3D ELECTRO-THERMAL MONTE CARLO STUDY OF TRANSPORT IN CONFINED SILICON DEVICES

BY

MOHAMED Y. MOHAMED

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Professor Umberto Ravaioli, Chair
Professor John Tucker
Professor Kanti Jain
Professor Kent Choquette

ABSTRACT

The simultaneous explosion of portable microelectronics devices and the rapid shrinking of microprocessor size have provided a tremendous motivation to scientists and engineers to continue the down-scaling of these devices. For several decades, innovations have allowed components such as transistors to be physically reduced in size, allowing the famous Moore's law to hold true. As these transistors approach the atomic scale, however, further reduction becomes less probable and practical. As new technologies overcome these limitations, they face new, unexpected problems, including the ability to accurately simulate and predict the behavior of these devices, and to manage the heat they generate.

This work uses a 3D Monte Carlo (MC) simulator to investigate the electro-thermal behavior of quasi-one-dimensional electron gas (1DEG) multigate MOSFETs. In order to study these highly confined architectures, the inclusion of quantum correction becomes essential. To better capture the influence of carrier confinement, the electrostatically quantum-corrected full-band MC model has the added feature of being able to incorporate subband scattering. The scattering rate selection introduces quantum correction into carrier movement.

In addition to the quantum effects, scaling introduces thermal management issues due to the surge in power dissipation. Solving these problems will continue to bring improvements in battery life, performance, and size constraints of future devices. We have coupled our electron transport Monte Carlo simulation to Aksamija's phonon transport so that we may accurately and efficiently study carrier transport, heat generation, and other effects at the transistor level. This coupling utilizes anharmonic phonon decay and

temperature dependent scattering rates. One immediate advantage of our coupled electro-thermal Monte Carlo simulator is its ability to provide an accurate description of the spatial variation of self-heating and its effect on non-equilibrium carrier dynamics, a key determinant in device performance. The dependence of short-channel effects and Joule heating on the lateral scaling of the cross-section is specifically explored in this work. Finally, this dissertation studies the basic tradeoff between various n-channel multigate architectures with square cross-sectional lengths ranging from 30 nm to 5 nm are presented.

ACKNOWLEDGMENTS

I take this opportunity to acknowledge my academic advisor and director of dissertation research, Prof. Umberto Ravaioli, for his invaluable mentoring, unwavering care, and support. Throughout my studies, he has challenged me as a student, motivated me in my research, but most importantly inspired me as a person. Individuals like Prof. Ravaioli are extremely rare, and I am blessed to have known him over the past many years. I also acknowledge Prof. John Tucker, Prof. Kanti Jain, and Prof. Kent Choquette, for providing their diverse expertise, sharing their wisdom, and for taking the time to be members of my dissertation committee. Substantial contributions were also made by my collaborators, of whom I especially acknowledge Dr. Zlatan Aksamija and Dr. Andres Godoy, whose extensive experience played a key role in the completion of this thesis. I would also like to acknowledge past and present colleagues in our lab. It is on their shoulders that I stood during my research. I am especially thankful for their friendliness and support and have truly enjoyed interacting and working with each one of them. I am very thankful to Dr. Nahil Sobh and the Network of Computational Nanotechnology (NCN) for partially supporting this work. I also acknowledge my friends who helped me endure this journey, especially Azeem Sarwar, Ray Liu, Naeem Sheikh, Fawad Ismail, and Kyu-Il Lee. Finally, I acknowledge my loving parents, my wife, and my siblings who commendably endured this long process by my side, and who gave me enormous mental, spiritual, and emotional support, allowing me to move forward. Above all, I thank and praise the Lord for His countless blessings and for putting me in the right place with the right people.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Scientific progress can be mapped by the watershed discoveries that open up entirely new frontiers of possibility. In biology, the structure of DNA unlocked many secrets of life; in physics, relativity and quantum mechanics transformed the contemporary understanding of the Universe; and in electronics, the single most important invention of the last century may have been the transistor. Fifty years later, John Bardeen's invention continues to transform the human experience at an accelerating rate. Intel founder Gordon Moore's 1965 prediction that the number of transistors in a given area of integrated circuit chips would double every two years has yet to be disproven. The continuation of this process however is facing physical barriers such as atomic size limitations, limits on the ability to manipulate materials, and thermodynamic effects. This dissertation proposes a simulation model of some of these electro-thermal effects relating to electron transport in silicon devices.

Though extensive research is being done to transcend the many challenges to scaling chip design, viable solutions are few. In MOSFET design, aggressive scaling causes increased power dissipation, thermal failures, and performance limiting short-channel effects (SCEs). With regards to interconnects, scaling down increases the time required for signal propagation; therefore, chip performance may actually decrease. Innovations in lithography have not delivered either. Among the four major innovations promising to provide next-generation lithography (NGL), three are behind schedule, and the capabilities for the fourth do not currently exist.

These challenges notwithstanding, however, advances are being made. Goals

set by the International Technology Roadmap for Semiconductors (ITRS), which works toward advancing high-performance, cost-effective semiconductor technology, have spurred research groups and initiatives to look beyond conventional methods for advanced semiconductor based device design. Within transistor design, scientists have begun to explore novel materials, quasi-one-dimensional (1D) nanowire transistor architectures, and innovative methods where strain is used to enhance current in bulk MOSFETs [1, 2, 3, 4]. Silicon-on-insulator (SOI) multigate nanowires have also been strongly pursued. SOI technology allows for suppression of SCEs inherent in bulk MOSFETs while introducing the least amount of change to the planar bulk MOSFET fabrication process.

Even these encouraging technologies, however, are limited by practical factors such as cost and manufacturing capabilities. When a silicon wafer can cost upwards of several thousands of dollars, experiments become expensive and cost-prohibitive very quickly. Manufacturers have often solved such problems using simulation methods. The fundamental challenge for improved simulation models, however, is an accurate representation illustrating the properties of carrier transport in an acceptable time frame. This dissertation will focus primarily on the electrical and thermal properties of one of these technologies: highly confined quasi-1D multigate MOSFETs, with an aim to provide a full-band Monte Carlo (MC) simulation model to study these properties.

Limiting the carrier movement in a quasi-single dimension causes quantum mechanical effects to arise in the 2D plane perpendicular to the carrier transport direction. The arising 2D properties of the carrier confinement have revealed unexpected physical properties. Attempts to accurately capture and understand these systems using classical or quantum mechanical simulation methods have been limited by the level of physical detail they are able to capture. Classical methods are not able to capture quantum mechanical effects of systems and thus require methods to include these quantum corrections in simulations. Conversely, a commonly used quantum mechanical approach, the

non-equilibrium Green's function (NEGF), only superficially treats or neglects altogether the scattering effects because of the high computational cost of including them. Using a classical method, then, it is necessary to carefully include the quantum mechanical effects along the 2D plane perpendicular to the direction of charge transport in the MC kernel. Prior efforts using MC models have not been able to provide 3D simulations with the appropriate quantum mechanical and electron transport effects. Along with [5, 6] this dissertation represents one of the most comprehensive models able to capture the effect of carrier confinement on transport. This work is the first to do this with 3D multi-subband, full-band ensemble MC with 2D quantum correction.

Aside from the challenge of capturing both classical and quantum mechanical effects in a single model, other challenges may be considered. Increased temperatures caused by high transistor packing densities, which contribute to chip-level temperatures as well, are a significant challenge in scaled device design. Heating, in fact, is often seen as the ultimate challenge to scaling. With regards to transistors, heat is manifested through electron phonon interaction: as phonons travel through the lattice, they increase its temperature and couple back to electrons through electron phonon absorption. Accurate simulation tools which thoroughly couple robust electrical and thermal transport models are consequently instrumental to the development of improved transistors.

Electron and thermal transport models have been extensively detailed and studied individually, but only a few studies have coupled both, especially within the device design community. Lai and Majumdar [7] , for example, simultaneously solved the electron and phonon transport problems with a hydrodynamic model for electron transport and energy conservation model for phonon transport. A solution by Sadi and coworkers [8] involves using a 2D MC with parabolic bandstructure coupled to a simple diffusion equation. Rowlette and his collaborators [9] solved both problems by pairing the frozen field 2D particle MC developed by Pop *et al.* [10] and the split flux solver developed by Sinha *et al.* [11]. Raleva and co-workers have also recently published several

articles in this area [12]. Their work is based on the use of a 2D electro-thermal simulator, coupling a more sophisticated electron MC with the hydrodynamic phonon model proposed by Lai and Majumdar [7] to solve the moment equation for acoustic and optical phonons.

Though these models have come closer to describing the interactions between the phonons and electrons, they leave details to be explored. The model presented here is the first coupled, quantum-corrected, 3D full-band particle MC with a phonon MC. The code for this model takes into account both the influence of anharmonic phonon decay on carrier transport and the temperature dependent scattering rate. The inclusion of anharmonic phonon decay, temperature dependent scattering rate, and full electron and phonon dispersion helps create a clear picture of heat transfer and provides the ability to generate a temperature map in nanoscale devices. The temperature map obtained from the phonon transport model is fed back to the electron transport MC with a temperature dependent scattering table in a self-consistent manner. This process continues until the convergence criterion is met.

This dissertation explores the investigation of electro-thermal behavior of SOI multigate devices at the limit of cross sectional scalability using the leading-edge 3D MC presented here. Specifically, the dependence of SCEs and Joule heating is explored in narrowly scaled cross sections. Additionally, the basic tradeoff between various n-channel multigate architectures with varying cross sections is presented.

CHAPTER 2

A PRELUDE TO DEVICE MODELING

In this chapter we briefly survey simulation approaches used to model the transport of semiconductor devices.

2.1 Introduction

As device dimension approaches the size of a molecule, the need to model and examine nanoscale structures is very compelling in order to understand the details of their fascinating physical properties. The modeling approach to a scientific problem since the advent of high-speed computers has paved the way for the ever-growing field of computer experimentation. Incidentally, the role of computer-aided design (CAD) simulation tools is ever increasing, as they provide physical predicative insight into the device physics and help in the design and production of new generations of device technologies. There are many levels of device modeling depending on the extent to which physics is used in the simulation. At the same time, it is important that these tools not require immense computation time. The hierarchy of device simulation is shown in Figure 2.1. It is arranged from physically more accurate (and computationally more intensive) at the top to physically less accurate (and computationally less intensive) at the bottom. At the top of the hierarchy are quantum models, which take into account quantum effects such as size quantization, quantum interference, and tunneling.

Quantum mechanics (QM) is fundamental to nanotechnology modeling, because many of the properties and materials that nanotechnology research

seeks to exploit can only be modeled using QM methods.The Schrödinger equation fully describes the electronic structure of atoms, molecules, and solids. The many-body Schrödinger equation is given by

$$H\Psi(\mathbf{r},\mathbf{R}) = E\Psi(\mathbf{r},\mathbf{R}) \tag{2.1}$$

$$\Psi(\mathbf{r},\mathbf{R}) = \Psi(r_1,...,r_N;R_1,...,R_M) \tag{2.2}$$

where r is a function of the spatial position and spin for the electron, R_M is the coordinate of the M^{th} ion and r_N is the coordinate of the N^{th} electron. The kinetic energy (T) and the potential energy (V) are related to the Hamiltonian as

$$H = T_n + \mathbf{V}_{nuc-nuc}(\mathbf{R}) + \mathbf{T}_{el} + \mathbf{V}_{nu-el}(\mathbf{x},\mathbf{R}) + \mathbf{V}_{el-el}(\mathbf{r}) \tag{2.3}$$

Although the above Schrödinger equation is precise within the non-relativistic regime, exact solution is not possible. To make the equation numerically friendly, we need to invoke approximations while retaining as much of the key physics as possible. We shall describe these approximations in the subsequent sections.

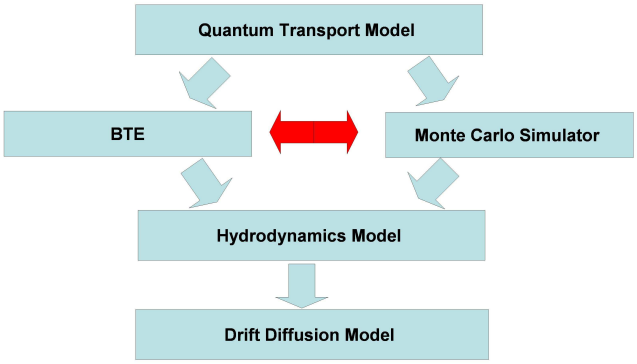


Figure 2.1: Levels of abstraction in device modeling.

2.1.1 Born-Oppenheimer adiabatic approximation: An attempt to simplify the Schrödinger equation

Dirac was the first to declare the powerful capability of the Schrödinger equation which ultimately describes completely all the interesting properties of nanostructures; however, he also attested how it is computationally complex and intractable. The Born-Oppenheimer adiabatic approximation is an important assumption in most many-body quantum problems which decouples the electronic and nuclear motion into two separate many-body problems; therefore, adding a bit of simplification to the complex many-body Schrödinger equation. The crux of this simplification is: Due to the difference in mass between the nucleus and electrons, the nucleus of an atom essentially moves much slower compared to its electrons. Therefore, due to the great disparities in the time scales of motion, the nucleus can be regarded as fixed (i.e. infinitely heavy). Hence, the electrons are assumed to follow the nuclear motion adiabatically (i.e. "dragged" with the nucleus instantaneously). The assumption that the electronic wavefunction adjusts virtually instantaneously to any changes in the nucleus allows us to simplify the Hamiltonian and cancel out the non-adiabatic (cross-coupling) electron-nuclear potential, which is expected to be quite small. Mathematically, this allows us to invoke the separation of variables on the many-body Schrödinger equation.

$$\Psi(\mathbf{r}, \mathbf{R}) = \Psi^{el}(r, R)\Psi^{nuc}(r, R) \tag{2.4}$$

$$H_{el}\Psi^{el}(r_1...r_n; R_1..., r_N) = E\Psi^{el}(r_1...r_n; R_1..., r_N) \tag{2.5}$$

$$H_{el} = \mathbf{T}_{el} + \mathbf{V}_{nu-el} + \mathbf{V}_{el-el} \tag{2.6}$$

However, although the electronic wavefunction is decoupled from its nuclear counterpart, it parametrically depends on the nucleus position R.

2.2 Ab initio Quantum Transport Approximation of the Schrödinger Equation

In this section we describe ab initio methods used to solve the many-body Schrödinger equation. The term ab initio refers to a direct solution using first principles with no empirical parameters.

2.2.1 Hartree-Fock method

Hartree-Fock theory was developed to solve the time-independent Schrödinger equation after invoking the Born-Oppenheimer approximation. The main assumption in the Hartree-Fock method is the independent-electron approximation which, in essence, removes the potential due to the electron-electron interaction and substitutes it with a self-consistent Hartree potential that no longer couples the individual motions with all the other electrons but instead depends simply upon the time-averaged electron distribution of the system. This would again allow us to separate the variables to further simplify the many-body Schrödinger equation. The electronic wavefunction is dependent on four N variables, 3 spatial and 1 spin, for each of the N particles. Hartree used the independent electron approximation to decouple the many-body wavefunction and expressed it as a Hartree product one-electron wavefunction.

$$\Psi(r_1, r_2, r_3, ...r_N) = \psi_1(r_1)\psi_2(r_2)...\psi_N(r_N) \tag{2.7}$$

Clearly this is a substantial approximation, which reduces the physics and accuracy of the problem. A shortcoming of the above Hartree product wavefunction is that it fails to take into account the Pauli exclusion principle which requires that two fermions do not exist at the same point in space with the same set of quantum numbers. Essentially, the inclusion of the Pauli exclusion principle would result in repulsion between any pair of identical

fermions possessing the same set of quantum numbers. Mathematically, the Pauli exclusion principle can be accounted for by ensuring that the wavefunction of a set of identical fermions is antisymmetric. The antisymmetry requirement is elegantly taken into account by writing the wavefunction as a single Slater determinant (Hartree-Fock method) [13], [14]:

$$\Psi(r_1, ..., r_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(r_1) & \psi_1(r_2) & \cdots & \psi_1(r_N) \\ \psi_2(r_1) & \cdots & \cdots & \psi_2(r_N) \\ \vdots & \ddots & \ddots & \vdots \\ \psi_N(r_1) & \psi_N(r_2) & \cdots & \psi_N(r_N) \end{vmatrix} \tag{2.8}$$

$$H_{HF}\psi_i = \xi_i\psi_i \tag{2.9}$$

$$H_{HF} = -\frac{\hbar}{2m}\nabla + c_1 \int \frac{n(r)}{|r - r'|}dr + E_x[n] + V_{ext} \tag{2.10}$$

where $E_x[n]$ is the exchange potential due to including the spin effect.

In order to solve the integro-differential Hartree-Fock equation numerically, one needs to express each individual electron wavefunction as a set of primitive functions called a basis set. This approach helps reduce the Hartree-Fock equation to a linear algebra problem.

$$\psi_j = \sum_{i=1}^{M} c_{ij}\varphi_i \tag{2.11}$$

In deciding upon a particular set of basis functions, one has to compromise between computational cost and accuracy. A large basis would provide a high-accuracy result and be computationally expensive. If the basis set ϕ_i is complete, the results would be the same as a direct numerical solution to the Hartree-Fock equation. But, for practical reasons, the set is always finite and far from the Hartree-Fock limit. The simplest form of basis set available is the minimal basis set, which includes as many basis sets as there are electrons for the individual ψ_j.

2.2.1.1 Limitation and computation cost of the Hartree-Fock method

There are two factors that limit the exactness of the Hartree-Fock method—the accuracy with which one actually solves the Hartree-Fock equations and the intrinsic limitation due to the simplification made in deriving the model. The accuracy with which one solves the Hartree-Fock equations is determined by the completeness of the expansion basis of the wavefunction. As we have seen, a complete Hartree-Fock calculation would require that the many-electron wavefunction should be represented by a linear combination of Slater determinant wavefunctions. For high accuracy the number of Slater determinant terms solved can go up to, for some calculation, 10^9, which scales horribly as we increase the number of electrons in the system [15]. Thus having the wavefunction as a key variable makes a complete Hartree-Fock model computationally expensive, if not intractable, especially if one tries to model real materials.

An intrinsic limitation of the Hartree-Fock model hinges upon the independent electron approximation, which allowed us to approximate the electron-electron potential by with an average field potential (self-consistent field) that essentially neglects the true effects of the rather important electron-electron repulsion in the system, which is called the electron correlation. The error in energy is due to not including the correlation effect is given by:

$$E_{corr}^{HF} = E - E_{HF} \tag{2.12}$$

Much work has been done to compensate this inherent shortcoming using post-Hartree-Fock models which include an approximation of the electron correlation effect in the Hartree-Fock model such as the many-body perturbation techniques and configuration interaction (CI). However, these approaches add much more computational cost to the already computationally demanding model.

2.2.2 Density functional theory (DFT)

In [15], Walter Kohn comments that many-body wavefunction method is not a legitimate scientific concept for more than 1000 particles. The density functional theory is a very elegant tool which help us overcome the complication of using the wavefunction as the primary variable in solving the Schrödinger equation. DFT utilizes the theorem by Hohenberg and Kohn which essentially legitimized the use of the electron density as the primary variable in describing the electronic system instead of the wavefunction itself [16]. The Hohenberg-Kohn theorem [16] states that the electron density minimizes the electronic energy and that there is a one-to-one correspondence between the ground-state electron-density and the ground-state wavefunction. Consequently, all other ground-state properties of the system are determined by the ground-state electron density. The Hohenberg-Kohn theorem however does not provide the details of such mapping and is only an existence theorem [13].

Kohn and Sham developed an approach which applied DFT to the Hartree model [17]. The resulting Kohn-Sham equation takes on the below form .

$$H_{KS}\psi_i = \xi\psi_i \tag{2.13}$$

$$H_{KS} = -\frac{\hbar}{2m}\nabla + c_1 \int \frac{n(r)}{|r - r'|} dr + E_{xc}[n] + V_{ext} \tag{2.14}$$

2.2.2.1 $E_{xc}[n]$: Success and limitations

Unlike the Hartree model, Kohn and Sham included both exchange and correlation effects in their model for all the many-particle interactions. In essence, the key challenge in DFT is finding a good approximation to $E_{xc}[n]$, which improves the accuracy of the model and constitutes a major area of development in modern DFT particularly since there is no approximation of $E_{xc}[n]$ that is valid for all cases, especially for problems with strong

Table 2.1: Summary of some of the advantages and disadvantages of DFT Method

Advantage	Disadvantage
Fast	Eigenvalues have no apriori meaning
Exact for total density and energy	Restricted for ground-states
Softwares freely available	Large deviation for bandgaps

electron-electron correlation.

The simplest approximation of $E_{xc}[n]$ is the local density approximation (LDA) which gives the functional dependence of the $E_{xc}[n]$ to the electronic density n [15]:

$$E_{xc}[n] \equiv \int e_{xc}(n(r))n(r)dr \tag{2.15}$$

where $e_{xc}(n(r))$ is the exchange–correlation energy per electron in a uniform electron gas of density n and can be computed exactly or through the Monte Carlo method depending on the problem.

Table 2.1 summarizes the advantages and disadvantages of the DFT method. As the table shows, an inherent problem with the DFT method is that it consistently underpredicts the electronic bandgaps due to it being restricted to only ground-state description. Figure 2.2 shows the bandgap computed using DFT for various semiconductor materials and compares it with the actual bandgap through the use of exact correlation.

Various methods have been used to overcome this hurdle such as the GWA, a hybrid method which incorporate DFT with Green's function.

2.3 Overview of Non-ab initio Quantum Transport Models

2.3.1 Non-equilibrium Green's function (NEGF)

A first-principles description of non-equilibrium systems is extremely complicated. Solving the time-dependent Schrödinger equation for the full many-particle system is not an option due to the large computational effort [14], [16]-[18]. Instead, Green's function techniques offer a natural and relatively simple method for describing non-equilibrium correlated many-particle systems. The non-equilibrium Green's function (NEGF) formalism provides a powerful means to handle open quantum systems which have non-vanishing boundary conditions for the Schrödinger wavefunctions. Given a partial differential equation of the form [19] , [20] :

$$(i\hbar\frac{\partial}{\partial t} - H(\mathbf{r}))\psi(r,t) = 0 \tag{2.16}$$

The time-independent Green's function is defined as the solution to

$$(i\hbar\frac{\partial}{\partial t} - H(\mathbf{r}))G(\mathbf{r},\mathbf{r}',t-t') = \hbar\delta(\mathbf{r}-\mathbf{r}')\delta(t-t') \tag{2.17}$$

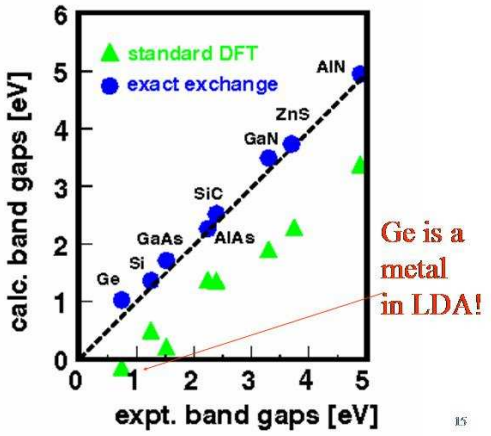


Figure 2.2: DFT bandgap calculation vs. calculation with exact exchange effect.

which satisfies the same solution and boundary condition of the original problem but with a source term added. Note that H is not a function of time. The Fourier transform of the energy dependent Green's function is give by:

$$G(\mathbf{r}, \mathbf{r}', t - t') = \frac{1}{2\pi} \int G(\mathbf{r}, \mathbf{r}', E) \exp(\frac{-iE(t - t')}{\hbar}) dE \tag{2.18}$$

Substituting Equation (2.13) back to (2.11) will give us the time-independent Green's function, which is practical for steady-state solution to the transport problem.

$$(E - H)G(\mathbf{r}, \mathbf{r}', E) = \delta(\mathbf{r} - \mathbf{r}') \tag{2.19}$$

In operator notation it can be written and expanded as

$$(H(\mathbf{r}) - E(\mathbf{r}))G(\mathbf{r}, \mathbf{r}', t - t') = -1 \tag{2.20}$$

$$G = \frac{1}{E - H} = \sum_n \frac{|n><n|}{E - E_n} + \int dn \frac{|n><n|}{E - E_n}$$

Note that the Green's function has a singularity at $E = E_n$ which thus prevents us from evaluating the integral given by Equation (2.18). However, we can use the path integral method on the complex z-place to solve this problem. Depending on the path we choose to evaluate our integral, we obtain either the retarded or advanced Green's function (see Figure 2.3) [19]:

$$G^R = \sum_n \frac{|n><n|}{E - E_n + i0^+} \tag{2.21}$$

$$G^A = \sum_n \frac{|n><n|}{E - E_n - i0^+} \tag{2.22}$$

where GA is the retarded Green's function and GR is the advanced Green's function. Since $G^A = G^R$, we only need to solve one equation. Further, when the

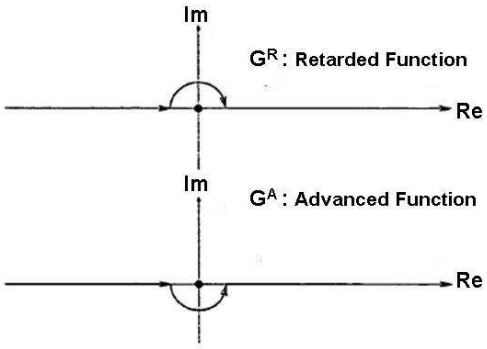


Figure 2.3: Calculating the Green's function using path integral with the assumption that only one singular point exists at the origin [19].

system involves some sort of interaction Σ, we can use Dayson's equation to relate the total Green's function of the full system (G) with the non-interacting original Green's function (G_0):

$$G = G_0 + G_0 \Sigma G \tag{2.23}$$

Self-energy can also be added to account for various interactions (i.e. scattering, open terminals, etc.). Using Σ to denote the various self-energy terms such as the coupling to the outside world through open boundary (Σ^B), the coupling due to phonon scattering (Σ^P) for phonon scattering, etc.

$$\Sigma = \Sigma^B + \Sigma^P + \Sigma^{e-e} + ... \tag{2.24}$$

2.3.1.1 Device modeling using NEGF

The main question when it comes to using NEGF to model nanostructures is whether it is plausible to incorporate all the correct interactions properly into NEGF. This indeed is the main hurdle that limits its potential to fully model the quantum mechanical effects of a nanostructure. As we have shown in the previous section, self-energy is added to NEGF to capture important interactions such as scattering. However, the main challenge to this approach is that in defining the interaction, a general self-energy is usually assumed for the scattering. This self-energy is developed using the Langreth theorem from the full two-particle interaction Green's function, but this theorem is only proven for near-equilibrium systems [21]. Incidentally, numerical tools which use NEGF do an excellent job describing quantum mechanical effects; however, they include only limited scattering processes in their model.

