SELF-HEATING AND SCALING OF THIN BODY TRANSISTORS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Eric Pop

December 2004

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

Kenneth E. Goodson
(Principal Co-Advisor)

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

Robert W. Dutton
(Principal Co-Advisor)

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

Krishna Saraswat

Approved for the University Committee on Graduate Studies.

Abstract

The most often cited technological roadblock of nanoscale electronics is the "power problem," i.e. power densities and device temperatures reaching levels that will prevent their reliable operation. Technology roadmap (ITRS) requirements are expected to lead to more heat dissipation problems, especially with the transition towards geometrically confined device structures (SOI, FinFET, nanowires), and new materials with poor thermal properties.

This work examines the physics of heat generation in silicon, and in the context of nanoscale CMOS transistors. A new Monte Carlo code (MONET) is introduced which uses analytic descriptions of both the electron bands and the phonon dispersion. Detailed heat generation statistics are computed in bulk and strained silicon, and within simple device geometries. It is shown that non-stationary transport affects heat generation near strongly peaked electric fields, and that self-heating occurs almost entirely in the drain end of short, quasi-ballistic devices. The dissipated power is spectrally distributed between the (slow) optical and (fast) acoustic phonon modes approximately by a ratio of two to one.

In addition, this work explores the limits of device design and scaling from an electrical and thermal point of view. A self-consistent electro-thermal compact model for thin-body (SOI, GOI) devices is introduced for calculating operating temperature, saturation current and intrinsic gate delay. Self-heating is sensitive to several device parameters, such as raised source/drain height and material boundary thermal resistance. An experimental method is developed for extracting via/contact thermal resistance from electrical measurements. The analysis suggests it is possible to optimize device geometry in order to simultaneously minimize operating temperature and intrinsic gate delay. Electro-thermal contact and device design are expected to become more important with continued scaling.

Acknowledgment

I would first like to thank my co-advisors, Profs. Ken Goodson (Mechanical Engineering) and Bob Dutton (Electrical Engineering) for agreeing to supervise this interdisciplinary project. I am also indebted to them for the many discussions about research, academia, and life in general, both during and after Stanford. Thanks also go out to Profs. Krishna Saraswat and KJ Cho for serving on my Orals committee, and for providing me with precise feedback during the course of this thesis.

My graduate career would not have happened at Stanford without the early support of Profs. Jim Harris (at Stanford), and Steve Senturia, Jesus del Alamo and Dimitri Antoniadis (at MIT). The latter two also served as role models while I was still at MIT, both having attended Stanford graduate school. My graduate studies were made possible by generous financial support from the Semiconductor Research Corporation (SRC) through an SRC/IBM Fellowship. Over the years, I have also enjoyed many technical discussions with various mentors from outside Stanford: Bob Miller, Max Fischetti, Steve Laux, Phil Oldiges and Keith Jenkins (IBM), Charvaka Duvvury (TI), Zoran Krivokapic (AMD), James Hutchby (SRC) and Profs. Mark Lundstrom (Purdue) and Umberto Ravaioli (UIUC).

Last but not least I would like to thank my research collaborators whom I have learned much from, Sanjiv Sinha and Chi On Chui, and our research group administrators, Cecilia "CC" Gichane-Bell and Fely Barrera. Finally, life in the Bay Area would have been a lot less fun without my many friends. Shout outs go to the KZSU 90.1 crew, the Thursday night crew, the Tahoe crew, the Swedish crew and the Spanish Armada (the last two having made me feel welcome as an unofficial member). This thesis is dedicated to my family, my sister Lia, and the memory of my grandparents.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The technological revolution that started with the introduction of the transistor just over half a century ago is without parallel in the way it has shaped our economy and our daily lives. The current trend toward nanoscale electronics is expected to have a similar impact into the third millennium. Commercial integrated circuits are currently available with transistors whose smallest lateral feature size is less than 100 nm and the thinnest material films are below 2 nm, or only a few atomic layers thick. Such miniaturization has led to tremendous integration levels, with a hundred million transistors assembled together on a chip area no larger than a few square centimeters. Integration levels are projected to reach the gigascale as the smallest lateral device feature sizes approach 10 nm.

The most often cited technological roadblock of this scaling trend is the "power problem," i.e. power densities, heat generation and chip temperatures reaching levels that will prevent the reliable operation of the integrated circuits. Chip-level power densities are currently on the order of 100 W/cm^2, and if the rates of integration and miniaturization continue to follow the ITRS (International Technology Roadmap for Semiconductors) guidelines [1], the chip-level power density is likely to increase even further [2], as illustrated in Fig. 1.1. Higher power densities will quickly drain the batteries of portable devices and render most advanced, future electronics unusable without significant cooling technology, or fundamental shifts in design. The situation is compounded by millimeter-scale hot spots

Material	κ_{th} ($Wm^{-1}K^{-1}$)
Si (bulk)	148
Ge (bulk)	60
Silicides	40
Si (10 nm)	13
$Si_{0.7}Ge_{0.3}$	8
SiO_2	1.4

Table 1.1: Thermal conductivities (κ_{th}) of a few materials used in semiconductor device fabrication. Phonon boundary scattering significantly reduces the thermal conductivity of a 10 nm thin silicon film. Note that phonons (lattice vibrations) are responsible for heat conduction in all dielectric materials listed, but electrons are the heat carriers in silicides (e.g. NiSi), which are metals.

on the chip, i.e. localized regions of higher heat generation rate per unit area, hence higher temperatures (e.g. near the clock drivers) [3]. The power problem has recently been addressed in several ways, mainly from a system design point of view [4]. A radical way to cool integrated circuits by running water microchannels on the backside of the chip was originally proposed [5], and has recently re-emerged in a sealed, compact package [6]. Circuit designers can selectively turn off different parts of the chip to save power, and clock frequencies can be adjusted on the fly, depending on the task at hand. Similarly, one can wonder if perhaps the circuit building blocks, i.e. the transistors themselves could be designed to be more power efficient, or less affected by their own self-heating. This dissertation has been geared toward answering the latter question, as well as obtaining a detailed physical picture of the origins of self-heating in silicon.

While the total heating rate of microprocessor chips has received much attention, a different thermal management challenge faces device and circuit designers at nanometer length scales, within individual transistors. Novel, complicated device geometries tend to make heat removal more difficult (Fig. 1.2) and most new materials being introduced in device processing have lower thermal conductivities than bulk silicon (Table 1.1). Modern device technologies already operate at length scales on the order of the electron and phonon[1]

[1]Phonons are discrete quanta of lattice vibrations, responsible for thermal energy transport in crystalline dielectrics. A good introduction to phonons can be found, e.g., in C. Kittel's classic text *Introduction to Solid State Physics* (Wiley) [7].

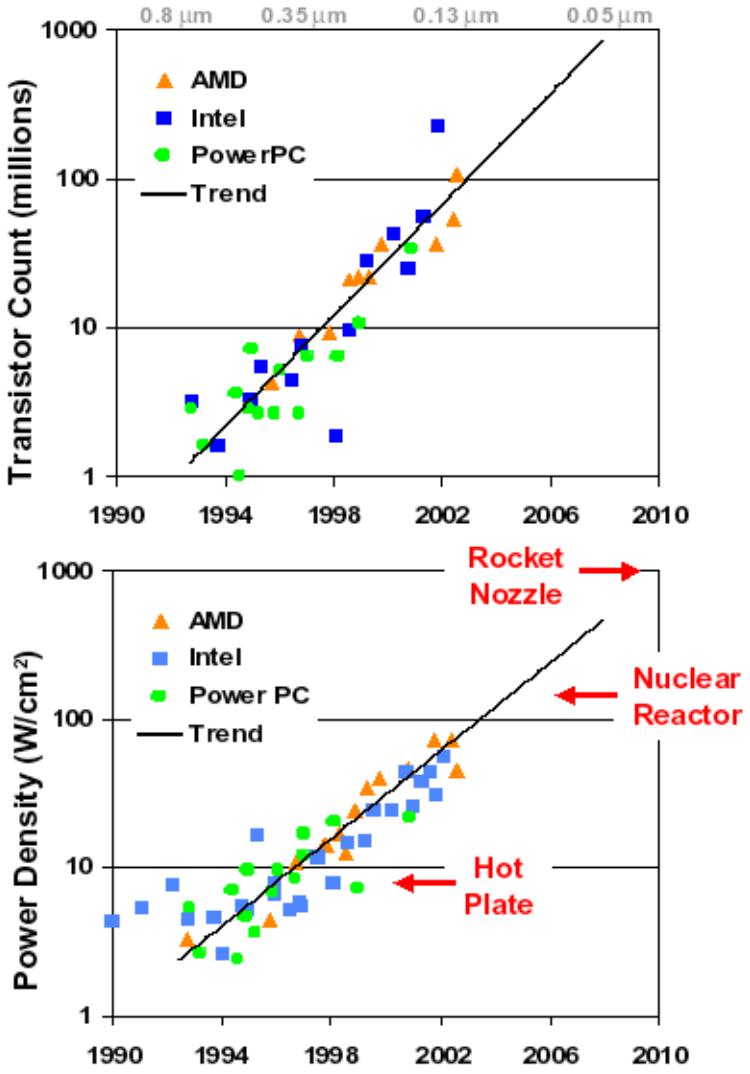

Figure 1.1: Trends of on-chip transistor count (top) and on-chip power density (bottom) for three of the main industry players, over the past 10+ years. Note the vertical axes are logarithmic and the horizontal ones (years) are linear. If the same trend is followed, impractical power densities could be reached in the near future, unless significant shifts in transistor and/or circuit design are implemented. (Figure data compiled by F. Labonte.)

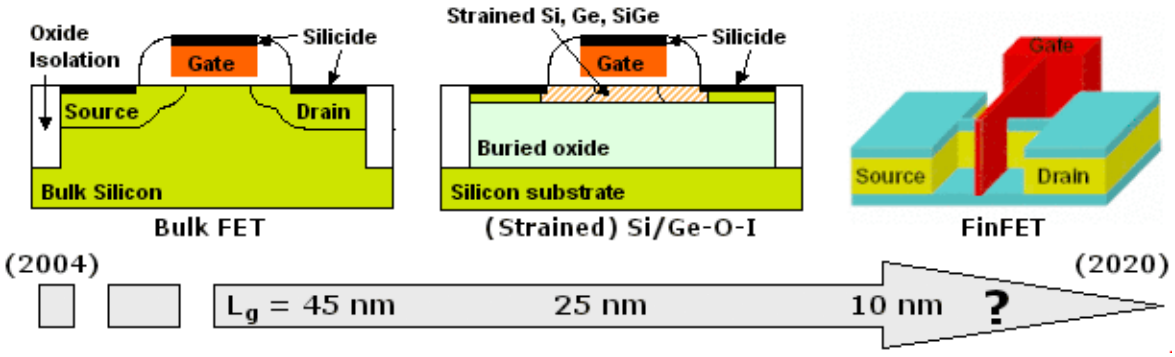

Figure 1.2: Evolution of transistor designs, from bulk FETs, toward transport-enhanced (strain, Ge channel) and thin body (SOI, FinFET) devices. Thin body devices all exhibit poorer thermal properties than bulk devices, owing to the confined geometry and oxide insulator.

mean free paths (approximately 5–10 and 200–300 nm in bulk silicon at room temperature, respectively [8, 9]), and future technologies are going to forge deeper into this sub-continuum regime. Ballistic conditions dominate both electron (current) and phonon (heat) transport at such length scales, leading to a non-equilibrium condition between the energy carriers. The electron-phonon interaction is neither energetically nor spatially uniform [10] and the generated phonons have widely varying contributions to heat transport: optical phonons make virtually no contribution to the thermal conductivity of silicon, which is dominated by acoustic phonon transport [9, 11].

In the context of a transistor, the applied voltage leads to a lateral electric field which peaks near the device drain. This field accelerates the free charge carriers (e.g. conduction band electrons in a n-MOSFET) which gain energy, therefore "heating up." Electrons can scatter with each other, with lattice vibrations (phonons), interfaces, imperfections or impurity atoms. Of these, electrons only lose energy by scattering with phonons, consequently heating up the lattice through the mechanism known as Joule heating. Other scattering mechanisms only affect the electron momentum [8]. The lattice absorbs the extra electron energy, heats up to a higher temperature (T), and in return affects the electronic transport properties of the material: the electron mobility in bulk (undoped) silicon decreases approximately as $T^{-2.4}$ around room temperature, approximately as $T^{-1.7}$ in highly doped bulk silicon, and $T^{-1.4}$ in nanometer-thin silicon layers [12, 13].

In silicon, as in most semiconductors, high field Joule heating is typically dominated by optical phonon emission. Optical phonons are slow and they make virtually no contribution to heat transport. Rather, they decay into the faster acoustic modes, which carry the energy away from the hottest regions. Optical-to-acoustic decay times are relatively long (on the order of picoseconds) compared to the electron-phonon scattering time (tenths of picoseconds) [14]. If the generation rate of optical modes due to Joule heating from current flow is higher than their rate of decay into acoustic modes, a phonon energy bottleneck is created and the optical mode density can build up, directly affecting electron transport.

1.1 Thermal Implications of Device Design

As devices are scaled to dimensions comparable to, or less than the mean free path of the thermal energy carriers (phonons), a number of new considerations come into play. The bulk properties of materials are modified at nanometer length scales, and the continuum classical model of heat conduction (Fourier law) must be replaced by a more sophisticated formulation which takes into account the "granularity" of heat transport via phonons. In addition, heat transfer through device boundaries and contacts, especially in the case of confined-geometry designs (e.g. ultra-thin body or FinFET devices) is expected to play an important, and possibly limiting role.

1.1.1 Bulk Silicon Transistors

Traditional, bulk transistor designs (first diagram in Fig. 1.2) typically incorporate only a few materials, most notably silicon, silicon dioxide and nitride insulators, and silicided (e.g. NiSi) contacts. The high thermal conductivity of bulk silicon facilitates heat transport from the transistor channel down to the backside of the chip, where it is usually removed with a heat sink. Thermal transport in bulk transistors has traditionally been modeled in the classical limit, as sub-continuum thermal effects can be neglected for device dimensions larger than the phonon mean free path. As devices are scaled below 100 nm, two sub-continuum effects are expected to play a role in bulk transistor thermal transport. The

small region of high electric field near the drain gives rise to a strongly localized hot spot, only a few tens of nanometers across, and hence much smaller than the bulk phonon mean free path [15]. This leads to ballistic phonon transport in the vicinity of the heat source, and higher temperatures than those predicted by classical diffusion theory [16]. In this situation, a solution of the phonon Boltzmann Transport Equation is more accurate than the classical heat diffusion equation [17].

The second sub-continuum thermal effect to be expected in ultra-scaled bulk FETs has to do with the non-equilibrium interaction between the generated optical and acoustic phonons. Since nearly-stationary optical phonons form the majority of the vibrational modes generated via Joule heating, they tend to persist in the hot spot region until decaying into the faster acoustic modes. This non-equilibrium scenario may become particularly relevant when device switching times approach the optical-acoustic decay times, on the order of several picoseconds. A careful transient solution of the phonon populations may be necessary to properly account for the non-equilibrium distributions [18].

Both sub-continuum effects in bulk nanotransistors are expected to take place in the drain region. Hence, their effect is more likely to be pronounced on device reliability, rather than reducing the device current drive, since the latter is thought to be determined by the electron injection velocity near the source [19]. However, some indications exist that in the limit of the smallest achievable bulk silicon MOSFETs (10 nm channel length) the optical phonons generated in the drain may reach the device source before decaying into acoustic phonons, and hence directly affect the source injection velocity [15].

1.1.2 Non-Traditional Transistors

Advanced, non-traditional device fabrication introduces a number of new materials with lower thermal conductivities than bulk silicon. The thermal properties of these materials are therefore expected to play a more significant role in device design and thermal behavior, as summarized in Table 1.1. Bulk germanium transistors, for example, would suffer from increased operating temperatures due to a substrate thermal conductivity approximately 60 percent lower than bulk silicon transistors. Strained silicon channel devices grown on

a graded $Si_{1-x}Ge_x$ buffer layer benefit from an increased mobility due to band degeneracy splitting and a lighter effective mass in the strained film. However, their thermal behavior is adversely affected by the lower thermal conductivity of the $Si_{1-x}Ge_x$ alloy layer. The situation is worse for ultra-thin body devices grown on a silicon dioxide film. From an electrical point of view, silicon-on-insulator (SOI) and FinFET devices (second and third diagrams in Fig. 1.2) benefit from lower capacitive coupling with the substrate and better gate control of the channel — hence increased switching speeds and better turn-off characteristics. Thermally, however, they are strongly affected by the very low thermal conductivity of the buried oxide layer, which is about two orders of magnitude less than that of silicon. The thermal conductivity of thin semiconductor films (thinner than the phonon mean free path) is also significantly reduced by phonon boundary scattering. For example, the thermal conductivity of a 10 nm thin silicon film is expected to be reduced by an order of magnitude from that of bulk silicon [20]. Although experimental data for such thin films does not yet exist, an estimate can be based on extrapolations to available data [21, 22], and also supported by recently measured reduced thermal conductivity in silicon nanowires [23]. For film (or nanowire) thicknesses on the order of a few nanometers, and therefore comparable to the phonon wavelength, phonon confinement effects may become important as well, further contributing to a degradation in thermal conductivity [24].

The small dimensions of future, confined-geometry device designs also imply a large surface-to-volume ratio, and hence a stronger effect of material boundary resistance. Very few estimates exist on the magnitude of the thermal boundary resistance between dissimilar materials, e.g. a dielectric and a metal. Some measurements indicate it is on the order of the thermal resistance of a 20 nm thick silicon dioxide film, and fairly independent of processing conditions or the specific type of metal and dielectric involved [25]. This is a significant value for nanoscale devices, and very important to understand. As more materials (e.g. high-k dielectrics, germanium, various silicides) are introduced in semiconductor processing, there is a growing need to understand the magnitude of boundary thermal resistance and its importance in future nanoscale device behavior. The boundary thermal resistance plays a significant role, for example, when a metal electrode is placed on top of the high-k gate

dielectric (as expected for threshold voltage control in future technologies), as well as for device metal contacts and interconnects. More measurements are needed in this area, and more available data on thermal boundary resistance would also help towards a better understanding and modeling of the atomic scale interaction at the interface between two materials. Part of Chapter 5 of this dissertation presents a method for characterizing the thermal interface resistance associated with the silicided contact and via geometry for deep submicron device designs. This contact and via thermal resistance may play a significant role in heat dissipation during normal operation as well as during Electrostatic Discharge (ESD) events of ultra-scaled thin-body transistors.

1.2 Heat Conduction in Semiconductors

Proper modeling of heat conduction in semiconductors and metals is essential for understanding their thermal behavior and enabling improved design of nanometer-scale transistors. As device dimensions and the body thickness (e.g. in ultra-thin body SOI and FinFET devices) are scaled to the order of tens of nanometers, or comparable to the mean free path of the energy carriers, sub-continuum effects are expected to become important. The carriers responsible for heat transport in metals are the nearly-free conduction electrons, whose thermal conductivity κ_e can be related to their electrical conductivity σ through the Wiedemann-Franz law [7]:

$$\kappa_e = \frac{\pi^2}{3}\left(\frac{k_B}{e}\right)^3 \sigma T \tag{1.1}$$

where k_B is the Boltzmann constant, e is the elementary charge and T is the absolute temperature. The energy carriers responsible for heat transport in crystalline dielectrics (semiconductors) are the lattice vibrations (phonons). Even in heavily doped semiconductors, the electronic contribution to the thermal conductivity is only on the order of one percent [26]. The thermal conductivity of a semiconductor can be written as [27]

$$\kappa_s = \frac{1}{3} C_s \overline{v} \Lambda \tag{1.2}$$

where C_s is the heat capacity per unit volume, $\overline{v}$ is the average phonon velocity and Λ is the average phonon mean free path. The classical, continuum, heat diffusion equation (Fourier Law)

$$C_s \frac{\partial T}{\partial t} = \nabla \cdot (\kappa_s \nabla T) + Q''' \tag{1.3}$$

where Q''' is the heat generation rate per unit volume, cannot properly resolve heat transfer problems on small time scales (on the order of the phonon relaxation times, i.e. picoseconds) or short length scales (tens of nanometers, or less than the acoustic phonon mean free path Λ). At such scales, the continuum heat diffusion theory must be replaced by a more sophisticated formulation which takes into account the "granularity" of heat conduction via discrete phonon modes. The phonon dispersion spectrum also comes into play (the phonon frequency ω being a function of wave vector), as transverse acoustic phonons are typically slower than longitudinal modes, and their group velocity is also a function of wave vector. At length scales shorter than the acoustic phonon mean free path (a few hundred nanometers), but larger than the phonon wavelength (a few nanometers), phonons can be treated as semi-classical particles and the Boltzmann Transport Equation (BTE) may be used [28, 29]:

$$\frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla f = \left(\frac{\partial f}{\partial t}\right)_{coll} + \left(\frac{\partial f}{\partial t}\right)_{g} \tag{1.4}$$

where $f(\mathbf{r}, \omega, t)$ is the phonon distribution function, $\mathbf{r}$ is the spatial coordinate and $\mathbf{v}$ is the phonon velocity. The $\partial \mathbf{k}/\partial t$ term usually present in the BTE has been omitted, since phonons, unlike electrons, are not influenced by external forces like the electric field [29]. The first term on the right hand side is due to phonon collisions, and the second term is due to phonon generation and annihilation. In the relaxation time approximation the collision term can be replaced by

$$\left(\frac{\partial f}{\partial t}\right)_{coll} = \frac{f_o - f}{\tau_{ph}} \tag{1.5}$$

where $f_o = 1/\left(\exp(\hbar\omega/k_B T) - 1\right)$ is the equilibrium Planck distribution at temperature T, and τ_{ph} is the average phonon scattering time, such that $\Lambda = \overline{v}\tau_{ph}$. With these approximations, the two sides of Eq. 1.4 can be integrated over the phonon frequency and density of

states, and the BTE can be rewritten in terms of the phonon energy density u as [17]:

$$\frac{\partial u}{\partial t} + \mathbf{v} \cdot \nabla u = \frac{u_o - u}{\tau_{ph}} + Q''' \tag{1.6}$$

where, as before, the term Q''' is the heat generation rate per unit volume. The current state of the art heat transport simulations all replace the heat diffusion equation with various solutions of the phonon Boltzmann Transport Equation. Mazumder and Majumdar [11] solved the BTE with the Monte Carlo method but only incorporated the acoustic phonon branches, and hence their approach is not directly applicable to the study of self-heating in electronic devices. Sverdrup *et al.* [17] solved the phonon BTE both for acoustic and optical phonons via the finite volume method inside a transistor. However, as with other previous work, they modeled the heat generation rate as the dot product of the electric field and current density, and assumed all energy to be dissipated to the optical phonon modes (which were approximated as stationary), hence overestimating the resulting lattice temperature. Such an approach does not provide enough information about the microscopic (non-local) details of self-heating in semiconductor devices, as described in the next section. It is also clear from a comparison of the BTE (Eq. 1.6) and the heat diffusion equation (Eq. 1.3), that irrespective of the complexity of the phonon transport model, the heat conduction problem in an electronic device intimately depends on identifying a proper treatment for the lattice heating term from electron-phonon interactions, Q'''.

1.3 Heat Generation in Semiconductors

The most basic method for calculating the total heat generation rate (power dissipated) within a lumped (semi)conducting element is to write it as the product of the current and voltage:

$$Q = I \times V. \tag{1.7}$$

Here, the voltage drop is that across the device alone, excluding its contacts. Hence, this formula must be applied with care in describing the power dissipated in a structure with

relatively large contact resistance, e.g. a nanotube or molecular device. This expression will also tend to overestimate the total heat dissipated in a quasi-ballistic device, i.e. one that is only a few electron mean free paths long. In such a device, electrons will gain energies comparable to qV but will generally not undergo enough inelastic scattering events to completely thermalize and give up this energy to the lattice (in the form of self-heating) by the time they exit. Hence, relatively hot electrons will escape through the contacts, and some portion of the $I \times V$ power will be deposited there instead [30]. In other words, the power dissipated inside the device is less than the above formula suggests, and the rest of power dissipation occurs in the contacts. In addition, the simple formula above only gives an estimate of the *total* power dissipated, not of the physical location of its peak (if any) or its make-up in terms of the emitted phonon frequencies. However this formula is very well suited for quick, first order estimates.

In the context of a semiconductor device simulator, heat generation due to electric current flow is most often calculated with the classical drift-diffusion approach. The main component of this heat generation expression is the dot product of the electric field $\mathbf{E}$ and current density $\mathbf{J}$, as computed at every grid node within the simulation [17, 26]:

$$Q''' = \mathbf{J} \cdot \mathbf{E} + (R - G)(E_g + 3k_B T) \tag{1.8}$$

where $\mathbf{J} = qn\mathbf{v_e}$, with n being the electron number density and $\mathbf{v_e}$ the average electron velocity.[2] Note the notation of Q''' (power density per unit volume, i.e. W/cm^3) versus Q in Eq. 1.7 (total power in Watts). The total power Q in this formulation can be recovered by integrating Eq. 1.8 over the device volume. The first term represents the Joule heating rate, which is usually positive (power generation) as electrons drift down the band structure slope under the influence of the electric field, and gradually lose energy through net phonon emission. It should be noted that Joule heat can also be negative (power consumption)

[2]Here it is assumed that electrons are the majority current carriers (n-type semiconductor), but the heating rate due to hole current can be similarly incorporated.

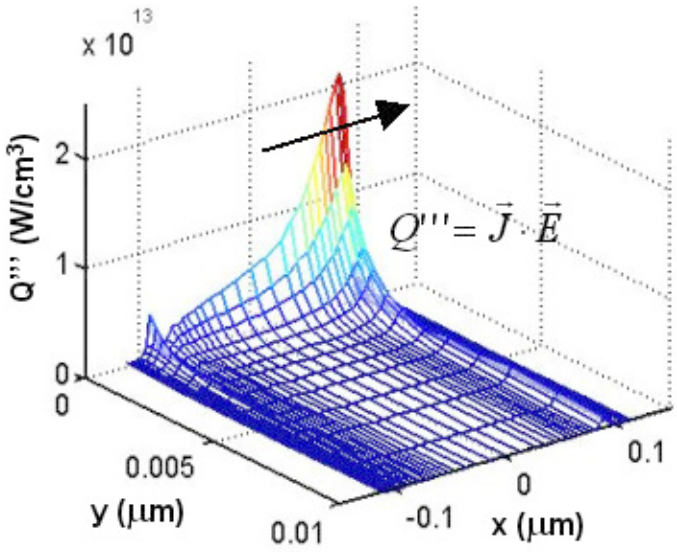

Figure 1.3: Heat generation rate computed in a 0.18 micron bulk nMOSFET device with the classic $\mathbf{J} \cdot \mathbf{E}$ approach (Eq. 1.8). This approximation does not capture non-local electron transport, and the arrow indicates the direction of heat generation beyond the peak of the electric field. Non-local heat generation effects become important at device channel lengths comparable to the electron mean free path, as described in Chapter 4.

when electrons diffuse *against* an energy barrier,[3] and the energy required to move up the conduction band slope is extracted from the lattice through net phonon absorption [29]. The second term of the above equation is the heat generation rate due to non-radiative electron and hole generation and recombination processes. When an electron and a hole, both with an average energy $(3/2)k_BT$ recombine, the excitation energy $E_g + 3k_BT$ is given off either directly to the lattice, or to another charge carrier (Auger transition). In the latter case, the excited particle eventually gives off the energy to the lattice by phonon emission as well. Equation 1.8 may include other higher order terms, accounting for electron drift along a temperature gradient or across a discontinuity in the band structure, e.g. a heterojunction like in a semiconductor laser [29]. Figure 1.3 shows the heat generation rate computed in a 0.18 micron gate length transistor with the approach described above, as implemented in the commercial simulator MEDICI. Unfortunately, this field-dependent method does not

[3]Such as in a forward-biased *pn* junction, or near the energy barrier at the injection point from the source into the channel of a MOSFET.

account for the microscopic non-locality of phonon emission near a strongly peaked electric field region, such as in the drain of the transistor. Although electrons gain most of their energy at the location of the electric field peak, they typically travel several mean free paths before releasing all of it to the lattice, in decrements of (at most) the optical phonon energy. In silicon transistors, for example, electrons can gain energies that are a significant fraction of an eV, while the optical phonon energy is only about 60 meV. Assuming an electron velocity of 10^7 cm/s (the saturation velocity in silicon) and an electron-phonon scattering time around 0.05–0.10 ps in the high-field region, the electron-phonon mean free path is then on the order of 5–10 nm. The full electron energy relaxation length is therefore even longer, on the order of several inelastic mean free paths. While such a discrepancy may be neglected on length scales of microns, or even tenths of a micron, it must be taken into account when simulating heat generation rates on length scales of 10 nm, as in a future generation transistor. The highly localized electric field in such devices leads to the formation of a nanometer-sized hot spot in the drain region, that is spatially displaced (by several mean free paths) from this drift-diffusion prediction. In addition, the $\mathbf{J} \cdot \mathbf{E}$ formulation of the Joule heating also does not differentiate between electron energy exchange with the various phonon modes, and does not give any spectral information regarding the types of phonons emitted.

The heating rate can also be computed with the more sophisticated hydrodynamic approach, as a function of the electron temperature and an average energy relaxation time [31]:

$$Q''' = \frac{3}{2} k_B \frac{n(T_e - T_L)}{\tau_{e-L}} + (R - G) \left[E_g + \frac{3}{2} k_B (T_e + T_L) \right] \tag{1.9}$$

where the holes have been assumed in thermal equilibrium with the lattice (T_L), but the electrons are described by their own temperature (T_e), energy relaxation time (τ_{e-L}), and number density (n). This is the situation in which electrons are the majority current carriers, but the holes and the hole temperature can be incorporated in a similar way. Unlike the $\mathbf{J} \cdot \mathbf{E}$ method, the hydrodynamic approach has been shown to be better suited for capturing

non-local transport effects near highly peaked electric field regions. However, the hydrodynamic approach suffers from simplifications inherent to using a single (averaged) carrier temperature and relaxation time, since scattering rates are strongly energy dependent. Similar to the previously described methods, this average carrier temperature-based approach also does not differentiate among electron energy exchange with the various phonon modes, and does not give information regarding the frequencies of phonons emitted. Such spectral information is important because it is well-known that the emitted phonons travel at different velocities and have widely varying contributions to heat transport [9, 11] and to device self-heating [32, 33].

The mechanism through which lattice self-heating occurs is that of electron scattering with phonons, and therefore only a simulation approach which deliberately incorporates all such scattering events will capture the full microscopic, detailed picture of self-heating. As such, the Monte Carlo method [34, 35], although originally developed for studies of hot electron effects [36], is ideally suited for computing a detailed picture of self-heating as well. This is the approach adopted and extended in this work, as described in Chapters 3 and 4. From a detailed physical point of view, self-heating starts when the nearly-free conduction band electrons in a semiconductor are accelerated by the electric field. The electrons gain energy from the field, then lose it by inelastically scattering with the lattice phonons, as all other scattering mechanisms (e.g. impurity or boundary scattering) are considered elastic (they affect the electron momentum, but not the energy) [8]. Electrons with energies below 50 meV tend to scatter mainly with acoustic phonons in silicon, while those with higher energy scatter strongly with the optical modes. The optical phonon branches have low group velocity (on the order of 1000 m/s) and their occupation number is also relatively low, hence they do not contribute to heat transport [9]. The primary heat carriers in silicon are the faster acoustic phonon modes, which are much more populated and have group velocities from 5000 (for transverse modes) to 9000 m/s (for longitudinal modes).[4] Optical phonons decay into acoustic modes, but over relatively long time scales (picoseconds) compared to

[4]The group velocity of a phonon branch is given by its slope on the dispersion relationship from, e.g., Fig. 3.4.

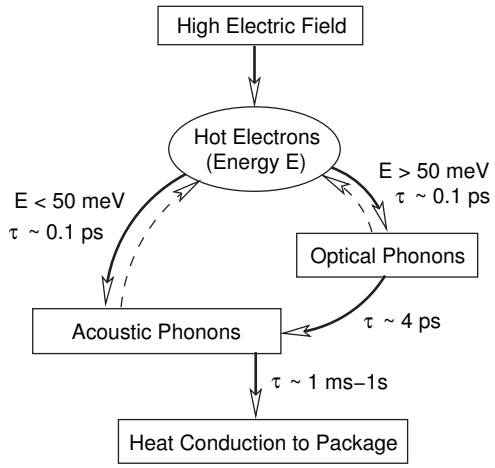

Figure 1.4: Diagram and characteristic time scales of the energy transfer processes in silicon. Scattering with low group velocity optical phonons is the dominant relaxation mechanism for electron energies above 50 meV. This may create a phonon energy bottleneck until the optical phonons decay into the faster acoustic modes.

the electron-optical phonon scattering time (on the order of tenths of picoseconds) [14]. This creates a phonon energy bottleneck which can cause the density of optical phonon modes to build up over time, leading to more scattering events and impeding electron transport [32]. These processes are symbolically illustrated in Fig. 1.4. The dotted lines represent the effect of the phonons on the electron population and hence, their transport. The details of the electron band structure and of the phonon dispersion, as well as those of the electron-phonon scattering rates can be readily incorporated into a Monte Carlo simulation, as described in Chapters 3 and 4 of this dissertation. The heat generation rate in a Monte Carlo simulation at steady state can then be computed as a sum over all phonon emission minus all phonon absorption events per unit time:

$$Q''' \sim \frac{1}{t_{sim}} \sum \left(\hbar\omega_{ems} - \hbar\omega_{abs} \right) \tag{1.10}$$

where t_{sim} is the simulation time. This approach can then be used to investigate the phonon generation spectrum (the generation rate as a function of phonon frequency and mode), as well as to study non-local heat generation near a strongly peaked electric field within a realistic device geometry.

