ABSTRACT

Title of dissertation: THERMAL AND PERFORMANCE MODELING OF NANOSCALE MOSFETS, CARBON NANOTUBE DEVICES AND INTEGRATED CIRCUITS

Akin Akturk, Doctor of Philosophy, 2006

Dissertation directed by: Professor Neil Goldsman
Department of Electrical and
Computer Engineering

We offer new paradigms for electronic devices and digital integrated circuits (ICs) in an effort to overcome important performance threatening problems such as self heating. To investigate chip heating, we report novel methods for predicting the thermal profiles of complex ICs at the resolution of a single device. We resolve device and IC temperatures self-consistently, with individual device performances, while accounting for IC layout and software application details. At the device level, we calculate performance and generated heat details. We then extend these performance figures to the overall chip using a stochastic or Monte Carlo type methodology. Next, at the IC level, we solve for the device temperatures using the chip's layout and application software details. Here, we apply our mixed-mode algorithm to two-dimensional (planar) and three-dimensional ICs. To relieve thermal stresses and performance degradation in specific areas of extreme heating or hot spots, we offer design strategies using thermal contacts or different IC layouts. Moreover, we

also show chips that we had designed and fabricated through IC fabrication clearing house MOSIS for experimental investigations.

We also investigate carbon nanotubes (CNTs) and CNT embedded MOSFETs as new device paradigms for future electronic circuits. To examine the effects of CNTs on device performance, we develop a CNT Monte Carlo simulator, and determine scattering rates and CNT electron transport. Here, we report position-dependent velocity oscillations and length effects in semiconducting single-walled zig-zag carbon nanotubes. Our calculated results indicate velocity oscillations in the Terahertz range, which approaches phonon frequencies. This may facilitate new high frequency RF device and circuit designs, opening new paradigms in communication networks. Furthermore, to obtain device performance figures for MOSFETs that embed CNTs in their channels, our device solver determines interactions between the CNT and silicon (Si) by obtaining quantization and transport effects on the tube and the Si, and at the CNT-Si barrier. We predict that the CNT-MOSFET yields a better performance than the traditional MOSFET. Especially, CNT-MOSFETs employing lower diameter tubes exhibit improved performance capabilities. We also perform similar analyses for CNT embedded SOI-MOSFETs.

THERMAL AND PERFORMANCE MODELING OF NANOSCALE MOSFETS, CARBON NANOTUBE DEVICES AND INTEGRATED CIRCUITS

by

Akin Akturk

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2006

Advisory Commmittee:

Professor Neil Goldsman, Chair/Advisor
Professor Martin Peckerar
Professor Chia-Hung Yang
Professor Christopher Davis
Associate Professor Michael Fuhrer

DEDICATION

To my parents, Fatma and Hüseyin Aktürk.

ACKNOWLEDGMENTS

I would like to thank my research advisor Prof. Dr. Neil Goldsman for guiding me in my research studies. His support made this work both possible and enjoyable for me. Also, I especially would like to thank Prof. Martin Peckerar for valuable discussions. In addition, I would like to thank the Department of Electrical and Computer Engineering of the University of Maryland College Park, LPS and George Metze for their supports.

Next, I would like to thank members of my dissertation committee. I would like to thank Prof. Chia-Hung Yang, Prof. Christopher Davis, Associate Prof. Michael Fuhrer, and once again Prof. Martin Peckerar and my advisor Prof. Neil Goldsman.

I also wish to thank my colleagues in my research group for their support and help: Zeynep Dilli, Dr. Gary Pennington, Dr. Chung-Kuang, Dr. Stephen Powell, Siddharth Potbhare, Bo Yang, Latise Parker, Datta Sheth, Yun Bai, Amrit Bandyopadhyay and Ben Funk.

Finally, I would like to thank my parents and Adam Markowski for their support and love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Motivation

As integrated circuits (ICs) become more densely packed with transistors and we approach the end of the semiconductor roadmap, manufacturers are facing several important problems threatening chip performance. To overcome these problems, investigators are exploring new paradigms for electronic devices and digital integrated circuits [1, 2].