2.3.2 Transport based on the Wigner distribution function

Classical and quantum mechanical formulation of statistical mechanics differ in structure. The relationship between a quantum mechanical description of particles and its classical counterpart has been the object of much discussion. Wigner presented a phase space quantum mechanical distribution function which in the proper limit, coincides with the classical phase space distribution function. The Wigner function is defined as the Fourier transform product of the overlap of the wavefunction with its mirror image [22].

$$f(r,k,t) = \frac{1}{\pi^3}\int \Psi(r-r',t)\Psi^*(r-r',t)\exp(i2r'k)dr' \tag{2.25}$$

From the Wigner function we can derive the quantum mechanical Wigner-Boltzmann transport equation as:

$$\frac{\partial f}{\partial t} + v.\nabla_r f + \frac{q}{\hbar}\sum_{\kappa=0}^{\infty}\frac{(-1)^{2\kappa}}{4^\kappa(2n+1)!}.\nabla_r^{2n+1}V(r).\nabla_k^{2n+1}f = \frac{\partial f}{\partial t}_C \tag{2.26}$$

At the limit of $\kappa = 0$ the quantum mechanical Wigner-Boltzmann transforms into the semi-classical Boltzmann transport equation which we will discuss in the next section. Further, using the moments method one can derive the quantum hydrodynamic and quantum drift diffusion model from the Boltzmann-Wigner equation, in the same way it is done for the semi-classical Boltzmann transport equation (BTE). Figure 2.4 shows prototype results obtained from Wigner method. Numerically, the Wigner transport can be implemented using the Monte Carlo method or some finite element scheme. Figure 2.5 shows a comparison between the Wigner MC approach and Green's function for a resonant tunneling diode (RTD).

A key limitation to the Wigner method is that the Wigner distribution function f is not a true distribution function, in that it can become negative. Much work on the Wigner approach has been explored. For a good review of this topic, see [23]- [24].

2.4 Overview of Semi-classical Transport Models

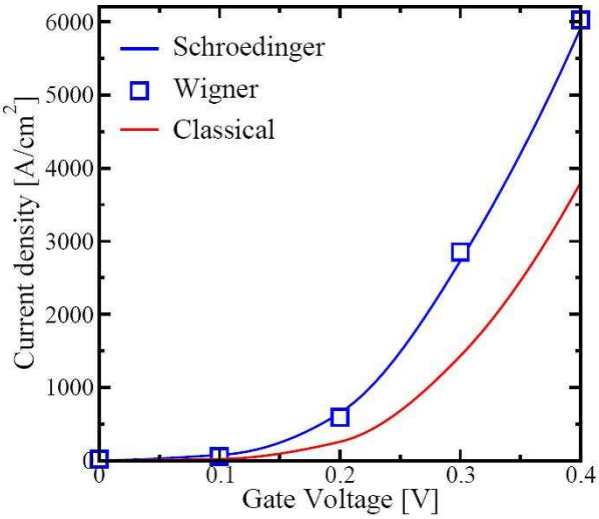
2.4.1 Boltzmann transport equation

The Boltzmann transport equation model has been the main tool used in the analysis of transport in semiconductors. The BTE equation is given by:

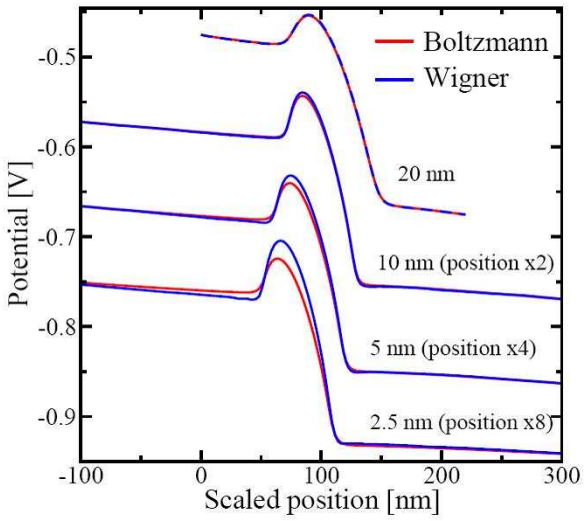
$$\frac{\partial f}{\partial t} + \frac{1}{\hbar}\nabla_k E(k).\nabla_r f + \frac{qF(r)}{\hbar}.\nabla_k f = [\frac{\partial f}{\partial t}]_{collision} \tag{2.27}$$

$$v = \frac{1}{\hbar}\nabla_k E(k) \tag{2.28}$$

The distribution function, f, is a dimensionless function which is used to extract all observables of interest and gives a full depiction of electron distribution in both real and k space. Further, it physically represents the probability of particle occupation of energy k at position r and time t. In



(a)



(b)

Figure 2.4: Comparison of the Schrödinger (a) and BTE with the Wigner model (b) [25].

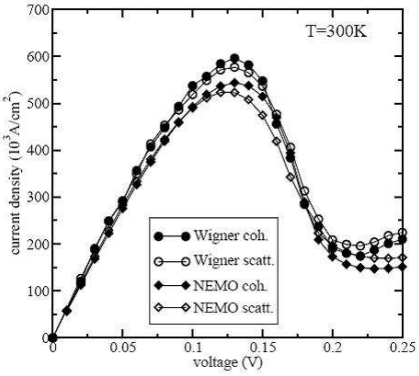


Figure 2.5: IV characteristics of an RTD device using Wigner MC and NEMO-1D (Green's function). Transport is coherent or dissipative [26].

addition, due to being a seven-dimensional integro-differential equation (six dimensions in the phase space and one in time) the solution to the BTE is cumbersome and can be solved in closed analytical form under very special restrictions. Numerically, the solution to the BTE is employed using either a deterministic or stochastic method. The deterministic method solution is based on a grid-based numerical method such as the spherical harmonics approach, whereas the Monte Carlo is the stochastic approach used to solve the BTE.

2.4.2 Monte Carlo solution of the BTE

The semi-classical Monte Carlo method is a statistical method used to yield exact solutions to the Boltzmann transport equation which includes complex bandstructure and scattering processes. This approach is semi-classical for the reason that scattering mechanisms are treated quantum mechanically using the Fermi golden rule, whereas the transport between scattering events is treated

using the classical particle notion. The Monte Carlo model in essence tracks the particle trajectory at each free flight and chooses a corresponding scattering mechanism stochastically. Two of the great advantages of semi-classical Monte Carlo are its capability to provide accurate quantum mechanical treatment of various distinct scattering mechanisms within the scattering terms, and the absence of assumptions about the form of carrier distribution in energy or k space. The semi-classical equation describing the motion of an electron is

$$\frac{dr}{dt} = \frac{1}{\hbar}\nabla_k E(k) \tag{2.29}$$

$$\frac{dk}{dt} = \frac{qF(r)}{\hbar} \tag{2.30}$$

where F is the electric field, $E(k)$ is the energy dispersion relation, and k is the momentum wave vector. To solve the above equations of motion, one needs strong knowledge of the bandstructure ($E(k)$). The $E(k)$ relation describes how the particle moves inside the device, in addition to depicting useful information necessary for transport such as the density of states (DOS) and the particle velocity. A full-band $E(k)$ relation can be obtained using the semi-empirical pseudopotential method.

2.4.3 Drift diffusion method

Both drift diffusion (DD) and the hydrodynamic models can be derived from the moments of the Boltzmann transport equation (BTE) using simplified approximation valid for long channel devices. The moment of BTE is defined as the integral over k space of the of the BTE multiplied by the weight function Υ. In general the moment of function (f) is:

$$\frac{1}{4\pi^3}\int \Upsilon f(k,r)dk \tag{2.31}$$

The drift diffusion model can be derived using the zeroth moment (continuity equation) and the first-order moment (current equation) of the BTE. The zeroth order moment is given by using $\Upsilon = 1$ and integrating the BTE over dk:

$$\int dk \frac{\partial f}{\partial t} + \int dk v.\nabla_r f + \int dk \frac{qF(r)}{\hbar}.\nabla_k f = \int dk [\frac{\partial f}{\partial t}]_{collision} + \int dk [\frac{\partial f}{\partial t}]_{G-R} \tag{2.32}$$

Where:

$$\int dk [\frac{\partial f}{\partial t}] = \frac{\partial n}{\partial t} \tag{2.33}$$

$$\int dk [v.\nabla_r f] = \nabla_r v.n = -\frac{1}{q}\nabla_r J \tag{2.34}$$

$$\int dk [\frac{qF(r)}{\hbar}.\nabla_k f] = 0 = \int dk [\frac{\partial f}{\partial t}]_{collision} \tag{2.35}$$

Putting all the above terms of the zeroth order moment, we finally get the continuity equation:

$$\frac{\partial n}{\partial t} - \frac{1}{q}\nabla_r J = [\frac{\partial f}{\partial t}]_{G-R} = G - R \tag{2.36}$$

Using the same process as above, the first-order moment is obtained by letting $\Upsilon = k$ and integrating over all possible wave-vectors. Doing this we get the current equation. Finally, the governing equations for the drift-diffusion model are obtained by writing the Poisson equation along with the current and continuity equation we obtained from the first two moments of the BTE for both electrons and holes:

$$\nabla_r(\varepsilon E) = \rho \tag{2.37}$$

$$J_n - \frac{\partial[\langle\tau\rangle J_n]}{\partial t} = qn\mu_n E + q\nabla_r(D_n n) \tag{2.38}$$

$$J_p - \frac{\partial[\langle\tau\rangle J_p]}{\partial t} = qn\mu_p E + q\nabla_r(D_p p) \tag{2.39}$$

$$\frac{\partial n}{\partial t} - \frac{1}{q}\nabla_r J_n = G(n) - R(n) \tag{2.40}$$

$$\frac{\partial p}{\partial t} + \frac{1}{q}\nabla_r J_p = G(p) - R(p) \tag{2.41}$$

The second term of Equations (2.38) and (2.39) is negligible unless the current density varies rapidly. The above equations are solved simultaneously to model carrier transport in the drift-diffusion limit. We also note that in the DD approach, the charge transit time is assumed to be very large in comparison to the energy relaxation time [27].

2.4.4 Energy balance/hydrodynamic model

The energy balance (hydrodynamic) equations consist of carrier, momentum, and energy balance equations and are derived using higher moments of the Boltzmann Transport Equation (BTE) [28],[29] . These higher moments add more equations to the set of equations used in DD. The second moment of the BTE, for example, can be derived by repeating the same process we used to get the continuity and the current equation. This time, we weight the BTE by Υ= Energy. The second moment will yield the energy flow equation (e.g. for electrons):

$$\frac{\partial \varepsilon_n}{\partial t} = \frac{1}{n} J_n.E - \frac{1}{n}\nabla_r.S_n \tag{2.42}$$

where the term $J_n.E$ is the well-known Joule heating, ε_n is the energy, and S_n is the entropy flux and is given by:

$$S_n = \frac{\hbar}{4\pi^3 m^*} \int dk[kE(k)f(k)] \tag{2.43}$$

The extra equations added by the second moment will allow us to capture and calculate physical details such as carrier heating and the velocity overshoot effect, which is not attainable using the DD model. Methods based on the first three to four moments are generally referred to as the energy transport (ET) model or the HD model. However, some authors employ higher-order moments to derive mathematically sophisticated hydrodynamic transport equations; nonetheless, the physical intuition attained from these moments become less clear. Needless to say, in the end, an accurate discretization method is required in HD simulation, since the governing equations are strongly coupled and one has to deal with a larger number of variables compared to the DD scheme.

2.4.5 Comparison of semi-classical models

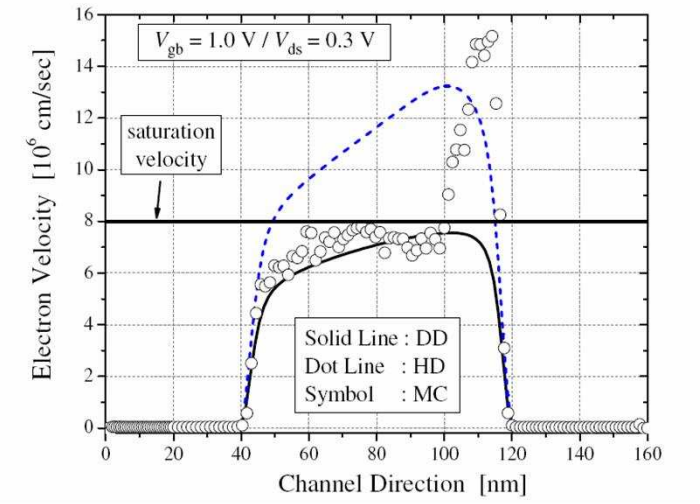
We compare the accuracy of semi-classical models based on the BTE by investigating how they treat the classical velocity overshoot problem, a key short-channel effect (SCE) in transistor structures. Essentially, velocity overshoot is a non-local effect of scaled devices, which is related to the experimentally observed increase in current drive and transconductance [30]. As the channel length becomes shorter, the velocity is no longer saturated in the high field region, but it overshoots the predicted saturation velocity. The cause of this phenomenon is that the carrier transit time becomes comparable to the energy relaxation time, and therefore the mobile carriers do not have enough time to reach equilibrium with the applied electric field by scattering in the short-channel devices [31].

The summary of simulation results in which we compare an MC model (UIUC MoCa) with DD and HD model is shown in Figure 2.6. In Figure 2.6(a) we see the case when the field is not high enough to cause the velocity overshoot effect

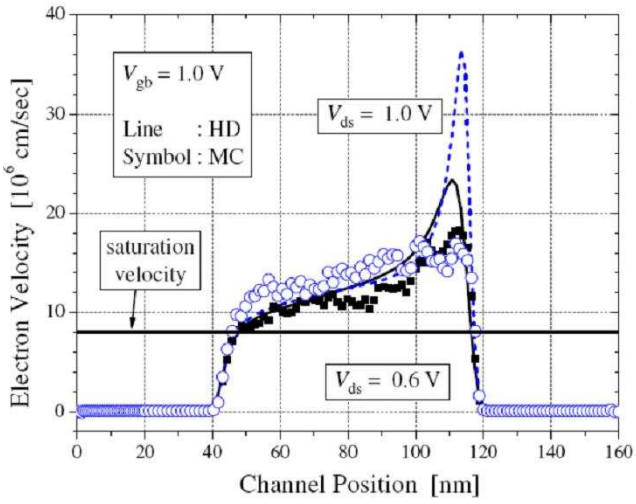
in the whole channel region. Note that at such a limit, the data from the DD model fit well to the MC model in the non-overshoot region, but the HD model overestimates the velocity in that region. The velocity overshoot is observed only near the drain junction in the MC data, and the HD model fits well in that region. From the MC data, it can be noticed that the velocity overshoot effect is abrupt in the high field region, which is not properly included in the HD model. For high field conditions as shown in Figure 2.6(b) velocity overshoot is observed almost all over the channel; in addition, the HD results and the MC results are very close in the channel region.

2.5 Summary

We have reviewed some of the basic theoretical approaches used to model the electronic properties of nanostructure materials and devices. The continuing advancement of the field of modeling and design has contributed in solving a diverse range of problems and aided experimentation as a predicative tool. As we are reaching the nanometer regime, modeling a system requires the inclusion of very detailed physical descriptions. Understanding the various theoretical methods used to approximate the many-body Schrödinger equation is only the beginning. A great deal of experimental and theoretical collaboration is necessary in order to develop a modeling tool which describes the essential physics as observed by experimental measurement.



(a)



(b)

Figure 2.6: Average carrier velocity for a 80 nm N-channel MOSFET comparing the various semi-classical simulation models (a) V_{ds}= 0.3 V (b) V_{ds}= 0.6 V.

CHAPTER 3

3D MULTI-SUBBAND MONTE CARLO WITH FULL ELECTRON AND PHONON DISPERSION

In this chapter we present the details of MoCa, the 3D particle Monte Carlo simulator developed at the University of Illinois. As was indicated in Chapter 2, the semi-classical Monte Carlo method is a statistical method used to yield exact solutions to the Boltzmann transport. A key strength of the MC method is the relative ease of including complex bandstructures and scattering processes. Furthermore, to extend the applicability of the MC model into the nanoscale regime, quantum correction becomes unavoidable. Here, we begin by reviewing details of the MC model. We also discuss the recent inclusion of subband scattering into our 3D MC kernel.

3.1 Details of the Monte Carlo Model

3.1.1 Bandstructure

Reproducing the E-k dispersion (Band structure) relation is the starting point in transport analysis. Equation (2.29) shows that carrier velocity has a strong dependence on the details of the band structure. In addition, bandstructure calculation readily renders the density of states (DOS), a key element essential in characterizing solids. The basic problem to be addressed by the determination of the electronic bandstructure is that of a many-electron system within a periodic potential. However, due to our current computational constraints, we often need to employ drastic simplifications to obtain the correct dispersion for a given material .

There are two classes of electronic bandstructure calculation: ab initio methods and empirical methods. Ab initio methods compute the material bandstructure from first principles without relying on any experimental inputs. Examples of common ab initio schemes include density functional theory (DFT) and Hartree-Fock. However, such methods are often limited to elucidating ground state properties of a given material or are used for thin (non-bulk) films. Empirical methods, on the other hand, rely heavily on the availability of experimental inputs for the material of interest. Examples of common empirical methods include the pseudopotential, k.p, tight binding, and orthogonalized plane wave (OPW). These methods are especially attractive because the many-body problem is replaced by a one-electron Schrödinger equation with empirical fitted potential. Consequently, empirical methods are computationally less expensive as compared to ab initio calculations and provide a relatively easy means of generating the electronic bandstructure.

Following the seminal paper by Cohen and Bergstresser [32] in which they computed the bandstructure for common diamond and zinc-blende materials using empirical deformation potential, the empirical pseudopotential method has gained popularity. In our 3D Monte Carlo simulator (MoCa), the numerical bandstructure of silicon is obtained from a local pseudopotential solution approach [33], and it is used to create tables for the energy dispersion, carrier group velocity, and the density of states. An empirical psuedopotential method (EPM) calculation of silicon bandstructure is provided in Figure 3.1 The choice of EPM is natural as it provides a stunning accuracy but yet is simple and easy to integrate into the Monte Carlo framework. Nonetheless, EPM is not very well suited to compute bandstructures of ultra-thin (confined) materials such as superlattices.

3.1.2 Scattering mechanisms

Electron transport in the diffusive regime is disturbed by scattering. Electrons may scatter due to interaction with the surface, other electrons, defects, and/or ionized impurities. It is important to incorporate such scattering mechanisms into the Monte Carlo simulation. In tackling the scattering problem, one often assumes that the effect of collision mechanism on carrier motion inside the crystal is weak and hence first-order perturbation theory maybe utilized. As such, the transition from one electron state to another due to scattering is treated via Fermi's golden rule:

$$S(k, k^{'}) = \frac{2\pi}{\hbar}|M_{k^{'},k}|^2\delta[E(k^{'}) - E(k) - \Delta(E)] \tag{3.1}$$

where $S(k, k^{'})$ denotes the transition rate, $M_{k^{'},k}$ is the matrix element, and $\Delta(E)$ is the change in energy caused by the scattering event. The term in the bracket ensures energy conservation. The above transition rate expresses the probability per unit time in which a carrier with an initial momentum k scatters

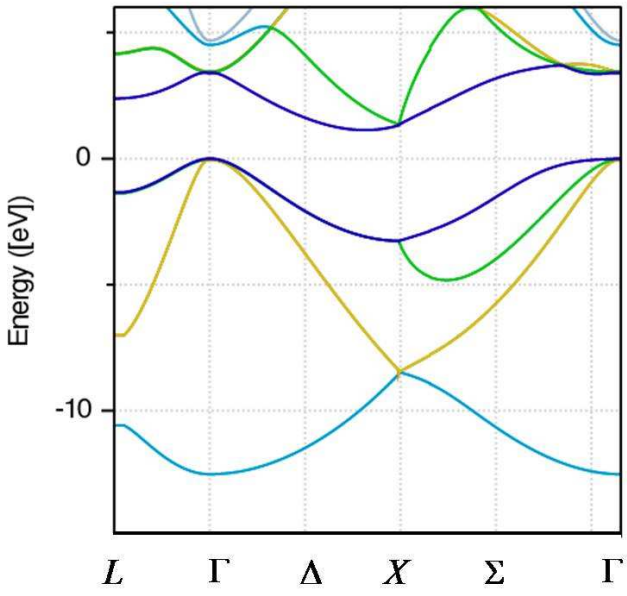


Figure 3.1: Empirical pseudopotential calculation of the electronic bandstructure for silicon.

to a state with final momentum $k^{'}$. It is widely known that scattering events in semiconductor devices is the main mechanism which tends to restore equilibrium distribution within the device, and therefore minimizes the prominence of non-local effects. However, scattering also poses a limit to device mobility and current. The mobility of device is given by:

$$\mu = \frac{\langle \mathcal{V} \rangle}{\mathbf{E}} = \frac{q\mathbf{E}}{m_e} \tag{3.2}$$

and

$$\frac{1}{\tau} = \sum_{k'} S(k, k') = \frac{\langle \mathcal{V} \rangle}{l}, \tag{3.3}$$

where l is the scattering length (mean free path), τ is the scattering time, and $\mathcal{V}$ is the drift velocity.

Within the Monte Carlo simulation, the durations of the carrier free flight and the scattering events are selected stochastically in agreement with the given distribution probabilities determined by the scattering probabilities. The scattering mechanism selects the momentum state after the scattering event for energy conservation, according to the relative probabilities of all possible scattering mechanisms. To ensure accuracy of results for phonon events, a full phonon dispersion is included in MoCa3D. The full phonon dispersion is calculated from an adiabatic bond charge model and tabulated for lookup [34]. Figure 3.2 shows the final state selection algorithm utilized for the full phonon dispersion. This algorithm is employed at each iteration to ensure energy conservation with the tabulated full phonon dispersion. MoCa3D accounts for various scattering mechanisms such as phonon, carrier-carrier, impact ionization, and surface roughness scattering. In addition, the total scattering rate is adjusted so that at high energies it follows the density of states.

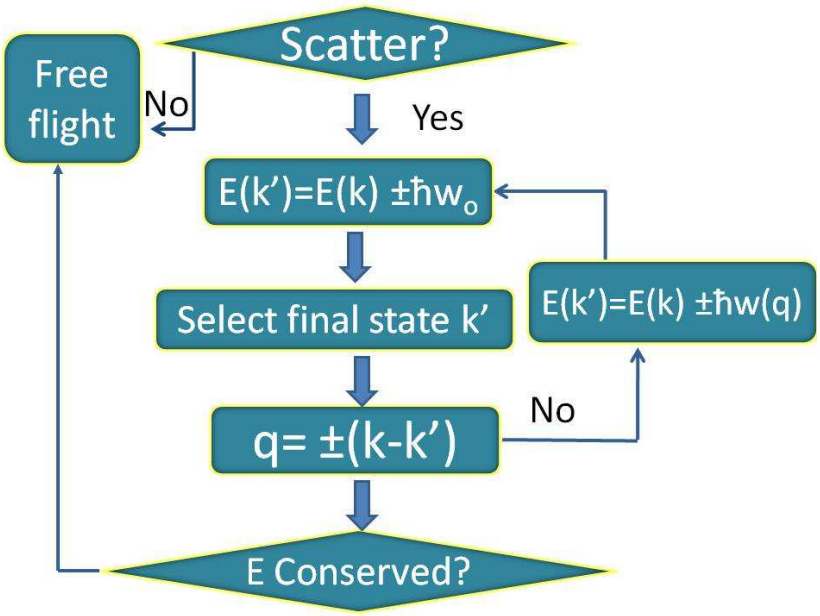


Figure 3.2: Flowchart showing the details of the final state selection with the full phonon dispersion. This algorithm is applied at each iteration in order to ensure energy conservation with the phonon dispersion which is tabulated from the adiabatic bond charge algorithm.

3.1.3 Electrostatic based quantum corrections for Monte Carlo simulation

The current trend of scaling down semiconductor devices has forced physicists to incorporate quantum mechanical issues in order to acquire a thorough understanding of device behavior. Simulating the behavior of nanoscale devices necessitates the use of a full quantum transport model especially for cases when the quantum effects cannot be ignored. For practical devices like the modern day MOSFET, the semi-classical transport model can be augmented to capture some of the relevant quantum effects. For example, electrostatic quantum corrections can be incorporated into a Monte Carlo simulator by simply introducing a quantum potential term which is superimposed onto the classical electrostatic potential seen by the simulated particles. Figure 3.3 shows the essential features of this technique. The various quantum approaches available for implementation are described next [35].

3.1.3.1 Wigner-based correction

The Wigner transport equation forms the basis for the Wigner-based quantum correction.

$$\begin{aligned} \frac{\partial f}{\partial t} &+ \mathbf{r}.\nabla_{\mathbf{r}} f - \frac{1}{\hbar}\nabla_{\mathbf{r}} V.\nabla_{\mathbf{k}} f \\ &+ \sum_{\alpha=1}^{\infty} \frac{(-1)^{\alpha+1}}{\hbar 4^{\alpha}(2\alpha+1)!} \times (\nabla_{\mathbf{r}}.\nabla_{\mathbf{k}})^{2\alpha+1} V f \\ &= \left(\frac{\partial f}{\partial t}\right)_c \end{aligned} \tag{3.4}$$

where, k is the crystal momentum, V is the classical potential, the term on the right-hand side (RHS) is the effect of collision, the fourth term on the left-hand side (LHS) represents non-local quantum mechanical effects. The standard Boltzmann Transport Equation is obtained when the non-local terms on the LHS disappear in the limit of slow spatial variations. The simplified (for $\alpha = 0$)

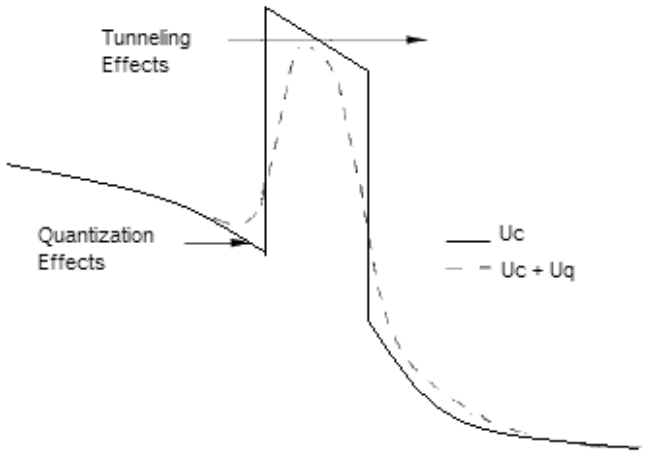


Figure 3.3: Effects of quantum correction.

quantum-corrected BTE then becomes

$$\frac{\partial f}{\partial t} + \mathbf{r}.\nabla_{\mathbf{r}} f - \frac{1}{\hbar}\nabla_{\mathbf{r}} V_w.\nabla_{\mathbf{k}} f + = \left(\frac{\partial f}{\partial t}\right)_c \tag{3.5}$$

where the quantum potential is contained in the term V_w.

3.1.3.2 Effective potential correction

This method of quantum correction was developed by Feynman and Hibbs [36]. In this method the effective potential is derived by calculating the contribution to the path integral of a particle's quantum fluctuations around its classical path. This calculation is undertaken by a variational method using a trial potential to first order. The effective classical potential on the average point on each path then becomes

$$\begin{aligned} V_{\text{eff}}(x) &= \frac{1}{\sqrt{2\pi a}} \int_{-\infty}^{\infty} V(x') e^{-\frac{(x'-x)^2}{2a^2}}\, dx' \\ a^2 &= \frac{\hbar^2}{12 m^* k_B T} \end{aligned} \tag{3.6}$$

3.1.3.3 Schrödinger-based correction

This approach involves periodical solving of a Schrödinger equation in a simulation with the input being the self-consistent electrostatic potential. The exact energy levels and wavefunctions relating to the electrostatic potential solution are employed to calculate the quantum potential. The quantum potential obtained from this approach produces a field which repels electrons from the interface to force the shape of the quantum density. The quantum correction potential $V_{qc}(y,z)$ and the quantum carrier density are related as:

$$V_{qc}(y,z) = -k_B T_t \log[n_q(y,z)] - V_p(y,z) + V_o \tag{3.7}$$

where $V_p(y,z)$ is obtained from the self-consistent solution of the Poisson equation, V_o is a reference potential point in the device where the quantum correction is fixed to zero, and $n_q(y,z)$ is the carrier density obtained from the solution of the Schrödinger equation. The total potential is related to the quantum potential as:

$$V_{total}(y,z) = V_p(y,z) + V_{qc}(y,z). \tag{3.8}$$

The corrected potential differs from the classical electrostatic potential mainly in the vicinity of the interfaces. Here, the strongly attractive nature of the electrostatic potential that would be applied in a classical model is transformed into a repulsive potential by the quantum correction. As a consequence, the charge density is shifted toward the center of the channel cross-section.