1.4 The Scope of this Work

The goals of this dissertation are two-fold. The first objective is to explore the microscopic details of self-heating in bulk and strained silicon. To this end, the Monte Carlo method was used, and a new model was implemented from the ground up. This approach is different from previous work (see Chapter 2) in its use of an analytic description for both the electron energy bands as well as the phonon dispersion. The method is also extended to simple, yet realistic device geometries where self-heating is computed and compared with traditional simulation methods. The Monte Carlo approach gives information on both the location and make-up of the heat generation region within the drain of a transistor.

The second objective of this dissertation is to analyze the design and scaling of confined-geometry (i.e. SOI, FinFET) transistors from an electro-thermal point of view. The Monte Carlo work shows that in the limit of the shortest possible devices, the heat is almost entirely generated in the device drain. A carefully calibrated compact model for the self-heating of such transistors is introduced and shown to yield device temperatures very sensitive to device geometry, as well as to contact resistance. Several device design guidelines are proposed, and the analysis indicates it is possible to optimize device geometry in order to simultaneously minimize operating temperature and intrinsic gate delay.

1.5 Organization

This dissertation is organized in six chapters and one appendix, as follows. Chapter 2 presents a brief introduction of the Monte Carlo (MC) method for transport in semiconductors, and specifically in silicon. This chapter also reviews several simulation approaches of various complexity, and places the current contribution in historical context.

Chapter 3 describes the details of the Monte Carlo implementation in this work. This includes a combination of analytic (ellipsoidal) electron energy bands and analytic (quadratic) phonon dispersion, an combined approach introduced for the first time in this work [37]. The electron-phonon scattering rates are re-derived taking into account the phonon dispersion, and a set of scattering deformation potentials is proposed which properly reproduces

experimental data in both bulk and strained silicon.[5] Chapter 3 also discusses the implementation of the Monte Carlo method within the context of a realistic 1- and 2-D device geometry, including impurity and boundary scattering, as well as a self-consistent solution of the Poisson equation.

Chapter 4 applies the Monte Carlo simulation method to compute detailed heat generation rates in both bulk and strained silicon. The heat generation spectrum (net emitted phonons as a function of frequency) is computed. The method is also applied to simple, yet relevant 1-D device geometries and the non-local nature of heat generation near strongly peaked electric fields is investigated. The study finds most of the heat generated in a short[6] device to occur in the device drain, and not in the channel where transport is quasi-ballistic. The centroid of this "hot spot" is displaced by several inelastic mean free paths from the peak of the electric field region.

Chapter 5 introduces a compact electro-thermal model suitable for analysis of confined-geometry devices (e.g. SOI, GOI or FinFET). It is shown that the design of the raised source/drain (i.e. its resistance) plays an important electrical as well as thermal role, since most of the heat is dissipated there. The interface thermal resistance, both at the gate (metal) to oxide boundary, as well as at the device contacts (silicide and via) is also shown to play an increasingly limiting role in the heat dissipation of devices near the end of the Technology Roadmap [1]. The compact model is also used to investigate the optimal device design space such as to minimize both operating temperature and intrinsic gate delay. Also, thin-body germanium-on-insulator (GOI) transistors are compared with equivalent SOI devices. It is shown that well-designed GOI devices are not expected to suffer from worse self-heating, despite their slightly higher thermal resistance, in part due to lower power dissipation and partly due to thin film germanium mobility being less sensitive to temperature.

Chapter 6 provides an overall conclusion of this thesis. The results are analyzed and several suggestions for future research directions are offered. It is suggested that similar

[5] Previous approaches have used separate sets of deformation potentials in the two cases, without being able to reconcile them.

[6] Compared to the electron mean free path, i.e. near 5-10 nm.

studies ought to be carried out for other confined geometry devices (nanotubes, nanowires), to understand the limiting effects of self-heating on device performance and design space. The role of thermal (and electrical) contact resistance must also be studied more carefully in such geometrically confined device designs. The appendix contains a brief user manual for MONET, the Monte Carlo code implemented during the course of this dissertation. The work described in the following chapters should provide enough detail to enable anyone with similar resources to duplicate the results of this thesis.

Chapter 2

The Monte Carlo Method for Transport in Semiconductors

The Monte Carlo (MC) method is regarded as the most comprehensive approach for simulating charge transport in semiconductors. An early standard was set by the work of Canali *et al.* [38] and Jacoboni *et al.* [34] using analytic, ellipsoidal descriptions of the energy band structure. Over the past two decades the research community has added numerous enhancements, including more comprehensive physical models, more efficient computer algorithms, new scattering mechanisms, boundary conditions, electrostatic self-consistency in device simulations, etc. A significant enhancement of the physical models was the introduction of full electron energy bands from empirical pseudopotential calculations [36, 39].

For device operating voltages near 5 Volts, the full band MC method has been very useful with high-energy transport simulations, including impact ionization [39, 40], where details of the full band structure are essential. As device dimensions are scaled into the nanometer range and supply voltages are reduced below the material's band gap (1.1 Volts for silicon) the role of impact ionization is greatly diminished. Transport at lower energies can be adequately simulated with analytic band models. Hence simpler, faster analytic band MC codes (including quantum mechanical corrections where required by confined dimensions) can be employed as engineering design tools for future nanoscale devices. In

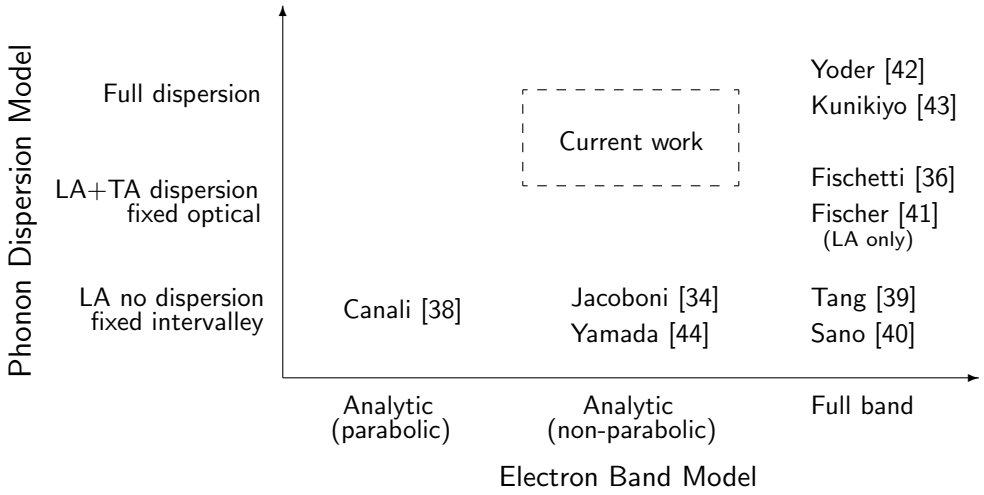

Figure 2.1: Historical context of various Monte Carlo models for electron transport in silicon. The computational burden increases for full band (and full dispersion) simulations.

addition, despite increasingly sophisticated treatment of the electron energy bands over the years, the phonon dispersion relation is still commonly simplified in practical device simulators. Electron-phonon scattering is usually computed with a single dispersionless acoustic mode and with one (or a few) fixed energy optical modes. This dissertation (in particular, Chapter 3) presents a new Monte Carlo model which uses complete analytical descriptions for both the electron band structure and the phonon dispersion relationship. The approach is computationally efficient on modern PC workstations and suitable for simulating electron transport in future, low-voltage technologies, while describing the electrons and phonons with comparable accuracy.

2.1 Historical Overview

Figure 2.1 shows a brief historical overview of various MC simulation methods for charge transport in silicon. Canali *et al.* [38] introduced the first multi-valley model with parabolic, ellipsoidal bands and phonon scattering with a single dispersionless longitudinal acoustic (LA) mode and six fixed-energy intervalley phonons. Jacoboni *et al.* [45] accounted for

analytic band non-parabolicity and slightly altered Canali's set of phonon deformation potentials. A few years later Brunetti *et al.* [46] introduced a new set of deformation potentials, more closely matching available data on the anisotropy of electron diffusion in silicon. This phonon model was used by Jacoboni *et al.* in an excellent and frequently referenced review of the MC method [34], and it subsequently became the set of phonon energies and deformation potentials most often employed in the literature over the past two decades. Other workers [44] also introduced scattering with first order intervalley phonons. Tang and Hess [39] were the first to incorporate the full band structure of silicon (computed from empirical pseudopotentials) for MC transport. However, they used the simple phonon model of Canali and Brunetti (dispersionless LA phonons, six fixed intervalley phonons), and the deformation potentials of Brunetti *et al.* [46]. Sano *et al.* introduced wave vector dependent impact ionization rates in a full band MC formulation [40], but computed phonon scattering rates with the multi-valley deformation potentials of Canali *et al.* [38].

Realistic device simulations using electrostatically self-consistent full band MC were first performed by Fischetti and Laux [36]. They were also the first to make the distinction between longitudinal (LA) and transverse acoustic (TA) intravalley scattering, using a simple analytic dispersion for both modes. Fischer *et al.* [41] pointed out the poor definition of energy "valleys" in the context of full band models, and used only two averaged deformation potentials: one for fixed-energy optical phonons and another for acoustic phonons (LA, but not TA), including their dispersion. The most sophisticated MC models for charge transport in silicon were developed by Yoder *et al.* [42] and Kunikiyo *et al.* [43]. They employed the full band structure computed from empirical pseudopotentials and the full (anisotropic) phonon dispersion obtained from an adiabatic bond-charge model. The electron-phonon scattering rates were calculated as a function of energy and wave vector, consistently with the band structure and phonon dispersion. In the absence of any adjustable parameters, mobilities computed with these *ab initio* models are typically less accurate than those computed using more empirical simulators. Such codes also present formidable computational burdens, rendering them impractical for simulations of realistic devices. Their only applications have been for very detailed bulk transport calculations.

Most MC codes found in practice today employ a sophisticated, full description of the electron energy bands (often including quantum effects [47]), yet scattering rates and energy exchange with the lattice are only computed with a simplified phonon dispersion [48, 49]. The phonon energies and deformation potentials in use most often are those originally introduced by Brunetti *et al.* [46]. Optical phonon dispersion is ignored and often only one acoustic branch (LA) is considered for intravalley scattering. Such models can lead to unphysical thresholds in the electron distribution function [41] and cannot be used to compute phonon generation rates for detailed phonon dynamics simulations (e.g. phonon Boltzmann transport or Molecular Dynamics). In a realistic electron device a full phonon dispersion is essential for extracting the correct phonon generation spectrum from Joule heating [10]. Use of the full phonon dispersion is also important in strained or confined materials and devices, where the dispersion relationship is altered from its bulk form.

Chapter 3 of this dissertation and Ref. [37] describe the implementation of a MC code which uses analytic descriptions for both the electron bands and the phonon dispersion. In the context of Fig. 2.1, the isotropic analytic phonon model described in this work lies on the vertical axis between the anisotropic bond-charge dispersion method [42, 43], and all other traditional approaches. This computationally efficient method is suitable for simulating low-voltage nanodevices, while treating the electron bands and phonon dispersion with equal attention.

2.2 General Monte Carlo Aspects

The general aspects of the Monte Carlo method for charge transport in semiconductors have been well described before [8, 34, 35]. This section provides but a brief overview of the MC algorithm, summarized with the diagram in Fig. 2.2. The ensemble MC approach used in this work preselects several tens of thousands "super-particles" to represent the mobile charge inside the semiconductor. This number is limited by computational (and to a lesser extent, today, by memory) constraints, but good statistics can be obtained if the simulation

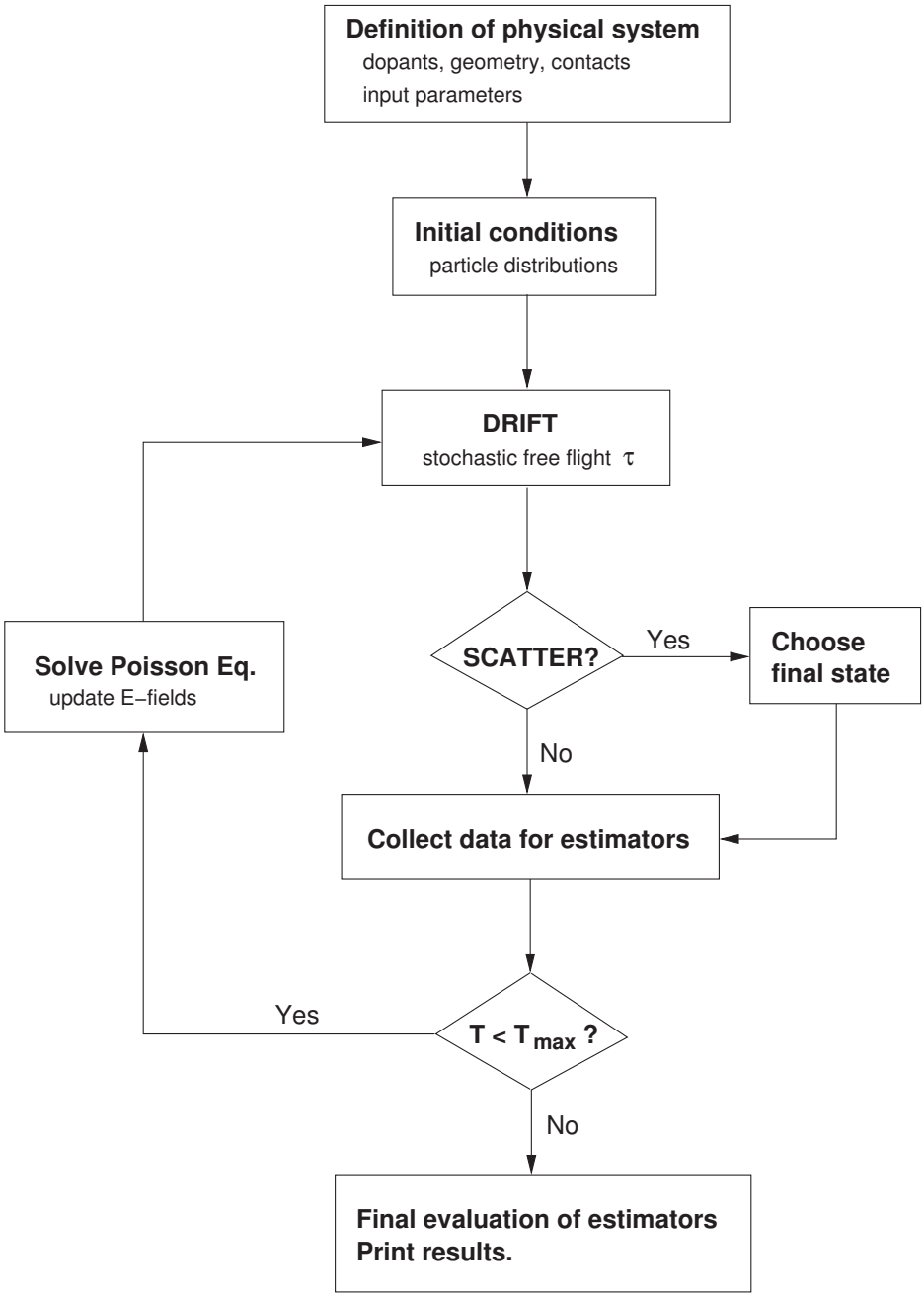

Figure 2.2: Basic Monte Carlo algorithm flowchart.

is run for an adequately long time. The particles are initialized with thermal energy distributions (average energy $3k_BT/2$) and with randomly oriented momenta. Spatially, in the case of a realistic device simulation (as opposed to modeling the transport properties of bulk silicon) the particles are initially distributed following the device doping profile or based on initial conditions read from, for example, a drift-diffusion device simulator. Once the simulation is started, the particles are allowed to drift for short periods of time (τ, shorter than the average time between collisions), then a scattering process (if any) is selected. A fictive "self-scattering" rate can be chosen in such a way that the sum of all scattering rates is constant (Γ_o) and independent of the carrier energy. The distribution of each particle's free flight time intervals (τ) is then directly related to this total scattering rate as [8]

$$\tau = -\frac{1}{\Gamma_o}\ln(r_1) \tag{2.1}$$

where r_1 is a random number uniformly distributed between 0 and 1. During its free flight, the carrier is allowed to drift under the influence of the electric fields, as dictated by Newton's laws of motion with an effective mass (as opposed to the free electron mass) which represents the collective influence of the lattice. Then another random number r_2 between 0 and 1 is drawn[1] and $r_2\Gamma_o$ is compared with cumulative probabilities of scattering which have been precomputed at the beginning of the simulation as a function of energy. A scattering mechanism (e.g. with impurities, acoustic or optical phonons) is selected in proportion to the strength of each process. If self-scattering is selected, the particle continues its free flight unimpeded. If a real scattering process is selected the particle's state after scattering is stochastically chosen taking into account both energy and momentum conservation, then another random time of flight is drawn. This procedure then repeats for all particles.

In the case of a realistic device simulation, the Poisson equation must be solved at every time step, to self-consistently update the electric fields as the mobile charge carriers move inside the device. The Monte Carlo simulation can also be run without the Poisson

[1] It is this stochastic nature of the Monte Carlo simulation method which provides its name, a reference to the gambling opportunities in the eponymous Mediterranean city.

equation, in post-processor mode on the fixed ("frozen") fields initially read from a drift-diffusion simulator, although extensive work has shown [50] that the results are less accurate and predictive, particularly for noise simulations. The super-particles are treated as single carriers during their free flights, and as charge clouds when the Poisson equation is solved. The cloud-in-cell method [35] is most often employed for assigning the super-particle charge to the grid nodes before Poisson's equation is solved. The charge on each super-particle is

$$\mathcal{Q} = e\frac{N}{N_{sim}} \tag{2.2}$$

where e is the elementary charge, N is the total number of mobile charges in the device and N_{sim} is the number of super-particles used in the simulation. It should be noted that the coupled solution to Poisson's equation yields a much more stringent requirement on the simulation time step, necessary to avoid charge imbalance due to plasma oscillations [8]. The Poisson equation therefore ought to be solved every

$$\Delta T < \frac{1}{2}\sqrt{\frac{\epsilon_s m^*}{e^2 n}} \tag{2.3}$$

where ϵ_s is the dielectric constant of the semiconductor, m^* is the lighter effective mass of the carrier in the material (the transverse mass m_t for electrons in silicon) and n is the mobile charge density. In the heavily doped contact regions of a device, where $n \simeq 10^{20}$ cm^{-3}, very short (and therefore time-consuming) time steps less than 1 fs are necessary. The charge density at the device contacts must also be updated at the end of each time step. This is done by injecting (or deleting) thermal electrons at the grid nodes adjacent to the contacts, to maintain charge neutrality there.

Ensemble averages are updated every time step, and statistics are gathered by sampling the super-particle system at regular time intervals, until reaching a targeted accuracy. The error margins are inversely proportional to the square root of the number of particles being simulated, $1/\sqrt{N_{sim}}$. The run of the algorithm ends when the total time alloted for the simulation ends (typically, on the order of tens or hundreds of picoseconds), or when enough

statistics have been gathered and the error margins of the sought-after ensemble averages are deemed appropriate. It should be noted that Monte Carlo simulations are not well-suited for low-field transport, where other, simpler but much faster methods may be preferred (e.g. drift-diffusion). However, the method represents the most physically comprehensive simulation approach for charge transport in semiconductors, and is usually the standard against which all other methods are judged. Several reference works have been dedicated to thorough reviews of the Monte Carlo method [8, 34, 35] and much more information can be gathered therein.

Chapter 3

Analytic Band and Dispersion Monte Carlo Implementation

This chapter describes the implementation of a new Monte Carlo model for electron transport, specifically developed to compute heat (phonon) generation rates in bulk and strained silicon, as well as in simple nanoscale device geometries. The model uses analytic, non-parabolic electron energy bands and an isotropic, analytic phonon dispersion model, which distinguishes between the optical/acoustic and longitudinal/transverse phonon branches. A new, unified set of deformation potentials for electron-phonon scattering is introduced and shown to yield accurate transport simulations (vs. the available data) in bulk *and* strained silicon across a wide range of electric fields and temperatures. The Monte Carlo model is then applied in the context of transport in one-dimensional (self-consistent with the Poisson equation) and two-dimensional device geometries.

3.1 Electron Energy Band Model

This work models the electron energy bands analytically, following Jacoboni [34], and including the non-parabolicity parameter α ($= 0.5$ eV^{-1} at room temperature). With $\alpha = 0$ the kinetic energy is purely parabolic and Canali's original model [38] is recovered. All six

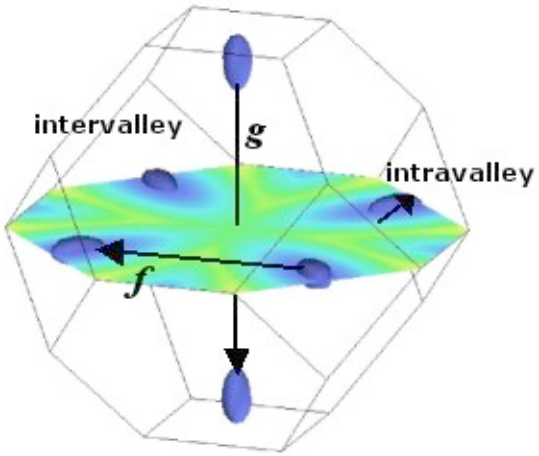

Figure 3.1: Three-dimensional view of the ellipsoidal conduction band valleys of silicon within the first Brillouin zone (momentum space). The arrows represent the type of electron-phonon scattering transitions, i.e. f- and g-type intervalley scattering, as well as intravalley scattering. (Original figure courtesy C. Jungemann.)

ellipsoidal, energetically equivalent conduction band valleys of silicon are explicitly included, as in Fig. 3.1. The non-parabolic band approximation represents a good description of electron transport at energies below approximately 1 eV, such as those of future low-voltage nanotechnologies, where impact ionization and high energy transport are not expected to play a significant role. Figure 3.2 shows a comparison between the total conduction band density of states (DOS) computed in the non-parabolic band approximation and the full band density of states. From the point of view of the DOS, which determines the scattering rates (as described in Section 3.3), the analytic band approximation is sufficient up to 1.5 eV in electron energy. This analytic, non-parabolic relationship between the electron energy E_k and the wave vectors k_i (i =1, 2 or 3, for the three Cartesian axes) is:

$$E_k(1+\alpha E_k) = \frac{\hbar^2}{2}\sum_{i=1}^{3}\frac{(k_i-\kappa_{vi})^2}{m_i} \tag{3.1}$$

where m_i is the component of the electron mass tensor along the i^{th} direction and κ_{vi} represents the coordinates of the respective conduction band minimum. Silicon has six equivalent conduction band minima near the X symmetry points, located at $\pm$85 percent of

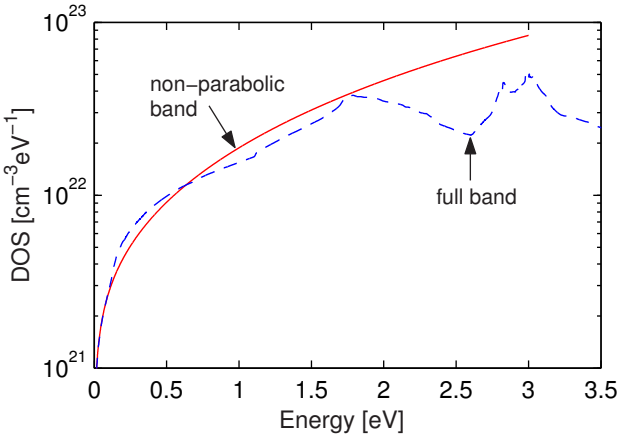

Figure 3.2: Conduction band density of states (DOS) in silicon from a full band calculation (courtesy C. Jungemann) vs. the DOS computed with the non-parabolic band approximation from Eq. 3.7.

the way to the edge of the Brillouin zone, along the three $\langle 100 \rangle$ axes (the Δ lines), as shown in Fig. 3.1. For example, the X-valley (sometimes also called the Δ-valley) along the $\langle 100 \rangle$ direction is centered at $(0.85, 0, 0)$G where $|\mathbf{G}| = 2\pi/a$ is the reciprocal lattice vector and $a = 5.431$ Å is the silicon lattice constant. The mass tensor components are the longitudinal mass $m_l/m_o = 0.916$ and the transverse mass $m_t/m_o = 0.196$ at room temperature, where m_o is the mass of the free electron. The temperature dependence of the band gap $E_g(T)$ is also included analytically, following the review of Green [51]:

$$E_g(T) = 1.1756 - 8.8131 \times 10^{-5} T - 2.6814 \times 10^{-7} T^2 \tag{3.2}$$

where T is the absolute temperature in degrees Kelvin. This dictates a slight temperature dependence of the transverse mass as $m_t/m_o = 0.196 E_{go}/E_g(T)$ and of the non-parabolicity parameter as $\alpha = 0.5 E_{go}/E_g(T)$ eV^{-1}, where E_{go} is the silicon band gap at room temperature [51]. Figure 3.3 shows a typical "snapshot" of the electron distribution in momentum space, as represented by the current work.

Since it uses the analytic non-parabolic band approximation, suitable for low-energy studies, the present work ignores the second conduction band (the L-valley) of silicon, which lies slightly more than 1 eV above the bottom of the X-valley. Hence the maximum

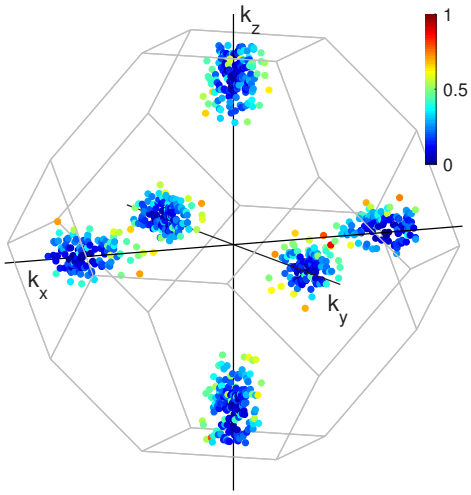

Figure 3.3: Electron distribution in momentum space, for an electric field of 50 kV/cm in the ⟨111⟩ direction, at 300 K. The color bar represents the electron energy, from 0 to 1 eV.

electron energy is usually limited to (or slightly above) 1 eV during the simulation. In fact, this upper limit on the electron energy is a convenient value for several reasons: (a) it is the approximate energy difference between the L and X valleys, (b) it is also about the value of the silicon energy band gap, which controls impact ionization, and (c) it conveniently corresponds to an upper limit on the maximum electron energy in low-voltage future nanodevices, which will operate at 1 Volt or below. Since electrons with energies larger than the band gap will be rare, impact ionization is not expected to play a significant role and consequently it can be safely neglected.

3.2 Phonon Dispersion Model

The present work treats all phonon scattering events inelastically, hence the electrons exchange the correct amount of energy (corresponding to the absorption or emission of a phonon) with each scattering event. Particular attention is paid to the treatment of inelastic acoustic phonon scattering, to properly account for energy dissipation at low temperatures and low electric fields. Treating the acoustic phonons inelastically is also important for heat

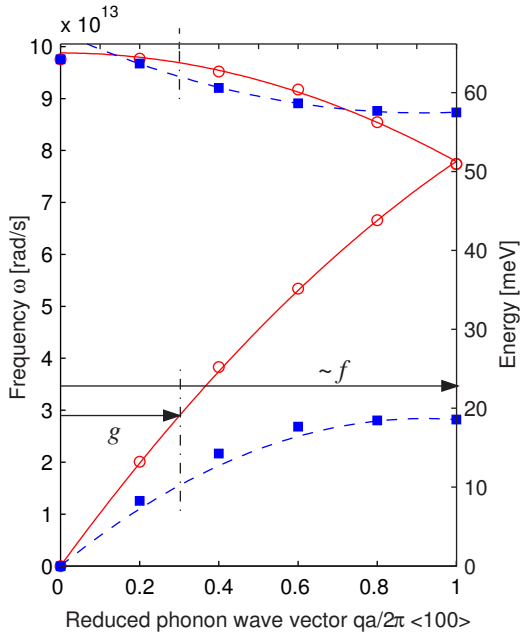

Figure 3.4: Phonon dispersion in silicon along the ⟨100⟩ direction, from neutron scattering data (symbols) [54]. The lines represent the quadratic approximation introduced in Ref. [37] and this work. The f and g phonons participate in the intervalley scattering of electrons [52].

generation spectrum calculations, as shown in Chapter 4 and Ref. [10]. Figure 3.1 illustrates the ellipsoidal conduction band valleys and the allowed phonon scattering transitions. As in the traditional analytic-band approach [34], scattering with six types of intervalley phonons is incorporated. Intervalley scattering can be of g-type, when electrons scatter between valleys on the same axis, e.g. from ⟨100⟩ to ⟨−100⟩, or of f-type when the scattering occurs between valleys on perpendicular axes, e.g. from ⟨100⟩ to ⟨010⟩. The phonons involved in these scattering transitions (three of f-type and three of g-type) can be determined from geometrical arguments [52] and are labeled in Fig. 3.4. Intravalley scattering refers to scattering within the same conduction band valley and usually involves only acoustic phonons [53].

Most typical MC codes [34, 40, 41, 44, 45, 46], both analytic- and full-band, treat intravalley scattering with a single kind of acoustic phonon. This simplification is accomplished by grouping the longitudinal acoustic (LA) and transverse acoustic (TA) branches

	ω_o	v_s	c
	10^{13} rad/s	10^5 cm/s	10^{-3} cm^2/s
LA	0.00	9.01	-2.00
TA	0.00	5.23	-2.26
LO	9.88	0.00	-1.60
TO	10.20	-2.57	1.12

Table 3.1: Quadratic phonon dispersion coefficients for each branch of the phonon spectrum: longitudinal acoustic (LA), transverse acoustic (TA), longitudinal optical (LO) and transverse optical (TO).

into a dispersionless mode with a single velocity and a single deformation potential. Historically, TA modes have been neglected because their matrix element is zero for intravalley scattering within a band located at the center of the Brillouin zone [34, 53]. This isn't the case for silicon, hence in a more comprehensive approach (where scattering with *all* phonon modes matters) intravalley scattering with TA modes should be considered. Unlike the traditional approach, this work considers scattering with LA and TA modes separately. Each phonon dispersion branch from Fig. 3.4 (including the optical modes) is treated with the isotropic approximation

$$\omega_q = \omega_o + v_s q + c q^2 \tag{3.3}$$

where ω_q is the phonon frequency and q the wave vector. For the acoustic phonons, the parameters v_s and c can be chosen to capture the slope of the dispersion near the Brillouin zone center and the maximum frequency at the zone edge, similar to Ref. [41]. The choice of parameters for longitudinal optical (LO) phonons insures that they meet the zone edge LA frequency. For both TA and transverse optical (TO) phonons the zone edge slope, i.e. their group velocity is fit to zero. The continuous (longitudinal) and dashed (transverse) lines in Fig. 3.4 represent these quadratic approximations, and the fitting coefficients are listed in Table 3.1. Quartic polynomials would offer a better fit in the $\langle 100 \rangle$ crystal direction but no advantage in the other directions, hence the quadratics are entirely sufficient for this isotropic approximation. They track the phonon dispersion data closely, especially in

the regions relevant to electron-phonon scattering in silicon: near the Brillouin zone center for long wavelength intravalley acoustic phonons, and near the frequencies corresponding to intervalley f- and g-type phonons. The quadratics are also easy to invert and, where needed, to extract the phonon wave vector as a function of frequency.

The same approach can be used to extend this phonon dispersion model to other materials or confined dimensions. Changes in the phonon dispersion due to strain or confinement (e.g. in nanostructures) can be easily included. The challenge in this case lies chiefly in determining the correct modified phonon dispersion to use in such circumstances. The electron-phonon scattering rates need to be numerically recomputed with the modified phonon description (as outlined below), which can be done efficiently if the dispersion is written as a set of analytic functions, like the polynomials in this work.

3.3 Electron-Phonon Scattering

Scattering by lattice vibrations (phonons) is one of the most important processes in the transport of carriers through a semiconductor. It is this scattering that limits the velocity of electrons in the applied electric field, and from this point of view transport can be seen as the balance between accelerative forces (the electric field) and dissipative forces (the scattering). The treatment of electron-phonon scattering in Monte Carlo simulations is based on the assumption that lattice vibrations cause small shifts in the energy bands, and this additional potential U causes the scattering process, with the matrix element

$$M(\mathbf{k}, \mathbf{k}') = \langle \mathbf{k}' | \, U \, | \mathbf{k} \rangle \tag{3.4}$$

between the initial state $\mathbf{k}$ and the final state $\mathbf{k}'$ [8, 14]. This matrix element contains the momentum conservation condition, $\mathbf{k}' = \mathbf{k} \pm \mathbf{q} + \mathbf{G}$, where $\mathbf{q}$ is the phonon wave vector, $\mathbf{G}$ is a reciprocal lattice vector, and the upper and lower signs correspond to the absorption and emission of a phonon. The electronic wave functions are typically taken to be Bloch functions that exhibit the periodicity of the lattice. The electron-phonon scattering

rate is based on Fermi's Golden Rule, which is derived from first-order time-dependent perturbation theory [8, 35] and gives the transition probability between the two eigenstates

$$P(\mathbf{k},\mathbf{k}') = \frac{2\pi}{\hbar}\left|M(\mathbf{k},\mathbf{k}')\right|^2 \delta(E_k - E_{k'} \pm \hbar\omega_q) \tag{3.5}$$

where the upper and lower signs have the same meaning as in the previous paragraph. It is assumed that the scattering potential is weak, such that it can be treated as a perturbation of the well-defined energy bands, and the δ-function ensures that two collisions do not "overlap" in space or in time, i.e. they are infrequent, or that the scattering time is much shorter than the time between collisions. The total scattering rate out of state $\mathbf{k}$ is obtained by integrating over all final states $\mathbf{k}'$ the electron can scatter into. Mathematically, this integration can be carried out over $\mathbf{k}'$ or $\mathbf{q}$ with the same result [14]. In those cases in which the matrix element is independent of the phonon wave vector, the matrix element can be removed from the integral, which leaves a total scattering rate directly dependent on the density of states:

$$\Gamma(\mathbf{k}) = \frac{2\pi}{\hbar}\left|M(\mathbf{k})\right|^2 g_d(E_k \pm \hbar\omega_q), \tag{3.6}$$

where $M(\mathbf{k})$ includes the dependence on the phonon occupation of states, on the wave function overlap integral and on the deformation potential characteristic of the particular phonon involved.[1] The dependence of the total scattering rate on the density of final states has a satisfying interpretation [14], as it gives us a means for comparing scattering rates in 1-, 2- or 3-dimensional systems. In three dimensions the electron-phonon scattering rate increases roughly as the square root of the electron energy, just like the density of states[2]

$$g_d(E_k) = \frac{(2m_d)^{3/2}}{2\pi^2\hbar^3}\sqrt{E_k(1+\alpha E_k)}(1+2\alpha E_k), \tag{3.7}$$

[1] As will be shown in the subsequent sections, deformation potentials are typically extracted empirically from comparison with low and high temperature electron mobility data.