For future integrated circuits, one especially important difficulty is chip heating [1]-[11]. Investigators have pointed out that towards the end of the semiconductor roadmap, there will be more devices per unit area due to scaling of physical device dimensions. This real estate crowding induces high temperatures, since power density can not be kept in line with the conventional scaling algorithm. The rule of thumb for the traditional scaling algorithm is that all relevant parameters are scaled by the same factor S, either proportionately or inversely, to keep the power density fixed. For example, physical dimensions and supply voltage are scaled downward by S, while frequency and capacitance per area are scaled upward by S. The power density, therefore, stays the same after scaling. However, voltage scaling will no longer be applicable for such small dimensions because of the intrinsic limitations of silicon (Si) bandgap and built-in voltages [1]-[5]. Therefore, IC manufacturers

deviate from the traditional scaling methods to guarantee good device and chip performances such as high on/off current ratios and subthreshold slopes. This results in clock frequencies and supply voltages that are higher than previously expected. In addition, nanoscale devices can not provide as much isolation between supply rails as previously employed longer channel devices. This leads to higher leakage levels. The chip also is likely to overheat faster than conventional cooling methods can account for. Thus power density per unit area keeps increasing exponentially for future electronic devices, making full-chip heating increasingly influential in the performance of next generation ICs. Hence, chip heating is considered as one of the major obstacles to be overcome for future IC designs [1]-[11].

To fully understand the chip-heating problem, researchers need tools to simulate and examine the phenomenon. These modeling tools can also be used to relieve heating problems by offering new design approaches to the chip layout. Preliminary research has been done to estimate the temperature profile for given chips [7]-[11]. Here, we develop a tool that establishes the necessary link between single device operation and full-chip heating for the first time in the literature.

We also explore new alternatives to conventional MOSFETs. Carbon nanotubes (CNTs) are being explored as a structure that may play a leading role in future electronic systems [12]-[15]. CNTs are planar graphite sheets (graphene) that are seamlessly wrapped into tubes. CNTs possess favorable electrical characteristics and can be fabricated in dimensions as small as 8Å in diameter. The electrical characteristics of CNTs vary with the diameter and the wrapping angle of the graphene [16]. Both the diameter and the wrapping angle can be described by the tube's

fundamental indices (l,m) (Standard notation uses (n,m). However, l is used in some chapters instead of n to avoid confusion with electron concentration). Theory indicates that CNTs can be metallic or semiconducting according to the fundamental tube indices (l,m), with the bandgap of the semiconducting tube depending on the CNT diameter. Analysis shows semiconducting CNTs have very high low-field mobilities, with peak electron drift velocities that can be as much as five times higher than that of silicon [17]-[21]. It has also been shown that tubes can be doped by donors and acceptors [22]-[24]. Experiments and calculations also indicate that CNTs may facilitate devices with large transconductances and high drive currents [20]-[39]. Experiments have demonstrated the viability of CNT-based FETs [34, 35], and CNT-SOI type MOSFETs [36, 37]. Moreover, we did preliminary research on modeling and design of CNT embedded bulk MOSFETs [30, 31].

Here, we investigate several CNT-MOSFET devices for the first time in the literature. Our calculations indicate that CNT-MOSFETs can have improved device performance over conventional MOSFETs [30, 31]. To investigate the potential attributes of the new designs, we developed a methodology for modeling nanoscale CNT-MOSFETs. We also used the same methodology to obtain device performance figures for SOI-MOSFETs that have CNTs embedded in their channels.

1.2 Device Modeling

To obtain performance details of devices like nanoscale MOSFETs and Silicon-On-Insulator (SOI) MOSFETs, several modeling methods either based on compact

analytical equations or physics can be utilized. Even though the ones based on compact analytical equations such as the SPICE model are useful for fast performance computation of devices and circuits containing several nodes, their results can not be extrapolated to predict performance details of smaller devices. The reason is that the parameters inherent to these models are empirically determined using experimental data or simulated (using a lower level device solver) device performance characteristics for that technology node. Therefore, their applicability to other, especially smaller, technology nodes might lack underlying physics. Also, these models can not be extended to unconventional device structures, since these novel structures are not geometrically relevant to those used to extract fitting parameters. In addition, they may not be governed by the same physical relations. Thus, to investigate nanoscale devices, researchers need to use physics based models.