Even though the above mentioned potentials for quantum correction differ in their method of calculation and their basic assumptions, yet when it comes to their inclusion into Monte Carlo simulation, they are all incorporated the same way.

3.2 Quantum Corrected MC with Subband Phonon Scattering

Current research in silicon device scaling is directed towards one-dimensional electron gases (1DEG) structures incorporating nanowire channels, like multigate MOSFET, nanoribbons, and nanowire MOSFET [37, 38, 39, 40]. One-dimensional electron gas based devices confine carriers in the two directions perpendicular to carrier transport. Low-dimensional semiconductors offer interesting physical phenomena, such as volume inversion, which affect carrier mobility and current. There are two opposing views on the mobility of carriers in quasi-1D structures. Sakaki [41] argued that the decrease in density of states (DOS) for scattering states enhances the mobility of nanowire as compared to

bulk structure. The other view attributes mobility reduction to strong electron-phonon wavefunction overlap which substantially influences the phonon scattering rate calculation as compared to the bulk scattering rate calculation.

This section introduces the inclusion of such a mechanism into our MC kernel. Previous attempts to carry out the multi-subband approach into the MC simulation have been limited to 1D transport simulations assuming infinitely long channel length [5], 2D MC simulation [42], or 3D simulation which does not take into consideration the wavefunction penetration into the oxide [43].

Since we are dealing with a one-dimensional (1D) electron gas, the expressions employed for the calculation of the scattering rates should be modified accordingly. In this respect, acoustic and intervalley scattering rates have been considered following a previous work [5] that demonstrated that the 1D density of states and the overlap integral between initial and final states are the dominant terms in the calculation of the phonon scattering rate. To our knowledge, this is the first 3D Monte Carlo simulator with full electron and phonon dispersion.

3.2.1 Acoustic phonon scattering rate in quasi-one-dimensional electron gas

The phonon mode interacting with the electron are assumed to be 3D and bulk in nature. The Fermi golden rule is utilized to compute the scattering rate. The interaction Hamiltonian used for acoustic scattering has been derived by the deformation potential theory of Bardeen and Schockley. This theory considers scattering due to long-wavelength acoustic modes and assumes that dilation and compression due to longitudinal acoustic modes distorts the crystal by shifting the energy band. The shift in energy band due to the electron-lattice interaction is expressed by means of a deformation Hamiltonian which takes into account the electron phonon coupling:

$$H_{el-ph}^{ac} = D_{ac}\frac{\delta V}{V} = D_{ac}\nabla.\mathbf{u}(r) \tag{3.9}$$

where D_{ac} is the acoustic deformation potential. The displacement of ions due to collective vibration of atoms around their equilibrium position is often approximated by summing terms, each representing a harmonic oscillator corresponding to normal modes. The lattice displacement at point r at time t is :

$$u(r) = i\sum_{q}\sqrt{\frac{\hbar}{2\rho V\omega_q}}e_q\left(a_q\mathrm{e}^{iq.r} + a_{-q}^{\dagger}\mathrm{e}^{-iq.r}\right) \tag{3.10}$$

where q is the wave vector, ω_q is the angular frequency, V is the volume of the crystal, a_q and a_{-q} are the phonon annihilation and creation operators and e_q is the polarization vector. Substituting Equation (3.10) into (3.9), the interaction Hamiltonian due to acoustic phonons becomes:

$$H_{ac} = \mathrm{D_{ac}}\sum_{q}\sqrt{\frac{\hbar}{2MN\omega_q}}e_q.q\left(a_q\mathrm{e}^{iq.r} + a_{-q}^{\dagger}\mathrm{e}^{-iq.r}\right) \tag{3.11}$$

We define the z coordinate as the direction of current transport and the y-z as the confinement plane. The usual plane wave basis is used to compute the initial and final wavefunction:

$$\Psi_n(x,y,z) = \frac{1}{L_x}\exp(ik_x x)\Psi_n(y,z) \tag{3.12}$$

$$\Psi_m(x,y,z) = \frac{1}{L_x}\exp(ik_x^{'} x)\Psi_m(y,z) \tag{3.13}$$

At this point, we are ready to evaluate the matrix element using the above perturbation Hamiltonian. The matrix element expressing the effect of the perturbing potential on the incident electronic wavefunction is:

$$\begin{aligned} M_{nm}(k_x, k_x) &= < k_x^{'}, n_q | H_{ac} | k_x, n_{q'} > \\ &= \mathrm{D_{ac}} \sum_q \sqrt{\frac{\hbar}{2MN\omega_q}} (ie_q.q) I_{nm}(q_y, q_z) \\ &\quad \sqrt{n_q + \frac{1}{2} \pm \frac{1}{2}} \delta(k_x + k_{x'} \pm q_x) \end{aligned} \tag{3.14}$$

where I_{nm} is given by:

$$I_{nm}(q_y, q_z) = \int \psi_n(y,z) \psi_m(y,z) \mathrm{e}^{i(q_y y + q_z z)} dydz \tag{3.15}$$

The scattering probability as given by the Fermi golden rule is:

$$S_{nm}(k_x, k_x^{'}) = \frac{2\pi}{\hbar} |M_{nm}|^2 \delta[E(k_x^{'}) - E(k_x) - \Delta(E)] \tag{3.16}$$

Once we compute the above scattering transition, we compute the expression for the acoustic scattering phonon scattering rate:

$$\Gamma_{nm}^{ac} = \frac{D_{ac}^2 k_B T \sqrt{2m^*}}{2\hbar\rho {v_s}^2} J_{nm} \frac{(1 + 2\alpha E_f)}{\sqrt{E_f(1 + \alpha E_f)}} \Theta(E_f) \tag{3.17}$$

where $\Theta(E_f)$ is the Heaviside step function and ensures energy conservation after scattering and J_{nm} is the overlap integral associated with the electron-phonon interaction and is given by:

$$J_{nm} = \int \int \psi_n^2(y,z) \psi_m^2(y,z) dydz \tag{3.18}$$

3.2.2 Intravalley phonon scattering rate

Intervalley phonon scattering rate is mediated by short wavelength acoustic and optical phonons with wave vectors near the zone boundary. The interaction Hamiltonian is assumed to take the following form:

$$H^{iv}_{el-ph} = D_{iv}.\mathbf{u}(r) \tag{3.19}$$

where D_{iv} is the intervalley deformation potential. Repeating the procedure we presented above for acoustic phonon, the scattering rate becomes

$$\Gamma^{iv}_{nm} = \frac{D^2_{iv} k_B T \sqrt{m^*}}{\sqrt{2}\hbar\rho {w_o}^2} J_{nm}(N_o + \frac{1}{2} \pm \frac{1}{2}) \frac{(1+2\alpha E_f)}{\sqrt{E_f(1+\alpha E_f)}} \Theta(E_f) \tag{3.20}$$

3.2.3 Multi-subband 3D Monte Carlo

Here we present the effect of size quantization in confined quasi-1D trigate MOSFETs. Figure 3.4 describes the implementation of subband scattering correction to the MC kernel. A schematic of a trigate device considered here is also presented in Figure 3.5. In our simulator, the carrier movement is supposed to be semi-classical along the x axis, and is separated from the y and z axes, along which the energy is quantized. The whole device is divided in N_x slices from source to drain where the time-independent 2D Schrödinger equation is solved. Their solutions provide the eigenpairs (E_n, ψ_n) associated with the subbands necessary to calculate the charge density used as input of a 3D Poisson equation. Both equations are self-consistently coupled and iterated to get the desired convergence criterion. Figure 3.6(a) shows a comparison of the subband scattering rate and the bulk scattering rate for a triple gate SOI nanowire whereas Figure 3.6(b) demonstrate the velocity profile.

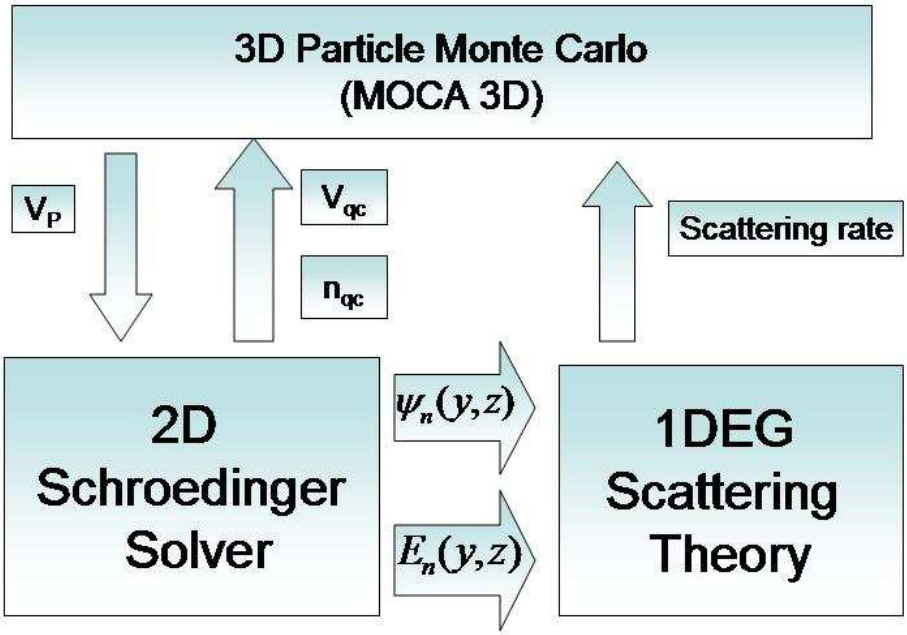


Figure 3.4: Flowchart describing the inclusion of multi-subband model into our 3D Monte Carlo simulator.

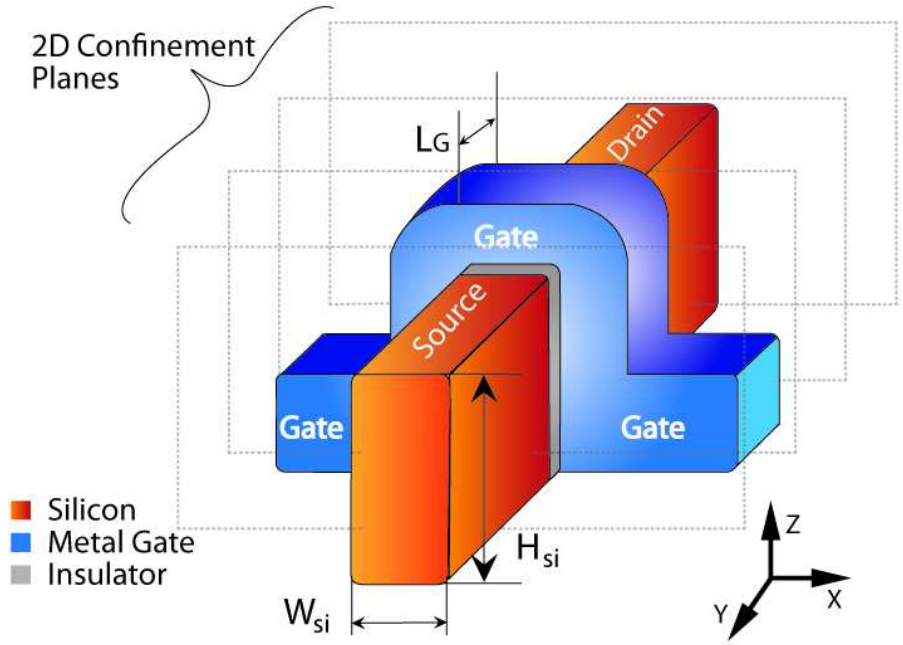
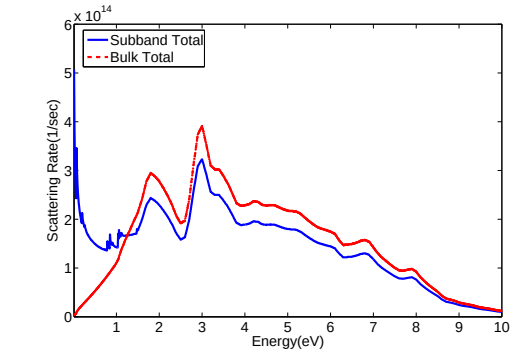
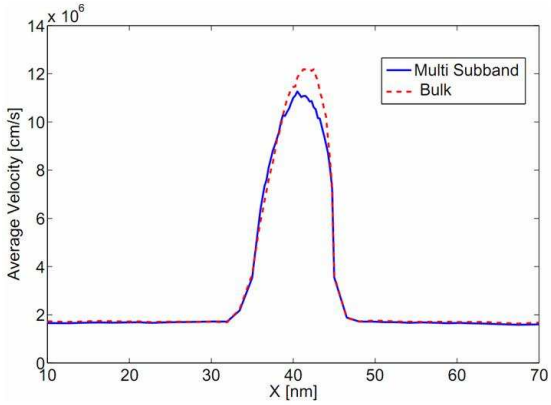


Figure 3.5: Application of multi-subband to a trigate MOSFET. The governing quantum mechanical equation is solved for each slice along the channel.



(a)



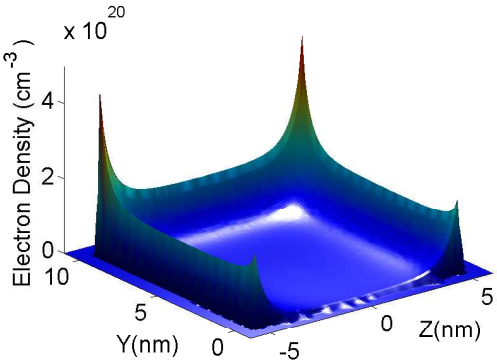
(b)

Figure 3.6: Comparison of the (a) scattering rate and (b) average velocity for quantum-corrected MC simulation of a trigate MOSFET with and without subband scattering. The reduction in carrier velocity is attributed to the increase in subband scattering rate as compared to bulk at energies below 1 V

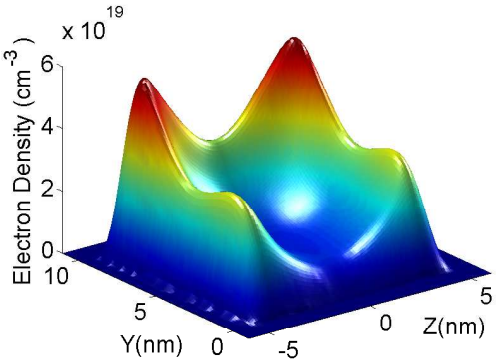
3.3 Size Quantization in Confined Structures

The inclusion of quantum effects on the physical description of semiconductor devices has important consequences when dimensions are reduced. Simulation results (not shown here) have verified that the importance of the quantum correction potential $V_{qc}(y, z)$ increases through the whole silicon slab, not only next to the interfaces as device dimensions are reduced. Figure 3.7 shows a prototypical electron density plot along the cross-section of a trigate (TG) MOSFET with square cross-section, $T_{Si} = H_{Si} = 10$ nm, under an applied bias $V_G = 0.5$ V. In Figure 3.7(a) the classical electron distribution reveals a maximum located at the insulator interface, especially at the corners, and an overestimation of the peak density is clearly observed when these results are compared with their corresponding quantum description depicted in Figure 3.7(b), which shows a shift of the maximum of carrier concentration away from the Si-SiO_2 interface. The average distance of this shift is a function of different parameters [44] and has been modeled in a parameter known as the inversion charge centroid (z_{ave}). As a consequence of this shift, the total gate capacitance is reduced since it is now the series combination of two capacitances: the gate insulator (ϵ_{ox}/t_{ox}) and the inversion charge capacitance (ϵ_{si}/z_{ave}). This effect has been thoroughly studied in bulk MOSFETs; in spite of that, there is still a lack of reliable models for this effect in multigate devices where the nonplanar structure of the devices precludes the use of the well-known bulk models. The reduction of the total gate capacitance leads to a reduction of the inversion charge for the same gate overdrive.

Moreover, it has been demonstrated that classical simulations of electron transport lead to a higher average velocity in the channel. This behavior can be traced to the energy distribution of particles. In the classical case, most of the electrons are found in the inversion region near the interface where the transverse fields are high, causing the particles to be more energetic. Phonon interactions redirect some of this energy along the longitudinal direction with



(a)



(b)

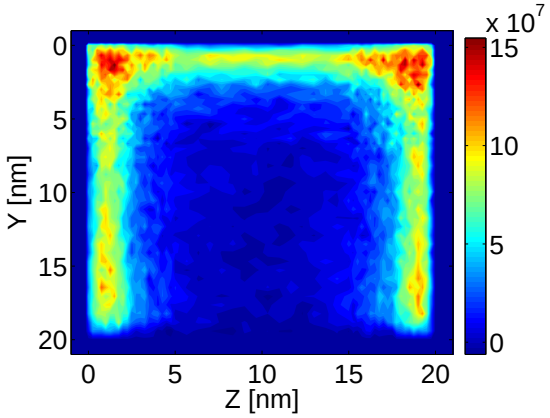
Figure 3.7: Electron density calculated for a TG MOSFET with square cross-section $T_{Si} = H_{Si} = 10$ nm. In (a), the classical electron distribution is shown for an applied gate voltage of 0.5 V and the peak density is reached at the Si-SiO_2 interface. (b) represents the quantum electron density at $V_G = 0.5$ V and the maximum is shifted at a certain distance from the interface.

corresponding higher average velocity in the channel. In contrast, for the quantum-corrected case, a large percentage of charge carriers flow closer to the middle of the cross-section with lower average energy [45].

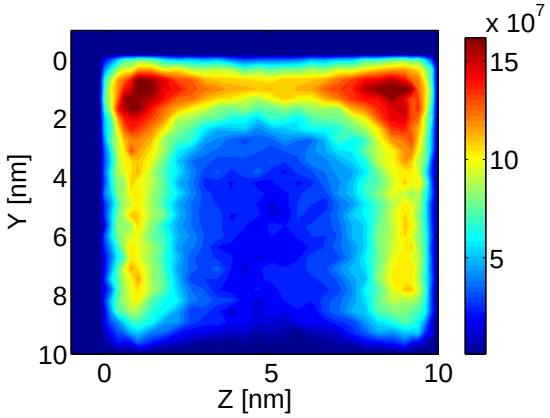
This effect is shown in Figure 3.8, where we show current density distribution in a TG (sliced at $x = L_g/2$) for square cross sections $T_{Si} = H_{Si} = 5, 10$, and 20 nm for $V_G = V_D = 0.5$ V. As the lateral dimensions are reduced, electrons from independent channels start to interact and more carriers are placed in the central region producing the so-called volume inversion effect [42, 46]. This effect is clearly amplified for the TG with the smallest cross-section where we observe that the magnitude of current flowing in the middle of the 5 nm x 5 nm fin is comparable to the magnitude of current flowing in the corner of the larger cross-sections. The main consequence of this phenomenon is that electrons on average are located farther from the insulator interface and, in this regime, quantum corrections are even more noticeable.

3.4 Volume Inversion and Charge Compensation

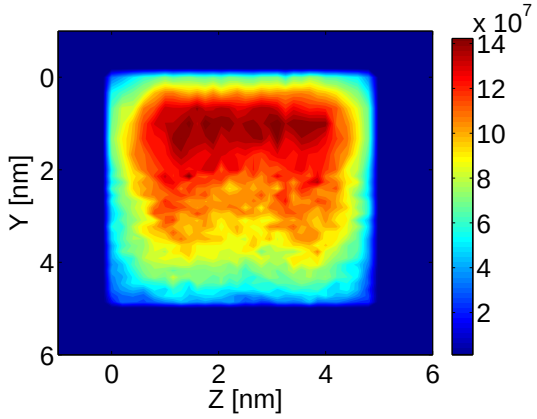
Figures 3.9 and 3.10 show the current variation along cross-section of the TG as we scale down the cross-section ($T_{Si} = H_{Si}$) from 30 nm to 10 nm. Here, we show both the raw drain current in amperes and the current normalized with respect to the perimeter, $W_{eff} = 2H_{Si} + T_{Si}$. We observe that quantization yields stronger coupling between the channels which in effect produces a volume inversion effects and the current is concentrated at the device corners irrespective of the device cross-section. Figures 3.11 and 3.12 depict the current dependence on the effective channel width. Note that the normalized device current drops non-linearly as we reduce the cross-section of the TG transistors. The reduction of the perimeter produces a lower value of the normalized current as can also be seen in Figure 3.12, which depicts the normalized drain current obtained in steady state for three different square cross-sections. The explanation to this is hinted in Figure 3.13 where we observe that as



(a)



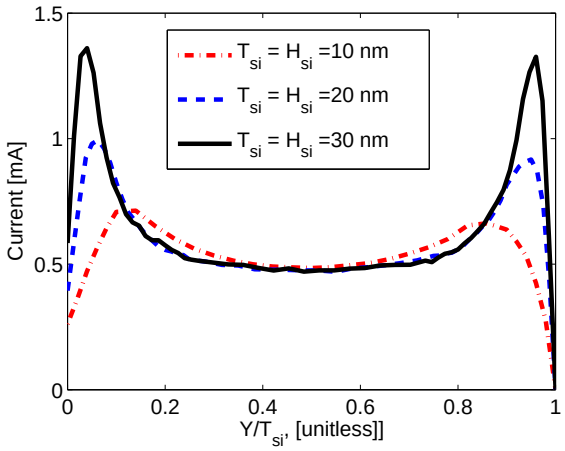
(b)



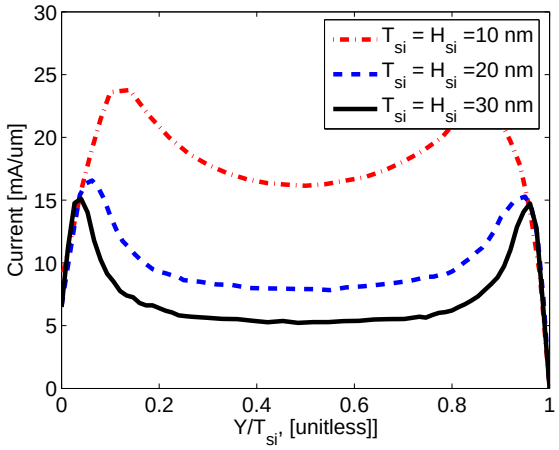
(c)

Figure 3.8: Contour plot showing the current density distribution, at $x = L_g/2$, along the cross-section of a 20 nm channel length TG with cross-sections $T_{Si} = H_{Si} = 5, 10$, and 20 nm at an applied bias $V_G = V_D = 1.0$ V.

cross-section decreases, the potential along different axes is modified, and as a consequence the charge distribution is also altered. This, in addition to the observation that volume inversion effect increases as we reduce the cross-section, suggests that the perimeter perhaps is not the most viable choice for effective width in the cross-sections considered in this work.

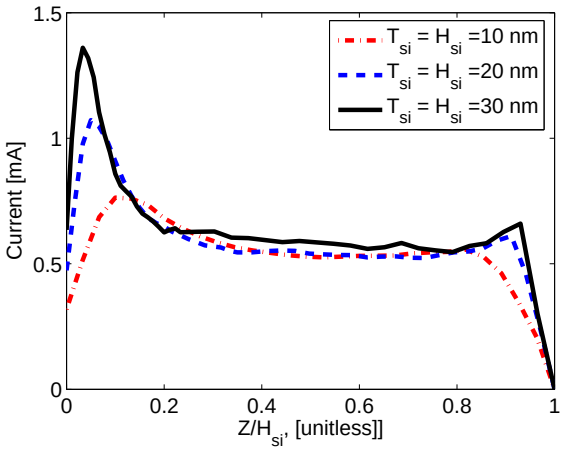


(a)

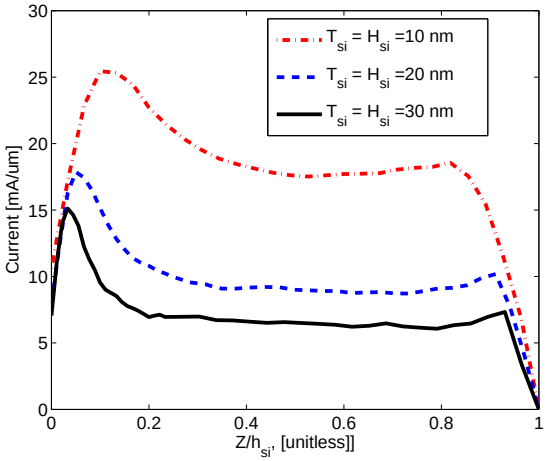


(b)

Figure 3.9: Evolution of drain current along the y direction for trigates of cross-section $T_{Si} = H_{Si} = 20$ nm. The bottom plot corresponds to the drain current normalized with respect to the perimeter, whereas the top figure shows the drain current of each trigate in amperes.



(a)



(b)

Figure 3.10: Evolution of drain current along the z direction for trigates of cross-section $T_{Si} = H_{Si} = 20$ nm. The bottom plot corresponds to the drain current normalized with respect to the perimeter, whereas the top figure shows the drain current of each trigate in amperes.

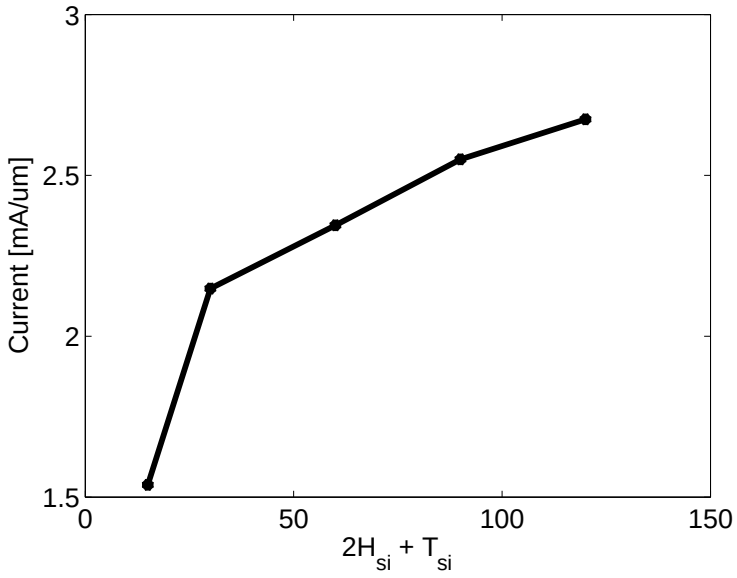


Figure 3.11: Normalized drain current as a function of the trigate perimeter defined as $2H_{Si} + T_{Si}$.