[2] This is the density of states per energy ellipsoid in silicon, including the factor of 2 for spin. Note this must be multiplied by a factor of six for all conduction band ellipsoids in silicon.

written here in the non-parabolic, analytic band approximation (Eq. 3.1) adopted in this work, where $m_d = (m_t^2 m_l)^{1/3}$ is the electron density of states effective mass.

3.3.1 Intravalley Scattering

Intravalley scattering refers to scattering within the same conduction band valley and it usually involves only acoustic phonons [53]. In this work, the total intravalley scattering rate is calculated separately with LA and TA phonons, as a function of the initial electron energy E_k:

$$\Gamma_i(E_k) = \frac{D_a^2 m_d}{4\pi\rho\hbar^2 k_s} \int_q \frac{1}{\omega_q} \left(N_q + \frac{1}{2} \mp \frac{1}{2} \right) \mathcal{I}_q^2 q^3 dq \tag{3.8}$$

where D_a is the respective deformation potential (D_{LA} or D_{TA}), and ρ is the mass density of silicon. The top and bottom signs refer to phonon absorption and emission, respectively. The electron wave vector is transformed to spherical Herring-Vogt [34, 55] space as:

$$k_s = \frac{\sqrt{2m_d E_k(1+\alpha E_k)}}{\hbar} \tag{3.9}$$

Since the scattering rates are numerically integrated at the beginning of the simulation, the correct phonon occupation can be incorporated as

$$N_q = \frac{1}{\exp(\hbar\omega_q/k_B T) - 1}, \tag{3.10}$$

without resorting to the equipartition or Joyce-Dixon approximations normally used [34]. The wave function overlap integral is included in the rigid ion approximation [56]:

$$\mathcal{I}_q = \frac{3}{(qR_s)^3} \left[\sin(qR_s) - qR_s \cos(qR_s) \right] \tag{3.11}$$

where $R_s = a[3/(16\pi)]^{1/3}$ is the radius of the spherical Wigner-Seitz cell, $R_s = 2.122$ Å for silicon. All quantities are numerically evaluated using the corresponding phonon dispersion. The scattering rate integral in Eq. 3.8 is carried out over all phonon wave vectors $\mathbf{q}$ that conserve both energy ($E'_k = E_k \pm \hbar\omega_q$) and momentum ($\mathbf{k}' = \mathbf{k} \pm \mathbf{q}$). These arguments can

be used to establish the range of q, as required by $|\cos(\phi)| \leq 1$ where

$$\cos(\phi) = \mp\frac{q}{2k_s} + \frac{m_d\omega_q}{\hbar q k_s}[1 + \alpha(2E_k \pm \hbar\omega_q)] \tag{3.12}$$

and ϕ is the angle between the phonon and the initial electron wave vector. As in the rest of this chapter, the top and bottom signs refer to phonon absorption and emission, respectively. The intravalley scattering rate typically cited in the literature [34] can be recovered by substituting the simple, dispersionless phonon frequency $\omega_q = v_s q$ (typically for LA phonons only), $\mathcal{I}_q = 1$ and using an approximation for N_q, which allows Eq. 3.8 to be integrated analytically.

The final state of the electron after scattering $|E'_k, \mathbf{k}'\rangle$ reflects both the energy and momentum exchange with the phonon, as follows: first the magnitude of the phonon wave vector $\mathbf{q}$ is selected within the allowed range using a rejection algorithm [34] applied to the integrand in Eq. 3.8, which includes the overlap integral. Then the magnitude of the electron wave vector $\mathbf{k}'$ after scattering is found by energy conservation, while the angle between $\mathbf{k}'$ and $\mathbf{k}$ is obtained by momentum conservation. The final electron state is only accepted if it falls within the first Brillouin zone, otherwise the rejection algorithm is repeated.

The intravalley deformation potentials have a general angular dependence which can be written as [55]:

$$\Xi_{LA}(\theta) = \Xi_d + \Xi_u \cos^2\theta \tag{3.13}$$

$$\Xi_{TA}(\theta) = \Xi_u \sin\theta\cos\theta \tag{3.14}$$

where θ is the angle between the phonon wave vector and the longitudinal axis of the conduction band valley, Ξ_u is the shear and Ξ_d is the dilation deformation potential. Detailed calculations have shown that the influence of this angular dependence on the electron transport is relatively small [57]. Hence the intravalley deformation potentials can be averaged over the angle θ, consistently with the general isotropic approach adopted in this work. The

isotropically averaged deformation potentials become

$$D_{LA} = \sqrt{\frac{\pi}{2}\left(\Xi_d^2 + \Xi_d\Xi_u + \frac{3}{8}\Xi_u^2\right)} \quad (3.15)$$

$$D_{TA} = \frac{\sqrt{\pi}}{4}\Xi_u \quad (3.16)$$

which are used for computing the intravalley scattering rates in Eq. 3.8. There is considerable variation in the values of the shear (Ξ_u) and dilation (Ξ_d) deformation potentials reported in the literature over the years. A good summary of these values can be found in Ref. [58]: various theoretical and empirical studies have estimated Ξ_u in the range of 7.3 to 10.5 eV, while Ξ_d has been previously cited both as -11.7 eV (Ref. [59]) and near 1.1 eV (Ref. [58]). Although, perhaps surprisingly, both values can be used to describe electron mobility (hence the original confusion over the correct choice), it was shown that only the latter ($\Xi_d = 1.1$ eV) yields the correct mobilities both for electrons and holes [58]. This is the value adopted in the current study. Then Ξ_u is used as a fitting parameter while calculating the low-field, low-temperature (T=77 K) electron mobility, a regime dominated by scattering with intravalley phonons. An empirical best-fit value of $\Xi_u = 6.8$ eV is found, in reasonable agreement with previous work. With these values of Ξ_d and Ξ_u the isotropically averaged deformation potentials are $D_{LA} = 6.39$ eV and $D_{TA} = 3.01$ eV. These are comparable with the value of 9 eV typically cited in the literature for MC models where scattering is only taken into account with the longitudinal modes [34].

3.3.2 Intervalley Scattering

As outlined in Section 3.2, intervalley scattering in silicon can take electrons between equivalent (g-type) and non-equivalent (f-type) valleys. Based on geometrical arguments [52], both f- and g-type scattering are Umklapp processes, involving a reciprocal lattice vector $|\mathbf{G}| = 2\pi/a$. Since the X-valley minima are located at 0.85 from the center to the edge of the Brillouin zone, the change required in electron momentum is $(0, 0.85, 0.85)$G for f-type scattering and $(1.7, 0, 0)$G for g-type scattering. Reduced to the first Brillouin zone, the

				Deformation potentials Δ_{if} (10^8 eV/cm)				
	Type	T (K)	E (meV)	Ref. [38]	Ref. [61]	Ref. [46]	Ref. [44]	This work
f_1	TA	220	19	0.15	–	0.3	2.5	0.5
f_2	LA/LO	550	51	3.4	4.3	2	–	3.5
f_3	TO	685	57	4	2	2	8	1.5
g_1	TA	140	10	0.5	0.65	0.5	–	0.3
g_2	LA	215	19	0.8	–	0.8	4	1.5
g_3	LO	720	62	3	7.5	11	8	6

Table 3.2: Summary of phonon energies and deformation potentials for intervalley electron-phonon scattering in silicon.

phonons involved are $(1, 0.15, 0.15)$G and $(0.3, 0, 0)$G respectively [52, 60]. The f-phonon is just 11° off the $\langle 100 \rangle$ direction, while the g-phonon is along $\langle 100 \rangle$, at 0.3G. These phonons are schematically drawn on the dispersion relation in Fig. 3.4. The g-phonon frequencies can be directly read off the $\langle 100 \rangle$ dispersion, while the f-phonons are typically assumed to be those at the edge of the Brillouin zone. In this work, ω_q is computed from the analytic phonon dispersion and the intervalley scattering rate between the initial (i) and final (f) valley can be written as follows [8, 34]:

$$\Gamma_{if}(E_k) = \frac{\pi \Delta_{if}^2 Z_f}{2\rho\omega_q} \left(N_q + \frac{1}{2} \mp \frac{1}{2} \right) g_{df}(E_k \pm \hbar\omega_q) \tag{3.17}$$

where Z_f is the number of available final valleys (4 for f-type and 1 for g-type scattering), $g_{df}(E_k)$ is the density of states in the final valley (Eq. 3.7), and other symbols are the same as previously defined. Intervalley scattering can also include an overlap factor, but its value is typically incorporated into the scattering constant Δ_{if}. The six phonons involved in intervalley scattering, along with their approximate energies, equivalent temperatures (as $T = \hbar\omega_q / k_B$), and deformation potential scattering constants are listed in Table 3.2.

Traditional MC models (apart from the *ab initio* approaches of Refs. [42] and [43]) assume the phonon energies involved in intervalley scattering are fixed at the values determined by transitions between the X-valley minima. Also, the state of the electron in the

final valley is computed isotropically [34]. These geometrical arguments only hold strictly for the lowest energy electrons at the bottom of the bands. This work takes into account the phonon dispersion for scattering with both optical and acoustic phonons when calculating the final state of the electron. After the type of intervalley scattering mechanism is determined, the state of the electron in the final valley is first chosen isotropically, as in the traditional approach. The phonon wave vector necessary for this transition can be calculated as $\mathbf{q} = \mathbf{k}' - \mathbf{k}$ because the initial state of the electron is known. The phonon is then reduced to the first Brillouin zone and its energy is obtained using the phonon dispersion described earlier. This procedure is applied to both acoustic and optical phonons. The phonons that do not satisfy both energy and momentum conservation within a certain tolerance are discarded with a rejection algorithm. This is a relatively inexpensive search which ends when a suitable phonon is found. The effect of this algorithm is to smear out any "hard" thresholds associated with intervalley phonon energies in the electron distribution. Figure 3.5 shows the low field (200 V/cm) electron distribution computed with this approach, compared to the typical models found in the literature [34, 46] where the intervalley phonon dispersion is not taken into account. Any unphysical threshold, e.g. at 62 meV due to g-type optical intervalley scattering, is removed when phonons of varying energies around this value (as given by the dispersion relation and by energy and momentum conservation) are allowed to participate. Such thresholds in the electron distribution are also present in full band MC models which use a single, fixed energy optical phonon [41]. The current model removes them in a computationally inexpensive way, while satisfying energy and momentum conservation for all scattering events.

Despite the added complexity of the full phonon dispersion, this analytic band code is more than an order of magnitude faster when compared to typical full band programs (using a simpler phonon description) doing the same velocity-field curve calculations, i.e. Fig. 3.7. A version of the code compiled using fixed phonon energy values and without the dispersion information (essentially identical to the one of Ref. [34]) was only a few percent faster than our model which includes the dispersion. Hence this work incorporates the phonon dispersion in an efficient way, giving significantly more physical insight than the

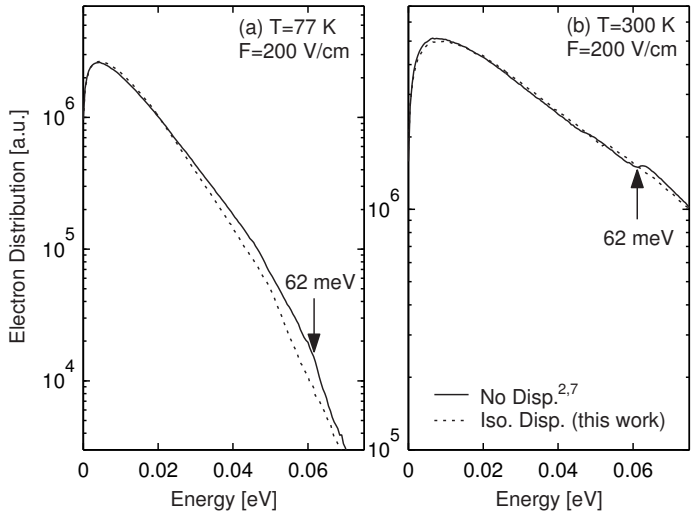

Figure 3.5: Electron distribution vs. energy at (a) 77 K and (b) 300 K with low applied electric field (200 V/cm). The typical dispersionless model [34, 46] is compared with the results of this work, which include the full isotropic dispersion. Note the vertical axes are not at the same scale.

typical analytic band code for very little computational overhead, while still being more than an order of magnitude faster than a typical full band code. The analytic phonon dispersion and analytic electron bands significantly speed up the calculations of the final electron state after scattering, compared to the look-up tables and interpolation schemes found in full band codes. Further speed improvements can be obtained by including an energy-dependent total scattering rate [62], which would significantly reduce the number of self-scattering events.

3.4 Electron-Ionized Impurity Scattering

Ionized impurity scattering must be taken into account for electron transport through the heavily doped regions (e.g. source or drain) of realistic devices. Unlike phonon scattering, ionized impurity scattering is an elastic process, meaning that it does not change the energy of the electron. However, the scattered electron momentum is altered, as indicated by the effect ionized impurities have on the electron mobility. The scattering potential due to an

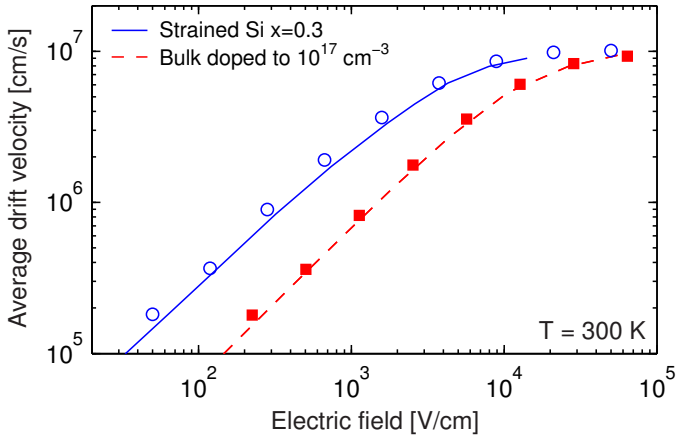

Figure 3.6: Electron velocity-field relationship in doped bulk and strained silicon. The dashed lines represent data for 10^{17} cm^{-3} doped bulk silicon, the solid lines are data for strained silicon on $x = 0.3$ substrate Ge fraction [66]. The symbols are our simulation results for the two respective cases.

impurity charge in a crystal is a screened Coulomb potential

$$U(r) = \frac{Ze^2}{4\pi\epsilon_s r} \exp(-r/L_D) \tag{3.18}$$

depending on how many free charge carriers are present. Here Ze is the net extra charge on the impurity atom,[3] ϵ_s is the dielectric constant of the semiconductor, r is the distance from the scattering center and $L_D = \sqrt{\epsilon_s k_B T/(e^2 n)}$ is the Debye length. where n is the free charge carrier (electron) density responsible for screening the potential in Eq. 3.18. Impurity scattering is a highly anisotropic process [8, 63], showing a strong preference for small scattering angles. Although physically sound, a direct implementation of this approach in a Monte Carlo technique would yield several problems. Many small-angle scattering events would have to be processed consuming computational time. Also, many short free-flight times would be obtained, further degrading the efficiency of the procedure. The scattering model proposed by Kosina [64, 65] avoids such pitfalls by reformulating impurity scattering as an isotropic process with the same momentum relaxation time as the anisotropic process. This work implements Kosina's model, including the screening function from Ref. [64]. The

[3]For example, $Z = 1$ for n-type dopants from Group V, like Arsenic or Phosphorus.

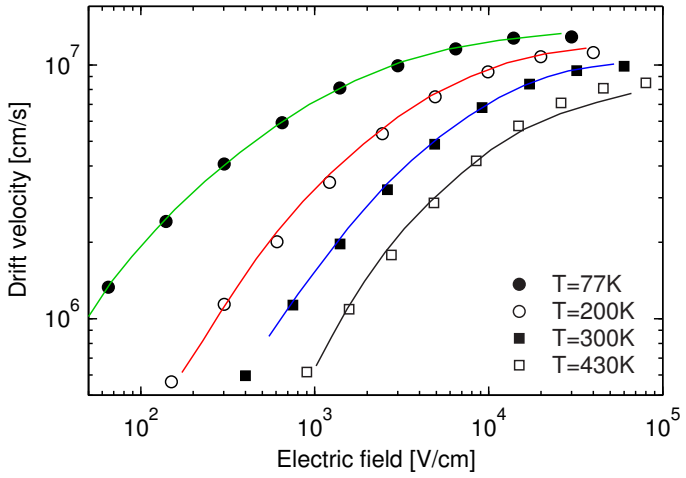

Figure 3.7: Electron drift velocity vs. electric field in unstrained silicon over a wide range of temperatures. Symbols are the Monte Carlo simulations of this work. The lines represent the time of flight experimental data of Canali *et al.* [38].

model has been shown to be adequate for doping concentrations up to 10^{20} cm^{-3}, with particularly notable improvements in efficiency at lower (less than 10^{17} cm^{-3}) doping levels. The dashed line and solid symbols in Fig. 3.6 show a comparison between velocity-field data obtained in 10^{17} cm^{-3} doped bulk silicon and Monte Carlo simulations using the isotropic scattering model. Good agreement is found over a wide range of electric fields. Similarly, the low-field mobility was computed over a wide range of doping densities and good agreement was found with available experimental data.

3.5 Transport Applications

Electron transport characteristics, as well as energy dissipation at moderate to high fields and for all but the lowest temperatures are determined by the choice of intervalley phonon coupling constants Δ_{if}. This choice also determines the relative strength of f- and g-type intervalley scattering. Several sets of coupling constants previously proposed are listed in Table 3.2. The parameter set introduced by Brunetti *et al.* [46] has been most commonly used in the literature over the past two decades, for both analytic and full band simulations. This parameter set strongly favors g-type scattering with the 62 meV LO phonon, while the

	Low Field		High Field	
	Ac (Ξ_a) $\uparrow$	Op (Δ_{if}) $\uparrow$	Ac (Ξ_a) $\uparrow$	Op (Δ_{if}) $\uparrow$
$\overline{E}$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
$\overline{v}$	$\downarrow$	$\downarrow$	$\downarrow$	$\Uparrow$

Table 3.3: Transport dependence on optical and acoustic scattering potentials. In general, increasing any coupling constant will decrease both the ensemble electron velocity ($\overline{v}$) and energy ($\overline{E}$), both in the low-field and high-field region. However, the average velocity has an opposite dependence on the optical intervalley scattering constants in the high-field region, as shown with the double arrow.

original set of Canali *et al.* [38] tends to favor f-type scattering with zone edge phonons. It should be noted that the zone edge f-phonon at 51 meV is typically classified as part of the LA branch, but this scattering can also happen with LO phonons [57], because the two branches meet at the zone boundary.[4] Since the current work takes into account both acoustic and optical dispersion, when this f-type scattering event is selected the participating phonon is assigned to the LO branch if $|\mathbf{k}' - \mathbf{k}| > 2\pi/a$, and to the LA branch otherwise.

3.5.1 Bulk Silicon Mobility

The intervalley scattering constants for the current model are derived starting from the set of Brunetti *et al.* [46]. To aid with parameter extraction, an inverse modeling code originally developed for doping profile extraction was modified and used [67]. The intervalley scattering parameters were extracted over a wide range of temperatures and electric fields, by comparison with available transport data [38, 51]. Low energy phonons typically control the low field and low temperature mobility. Increased coupling constants with low energy phonons lead to lower drift velocities and lower electron energies, both in the low field (linear) and the high field (saturation) region. The effect is the same in the low field region when increasing the high energy (optical or f-type LA) coupling constants. On the other hand, increased coupling constants with high energy phonons leads to *higher* drift velocities in the high field region, while the average energy decreases. In other words, cooling the electron distribution through high energy phonon emission leads to higher velocities because

[4] See the dispersion relationship in Fig. 3.4.

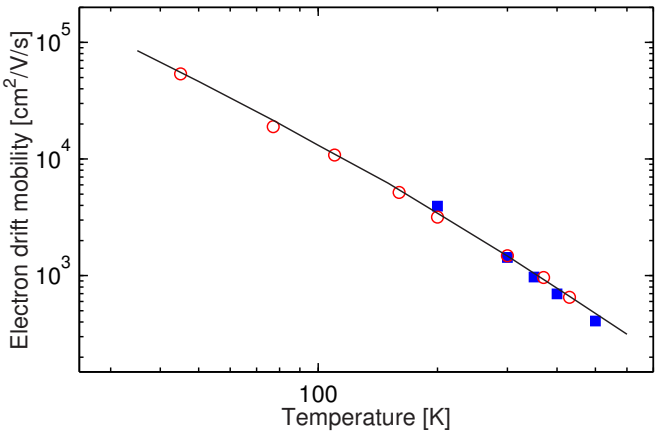

Figure 3.8: Electron drift mobility simulation and data over a wide range of temperatures. Open symbols are data from Canali [38], closed symbols are data from Green [51]. The solid line was simulated with the current Monte Carlo method.

at higher energies the electron velocity is curtailed by non-parabolicity, which increases the effective mass by a factor of $(1 + 2\alpha E_k)$ [8]:

$$v_i = \frac{\hbar k_i}{m_i(1 + 2\alpha E_k)} \tag{3.19}$$

for each component of velocity, wave vector and effective mass along the i^{th} Cartesian direction, as in Eq. 3.1. The low and high energy intervalley coupling constants have the same effect on the drift velocity at low fields, but opposing effects at high fields. The qualitative dependence of velocity and energy on the scattering constants is illustrated in Table 3.3. This opposite dependence of velocity on low and high energy intervalley phonons determines the "shape" (i.e. the "bend") of the velocity-field curves (see Fig. 3.7) and can be used to fine-tune the coupling constants. Since phonons involved in intervalley scattering have different energies, the inverse modeling method can distinguish between the contribution of the various parameters to the velocity-field curves. Most notably, a smaller contribution of the g-type LO phonon is found, with a deformation potential approximately 40 percent lower than the value reported by Brunetti *et al.* [46] (see Table 3.2). For f-type scattering the deformation potential of LA/LO is found to be stronger than that of TO phonons, which is consistent with *ab initio* calculations [57].

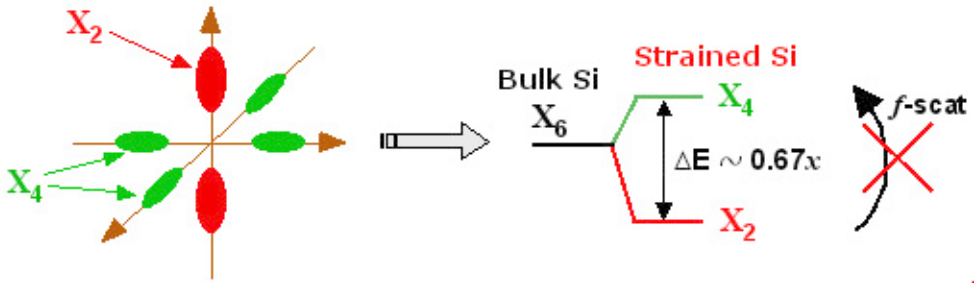

Figure 3.9: Conduction band degeneracy splitting due to strain. The band splitting is proportional to the fraction x of Ge in the $Si_{1-x}Ge_x$ buffer substrate. A large enough splitting ($x > 0.15$) will almost completely suppress f-type intervalley scattering between the two lower (X_2) and four upper (X_4) valleys.

The temperature dependence of the low field mobility can be used to fine-tune the low energy intervalley phonon parameters, assuming impurity scattering can be neglected. Figures 3.7 and 3.8 show the results of transport simulations using the current set of parameters, which are listed in Table 3.2. Note the wide range of electric fields and temperatures (from 30 K to 600 K) covered by the simulations and their comparison with the transport data. The current model agrees with this data well within experimental error.

3.5.2 Strained Silicon Mobility

Strained silicon transport data was not available when the original sets of intervalley coupling constants listed in Table 3.2 were proposed. As the technology for growing defect-free strained silicon layers on top of $Si_{1-x}Ge_x$ buffers was perfected, record mobilities have been measured. Electron drift mobilities near 3000 $cm^2/V\,s$ at room temperature have been reported in strained silicon modulation-doped structures (MODFETs) [66, 68]. These mobilities are phonon-limited at room temperature, since remote impurity scattering only plays a role at much lower temperatures. Also, the lattice-matched strained silicon layer guarantees a lack of surface roughness scattering on both its sides, unlike in strained MOS inversion layers, where surface scattering with the oxide interface dominates. It is this lack of direct impurity scattering and of interface scattering which makes such modulation-doped

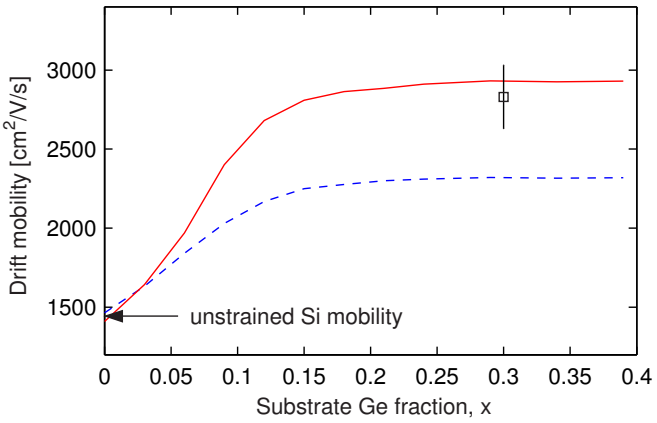

Figure 3.10: Room temperature electron mobility in strained silicon grown on $Si_{1-x}Ge_x$. Mobilities computed with this model (solid line), with the parameter set of Ref. [46] (dashed line) and the record phonon-limited mobility data from Ismail, Nelson and co-workers [66, 68].

structures ideal for exploring phonon scattering in strained silicon. The high mobilities observed in such MODFETs cannot be explained with the intervalley scattering parameters of Ref. [46], as they require a stronger f-type intervalley coupling [58].

Incorporating strained silicon in the MC simulation is relatively straightforward. The biaxial strain removes the degeneracy of the conduction band, lifting four of the six X-valleys by $\Delta E \simeq 0.67x$ where x is the Ge fraction in the $Si_{1-x}Ge_x$ buffer substrate [44]. The in-plane conductivity effective mass of the two lower valleys is the lighter transverse mass m_t of silicon. The difference in energy between the non-equivalent valleys also means that f-type intervalley scattering is strongly reduced as the fraction x increases. For $x > 0.15$, the energy splitting is large enough to almost completely suppress f-type scattering between the lower and upper valleys at room temperature, and the strained silicon mobility enhancement is dominated by conduction via the two lower valleys with the lighter transverse mass. This explains the apparent "saturation" of the mobility for values of $x > 0.15$ in Fig. 3.10. It should be noted that the transverse electron mass is slightly increased by the presence of strain [69], e.g. $m_t = 0.199m_o$ at $x = 0.3$, and this is taken into consideration in the current model. The mobility enhancement of strained compared to that of unstrained silicon is illustrated in Fig. 3.10. The "usual" parameter set [46] cannot account for the mobil-

ity enhancement observed experimentally. The strained mobility data suggests a stronger coupling with f-phonons, and consequently a weaker g-phonon coupling. This, along with the fine tuning explained earlier, ultimately narrows down our choice of parameter sets to that listed in the last column of Table 3.2, which was used to generate all figures. The current choice of intervalley parameters is also in close agreement with recently reported deformation potentials [53] from comprehensive theoretical calculations.[5]

3.6 One-Dimensional Device Applications

In the ensemble Monte Carlo method for device simulation, several things must be taken into account in addition to the ensemble Monte Carlo method for bulk semiconductors (described in the previous sections). One is that the motion of the particles is spatially restricted to the device domain, hence suitable boundary conditions must be set up. Another is that the impurity concentration, and hence the impurity scattering rate is dependent on position, i.e. on the doping profile. Finally, the electric fields must be updated self-consistently with the motion of the charged particles, through repeated solutions of the Poisson equation (at every time step) with appropriate boundary conditions, which are consistent with the boundary conditions applied to the carrier dynamics.

The most frequently studied, realistic, one-dimensional device in the Monte Carlo and device transport community is the n^+nn^+ structure, sometimes referred to as a "ballistic" diode [70, 71]. The energy band diagram of the ballistic diode is such that it represents a simple model for a cross-section through the channel of a MOSFET transistor, as shown in Fig. 3.11. The n^+nn^+ band diagram has similar features, like the voltage-controlled injection barrier at the beginning of the "channel," followed by a steep drop in potential (i.e. highly peaked lateral electric field). Charge transport may be quasi-ballistic across the channel region, provided it is short enough compared to the electron mean free path. This device structure is also ideal as a testbed for the comparison of various simulation approaches (e.g.

[5] Good agreement with the theoretical results was serendipitous, as the current empirical set of deformation potentials had already been settled upon when the results of the *ab initio* calculations came to our attention.

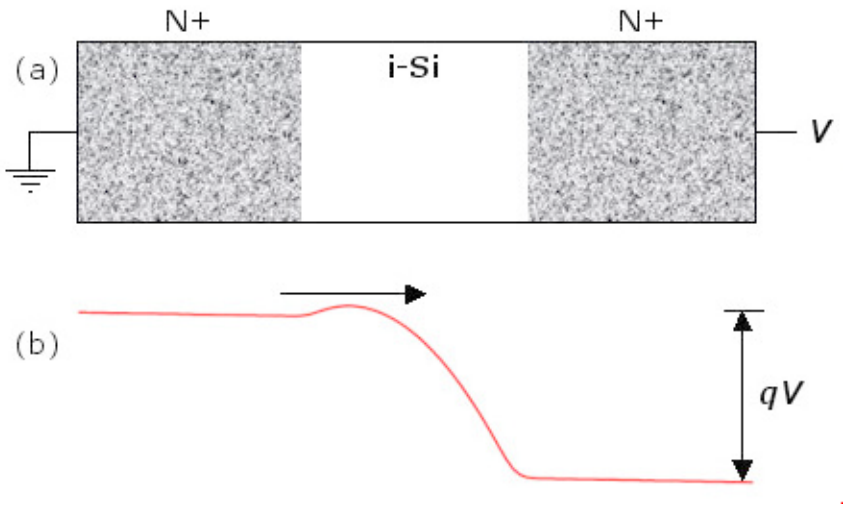

Figure 3.11: Ballistic diode physical structure (a) and energy band diagram (b). The "source" and "drain" end regions are heavily doped ($10^{19} - 10^{20}$ cm^{-3}) whereas the middle region is lightly doped (e.g. 10^{16} cm^{-3}) or almost intrinsic ("i"). This yields the band diagram in subplot (b), which is similar to that along the channel of a MOSFET.

drift-diffusion, energy balance or Monte Carlo) since it incorporates impurity scattering, charge transport (with likely velocity overshoot) and realistic boundary conditions. On the other hand, transport in a ballistic diode is not complicated by two-dimensional potential or quantum confinement effects (both present in the channel of a MOSFET), which allows for the other transport features mentioned above to be better isolated and understood. The program code described in this dissertation has been implemented to simulate any one-dimensional electron device, but focus in this section will be given to the ballistic diode because of its relevance to a variety of transport problems. The code was named MONET[6] and it is occasionally referred to as such in the remainder of this manuscript.

3.6.1 Self-Consistent Poisson Equation

The Monte Carlo modeling of a one-dimensional device, such as a ballistic diode, requires the use of a simulation grid since the doping, electric field, potential and carrier profiles will all be dependent on position. In this work, as is often the case in Monte Carlo simulation,

[6]Perhaps because of the author's admiration for the artists's work, or perhaps because MONET could stand for, e.g., MONte carlo for Electron Transport.

the grid is chosen to be uniform. This is done to simplify charge assignment on the grid nodes and to eliminate spurious "self-forces" [72, 73].

As mentioned in Chapter 2, the ensemble Monte Carlo method models the entire mobile charge inside the semiconductor device with a few thousand (e.g. ten to twenty thousand) particles. These "super-particles" are treated as individual charge carriers while they drift, but as clouds of charge when the simulation is stopped and the Poisson equation is solved. The amount of charge then assigned to each super-particle is given by (from Chapter 2) $Q = eN/N_{sim}$ where e is the elementary charge, N is the total number of mobile charges expected in the real device and N_{sim} is the number of super-particles used in the simulation. Charge assignment on the device grid is done with the cloud-in-cell method, with

$$w_1 = (x - X_i)/(X_{i+1} - X_i) \tag{3.20}$$

$$w_2 = 1 - w_1 \tag{3.21}$$

which are weights used in a simple linear interpolation of the charge Q at position x, onto grid nodes at locations X_i and X_{i+1} (where $X_i < x < X_{i+1}$). The charge assigned to the grid nodes is then given by $w_1 Q$ for grid node i and $w_2 Q$ for grid node $i + 1$.

