1998.

[16] V.V.Zavyalov, J.S. McMurray, and C.C. Williams. Advances in Experimental Technique for Quantitative Two-dimensional Dopant Profiling by Scanning Capacitance Microscopy. *Review of Scientific Instruments*, 70(1):158–164, 1999.

[17] J. Hilibrand and R.D. Gold. Determination of the Impurity Distribution in Junction Diodes from Capacitance-Voltage Measurements. *RCA Review*, 21:25–252, 1960.

[18] G.J.L Ouwerling. *Nondestructive One- and Two-Dimensional Doping Profiling by Inverse Methods.* PhD thesis, Delft University of Technology, 1989.

[19] N. Khalil, J. Faricelli, C.-L. Huang, and S. Selberherr. Two-Dimensional Dopant Profiling of Submicron Metal-Oxide-Semiconductor Field-Effect Transistor Using Nonlinear Least Squares Inverse Modeling. *J. Vac. Sci. Technology B 14(1)*, 1996.

[20] Zachary K. Lee. *A New Inverse-Modeling-Based Technique for Sub-100-nm MOSFET Characterization.* PhD thesis, Massachusetts Institute of Technology, 1998.

[21] Z.K. Lee, M.B. McIlrath, and D.A. Antoniadis. Inverse Modeling of MOSFETs using I-V Characteristics in the Subthreshold Region. *IEDM Tech. Dig.*, 1997.

[22] Narain Arora. *MOSFET Models for VLSI Circuit Simulation.* Springer-Verlag, 1993.

[23] T.Y. Chan, P.K. Ko, and C. Hu. Dependence of Channel Electric Field in Device Scaling. *IEEE Electron Device Letters*, EDL-6:551–553, 1985.

[24] J. Chung, M.C. Jeng, G. May, P.K. Ko, and C. Hu. Hot-Electron Currents in Deep-Submicrometer MOSFETs. *IEDM Tech. Dig.*, pages 200–203, 1988.

[25] A. S. Grove. *Physics and Technology of Semiconductor Devices*, chapter 6.5. John Wiley and Sons, 1967.

[26] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C, 2nd. edition.* Cambridge University Press, 1992.

[27] Technology Modeling Associates. *MEDICI Manual, version 2.3.1*, 1996.

[28] R.S. Muller and T.I. Kamins. *Device Electronics for Integrated Circuits.* John Wiley & Sons, 2nd edition, 1986.

[29] J. W. Slotboom. The pn-Product in Silicon. *Solid-State Electronics, Vol. 20*, 1977.

[30] E. H. Nicollian and J. R. Brews. *MOS Physics and Technology.* John Wiley and Sons, 1982.

[31] T. Y. Chan, A. T. Wu, P. K. Ko, and C. Hu. A Capacitance method to Determine Gate-to-Drain/Source Overlap Lengths of MOSFETs. *IEEE Electron Device Letters*, EDL-8, 1987.

[32] C.S. Oh, W. H. Chang, B. Davari, and Y. Taur. Voltage Dependence of the MOSFET Gate-to-Source/Drain Overlap. *Solid-State Electronics, Vol. 33*, 1990.

[33] C.H. Wang. Identification and Measurement of Scaling-Dependent Parasitic Capacitances of Small-Geometry MOSFETs. *IEEE Transactions on Electron Devices*, ED-43(6):965–972, June 1996.

[34] R. Shrivastava and K. Fitzpatrick. A Simple Model for the Overlap Capacitance of a VLSI MOS Device. *IEEE Transactions on Electron Devices*, ED-29(12):1870–1875, December 1982.

[35] V.I. Koldyaev and L. Deferm. Closed-Form Model of the Sub-Half Micrometer LDD MOSFET Overlap Capacitance. In *Proceedings of the 27th European Solid-State Device Research Conference*, pages 668–671. ESSDERC, September 1997.

[36] International Business Machines Corporation. *DOPING User's Guide, version 4.2.0*, 1992.

[37] IBM J.B. Johnson. *REGRID Manual, version 2.15*, 1998.

[38] S.S. Furkay, J.B. Johnson, G. Fiorenza, and L.R. Logan. *FIELDAY II Manual, version 7.5.0*, 1995.

[39] S. Selbherherr, W. Hansch, M. Seavey, and J. Slotboom. The Evolution of the MINIMOS Mobility Model. *Archiv Fur Elek. and Ubertragung*, 44:161, 1990.

[40] E. Takeda and N. Suzuki. An Empirical Model for Device Degradation Due to Hot-Carrier Injection. *IEEE Electron Device Letters*, EDL-4(4):111–113, 1983.

[41] D.J. Robbins and P.T. Landsberg. Impact Ionization and Auger Recombination Involving Traps in Semiconductors. *J. Phys. C*, 13:2425–2439, 1980.

[42] W. Shockley. Problems Related to p-n Junctions in Silicon. *Solid-State Electronics*, 2(1):35–67, 1961.