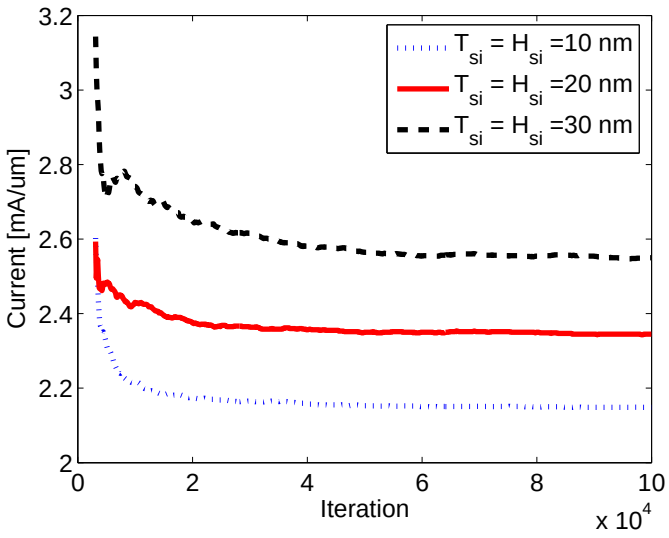
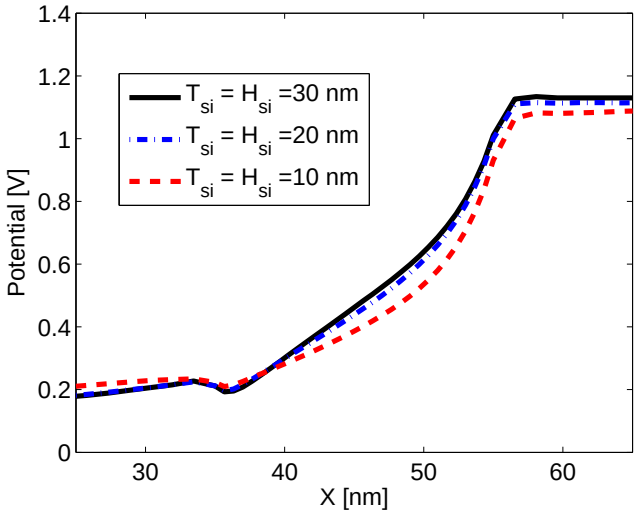
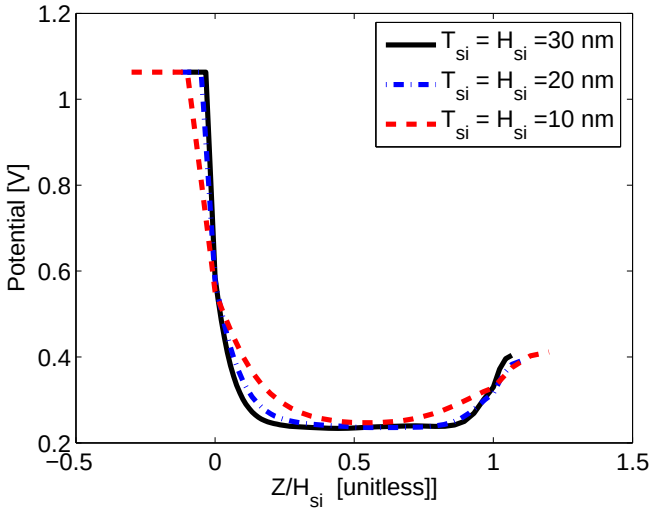


Figure 3.12: Steady state drain current obtained from MoCa normalized with respect to the perimeter $(2H_{Si} + T_{Si})$ of each structure.



(a)



(b)

Figure 3.13: Potential distribution along the x direction (a) and z direction (b) for a TG with cross-section $T_{Si} = H_{Si} = 20$ nm at an applied bias $V_d = V_g = 1.0$ V.

3.5 Summary

We have have presented the details of our in-house 3D quantum-corrected particle Monte Carlo simulator. Predictive simulation tools, which capture both relevant quantum mechanical effects and transport properties, play a vital role in the development and evaluation of highly confined quasi-one-dimensional devices. To extend the validity of our 3D Monte Carlo simulation, we have augmented the electrostatic Schrödinger based correction by introducing a multi-subband quantum correction approach which in essence adds quantum correction to carrier transport via scattering. Using our 3D Monte Carlo, we demonstrated the influence of size quantization on quasi-1D silicon-on-insulator TG MOSFETs. We have shown that the interplay between carrier confinement, and the coupling between the top and bottom gates leads to field compensation which as we will see in the next chapter limits the overall device performance on multigate structures.

CHAPTER 4

ISO-THERMAL SIMULATION OF MULTIGATE NANOWIRE MOSFET

4.1 Overview

In this chapter we employ quantum-corrected 3D full-band Monte Carlo simulation coupled with full phonon dispersion to investigate the electro-thermal behavior of silicon-on-insulator (SOI) multigate devices at the limit of cross-section scalability. We particularly explore the dependence of short-channel effects and Joule heating on the lateral scaling of the cross-section. The basic tradeoff between n-channel double-gate (DG), trigate (TG), and gate-all-around (GAA) transistors with square cross-section lengths varying from 30 nm down to 5 nm are presented.

Figure 4.1 schematizes the cross-section layout of the devices considered in this chapter. A summary of the device parameters is also shown in Table 4.1. We have simulated DG, TG, and GAA n-MOSFET structures with a square cross-section length (channel thickness (T_{Si}) equal to the channel height (H_{Si})) varying from 5 to 30 nm . The substrate was lightly doped (P type) with $N_A = 10^{16}$ cm^{-3}. Unless otherwise stated, the devices have a gate length (L_g) of 20 nm and a gate oxide thickness (t_{ox}) of 10 Å. The source (S) and drain (D) were n doped with 10^{20} cm^{-3} impurities with the S/D junctions ending abruptly at the gate edge. The channel region starts at $x = 35$ nm for all the results shown in this chapter.

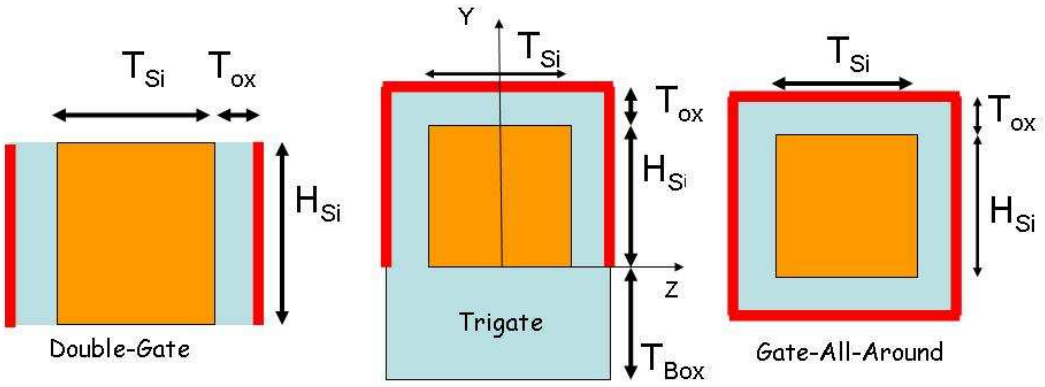


Figure 4.1: Schematic diagram showing the cross-section of DG, TG, and GAA transistors considered in this chapter.

Table 4.1: Device Parameters of Simulated MOSFETs

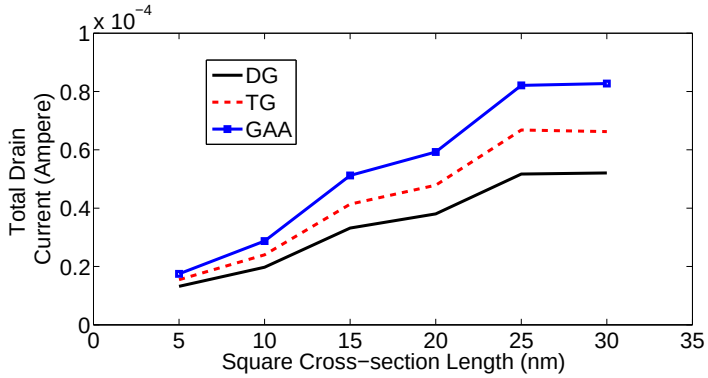
L_g	t_{ox}	$T_{Si} = H_{Si}$	S/D Doping	Channel Doping
5 - 20	1	5 - 30	1×10^{20}	1×10^{16}
[nm]	[nm]	[nm]	[cm^{-3}]	[cm^{-3}]

4.2 Electrical Performance of Multigate Structures

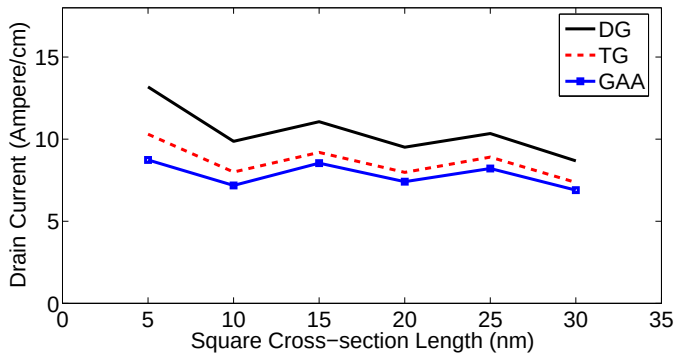
4.2.1 Influence of cross-section scalability on drain current

Within the limits of a large cross-section, we expect the total charge obtained in the device to resemble that of a bulk MOSFET with an effective width defined by the perimeter of the gate αT_{Si}, where α amounts to the number of gates. As such, one would ideally expect TG and GAA to yield, respectively, 50% and 100% current enhancement relative to a DG MOSFET with all other device parameters being equal, since each added gate effectively introduces a new inversion channel. In this section, we look at how current drive in multigate architectures deviates from this ideal behavior at the limit of small cross-sections.

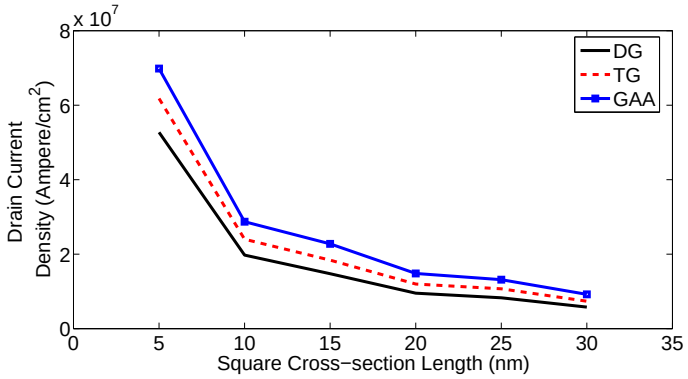
The results obtained are summarized in Figures 4.2 and 4.3, where we show the change in drain current and inversion charge as the square cross-section length ($T_{Si} = H_{Si}$) is varied from 5 to 30 nm. All findings shown in this section were computed for a 20 nm channel length at an applied bias $V_G = V_D = 0.5$ V. Figure 4.2(a) presents drain current variation as a function of the cross-section length. We observe that the total drain current increases with lateral scaling of the cross-section; nonetheless, the relative drain current difference among the multigate structures reduces with scaling. Additionally, it is interesting to note how normalization with respect to the perimeter (Figure 4.2(b)) reveals that the drain current per channel is most efficient for the structure with the least number of gates. In Figure 4.4, we have calculated the linear density obtained by integration of the electron density at $x = L_g/2$ along the cross-section. There is an apparent parallelism between Figure 4.2(a) and Figure 4.4 where we see that the boost in both the inversion charge and drain current, because of using more gates, becomes less prominent as we reduce the cross-section. This effect is stressed in Figure 4.3 which demonstrates that although GAA and TG MOSFETs provide an unconcealed current improvement as compared to DG,



(a)



(b)



(c)

Figure 4.2: Cross-sectional variation of the (a) total drain current, (b) drain current normalized with the perimeter, and (c) drain current normalized with respect to the cross-section for $L_g = 20$ nm and $V_D = V_G = 0.5$ V.

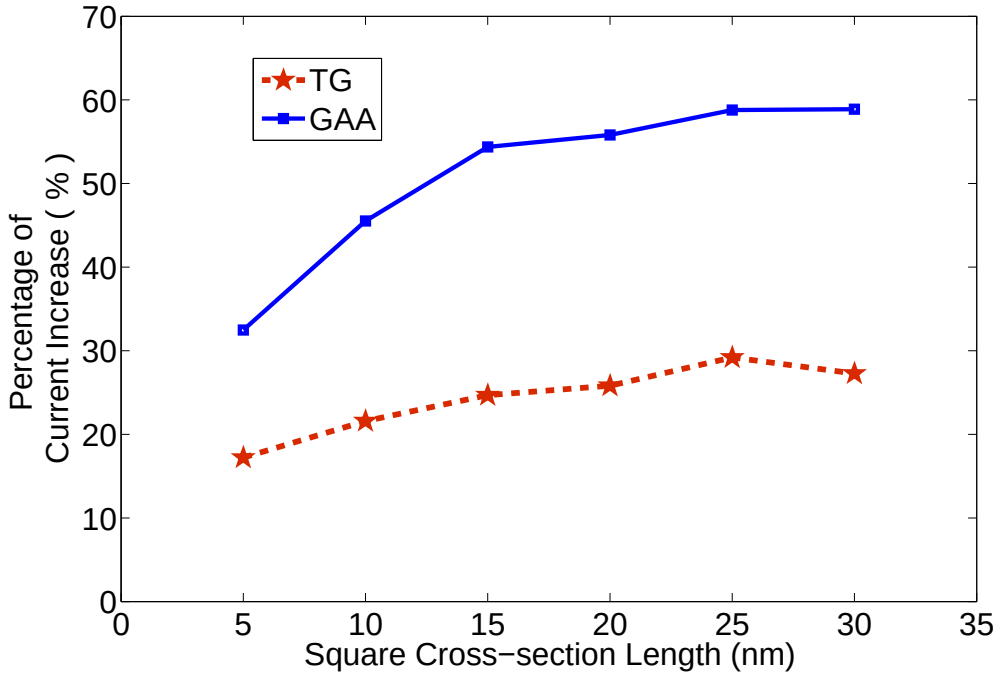


Figure 4.3: Performance enhancement of GAA and TG with respect to DG. Note that the gain in current is well below the ideal 100% and 50% current improvement for GAA and TG, respectively. This gain is reduced as we downsize their cross-section.

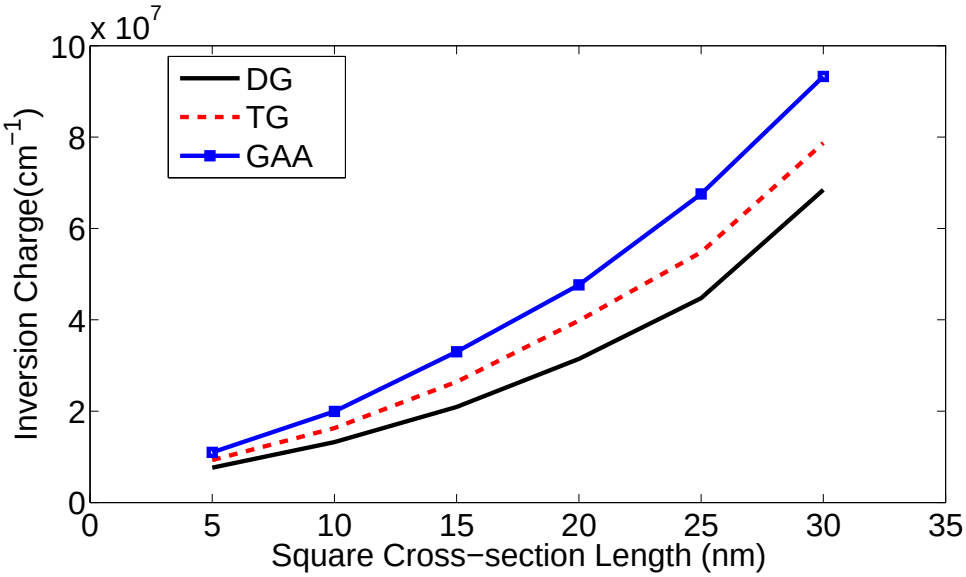


Figure 4.4: Inversion charge along the cross-section for DG,TG, and GAA. All three devices were simulated under $V_D = V_G = 0.5$ V and had a channel length of 20 nm. The gap in charge density between the simulated structures decreases as we shrink the cross-section.

this gain is substantially below their ideal 100% and 50% aptitude, respectively. Furthermore, the reduction of drive currents below their expected ideal limits is even more pronounced as we shrink the cross-section. The main inference obtained from these plots is that shrinking the lateral cross-section to the nanometer scale results in a considerable reduction of the relative differences between multigate devices. Accordingly, to gauge the effectiveness of DG, TG, and GAA at such limits, one has to assess the gain in current improvement obtained from each of these structures with the overall cost and ease of fabrication associated with them.

4.2.2 Comparison of device performance and short-channel effects

With the scaling of the cross-section, the electrical coupling between the silicon channel and the gate increases as we add more gates. Consequently, we accomplish better suppression of SCEs and a promise of device scalability, provided that technological challenges to fabrication are met. In this section, we investigate the effect of the nonlinearity of electron current gain due to lateral scaling of the cross-section on device performance. Most of the analysis presented will focus on devices with the 5 nm square cross-section length to highlight the differences between multigate architectures at the edge of lateral scalability.

Our MC simulator allows the calculation of the linear charge density through the whole device for an arbitrary bias as it is shown in Figure 4.5 for DG, TG, and GAA transistors with a silicon fin size of 5 nm. We observe an important distinction between the line charge profile inside the channel, due to the disparity between the number of gates in these devices. The carrier concentrations inside the channel are largest for GAA with its four gates, followed by the TG with three gates, and finally the DG with two gates. Nevertheless, we note that there is no distinction in the electron density along the source and drain regions between the various multigate structures. Similar

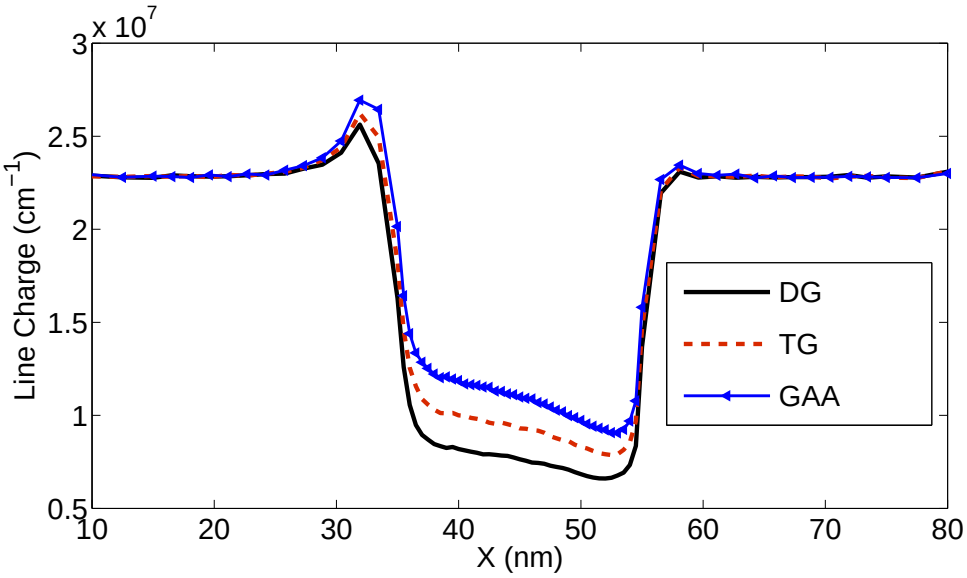


Figure 4.5: Spatial variation of the line charge profile along the channel for DG, TG, and GAA with the following parameters: $T_{Si} = H_{Si} = 5$ nm, $L_g = 20$ nm and $V_D = V_G = 0.5$ V. Note that the line charge profile is the same for all structures in the source and drain regions.

results were obtained for the all other lateral dimensions. This point imposes important difference in the average velocity profile of DG, TG, and GAA along the S/D reservoir, since the charge density in those regions is the same but their respective drive current is different. Figure 4.6 illustrates the clear distinction between the average velocity profile of DG, TG, and GAA inside the source and drain regions. It is interesting to highlight how this trend is inverted along the channel where electrons in the DG show higher average velocity than in the TG and GAA; this is exactly the opposite of what was previously described in the contacts. Furthermore, our simulations consistently reveal that the disparity in average velocity between these devices becomes increasingly explicit with lateral scaling of cross-section.

In order to explain the implication of higher velocity in the source and drain regions on device performance, the average conduction band along the channel for the three devices is presented in Figure 4.7. We observe a voltage drop in the source and drain because of the velocity profile in those regions. We have chosen the 5 nm by 5 nm cross-section to highlight the potential drop along the highly

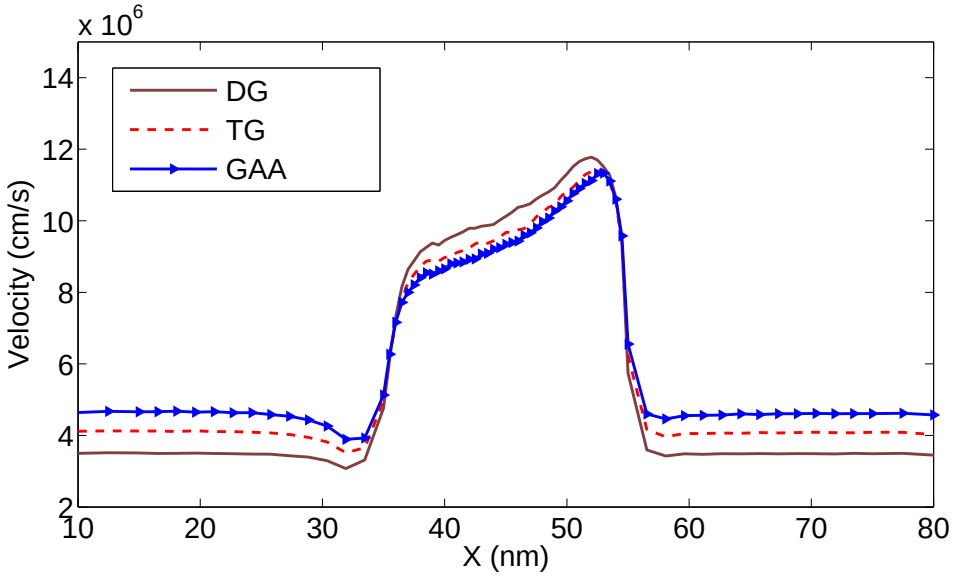


Figure 4.6: Average velocity profile for DG, TG, and GAA with $T_{Si} = H_{Si} = 5$ nm, $L_g = 20$ nm and $V_D = V_G = 0.5$ V.

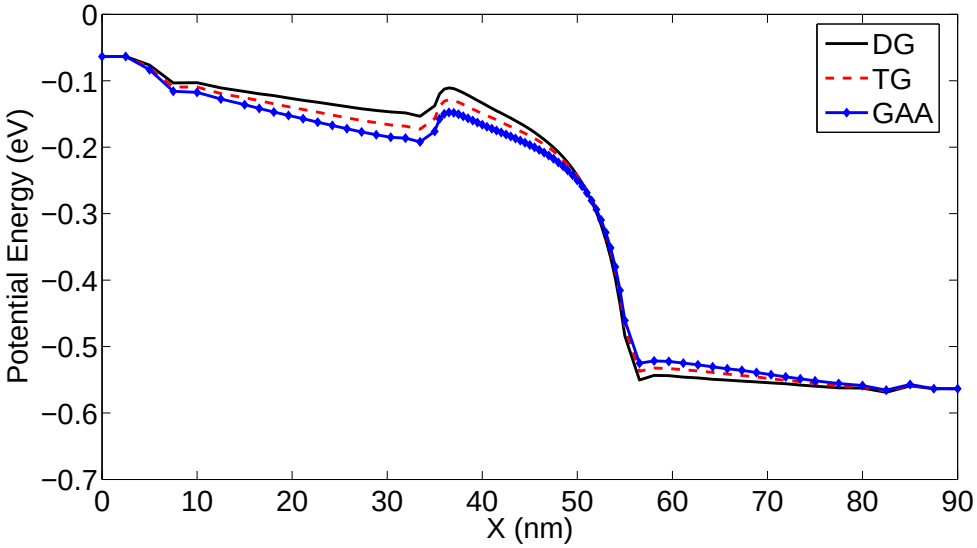


Figure 4.7: Conduction band profile sliced at $y = z = T_{Si}/2$ along the channel for DG, TG, and GAA MOSFETs with the following parameters:
$V_D = V_G = 0.5$ V, $T_{Si} = H_{Si} = 5$ nm and $L_g = 20$ nm. Potential drop in the source and drain regions increases with the number of gates.

Table 4.2: Series Resistance ($V_G = V_D = 0.5$ V)

	$\triangle V_S$	$\triangle V_D$	I	R_S	R_D
GAA	129	38.3	8.7	147.7	43.9
TG	110	26.1	10.3	106.7	25.3
DG	90	13.1	13.2	68.3	9.93
	[mV]	[mV]	[A / cm]	[$\Omega\mu$m]	[$\Omega\mu$m]

doped source side (from 10 nm to 35 nm) and also because higher differences in the average electron velocity were found for this case. In our simulations the contribution of the contact resistance is neglected; nonetheless, our results systematically reveal that for a given cross-section, both the voltage drop and the computed series resistance is largest for GAA, followed by TG and then DG. An estimation of the series source/drain resistance for the three devices is shown in Table 4.2. Values higher than 140 $\Omega\mu$m (as R_S for the GAA) are against ITRS recommendations. Since the electrons reach the drain region with high energies, the drain contact appears to be less resistive than the source [47]. In order to alleviate the influence of series resistance, doping density may be increased or the length of the source and drain regions may be reduced. However, this may degrade other aspects of device performance, and the tradeoffs should be considered. Similar inferences may be obtained by looking at the longitudinal electric field profile along the channel shown in Figure 4.8, where we observe that the electric field profile in the S/D region increases as we add more gates. Meanwhile, since increasing the number of gates would naturally enhance gate control over the channel, we see this relation is inverted along the channel.

Figure 4.9(a) shows the I_{DS}–V_{DS} characteristic of a 20 nm channel length DG and GAA MOSFET with cross-section size of $5, 10$, and 20 nm. We observe considerable deterioration of the drain conductance, g_D, regardless of the number of gates considered, when $L_g = T_{Si} = H_{Si} = 20$ nm. This illustrates the unsubtle need for multigate structures to obey design criteria in order to subdue short-channel effects. A rather transparent design rule to enhance gate control

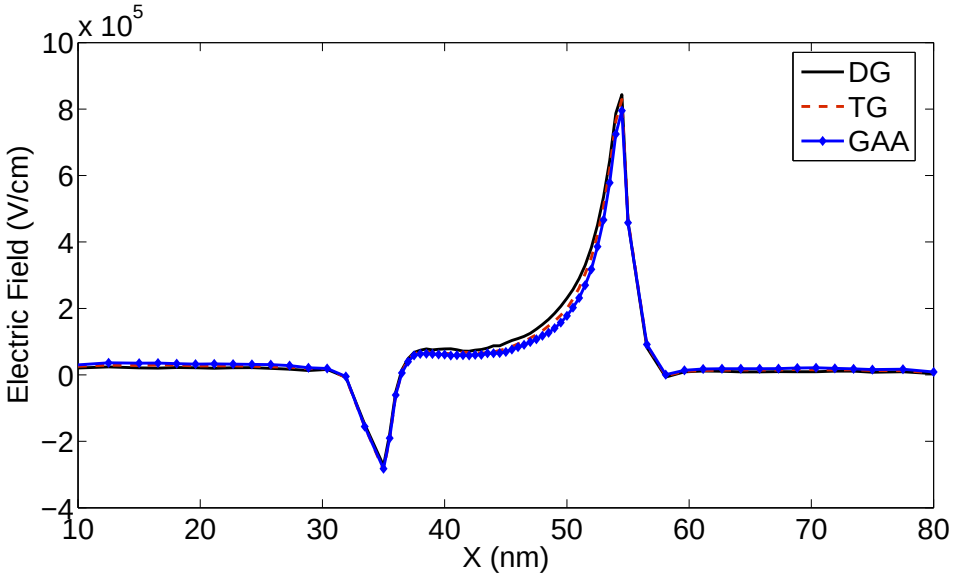


Figure 4.8: Electric field profile along the channel for a DG, TG, and GAA with square cross-section $T_{Si} = H_{Si} = 5$ nm, $L_g = 20$ nm and $V_G = V_D = 0.5$ V. As the number of gates is increased the electric field inside the channel is reduced.