In order to self-consistently update the electric field as the mobile charge moves during the simulation, Poisson's equation must be solved at every time step ΔT. In other words the mobile charge is allowed to drift under the influence of the electric fields for ΔT seconds (an upper limit on this time step being given by the plasma oscillation period, as explained in Chapter 2), then the simulation is stopped, the mobile charge is assigned to the grid nodes, and the Poisson equation is solved in order to update the electric fields. The Poisson equation may be written as

$$\nabla^2 \Phi(x) = -\frac{\rho(x)}{\epsilon_s} = -\frac{e}{\epsilon_{si}} \left[p(x) - n(x) + N_d(x) - N_a(x) \right] \tag{3.22}$$

where Φ is the potential, ρ is the net charge density and ϵ_s is the dielectric constant of the semiconductor. The mobile charge densities (after charge assignment with the cloud-in-cell

method) are given by n and p for electrons and holes, while the fixed charge is determined by N_d and N_a, the donor and acceptor doping profiles. In simulations of ballistic diodes with MONET, the acceptor and hole densities are zero, since MONET only simulates electron, not hole, transport. The Poisson equation can be discretized in general as:

$$-\frac{1}{h_i}\Phi_{i-1}+\left[\frac{1}{h_i}+\frac{1}{h_{i+1}}\right]\Phi_i-\frac{1}{h_{i+1}}\Phi_{i+1}=\frac{e}{2\epsilon_s}(N_{di}-n_i)(h_i+h_{i+1}) \tag{3.23}$$

where $h_i = x_i - x_{i-1}$ and $h_{i+1} = x_{i+1} - x_i$ and these differences in x become simply Δx on a uniform grid. The discretized Poisson equation can then be written as a set of linear algebraic equations which can be easily solved through conventional means, e.g. tridiagonal elimination [35, 74]. Once the potential is found, the electric field is written as its negative derivative through centered differencing [75]:

$$E_i=-\frac{d\Phi}{dx}\simeq\left[\frac{h_{i+1}/h_i}{h_i+h_{i+1}}\right]\Phi_{i-1}+\left[\frac{1}{h_{i+1}}-\frac{1}{h_i}\right]\Phi_i-\left[\frac{h_i/h_{i+1}}{h_i+h_{i+1}}\right]\Phi_{i+1} \tag{3.24}$$

where $h_{i,i+1}$ are as defined above and, in the case of uniform grid spacing Δx, reduces to

$$E_i=-\frac{\Phi_{i+1}-\Phi_{i-1}}{2\Delta x}. \tag{3.25}$$

Particular care must be taken near the device boundaries and the following approach is adopted in this work. The potential at the two boundaries (grid nodes 1 and n) is assumed fixed, set by the applied voltage V, such that $\Phi_n - \Phi_1 = V$ (the initial potential profile "guess" is actually read at the beginning of the simulation from a previous simulation run done with a commercial drift-diffusion code, like Medici). The electric field for the two boundary nodes is then found through off-centered differencing as [75]:

$$E_1 = -\frac{-3\Phi_1+4\Phi_2-\Phi_3}{X_3-X_1} \tag{3.26}$$

$$E_n = -\frac{3\Phi_n-4\Phi_{n-1}+\Phi_{n-2}}{X_n-X_{n-2}} \tag{3.27}$$

where the denominator, in both cases, is equal to $2\Delta x$ for a uniformly spaced grid. After

the electric field is found, the simulation resumes and particles are allowed to drift under the influence of the new field distribution for another ΔT seconds, after which this process repeats (see Fig. 2.2).

3.6.2 Contact Boundary Conditions

In the case of one-dimensional simulation only two boundaries are present, which are the contacts where the voltage is applied. In general, these contacts are unions of mesh nodes where the device domain touches an ideal source/sink of carriers. In most Monte Carlo simulations these boundaries are treated as ideal ohmic contacts, absorbing all incident electrons that actually reach them, and emitting (as neccesary, and explained further below) only electrons in thermal equilibrium with the contact temperature [71]. The boundary conditions for particle transport must be consistent with those for the electric field and potential. There are two ways which are usually employed to treat the particle flux at boundaries within Monte Carlo simulation. They have both been implemented within MONET, the code developed during this dissertation, and one or the other can be selected when the code is compiled. The simplest way to model the two contacts is to assume periodic boundary conditions, that is, particles which escape from one contact are reinjected at the other with thermal energy, and with a momentum component weighed toward the inside of the device as [8]:

$$p_x = \sqrt{-2m_x k_B T \ln(r)} \tag{3.28}$$

where m_x is the conduction band effective mass along the injection direction and r is a uniformly distributed random number between 0 and 1. This method conserves the particle flux (current continuity) at the boundaries, but it is only suitable for one-dimensional simulation, and not for devices with three or more contacts (e.g. a bipolar junction transistor). The particle current can be computed, for example as

$$I = \frac{1}{t_{sim}} Q(N_{right} - N_{left}) \tag{3.29}$$

where Q is the super-particle charge (Eq. 2.2), t_{sim} is the simulation time and the term in parenthesis is the difference between the number of particles that exit through the right versus the left contacts. The instantaneous current (e.g. during transients) can be similarly computed by counting particles exiting through the contacts during shorter periods of time, e.g. only a few time steps ΔT.

The other method for treating device boundaries is more frequently employed in the literature because it can be extended to devices with an arbitrary number of contacts. It involves maintaining local charge neutrality at the grid nodes adjacent to the contact, which is done as follows. At the beginning of the simulation a target super-particle density is calculated at each contact, as consistent with local charge neutrality. During the simulation the particles which exit through the contacts are deleted and tallied as current. Within MONET, this is done by copying the information of the last particle in the array where particles are stored on top of the i^{th} particle to be deleted, then shrinking the array size by one. After each time step ΔT, just before the Poisson equation is solved, the program examines the super-particle count at each contact node and determines how many particles should be injected or deleted to reach the charge-neutral target initially determined. The injected particles are assumed to have thermal energy, and a momentum component forward-weighed into the device, as previously described (Eq. 3.28). This velocity weighing is essential, as it accounts for the higher probability of a "fast" particle entering the device from the conceptual thermal carrier gas considered touching the contact. Every particle injected or deleted is tallied as current too. Note that with this second method for modelling device contacts, the number of super-particles present in the device at any given time during the simulation is not constant. This is also the method preferred for Monte Carlo noise simulations [50, 71].

3.6.3 Ballistic Diode Simulation Results

To illustrate the one-dimensional device applications of the Monte Carlo code MONET, a n^+nn^+ ballistic diode was simulated. The results are shown in Fig. 3.12 for the potential, electric field, average electron velocity and density (solid lines) — and they are compared

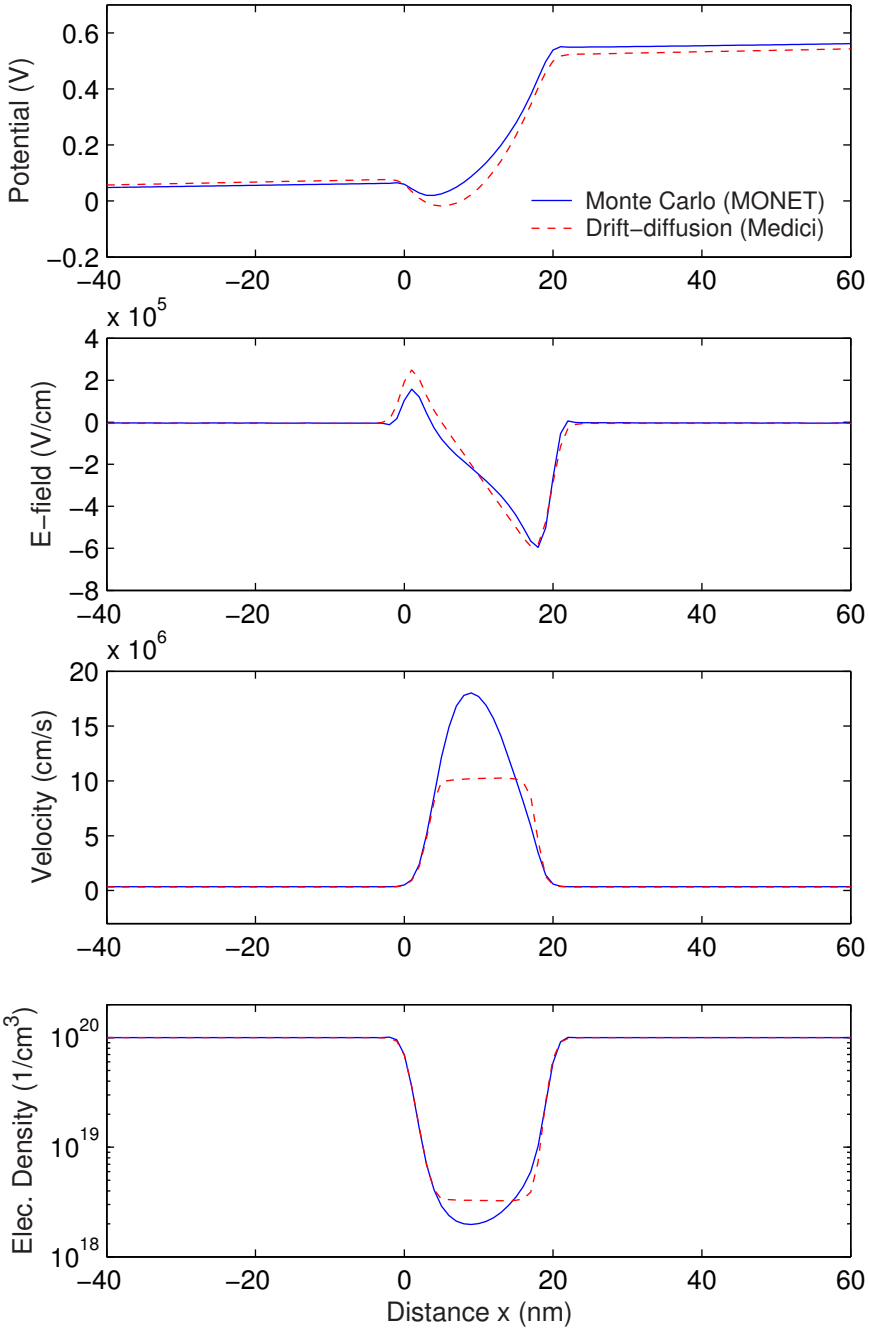

Figure 3.12: Ballistic diode with 20 nm long middle "n" region (doped 10^{16} cm^{-3}), as simulated with the drift-diffusion code Medici (dashed lines) and the Monte Carlo program developed in this thesis, MONET (solid lines). The applied bias is 0.6 V, and the "n$^+$" regions are not entirely shown (they were 100 nm long, doped 10^{20} cm^{-3}). Note the Monte Carlo code indicates significant velocity overshoot.

with the results of the commercial drift-diffusion code Medici (dashed lines). The n^+nn^+ diode has a "channel" length of 20 nm and source and drain lengths of 100 nm (although only 40 nm of each are shown in the plots). The source/drain doping is 10^{20} cm^{-3} and the channel doping is 10^{16} cm^{-3}. The applied voltage for the simulations in the figure was 0.6 V. The one-dimensional device structure was first "built" and simulated with the commercial code Medici, with a uniform grid spacing. The resulting grid, charge, potential and electric field distributions were then saved and imported into MONET, where they served as the initial conditions. The Poisson equation was self-consistently solved along with the Monte Carlo transport of charge. Several similarities and differences can be pointed out between the drift-diffusion code and the Monte Carlo results. As can be seen from the plots, the potential and electric field distributions are very similar. The Monte Carlo code, however, predicts significant velocity overshoot in the short "channel" region, whereas the average velocity predicted with the model selected in Medici plateaus at 10^7 cm/s, the saturation velocity in silicon. Moreover, the influence of the heavily doped drain region (which injects cool, slow electrons) is clearly seen in the velocity distribution computed by the Monte Carlo method, which is slightly skewed toward the source side. It is also clear that the average electron velocity is not at all a local function of the electric field. The differences in the particle density distributions are consistent with the differences in the average velocity between the two computational methods, as the net current density (proportional to $n \times \overline{v}$) is the same, and constant through the one-dimensional profile, as required by current continuity. This example shows the applicability of MONET to one-dimensional transport problems, including self-consistently computed electric field distribution, spatially-varying doping profile and realistic device contacts.

3.7 Two-Dimensional Device Applications

The scope of this dissertation work was not to develop a new Monte Carlo simulator from the ground up,[7] but rather to enhance the current transport models by including an analytic

[7]Such efforts have taken teams of full-time researchers years to complete [36, 73].

phonon dispersion and to study heat generation in bulk and strained silicon, and in simple device geometries. MONET can be used to simulate transport in two-dimensional device geometries, but more care must be exercised in interpreting the results of such simulations. The two-dimensional grid (including electric fields, doping, and device boundaries) must be imported from a previous drift-diffusion simulator run (e.g. Medici). MONET does not solve the two-dimensional Poisson equation, and hence the electric fields it computes the particle motion on are the "frozen" ones imported at the beginning of the simulation. This is the so-called non-self-consistent approximation, which has limited applications, and has been shown [50] (as it might be expected) to not yield significant improvements in accuracy over the drift-diffusion approach. This approach is also not suitable for device noise simulations [50]. However, the results of such non-self-consistent simulations can still be appreciated and interpreted qualitatively. Figure 3.13 illustrates the three-step process by which MONET can be used to perform such simulations: the mesh (top subplot) and electric field distribution (middle subplot) are imported from a drift-difussion simulation with Medici, with voltages applied as necessary. MONET initially distributes particles in proportion with the charge density (not the doping density) imported from Medici. These particles are first assigned thermally distributed energies and randomly oriented momenta. Then, the particles are allowed to drift under the influence of the electric field grid, but the electric fields are not updated as the charge moves around. Boundary conditions at the source and drain electrodes are similar to those described in the previous section. Scattering with the other surfaces (e.g. between silicon and silicon dioxide) reflects the particles back into the simulation domain, with unchanged energy, but newly oriented momenta. This scattering can be either specular (the reflection angle is the same as the incident angle) or diffuse (randomly chosen reflection angle). A specularity parameter is used to choose between the two types of surface scattering, and the ratio of diffuse to specular scattering is set at 0.15 [76] (which can be used-adjusted, as described in Appendix A). The bottom subplot in Fig. 3.13 shows a snapshot of such a Monte Carlo simulation with only a few hundred super-particles shown, for clarity. The device being simulated is an 18 nm gate length thin-body SOI with 10^{20}cm^{-3} doped source and drain, undoped body

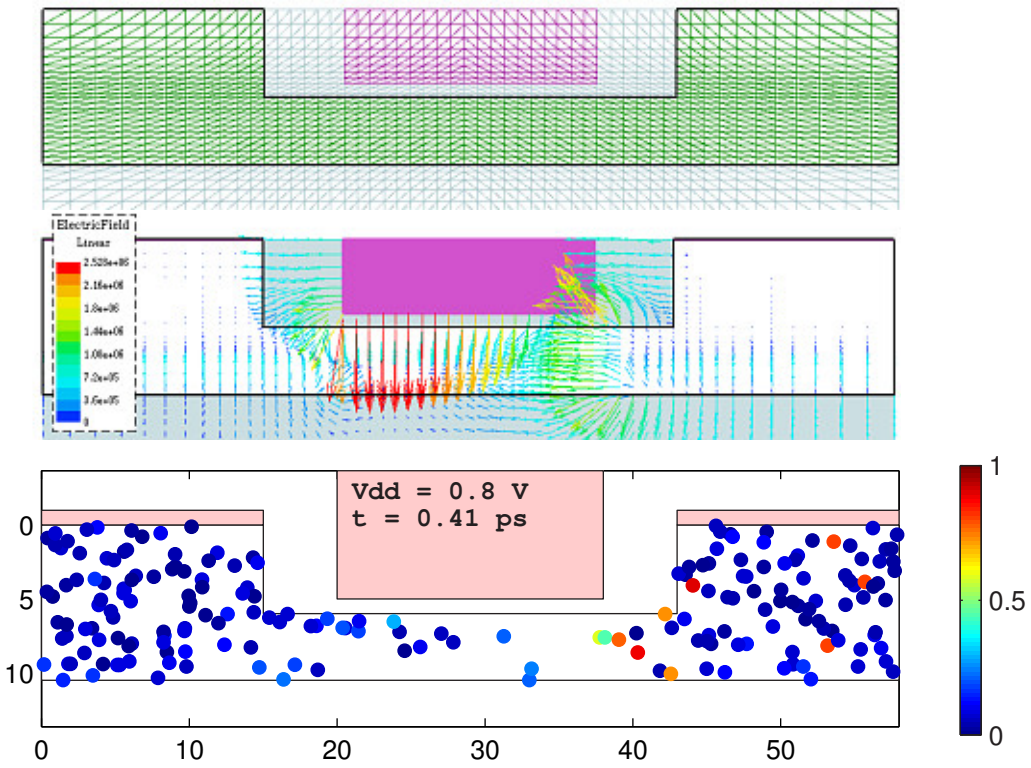

Figure 3.13: Mesh layout (top), electric fields (middle) and Monte Carlo simulation snapshot (bottom) of an 18 nm gate length thin-body SOI device. The mesh and electric field distribution are imported from a drift-diffusion simulation with Medici. The Monte Carlo simulation only shows a few hundred particles, for clarity. The vertical color bar is the electron energy scale in eV, the physical axes are in nm.

and Molybdenum gate. The body thickness is 4.5 nm. The on/off current ratio predicted by Medici for this layout is 1000/1 (μA/μm). Qualitatively, a few observations can be made based on this non-self-consistent simulation. One can note the presence of hot electrons almost entirely in the drain of the device. This indicates that (a) transport across the short channel is nearly ballistic, and that (b) energy relaxation of the carriers, and therefore Joule heating of the lattice happens entirely in the drain region of the device. This point will be discussed in more detail in the next chapter, and the exact location of the heat generation region will be analyzed with electrostatically self-consistent one-dimensional simulations.

3.8 Summary

This chapter introduced a Monte Carlo (MC) simulation approach which fills the gap of computational tools between simple analytic-band MC codes [34, 44] and more complex full-band simulators [36, 43]. The emphasis of this work is on sophisticated physical modeling within a computationally efficient framework. The use of analytical electron bands and phonon dispersion enables simulations which are more than an order of magnitude faster than full-band techniques, and very accessible on modern desktop computers. This method can be applied to the engineering of low-voltage nano-devices and materials that require detailed knowledge of electron-phonon coupling. The generated phonon distributions can be extracted [10] and used as inputs to a phonon transport solver [18].

A new, unified set of deformation potentials for intervalley scattering was also introduced which enables more accurate electron transport simulations in both strained and unstrained silicon. The empirically fine-tuned coupling constants were extracted consistently with the band and phonon structure. This work represents a new approach to analytic-band Monte Carlo codes because it distinguishes between intravalley scattering with LA and TA phonons, and includes an analytic dispersion for *all* phonon modes involved. The work can be extended beyond silicon, to other materials (like germanium) or to strained or confined nanostructures. Hole transport could also be simulated accordingly by modeling the valence bands like in Ref. [34] and including the full phonon dispersion.

Comprehensive and electrostatically self-consistent one-dimensional device simulation capability was demonstrated. Two-dimensional device simulations were shown to be adequate for qualitative analysis. Appendix A contains a user manual for MONET, the Monte Carlo code described in this chapter. Further documentation, results, and examples are shared online [77].

Chapter 4

Heat Generation in Silicon and in Simple Device Geometries

This chapter examines the detailed spectral make-up of Joule heating (phonon emission) in silicon with the aid of the Monte Carlo technique developed earlier, including acoustic and optical phonon dispersion. It is found that a significant portion of the generated phonons have low group velocity, like optical modes, or acoustic modes near the edge of the Brillouin zone. The generated phonon spectrum in strained silicon is different from bulk silicon at low electric fields due to band splitting and scattering selection rules which favor g-type and reduce f-type phonon emission. However, heat generation is essentially the same in strained and bulk silicon at high fields, when electrons have enough energy to emit across the entire phonon spectrum despite the strain-induced band splitting. Heat generation in short devices (ballistic diodes) is found to occur almost entirely in the drain, beyond the peak electric field region, as transport across the channel is quasi-ballistic. The heat generation region ("hot spot") extends deep into the drain, because the energetic electrons take a relatively long time (and space) to fully relax their energy to the lattice.

4.1 Introduction

Understanding heat generation in silicon is of great physical interest and particularly relevant to the self-heating and reliability of nanoscale and thin-film transistors. As previously mentioned in Section 1.3 ("Heat Generation in Semiconductors"), Joule heating in the context of a semiconductor device is often simulated with the classical drift-diffusion approach, whose main component is the dot product of the electric field ($\mathbf{E}$) and current density ($\mathbf{J}$) [17, 26, 29]:

$$Q''' = \mathbf{J} \cdot \mathbf{E} + (R - G)(E_g + 3k_B T) \tag{4.1}$$

where $(R - G)$ is the net (non-radiative) recombination rate, E_g is the semiconductor band gap and T is the lattice temperature. Details on the significance of each term are given in Section 1.3. The heating rate can also be computed with the more sophisticated hydrodynamic approach, as a function of the electron temperature (T_e) and an average electron energy relaxation time (τ_{e-L}) [31]:

$$Q''' = \frac{3}{2} k_B \frac{n(T_e - T_L)}{\tau_{e-L}} + (R - G)\left[E_g + \frac{3}{2} k_B (T_e + T_L)\right] \tag{4.2}$$

where n is the electron density and the subscript L denotes the semiconductor lattice. As mentioned earlier, the field-dependent drift-diffusion approach does not account for the non-local nature of transport and phonon emission near strongly peaked electric field regions. The hydrodynamic approach suffers from simplifications inherent to using a single (averaged) carrier temperature and relaxation time, since scattering rates are strongly energy dependent. Neither method differentiates among electron energy exchange with the various phonon modes, nor do these methods give information regarding the frequencies of phonons emitted. Such spectral information is important because the emitted phonons travel at different velocities and have widely varying contributions to heat transport [9, 11] and device self-heating [32, 33]. This chapter addresses these issues by using a Monte Carlo (MC) simulation method to compute detailed phonon generation rates at various electric fields, in technologically relevant doped bulk and strained silicon. The heating generation

rate is also computed in a few simple (one-dimensional) device geometries, and the role of non-local phonon emission near strongly peaked electric fields is investigated.

4.2 Implementation

The details of this Monte Carlo implementation have been described in Chapter 3 and elsewhere [37]. The electron energy bands are modeled with the analytic non-parabolic band approximation, including the six ellipsoidal conduction X-valleys of silicon [34]. This is a good approximation for device voltages near or below the silicon band gap (1.1 eV), such as those of future nano-technologies, and it is significantly faster than the full-band MC method [36, 57]. In addition, the low electron energy range means that both impact ionization (which is controlled by the band gap) and interband scattering with higher energy bands can be safely neglected (the X-L band separation in silicon is about 1.2 eV).

As in the typical analytic-band approach [34], inelastic scattering with six types of intervalley phonons is incorporated. Three of these are of g-type, assisting electron scattering between valleys on the same k-axis, and three are of f-type, when the scattering occurs between valleys on orthogonal axes. Intravalley scattering refers to scattering within the same conduction band valley and usually involves only acoustic phonons [53]. Ionized impurity scattering is included with an efficient model which reduces the number of time-consuming small angle scattering events [65] (see Section 3.4). Electron transport in strained silicon is incorporated in the MC simulation as follows (also see Section 3.5.2). The biaxial strain of silicon grown on a $Si_{1-x}Ge_x$ substrate removes the degeneracy of the conduction bands, lifting four of the six valleys by $\Delta E \simeq 0.67x$, where x is the Ge fraction of the substrate [44]. The in-plane conductivity effective mass of the two lower valleys is the lighter transverse mass m_t of silicon. The strain-induced energy splitting between the valleys also reduces f-type intervalley scattering, since the maximum available f-type phonon energy in silicon is only about 50 meV (see the phonon dispersion plot in Figure 3.4). For $x > 0.15$ the energy splitting is large enough to essentially suppress f-type scattering between the lower and upper valleys, and the strained silicon mobility enhancement is dominated by conduction via

the two lower valleys with the lighter transverse mass. Figure 3.6 compares data reported by Ismail *et al.* [66] with our simulations for electron drift velocity over a wide range of electric fields. The simulation for strained silicon assumes no impurity concentration, while the bulk sample is doped to 10^{17} cm^{-3}. The strained silicon data of Ref. [66] were taken on modulation-doped structures, where transport is phonon-limited at room temperature. Our simulations agree well with the data within MC and experimental error, over a large range of electric fields, and in both bulk and strained silicon. The scattering potentials used in the simulations have also been carefully calibrated over a wide range of temperatures, as outlined in Section 3.3 and Ref. [37].

Typical Monte Carlo codes treat phonon scattering with a simplified dispersion relationship, often grouping the longitudinal (LA) and transverse acoustic (TA) branches into a dispersionless mode with constant group velocity [34], and assuming a single, fixed optical phonon energy [36]. Since the intentions of this chapter (and this work) are to explore the details of the phonon generation spectrum, scattering with all phonon modes must be treated individually, and taking into account the phonon frequency dependence on wave vector. Each branch of the phonon dispersion is modeled with the analytic approximation

$$\omega_q = \omega_o + v_s q + c q^2, \tag{4.3}$$

where ω_q is the phonon frequency and q is the wave vector [37]. The continuous (longitudinal) and dashed (transverse) lines in Fig. 4.1(a) represent this quadratic model, and the fitting coefficients are listed in Table 3.1. Our dispersion model is otherwise assumed isotropic, since complex *ab initio* studies have shown that the anisotropic effect of the phonon dispersion is rather small [43, 57].

The electron-phonon scattering rate is calculated in the typical fashion, based on the Fermi Golden Rule [8, 34]. However unlike the typical approach, the intravalley scattering rate is computed separately with LA and TA phonons. Intervalley scattering is computed taking into account the six phonons (three of f-type, three of g-type) which satisfy the conservation of crystal momentum. Both the intervalley and intravalley scattering rates

are computed taking into account the isotropic phonon dispersion model, as described in Section 3.3. The simulation treats all phonon scattering events as inelastic, and electrons exchange energy with the lattice as determined by the phonon dispersion and scattering selection rules. Scattering with intravalley LA and TA phonons, as well as with intervalley longitudinal (LO) and transverse optical (TO) phonons is considered individually. The phonon dispersion is also used when computing the final electron state after scattering, taking into account both momentum and energy conservation. This allows a range of phonon wave vectors and energies *around* the six typical f- and g-type phonons to participate in scattering. This is an innovative, efficient and physically realistic approach introduced for the first time in this dissertation and in Ref. [37]. During the simulation all phonons absorbed and emitted are tallied, and full phonon generation statistics can be computed. The total heat generation rate can be obtained from the sum of all phonon emission events minus all phonon absorption events per unit time and unit volume:

$$Q''' = \frac{A}{\Delta V \Delta t} \sum (\hbar\omega_{ems} - \hbar\omega_{abs}) \tag{4.4}$$

where $A = N/N_{sim}$ is the scaling constant (ratio) between the total number of mobile charges in the device, N, vs. the number of super-particles used in the simulation, N_{sim} (also see Eq. 2.2). The steady-state heat generation rate on a particular device grid can be obtained by replacing Δt by the total simulation time after the transient is assumed to die out, while ΔV is the volume element at each grid node. As shown in the next sections, this approach can also be used to investigate the phonon generation spectrum (the generation rate as a function of phonon frequency and mode), as well as to study non-local heat generation near a strongly peaked electric field within a realistic device geometry.

4.3 Heat Generation in Bulk and Strained Silicon

This section explores the heat generation spectrum in doped bulk (and strained) silicon with a constant applied electric field, i.e., essentially in the context of an infinite resistor.

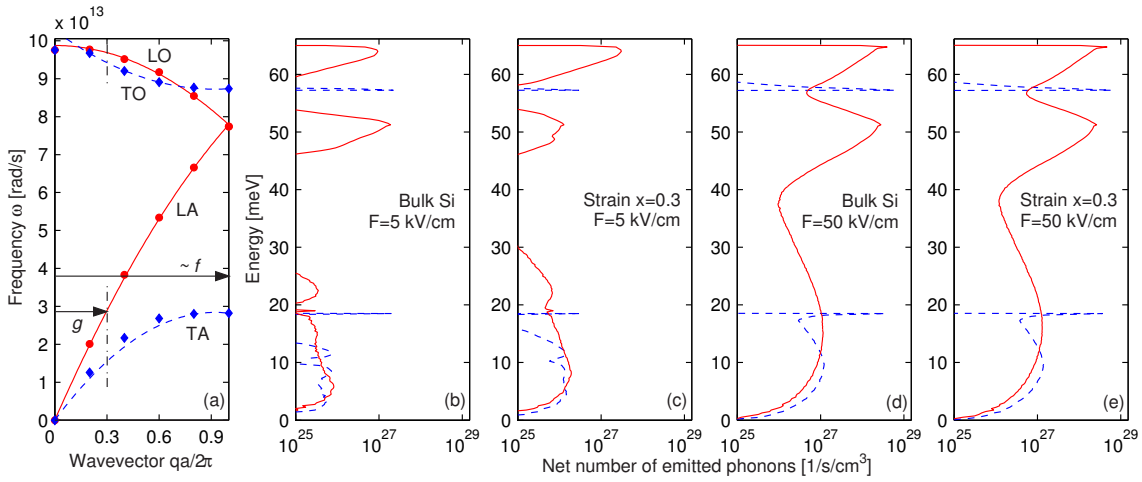

Figure 4.1: Phonon dispersion in silicon (a) and computed net phonon generation rates (emission minus absorption) with low field (b,c) and high field (d,e) in strained and bulk silicon doped to 10^{17} cm^{-3}, at T=300 K. Subplot (a) shows the dispersion data of Ref. [54] (symbols), our quadratic approximation (lines), and the vector magnitude of f- and g-type intervalley phonons. Dashed lines represent transverse, while solid lines represent longitudinal phonons throughout.

This is the "zero-dimensional" mode of operation of MONET, as described in more detail in Appendix A. In this situation, the Monte Carlo program can be used to examine the details of net phonon generation as a function of phonon frequency, in order to find out exactly which branches of the phonon dispersion are excited when current flows in a constant applied electric field. This situation resolves how many acoustic vs. optical and longitudinal vs. transverse modes are generated through Joule heating in silicon.

Figure 4.1 shows the computed generation spectrum in 10^{17} cm^{-3} doped bulk and strained silicon with both a lower (5 kV/cm) and higher (50 kV/cm) applied electric field. These electric field values were chosen from two regions of Fig. 3.6 such that the mobility enhancement in strained silicon is maintained at the lower field value, but not at the higher field. To facilitate comparison, Fig. 4.1 subplots (b)-(e) are drawn such that the vertical axes with energy units in meV match the vertical frequency axis of the phonon dispersion in subplot (a), with units in rad/s, as given by $E = \hbar\omega$. Note the cutoff energies of the various emitted phonon populations as required by their respective dispersion relation. Few acoustic phonons are generated through intravalley scattering at low energies because the

three-dimensional phonon density of states vanishes near the Brillouin zone (BZ) center, where the phonon wave vector $q \rightarrow 0$, as [7]:

$$g_p(\omega) = \frac{\partial N_s}{\partial \omega} = \frac{q^2}{2\pi^2}\left(\frac{dq}{d\omega}\right) \tag{4.5}$$

where N_s is the total number of phonon states up to the frequency ω and $dq/d\omega = 1/v_s$ is the inverse of the phonon group velocity near the BZ center (see Table 3.1). Intravalley emission also decreases at higher frequencies (higher wave vectors) since fewer electrons with large enough momentum are available to emit phonons of larger wave vector. This behavior limits the intravalley phonon emission spectrum, both for LA and for TA phonons.

The sharp peaks in the phonon generation plots occur due to intervalley scattering with the three g-type (TA, LA and LO, at 0.3 of the distance to the edge of the BZ) and three f-type (TA, LA/LO and TO, at the edge of the BZ) phonons. The momenta and hence the location within the BZ of these six intervalley phonons are given by scattering selection rules [52]. The relative magnitude of their generation rates depends on the choice of scattering deformation potentials, which have been carefully calibrated in Chapter 3 and Ref. [37]. While the deformation potentials used in this work may not be unique,[1] the values empirically determined and used in this dissertation are the only ones currently available in the literature which reproduce the experimental mobility data for both strained and bulk (doped) silicon. Previous such mobility simulations have used two separate sets of deformation potentials for bulk and for strained silicon, a situation acknowledged [58] to be unphysical. Figures 4.1(b) and (c) highlight the difference in the phonon emission spectrum between strained and bulk silicon at low electric fields. The strain-induced band splitting suppresses f-type phonon emission between the two lower and four upper valleys. However, since most conduction electrons in strained silicon are confined to the two lower valleys (of lighter mass m_t), they quickly gain energy and g-type emission between the lower valleys is enhanced. Comparing Figs. 4.1(d) and (e), it can be noted that phonon generation in strained and bulk silicon at high field is essentially identical, when electrons have enough

[1] Other possible sets are summarized in Table 3.2.

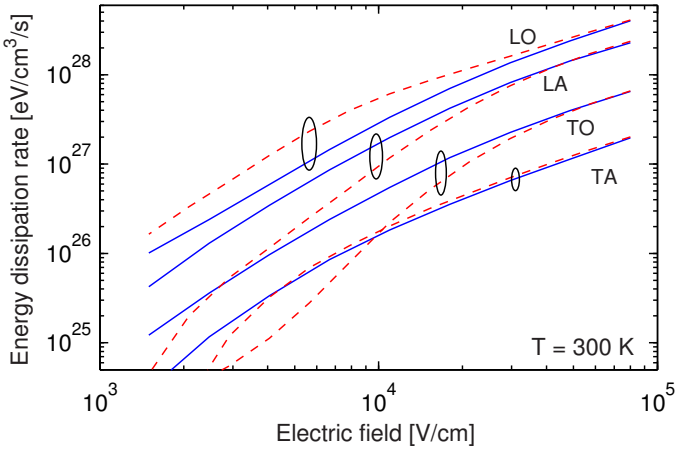

Figure 4.2: Heat generation rates for each phonon mode as a function of applied steady-state electric field. Dashed lines are for strained silicon ($x = 0.3$ substrate Ge fraction), solid lines are for bulk silicon, both doped to 10^{17} cm^{-3}.

energy to emit across the entire phonon spectrum despite the strain-induced band splitting. This is consistent with the observation of similar saturation velocity in strained and bulk silicon (Fig. 3.6).