Depending on the number of details and assumptions included, physical models can be divided up into three main categories:

- Classical Model
- Semiclassical Model
- Quantum Model

The classical models are the moments of the Boltzmann Transport Equation (BTE). On the other hand, the BTE is a semiclassical model. Moreover, the BTE is a continuity equation for electrons, and keeps track of the electron distribution in real space, momentum space and time. According to carrier continuity, conservation of carriers requires that the change in carrier distribution in time should differ

from the gradient of the net flux in real space and momentum space by the net contribution of "sources" and "sinks" in the enclosed momentum and real space volume. More specifically, if $f(r,p,t)$ is the distribution function that gives the probability of finding a particle at position r and time t with momentum p, then particle continuity equation is the following.

$$\left(\frac{\partial}{\partial t}+\frac{\partial \vec{r}}{\partial t}\frac{\partial}{\partial r}+\frac{\partial \vec{p}}{\partial t}\frac{\partial}{\partial p}\right)f = G - R \tag{1.1}$$

Here, G and R are generation ("source") and recombination ("sink") terms, respectively. The net generation-recombination rate in real space is determined by the net scattering rate $s(r,p,t)$ that accounts for phenomena such as photogeneration and recombination through traps or from band-to-band (Auger recombination). In momentum space, the collision ratio $\left.\frac{\partial f}{\partial t}\right|_{\text{coll}}$ gives the net generation-recombination rate, which accounts for phenomena such as reflections and transmissions. Using the two terms for the net generation-recombination rate in real and momentum spaces, and replacing $\frac{\partial \vec{r}}{\partial t}$ and $\frac{\partial \vec{p}}{\partial t}$ with $\vec{v}$ (velocity) and $\vec{F}$ (force, which is equal to $-q\vec{E}$ in an electric field $\vec{E}$), respectively, the BTE becomes:

$$\frac{\partial f}{\partial t}+\vec{v}\cdot\vec{\nabla_r}f+\vec{F}\cdot\vec{\nabla_p}f = s(r,p,t)+\left.\frac{\partial f}{\partial t}\right|_{\text{coll}} \tag{1.2}$$

To find the spatial and time distribution of a physical parameter $\zeta(r,t)$, the distribution function $f(r,p,t)$ is first scaled by associated weighting coefficients $w(p)$ and normalization factor Ω, and then integrated over the momentum space, as written below:

$$\zeta(r,t)=\frac{1}{\Omega}\int w(p)f(r,p,t)dp \tag{1.3}$$

Above, substituting $w(p)$ with 1, p (momentum) or ε (energy) respectively gives electron concentration n, average momentum P or average energy W for $\zeta(r,t)$. To obtain values for these physical parameters, we first need to calculate $f(r,p,t)$. This requires finding a solution for the BTE [40, 41] for the physical parameter of interest. Below, we include the continuity equation for any physical parameter, which is obtained by multiplying Eqn. 1.2 by $\frac{1}{\Omega}\int w(p)dp$:

$$\frac{1}{\Omega}\int w(p)\frac{\partial f}{\partial t}dp + \frac{1}{\Omega}\int w(p)\vec{v}\cdot\vec{\nabla_r}f\,dp - q\frac{1}{\Omega}\int w(p)\vec{E}\cdot\vec{\nabla_p}f\,dp = \frac{1}{\Omega}\int w(p)s(r,p,t)dp + \frac{1}{\Omega}\int w(p)\left.\frac{\partial f}{\partial t}\right|_{\text{coll}}dp \quad (1.4)$$