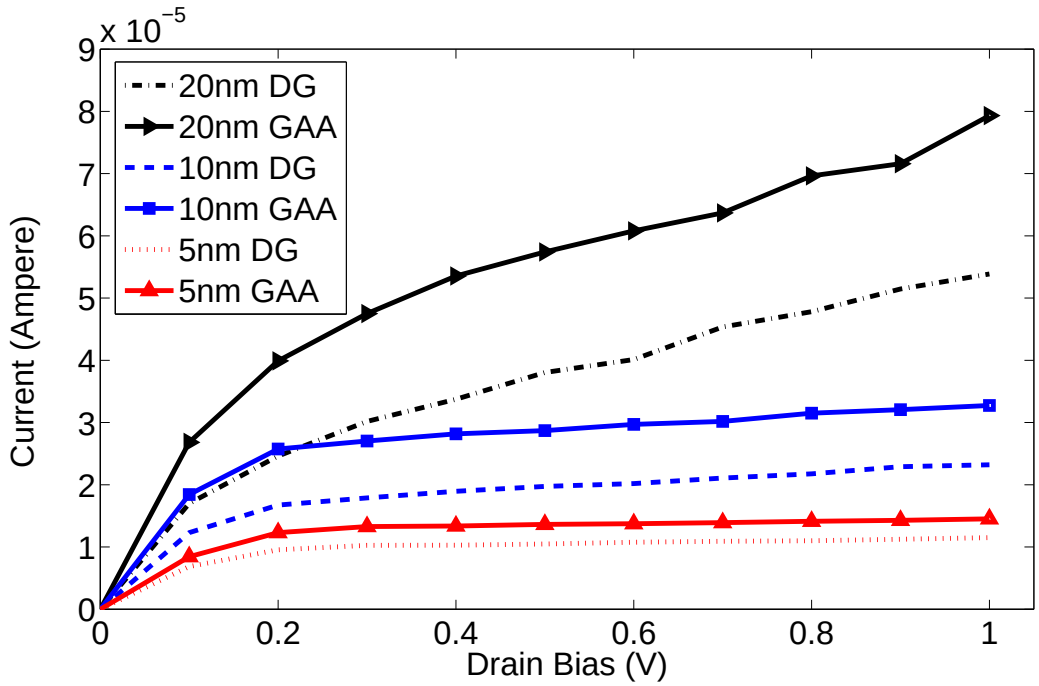
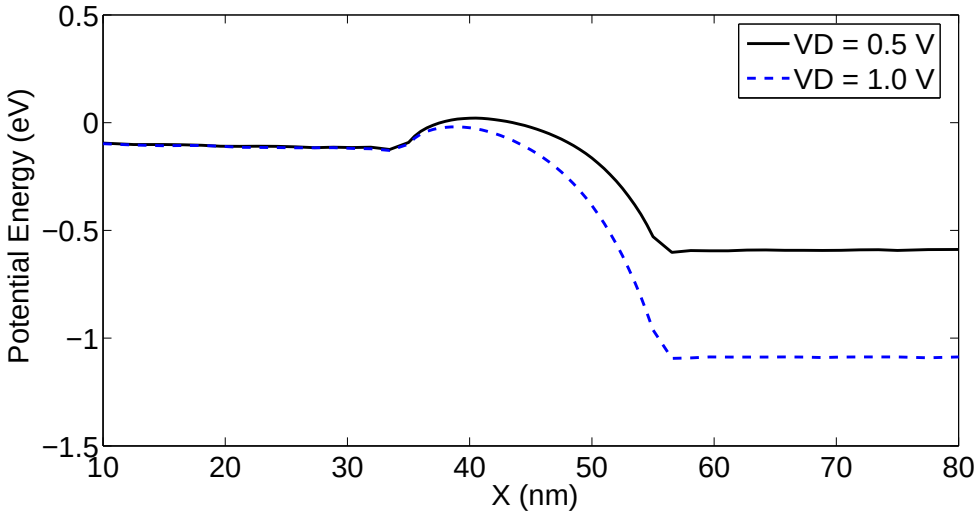


Figure 4.9: I-V characteristics of $T_{Si} = H_{Si} = 5$ nm, 10 nm, and 20 nm DG and GAA MOSFETs with $V_D = V_G = 0.5$ V.

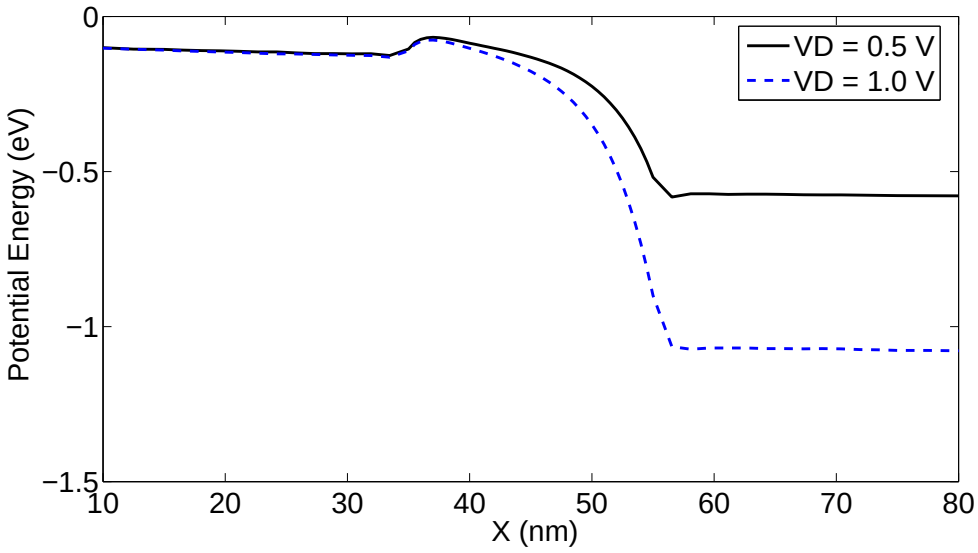
over the channel is to ensure that $T_{Si} \leq L_g/2$. We see substantial improvement of g_D as we reduce the cross-section as shown for the 10 nm and 5 nm silicon fins. Interestingly, the drain conductance calculated with two or four gates is quite similar and g_D appears to be independent of the multigate structure considered. Finally, in order to estimate the drain-induced barrier lowering (DIBL) in these structures, we show a comparison of the conduction band edge profile for a GAA MOSFET with cross-sections of 20 nm and 10 nm in Figure 4.10. As anticipated, we found clear reduction in DIBL when the cross-section is reduced to 10 nm. Similar results where obtained for other structures.

4.2.3 Ballistic limit statistics: 2D example

In this section, our aim is to investigate the ballisticity of deeply scaled devices. We utilize a 2D full-band Monte Carlo (MoCa2D) with quantum correction to monitor particle scattering in a DG MOSFET with a silicon thickness of 2 nm and a channel length varied from 5 to 20 nm. Each particle entering the channel from the source region may either backscatter to the source, cross through the channel with no scattering, or experience one or more scatterings before it gets collected in the drain. The acquired statistics are summarized in Table 4.3, which was computed by tracing the trajectories of individual electrons inside the channel for a gate and drain bias of 0.5 V. It is evident from the table that the number of scattering events decreases as a function of channel length; however, even at such very small dimensions the devices are not entirely ballistic. We have also collected statistics at $V_G = V_D = 0.8$ V and we found a slight increase in the percentage of ballistic electrons. The backscattering statistics for a gate and drain bias of 0.5 V is shown in Table 4.4 and, interestingly, we see that a large fraction of the electrons do not backscatter (first row of the table). In addition, only a small fraction of the overall particles experience more than two backscattering events. Simulations for $V_G = V_D = 0.8$ V produced similar results.



(a)



(b)

Figure 4.10: An example of the average conduction band edge profile obtained for a 20 nm channel length GAA MOSFET under an applied bias $V_D = V_G = 0.5$ V. We see that in (a) the device clearly suffers from DIBL effect when $T_{Si} = H_{Si} = L_g = 20$ nm. In (b) we observe the DIBL is alleviated as we reduce the cross-sectional length to $T_{Si} = hsi = L_g/2$.

Table 4.3: Overall Particle Scattering Statistics ($V_G = V_D = 0.5$ V)

No. of Scatterings	L_g [nm]			
	20	15	10	5
0	43.5377	48.2446	54.7889	68.0814
1	24.0662	25.9196	26.5784	22.4583
2	15.5195	13.7623	11.2084	6.8050
3	8.3519	6.5113	4.5793	1.8657
4	4.3071	3.0114	1.8116	0.5144
5+	4.2176	2.5509	1.0334	0.2751
Total	100.0000	100.0000	100.0000	100.0000
	unit: %			

Table 4.4: Particle Backscattering Statistics ($V_G = V_D = 0.5$ V)

No. of Scatterings	L_g [nm]			
	20	15	10	5
0	65.0070	66.2766	68.8277	74.4315
1	20.7797	21.3919	21.5091	19.5449
2	7.4352	7.0822	6.2167	4.4069
3	3.3260	2.8607	2.0930	1.1469
4	1.6332	1.2566	0.8052	0.3218
5+	1.8188	1.1321	0.5482	0.1481
Total	100.0000	100.0000	100.0000	100.0000
	unit: %			

4.3 Joule Heating in Multigate Devices

The generation of heat in silicon MOSFETs is a hot topic due to the continued scaling of device dimensions, which allows increasing packing densities of transistors and switching at higher frequencies. As a consequence, device performance and circuit reliability may be degraded due to high temperatures effects. Joule heating is caused by emission of phonons as electrons traverse through the semiconductor channel. In common silicon MOSFETs, most of the emission is concentrated in the small region where the channel meets the drain. In this section, we have employed a 3D self-consistent Monte Carlo device simulator with full electron bandstructure and a full phonon dispersion relationship to analyze the influence of size effects on thermal emission.

Data on scattering events were obtained from a Monte Carlo simulation with an iterative algorithm devised to make all scattering events involving phonon energy and momentum conserving with the full phonon dispersion relationship. The algorithm starts at each scattering event with an estimate for the energy of the phonon involved. This energy can be calculated from the momenta involving transitions between bottoms of the X and L equivalent valleys in silicon [48]. Then the final state can be looked up from a table of electron energies and momenta. Finally, this result is checked with the phonon dispersion to ensure energy conservation. If energy is not conserved to within a small tolerance dictated by collision broadening [49, 50, 51], then the final state is rejected and a new state is sought. To ensure fast convergence, the new state is calculated based on the previous estimate of the phonon energy. After several iterations, a final state is found that satisfies both momentum and energy conservation. This gives us a more accurate value for the phonon momentum and energy. Another consequence of this process is ensuring that all possible transitions are well represented, not just those with a fixed, predetermined phonon energy.

One additional quantum effect considered in the simulation is that of finite state lifetime and collisional broadening. Due to high scattering rates present in

such scaled devices, the energy shift due to scattering becomes comparable to the thermal energy and the impact of quasi-particle states needs to be taken into account. In the present implementation, this is accomplished by selecting the final state after each phonon scattering from a Lorentzian distribution instead of strict energy conservation [49]. In order to avoid unphysical accumulation of broadening, the non-accumulated broadening (NAB) algorithm due to Register and Hess [51] is employed. This allows each particle to take on a full range of quasi-particle states without conflicting with the full phonon dispersion.

Once the simulation run is complete, data on all phonon events that occurred are tabulated and total net emission per electron is obtained by summing the energy of all the emitted phonons, subtracting the energy of all the absorbed phonons, and finally dividing by the total number of simulated electrons. This provides a straightforward way of analyzing the thermal emission of a simulated device. Figure 4.11 summarizes the effect of reducing the cross-section on energy loss rate for all three multigate architectures. The energy loss rate increases as the cross-section is reduced and as more gates are added. This is intuitive since, classically, Joule heating is defined by the product of current density with electric field. Consequently an increase in current drive capability will naturally lead to an increase in power dissipation. This statement is demonstrated by Figure 4.12 and 4.13 where the trends of the energy loss rate of the multigate devices as a function of drain bias and channel length are plotted. In both cases, the behavior is intuitive since the electric field increases when drain voltage grows or the channel length is shrunk, providing more energy to electrons crossing the channel. Accordingly, thermal and electrical design considerations may appear at first to be conflicting, since an increase in current density, which is desirable electrically for enhancing device performance, would result in an increase in Joule heating. This is made apparent from the average energy plots shown in Figure 4.14 and 4.15 where, although the average energy inside the channel is reduced for a higher number of gates, the trend inside the drain region is reversed due to the current drive difference. This behavior is consistent

with results shown in Figure 4.6 where we observe that the average electron velocity along both the channel and source and drain regions follows the same trend as that of the energy.

4.4 Summary

We have studied the effect of scaling the cross-section of DG, TG, and GAA MOSFETs on device performance and heat generation rate using a self-consistent quantum-corrected 3D Monte Carlo device simulator with full electron bandstructure and a full phonon dispersion. Our calculations show that the variation of current gain in these structures deteriorates when cross-sections are reduced, and the number of gates per se is not a definite indication of the improvement of device performance. Further, more energy per electron is dissipated when both the cross-section and channel length are reduced as Moore's law demands. In addition, it was shown that a higher number of gates increases the amount of energy dissipated. As such, for comparative purposes, a high Joule heating rate in and by itself is not ill-favored, unless it is coupled with SCEs and fabrication concerns. Consequently, for an optimal device design one has to weigh collectively both the thermal and electrical issues of the device under consideration.

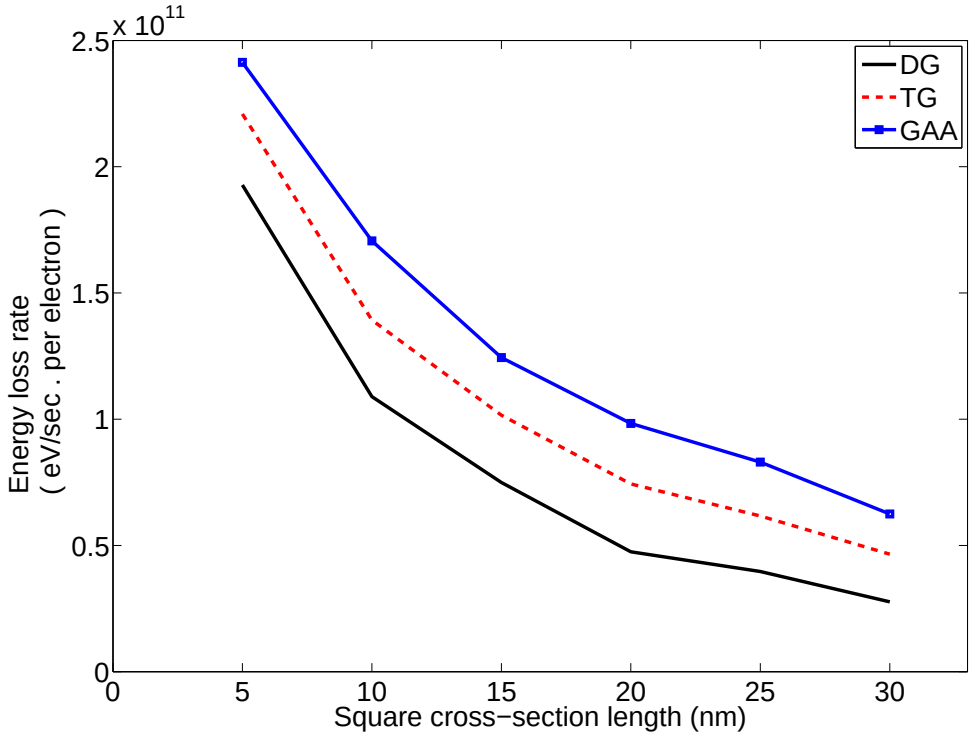


Figure 4.11: Variation of phonon energy dissipation as a function of square cross-section length for three multigate MOSFETs with $L_g = 20$ nm.

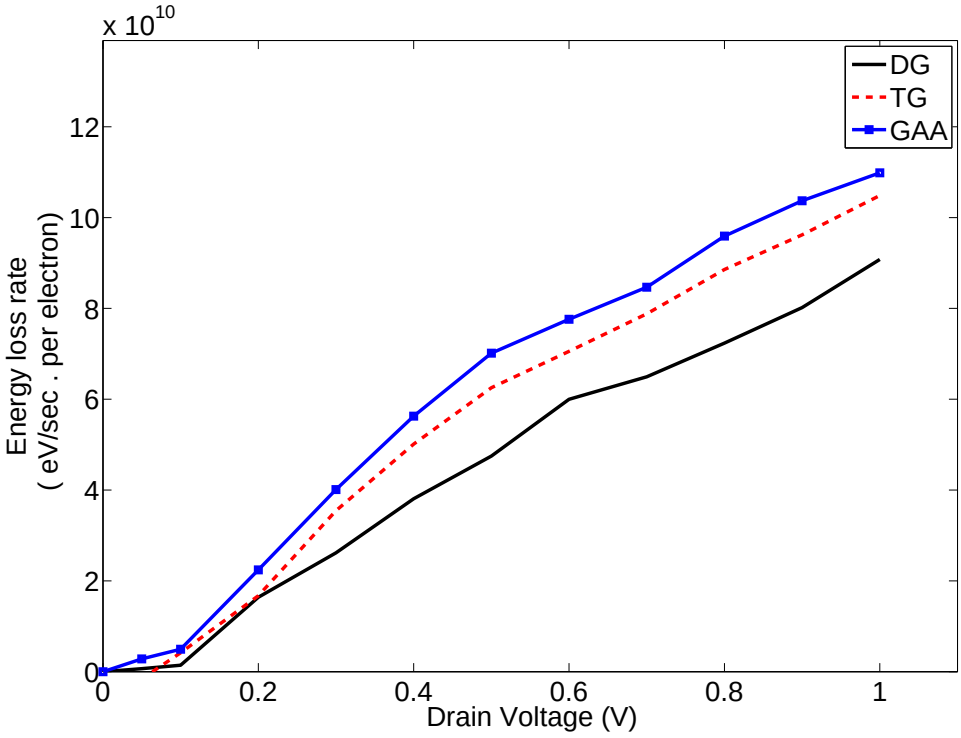


Figure 4.12: Variation of phonon energy dissipation as a function of drain voltage for three multigate transistors with $L_g = 20$ nm.

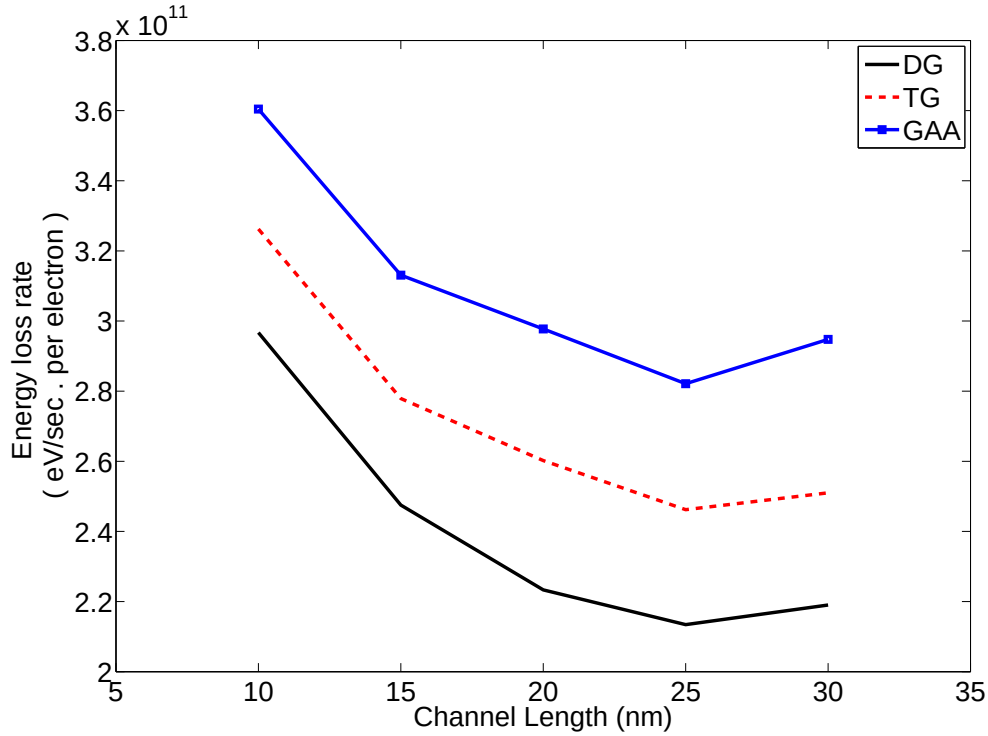


Figure 4.13: Phonon energy dissipation as a function of channel length for three different devices at an applied bias of $V_D = V_G = 0.5$ V.

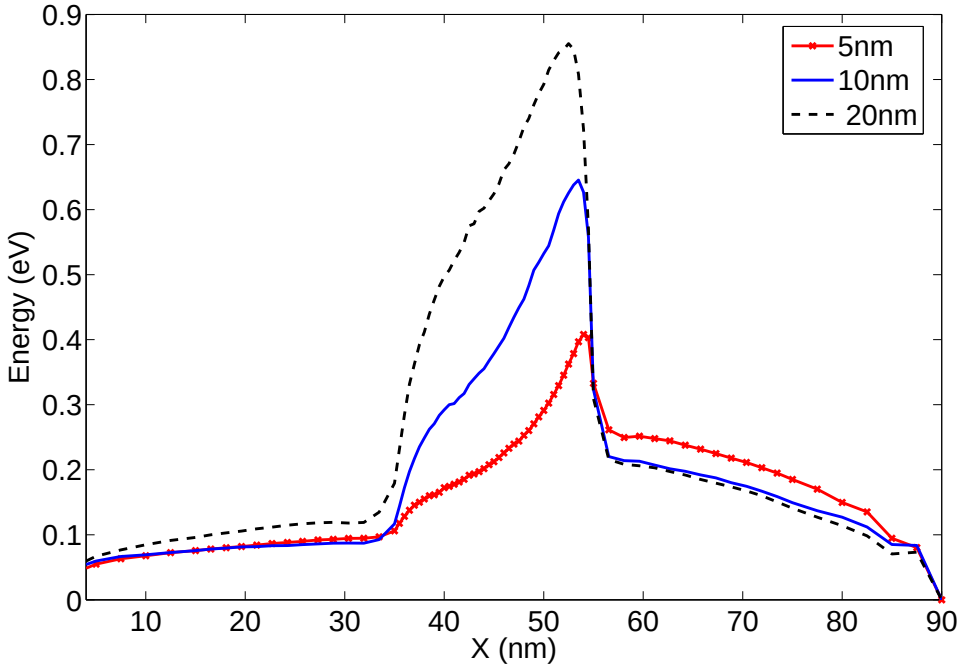


Figure 4.14: Spatial variation of the average energy along the channel for a TG MOSFET with $L_g = 20$ nm and $V_d = V_g = 0.5$ V for three cross-sections.

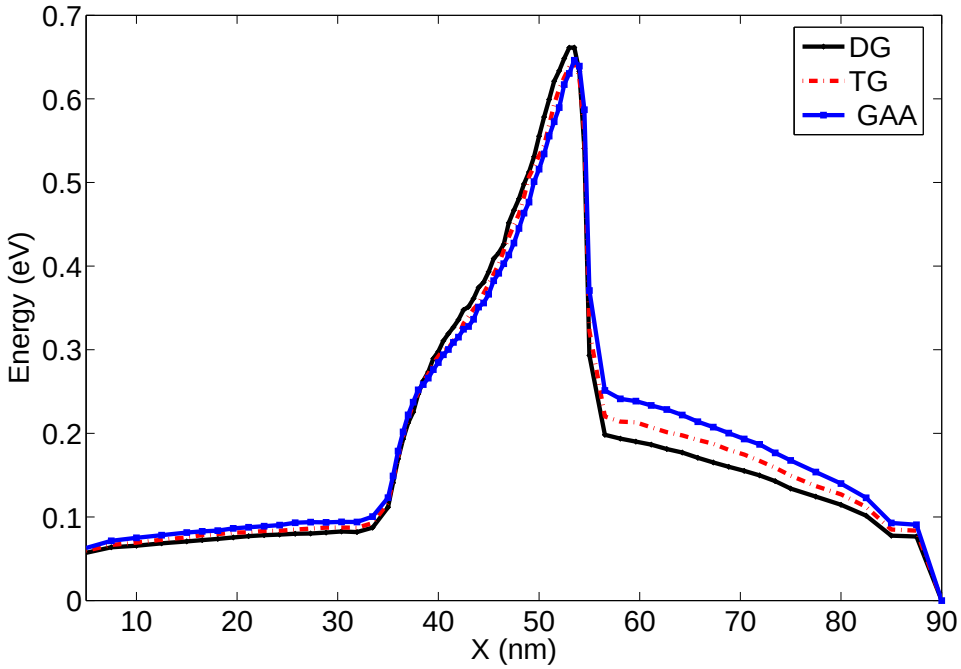


Figure 4.15: Comparison of the spatial variation of average energy along the channel for DG, TG, and GAA MOSFET with $L_g = 20$ nm, $T_{Si} = H_{Si} = 10$ nm, and $V_d = V_g = 0.5$ V.

CHAPTER 5

LATTICE DYNAMIC AND TRANSPORT

5.1 Overview

This chapter introduces the principles of phonons and the various approaches used to model phonon transport. Phonons are the dominant carriers of heat energy in most semiconductors and dielectric materials. Understanding phonon physics and transport is essential to predicting thermal transport coefficients and in elucidating the details of heat flow in nanoscale devices. We will apply some of the theories presented here to compute transport coefficients of silicon nanowire structures with various cross-sections. The foundations developed in this chapter will be applied to chapter 6, where we will couple electron and phonon transport.

5.2 Phonons as Heat Particles

It is well established that atoms in a crystal oscillate about their equilibrium lattice position. These collective oscillations induce a propagating and energy carrying vibrational wave. It is possible to resolve these vibrational waves as linear combination of normal modes (i.e. collective oscillations with well-defined frequency and wavelength). However, unlike classical normal modes which may have any amplitude and energy, the allowed frequencies of oscillation are quantized similarly to those of an electromagnetic field in a vacuum. The constraints set by the quantum mechanical equation quantize the normal mode energies. The quantization makes the lattice excitation particle-like (each

particle carries a discrete amount of energy). This picture applies to both photons and phonons. A quantum of lattice vibration has an energy which is a multiple of $\hbar(\omega + \frac{1}{2})$ and is called phonon. Simply put, phonons are the quantized normal modes of a particle oscillator having many modes. If many quanta are excited, the mode will look classical. Many physical properties can be accounted for by classical treatment if the effect of quantization is small (usually the case at high temperatures). The particle nature of phonons is justified as long as the energy transfer with electrons (or matter) involves the exchange of a single quanta.

5.3 Lattice Dynamics

The static model of crystals has been widely utilized to study many material-related properties including hardness, chemical properties, and electronic structures and properties [52]. However, one must also take into account that the motion of ions is essential to defining temperature and describing thermal-related coefficients and phenomena in solids such as heat capacity, thermal conductivity, and expansion. The goal here is to understand how the vibrational normal modes (phonon dispersion) are computed under the harmonic approximation. The calculation of phonon dispersion is fundamental to understanding thermal properties and transport coefficients. In the spirit of Born-Oppenheimer, the total Hamiltonian of the ions is given by:

$$H = \sum_i \left(\frac{P_i^2}{2M_i} \right) + V\left(r_1, r_2, ..., r_N\right) \tag{5.1}$$

where the summation is over all ions and $V\left(r_1, r_2, ..., r_N\right)$ represent the ion potential energy.