Figure 4.2 compares the integrated net energy dissipation rates with each branch of the phonon spectrum, at various steady-state electric fields. Dashed lines are used for strained silicon, and solid lines are for bulk silicon. Note the enhancement of g-type LO emission and the reduction in energy relaxation through f-type LA and TO phonons in strained silicon at lower electric fields, as expected from the earlier discussion. However, at fields larger than 30–40 kV/cm, the average electron energy (in excess of 0.25 eV) becomes comparable to or larger than the 0.2 eV strain-induced band-splitting (for Ge fraction $x = 0.3$), and heat generation rates are the same in bulk and strained silicon. The total Joule power (per unit volume) dissipated to the lattice, over all four phonon branches, conveniently sums up to

$$Q''' = \mathbf{J} \cdot \mathbf{E} = en\overline{v}|\mathbf{E}| \tag{4.6}$$

as anticipated, where J is the current density, E the electric field, e the elementary charge, n the electron density and $\overline{v}$ the average drift velocity (as calculated, e.g., in Fig. 3.6). This

	Bulk (all fields) and high-field strained Si	Low-field strained Si
TA	< 0.03	0.02
LA	0.32	0.08
TO	0.09	< 0.01
LO	0.56	0.89

Table 4.1: Approximate fractions of the total Joule heating rate for each branch of the phonon dispersion, based on Fig. 4.2. Note that each column adds up to unity. LO (mostly g-type) phonon emission dominates low-field heat dissipation in strained silicon, but it accounts for only slightly over half the heat dissipation in bulk silicon (all fields) and high-field strained silicon.

relationship holds for both bulk and strained silicon. Note the use of electron-volt units (the elementary charge is taken to be unity) instead of Joules in Fig. 4.2. It should also be noted that intervalley f-type scattering near the 50 meV phonon energy can be satisfied both by LA and LO phonons [57] (where the two branches meet), so the distinction between the two modes is somewhat arbitrary. In this work, a phonon is assigned to the LA branch when its wave vector $|\mathbf{q}| < |\mathbf{G}| = 2\pi/a$, and to the LO branch otherwise, where $\mathbf{G}$ is the reciprocal lattice vector and $a = 5.431$ Å is the silicon lattice constant. Also note that Fig. 4.1 shows generated phonon numbers, whereas Fig. 4.2 displays the phonon energy generation (electron energy dissipation) rates. Even a quick examination of Figs. 4.1 and 4.2 reveals that, unlike it is usually assumed [17, 31], *not* all electron energy is dissipated into the optical phonon modes. Table 4.1 summarizes the relative proportion of energy dissipated into each phonon mode, in bulk and strained silicon, at low and high electric fields. The longitudinal optical (LO) modes make up the majority of phonon emission especially in strained silicon, owing to strong electron coupling with the g-type phonon near 62 meV, but they account for just over half of the total heat generation rate in bulk silicon at most practical electric fields. Strong intervalley scattering with the f-type LA phonon at 50 meV is also a good mechanism for electron energy relaxation, especially in bulk silicon, where it accounts for about one third of the total heat generation rate. The remaining energy dissipation occurs through transverse optical (TO) and acoustic (TA) modes. The TO phonons involved in

f-type intervalley scattering are relatively energetic, but their coupling constants with the electrons (see Table 3.2), and thus their scattering rates, are weak, making them account for just under one tenth of the energy relaxation at high fields. Finally, transverse acoustic (TA) phonons have relatively low energy and low coupling constants, and consequently are the weakest energy relaxation (heat generation) mechanism. Note that both the f-type TA and TO phonons are generated at the edge of the Brillouin zone (BZ), where their group velocity is nearly zero, making them essentially stationary. However, their density of states is large at the edge of the BZ (large wave vector, large $dq/d\omega$ in Eq. 4.5), implying that their occupation N_q is likely to be small. Hence, an overpopulation of zone-edge TA and TO modes in a silicon-based electronic device is unlikely.

In strained silicon at low electric fields, all f-type intervalley phonon emission mechanisms (TA, LA and TO) are strongly suppressed by the strain-induced band splitting.[2] Transport in this case is dominated by conduction via the two lower band valleys, and intervalley scattering with the g-type LO phonon is the dominant means for energy relaxation, accounting for all but about ten percent of the energy released to the lattice (Table 4.1). In other words, the conventional assumption of Joule heat being dissipated primarily into the (nearly stationary) optical modes [17, 31] is a good approximation only in strained silicon at low electric fields. Otherwise, care must be exercised when solving the phonon BTE to compute self-heating in a silicon transistor operating under high field conditions, and the entire silicon dispersion ought to be used.

4.4 Heat Generation in Ballistic Diodes

This section investigates heat (phonon) generation in one-dimensional devices, and specifically in the case of the ballistic diode first introduced in Section 3.6. Three ballistic diode scenarios are first considered, with "channel" lengths of 500, 100 and 20 nm. The source and drain regions are assumed doped to 10^{18}, 10^{19} and 10^{20} cm^{-3}, and the applied voltages

[2] As long as the strain-induced band splitting ($\Delta E = 0.67x$) is several times larger than the largest available f-type phonon, i.e. the TO mode near 50 meV (hence, for Ge fraction values $x > 0.15$).

are 2.5, 1.2 and 0.6 Volts, respectively. The latter are roughly equivalent to the operating voltages recommended by the ITRS guidelines [1] for CMOS devices of similar channel lengths. The middle (channel) region is assumed doped to 10^{16} cm^{-3} throughout. Monte Carlo simulations of heat generation using the code developed in this thesis, MONET, are compared to the heat generation rate computed using the commercial drift-diffusion simulator MEDICI, with the $\mathbf{J} \cdot \mathbf{E}$ approach of Eq. 1.8. In general, Monte Carlo simulation results are expected to be similar to those of the drift-diffusion calculations for "long" devices (long compared to the electron-phonon mean free path), i.e. in the continuum approximation. This limit provides a check on the accuracy of the Monte Carlo simulation, and enables a study of the conditions under which the drift-diffusion heat generation calculations break down. The Monte Carlo results are expected to differ from (and be more physically accurate than) the drift-diffusion results in the limit of short channel lengths (comparable to the electron-phonon mean free path), where velocity overshoot and other non-equilibrium transport effects are expected to dominate. This is the limit under which the "granularity" of charge transport and phonon emission becomes important, and the continuum approximation of the drift-diffusion method breaks down.

4.4.1 Joule Heating of the Drain

Figure 4.3 displays heat generation rates computed along the three n^+nn^+ ballistic diodes of varying channel lengths. Both the drift-diffusion (Medici) and Monte Carlo (MONET) simulations are solved self-consistently with the Poisson equation, as described in Section 3.6.1. As expected, the two approaches give very similar results for the longest simulated device, with channel length (500 nm) much greater than the average electron-phonon scattering length (5–10 nm). This is essentially still in the continuum limit, and the drift-diffusion simulation approach is adequate. However, for the two shorter (100 nm and 20 nm) diodes the heat generation rates computed by the Monte Carlo approach are seen to differ significantly from the drift-diffusion results. The peak of the Monte Carlo heat generation is "displaced" from the peak of the drift-diffusion heat generation. This outcome is qualitatively expected, and an explanation for it was already given in Section 1.3: electrons gain

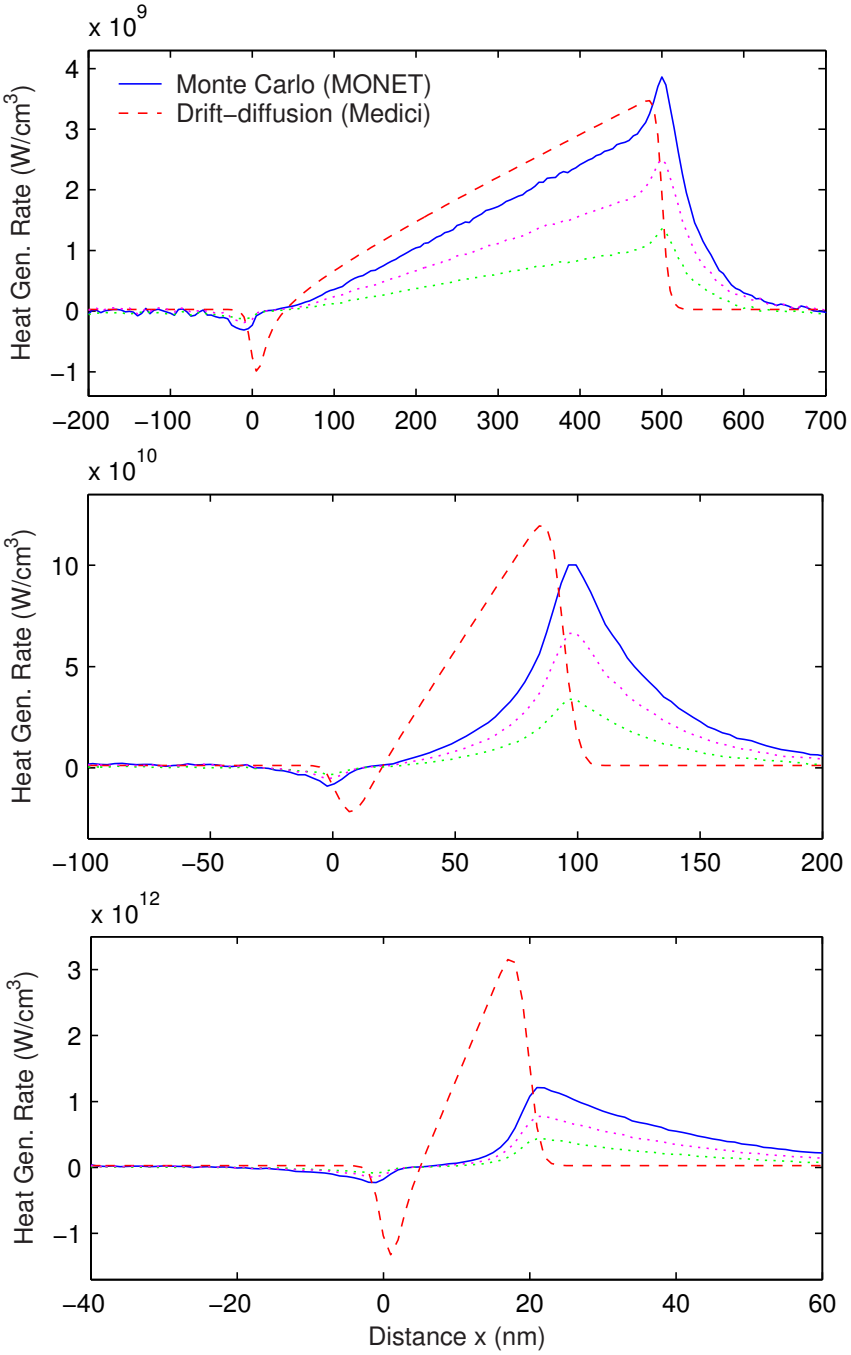

Figure 4.3: Heat generation along three different ballistic diodes with middle ("channel") regions of length 500 nm (top), 100 nm (middle) and 20 nm (bottom) and applied voltages of 2.5, 1.2 and 0.6 V, respectively. The solid line is the result of Monte Carlo simulations with MONET, while the dashed line is taken from drift-diffusion calculations using Medici. The dotted lines represent the optical (upper) and acoustic (lower) phonon heat generation rates, as computed by MONET.

most of their energy at the location of the peak electric field, yet they travel several mean free paths until they release this energy back to the lattice. Note that since the transport is one-dimensional, the current density $J = en\overline{v}$ is constant along the length of the diode, and the heat generation rate computed by the drift-diffusion ($\mathbf{J} \cdot \mathbf{E}$) approach peaks at the location of the electric field maximum. By comparison to the channel length L, the "non-local" errors incurred by using the drift-diffusion vs. the Monte Carlo approach when finding the location of the peak heat generation rate are $\Delta L/L = 0.10$, 0.38 and 0.82 for the three diode lengths $L = 500$, 100 and 20 nm. Another observation can be made about the "shape" of the heat generation in the drain region of the diode, downstream from the electric field. Since in reality electrons can only release energy in discrete packets (phonons) of at most 50–60 meV (the energy range of the optical phonons in silicon), the heat generation region computed by the (physically correct) Monte Carlo approach spreads deep into the device drain, as electrons drift toward the contact. This situation is particularly noticeable for the shortest device (20 nm), where transport in the channel is nearly ballistic, and almost the entire heat generation occurs in the drain. Note that the Monte Carlo method also computes the integrated optical and acoustic phonon generation rates, with dotted lines in Fig. 4.3. It can be seen that about twice as much energy is deposited in the optical (LO and TO) compared to the acoustic (LA and TA) modes, along the length of the simulated ballistic diodes. This is consistent with the observations summarized in Table 4.1, for Joule heating in bulk silicon due to current flow at most practical electric fields.

Figure 4.4 explores heat generation in the 20 nm diode in more detail. Several voltages are considered, from 0.2 to 1.0 Volts for the self-consistent Monte Carlo analysis. It can be easily seen that the maximum heat generation rate scales linearly with the potential drop across the channel, hence essentially with the applied voltage. The maximum average electron energy also scales linearly with the applied voltage V, approximately as $e \times 0.4V$, where e is the elementary charge. However, the characteristic (exponential) decay length of the heat generation region in the drain is always approximately $\Lambda_h = 20$ nm, regardless of the applied voltage. This can be qualitatively understood because electrons lose $\hbar\omega$ (a

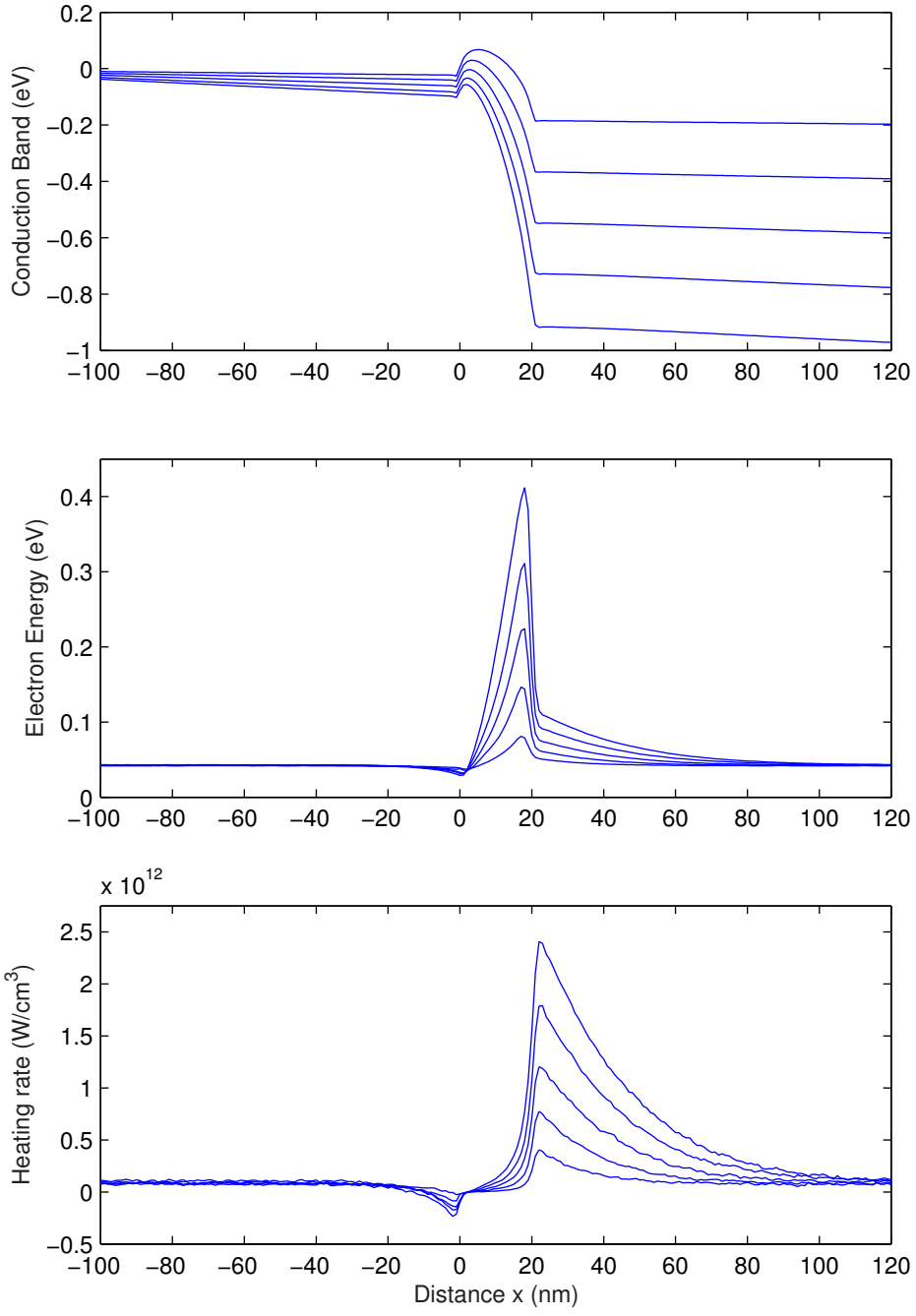

Figure 4.4: Monte Carlo simulations results for a short channel (20 nm) ballistic diode with applied voltages of 0.2, 0.4, 0.6, 0.8 and 1.0 V. The n^+ regions are doped 10^{20} cm^{-3}, the "channel" region is 10^{16} cm^{-3}. The edges of the channel are at 0 and 20 nm respectively. The top plot is the conduction band (increasing voltage from top to bottom), the middle plot is the average electron energy, and the bottom plot is the net heat generation rate (increasing with voltage from bottom).

phonon of) energy approximately every $v_e\tau_o$, the inelastic scattering length. Neglecting non-parabolicity, the electron velocity v_e scales as the square root of energy, while the inelastic (phonon) scattering time τ_o scales as $1/\sqrt{E}$ because the phonon scattering rate ($1/\tau_o$) scales with $\sqrt{E}$ from the density of states (Eq. 3.7). Therefore, the inelastic scattering length is relatively independent of the electron energy and of the applied voltage.

The length of the heat generation region in the drain can be understood in more detail as follows. The electrons present in the drain are a heterogeneous mixture of two populations, one being the "hot" electrons injected across the channel, and another made up of the many "cold" electrons already present there due to the high doping. The cold electrons have an average energy of $3k_BT/2$ (the thermal average) and they do not contribute to any net heat generation. Hence, the heat generation (bottom plot in Fig 4.4) is entirely caused by the hot electrons injected almost ballistically across the channel. While crossing the channel, these electrons acquire an amount of energy that is a significant fraction of the applied voltage. This energy is then released, in discrete amounts of $\hbar\omega$ (the phonon energy) to the lattice in the drain. Assuming an average inelastic scattering time $\tau_o = 0.05$–0.1 picoseconds (based on the Monte Carlo scattering rates computed in Chapter 3) and an average injected electron velocity $v_e = 10^7$ cm/s, the inelastic scattering length is about 5–10 nm. Since an electron of energy E must release multiple phonons to relax its energy entirely down to the thermal average, the total length of the heat generation in the drain can be written approximately as

$$L_h \simeq \frac{E - \frac{3}{2}k_BT}{\hbar\overline{\omega}}\ v_e\tau_o \tag{4.7}$$

where $\hbar\overline{\omega}$ is the average emitted phonon energy. The average energy of the hot electrons injected across the drain scales linearly with the applied voltage and it is a significant fraction of it ($E \sim \alpha V$). Furthermore, if the electron energy is significantly larger (several tenths of an eV) than $3k_BT/2$ (39 meV at room temperature), the multiplying fraction in Eq. 4.7 can be reduced to $E/(\hbar\overline{\omega})$. If the average emitted phonon (including acoustic and optical modes) has an energy about $\hbar\overline{\omega}$=50 meV, the multiplying factor $E/(\hbar\overline{\omega})$ is approximately 10–20 at biases near 1 Volt. Hence the length of the heat generation region

in the drain is on the order of $L_h \simeq 100$ nm, which is consistent with our findings via Monte Carlo simulation, as shown at the bottom of Fig. 4.4. Equation 4.7 is a crude approximation, but it gives a good order of magnitude estimate and correctly explains the long (much longer than the channel length when quasi-ballistic transport dominates) heat generation region in the device drain. These findings are consistent with the work of Lake and Datta [30], implying that heat dissipation in mesoscopic devices occurs in or near the contacts rather than in the active device region, i.e. when the length of the active region is on the order of the inelastic mean free path.

4.4.2 Thermoelectric Cooling of the Source

Unlike in the drain, the electrons in the source region are essentially in thermal equilibrium with the lattice temperature. However, a careful examination of both Figs. 4.3 and 4.4, reveals a small, but consistent negative heat generation region (lattice cooling) at the beginning of the channel. This is a thermoelectric effect due to the presence of the potential barrier at the injection point from the source into the channel. The situation is analogous to the Peltier effect in macroscopic junctions made of dissimilar materials which are used to make thermoelectric coolers. To understand the cooling effect when current flows over the potential barrier into the channel, consider the electron energy distribution just to the left of the barrier. The electrons in the source are essentially in thermal equilibrium and the distribution is a Maxwellian (or Fermi-Dirac function, if the doping level is degenerate) at temperature T. From this distribution, only the electrons with forward-oriented momenta and energies larger than the barrier height are going to travel into the channel. Since a portion of the high energy tail of the distribution is able to leave, the remaining electrons have an average energy below $3k_BT/2$. By the principle of detailed balance, these electrons will, on average, absorb more phonons than they emit, which contributes to a net effective cooling of the lattice. This thermoelectric cooling effect as current flows over an energy barrier can also be explained from the more classical approach of Eq. 1.8 (the $\mathbf{J} \cdot \mathbf{E}$ approach) and the discussion surrounding it. The electric field and the direction of current

flow are pointing in opposite directions at the beginning of the energy barrier into the channel, hence the $\mathbf{J} \cdot \mathbf{E}$ product is negative, and so is the heat generation rate. In other words, electrons diffusing *against* an energy barrier extract the energy required to move up the conduction band slope (against the electric field) from the lattice, through net phonon absorption. This phenomenon has been exploited in the design of heterojunction laser diodes, where the energy barriers introduced by band structure offsets can be optimized to provide internal thermoelectric cooling near the active laser region [78].

4.5 Summary

This chapter investigates the details of Joule heating (phonon emission) in silicon with the aid of Monte Carlo simulations including acoustic and optical phonon dispersion. Phonon generation is examined both in bulk and strained silicon, as well as in simple device geometries. The generated phonon distributions are different in bulk and strained silicon at low fields, but they are essentially the same at high fields. This difference is due to band splitting and scattering selection rules in strained silicon which favor g-type and reduce f-type phonon emission. It is found that the usual assumption of Joule heat being exclusively deposited in the optical modes only holds in strained silicon at low electric fields. In most practical situations (i.e. high electric fields, and bulk silicon at all fields) the emitted phonon spectrum is more evenly distributed among the phonon modes, although about half of it is relaxed into longitudinal optical phonons. The heat generation spectrum in strained and bulk silicon is similar at high fields, because electrons have enough energy to emit across the entire phonon spectrum despite the strain-induced band splitting.

Heat generation is also explored in simple device geometries, using Monte Carlo transport simulations self-consistent with the electric fields, which are updated through repeated solutions of the Poisson equation. Joule heating in short ballistic diodes is found to occur almost entirely in the drain, beyond the peak electric field region, as transport across the channel is quasi-ballistic. The heat generation region ("hot spot") extends deep into the drain, because the energetic electrons only relax their energy to the lattice over the length

of several inelastic mean free paths. A small heat absorption region is also found near the potential barrier to electron injection from the source into the channel. This is shown to be thermoelectric effect, consistent with the observations of previous researchers.

This work advances the state-of-the-art understanding of self-heating in silicon, and has applications to the engineering of devices that may require detailed knowledge of the heat generation spectrum. The approach can be similarly extended to other materials (e.g., germanium) or to low-dimensional structures (e.g., nanowires).

Chapter 5

Analysis of Thin-Body Device Scaling Including Self-Heating

This chapter explores the effects of confined dimensions and complicated geometries on the self-heating of ultra-thin body silicon-on-insulator (SOI) and germanium-on-insulator (GOI) devices near the limits of conventional scaling [1]. An electro-thermally self-consistent compact model is introduced for calculating device temperature, saturation current and intrinsic gate delay. The compact model also enables quick engineering estimates on the effect of various parameters (e.g. device geometry, interface thermal resistance) on device performance. The analysis assumes the heat is dissipated entirely in the drain for devices with very short (quasi-ballistic) channel lengths, a result of previous Monte Carlo simulations, The device operating temperature is found to be very sensitive to the choice of drain and channel extension dimensions, and to material boundary resistance. The analysis indicates that the raised device source/drain can be designed to simultaneously lower device temperature and parasitic capacitance, such that the intrinsic gate delay (CV/I) is optimal. It appears that a raised source/drain height approximately three times the channel thickness would be desirable both from an electrical and a thermal point of view. Furthermore, optimized GOI devices could provide at least 30 percent performance advantage over similar SOI devices, despite the lower thermal conductivity of the germanium layer.

5.1 Introduction

Ultra-thin body, fully depleted silicon-on-insulator (SOI) devices offer great promise for scaling near the end of the roadmap [1] due to better control of short channel effects, more immunity to latch-up and radiation effects, and lower parasitic capacitance [79, 80, 81, 82]. Such devices are built in a thin silicon layer on top of a thicker silicon dioxide layer (see, e.g., Fig. 1.2, Fig. 5.1 or Fig. 5.7). The buried oxide layer minimizes the depletion capacitance typically associated with the source/drain of a bulk device. The lower depletion capacitance enables faster switching speeds, as well as a better sub-threshold slope

$$S = \frac{k_B T}{e} \log\left(1 + \frac{C_d}{C_{ox}}\right) \tag{5.1}$$

where e is the elementary charge, C_{ox} is the capacitance of the gate oxide and C_d is the depletion capacitance. The use of a fully depleted transistor body and of the buried oxide layer helps decrease C_d, which consequently decreases the sub-threshold slope toward the theoretical minimum value of 60 mV/decade. The lower sub-threshold slope allows a decrease of the transistor threshold voltage (hence increasing the gate overdrive, and the drive current) while lowering the Drain Induced Barrier Lowering (DIBL) effect. Unfortunately, besides being a very good electrical insulator, the buried silicon dioxide layer is also a very good thermal insulator. Since its thermal conductivity is two orders of magnitude less than that of silicon (Table 1.1), SOI devices have usually been associated with self-heating problems [83]. This is somewhat alleviated if the buried oxide becomes thinner with scaling. However, confined dimensions and more complicated device geometries can still lead to significant self-heating, as the analysis shows in the rest of this chapter.

Very recently germanium-on-insulator (GOI) structures and devices have been reported [85, 86, 87], that could be even more attractive because germanium offers a mobility enhancement up to 2× compared to silicon, for both electrons and holes. However, the thermal conductivity of bulk germanium is only 40 percent as large as that of silicon, which combined with the poor thermal conductivity of the buried oxide may lead to worse thermal problems for GOI than those already well documented for SOI [20, 83]. The analysis in this

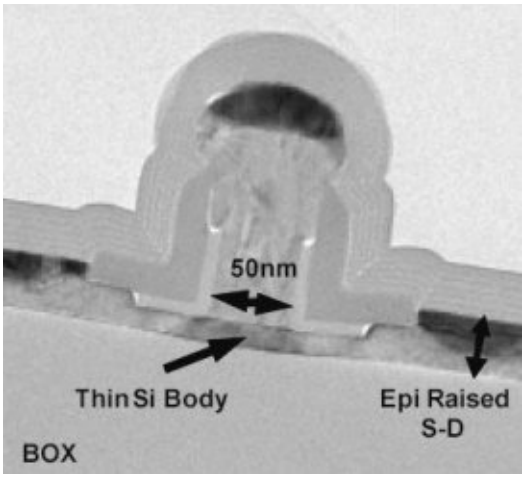

Figure 5.1: TEM (Transmission Electron Microscopy) image of a typical fully-depleted SOI transistor. The thin body is about 3-4× thinner than the gate length. The source and drain have been epitaxially raised to lower series resistance. The buried oxide (BOX) is not fully shown. Image courtesy Intel Corp. [84].

chapter investigates self-heating trends in similarly "well-behaved" SOI and GOI devices near the limits of scaling, and shows that despite the lower thermal conductivity of germanium, the temperature rise in GOI may be comparable to that in similar SOI devices, owing mainly to reduced power dissipation. The analysis also indicates that ultra-thin body GOI and SOI devices can be designed to provide optimal performance, taking self-heating into account self-consistently. The next few sections introduce several heat transfer issues in confined-geometry thin body transistors, which serve as ingredients for the self-consistent electro-thermal compact model presented in the latter part of the chapter.

5.2 Thin Film Thermal Conductivity

Previous work [15, 17] has suggested that sub-continuum heat conduction effects play a role in bulk nanotransistors because the area of most intense heat generation (i.e. the phonon "hot spot") is much smaller than the phonon mean free path ($\Lambda_b \sim 100$ nm in bulk silicon). In thin body transistors the phonon mean free path is limited by the body thickness (t_{si}), hence the sub-continuum effect predicted for bulk devices is of less importance. However,

the thermal conductivity of doped ultra-thin films is strongly reduced from the bulk value (148 $\mathrm{Wm^{-1}K^{-1}}$ for silicon) by phonon boundary and impurity scattering. From kinetic theory arguments, the thermal conductivity can be written as [7, 27]:

$$\kappa = \frac{1}{3} C_s \overline{v} \Lambda \tag{5.2}$$

where C_s is heat capacity per unit volume, $\overline{v}$ is the average phonon velocity and Λ is an appropriately averaged phonon mean free path, which is affected by boundary and impurity scattering. Note that the lattice vibrations (phonons) are the dominant heat carriers in semiconductors, since electrons have a negligible contribution even in heavily doped silicon [26]. Ju and Goodson [9] measured the thermal conductivity of undoped crystalline silicon to be halved in films of thickness $t_{si} \sim 100$ nm. This is due to phonon boundary scattering becoming the limiting factor of thermal conductivity. Other recent experimental work by Asheghi *et al.* [21] found the thermal conductivity decreasing about 30 percent in highly doped ($> 10^{19}$ $\mathrm{cm^{-3}}$) bulk silicon characteristic of modern devices. The combined thin film and impurity effects can be approximated by writing the phonon mean free path through Matthiessen's rule as

$$\frac{1}{\Lambda} = \frac{1}{\Lambda_b} + \frac{1}{t_{si}} + \frac{1}{\Lambda_{imp}} \tag{5.3}$$

where Λ_b is the phonon mean free path in undoped, bulk silicon (controlled only by anharmonic phonon-phonon interactions, and hence by temperature), and Λ_{imp} is the mean phonon-impurity scattering length, chosen to fit the data of Ref. [21]. This simple approach can be extrapolated to ultra-thin films that have not yet been experimentally measured. For example, the thermal conductivity of a highly doped 10 nm thin silicon film is estimated at 13 $\mathrm{Wm^{-1}K^{-1}}$, less than 9 percent of the bulk value. This is consistent with recent measurements on 20 nm thin silicon films that found $\kappa_{si} = 22$ $\mathrm{Wm^{-1}K^{-1}}$ [22], as well as with recent data on silicon nanowires [23], suggesting their thermal conductivity is on the order of 5 $\mathrm{Wm^{-1}K^{-1}}$ for a 20 nm diameter (which, as expected, is lower than the thermal conductivity of comparably thin silicon films). Some theoretical work has also

proposed that in ultra-thin films[1] changes in the phonon dispersion may reduce the phonon velocity and further decrease their thermal conductivity [24], although experimental results have not yet conclusively supported this. Figure 5.2 plots the thermal conductivity of undoped ultra-thin silicon and germanium films based on this Matthiessen's rule estimate for the phonon mean free path. The theoretically estimated thermal conductivity for thin silicon films is consistent with the experimental values mentioned earlier, but no thermal conductivity data for thin germanium films is yet available. The estimated ratio of the thermal conductivities, κ_{ge}/κ_{si}, is closer to unity (higher) in ultra-thin films than in bulk, where germanium (60 $\mathrm{Wm^{-1}K^{-1}}$) is only 40 percent as thermally conductive as silicon (148 $\mathrm{Wm^{-1}K^{-1}}$). In other words, germanium suffers a proportionally smaller decrease of thermal conductivity in thin films (a lesser "size effect") due to its shorter bulk phonon mean free path.[2] It should be noted that when heat transport is limited by phonon boundary scattering with the film thickness, the ultra-thin film thermal conductivity is largely independent of temperature [22].

5.3 Material Interface Thermal Resistance

When a heat flux flows across the interface between two dissimilar materials, a temperature difference usually develops. This can be modeled in terms of a boundary thermal resistance. If both materials are dielectrics (i.e. phonons are the heat flux carriers) this can be understood by considering that when the phonons reach the interface, some of them are transmitted through, while the remainder are reflected. There are two generally accepted models for this situation. The acoustic mismatch model (AMM) assumes that all phonons have a specular interaction with the interface [90], while the diffuse mismatch model (DMM) assumes phonons scatter diffusely with it [91]. The AMM is analogous to the impedance mismatch due to the refractive indices of two optically different materials,

[1] Of thickness 5 nm or less, i.e. comparable to the phonon wavelength.

[2] The phenomenon is similar to the size effect observed on the electrical conductivity of aluminum and copper lines of width comparable to the electron mean free path [88, 89]: bulk aluminum has a shorter electron mean free path (16 nm) than copper (38 nm), but in metal lines thinner than about 50 nm the electrical conductivity of aluminum is less affected by boundary scattering.

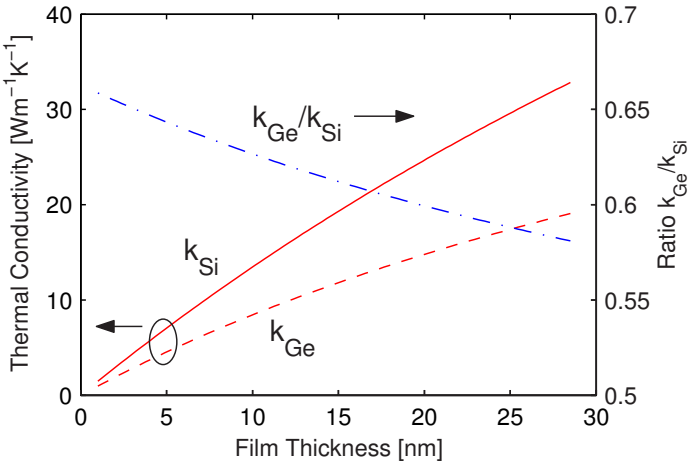

Figure 5.2: Estimated thermal conductivity of thin Si and Ge layers. As the film is thinned, the thermal conductivity decreases due to phonon boundary scattering, but it decreases less (vs. bulk) for Ge films due to the shorter phonon mean free path of this material. In bulk form, the thermal conductivity ratio is $\kappa_{ge}/\kappa_{si} = 60/148 \simeq 0.40$, but this fraction is closer to unity for ultra-thin films.