[43] A.G. Chynoweth. Ionization Rates for Electrons and Holes in Silicon. *Physical Review*, 109(5):1537–1540, 1958.

[44] G.A. Baraff. Distribution Functions and Ionization Rates for Hot Electrons in Semiconductors. *Physical Review*, 128(6):2507–2517, 1962.

[45] L.V. Keldysh. Concerning the Theory of Impact Ionization in Semiconductors. *Soviet Physics JETP*, 21(6):1135–1144, 1965.

[46] J. Bude, K. Hess, and G.J. Iafrate. Impact Ionization: Beyond the Golden Rule. *Semiconductor Science and Technology, vol. 7, no. 3B*, 1992.

[47] J. Bude and K. Hess. Impact Ionization in Semiconductors: Effects of High Electric Fields and High Scattering Rates. *Physical Review B, vol. 45, no. 19*, 1992.

[48] R.C. Woods. "Soft" Energy Thresholds in Impact Ionization: a Classical Model. *IEEE Transactions on Electron Devices, vol. ED-34*, 1987.

[49] M.V. Fischetti and S.E. Laux. Monte Carlo Analysis of Electron Transport in Small Semiconductor Devices Including Band-Structure and Space-Charge Effects. *Physical Review B, vol. 38, no. 14*, 1988.

[50] F. Venturi, E. Sangiorgi, R. Brunetti, W. Quade, C. Jacoboni, and B. Ricco. Monte Carlo Simulations of High Energy Electrons and Holes in Si-n-MOSFET's. *IEEE Transactions on CAD*, 10(10):1276–1286, 1991.

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

[52] S. Saha, C.-S. Yeh, and B. Gadepally. Impact Ionization Rate of Electrons for Accurate Simulation of Substrate Current in Submicron Devices. *Solid-State Electronics*, 36(10):1429–1432, 1993.

[53] M. Knaipp and S. Selberherr. A Physically Based Substrate Current Simulation. *5th International Conference on VLSI and CAD*, pages 558–560, 1997.

[54] W. Quade, M. Rudan, and E. Schöll. Hydrodynamic Simulation of Impact-Ionization Effects in P-N Junctions. *IEEE Transactions on Computer-Aided Design*, 10(10):1287–1294, 1991.

[55] M. Rudan, F. Odeh, and J. White. Numerical Solution of the Hydrodynamic Model for a One-dimensional Semiconductor Device. *COMPEL*, 6:151–170, 1987.

[56] A. Pierantoni, P. Ciampolini, A. Gnudi, and G. Baccarani. Three-Dimensional Evaluation of Substrate Current in Recessed-Oxide MOSFETs. *IEICE Trans. Electron.*, E75-C(2):181–188, 1992.

[57] J.P. Nougier, J.C. Vaissiere, and D. Gasquet. Determination of Transient Regime of Hot Carriers in Semiconductors, Using the Relaxation Time Approximation. *J. Appl. Phys.*, 52(2):825–832, 1981.

[58] P. Scrobohaci and T. Tang. Modeling of the Hot Electron Subpopulation and its Applications to Impact Ionization in Submicron Silicon Devices — Part I: Transport Equations. *IEEE Transactions on Electron Devices*, 41(7):1197–1212, 1994.

[59] J.-G. Ahn, C.-S. Yao, Y.-J. Park, H.-S. Min, and R. Dutton. Impact Ionization Modeling Using Simulation of High Energy Tail Distributions. *IEEE Electron Device Letters*, 15(9):348–350, 1994.

[60] C.-S. Yao. *Impact Ionization Modeling and High Energy Tail Electron Transport.* PhD thesis, Stanford University, 1995.

[61] C.-S. Yao, J.-G. Ahn, Y.-J. Park, H.-S. Min, and R. Dutton. Formulation of a Tail Electron Hydrodynamic Model Based on Monte Carlo Results. *IEEE Electron Device Letters*, 16(1):26–29, 1994.

[62] W. Quade, E. Schöll, and M. Rudan. Impact Ionization within the Hydrodynamic Approach to Semiconductor Tranport. *Solid-State Electronics*, 36(10):1493–1505, 1993.

[63] A. Duncan, U. Ravaioli, and J. Jakumeit. Full-Band Monte Carlo Investigation of Hot Carrier Trends in the Scaling of Metal-Oxide-Semiconductor Field-Effect Transistors. *IEEE Transactions on Electron Devices*, 45(4):867–876, 1998.

[64] M.V. Fischetti and S.E. Laux. Monte Carlo Study of Sub-Band-Gap Impact Ionization in Small Silicon Field-Effect Transistors. *IEDM Tech. Dig.*, pages 305–308, 1995.

[65] A. Toriumi, M. Yoshimi, M. Iwase, and K. Taniguchi. Experimental Determination of Hot-Carrier Energy Distribution and Minority Carrier Generation Mechanisms due to Hot-Carrier Effects. *IEDM Tech. Dig.*, pages 56–59, 1985.

[66] D. Bertsekas. *Nonlinear Programming.* Athena Scientific, 1995.