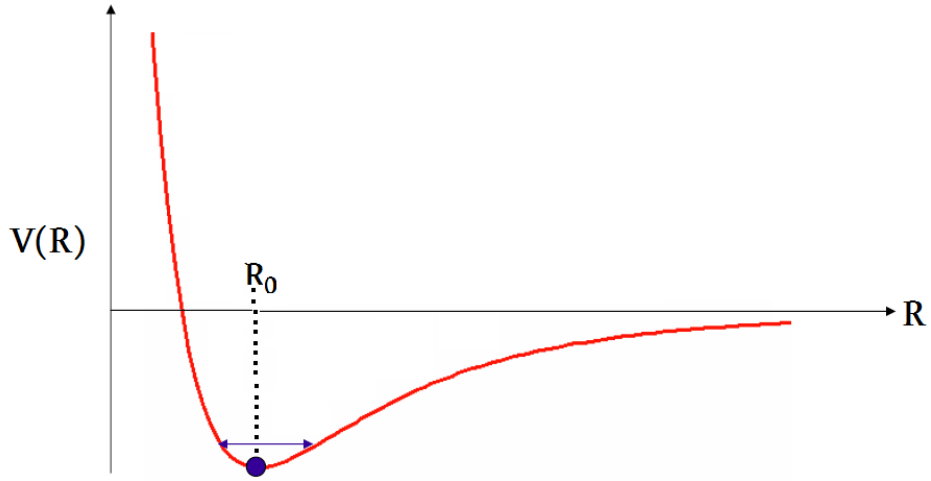


Figure 5.1: The interatomic potential as a function of the spacing between atoms

5.3.1 Harmonic approximation

Under the harmonic approximation, the atoms are slightly displaced from their equilibrium position. The total potential energy of the crystal can be described by the interatomic potential between a pair of atoms as shown in Figure 5.1. Under the harmonic approximation, the variations of the atoms from the equilibrium is small and, therefore neglecting cubic and higher-order terms, we can expand the potential variation from the equilibrium lattice point as:

$$\begin{aligned} V(R_1, R_2, ..., R_N) &= V\left(R_{1_0}, R_{2_0}, ..., R_{N_0}\right) + \sum_i \left(\frac{1}{2}\right) \left(\frac{\partial V}{\partial R_i}\right)_0 \delta r_i \\ &+ \sum_{i,j} \left(\frac{\partial^2 V}{\partial R_i \partial R_j}\right)_0 \delta r_i \delta r_j + ... \end{aligned} \tag{5.2}$$

where $R_{i,o}$ represents the equilibrium position of the i^{th} atom while the deviation of the i^{th} atom from its equilibrium position is represented by $\delta r = R_i - R_o$. The subscripts i and j indicate the three Cartesian coordinates. The second term of the Equation (5.2) is zero and we can rewrite the equation as:

$$V(R_1, R_2, ..., R_N) = V\left(R_{1_0}, R_{2_0}, ..., R_{N_0}\right) + \sum_{i,j} \gamma_{i,j} \delta r_i \delta r_j \tag{5.3}$$

with $\gamma_{i,j}$ denoting the force constant:

$$\gamma_{i,j} = \left(\frac{\partial^2 V}{\partial R_i \partial R_j} \right)_0 \tag{5.4}$$

This allows us to represent the lattice as a system of n-coupled quantum harmonic oscillators where each pair of atoms are connected by a spring. The main difficulty is in the form of the force constant. With the force constant being the unknown parameter, we can rewrite the overall Hamiltonian and the equation of motion as:

$$H = \sum_i \left(\frac{P_i^2}{2M_i} \right) + \sum_{i,j} \gamma_{i,j} \delta r_i \delta r_j \tag{5.5}$$

$$\frac{\partial p}{\partial t} = -\frac{\partial H}{\partial u_i} = -\sum_j \gamma_{i,j} \delta r_j \tag{5.6}$$

$$\frac{\partial \delta r}{\partial t} = -\frac{\partial H}{\partial P_i} \tag{5.7}$$

Newton's equation can be obtained by combining Equations (5.6) and (5.7):

$$M \frac{\partial^2 r}{\partial t^2} = -\frac{\partial H}{\partial u_i} = -\sum_j \gamma_{i,j} \delta r_j \tag{5.8}$$

5.3.2 Phonon bandstructure calculation

A key challenge in computing the vibrational modes of a solid is how accurately to describe the atomic potentials. Once the potential is known, the forces on the atoms can be easily computed. Here we focus on surveying (semi)empirical models used for computing semiconductor phonon dispersion. Generally, the phonon frequencies can be computed using either phenomenological models or computationally intensive models based on first principles. Phenomenological models describe the problem of lattice dynamics in terms of charges and force constants, which are then used to construct the relevant dynamical matrix. We

can classify the phenomenological models into three groups: models of interatomic forces, rigid-ion model and dipolar model [53]. The parameters of these phenomenological models are determined empirically. A brief review of these models follows.

5.3.2.1 Force constant models

Many parametrized and phenomenological models for the force constant have been proposed. Within the force constant model, all force constants are treated as independent parameters. Examples include the Born-von K*á*rm*á*n model and the valence force field model. The Born-von K*á*rm*á*n model is the simplest and takes into account the nearest neighbor interaction. In this model, the atoms are assumed to be hard spheres connected by springs, and the spring constant can be determined empirically. This model has been applied by Hsieh [54] to calculate the phonon dispersion of silicon. The calculated curve failed to fit the experimental dispersion curve at short wavelengths since this model did not accurately capture long-range interatomic interaction. Herman [55] extended this to include the fifth-nearest neighbor interaction and was able to reproduce the phonon dispersion of germanium using 15 force constants. However, the inclusion of higher-order interactions demonstrates that the interaction potential became larger for distant neighbors. As such, the presence of long-range forces in a semiconductor restricts the Born-von K*á*rm*á*n applicability to most semiconductors. In the valence force field model, the interatomic potential is written in terms of bond lengths and angles. Among the various valence force field approaches, the model utilizing Keating's potential with two parameters (harmonic bond bending and bond stretching term) has become the most popular [56].

5.3.2.2 Rigid-ion model

The rigid-ion model (RIM) was initially used to study the vibrational frequency in ionic crystals and was latter applied to ionic III-V and II-VI semiconductors [53, 57]. In this model, the rigid-ions are connected to adjustable springs which represent the interatomic forces. The ions are presumed to move as non-deformable point charges, and the only allowed interactions are (i) short-range forces and (ii) long-range interactions due to the Coulomb forces between the ions. For the Coulomb interaction, the charges in this model are assumed to be unpolarized point charges. Despite its simplistic picture, the model depends on a substantial number of parameters in order to attain reasonably good results. In addition, RIM-based models fail to describe the electronic polarization induced by a charge on the surrounding lattice, which can change the energy spectrum of the crystal.

5.3.2.3 Dipole model

Dipolar methods which take into account the effect of polarization on the atoms include the shell and bond charge models. Shell models of varying complexity have been proposed [58, 59] which separate the core ion and the valence electrons as separate shells. These models are successful because they take into account electronic polarizability and can be used for a wider selection of materials, particularly ionic and covalent crystals. The basic assumptions of the shell and the rigid-ion models are the same except the point charges of the ions are replaced by a polarizable ion in the shell model. The idea behind the shell models is shown in Figure 5.2. Each atom is described in terms of a positively charged ion core and a negatively charged shell consisting of polarizable electrons which are bound by spring-like harmonic forces. These forces model the dipole polarization of the atom by the force constants. As shown in figure 5.2, six types of interactions are captured in this model: two are due to the coupling between each core and the neighboring shell, another two due are to

the coupling of a particular shell with its own core, and one each to the coupling between the neighboring shell and between the neighboring cores. These interactions capture both short-range and long-range Coulomb forces. The dynamical matrix of the shell model is the same as that of the rigid-ion model plus the polarization term arising from the induced dipolar forces.

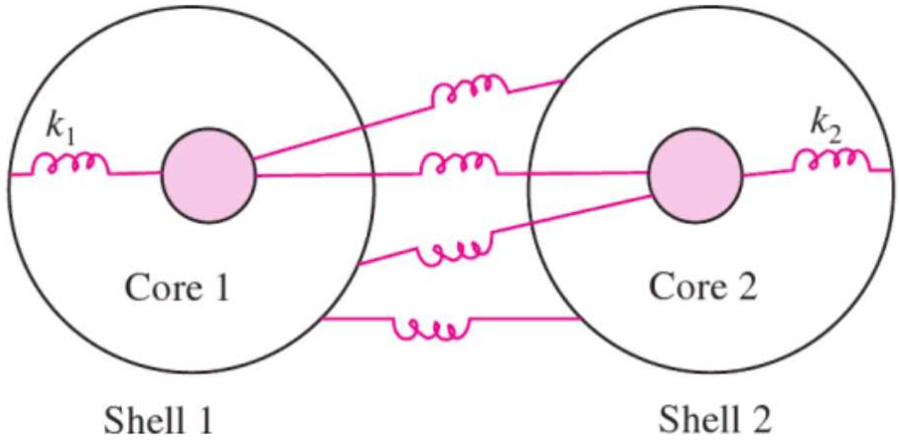


Figure 5.2: The interactions captured by the shell models [60]

A variation of the shell model is the bond charge model which was first proposed by Phillips [61]. In this model, the induced dipole can be viewed as being due to the displacement of the electronic charges situated on the bond (bond-charge) Martin [62] modified the Phillips model and assumed a free-electron-like screening potential which he computed using first principle pseudopotential calculations. Weber [63] later modified Martin's approach and allowed the bond-charges to be particle-like and move adiabatically from the midpoints of the bond. In addition, he treated the coupling forces as adjustable parameters. Weber's bond charge model, widely known as the adiabatic bond charge model, has been one of the most popular semi-empirical methods used to compute phonon dispersion and is applied to a wide range of materials including elemental group IV semiconductors, III-V zinc-blende structures, II-VI semiconductors, wurtzite and superlattices [64, 65, 66]. The bond charge model takes into account the bond forces due to the electronic charge between nearest

neighbors in a covalently bonded crystal. The accumulated charge distribution (i.e. bond charge) is taken into account by placing a negative point charge between each pair of nearest neighbors. A schematic of the bond charge model is shown in Figure 5.3. The potential between the nearest-neighbor ions is given by φ_{i-i} while φ_{i-b} represents the potential between the bond-charge and the ion. $-Z$ represents the ion charge while V_K and $2Z$ are the bond potential and charge. The forces captured in this model include: (i) long-range Coulombic interaction between the ions and the bond charges, (ii) short-range central potential ion-ion interaction between nearest neighbors, (iii) short-range ion-bond charge central interaction, and (iv) short-range noncentral Keating-like bond charge to bond charge interaction. A result for the phonon dispersion curve using the adiabatic bond charge model is shown in Figure 5.4. There is a good agreement between the theory and the experimental results (represented by dots in the graph).

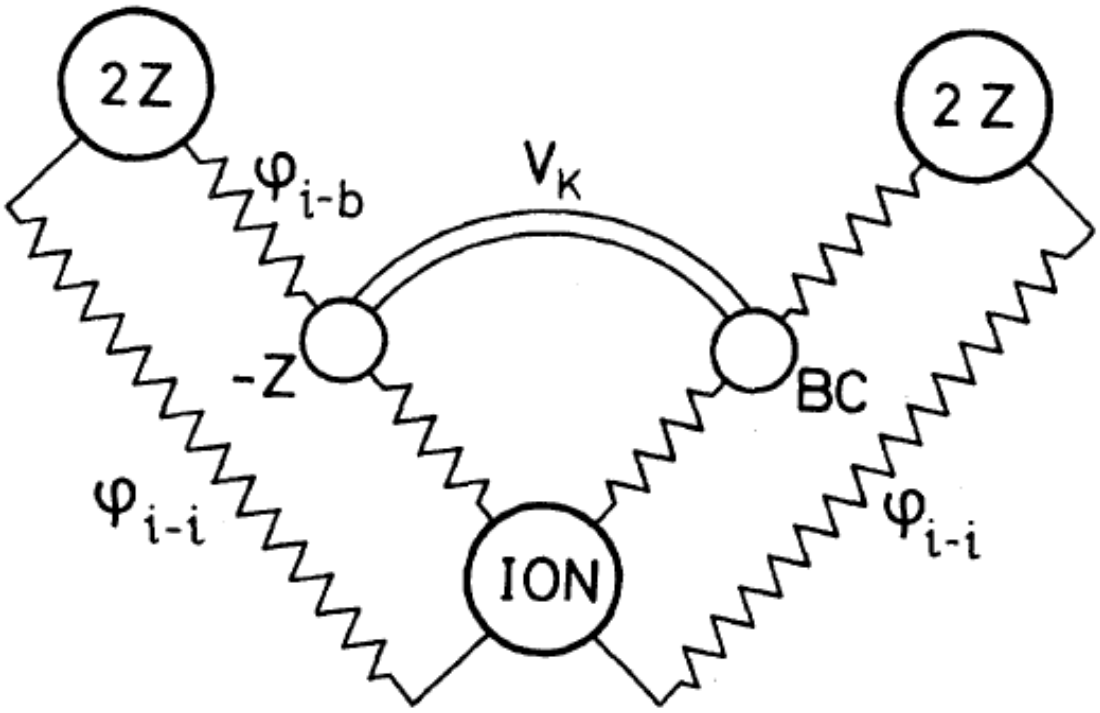


Figure 5.3: Schematic of Weber's adiabatic bond charge model [64].

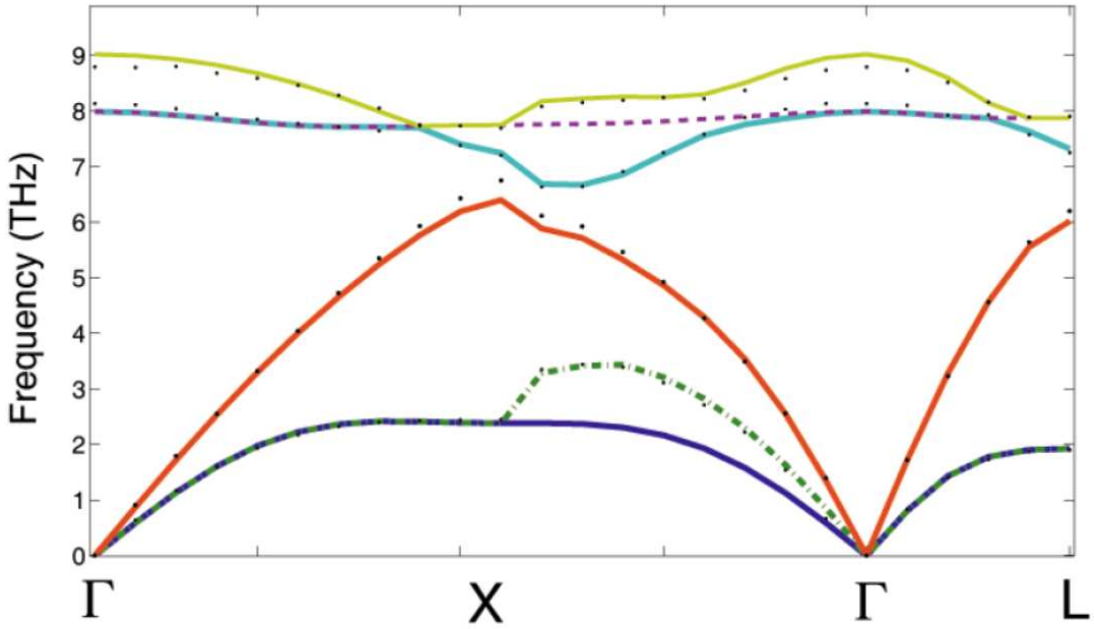


Figure 5.4: Phonon Dispersion of GaAs computed using adiabatic bond charge model.

5.4 Phonon Transport

Generally, transport describe processes in which mass, momentum, angular momentum, and energy move about in matter. Thus it includes diffusional phenomena, fluid dynamics, and heat transport. Here we briefly review some of the commonly used scheme to model phonon transport.

5.4.1 Molecular dynamic

Molecular dynamic (MD) was first introduced by the seminal paper of Alder and Wainraight in which they studied interaction of hard spheres [67]. Since then, this approach has emerged as a popular modeling technique to characterize materials, molecules, and biological properties. MD simulation describe the interaction of all the atoms or ions in a system using Newton's law of motion for

a given interatomic potential. A typical MD simulation numerically solves Newton's equations of motions periodically for each atom and computes their corresponding trajectory and velocity as a function of time:

$$\frac{dp_j}{dt} = F_j \tag{5.9}$$

$$\frac{dr_j}{dt} = \frac{p_j}{m_j} \tag{5.10}$$

where r_j and p_j describe the position and momentum of the j^{th} atom. The force term (F_j) captures the interatomic forces between the atom and is related to the total potential energy of the system:

$$F_j = -\nabla_{r_i} V(r_1, r_2, ..., r_N) \tag{5.11}$$

The interaction potential is a key input to the simulation, which allows us to solve the forces on the atoms (or ions) and takes into account all possible interactions (e.g. electron-ion interaction, electron-electron interaction, etc.) except for the kinetic energy of the atom (or ion), which we solve directly using the particle velocities. As such, a key approximation in a typical MD simulation lies in the choice of the interatomic potential, which is obtained either empirically or from ab initio methods such as the density functional theory and quantum Monte Carlo. The use of ab initio methods to compute the interatomic potentials introduces computational burden into our simulation and hence limits the problems we can compute to a few ions. In the end, the essence of MD simulation is to use the computed dynamic variables and the interatomic potential to determine the macroscopic properties of a system, such as pressure and temperature, using the formalism averages and fluctuations obtained from statistical mechanics. For example, the thermal conductivity which is a key parameter in heat transfer problems can be derived from the MD simulation by using the fluctuation dissipation theorem.

5.4.2 Boltzmann transport equation

The Boltzmann transport equation (BTE) can be applied to study phonon transport. The wave nature of the phonons can be neglected when the length scale of the system is much larger than the wavelength of the phonon. The phonon BTE is written as:

$$\frac{\partial f}{\partial t} + \frac{1}{\hbar}\nabla_k E(k).\nabla_r f + \frac{qF(r)}{\hbar}.\nabla_k f = [\frac{\partial f}{\partial t}]_{collision} \tag{5.12}$$

where f is the phonon distribution and $E(k)$ is generally obtained from the full phonon dispersion. The right-hand side as usual represents the scattering term. Some of the relaxation time approximation methods used to solve the phonon BTE problem are summarized below. We will demonstrate the Monte Carlo solution to the phonon BTE in the next chapter.

5.4.2.1 Gray and semi-Gray BTE

In the gray BTE method, the phonons are grouped together in one mode and are characterized by a single group velocity and relaxation time. No distinction is made between different phonon modes [68, 69].Another common approach used to simplify the BTE is using semi-gray BTE which in essence categorizes phonons into two modes: propagating (i.e. carrying energy) or stationary. Both phonon groups are allowed to exchange energy [68, 69].

5.4.2.2 Phonon radiative approach equation

Majumdar proposed a phonon radiative transport equation for phonons employing the relaxation time approximation and demonstrated that phonon transport at the ballistic limit could be studied using radiative phonon transfer [70]. The radiative phonon transfer equation maybe derived from the Boltzmann transport equation by weighting the BTE by $v \cdot \hbar\omega \cdot D$. This term is related to the phonon intensity (I_k) as:

$$I_k\left(r,k,s,t\right)=v\left(k,s\right)\cdot f\left(r,k,s,t\right)\cdot\hbar\omega\left(k,s\right) \tag{5.13}$$

The resulting phonon radiative heat transfer equation can be written as:

$$\frac{\partial I_k}{\partial t}+v\cdot\nabla_k I_k=\left(\frac{\partial I_k}{\partial t}\right)_{collision} \tag{5.14}$$

5.4.3 Fourier law and heat conduction

The most basic model for heat flow was introduced by Fourier, who introduced a rate equation which determines heat flux in a medium based on the temperature distribution:

$$q^{''}=-\kappa\nabla T \tag{5.15}$$

The Fourier equation like other parabolic equations provides an infinite propagation speed of heat. Realistically, thermal energy propagation is dependent on the finite velocity of the associated phonon mode. In addition the diffusive nature of heat transfer is questionable when the phonon mean free path is comparable or larger than the size of the device structure. One way to overcome the limitation of Fourier's law is to use the hyperbolic wave equation, which demonstrates finite propagation speed. This equation can be derived from the phonon BTE under the absence of external force in a manner similar to the one we used to derive the hydrodynamic equation from the electron BTE. In this case, the phonon BTE is weighted by the function $v\cdot\hbar\omega\cdot D(\omega)$ and integrated over all frequency ω:

$$\int d\omega\left(v\cdot w\cdot D(\omega)\right)\frac{\partial f}{\partial t}+\int d\omega\left(v\cdot w\cdot D(\omega)\right)v.\nabla_r f=\int d\omega\left(v\cdot w\cdot D(\omega)\right)[\frac{\partial f}{\partial t}]_{collision} \tag{5.16}$$

$$q = \int d\omega v(\omega)\hbar\omega f(r, w, t)D(\omega) \tag{5.17}$$

where q is the heat flux, $D(\omega)$ is the density of states, v is the velocity. Utilizing an energy independent relaxation time approximation and expanding $\frac{\partial f^0}{\partial x} = \frac{\partial f^0}{\partial T}\frac{\partial T}{\partial x}$ we obtain the Cattaneo equation [71, 72]:

$$\tau\frac{\partial q}{\partial t} + q = -\kappa\frac{\partial T}{\partial x} \tag{5.18}$$

where τ is the average relaxation time and κ is the thermal conductivity. Cattaneo's equation demonstrates that heat propagates with finite speed and reduces to the Fourier equation when the relaxation time τ is negligibly small. The hyperbolic equation can be easily obtained by coupling the Cattaneo equation with the energy conservation equation:

$$\frac{\partial T}{\partial t} = -C\frac{\partial q}{\partial X} \tag{5.19}$$

The specific heat term is represented by C. Using the above two equations the hyperbolic wave equation can be written as:

$$\frac{\partial^2 T}{\partial t^2} + \frac{\partial T}{\partial t} = -\frac{k}{C}\frac{\partial^2 T}{\partial x^2} \tag{5.20}$$

The above equation demonstrates a wave propagating at a speed $\sqrt{\frac{\kappa}{C\tau}}$. This model is non-local in time but local in space since the temperature represents a spatially localized thermodynamic equilibrium [72]

CHAPTER 6

COUPLED ELECTRO-THERMAL MONTE CARLO ANALYSIS OF SELF-HEATING IN NANOWIRES

6.1 Overview

Thermal management and power dissipation in microprocessors are quickly emerging as the ultimate bottleneck to improving the performance of consumer/commercial electronics. Controlling device temperature and power dissipation is crucial to sustaining electronic devices with longer battery life and improving the overall device reliability and life expectancy. As such, thermally conscious blueprints have become recently the focal concern of semiconductor roadmaps, at all design phases, as the temperature at the chip level and within a single transistor rises for future electronic devices [73]. Without proper thermal management, inordinate power dissipation can potentially halt integrated circuit functionality.

This chapter explores the electro-thermal behavior of silicon-on-insulator (SOI) multigate (MG) nanowires. Our approach is unique in that we couple our in-house 3D quantum-corrected Monte Carlo with a phonon model explicitly taking into account the anharmonic decay. The anharmonic three-phonon decay and the use of full dispersion facilitate a detailed description of heat transfer and the determination of the temperature map in nanoscale devices.

6.2 3D Electro-Thermal Simulator

The hierarchy of electron and phonon simulation models is shown in Figure 6.1. The most detailed level of simulation is to solve the wave (Schrödinger)

equations when device dimensions are comparable to the particle (electron/hole) wavelength. However, although the electron community has utilized Schrödinger based solutions to device modeling, the phonon community is still lagging behind. For heat transfer, the most detailed level of simulation is the direct solution of the lattice dynamical equations or the equivalent treatment of the solid as an elastic continuum. If phase/coherence effects can be neglected, as they can in many applications with diffuse interfaces and boundaries, it is possible to treat phonon transport under a particle framework. Under this framework, Monte Carlo simulation offers the most comprehensive simulation tool.

Here we describe the thermal transport model used in this work and the coupling between the electron (described in the previous chapters) and the phonon model.

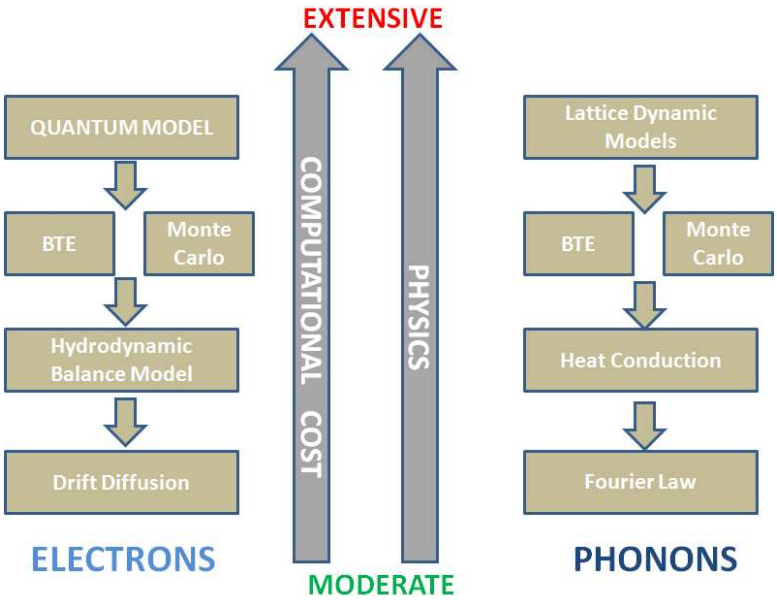


Figure 6.1: Hierarchy of electron and phonon transport models.

6.2.1 Thermal transport

Thermal energy is largely transported in semiconductors by the collective vibrational motion of ions. The collective vibration waves of a periodic solid are made of quantized units of vibrational energy, referred to as phonons. In order to capture the details of self-heating in nanoscale devices, we need to carefully model phonon generation and transport. At length scales between the phonon mean free path ($\Lambda_{si,bulk}$ 300 nm at room temperature) and phonon wavelength (1 nm at room temperature) continuum heat transfer models (e.g. Fourier law) are inadequate. At this limit, the Boltzmann transport equation provides the most appropriate means to study phonon transport. The main difference between the electron and the (uncharged) phonon BTE is the irrelevance of Lorentz force to phonons :

$$\frac{\partial f_{_{ph}}}{\partial t} + \upsilon(q)\frac{\partial f_{_{ph}}}{\partial r} = \frac{\partial f_{_{ph}}}{\partial t}\mid_{_{el\text{-}ph}} - \frac{\partial f_{_{ph}}}{\partial t}\mid_{_{anharmonic}} \tag{6.1}$$

The right-hand side identifies two terms which cause a variation in the phonon distribution function: the first term represents the electron phonon scattering events and the second term comprises anharmonic phonon decay processes which are treated as a scattering mechanism.