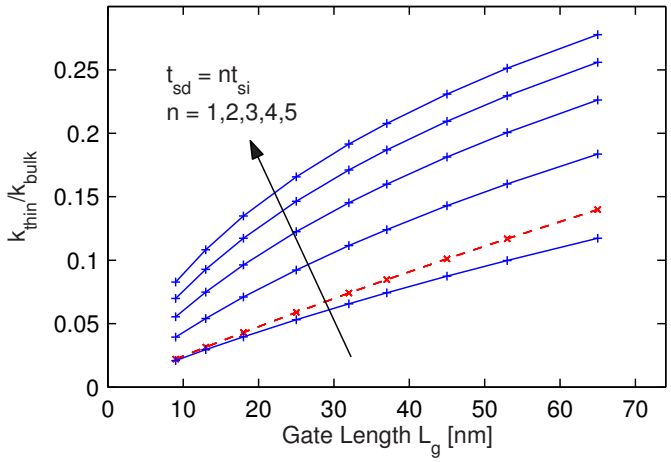

Figure 5.3: Thermal conductivity reduction of ultra-thin device layers, as a fraction of the bulk silicon thermal conductivity. The dashed line represents the thermal conductivity of the undoped channel (assumed to scale as $t_{si} = L_g/4$). The five solid lines are the thermal conductivity of the highly doped source/drain regions, with varying thicknesses from $t_{sd} = t_{si}$ (bottom) to $t_{sd} = 5t_{si}$ (top). For the shortest devices, the thermal conductivity of the thin layers can drop well below 10 percent of its value in bulk silicon (148 Wm^{-1}K^{-1}).

and it is generally appropriate at low temperatures, when the phonon wavelengths are much larger than the characteristic size of the interface roughness. On the other hand, the DMM is more adequate at room temperature, when the dominant transport phonons have a much shorter wavelength. In the latter case, the probability of phonon transmission across the interface depends on the ratio of phonon density of states between the two materials, and the maximum phonon frequency transmitted is limited by the material with the lower Debye temperature [17, 91]. While both models provide useful reference calculations, neither captures the complexity of the interaction between phonons and real interfaces, much of which is still not understood [92].

In the case of the interface between a metal and a dielectric, heat conduction is dominated by electrons in the metal and by phonons in the dielectric. Hence, for heat transport to occur across a metal-dielectric interface energy must be transferred between the electrons and the phonons. There are two possible scenarios, namely: (i) coupling between metal electrons and dielectric phonons through anharmonic interactions at the interface and (ii) coupling between electrons and phonons within the metal, followed by coupling between phonons in the metal and the dielectric. The role of the electron-phonon coupling at the metal-dielectric interface was only recently explored at room temperature [93]. The total interface thermal resistance is the sum, i.e. the series combination of the electron-phonon (in the metal) and the phonon-phonon (across the interface) thermal resistances. The latter can be estimated from the DMM, while the former was found to scale as $(C_e \kappa_p / \tau_{ep})^{-1/2}$, where C_e is the heat capacity of the metal electrons per unit volume, κ_p is the phonon thermal conductivity in the metal and τ_{ep} is the electron-phonon scattering time [93].

5.3.1 MOS Thermal Boundary Resistance

The room temperature thermal conductivity of bulk silicon dioxide ($\kappa_{ox} = 1.38$ $\mathrm{Wm^{-1}K^{-1}}$) is about two orders of magnitude less than that of silicon. Measurements on metal-oxide-silicon (MOS) structures have found that the thermal resistance of very thin oxide films does not scale linearly with their thickness and is larger than expected by a constant offset. This can been explained if it is assumed that the two interfaces between the MOS materials

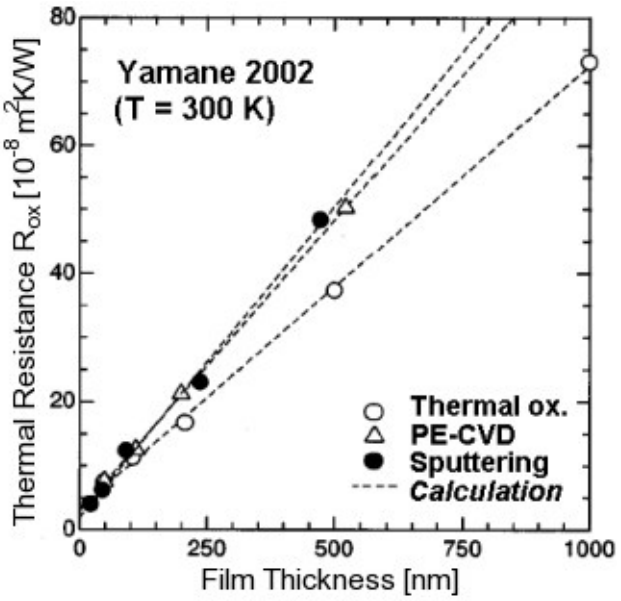

Figure 5.4: Measured metal-oxide-silicon thermal resistance, for various processing conditions, as a function of oxide film thickness (reproduced after Yamane *et al.* [25]).

have a non-negligible thermal resistance. The data of Refs. [25] and [94] can be interpreted if the total MOS thermal resistance is written as

$$R_{ox} = \frac{1}{A}\left(R_i + \frac{t_{ox}}{\kappa_{ox}}\right) \tag{5.4}$$

where t_{ox} is the thickness, A is the area of the oxide, and the total interface resistance R_i (including both interfaces) is about 2×10^{-8} $\mathrm{m^2KW^{-1}}$. The origin of this value is not well understood, since it is about an order of magnitude larger than the DMM theoretical predictions based on phonon dispersion mismatch at the boundaries [91]. This theoretical lower bound is only slightly enhanced by heat carrier (electron-phonon) conversion at the metal/oxide interface [93]. More subtle boundary effects may be playing a significant role, e.g. near-interfacial disorder in the metal, such as porosity or small grain size. However, the value of R_i was reported to be relatively independent of processing conditions (see Fig. 5.4) and is approximately equivalent to the thermal resistance of a 20 nm oxide film. This implies that the thermal resistance of the nanometer-thin silicon dioxide films

found in practical MOSFET devices with a metal gate is essentially independent of the oxide thickness. In other words smaller, thinner FETs with metal gates do not have an advantage of increased cooling through the gate oxide, despite a thinner insulator. Some high-k dielectrics have a higher thermal conductivity than SiO_2, but similar interface issues are likely to dominate their MOS thermal resistance as well, although no experimental data for them exists yet. The thermal resistance of a polysilicon-oxide-silicon stack is likely to be somewhat smaller because phonons are the heat carriers across both boundaries, and the additional electron-phonon scattering impedance in the metal is lacking. However, the polysilicon gate introduces a larger thermal resistance itself (compared to a metal gate), since the thermal conductivity of polysilicon is typically lower than that of most metals. Moreover, with the transition toward metal gates in future technologies, the metal-oxide-silicon thermal impedance is likely to remain an issue. In the remainder of this study, all devices being considered are near the limits of conventional scaling and they are assumed to have metal gates. Hence, the combined metal-oxide-silicon interface resistance from Eq. 5.4 and the discussion above is assumed to be present and is incorporated in the model.

5.3.2 Contact and Via Thermal Resistance

The source and drain contacts of a modern device are layered structures consisting of several dissimilar materials: silicon, silicide (e.g. $CoSi_2$) and the via metals (e.g. tungsten or copper). The copper used in vias and interconnects can diffuse into silicon dioxide and silicon, hence additional barrier metal liners (e.g. TiN) must be present around the vias and interconnects. The presence of these various material interfaces introduces not only an additional electrical resistance, but a thermal boundary resistance as well. The latter is typically ignored in thermal interconnect studies [95], and although the electrical properties of the contacts have been studied exhaustively [89, 96, 97], only one known (and currently outdated) study is available on the magnitude of their thermal resistance [98]. With continued device scaling, the thickness of the metal barrier cannot be reduced as rapidly as the interconnect dimensions because of reliability constraints, hence its effect on via electrical and thermal resistance is expected to be enhanced. Furthermore, for the

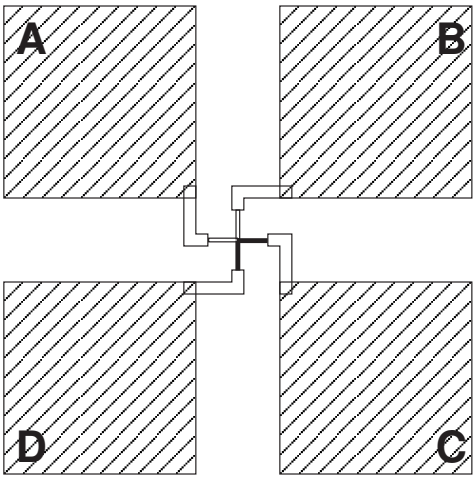

Figure 5.5: Typical Kelvin probe structure layout (top view), such as the one used in this work. Pads A and B are connected to the metal lines on top, C and D to the active (doped) region beneath. To obtain the contact resistance, a current is forced through pads A and C and the voltage drop is measured across B and D.

smallest devices the proximity of the contact to the active (channel) region will make both electrical and thermal transport at the contact a relevant part of the overall device behavior. As the heat generation region in the drain gets closer to the contact (see Section 4.4.1), thermal transport through the contact, and the possibility of device cooling through its vias may also become important.

In order to resolve the magnitude of the contact/via thermal resistance in a modern technology, several wafers were obtained from Texas Instruments.[3] Typical Kelvin probe structures were selected for contacts of diameter 0.13 μm and their electrical resistance was measured with the four-point probe technique. As shown in Fig. 5.5, a current was forced through pads A and C and the voltage drop was measured across pads B and D. The pads were then reversed (current forced through pads B and D, voltage measured across A and C) and the values of the obtained resistance were averaged. This measures the total electrical resistance of the copper via, in series with the contact resistance between via, liner metal and silicide, and with the contact resistance between silicide and the (highly n+ doped) silicon diffusion region. The metal and silicide resistances have a linear (monotonically

[3]The author is much indebted to Dr. Charvaka Duvvury.

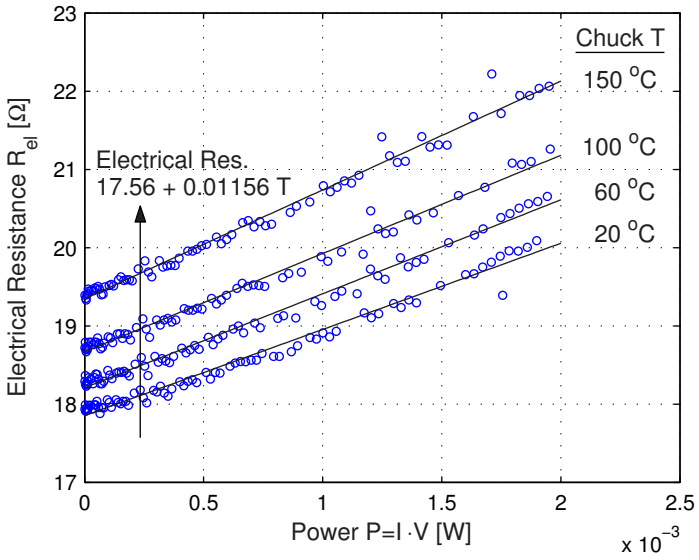

Figure 5.6: Electrical resistance measurement of a 0.13 μm diameter via and contact. The resistance is measured with the four-point probe technique both as a function of temperature and as a function of input power ($P = IV$).

increasing) dependence on temperature, while the silicide/silicon contact resistance has a more complicated dependence, which is roughly inversely proportional to temperature [96]. This is the case because thermionic emission over the silicide/silicon energy barrier is enhanced as the temperature increases, whereas tunneling through the barrier is not strongly affected. The metal resistance is linearly dependent on temperature due to a reduction of the electron-phonon scattering length as the temperature (and phonon occupation) increases. The four-point probe electrical resistance was measured as a function of chuck temperature, and a linear overall dependence was found:

$$R_{el}(T) = R_o\,[1 + \alpha(T - T_o)] \tag{5.5}$$

where the subscript o refers to the reference temperature (taken to be zero degrees Celsius here), and α is the temperature coefficient of resistance (TCR). Since the overall dependence of the measured resistance is linear with temperature, it is possible to surmise that the combined resistance of the metals (via, barrier and silicide) dominates the total via/contact

resistance. The experimental TCR is $\alpha \simeq 6.6 \times 10^{-4}$ K^{-1}, almost seven times smaller than the via (copper) TCR which is 4.5×10^{-3} K^{-1}. However, detailed knowledge of the various components of the electrical resistance is not necessary to determine the lumped *thermal* resistance of the via/contact. What is needed is simply the temperature dependence of this resistance (its calibration), which can then be used for electrical thermometry. If the electrical power ($P = IV$) flowing through the via/contact is increased, the contact is expected to self-heat and its electrical resistance will change accordingly. In other words, if $R_{el}(T)$ and $R_{el}(P)$ are both known, then a relationship between the temperature T and power P can be inferred, and consequently the *thermal* resistance is simply written as $R_{th} = T/P$. The results of such measurements, with electrical resistance as a function of varying input power, at different chuck temperatures, are summarized in Fig. 5.6. The dependence of electrical resistance on power is linear, just like its dependence on temperature. This implies another linear relationship between temperature and input power, the ratio of which is the thermal resistance:

$$\frac{dR_{el}}{dP} = \left(\frac{dR_{el}}{dT}\right)\left(\frac{dT}{dP}\right) = \alpha R_{th} \tag{5.6}$$

It should be noted that this linear relationship holds as long as resistance increase due to current crowding can be neglected, i.e. at the moderate current/power levels shown in Fig. 5.6. At higher current levels the electrical resistance rises faster (almost quadratically) with input power, and some of this increase is owed to current crowding in the small contact, not due to self-heating alone. The extracted thermal resistance itself has a small dependence on temperature, which is expected since the thermal conductivities of all materials involved decrease somewhat as the temperature is increased. The lumped thermal resistance of the 0.13 μm via/contact is extracted to be approximately $R_{th} \simeq 9.54 \times 10^4$ K/W at room temperature and 1.08×10^5 K/W at 100 °C. Part of this is a contribution from the spreading thermal resistance from the contact to the heat sink (chuck) at the back of the wafer, which can be estimated as

$$R_{sp} \simeq \frac{1}{2\pi\kappa_s d} \simeq 1.4 \times 10^4 \text{ K/W} \tag{5.7}$$

where κ_s is the thermal conductivity of the silicon wafer and $d = 0.13$ μm is the contact

diameter. This is the only thermal resistance component which would be different for a SOI wafer (rather than the bulk wafer these measurements were made on) hence it must be subtracted out. Without the spreading component, the remaining lumped via/contact thermal resistance is approximately in the range of $8.0 - 9.5 \times 10^4$ K/W for the normal operating temperature range of a transistor device.[4] This is the value for a 0.13 μm diameter contact, which is then scaled with the contact area and used in the compact thermal model introduced in the following sections. Assuming a circular area, the lumped thermal resistivity of the contact is approximately 1.6×10^{-9} $\mathrm{m^2KW^{-1}}$. This is about one order of magnitude less than the contact thermal resistivity for the metal-oxide-silicon (MOS) stack described in the previous section, but comparable to recently measured data on TiN/MgO interfaces [99]. Although relatively brief in its scope, the present analysis is the first study to explore the via/contact thermal resistance for modern, cobalt silicided contacts with copper vias. Especially with device scaling, the via/contact thermal resistance is expected to play a more important role not only for thermal device behavior under normal operating conditions (see below) but also during Electrostatic Discharge (ESD) events.

5.4 Ultra-Thin Body Device Thermal Model

A typical ultra-thin body (UTB) SOI or GOI device can be modeled using the thermal circuit in Figs. 5.7 (top) and 5.8 [20, 100]. All dimensions, thermal resistances and temperatures of interest are labeled on the two diagrams. A thermal circuit can be solved similarly to an electrical (resistive) network, with temperature being the equivalent of voltage and power (heat flow) the equivalent of current. The thermal circuit has an equivalent "Ohm's Law," i.e. $T = PR_{th}$, comparable with the electrical version where $V = IR_{el}$. A similar thermal circuit can be used for a FinFET device, but with the gate being wrapped over the channel such that the gate and gate oxide thermal resistances (R_g and R_{ox} respectively) are lower due to a larger surface area [20]. The FinFET body ("fin") resistance is also different due to additional boundary scattering from the limited fin height, itself typically less than

[4]This value is comparable to the thermal resistance found for 0.35 μm titanium silicided contacts with tungsten vias, in the only previously published work known to have measured it [98].

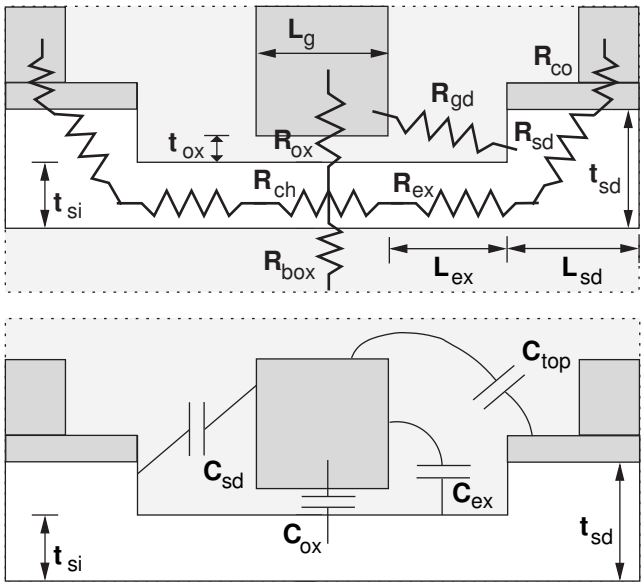

Figure 5.7: Ultra-thin body MOSFET and the thermal resistances (top) and parasitic capacitances (bottom) used in this model. The dark gray represents the metalized gate and contacts, and the light gray is the surrounding oxide insulator. The image is not drawn to scale.

the bulk silicon phonon mean path.

The thermal resistance of each device portion is written as $R = L/(\kappa A)$ where the material thermal conductivity κ is adjusted for boundary and impurity scattering, L is the dimension along the heat conduction direction and A is the cross-sectional area of heat flux perpendicular to it. The thermal circuit layout of the model is shown in Fig. 5.8. The thermal resistances are as defined in Fig. 5.7, and the subscript notation is consistent throughout. The resistances R_{xs} and R_{xd} are the source- and drain-side series resistances, including the thin channel extension. For example, $R_{xs} = R_{ex} + R_{sd}$ where $R_{ex} = L_{ex}/(\kappa_{si} t_{si} W)$ and the subscript si refers to the thin silicon body. Since the current specified by the ITRS guidelines is given per device width, and because all thermal resistances scale as the device width W, the entire problem can be scaled with this width, which can be safely dropped from the math when the thermal circuit solution is sought. The temperature T_d is defined at the point of maximum heat generation, which makes it the highest temperature, i.e. the worst-case scenario for a given device design. The temperature T_s is defined at the point directly under the source-side edge of the gate, making it the temperature which directly affects

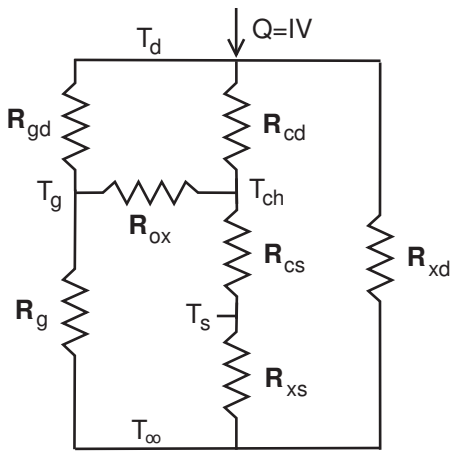

Figure 5.8: Equivalent thermal circuit of the thin-body FET. R_g is the gate thermal resistance, R_{cd} and R_{cs} are the drain- and source-side thermal resistance of the thin body channel. R_{xd} and R_{xs} contain R_{sd} from Fig. 5.7 in series with the drain- and source-side component of the thin channel extension, respectively. Other thermal resistances are defined in Fig. 5.7.

the channel injection velocity, backscattering coefficient and the maximum current drive of the device. The next section contains more details on how T_s is used to self-consistently compute the current drive of the device. Both T_d and T_s affect the temperature-dependent series resistance of the device.

This compact thermal model correctly reproduces the experimentally observed DC (device fully on) temperature rise in 100 nm channel length SOI devices [101]. Polonski and Jenkins [101] found a peak (steady-state) temperature rise of 96 ^{o}C above ambient for an input power level of 1.5 mW/μm, and a device thermal constant on the order of 100 ns, results which are consistent with their earlier electrical measurements [102]. Ref. [101] is the most recent known publication to have directly measured the temperature rise in modern devices, although unfortunately no results are yet available for sub-100 nm channel lengths. The model described in this dissertation is extended and applied precisely to this range for end-of-roadmap SOI and GOI devices. The design space for such devices is easier to explore from a modeling point of view, since hardware is difficult and expensive to manufacture, and some issues (e.g. ultra-thin germanium films, proper selection of gate material work-function) have not yet been fully addressed experimentally. This is a situation in which

modeling can guide experimental work, by providing and narrowing down the device design space and taking into consideration, for the first time, both electrical and thermal effects self-consistently.

The gate length (L_g), saturation current (I_d), nominal voltage (V_{dd}) and gate oxide thickness (t_{ox}) used in this study follow the most recent ITRS guidelines [1]. Other assumptions made regarding the device geometry are as follows. The SOI body thickness needed to ensure good electrostatics scales as $t_{si} = L_g/4$ [79]. The GOI body thickness should then scale by a factor of the material permittivity ratio, as $t_{ge} = \epsilon_{si} t_{si}/\epsilon_{ge} = 3t_{si}/4$. The buried oxide thickness scales as $t_{BOX} = 2L_g$ [1]. The thermal resistance of the buried oxide can be approximated as [83]

$$R_{BOX} \simeq \frac{1}{2W} \left(\frac{t_{BOX}}{\kappa_{ox} \kappa_{si} t_{si}} \right)^{1/2} \tag{5.8}$$

where W is the width of the device and κ_{si} is the thermal conductivity of the thin body of silicon, significantly reduced by boundary scattering (Fig. 5.2). The thin body is assumed to be essentially undoped to prevent dopant fluctuation effects on the threshold voltage. The threshold voltage is then mainly determined by the choice of gate metal workfunction, which in this work is taken to be a metallic alloy with a thermal conductivity of 40 $\mathrm{Wm^{-1}K^{-1}}$, typical of silicides. All other silicon regions (channel extension L_{ex}, source and drain L_{sd}) are highly doped to reduce series resistance, and their thermal conductivity is adjusted accordingly.

Electron mobility in thin germanium layers is about 2× higher than in thin silicon layers near room temperature [86, 103]. Recent devices built by Yu *et al.* [86] indicate this mobility enhancement means GOI devices can carry the same saturation current (I_d) at 40 percent lower voltage (V_{dd}) than comparable SOI transistors. This is the assumption used in the current work when comparing otherwise similar SOI and GOI transistors. However, since the FETs in Ref. [86] are large ($L_g = 10$ μm), this may be a conservative estimate for very small devices, where velocity saturation is less important and the 2× mobility advantage of germanium could play a stronger role [19, 104]. With the assumption used in this work, a

"well-behaved" GOI device dissipates 40 percent less power ($P = I_d V_{dd}$) than an equivalent SOI device, while generating the same drive current, as specified by the ITRS guidelines [1].

5.5 Temperature Dependence of Saturation Current

To estimate the temperature dependence of the saturation current (per unit width) for devices near the limit of scaling, the following simple model is employed [19]:

$$I_d = v_T \frac{\lambda}{2l + \lambda} C_{ox}(V_{gs} - V_t) \tag{5.9}$$

where v_T is the unidirectional thermal velocity, λ the electron mean free path (both at the source), l is the distance of the first $k_B T/e$ potential drop in the channel, C_{ox} the gate oxide capacitance per unit area and V_t is the threshold voltage. The various temperature dependencies are [19, 100, 105]:

$$v_T = v_{T_o}(T/T_o)^{1/2} \tag{5.10}$$

$$\lambda = \lambda_o(T/T_o)^{1/2+\alpha} \tag{5.11}$$

$$l = l_o(T/T_o) \tag{5.12}$$

$$V_t = V_{to} + \eta(T - T_o) \tag{5.13}$$

$$\mu = \mu_o(T/T_o)^{\alpha} \tag{5.14}$$

where the subscript o denotes the value at room temperature. Electron mobility in ultra-thin silicon layers has been experimentally found to vary as $T^{-1.4}$ ($\alpha = -1.4$) near room temperature, and to be largely independent of the layer thickness [12, 13]. This temperature coefficient of mobility enters the mean free path from $\lambda = 2\mu(k_B T/e)/v_T$ [105]. No data is yet available on the temperature dependence of electron mobility in ultra-thin germanium layers. However, it is well known that electron mobility in bulk (undoped) germanium is less temperature sensitive ($T^{-1.7}$) than in bulk (undoped) silicon ($T^{-2.4}$), due to the lower optical phonon energy in germanium. By extension, in this work the assumption is made

that the thin layer germanium mobility has a T^{-1} dependence.

The threshold voltage of ultra-thin body fully depleted devices varies linearly with temperature, with a coefficient η, which can be approximated as [106, 107]:

$$\eta \simeq \frac{\partial \phi_F}{\partial T} = \frac{k_B}{e}\left[\ln\left(\frac{N_a}{\sqrt{N_c N_v}}\right) + \frac{1}{2k_B}\frac{\partial E_g}{\partial T} - \frac{3}{2}\right] \tag{5.15}$$

where e is the elementary charge, N_a is the body doping, N_c and N_v are the conduction and valence band effective density of states and E_g is the band gap [107]. The temperature dependence of the band gap for silicon can be written empirically as in Eq. 3.2, and a similar relationship exists for germanium [107]. Recent experimental work [106] has found $\eta \simeq -0.7$ mV/K for fully depleted thin-body SOI devices. Although such data is not yet available for similar GOI devices, a quick estimate (accounting for the smaller germanium band gap, different conduction and valence band effective density of states) yields a nearly identical value of η, which will be used in this study.

Taking the above into account, a relationship for the temperature dependence of the saturation current for quasi-ballistic devices near the limit of scaling can be extracted:

$$\frac{\Delta I_d}{I_{do}} = \left[\frac{1}{T_o}\left(\frac{1}{2} + \frac{2\alpha - 1}{2 + \lambda_o/l_o}\right) - \frac{\eta}{V_{gs} - V_{to}}\right]\Delta T, \tag{5.16}$$

which is a generalization of the expression in Ref. [105]. All values with subscript o are taken to be at room temperature ($T_o = 300$ K), and in the rest of this work I_{do} and V_{to} are assumed to be the values of saturation current and threshold voltage, respectively, targeted by the ITRS guidelines [1]. It should be noted that the equation has two distinct components, one from the channel backscattering coefficient (i.e. the low-field mobility at the point of injection into the channel) and the other from the temperature dependence of the threshold voltage. The two components have opposite effects on the device saturation current. As the temperature increases, the mobility decreases due to additional scattering at the beginning of the channel (shorter mean free path, larger backscattering coefficient), which leads to a decrease in the saturation current. On the other hand, the threshold voltage is lowered with

an increase in temperature ($\eta < 0$), which gives a larger overdrive for a given gate voltage ($V_{gs} - V_t$) and therefore has a positive effect on the saturation current. This simple argument explains the experimentally observed dependence of saturation current on mobility (and on temperature), which is not a one-to-one relationship [102, 106].

The temperature rise due to self-heating used in the model above, $\Delta T = T_s - T_\infty$, is assumed to be that at the source end of the channel, since this is the region which affects the injection velocity, mean free path and threshold voltage in Eq. 5.9 and in the rest of this model. This temperature is computed with respect to the background circuit temperature T_∞, which is generally a function of device density, layout, surrounding circuit activity [108] and cooling technology in the packaging. In a modern chip, T_∞ can reach 360–380 K near the clock [3, 109] and this figure is expected to increase unless significant improvements are made in thermal packaging or circuit layout [4].

5.6 Self-Consistent Current Estimate

A self-consistent iterative solution of the device temperature and current based on the model in Fig. 5.7 and the discussion above was implemented. The total dissipated power ($I_d \times V_{dd}$) is assumed to be entirely generated in the device drain, based on the previous Monte Carlo simulation results (Section 4.4.1). This power is input to the thermal resistance model assuming (at first) the current to be the room-temperature value (I_{do}) targeted by the ITRS. The model yields a temperature rise at the source end of the channel (ΔT) which is used to adjust the current based on Eq. 5.16. The device power is reevaluated with the new current level as $P(T) = I(T)V_{dd}$, after which it is used again to solve for the device temperature, and this loop is repeated until the temperature and current are obtained self-consistently.

Figure 5.9 shows the calculated average temperature rise at 20 percent duty factor for SOI and GOI devices along the technology roadmap. The relationship between maximum (DC) temperature and the average temperature for a given duty factor f can be written as $T_{avg} = fT_{dc}$, since device thermal time constants (tens of nanoseconds) are much longer

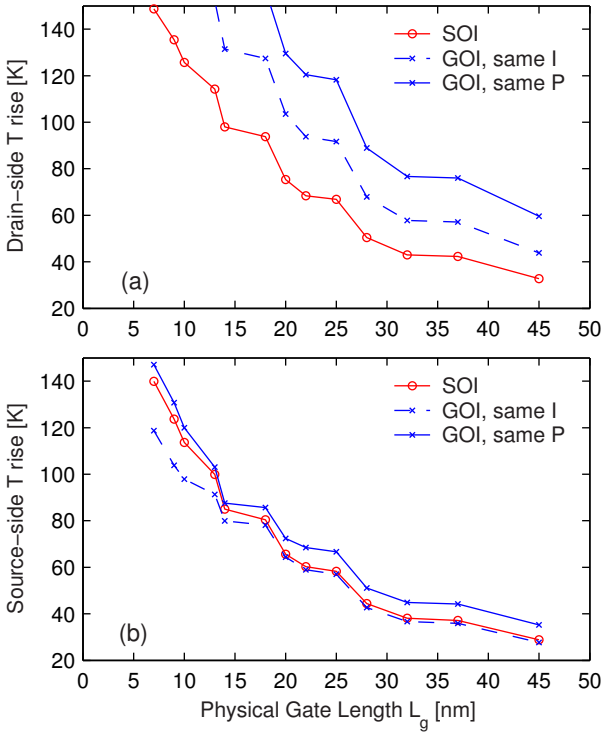

Figure 5.9: Self-consistently computed average drain- (a) and source-side (b) temperature rise in SOI and GOI devices operated with a duty factor of 20 percent. Two GOI cases are shown, one with the same current (but 40 percent lower V_{dd}) as the SOI, and one with the same power as the SOI. The raised SD thickness scales as $t_{sd} = 3t_{si}$ and the channel extension as $L_{ex} = L_g/2$.

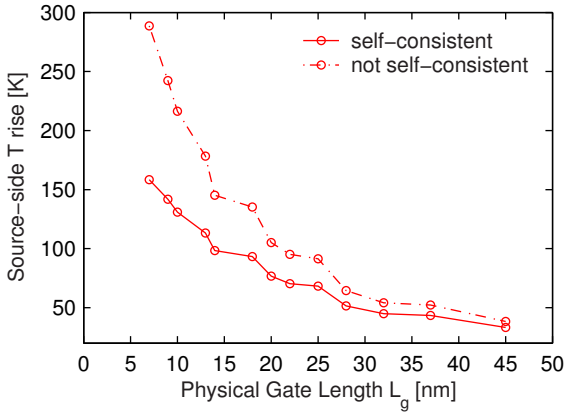

Figure 5.10: Comparison of SOI source-side temperature estimate obtained from the self-consistent temperature-current calculation (solid line) and a calculation where the current is not iteratively adjusted for changes in temperature (dash-dotted line). The temperature-current consistency is important, especially for the smallest devices where the error is near 100 percent.

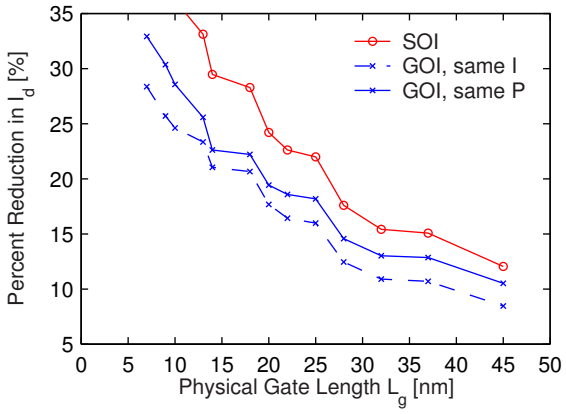

Figure 5.11: Self-consistently computed percentage decrease in drain current due to self-heating (vs. the ITRS-targeted current), for the same cases as in Fig. 5.9.

than device switching times (tens of picoseconds) [101]. Both same-current (but lower V_{dd}, hence lower power) and same-power scenarios are compared for GOI and SOI in Fig. 5.9. The drain temperature rise of GOI is expected to be higher in either case, due to the lower overall thermal conductivities. However, the source temperature rise is generally comparable, and even slightly lower for the same-current GOI vs. SOI case. This is due to the larger GOI channel thermal resistance, along with the lower dissipated power. Self-consistency is important in these calculations, since without it the temperature may be overestimated by close to 100 percent for the smallest devices, as shown in Fig. 5.10. Owing to their less temperature-sensitive mobility, GOI devices show less current degradation due to self-heating, as shown in Fig. 5.11.