6.2.1.1 Phonon Monte Carlo

The application of MC to particle transport is relatively simple to understand. Assuming the carrier motion consists of free flights interrupted by scattering mechanisms, a computer is utilized to simulate the trajectories of particles as they move across the device. The scattering events and the duration of particle flight are determined using random numbers. The general flow chart for the phonon Monte Carlo employed here is shown in Figure 6.2. A fundamental step in the phonon MC used in this work is that it takes as an input the detail of phonon generation (i.e. iteration number in which the scattering event occurred,

phonon branch/position/momentum, etc.) computed from the electron MC. The first step to simulate heat transport consists of using the data obtained from the electrical Monte Carlo to calculate the average lifetime (τ_{anh}) of the emitted phonon due to anharmonic decay for each phonon energy and branch [9]. The average lifetime is then used to stochastically compute phonon decay time. This is usually achieved utilizing

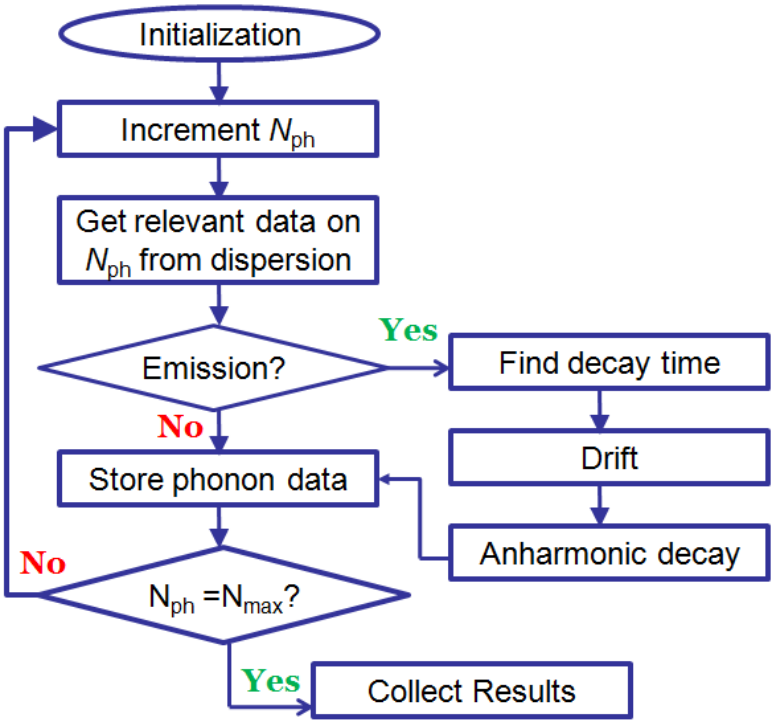


Figure 6.2: The flow chart of the phonon Monte Carlo used in this work

$$\tau_{decay} = ln(r)\tau_{anh} \tag{6.2}$$

Here r is a random number uniformly distributed between 0 and 1. At this point, the emitted phonon is allowed to drift from its original position r_{ems} up to the position r_{decay} in which it will decay into two other phonon modes:

$$r_{decay} = r_{ems} + \upsilon dt \tag{6.3}$$

where v is velocity of the emitted phonon extracted from the phonon dispersion. The next step is to determine a pair of final states for the emitted phonon. This step is accomplished using the rejection algorithm on the probability distribution for anharmonic decay including the matrix element [74]. The probability distribution for anharmonic decay was obtained by Klemens using time-dependent perturbation theory as the product of the anharmonic matrix element for the three phonon processes and a time-dependent resonance factor:

$$P(q, q') = |\langle q|H'|q'\rangle|^2 \frac{1 - \cos(\triangle w t)}{\triangle^2 t} \tag{6.4}$$

where $\triangle w = w - w' - w''$ is the net energy exchange between the initial phonon momentum (q) and the final states (q', q'') and H' is the perturbing Hamiltonian due to cubic anharmonicity. The rejection algorithm starts by estimating a possible pair of final states generated by the anharmonic decay processes; q' is chosen randomly from a uniform distribution in the first Brillouin zone, while the second final state (q'') is determined by ensuring momentum conservation:

$$q'' = q - q' \pm G \tag{6.5}$$

To determine whether the pair of of final states is accepted, a random number (r_{rej}) uniformly distributed in the unit interval is generated and compared to the probability $P(q, q')$. The pair of final state is rejected if $P(q, q') < r_{rej}$, and a new pair is searched. Finally, once the final state is determined, the phonons are transported until the end of simulation without undergoing any further anharmonic decay.

6.2.2 Coupling electron and thermal transport

In order to capture the details of self-heating in nanoscale devices, we need to carefully model phonon generation and transport. In the transistor level, heating is established when immensely energetic electrons, exiting the highly resistive

channel, relax by giving off their excess energy to the crystal through electron-phonon interactions. Particularly, the effectual energy relaxation mechanism for energetic electrons is to couple with high frequency optical phonon [75, 10]. Optical phonons, however, have insufficient group velocity imperative for efficient heat transport and hence this process promotes the confinement of thermally nonequilibrium distribution of hot optical phonon at the vicinity of the source-drain region. The excess phonon population eventually diffuse diffuse out as the optical modes slowly decay into acoustic phonon mode with higher group velocity. As such, the coupling between electrons and optical phonons, and the subsequent decay of optical phonons towards equilibrium, are found to play a large role in determining the temperature distribution in silicon devices. Figure 6.3 and 6.4 summarize the electro-thermal coupling process. As shown in the flow diagram in Figure 6.3(a) the self consistent coupling entails feeding the heat generation obtained from our 3D quantum-corrected Monte Carlo to a thermal Monte Carlo which dissipates the anharmonic phonon processes. The anharmonic three-phonon decay and the use of full dispersion (Figure 6.4(a) facilitate a detailed description of heat transfer and the determination of temperature map in nanoscale devices. The temperature map obtained from the heat transport model is fed back to the electron Monte Carlo, which uses temperature-dependent scattering table (Figure 6.4(b) in a self-consistent manner until the convergence criterion is met.

6.2.3 Temperature distribution

One approach to obtaining the temperature map uses the data from the phonon simulation and performing direct conversion between the lattice energy and temperature. The method entails careful bookkeeping of phonon emission and absorption along each cubic mesh simulated. For each cube, the energies of the phonon falling into it are either summed (for emission) or subtracted (for absorption) from the initial energy. Phonons which leave the region of interest

during drift and anharmonic decay process are ignored. The updated energy map represents the final lattice thermal energy. At the end of this process, reverse temperature look-up is performed utilizing the energies of each cube. This method of computing the temperature map, although conceptually simple, places an unacceptable computational burden on the thermal simulation due to the great amount of phonon data involved. To surmount this, we employ a second approach that restricts our analysis only to the phonons generated at the end of the electrical simulation. This is well justified since we generally run the simulation long enough until steady state is met. As such, it is sufficient to consider only phonons generated ensuing steady state. With this approach, the energy of the phonon is not added to the energy map, but is considered by itself. Nonetheless, in order to make the results of this approach independent of the number of phonons processed, phonon contribution is normalized by the appropriate time window reflecting the iteration of the first phonon processed to

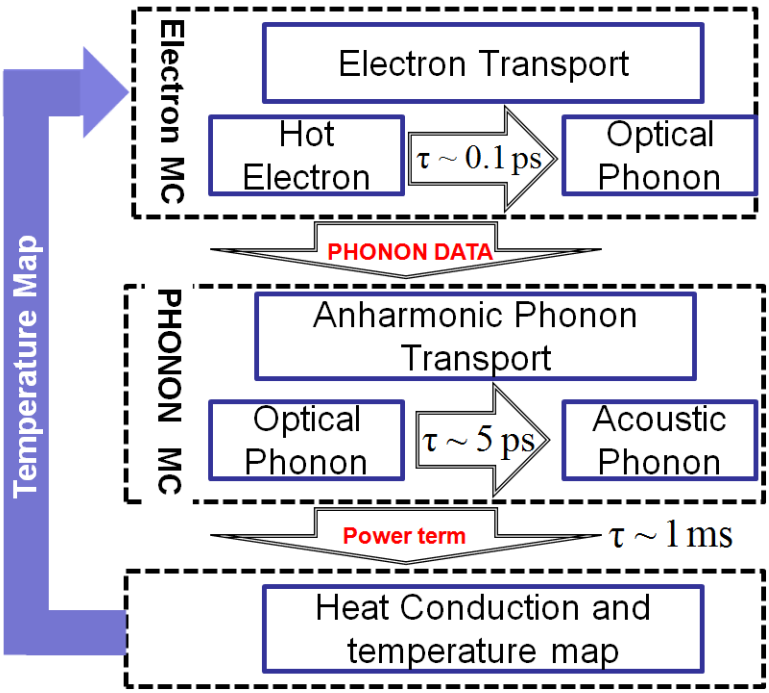
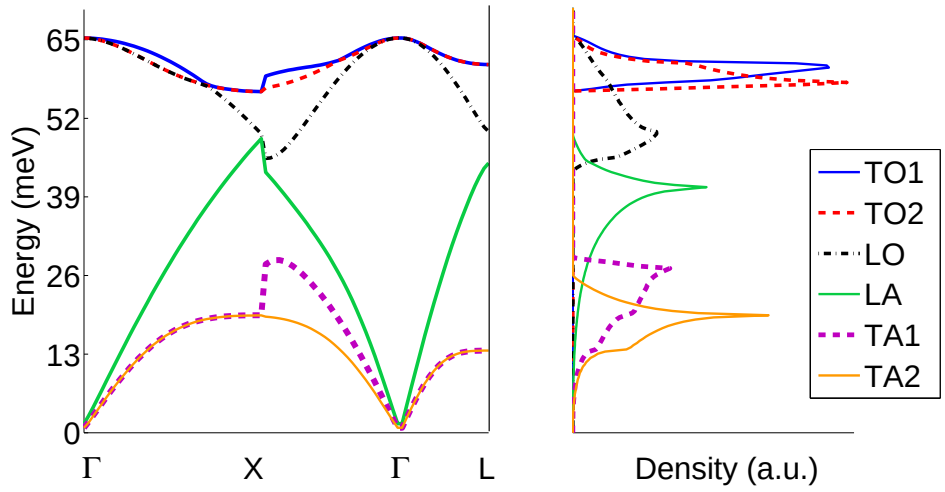
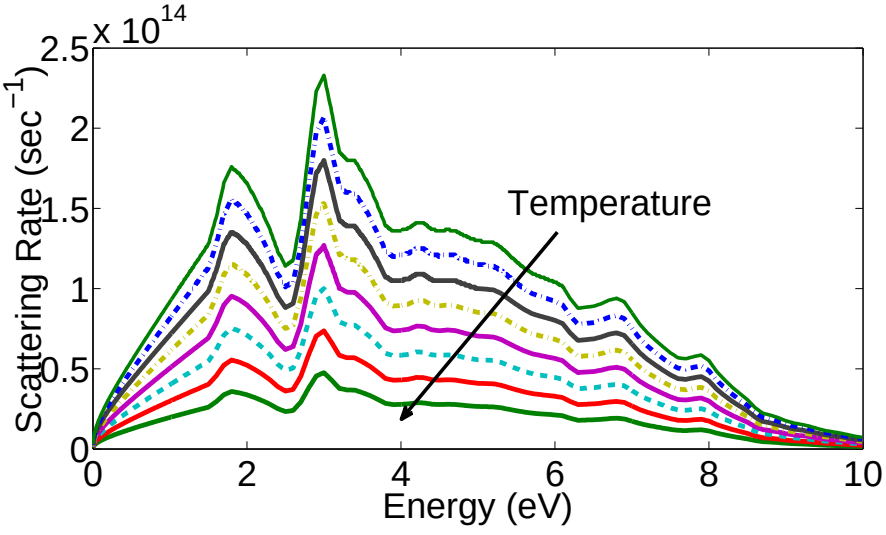


Figure 6.3: Flow chart demonstrating the approach we use in thermal MC to couple both the electrical and phonon transport models.



(a)



(b)

Figure 6.4: The coupled electro-thermal simulation uses full phonon dispersion data (a) and temperature-dependent scattering rates (b).

the final iteration.

6.3 Results and Discussion

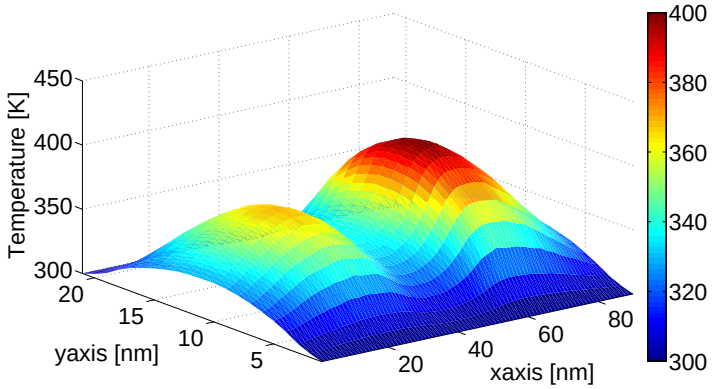
The performance gains and short-channel effects of multigate architectures have been discussed in Chapter 4. We have seen that these architectures lack true immunity from short-channel effects and a need to scale the cross-sectional area along with the channel length itself is required. Particularly, to suppress short-channel effects in these structures, we need to facilitate a behavior which nearly follows the ideal gradual channel approximation (GCA)—i.e. the electric field induced by the gate (V_G/L_{si}) must be much stronger then the electric along the channel ((V_D/Lg)) . Here V_D is the applied drain voltage, V_G is gate voltage, while L_G and L_{si} represent the channel length and the cross-sectional length (T_{si}=H_{si}), respectively. Because the applied drain and gate voltages are comparable, GCA-like behavior is realized if the cross-sectional length (L_{si}) is much smaller than the channel length (L_g). This design rule allows the gate to have better control of switching the transistor on or off, while the source and drain voltages serve only to transport carriers across the channel. Furthermore, with increased scaling of the cross-sectional length, the performance differential between different multigate architectures diminishes at the limit of small cross-sections. Therefore, practical issues such as ease and cost of fabrication become much more relevant in choosing which structure to use. In addition, the number of gates is not as strongly correlated with performance enhancements as previously expected.

In this section we explore multigate devices further, adding to the discussion the influence of the thermal effects on device performance by utilizing the coupled 3D electro-thermal Monte Carlo simulator (also referred to as thermal MoCa) introduced in this chapter and compare it to the 3D electrical Monte Carlo (which we refer to as MoCa or electrical MoCa here). For the purposes of this discussion, a GAA multigate device is considered with the following

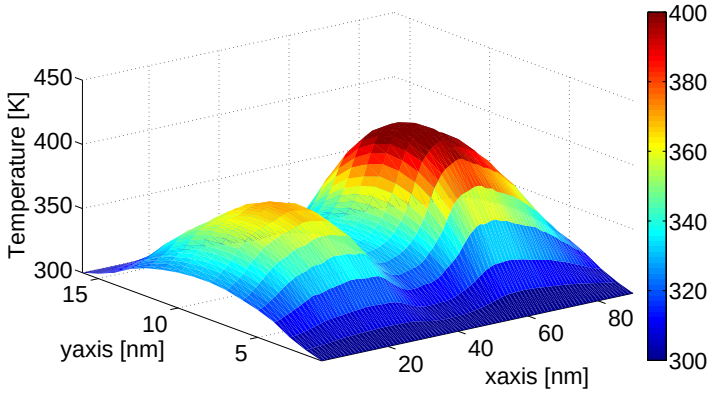
parameters: square cross-sections (T_{si}=H_{si}) varying from 5 to 35 nm, constant gate length of 20 nm, substrate doping at 10^{16} cm^{-3}, and fixed oxide thickness of 1.5 nm. The device geometry and doping is in accordance with those presented in Chapter 4. The carriers are confined along the y-z plane and the direction of current flow is along the x axis. The source and drain edge of the 20 nm channel is at $x = 35$ nm and $x = 55$ nm. Except in instances where we show how the cross-sectional scaling influences device behavior, the analysis presented here picks the smallest cross-section (T_{si}=H_{si}=5 nm) in order to ensure immunity to short-channel effects.

The temperature profile of the GAA structure using thermal MoCa in Figure 6.5 illustrates the effect of scaling the cross-sectional length on temperature distribution within the device. The three plots show that as the square cross-sectional length is reduced from 20 nm, 10 nm, and 5 nm, the temperature distribution of the device increases. This is expected as the reduced cross-sections trap more nonequilibrium phonons which are unable to escape from the channel due to the low conductivity of the surrounding oxide region. This effect is further illustrated in Figure 6.6 where the peak temperature consistently increases with decreasing cross-sectional area. This, again, is due to the increased thermal resistance within the small cross-sectional area. It should also be noted that the hotspot in Figure 6.5 is not localized to the drain region and extends into almost half the channel.

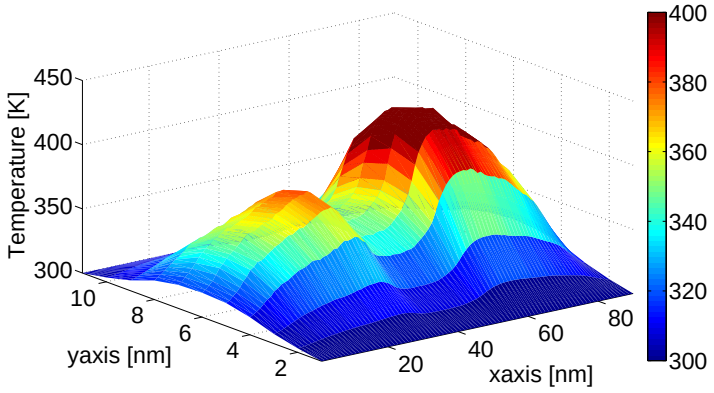
In Figure 6.7 we demonstrate how the overall temperature distribution relates directly to the drain voltage. This result is expected, since higher drain voltages increase the energy of the electrons and promote higher optical phonon emission rates. The corresponding peak temperature is shown in Figure 6.8 as the drain bias is ramped from 0.1 V to 1.2 V.



(a)



(b)



(c)

Figure 6.5: Temperature distribution of a GAA structure with square cross-section of (a) 20 nm, (b) 10 nm, and (c) 5 nm TG at an applied bias $V_G = V_D = 0.5$ V. The channel region range is from $x = 35$ to $x = 55$ nm.

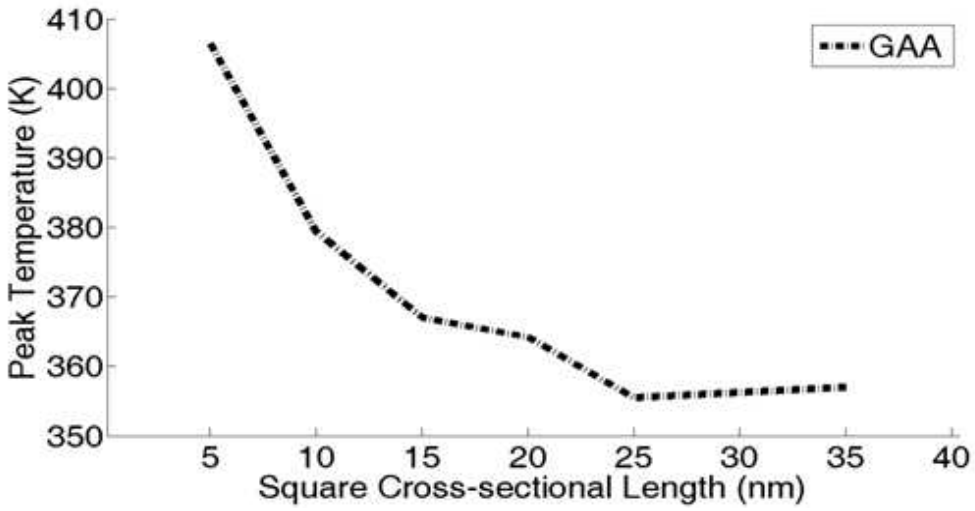
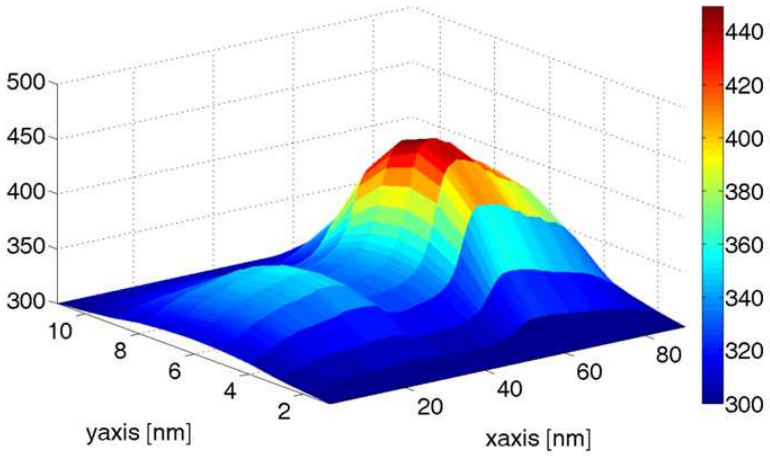
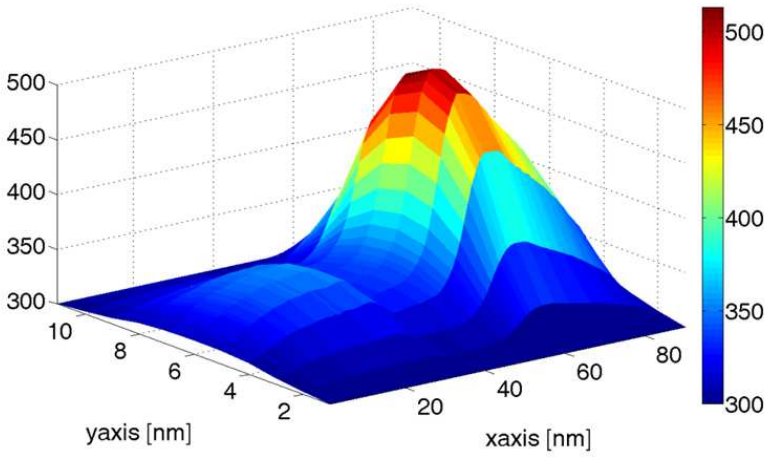


Figure 6.6: Variation of temperature as a function of the square cross-section bias for a GAA MOSFET with $L_g = 20$ nm, $T_{Si} = H_{Si} = 5$ nm, and $V_d = V_g = 0.5$ V.

Figure 6.9 shows a comparison between the drive current obtained using the thermal and the electrical MoCA simulators. The result clearly shows that thermal effects degrade the drive current due to the influence of temperature dependent scattering rate on carrier transport. In the electrical MoCa, the scattering rates are calculated at a fixed temperature (300 K), without consideration of the spatial variations of temperature within the channel. In thermal MoCa, however, the temperature distribution (obtained from the phonon transport) is used to determine the temperature dependent scattering rates, providing a much more accurate illustration of the device performance. In Figure 6.10, we further emphasize this result by showing the impact of cross-sectional area on drain current. The results shown in Figure 6.10 is for a GAA MOSFET with a cross-sectional length of 5 nm and an applied bias $V_D = V_G = 0.5$ V. Figure 6.10(a) seems to suggest that thermal effects are less important at small cross-sections since the difference between the total drive currents obtained MoCa and thermal MoCa is not significant. However, this is an incorrect assessment and is due to the fact that the plots are shown in linear scale and in the y axis and not logarithmic. A clearer picture is shown in Figure



(a)



(b)

Figure 6.7: Temperature map of a 5 nm x 5 nm GAA MOSFET for (a) $V_d = V_g = 0.7$ V and (b) $V_d = V_g = 1.0$ V.

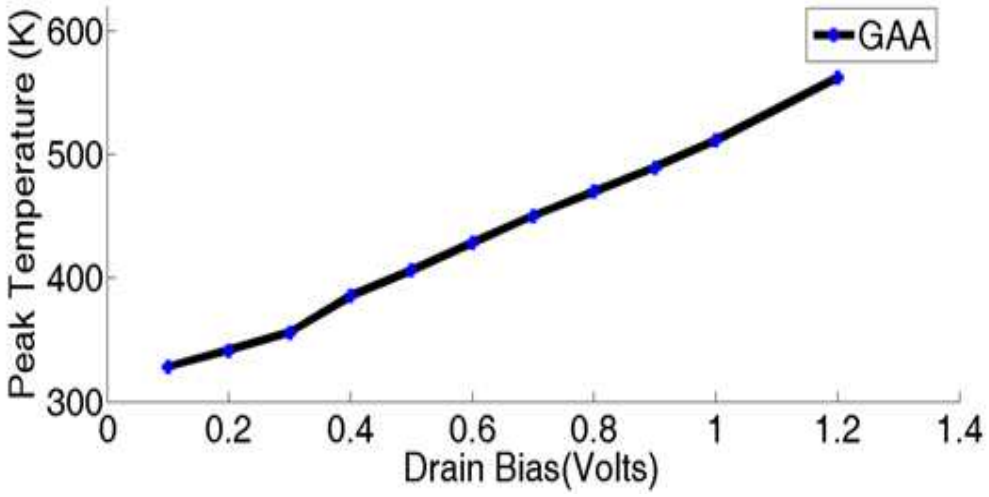


Figure 6.8: Variation of temperature as a function of the applied drain bias for a GAA MOSFET with $L_g = 20$ nm, $T_{Si} = H_{Si} = 5$ nm, and $V_d = V_g = 0.5$ V.

6.10(b) in which we compare the percentage of current enhancement in GAA relative to DG using both the electrical and electro-thermal Monte Carlo models. We find that there is indeed a significant difference between the current increase under the electro-thermal model relative to the electrical model even at the 5 nm x 5 nm cross-sections.

In order to demonstrate the cause of current reduction, we examine the difference between the electron concentration and average velocity obtained from MoCa and thermal MoCa. Figure 6.11(b) clearly shows that the electron concentration does not considerably change between the MoCa and thermal MoCa. Figure 6.11(a), however, shows that the average velocity is significantly changed between the two models. Again, this is because the electrical MoCa simulation is run under a fixed temperature. This is a gross simplification of what happens in real devices as a device would heat up during operation, which would result in increased scattering rates. In thermal MoCa, the influence of self-heating on the electrical performance is readily captured by first obtaining a spatially varying temperature distribution along the device from the phonon simulator and then utilizing the temperature dependent scattering rate in the electron transport. The higher scattering rate results in further reduction of

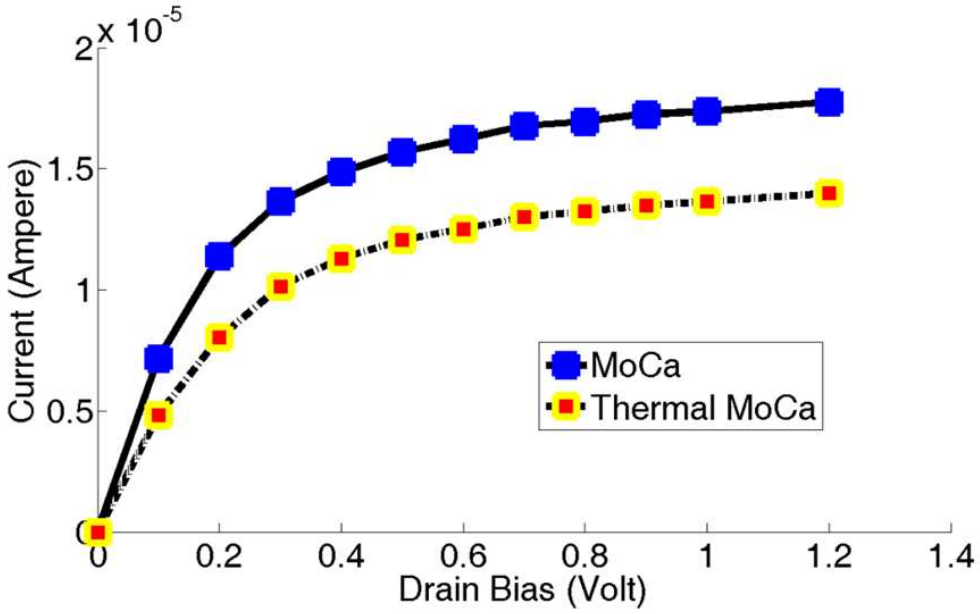
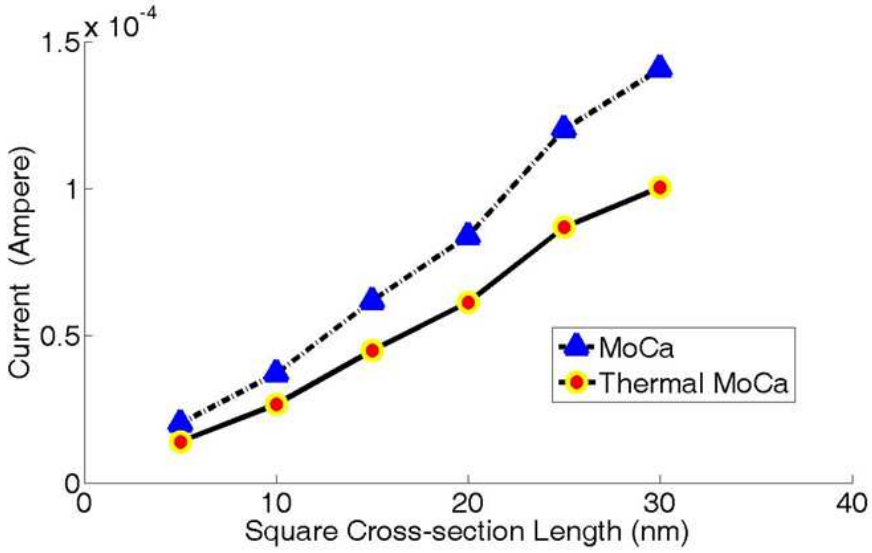


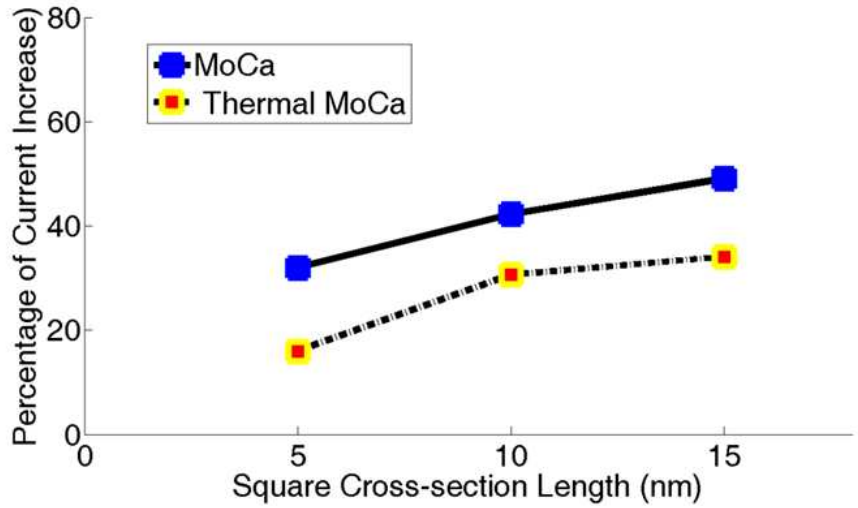
Figure 6.9: Comparison of GAA I_{DS}-V_{DS} characteristic obtained from a coupled electro-thermal MC simulation and 3D electrical MC simulation. The electro-thermal MoCa simulation shows current degradation as a consequence of heating.

average velocity of the carriers trying to move through a device hotter than the room temperature of 300 K. It also worth noting that there is little change in the carrier velocity between the source edge ($x = 35$ nm) and drain edge ($x = 55$ nm) of the channel. The slow variation of the velocity profile along the channel indicates that highly energetic (hot) electrons are distributed across the channel and that these hot electrons are not simply concentrated in the vicinity of the drain. The higher distribution of hot electrons, in turn, may help explain why the temperature distribution seen earlier extends to half the channel rather than being concentrated near the drain edge.

Finally, we show the distribution of the electron kinetic energy as we move from the source to the drain in Figure 6.12. It is evident that the distribution progressively departs further from the Maxwellian source distribution as we move towards the drain. In addition, the distribution becomes more and more asymmetric with respect to the maximum, especially in the drain region where it is rather flat for energies up to the maximum, with a thermal equilibrium tail for

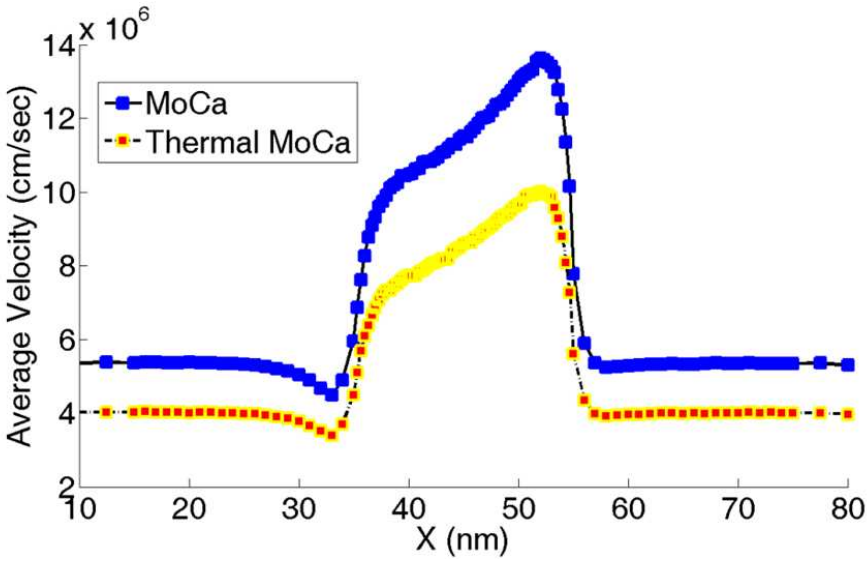


(a)

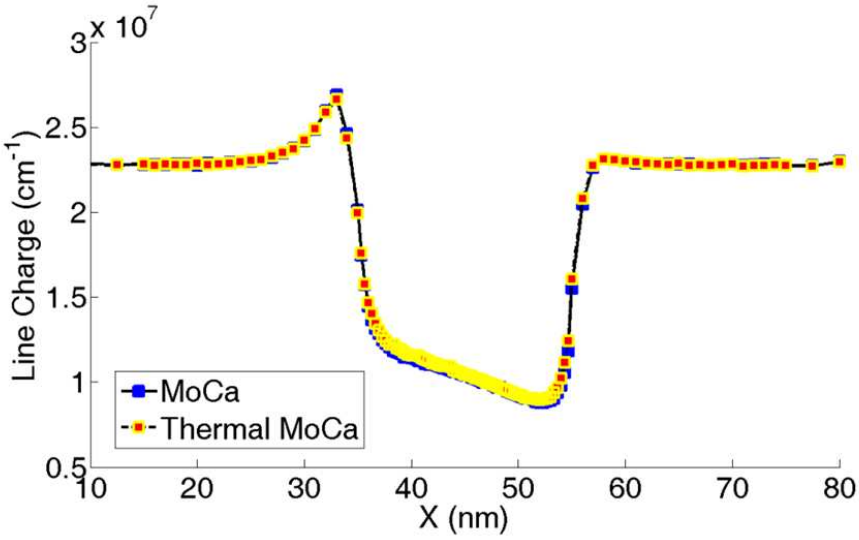


(b)

Figure 6.10: Comparison of the results obtained from thermal MoCa simulation and electrical MoCa simulation for a GAA structure. The variation of the drain current as we reduce the cross-section is shown in (a) while (b) shows the reduction in current gain.



(a)



(b)

Figure 6.11: Comparison of the (a) average velocity profile and (b) the electron density obtained using a 3D quantum-corrected Monte Carlo simulator with and without electro-thermal coupling.

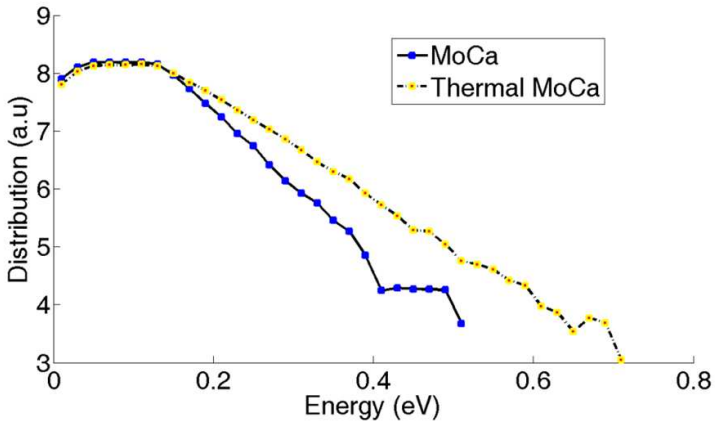
larger kinetic energies corresponding to the lattice temperature.

6.4 Summary

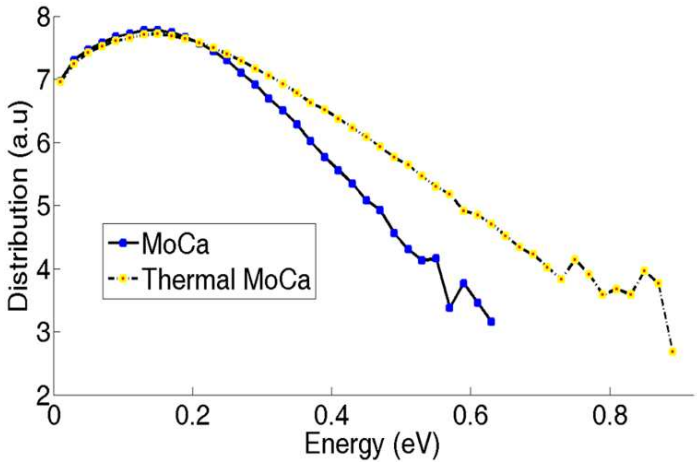
Although the planar bulk MOSFET technology has dominated chip design for many years, as we reach the limits of the capabilities of this technology, new architectures are constantly being explored. Heat has risen to be one of the most important considerations in assessing novel 3D nonplanar nanowire-like devices. In addition, 3D simulation tools which accurately couple electrical and thermal transport models are becoming increasingly important in order to validate these novel structures.

In this chapter, we have introduced our newly coupled 3D quantum corrected electro-thermal Monte Carlo simulation tool. Our model incorporates anharmonic phonon decay and temperature dependent scattering rates which are essential to obtaining the temperature distribution within the device. Our simulator readily provides all the relevant phonon statistics. The temperature hot spots, current degradation, and transport coefficients obtained can be used to compare various device architectures. We have demonstrated the impact of this tool by using it to study a test case of an SOI GAA structure. Our results have shown that current enhancement is decreased due to thermal effect, further reducing the limited gain from these devices, as has been discussed in Chapter 4. These results also indicate the promise of these tools to explore different areas of device design, including environmental impact. By optimizing heat issues at the transistor level, and consequently at all higher design levels, many advantages are gained. By using electro-thermal simulation and having a better understanding and control over the thermal issues within devices, designs can be more thermally conscious devices. Until recently, device design has been primarily driven by performance considerations, but this development allows them to optimize their design with respect to environmental issues, as well as to heat management and energy efficiency at the nanoscale level, and to packaging

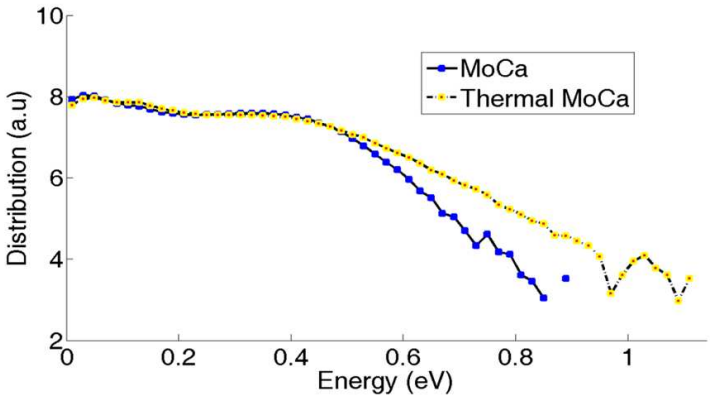
and material selection at other levels. The use of electro-thermal device simulators will help create a cycle which will push designs to have lower carbon footprints and be more environmentally conscious.



(a)



(b)



(c)

Figure 6.12: Electron distribution as a function of kinetic energy along the (a) source , (b) channel and (c) drain. The electron distribution progressively departs from an equilibrium distribution as we move towards the drain region.

REFERENCES

[1] S. Flachowsky, A. Wei, R. Illgen, T. Herrmann, J. Hontschel, M. Horstmann, W. Klix, and R. Stenzel, "Understanding strain-induced drive-current enhancement in strained-silicon n-MOSFET and p-MOSFET," *IEEE Transactions on Electron Devices*, vol. 57, no. 6, pp. 1343–1354, 2010.

[2] R. Delhougne, G. Eneman, M. Caymax, R. Loo, P. Meunier-Beillard, P. Verheyen, W. Vandervorst, K. D. Meyer, and M. Heyns, "Selective epitaxial deposition of strained silicon: A simple and effective method for fabricating high performance MOSFET devices," *Solid-State Electronics*, vol. 48, no. 8, pp. 1307–1316, 2004.

[3] Y.-M. Lin, H.-Y. Chiu, K. Jenkins, D. Farmer, P. Avouris, and A. Valdes-Garcia, "Dual-gate graphene FETs with f_T of 50 GHz," *IEEE Electron Device Letters*, vol. 31, no. 1, pp. 68–70, 2010.

[4] J. Xiang, W. Lu, Y. H. amd Y. Wu, H. Yan, and C. Lieber, "Ge/Si nanowire heterostructures as high-performance field-effect transistors," *Nature*, vol. 441, no. 7092, pp. 489–493, 2006.

[5] A. Godoy, F. Ruiz, C. Sampedro, F. Gamiz, and U. Ravaioli, "Calculation of the phonon-limited mobility in silicon gate all-around MOSFETs," *Solid-State Electronics*, vol. 51, no. 9, pp. 1211–1215, 2007.

[6] H.-S. Hahm, "Enhancement of Monte Carlo simulations on three-dimensional nanoscale semiconductor devices," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2010.

[7] J. Lai and A. Majumdar, "Concurent thermal and electrical modeling of submicrometer silicon devices," *Journal of Applied Physic*, vol. 79, pp. 7353 – 7361, 1996.

[8] T. Sadi, R. Kelsall, and N.J.Pilgrim, "Electrothermal Monte Carlo simulation of submicrometer Si/SiGe MODFETs," *IEEE Trans. on Electron Devices*, vol. 54, pp. 332 – 339, 2007.

[9] J. Rowlette and K. Goodson, "Fully coupled nonequilibrium electron-phonon transport in nanometer-scale silicon FETs," *IEEE Transactions on Electron Devices*, vol. 55, pp. 220–232, 2008.

[10] E. Pop, R. Dutton, and K. Goodson, "Monte Carlo simulation of Joule heating in bulk and strained silicon," *Applied Physics Letters*, vol. 86, p. 082101, 2005.

[11] S. Sinha, E. Pop, R. Dutton, and K. Goodson, "Non-equilibrium phonon distributions in sub-100 nm silicon transistors," *Journal of Heat Transfer*, vol. 128, pp. 638 – 647, 2006.

[12] K. Raleva, D. Vasileska, S. Goodnick, and M. Nedjalkov, "Modeling of thermal effects in nanodevices," *IEEE Trans. Electron Devices*, vol. 55, pp. 1306 – 1316, 2008.

[13] R. Parr and W. Yang, *Density Functional Theory for Atoms and Molecules*. Oxford: Oxford University Press, 1989.

[14] J. C. Inkson, *Many-body Theory of Solids: An Introduction*. New York: Plenum Press, 1984.

[15] W. Kohn, "Nobel lecture: Electronic structure of matter-wave functions and density functionals," *Reviews of Modern Physics*, vol. 71, pp. 1253–1266, 1999.

[16] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Physical Review*, vol. 136, pp. B864–B871, 1964.

[17] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation," *Physical Review*, vol. 140, no. 4A, pp. A1133–A1138, 1965.

[18] G. Roach, *Green's Functions*. New York: Cambridge University Press, 1982.

[19] D. Guan, "Study of quantum transport in nanoscale structures using non-equilibrium Green's function," Ph.D. dissertation, University of Illinois at Urbana, 2002.

[20] E. Economou, *Green's Functions in Quantum Physics*. Berlin: Springer, 1979.

[21] D. K. Ferry and H. L. Grubin, "Modeling of quantum transport in semiconductor devices," *Journal of Solid State Physics*, vol. 49, pp. 283–448, 1995.

[22] A. Gehring, "Simulation of tunneling in semiconductor devices," Ph.D. dissertation, Vienna University of Technology, 2003.

[23] E. Wigner, "On the quantum correction for thermodynamic equilibrium," *Phys. Rev*, vol. 40, pp. 749–759, 1932.

[24] E. Cancellieri, P. Bordone, A. Bertoni, G. Ferrari, and C. Jacoboni, "Wigner function for identical particles," *Journal of Computational Electronics*, pp. 411–415, 2005.

[25] H. Kosina, "Numerical simulation of quantum devices utilizing the Wigner function formalism," Presentation: Institute for Microelectronics Technical University Vienna, Austria, 2006.

[26] H. Kosina, G. Klimeck, M. Nedjalkov, and S. Selberherr, "Comparison of numerical quantum device models," in *Proceedings of the International Conference on Semiconductor Processes and Devices (SISPAD)*, Boston, MA, USA, September 2003.

[27] S. M. Sze, *Physics of Semiconductor Devices*. New York: John Wiley and Sons, Inc, 1981.

[28] W.-S. Choi, J.-K. Ahn, Y.-J. Park, H.-S. Min, and C.-G. Hwang, "A time dependent hydrodynamic device simulator SNU-2d with new discretization scheme and algorithm," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 13, pp. 899 – 908, 1994.

[29] A. Forghieri, R. Guerrieri, P. Ciampolini, A. Gnudi, M. Rudan, and G. Baccarani, "A new discretization strategy of the semiconductor equations comprising momentum and energy balance," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 7, pp. 231–242, 1988.

[30] G. A. Sai-Halasz, M. R. Wordeman, D. P. Kern, S. Rishton, and E. Ganin, "High transconductance and velocity overshoot in NMOS devices at the 0.1 μm gate-length level," *IEEE Electron Device Letter*, vol. 9, pp. 464–466, 1988.

[31] J. Song, Y. Park, and H. Min, "Drain current enhancement due to velocity overshoot effects and its analytic modeling," *IEEE Transaction on Electron Devices*, vol. 43, pp. 1870 – 1875, 1996.

[32] M. Cohen and T. Bergstresser, "Band structures and pseudopotential form factors for fourteen semiconductors of the diamond and zinc-blende structures," *Physical Review*, vol. 141, no. 2, pp. 789–796, 1966.

[33] 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. 867–76, 1998.

[34] Z. Aksamija and U. Ravaioli, "Joule heating and phonon transport in silicon MOSFETs," *Journal of Computational Electronics*, vol. 5, pp. 431–434, 2006.

[35] H. Tsuchiya, B. Winstead, and U. Ravaioli, "Quantum potential approaches for nano-scale device simulation," *VLSI Design*, vol. 13, pp. 335–340, 2001.

[36] R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals*. New York: McGraw Hill.

[37] G. Pei, J. Kedzierski, P. Oldiges, M. Ieong, and E.-C. Kan, "FinFET design considerations based on 3-d simulation and analytical modeling," *IEEE Tranactions. on Electron Devices*, vol. 49, pp. 1411–1419, 2002.

[38] G. Kathawala, B. Winstead, and U. Ravaioli, "Monte Carlo simulations of double-gate MOSFETs," *IEEE Tranactions on Electron Devices*, vol. 50, no. 12, pp. 2467–2473, 2003.

[39] B. Doyle, S. Datta, M. Doczy, S. Hareland, B. Jin, J. Kavalieros, T. Linton, A. Murthy, R. Rios, and R. Chau, "High performance fully-depleted tri-gate CMOS transistors," *IEEE Electron Device Letters*, vol. 24, no. 5, pp. 263–265, 2003.

[40] A. Sharma, S. Zaidi, S. Lucero, S. Brueck, and N. Islam, "Mobility and transverse electric field effects in channel conduction of wrap-around-gate nanowire MOSFETs," *IEE Proceedings Circuits, Devices and Systems*, vol. 151, no. 5, pp. 422–430, 2004.

[41] H. Sakaki, "Scattering suppression and high-mobility effect of size-quantized electrons in ultrafine semiconductor wire structures," *Japanese Journal of Applied Physics*, vol. 19, no. 12, pp. L735–L738, 1980.

[42] F. Balestra, S. Cristoloveanu, M. Benachir, J. Brini, and T. Elewa, "Double-gate silicon-on-insulator transistor with volume-inversion: A new device with greatly enhanced performance," *Electron Device Letter*, vol. EDL-8, no. 9, pp. 410–412, 1987.

[43] D. Querlioz, J. Saint-Martin, K. Huet, A. Bournel, V. Aubry-Fortuna, C. Chassat, S. Galdin-Retailleau, and P. Dollfus, "On the ability of the particle Monte Carlo technique to include quantum effects in Nano-MOSFET simulation," *IEEE Transactions on Electron Devices*, vol. 54, no. 9, pp. 2232–2242, 2007.

[44] J. Lopez-Villanueva, P. Cartujo-Cassinello, F. Gamiz, J. Banqueri, and A. Palma, "Effects of inversion-layer centroid on the performance of double-gate MOSFETs," *IEEE Trans. on Electron Devices*, vol. 47, no. 1, pp. 141–146, 2000.

[45] G. A. Kathawala and U. Ravaioli, "3-D Monte Carlo simulations of FinFETs," in *Technical Digest-International Electron Device Meeting (IEDM)*, Washington, DC, USA, 2003, pp. 683–686.

[46] F. Gamiz and M. Fischetti, "Monte Carlo simulation of double-gate silicon-on-insulator inversion layers: The role of volume inversion," *Journal of Applied Physics*, vol. 89, no. 10, pp. 5478–5487, 2001.

[47] J. Saint-Martin, A. Bournel, and P. Dollfus, "Comparison of multiple-gate MOSFET architectures using Monte Carlo simulation," *Solid-State Electronics*, vol. 50, pp. 94–101, 2006.

[48] C. Jacoboni and L. Reggiani, "The Monte Carlo method for solution of charge transport in semiconductor with application to covalent materials," *Reviews of Modern Physics*, vol. 55, pp. 645–705, 1983.

[49] K. Kim, B. Mason, and K.Hess, "Inclusion of collision broadening in semiconductor electron-transport simulations," *Physical Review B*, vol. 36, pp. 6547–6550, 1987.

[50] Y.-C. Chang, D. Tang, J. Tang, and K. Hess, "Monte Carlo simulation of impact ionization in GaAs including quantum effects," *Applied Physics letters*, vol. 42, no. 1, pp. 76–78, 1983.

[51] L. Register and K. Hess, "Improved algorithm for modeling collision broadening through a sequence of scattering events in semiclassical Monte Carlo," *Journal of Applied Physics*, vol. 87, pp. 303–311, 2000.

[52] M. Dove, *Introduction to Lattice Dynamics*. Cambridge: Cambridge University Press, 2005.

[53] G. Srivastava, *The Physics of Phonons*. Bristol: IOP Publishing Ltd., 1990.

[54] Y. Hsieh, "The vibrational spectrum and the specific heat of germanium and silicon," *The Journal of Chemical Physics*, vol. 22, pp. 306–311, 1954.

[55] F. Herman, "Lattice vibrational spectrum of germanium," *Journal of Physics and Chemistry of Solids*, vol. 8, pp. 405–418, 1959.

[56] P. Keating, "Effect of invariance requirements on the elastic strain energy of crystals with application to the diamond structure," *Physical Review*, vol. 145, no. 2, pp. 637–645, 1966.

[57] E. W. Kellermann, "Theory of the vibrations of sodium chloride lattice," *Philosophical Transacations of the Royal Society London*, vol. 238, pp. 513–548, 1940.

[58] W. Cochran, "Theory of the lattice vibrations of germanium," *Proceeding of the Royal Society*, vol. 253, pp. 260–276, 1959.

[59] W. Kress, "Lattice dynamics of highly polarizable homopolar crystals with diamond structure," *Physica Status Solidi (b)*, vol. 49, pp. 235–245, 1972.

[60] P. Yu and M. Cardona, *Fundamentals of Semiconductors: Physics and Materials Properties*. Berlin: Springer, 2010.

[61] J. Phillips, "Covalent bond in crystals. I. Elements of a structural theory," *Physical Review*, vol. 166, pp. 832–838, 1968.

[62] M. Martin, "Dielectric screening model for lattice vibrations of diamond-structure crystals," *Physical Review*, vol. 186, pp. 871–884, 1969.

[63] W. Weber, "New bond-charge model for the lattice dynamics of diamond-type semiconductor," *Physical Review letters*, vol. 33, p. 371374, 1974.

[64] W. Weber, "Adiabatic bond charge model for the phonons in diamond, Si, Ge, and α-Sn," *Physics Reviews B*, vol. 15, pp. 4789–4803, 1977.

[65] H. M. T. tütüncü and G. P. Srivastava, "Phonons in zinc-blende and wurtzite phases of GaN, AlN, and BN with the adiabatic bond-charge model," *Physical Review B*, vol. 62, pp. 5028–5035, 2000.

[66] S. kit Yip and Y.-C. Chang, "Theory of phonon dispersion relations in semiconductor superlattices," *Physical Review B*, vol. 30, pp. 7037–7059, 1984.

[67] S. Volz and G. Chen, "Molecular dynamics simulation of thermal conductivity of silicon nanowires," *Applied Physics Letters*, vol. 75, pp. 2056–2058, 1999.

[68] S. J. Narumanchi, J. Y. Murthy, and C. H. Amon, "Comparison of different phonon transport models for predicting heat conduction in silicon-on-insulator transistors," *Journal of Heat Transfer*, vol. 127, pp. 713–723, 2005.

[69] S. J. Narumanchi, J. Y. Murthy, and C. H. Amon, "Boltzmann transport equation-based thermal modeling approaches for hotspots in microelectronics," *Heat Mass Transfer*, vol. 42, pp. 478–491, 2006.

[70] A. Majumdar, "Microscale heat conduction in dielectric thin films," *Journal of Heat Transfer*, vol. 115, pp. 7–16, 1993.

[71] D. D. Joseph and L. Preziosi, "Heat waves," *Review of Modern Physics*, vol. 61, pp. 41–73, 1989.

[72] C.-L. Tien, A. Majumdar, and F. Gerner, *Microscale Energy Transport*. London: Taylor and Francis, 1998.

[73] "International technology roadmap for semiconductors (ITRS) executive summary," 2009. [Online]. Available: http://public.itrs.net/

[74] M. Hjelm and H. E. Nilsson, "Methods for anisotropic selection of final states in the full band ensamble Monte Carlo simulation framework," *Simulation Practice and Theory*, vol. 9, pp. 331–332, 2002.

[75] D. G. Cahill, W. Ford, K. Goodson, G. Mahan, A. Majumdar, H. Maris, R. Merlin, and S. Phillpot, "Nanoscale thermal transport," *Journal of applied physics*, vol. 93, pp. 793–818, 2003.