5.7 Design Considerations

Since a thin device body and a relatively thicker buried oxide are required by electrostatics [79], and because heat transfer through the gate oxide and contacts is limited by interface thermal resistance, another way to ensure heat is more easily transferred out is to lower the thermal series resistance of the source and drain regions [20]. In other words, a (e.g., epitaxially) raised source/drain (SD) and shorter extension length L_{ex} are essential not only to reduce the electrical series resistance, but also to reduce the thermal series resistance of a

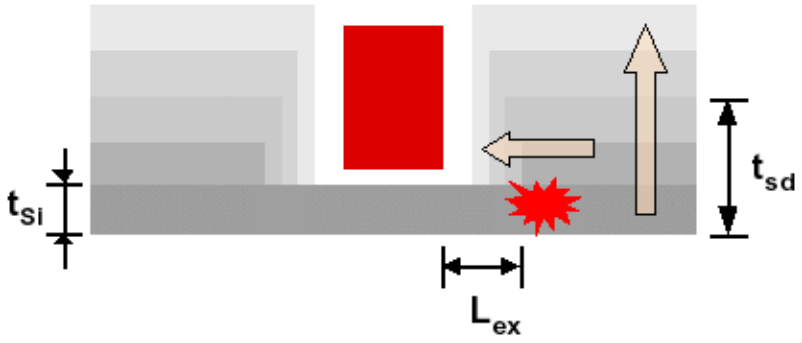

Figure 5.12: Possible changes in device source/drain geometry to reduce both its electrical and thermal series resistance, and hence lower its operating temperature. The extension length L_{ex} can be shortened and the source/drain t_{sd} can be epitaxially raised. The heat generation region ("hot spot") in the drain is also illustrated.

device, and therefore lower its operating temperature. Possible device design variations are illustrated in Fig. 5.12. Raising the source/drain height t_{sd} has a double thermal benefit: it improves the thermal conductivity of the region by allowing a larger phonon mean free path (Eq. 5.3), and it lowers the thermal resistance by offering a larger area for lateral heat flow. Shortening the extension length L_{ex} also helps move the heat generation region farther into the (larger) drain, away from the channel, and closer to the contact where it may be more easily dissipated. Among other constraints, the smallest achievable L_{ex} may in practice be limited by technological control over the spacer width. The choice of source/drain dimensions must also account for the geometry effect on dopant diffusion into the channel extension and possibly under the gate. From an electrostatic point of view, a raised source/drain and shorter L_{ex} will also increase the gate fringing capacitance (see bottom of Fig. 5.7). Hence, a clear trade-off exists between changes in geometry which may enhance the saturation current (by lowering device temperature) and the effect the same changes have on the device parasitic capacitance.

The performance impact (or penalty) of the modified source/drain can be quantified by estimating the intrinsic gate delay, $CV_{dd}/I_d(T)$. The gate capacitance components are shown in the bottom of Fig. 5.7, and modeled as in Ref. [110]. For example, the fringing

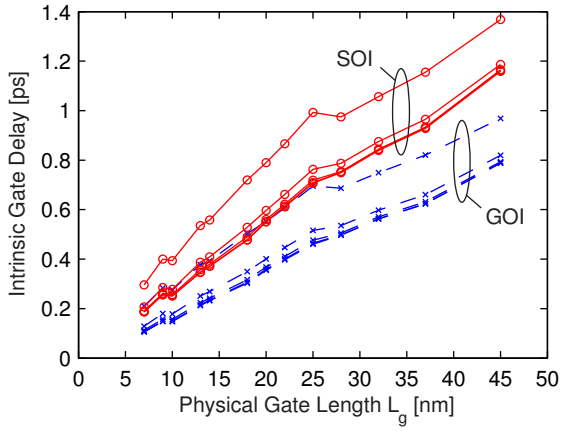

Figure 5.13: Self-consistently computed intrinsic delay for SOI and GOI devices in the same-current scenario. The SD height t_{sd} is varied as a parameter, from t_{si} (no raised SD, top line in each set of curves) to $5t_{si}$. The extension length is assumed constant at each node, $L_{ex} = L_g/2$. The intrinsic delay is not reduced significantly for $t_{sd} > 3t_{si}$.

component C_{ex} can be written as:

$$C_{ex} = \frac{2\beta\epsilon_{sw}}{\pi} \ln\left(1 + \frac{L_{ex}}{t_{ox}}\right) \tag{5.17}$$

where ϵ_{sw} is the dielectric constant of the sidewall material (assumed to be oxide in this study) and $\beta \simeq 0.8$ is a geometrical shape factor [110]. Figure 5.13 shows the computed intrinsic delay for SOI and GOI devices with the same drive current, but implicitly lower V_{dd} for GOI (as discussed earlier). The elevated source/drain lowers the device temperature and thus improves the drain current I_d, but at the same time increases the capacitance between gate and drain. For this reason, it appears that raising the source/drain thickness t_{sd} much beyond $3{\times}t_{si}$ does not result in significant additional speed gain: the trend lines for $t_{sd} = 3, 4$ and $5 \times t_{si}$ are essentially superposed both for SOI and GOI, at all device dimensions along the far-term roadmap. The intrinsic gate delay in this case was computed assuming the channel extension scales linearly with gate length, as $L_{ex} = L_g/2$.

Using a similar approach, the extension length and the source/drain height can also be optimized simultaneously for a given device gate length. Otherwise "well-behaved" 18 nm SOI (top) and GOI (bottom) devices are considered in Fig. 5.14, and the extension

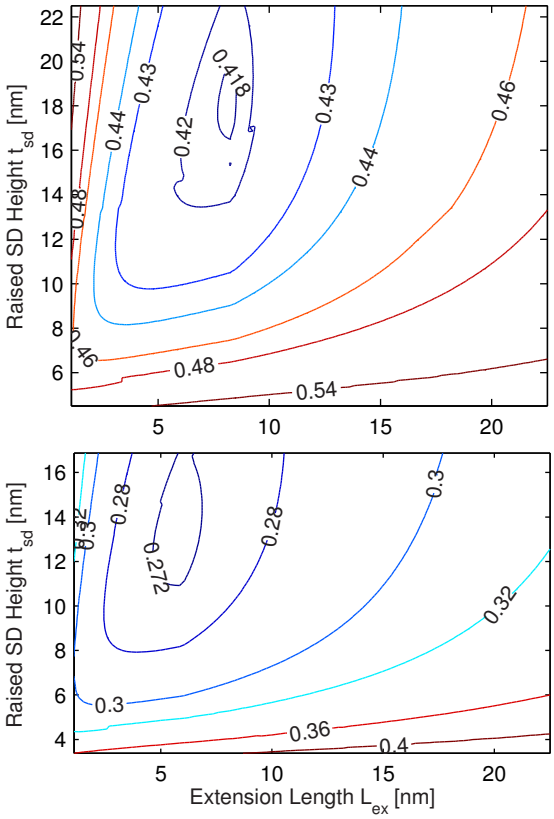

Figure 5.14: Geometry optimization to minimize intrinsic delay for a SOI (top) and GOI device (bottom) with L_g = 18 nm and t_{si} = 4.5 nm, assuming the GOI device provides the same current at 40 percent less V_{dd}. The results are expressed as contour plots of the delay (in picoseconds) with the extension length (L_{ex}) and SD thickness (t_{sd}) as parameters.

length and source/drain height are used as optimization parameters. Following an earlier discussion, the silicon and germanium body thickness are assumed to be t_{si}=4.5 nm and $t_{ge} = 3/4t_{si}$. The computed intrinsic delay (taking into account self-consistently the geometry effect on temperature and hence on current — and on parasitic capacitance) is plotted as a series of contours. The delay contours again suggest an optimal source/drain height around 3-4×t_{si} and an extension length approximately $L_g/3$ for GOI and closer to $L_g/2$ for the SOI device. Both of these targets may be practically achievable, although their exact effect on device self-heating remains to be seen experimentally. However, this theoretical scaling study provides a reasonable set of guidelines for future experimental device research

and fabrication.

Significant improvement in device heat dissipation and lowering the operating temperatures could also be achieved by lowering the contact thermal resistance, or the power input. However, aside from the study presented in Section 5.3.2 of this dissertation, little data is available on the thermal resistance of modern contact technologies, and it is still difficult to speculate how they may change in the future. It should be also be noted that throughout this study, the average temperature is seen to increases with decreasing device dimensions, despite the planned power reduction in the ITRS guidelines. This occurs primarily because the device volume scales faster, i.e. quadratically or cubically with the gate length (depending on device width) than the power $P = I_d V_{dd}$, which is planned to scale about linearly with the gate length and technology node [1]. Instead, a quadratic power scaling rule, e.g. $Q = -0.17L_g^2 + 27.5L_g$ with L_g in nm can be used for near-isothermal scaling of UTB-SOI devices down to the shortest gate lengths. The power budget for thin-body devices near 10 nm gate lengths may need to be as low as 250 μW/μm, suggesting operating voltages near 0.25 V for drive currents of 1000 μA/μm. All else being equal, dramatically increasing power densities will probably require a reduction in the ITRS power guidelines to achieve near-isothermal scaling of thin-body devices, especially below the 25 nm node.

5.8 Summary

This chapter compares the electro-thermal behavior of ultra-thin body fully depleted GOI and SOI devices near the limits of scaling. The thermal conductivity of ultra-thin silicon and germanium layers is investigated, and it is shown that for the thinnest technologically relevant films (a few nm thick) the thermal conductivity reduction of germanium is less severe than for silicon, owing to the former's shorter bulk phonon mean free path. Via and contact thermal resistance are measured experimentally, providing the first such study for modern device architectures (copper via, cobalt silicided diffusion), and a simple method for the thermal characterization of contacts, using already existing tools and test structures.

A self-consistent model for calculating device temperature, current and intrinsic gate

delay is introduced. It is shown that device temperatures are very sensitive to the drain and extension dimensions, and to material boundary resistance. Lowering the drain region's thermal resistance (e.g. by epitaxially raising it) can aid heat dissipation. While contact and material interface thermal properties are difficult to manage and predict, the device source/drain geometry could be designed to simultaneously minimize device temperature and parasitic capacitance, such that the intrinsic gate delay is optimal. A source/drain thickness that scales roughly as three times the body thickness seems optimal both from an electrical and a thermal point of view. In order to manage power dissipation at the smallest device lengths, ultra-thin body SOI and GOI devices ought to be used sparingly, in circuits with low duty factor, or with an operating power significantly lower (e.g. quadratically scaled with technology node) than current ITRS guidelines. Finally, it is found that optimized GOI devices could provide at least 30 percent performance advantage over similarly "well-behaved" SOI devices, even when self-heating is taken into account. There are strong indications that the ultimate device parameter design choices will need to involve thermal as well as electrical and technological considerations.

Chapter 6

Conclusions

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

6.1 Summary

This dissertation explored the details of Joule heat generation in silicon, and analyzed the possible scaling options of confined geometry (ultra-thin body) transistors, taking into account their self-heating. There are several important contributions of this work. First, a new Monte Carlo (MC) code named MONET was developed from the ground up, specifically aimed at computing heat (phonon) generation rates in bulk and strained silicon, and in simple mesoscopic device geometries. The model is different from previous work in that it uses an analytic (non-parabolic) description of the electron energy bands combined with an analytic (quadratic) phonon dispersion model, which distinguishes between the four phonon branches. This approach bridges the gap of computational tools between simple analytic-band MC codes [34, 44] and more complex full-band simulators [36, 43]. The implementation is computationally efficient (an order of magnitude faster than full-band MC), yet physically sophisticated, best suited for sub-band gap (1.1 V) device operating voltages, such as those of future technologies. A new, unified set of deformation potentials for electron-phonon scattering was introduced and shown to properly reproduce experimental transport data

in both bulk *and* strained silicon across a wide range of electric fields and temperatures (previous MC approaches have used separate sets of deformation potentials for bulk and strained silicon, without being able to reconcile them). The empirically fine-tuned coupling constants were extracted consistently with the band and phonon structure.

Comprehensive and electrostatically self-consistent one-dimensional device simulation capability was demonstrated. Two-dimensional device simulations were shown to be adequate for qualitative analysis. The current Monte Carlo approach gives information on both the location and spectral make-up of the heat generation region in a mesoscopic device. It was found that heat is dissipated almost entirely in the drain, near the contact, and not in the active region of short devices under quasi-ballistic transport conditions. This finding has implications for the design and engineering of mesoscopic devices and materials where self-heating is expected to play important role. The generated phonon distributions can be extracted (as demonstrated in Chapter 4) and used as inputs to a phonon transport solver [18]. The approach can be extended beyond silicon, to other materials (like germanium) or to strained or confined nanostructures, as long as the electron band and phonon dispersion relation are replaced by appropriate analytic expressions.

Another objective of this dissertation was to analyze the design and scaling of confined-geometry (ultra-thin body) transistors from an electro-thermal point of view. Several restrictions to heat flow in such devices were identified, such as thermal conductivity reduction due to phonon boundary scattering in ultra-thin layers, and interface thermal resistance between the different materials used. It was shown that the "size effect," i.e. thermal conductivity reduction in ultra-thin germanium films is less severe than in equally thin silicon films, due to the shorter bulk phonon mean free path of the former material. An experimental study was designed to extract the thermal resistance of via/contact structures for modern device technologies. The method uses standard test structures (Kelvin probes) and could be generalized and employed *in situ* during wafer testing. The temperature dependence of the contact resistance could also be used to determine limiting factors of the via/contact processing conditions. A carefully calibrated compact model for the self-heating of ultra-thin body transistors was introduced, incorporating the most recent understanding of heat

generation and flow at such dimensions. The model is electro-thermally self-consistent and useful for quick estimates on the effect of various device parameters (e.g. geometry, interface thermal resistance) on device performance. In particular, the device operating temperatures were found to be very sensitive to the choice of drain and channel extension dimensions. However, the self-consistent analysis indicated that the elevated device source/drain could be designed to simultaneously lower device temperature and parasitic capacitance, such that the intrinsic gate delay (CV/I) is optimal. It appears that a raised source/drain height approximately three times the channel thickness would be desirable both from an electrical and a thermal point of view in aggressively scaled, thin-body transistors. Furthermore, "well-behaved," optimized germanium-on-insulator (GOI) devices could provide at least 30 percent performance advantage over similar silicon-on-insulator (SOI) devices, despite the lower thermal conductivity of the thin germanium layer. This study could be used to guide future experimental device work, by providing and narrowing down the device design space and taking into consideration, for the first time, both electrical and thermal effects self-consistently.

6.2 Discussion and Suggestions for Future Work

Several unknowns relating to the electrical and thermal behavior of semiconductors at mesoscopic length scales have been identified during the course of this work. A deeper understanding of this research area is important from a physical point of view, while certain deficiencies may also be a limiting factor which could impede future nanoscale device design. Some of these issues are related to the electrical and thermal properties of nanometer-thin semiconductor films, still a very recent and quickly evolving research area. While the mobility and thermal conductivity of thin silicon films has been recently characterized [12, 22], no such data is yet available for thin germanium films. In this study, the thin film germanium mobility was assumed to scale as T^{-1} based on simple theoretical arguments, and its thermal conductivity was based on a Matthiessen's rule estimate, previously calibrated against the silicon data. Experimental studies must be carried out to find the validity of

these assumptions and to better understand the electro-thermal behavior of such thin films. More comprehensive Monte Carlo studies (possibly based on an extension of the method introduced in Chapter 3) could also be carried out to verify the temperature dependence of mobility [13].

6.2.1 Experimental Source/Drain Design

A number of other assumptions and findings relating to the compact thermal model introduced in this dissertation could be verified experimentally, especially if this work is used to guide future device design. Studies similar to those of Jenkins *et al.* [101, 102] could be carried out to probe the steady-state temperature rise of thin-body device geometries with raised source/drain. Similarly, more research must be done to find the thermal side-effects (if any) of replacing polysilicon with metal gates for future device designs. The metal may introduce an additional thermal interface resistance with the high-k dielectric, but its higher thermal conductivity (higher than polysilicon) may, on the other hand, compensate for it. Schottky (metal) source/drain devices may also benefit from the higher thermal conductivity of the metal, but no information is yet available on their thermal behavior. Similarly, since heat is almost entirely dissipated in the drain of ultra-scaled devices, any alternative drain designs are likely to affect the device thermal behavior as well, including its reliability. Heavily silicided or $Si_{1-x}Ge_x$ source/drain designs [111] may also have an impact, as silicides are comparable thermal conductors with thin-film silicon, while $Si_{1-x}Ge_x$ alloys are generally worse (Table 1.1).

6.2.2 Contact and Thermal Interface Resistance

With decreasing device dimensions, and even more rapidly decreasing device volume for heat dissipation, the interfaces between the device and its surroundings are expected to play a more important role, electrically as well as thermally. Interface effects can already be seen, such as carrier-boundary scattering which is responsible for both electrical as well as thermal conductivity reduction in ultra-thin films and wires. The larger surface-to-volume ratio also

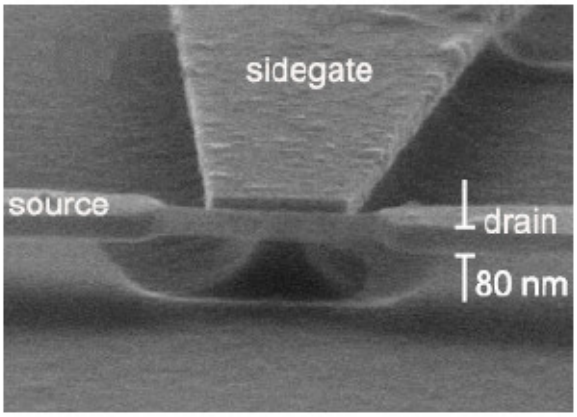

Figure 6.1: Typical suspended silicon nanowire field-effect device fabricated with electron beam lithography and a buffered HF underetch. The cross-section of this wire is 23×80 nm. The confined dimensions significantly alter both current and heat transport through the wire.

implies a stronger effect of material boundary resistance, while very little (aside from the studies reviewed or presented in this dissertation) is known about it. As more materials (e.g. high-k dielectrics, germanium, various silicides) are introduced in semiconductor processing, there is a growing need to understand the magnitude of boundary resistance between them, and its significance in future nanoscale device behavior. Approaches like the 3-ω method [25] can be used to investigate, for example, the thermal interface resistance between new materials immediately relevant to the semiconductor industry (e.g. metal gate with high-k dielectric, silicide with silicon). The via/contact resistance for future technologies should also be explored, possibly through the method proposed in this thesis. Additional data on contact and thermal boundary resistance would also help towards a better understanding (and modeling) of the atomic scale interaction at the interface between two materials.

6.2.3 Electro-Thermal Properties of Nanowires

Transistors made from semiconductor nanowires (NWs) may be the ultimate, electro-thermally confined, limiting case of confined geometry semiconductor devices. They have received a lot of attention for their current-carrying capabilities, low synthesis temperature

(as low as 275 ^{o}C for CVD-grown germanium NWs [112]), and relative ease of fabrication and integration with currently existing technologies [113, 114, 115]. Nanowires also present an ideal vehicle for studying (low-dimensional) electronic and thermal transport at nanometer length scales, as well as coupled electro-thermal transport. There are few data available on the mobility and thermal conductivity of NWs, but it is strongly believed that confined electron and phonon conduction play an important role in these devices. Recently measured silicon NW thermal conductivity suggests it is on the order of 5 $\mathrm{Wm^{-1}K^{-1}}$ for wires of 20 nm diameter [23], which (as expected) is lower than the thermal conductivity of comparably thin silicon films. The theoretical understanding of such transport is still poor, and key simulation tools are not yet available. This is but one area where intensive research must still be carried out. The low NW thermal conductivity combined with their good current-carrying ability imply tight and possibly limiting coupling (especially at high current levels) between electrons and the lattice. In other words, although such devices are known to be adversely affected by poor contact resistance, in practice their performance may be ultimately limited by self-heating. The design space for NW-based devices ought to be charted, taking into account self-heating effects, as well as the physics of transport in the confined geometry (quantized electron as well as phonon states). Only with a solid theoretical understanding from an electro-thermal point of view (and controlled growth on a large scale), could NWs perhaps complement (although not necessarily displace) currently existing CMOS technology.

6.2.4 Carbon Nanotubes

Unlike nanowires, carbon nanotubes (CNTs) cannot be synthesized at relatively low temperatures, but rather only in the 550-1100 ^{o}C range, depending on the method and catalyst used. However, both their electrical and thermal properties are outstanding, such that CNT-based devices may well represent the ultimate limit of nano-transistors. CNTs exhibit quantized, ballistic transport at low temperature and low bias, and extremely high current-carrying ability at high bias, up to 70 mA through a single-wall tube of diameter

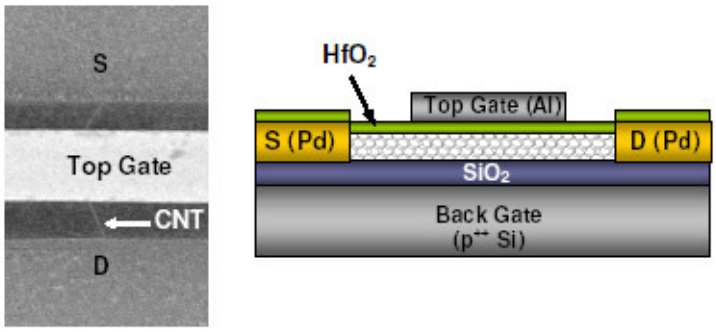

Figure 6.2: Carbon nanotube (CNT) transistor including low-resistivity (Ohmic) Pd contacts, high-k dielectric and dual gate control [116]. (Image and diagram courtesy A. Javey.)

1.4 nm [116]. Until recently, CNT performance was significantly hindered by large contact resistance; however with the introduction of low-resistivity Pd contacts, such devices are one step closer to realizing their potential applications [116]. The same work has also shown that for nanotubes much longer than the electron-optical phonon mean free path (MFP), high-bias electron transport is not truly ballistic. Despite the ballistic nature of CNTs at low-bias (where the intrinsic resistance is close to $h/4e^2$ or 6.5 kΩ), the electron-optical phonon emission dominates high-field transport, when electrons are able to gain energies more than 0.16 eV (the zone-boundary optical phonon energy in CNTs) [117]. In general, the electron-acoustic phonon interaction in CNTs is assumed to be elastic (electrons are simply backscattered, but their energy is not changed) while the electron-optical (zone center or zone edge) interaction is assumed to be strongly inelastic (electrons lose $\hbar\omega$ = 0.16-0.20 eV energy when emitting a high-energy phonon) [117]. The electron-optical phonon MFP from theoretical estimates is on the order of 30 nm [118], but the empirically extracted MFP is closer to 10 nm [116]. The two values have not yet been reconciled, but it is suspected that optical phonon (re)absorption may play a role in the MFP reduction. In other words, in spite of their high thermal conductivity [119], CNTs may still suffer from non-equilibrium self-heating effects, dependent on the overpopulation of optical modes and their decay into the faster acoustic modes. Optical phonon absorption or stimulated emission may be important if phonons cannot rapidly escape into the substrate, and if the

optical-acoustic decay times (currently, not well understood) are long enough under some conditions. Measurements at high bias on suspended CNTs could shed light on this possible non-equilibrium self-heating issue. Similarly, a more sophisticated high-bias transport theory including self-heating and contact effects must be introduced.

6.3 Epilogue

Future research must contribute both theoretically and experimentally to the body of knowledge on electro-thermal transport at mesoscopic length scales, and across material interfaces. The ultimate goals are to study limiting experimental scenarios and develop theory, models and code which are of both fundamental relevance and practical use. Much interesting work is yet to be done, and without a doubt this research area will remain "hot" (pun intended) for years to come.

Appendix A

MONET User Manual

The implementation of the Monte Carlo (MC) code MONET was described in Chapter 3 of this work. The program was written from the ground up, without borrowing any lines of code from previous MC programs (e.g. DAMOCLES, MOCA). Only the random number generator was copied from the Numerical Recipes book [74], but converted from type *float* to *double*. The rest of the implementation however follows that described with good detail in comprehensive reference works by Jacoboni [34] and Tomizawa [35]. The latter, in particular, is a great guide for the technical implementation of a MC code, featuring a lot of helpful examples written in pseudo-code.

MONET was written in C and compiled on the Linux operating system with the freely available Intel C Compiler (`icc`). The executable is run from the command line and can take a few options, a summary of which can be obtained with `monet -h`:

```
Usage: monet [options] infile logfile
  -d        Append date stamp to output directory
  -h        Print this help
  -r        Read scattering rates from scatrates.txt
  infile    <monet.in>
  logfile   <monet.log>
```

The output files are written in a directory called `out`, unless the `-d` switch is given, in

which case the output directory could be called, for example, `out-18Jul04-182833` if the simulation was started on July 18th, 2004 at 18:28 hours and 33 seconds.[1] Typically, MONET will compute the electron scattering rates at the beginning of every simulation, and will save them in a text file called `scatrates.txt` in the directory where MONET was called from. If several runs are made where the scattering rates should not change (i.e. the temperature, doping, strain and deformation potentials remain the same) it is useful, in the interest of saving some computational time, to re-read the scattering rates from the previous run. This can be done with the `-r` switch, which will read the scattering rates from the `scatrates.txt` file, if one exists from a previous run. Finally, MONET reads all other runtime parameters from a file called `monet.in`. The name of this file can be changed if it is specified on the command line when MONET is run, e.g., `monet otherinputfile`. The evolution of ensemble averages (energy, momentum, etc.) is written to the screen during the run of the code, as well as in a log file. The log file is typically called `monet.log` and it is written in the output directory. The name of the log file can also be overwritten from the command line, when MONET is invoked. A copy of the input file is saved in the output directory as well, as a reminder of what input parameters the respective results were obtained with.

A MONET simulation can be performed in zero, one and two dimensions. In "zero" dimensions it computes ensemble averages in an infinite resistor with constant applied electric field. This is useful for computing electron velocity-field curves, mobilities (in bulk or strained silicon) or heat generation rates for each phonon mode at a given electric field. MONET can also be used on a one (1-) or two dimensional (2-D) realistic device grid, with a few caveats. First, the grid must be equally spaced and rectangular (although in two dimensions the x-grid spacing is not required to equal the y-grid spacing). Second, in 2-D the Poisson equation is *not* solved along with the equations of motion, so the Monte Carlo results should only be regarded qualitatively. However, when simulating a 1-D cross-section of a device (e.g. a n^+nn^+ ballistic diode) the simulation is realistic, self-consistently solving

[1] This becomes useful when several runs are made with slightly modified input parameters, since it automatically avoids overwriting the `out` directory results from a previous run.

the Poisson equation every time step, as well as accounting for impurity scattering which varies with the doping concentration at each grid node. The 1- and 2-D grid simulations require an input file (see the parameter FIELDFILE below) with the grid coordinates, doping profiles and initial electric fields typically obtained from another device simulator run (e.g. the commercial code MEDICI). A PERL script is available to extract grid node information from MEDICI and translate it to the column input format that MONET requires. In this sense, the 2-D MONET simulation is essentially just a post-processor for a MEDICI run on the same grid.

A.1 Input File Description

The MONET input file (called `monet.in` by default) is read by the program at the beginning of each run. This file must be found in the directory MONET is invoked from, and it contains a number of parameters that can be modified without having to recompile the code. These parameters are listed below, with their description followed by the default value and units (if any) in parenthesis. The input file is a list of `PARAMETER = value` pairs, where the equal sign is mandatory, but the number of white spaces surrounding it (none or more) may vary. Any line beginning with a hash (`#`) or percent (`%`) sign is considered a comment and ignored. Similarly, any portion of a line following such a sign is also ignored.

A.1.1 Parameters of Relevance in All Dimensions

TEMP = ambient lattice temperature. (300 K)

DT = simulation time step. The simulation runs faster (fewer interruptions) when DT is larger, but its upper value is limited by the minimum grid spacing and by the maximum doping (plasma oscillations). The program will warn if either of these is exceeded. When the Poisson equation is self-consistently solved in 1-D to update the electric fields every DT (see the input parameter POISSON), this time step has to be reduced significantly, typically to fractions of a femtosecond. (10^{-14} sec)

SUBHIST = the number of time steps DT over which to compute intermediate ensemble

averages. It is assumed these "sub-histories" are roughly independent, so this number should be chosen such that the sub-history duration (SUBHIST × DT) is at least on the order of tenths of a picosecond. (50)

NELEC = number of electrons to simulate. (10000)

TTRANS = time (in seconds) after which transients are assumed to die and ensemble averages start to be computed. During TTRANS no averages are computed (i.e. the code runs somewhat faster), nothing is written to the log file and only dots are printed to the screen, one for every SUBHIST. If TTRANS is not user-specified in the input file, the program uses an internal algorithm to make an (admittedly, fairly rudimentary) attempt at estimating it. If possible, it is best advised for the user to make a test run with TTRANS = 0 and watch the ensemble averages printed to the screen, make a note of the approximate time it takes them to converge, then use that as the value for TTRANS in the next run(s).

TTOT = total time until the simulation ends. Note that final ensemble averages are computed over the time TTOT − TTRANS and sampled every SUBHIST × DT seconds.

EMAX = maximum electron energy allowed in the simulation. This should not be set much higher than 1.1 eV (the band gap energy) for an analytic-band code like MONET, which ignores impact ionization and high energy (e.g. L-valley) transport. If an electron exceeds this value, its energy is reset to EMAX and its momentum components appropriately rescaled. Also see Fig. 3.2, a comparison of the analytic band and full band density of states (DOS), indicating that EMAX < 2.0 eV is reasonable from the point of view of the DOS. (1.1 eV)

SEED = sets the starting seed (integer value) of the random number generator. If this is not set, MONET uses a value which is the (somewhat random) product of the process ID and the local time in milliseconds, at the moment when the command was invoked. When SEED is set, MONET will generate the same sequence of pseudo-random numbers, and hence the exact same results (ensemble averages) in consecutive runs, as long as no other input parameters are changed.

INPUTFILE = if this is set, MONET will read the starting electron distribution (position, momentum x y z components, energy, integer valley index) from the six columns of a

file with this name. If the simulation is on a 2-D grid, the input file has seven columns, with the first two being the x and y coordinates of the electrons.[2] Otherwise MONET initializes the electrons based on a Maxwellian distribution with average thermal energy ($3k_BT/2$) and randomly oriented momenta. Similarly, if INPUTFILE is not set and the simulation is to be performed in 1- or 2-D, MONET will distribute the initial particles in proportion to the doping density on the grid (or in proportion to the charge distribution read from MEDICI, if any).

OUTPUTFILE = if this is set, MONET will write the final electron distribution (position, momentum x y z components, energy, integer valley index) to a file with this name. This file can be used to save the status of the particle distribution for analysis, or such that a future run of the code can use it as its input (INPUTFILE) and thus avoid spending time going through the particle transients.

ALPHA = non-parabolicity parameter used in the description of the analytic electron energy bands, as in Eq. 3.1. (0.5 eV^{-1})

DTAf = deformation potential for f-type intervalley scattering with Transverse Acoustic (TA) phonons of energy near 19 meV. (5×10^7 eV/cm)

DLAf = deformation potential for f-type intervalley scattering with Longitudinal Acoustic and Optical (LA/LO) phonons of energy near 51 meV. Note that both LA and LO phonons can assist with this type of scattering near the edge of the Brillouin zone, where their dispersion meets. (3.5×10^8 eV/cm)

DTOf = deformation potential for f-type intervalley scattering with Transverse Optical (TO) phonons of energy near 57 meV. (1.5×10^8 eV/cm)

DTAg = deformation potential for g-type intervalley scattering with Transverse Acoustic (TA) phonons of energy near 10 meV. (3×10^7 eV/cm)

DLAg = deformation potential for g-type intervalley scattering with Longitudinal Acoustic (LA) phonons of energy near 19 meV. (1.5×10^8 eV/cm)

[2]The format is exactly the same as that of the OUTPUTFILE. For an example, set the OUTPUTFILE and look at the output, then use that as the INPUTFILE.

DLOg = deformation potential for g-type intervalley scattering with Longitudinal Optical (LO) phonons of energy near 62 meV. (6×10^8 eV/cm)[3]

XU = shear deformation potential (Ξ_u) used to compute intravalley acoustic phonon scattering. (6.8 eV)

XD = dilatation deformation potential (Ξ_d) used to compute intravalley acoustic phonon scattering. (1.1 eV)[4]

DIM = dimensionality of the problem (0, 1 or 2). Set DIM=0 to calculate transport parameters in a steady-state electric field (e.g. in an infinite resistor), such as the velocity-field curve, mobility, or heat generation rates for each phonon mode. Set DIM=1 or 2 to calculate electron transport and heat generation on a 1- or 2-D device grid. Note that the Poisson equation (see parameter POISSON below) can be solved self-consistently in 1-D, but not in 2-D. (0)

A.1.2 Parameters of Relevance in DIM=0 Simulations

EFSTART = starting value for the electric field (in V/cm) for a velocity-field, mobility, or steady-state heat generation calculation. This can also be a comma-separated list of electric fields, in which case the next 3 parameters (EFSTOP, EFSTEPS and LOGSTEPS) are ignored.

EFSTOP = last value of the electric field range (in V/cm) for a velocity-field or steady-state heat generation calculation. If EFSTART is a comma-separated list of electric field values then this parameter is ignored.

EFSTEPS = the (integer) number of steps MONET should take between the EFSTART and EFSTOP electric fields. The number of field values MONET runs simulations for is always EFSTEPS+1. For example, if EFSTEPS=4 then the program computes velocities, energies and other ensemble averages at the EFSTART, EFSTOP fields, and at another three values in between (see LOGSTEPS for how these are chosen). If EFSTEPS=0 then

[3]The six intervalley deformation potentials described above are summarized in Table 3.2 and discussed in Section 3.3.2.

[4]The intravalley deformation potentials XU and XD are discussed in Section 3.3.1.

MONET computes ensemble averages only for the EFSTART field value and ignores EFSTOP and LOGSTEPS. In this case of single field computation, the program also saves detailed heat generation statistics for each phonon polarization in the files `lphon.txt` (for the longitudinal acoustic and optical modes) and `tphon.txt` (for the transverse acoustic and optical modes). These files contain the net (emission minus absorption) number of phonons generated as a function of phonon energy, over the range of longitudinal and transverse phonon energies available. Both files can be found in the output directory, and can be used to create heat generation diagrams like those in Fig. 4.1. If EFSTART is a comma-separated list of electric field values then this parameter is ignored.

LOGSTEPS = used to determine either linear or logarithmic spacing of electric field values between EFSTART and EFSTOP. Set it to 0 for linear spacing (equally spaced fields) or to 1 for logarithmic spacing. If EFSTART is a comma-separated list of electric field values then this parameter is ignored.

EFDIR = direction of the applied electric field, e.g. 100, 110 or 111.

VERROR = ratio of velocity standard deviation to ensemble velocity average ($\Delta v/\overline{v}$, as computed over each simulation sub-history), used as a criterion for convergence. The simulation is ended early (before TTOT is reached) if $\Delta v/\overline{v} <$ VERROR. Set this value to zero if computing ensemble averages until TTOT is reached on every simulation run is desired.

DOPING = value of ionized impurity concentration (in cm^{-3}) to be used when calculating the impurity scattering rate in the sample being simulated. If not set, the sample is assumed undoped.

MOBCALC = enables (if set to 1) or disables (if set to 0) a detailed mobility calculation. If set to 1, the mobility is obtained around the (first) field value specified by EFSTART. This is done by calculating the average drift velocity at five field values centered around EFSTART, two slightly above this value and two slightly below, then interpolating to find an average mobility. (0)

SIGEX = sets the Ge percentage (x) of the $\mathrm{Si}_{1-\mathrm{x}}\mathrm{Ge}_{\mathrm{x}}$ buffer substrate for strained silicon mobility calculations. See Section 3.5.2 for a detailed discussion of transport in strained

silicon. (0)

A.1.3 Parameters of Relevance in 1-D or 2-D Simulations

FIELDFILE = the name of the file where the 1- or 2-D device grid is read from. The grid node position, electric field, mobile charge density, potential and doping density are stored in separate columns as follows:

`x (cm)` `Field (V/cm)` `Nelec (cm`$^{-3}$`)` `Pot (V)` `Doping (cm`$^{-3}$`)`

For a 2-D simulation an additional column must exist with the `y` grid coordinate (also in `cm`) between the position and electric field columns. This file can be extracted from a previous simulation with a commercial code like MEDICI, and a PERL script exists to convert the MEDICI output. Any line beginning with a hash (`#`) or percent (`%`) sign is considered a comment and ignored. If not user-supplied, the default FIELDFILE name is `efield1.txt` for 1-D simulations and `efield2.txt` for 2-D simulations.

POISSON = enables (if set to 1) or disables (if set to 0) the self-consistent solution of Poisson's equation in simulations on a 1-D device grid. If this is enabled, the Poisson equation is solved and the electric field is updated every time step DT. This maintains consistency between the motion of mobile charge in the device and the electric field, at the expense of some computational time and a tighter restriction on the choice of DT (generally less than a femtosecond) in order to avoid plasma oscillations. If POISSON is not enabled, the electric fields are read once from FIELDFILE and charge motion is simulated on this "frozen" field distribution. (0)

SURFSCAT = the ratio of diffuse to specular scattering events used for computing surface roughness scattering in a 2-D simulation. (0.15)

A.2 Example: Heat Generation in Bulk Si

The following input file can be used to compute the heat (phonon) generation profile in a bulk silicon sample doped to 10^{17} cm^{-3}, at an electric field of 50 kV/cm. The phonon

generation results are stored in the files `lphon.txt` and `tphon.txt` in the output directory, and plotting them will reproduce the result in Fig. 4.1(d). The phonon generation files are only written to the output directory when EFSTEPS=0 and the computation is done for a single field value.

```
TEMP = 300          # ambient lattice temperature
DT = 10.e-15        # time step
SUBHIST = 75        # sub-history for averaging (x DT)
NELEC = 10000       # number of electrons
TTRANS = 5e-12      # time after which transients are assumed to die
TTOT = 50e-12       # total simulation time
EMAX = 1.2          # maximum electron energy in the simulation
DIM = 0             # dimensionality of the problem
EFSTART = 50000     # electric field value
EFSTEPS = 0         # how many E-field steps (int)
EFDIR = 100         # direction of E-Field
VERROR = 0.002      # fraction Vstd/Vmean for convergence
DOPING = 1e17       # ionized impurity concentration
```

Note that turning on MOBCALC=1 and using, e.g., EFSTART=500 would enable the computation of low-field electron mobility. The program then calculates the mobility at five electric field values around 500 V/cm and reports the average. The simulation progress is printed to the screen as well as to the log file `monet.log` (along with the mobility value), which is stored in the output directory.

A.3 Example: Strained Si Velocity-Field Curve

The following input file can be used to compute the average electron velocity as a function of electric field for transport in a strained silicon layer grown on top of a $Si_{0.7}Ge_{0.3}$ substrate (x = 0.3).

```
TEMP = 300          # ambient lattice temperature
DT = 10.e-15        # time step
SUBHIST = 75        # sub-history for averaging (x DT)
NELEC = 10000       # number of electrons
TTRANS = 10e-12     # time after which transients are assumed to die
TTOT = 50e-12       # total simulation time
EMAX = 1.2          # maximum electron energy in the simulation
DIM = 0             # dimensionality of the problem
EFSTART = -50000    # can be comma-separated list too
EFSTOP = -50        # last value in E-field range
EFSTEPS = 8         # how many E-field steps (int)
LOGSTEPS = 1        # 1 log scale, 0 linear scale
EFDIR = 100         # direction of E-Field
VERROR = 0.0002     # fraction Vstd/Vmean for convergence
SIGEX = 0.3         # Si(1-X)Ge(X) percentage
```

The simulation starts at a field of -50 kV/cm and ends with a field of -50 V/cm, after taking eight logarithmically spaced steps.[5] This input file will generate the strained silicon velocity-field curve (open circles) from Fig. 3.6. The results are stored in the file `monet.log`, which also contains the net amount of phonons generated per branch of the spectrum, at each value of the electric field.

A.4 Example: 1-D Device Simulation

The following input file sets up a realistic 1-D device simulation, including impurity scattering and self-consistent updates of the electric field (Poisson equation) at every time step. The device grid, doping, starting potential and electric field values are read from the FIELDFILE `nin100nm1916.txt`, which (in this case) contained the layout of a n^+nn^+

[5] The choice of electric field sign is not essential, but a negative value will result in positive average electron velocities.

ballistic diode with a 10^{16} cm^{-3} doped and 100 nm long "channel" region, and 10^{19} cm^{-3} doped n^{+} regions.

```
TEMP = 300              # ambient lattice temperature
DT = 0.25e-15           # time step
SUBHIST = 400           # sub-history for averaging (x DT)
NELEC = 40000           # number of electrons
TTRANS = 5.0e-12        # time after which transients are assumed to die
TTOT = 55.0e-12         # total simulation time
EMAX = 1.5              # maximum electron energy in the simulation
OUTPUTFILE = monet.out          # write output electron distribution
DIM = 1                 # dimensionality of the problem
FIELDFILE = nin100nm1916.txt    # where to read device grid from
POISSON = 1             # calculate self-consistent field
```

Note the very short (less than a femtosecond) time step, to avoid plasma oscillations in the highly doped regions. Also note that the final electron distribution (position, momenta, energies) is written to the output file `monet.out`. Otherwise, the ensemble averages (electron density, velocity, potential, electric fields, at each grid node) are stored in the file `avgs.txt` in the output directory. This file is updated every SUBHIST time steps, and it can be replotted in "real time" to observe the progress and convergence of the simulation. The `avgs.txt` file is column-delimited, with the information stored as follows:

```
X [cm]  N [cm^-3]  E [eV]  V [cm/s]  Pot [V]  Fld [V/cm]  AcPhon  OpPhon
```

where the quantities are averaged at each grid node, over the simulation time that has passed since the end of TTRANS. The quantities AcPhon and OpPhon are the phonon energy generation rates at each grid node, in units of eV/cm^3/s. Plot and compare these (in particular, their sum) with the product $\mathbf{J} \cdot \mathbf{E}$, which is the classical way of computing heat generation inside a device. The results of such a simulation and comparison can be seen in Fig. 4.3.

Bibliography

[1] International Technology Roadmap for Semiconductors (ITRS). [Online]. Available: http://public.itrs.net

[2] S. Borkar, "Low power design challenges for the decade," in *Design Automation Conference*, Feb. 2001, pp. 293–296.

[3] C. C. Liu, J. Zhang, A. K. Dutta, and S. Tiwari, "Heating effects of clock drivers in bulk, SOI, and 3-D CMOS," *IEEE Electron Device Lett.*, vol. 23, p. 716, 2002.

[4] R. Mahajan, R. Nair, V. Wakharkar, J. Swan, J. Tang, and G. Vandentop, "Emerging directions for packaging technologies," *Intel Technology J.*, vol. 6, p. 62, 2002.

[5] D. B. Tuckerman and R. F. W. Pease, "High-performance heat sinking for VLSI," *IEEE Electron Device Lett.*, vol. 2, no. 5, pp. 126–129, 1981.

[6] J.-M. Koo, S. Im, L. Jiang, T. W. Kenny, J. G. Santiago, and K. E. Goodson, "VLSI hotspot cooling using two-phase microchannel convection," in *IMECE Tech. Dig.*, Nov. 2002.

[7] C. Kittel, *Introduction to solid state physics*, 7th ed. John Wiley & Sons, 1995.

[8] M. Lundstrom, *Fundamentals of carrier transport*, 2nd ed. Cambridge University Press, 2000.

[9] Y. S. Ju and K. E. Goodson, "Phonon scattering in silicon films with thickness of order 100 nm," *Appl. Phys. Lett.*, vol. 74, no. 20, pp. 3005–3007, 1999.

[10] E. Pop, R. W. Dutton, and K. E. Goodson, "Detailed heat generation simulations via the Monte Carlo method," in *SISPAD Tech. Dig.*, Sept. 2003, pp. 121–124.

[11] S. Mazumder and A. Majumdar, "Monte Carlo study of phonon transport in solid thin films including dispersion and polarization," *J. Heat Transfer*, vol. 123, pp. 749–759, 2001.

[12] D. Esseni, M. Mastrapasqua, G. K. Celler, C. F. L. Selmi, and E. Sangiorgi, "Low field electron and hole mobility of SOI transistors fabricated on ultrathin silicon films for deep submicrometer technology application," *IEEE Trans. Electron Devices*, vol. 48, no. 12, pp. 2842–2850, 2001.

[13] F. Gamiz, "Temperature behaviour of electron mobility in double-gate silicon on insulator transistors," *Semicond. Sci. Technol.*, vol. 19, pp. 113–119, 2004.

[14] D. K. Ferry, *Semiconductor transport.* Taylor & Francis, 2000.

[15] E. Pop, K. Banerjee, P. G. Sverdrup, R. W. Dutton, and K. E. Goodson, "Localized heating effects and scaling of sub-0.18 micron CMOS devices," in *IEDM Tech. Dig.*, Dec. 2001, pp. 677–680.

[16] G. Chen, "Nonlocal and nonequilibrium heat conduction in the vicinity of nanoparticles," *J. Heat Transfer*, vol. 118, pp. 539–545, 1996.

[17] P. G. Sverdrup, Y. S. Ju, and K. E. Goodson, "Sub-continuum simulations of heat conduction in silicon-on-insulator transistors," *J. Heat Transfer*, vol. 123, pp. 130–137, 2001.

[18] S. Sinha, E. Pop, and K. E. Goodson, "A split-flux model for phonon transport near hotspots," in *IMECE Tech. Dig.*, Nov. 2004.

[19] M. S. Lundstrom, "On the mobility versus drain current relation for a nanoscale MOSFET," *IEEE Electron Device Lett.*, vol. 22, no. 6, pp. 293–295, 2001.

[20] E. Pop, R. W. Dutton, and K. E. Goodson, "Thermal analysis of ultra-thin body device scaling," in *IEDM Tech. Dig.*, Dec. 2003.

[21] M. Asheghi, K. Kurabayashi, R. Kasnavi, and K. E. Goodson, "Thermal conduction in doped single-crystal silicon films," *J. Appl. Phys.*, vol. 91, no. 8, pp. 5079–5088, 2002.

[22] W. Liu and M. Asheghi, "Thermal conductivity of ultra-thin single crystal silicon layers, part I - experimental measurements at room and cryogenic temperatures," *J. Heat Transfer*, 2004.

[23] D. Li, Y. Wu, P. Kim, L. Shi, P. Yang, and A. Majumdar, "Thermal conductivity of individual silicon nanowires," *Appl. Phys. Lett.*, vol. 83, no. 14, pp. 2934–2936, 2003.

[24] A. Balandin and K. L. Wang, "Significant decrease of the lattice thermal conductivity due to phonon confinement in a free-standing semiconductor quantum well," *Phys. Rev. B*, vol. 58, pp. 1544–1549, 1998.

[25] T. Yamane, N. Nagai, S. Katayama, and M. Todoki, "Measurement of thermal conductivity of silicon dioxide thin films using a 3ω method," *J. Appl. Phys.*, vol. 91, no. 12, pp. 9772–9776, 2002.

[26] G. K. Wachutka, "Rigorous thermodynamic treatment of heat generation and conduction in semiconductor device modeling," *IEEE Trans. Electron Devices*, vol. 9, no. 11, pp. 1141–1149, 1990.

[27] J. Ziman, *Electrons and Phonons.* Oxford Press, 1960.

[28] A. A. Joshi and A. Majumdar, "Transient ballistic and diffusive phonon heat transport in thin films," *J. Appl. Phys.*, vol. 74, no. 1, pp. 31–39, 1993.

[29] U. Lindefelt, "Heat generation in semiconductor devices," *J. Appl. Phys.*, vol. 75, no. 2, pp. 942–957, 1994.

[30] R. Lake and S. Datta, "Energy balance and heat exchange in mesoscopic systems," *Phys. Rev. B*, vol. 46, no. 8, pp. 4757–4763, 1992.

[31] J. Lai and A. Majumdar, "Concurrent thermal and electrical modeling of sub-micrometer silicon devices," *J. Appl. Phys.*, vol. 79, no. 9, pp. 7353–7361, 1996.

[32] M. Artaki and P. J. Price, "Hot phonon effects in silicon field-effect transistors," *J. Appl. Phys.*, vol. 65, no. 3, pp. 1317–1320, 1989.

[33] P. Lugli and S. M. Goodnick, "Nonequilibrium longitudinal-optical phonon effects in GaAs-AlGaAs quantum wells," *Phys. Rev. Lett.*, vol. 59, no. 6, pp. 716–719, 1987.

[34] C. Jacoboni and L. Reggiani, "The Monte Carlo method for the solution of charge transport in semiconductor with applications to covalent materials," *Rev. Mod. Phys.*, vol. 55, no. 3, pp. 645–705, 1983.

[35] K. Tomizawa, *Numerical simulation of submicron semiconductor devices.* Artech House, 1993.

[36] M. V. Fischetti and S. E. Laux, "Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects," *Phys. Rev. B*, vol. 38, no. 14, pp. 9721–9745, 1988.

[37] E. Pop, R. W. Dutton, and K. E. Goodson, "Analytic band Monte Carlo model for electron transport in Si including acoustic and optical phonon dispersion," *J. Appl. Phys.*, vol. 96, no. 7, Oct. 2004.

[38] C. Canali, C. Jacoboni, F. Nava, G. Ottaviani, and A. Alberigi-Quaranta, "Electron drift velocity in silicon," *Phys. Rev. B*, vol. 12, no. 4, pp. 2265–2284, 1975.

[39] J. Y. Tang and K. Hess, "Impact ionization of electrons in silicon (steady state)," *J. Appl. Phys.*, vol. 54, no. 9, pp. 5139–5144, 1983.

[40] N. Sano, T. Aoki, M. Tomizawa, and A. Yoshii, "Electron transport and impact ionization in Si," *Phys. Rev. B*, vol. 41, no. 17, pp. 12 122–12 128, 1990.

[41] B. Fischer and K. R. Hofmann, "A full-band Monte Carlo model for the temperature dependence of electron and hole transport in silicon," *Appl. Phys. Lett.*, vol. 76, no. 5, pp. 583–585, 2000.

[42] P. D. Yoder and K. Hess, "First-principles Monte Carlo simulation of transport in Si," *Semicond. Sci. Technol.*, vol. 9, pp. 852–854, 1994.

[43] T. Kunikiyo, M. Takenaka, Y. Kamakura, M. Yamaji, H. Mizuno, M. Morifuji, K. Taniguchi, and C. Hamaguchi, "A Monte Carlo simulation of anisotropic electron transport in silicon including full band structure and anisotropic impact-ionization model," *J. Appl. Phys.*, vol. 75, no. 1, pp. 297–312, 1994.

[44] T. Yamada, J.-R. Zhou, H. Miyata, and D. K. Ferry, "In-plane transport properties of $Si/Si_{1-x}Ge_x$ structure and its FET performance by computer simulation," *IEEE Trans. Electron Devices*, vol. 41, pp. 1513–1522, 1994.

[45] C. Jacoboni, R. Minder, and G. Maini, "Effects of band non-parabolicity on electron drift velocity in silicon above room temperature," *J. Phys. Chem. Solids*, vol. 36, pp. 1129–1133, 1975.

[46] R. Brunetti, C. Jacoboni, F. Nava, and L. Reggiani, "Diffusion coefficient of electrons in silicon," *J. Appl. Phys.*, vol. 52, no. 11, pp. 6713–6722, 1981.

[47] B. Winstead and U. Ravaioli, "A quantum correction based on Schrodinger equation applied to Monte Carlo device simulation," *IEEE Trans. Electron Devices*, vol. 50, no. 2, pp. 440–446, 2003.

[48] 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 Trans. Electron Devices*, vol. 45, no. 4, pp. 867–876, 1998.

[49] F. M. Bufler, Y. Asahi, H. Yoshimura, C. Zechner, A. Schenk, and W. Fichtner, "Monte Carlo simulations and measurement of nanoscale n-MOSFETs," *IEEE Trans. Electron Devices*, vol. 50, no. 2, pp. 418–424, 2003.

[50] C. Jungemann and B. Meinerzhagen, "On the applicability of nonself-consistent Monte Carlo device simulation," *IEEE Trans. Electron Devices*, vol. 49, no. 6, pp. 1072–1074, 2002.

[51] M. A. Green, "Intrinsic concentration, effective densities of states, and effective mass in silicon," *J. Appl. Phys.*, vol. 67, no. 6, pp. 2944–2954, 1990.

[52] D. Long, "Scattering of conduction electrons by lattice vibrations in silicon," *Phys. Rev.*, vol. 120, no. 6, pp. 2024–2032, 1960.

[53] C. Hamaguchi, *Basic semiconductor physics.* Springer, 2001.

[54] G. Dolling, "Lattice vibrations in crystals with the diamond structure," in *Symposium on Inelastic Scattering of Neutrons in Solids and Liquids*, Sept. 1963, pp. 37–48.

[55] C. Herring and E. Vogt, "Transport and deformation-potential theory for many-valley semiconductors with anisotropic scattering," *Phys. Rev.*, vol. 101, no. 3, pp. 944–961, 1956.

[56] A. Haug, *Theoretical solid state physics.* Pergamon Press, 1972, vol. 2.

[57] H. Mizuno, K. Taniguchi, and C. Hamaguchi, "Electron-transport simulation in silicon including anisotropic phonon scattering rate," *Phys. Rev. B*, vol. 48, no. 3, pp. 1512–1516, 1993.

[58] M. V. Fischetti and S. E. Laux, "Band structure, deformation potentials, and carrier mobility in strained Si, Ge, and SiGe alloys," *J. Appl. Phys.*, vol. 80, no. 4, pp. 2234–2252, 1996.

[59] ——, "Monte Carlo study of electron transport in silicon inversion layers," *Phys. Rev. B*, vol. 48, no. 4, pp. 2244–2274, 1993.

[60] P. Y. Yu and M. Cardona, *Fundamentals of Semiconductors.* Springer, 1996.

[61] M. H. Jørgensen, "Electron-phonon scattering and high-field transport in n-type Si," *Phys. Rev. B*, vol. 18, no. 10, pp. 5657–5666, 1978.

[62] E. Sangiorgi, B. Ricco, and F. Venturi, "MOS^2: an efficient MOnte Carlo simulator for MOS devices," *IEEE Trans. Computer-Aided Design*, vol. 7, pp. 259–271, 1988.

[63] B. Ridley, "Reconciliation of the Conwell-Weisskopf and Brooks-Herring formulae for charged-impurity scattering in semiconductors: third-body interference," *J. Phys. C: Solid State Phys.*, vol. 10, pp. 1589–1593, 1977.

[64] H. Kosina and G. Kaiblinger-Grujin, "Ionized-impurity scattering of majority electrons in silicon," *Solid-State Electronics*, vol. 42, no. 3, pp. 331–338, 1998.

[65] H. Kosina, "A method to reduce small-angle scattering in Monte Carlo device analysis," *IEEE Trans. Electron Devices*, vol. 46, no. 6, pp. 1196–1200, 1999.

[66] K. Ismail, S. Nelson, J. Chu, and B. Meyerson, "Electron transport properties of Si/SiGe heterostructures: measurements and device implications," *Appl. Phys. Lett.*, vol. 63, no. 5, pp. 660–662, 1993.

[67] E. Pop, "CMOS inverse doping profile extraction and substrate current modeling," Master's thesis, M.I.T., 1999.

[68] S. Nelson, K. Ismail, J. Chu, and B. Meyerson, "Room-temperature electron mobility in strained Si/SiGe heterostructures," *Appl. Phys. Lett.*, vol. 63, no. 3, pp. 367–369, 1993.

[69] M. M. Rieger and P. Vogl, "Electronic-band parameters in strained $Si_{1-x}Ge_x$ alloys on $Si_{1-y}Ge_y$ substrates," *Phys. Rev. B*, vol. 48, no. 19, pp. 14 276–14 287, 1993.

[70] D. Chen, E. Sangiorgi, M. R. Pinto, E. C. Kan, U. Ravaioli, and R. W. Dutton, "Analysis of spurious velocity overshoot in hydrodynamic simulations," in *NUPAD Tech. Dig.*, Sept. 1992, pp. 109–114.

[71] D. L. Woolard, H. Tian, M. A. Littlejohn, and K. W. Kim, "The implementation of physical boundary conditions in the Monte Carlo simulation of electron devices," *IEEE Trans. Computer-Aided Design*, vol. 13, pp. 1241–1246, 1994.

[72] R. W. Hockney and J. W. Eastwood, *Computer simulation using particles.* IOP Publishing, 1988.

[73] K. Hess, Ed., *Monte Carlo device simulation: full band and beyond.* Kluwer, 1991.

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

[75] J. H. Ferziger, *Numerical methods for engineering applications*, 2nd ed. Wiley, 1998.

[76] C. Jungemann, A. Emunds, and W. L. Engl, "Simulation of linear and nonlinear electron transport in homogenous silicon inversion layers," *Solid-State Electronics*, vol. 36, no. 11, pp. 1529–1540, 1993.

[77] MONET. [Online]. Available: http://nanoheat.stanford.edu

[78] K. P. Pipe, R. J. Ram, and A. Shakouri, "Internal cooling in a semiconductor laser diode," *IEEE Phot. Tech. Lett.*, vol. 14, no. 4, pp. 453–455, 2002.

[79] H.-S. P. Wong, D. J. Frank, and P. M. Solomon, "Device design considerations for double-gate, ground-plane and single-gated ultra-thin SOI MOSFET's at the 25 nm channel length generation," in *IEDM Tech. Dig.*, Dec. 1998, pp. 407–410.

[80] B. Doris, M. Ieong, H. Zhu, Y. Zhang, M. Steen, W. Natzle, S. Callegari, V. Narayan, J. Cai, S. H. Ku, P. Jamison, Y. Li, Z. Ren, V. Ku, D. Boyd, T. Kanarsky, C. D. Emic, M. Newport, D. Dobuzinsky, S. Deshpande, J. Petrus, R. Jammy, and W. Haensch, "Device design considerations for ultra-thin SOI MOSFETs," in *IEDM Tech. Dig.*, Dec. 2003, pp. 631–634.

[81] T. Ernst, S. Cristoloveanu, G. Ghibaudo, T. Ouisse, S. Horiguchi, Y. Ono, and Y. Takahashi, "Ultimately thin double-gate SOI MOSFETs," *IEEE Trans. Electron Devices*, vol. 50, no. 3, pp. 830–838, 2003.

[82] J.-P. Colinge, *Silicon-on-Insulator Technology: Materials to VLSI*, 3rd ed. Kluwer Academic, 2004.

[83] L. T. Su, J. E. Chung, D. A. Antoniadis, K. E. Goodson, and M. I. Flik, "Measurement and modeling of self-heating in SOI nMOSFET's," *IEEE Trans. Electron Devices*, vol. 41, no. 1, pp. 69–75, 1994.

[84] B. Doyle, R. Arghavani, D. Barlage, S. Datta, M. Doczy, J. Kavalieros, A. Murthy, and R. Chau, "Transistor elements for 30 nm physical gate length and beyond," *Intel Technology J.*, vol. 6, pp. 42–54, 2002.

[85] N. A. Bojarczuk, M. Copel, S. Guha, V. Narayanan, E. J. Preisler, F. M. Ross, and H. Shang, "Epitaxial silicon and germanium on buried insulator heterostructures and devices," *Appl. Phys. Lett.*, vol. 83, no. 26, pp. 5443–5445, 2003.

[86] D. S. Yu, C. H. Huang, A. Chin, C. Zhu, M. F. Li, B. J. Cho, and D.-L. Kwong, "$Al_2O_3 - Ge - On - Insulator$ n- and p-MOSFETs with fully NiSi and NiGe dual gates," *IEEE Electron Device Lett.*, vol. 25, no. 3, pp. 138–140, 2004.

[87] D. S. Yu, K. C. Chiang, C. F. Cheng, A. Chin, C. Zhu, M. F. Li, and D.-L. Kwong, "Fully silicided $NiSi : Hf - LaAlO_3/SG - GOI$ n-MOSFETs with high electron mobility," *IEEE Electron Device Lett.*, vol. 25, no. 8, pp. 559–561, 2004.

[88] W. Steinhögl, G. Schindler, G. Steinlesberger, and M. Engelhardt, "Size-dependent resistivity of metallic wires in the mesoscopic range," *Phys. Rev. B*, vol. 66, p. 75414, 2002.

[89] P. Kapur, J. P. McVittie, and K. C. Saraswat, "Technology and reliability constrained future copper interconnects – part I: resistance modeling," *IEEE Trans. Electron Devices*, vol. 49, no. 4, pp. 590–597, 2002.

[90] W. A. Little, "The transport of heat between dissimilar solids at low temperatures," *Can. J. Phys.*, vol. 37, pp. 334–349, 1959.

[91] E. T. Swartz and R. O. Pohl, "Thermal boundary resistance," *Rev. Mod. Phys.*, vol. 61, no. 3, pp. 605–668, 1989.

[92] D. G. Cahill, W. K. Ford, K. E. Goodson, G. D. Mahan, A. Majumdar, H. J. Maris, R. Merlin, and S. R. Phillpot, "Nanoscale thermal transport," *J. Appl. Phys.*, vol. 93, no. 2, pp. 793–818, 2003.

[93] A. Majumdar and P. Reddy, "Role of electron-phonon coupling in thermal conductance of metal-nonmetal interfaces," *Appl. Phys. Lett.*, vol. 84, no. 23, pp. 4768–4770, 2004.

[94] S.-M. Lee and D. G. Cahill, "Heat transport in thin dielectric films," *J. Appl. Phys.*, vol. 81, no. 6, pp. 2590–2595, 1997.

[95] T.-Y. Chiang, K. Banerjee, and K. C. Saraswat, "Analytical thermal model for multilevel VLSI interconnects incorporating via effect," *IEEE Electron Device Lett.*, vol. 23, no. 1, pp. 31–33, 2002.

[96] K.-H. Oh, J.-H. Chun, K. Banerjee, C. Duvvury, and R. W. Dutton, "Modeling of temperature dependent contact resistance for analysis of ESD reliability," in *Proc. IRPS*, Dec. 2003, pp. 249–255.

[97] C.-N. Liao and K.-C. Chen, "Current crowding effect on thermal characteristics of Ni/doped-Si contacts," *IEEE Electron Device Lett.*, vol. 24, no. 10, pp. 637–639, 2003.

[98] K. Banerjee, A. Amerasekera, G. Dixit, and C. Hu, "Temperature and current effects on small-geometry-contact resistance," in *IEDM Tech. Dig.*, Dec. 1997, pp. 115–118.

[99] R. M. Costescu, M. A. Wall, and D. G. Cahill, "Thermal conductance of epitaxial interfaces," *Phys. Rev. B*, vol. 67, p. 54302, 2003.

[100] E. Pop, C. O. Chui, S. Sinha, R. W. Dutton, and K. E. Goodson, "Electro-thermal comparison and performance optimization of thin-body SOI and GOI MOSFETs," in *IEDM Tech. Dig.*, Dec. 2004.

[101] S. Polonsky and K. A. Jenkins, "Time-resolved measurements of self-heating in soi and strained-silicon mosfets using photon emission microscopy," *IEEE Electron Device Lett.*, vol. 25, no. 4, pp. 208–210, 2004.

[102] K. A. Jenkins and K. Rim, "Measurement of the effect of self-heating in strained-silicon MOSFETs," *IEEE Electron Device Lett.*, vol. 23, pp. 360–362, 2002.

[103] A. Khakifirooz and D. A. Antoniadis, "On the electron mobility in ultrathin SOI and GOI," *IEEE Electron Device Lett.*, vol. 25, no. 2, pp. 80–82, 2004.

[104] S. Takagi, "Re-examination of subband structure engineering in ultra-short channel MOSFETs under ballistic carrier transport," in *IEEE Symp. VLSI Tech.*, June 2003, pp. 115–116.

[105] M.-J. Chen, H.-T. Huang, K.-C. Huang, P.-N. Chen, C.-S. Chang, and C. H. Diaz, "Temperature dependent channel backscattering coefficients in nanoscale MOSFETs," in *IEDM Tech. Dig.*, Dec. 2002, pp. 39–42.

[106] L. Vancaillie, V. Kilchytska, P. Delatte, L. Demeus, H. Matsuhashi, F. Ichikawa, and D. Flandre, "Peculiarities of the temperature behavior of SOI MOSFETs in the deep submicrom area," in *IEEE Int. SOI Conf.*, 2003, pp. 78–79.

[107] Y. Taur and T. Ning, *Fundamentals of Modern VLSI Devices.* Cambridge University Press, 1998.

[108] K. E. Goodson and M. I. Flik, "Effect of microscale thermal conduction on the packing limit of silicon-on-insulator electronic devices," *IEEE Trans. Comp., Hybrids, Manufact. Technol.*, vol. 15, no. 5, pp. 715–722, 1992.

[109] A. Akturk, N. Goldsman, and G. Metze, "Coupled modeling of time-dependent full-chip heating and quantum non-isothermal device operation," in *SISPAD Tech. Dig.*, Sept. 2003, pp. 311–314.

[110] N. R. Mohapatra, M. P. Desai, S. G. Narendra, and V. R. Rao, "Modeling of parasitic capacitances in deep submicrometer conventional and high-k dielectric MOS transistors," *IEEE Trans. Electron Devices*, vol. 50, pp. 959–966, 2003.

[111] M. C. Öztürk, J. Liu, H. Mo, and N. Pesovic, "Advanced $Si_{1-x}Ge_x$ source/drain and contact technologies for sub-70 nm CMOS," in *IEDM Tech. Dig.*, Dec. 2002, pp. 375–378.

[112] D. Wang and H. Dai, "Germanium nanowire by chemical vapor deposition," *Angew. Chem. Int. Ed.*, vol. 41, no. 24, pp. 4783–4786, 2002.

[113] D. Wang, Q. Wang, A. Javey, R. Tu, H. Dai, H. Kim, P. C. McIntyre, T. Krishnamohan, and K. C. Saraswat, "Germanium nanowire field-effect transistors with SiO_2 and high-k HfO_2 gate dielectrics," *Appl. Phys. Lett.*, vol. 83, no. 12, pp. 2432–2434, 2003.

[114] Y. Cui, Z. Zhong, D. Wang, W. U. Wang, and C. M. Lieber, "High performance silicon nanowire field effect transistors," *Nano Lett.*, vol. 3, no. 2, pp. 149–152, 2003.

[115] L. Pescini, A. Tilke, R. H. Blick, H. Lorenz, J. P. Kotthaus, W. Eberhardt, and D. Kern, "Suspending highly doped silicon-on-insulator wires for applications in nanomechanics," *Nanotechnology*, vol. 10, pp. 418–420, 1999.

[116] A. Javey, J. Guo, M. Paulsson, Q. Wang, D. Mann, M. Lundstrom, and H. Dai, "High-field quasiballistic transport in short carbon nanotubes," *Phys. Rev. Lett.*, vol. 92, no. 10, pp. 106 804–106 807, 2004.

[117] Z. Yao, C. L. Kane, and C. Dekker, "High-field electrical transport in single-wall carbon nanotubes," *Phys. Rev. Lett.*, vol. 84, no. 13, pp. 2941–2944, 2000.

[118] J.-Y. Park, S. Rosenblatt, Y. Yaish, V. Sazonova, H. Ustinel, S. Braig, T. A. Arias, P. W. Brouwer, and P. L. McEuen, "Electron-phonon scattering in metallic single-walled carbon nanotubes," *Nano Lett.*, vol. 4, no. 3, pp. 517–520, 2004.

[119] J. Hone, M. C. Llaguno, M. J. Biercuk, A. T. Johnson, B. Batlogg, Z. Benes, and J. E. Fischer, "Thermal properties of carbon nanotubes and nanotube-based materials," *Appl. Phys. A*, vol. 74, pp. 339–343, 2002.