To derive continuity equations, we substitute $w(p)$ with 1, which is equal to p^0. Thus, the drift-diffusion model is a classical model, and can be derived using the zeroth and first moments of the BTE. Moreover, the electron current continuity equation is derived using Eqn. 1.4 and the aforementioned expansion in p (p^0). Next, the integrals in Eqn. 1.4 are evaluated, and the following expressions are substituted for each term above in the given order:

$$\frac{1}{\Omega}\int\frac{\partial f}{\partial t}dp = \frac{\partial n}{\partial t} \quad (1.5)$$

$$\frac{1}{\Omega}\int\vec{v}\cdot\vec{\nabla_r}f\,dp = \vec{v}\cdot\vec{\nabla_r}n \quad (1.6)$$

$$-q\frac{1}{\Omega}\int\vec{E}\cdot\vec{\nabla_p}f\,dp = -\frac{qEf}{\Omega} \quad (1.7)$$

$$\frac{1}{\Omega}\int s(r,p,t)dp = G_n - R_n \quad (1.8)$$

$$\frac{1}{\Omega}\int\left.\frac{\partial f}{\partial t}\right|_{\text{coll}}dp = -\frac{\Delta n}{\langle\tau_n\rangle} \quad (1.9)$$

Here, we have change in carrier (electron n) concentration in time, Eqn. 1.5, divergence of drift flux, Eqn. 1.6, field generation rate, Eqn. 1.7, net generation-

recombination term, Eqn. 1.8, and divergence of carrier diffusion flux, Eqn. 1.9. In addition, $G_n - R_n$ is the net generation-recombination rate and $\langle \tau_n \rangle$ is the ensemble relaxation rate. Furthermore, since f approaches zero fast, Eqn. 1.7 can be approximated by zero for low fields. Also, we further simplify the divergence of the diffusion flux using the diffusion constant D_n:

$$\frac{\Delta n}{\langle \tau_n \rangle} = \frac{\Delta r^2}{\langle \tau_n \rangle} \frac{\Delta n}{\Delta r^2} \cong D_n \nabla_r^2 n \tag{1.10}$$

Combining drift and diffusion fluxes together, we write the electron current continuity equation in a familiar form, as follows:

$$\frac{\partial n}{\partial t} = -\nabla_r \cdot (\vec{v} n + D_n \vec{\nabla_r} n) + G_n - R_n \tag{1.11}$$

Likewise, substituting $w(p)$ with p in Eqn. 1.4, we derive the momentum balance equation. In this case, we have the following terms for the BTE (in one dimension):

$$\frac{1}{\Omega} \int p \frac{\partial f}{\partial t} dp = \frac{\partial P}{\partial t} \tag{1.12}$$

$$\frac{1}{\Omega} \int p \vec{v} \cdot \vec{\nabla_r} f dp = 2 \nabla_r \cdot \vec{W} \tag{1.13}$$

$$-q \frac{1}{\Omega} \int p \vec{E} \cdot \vec{\nabla_p} f dp = -\frac{qEpf}{\Omega} - qnE \tag{1.14}$$

$$\frac{1}{\Omega} \int p \left. \frac{\partial f}{\partial t} \right|_{\text{coll}} dp = -\frac{\Delta P}{\langle \tau_p \rangle} \tag{1.15}$$

We have change in carrier momentum in time, Eqn. 1.12, divergence of average energy ($\frac{1}{\Omega} \int p \vec{v} \cdot \vec{\nabla_r} f dp = \frac{1}{\Omega} \int 2\varepsilon \cdot \vec{\nabla_r} f dp = 2W$), Eqn. 1.13, field generation rate, Eqn. 1.14, and collision term, Eqn. 1.15. We do not have the scattering term as the one in Eqn. 1.8, because that does not contribute to a change in momentum space.

Additionally, in Eqn. 1.14, $-qEpf \cong 0$ because f approaches zero fast for large p. Therefore, we can write a compact momentum balance equation, which is the first moment of the BTE, as shown below:

$$\frac{\partial P}{\partial t} = -2\nabla_r \cdot \vec{W} - qnE - \frac{\Delta P}{\langle \tau_p \rangle} \tag{1.16}$$

The second moment of the BTE is the energy balance equation. It is derived by substituting $w(p)$ with ε ($= p^2/2m^*$ for parabolic bands, where m^* is the effective mass) in Eqn. 1.4. This gives the following terms for the energy balance equation:

$$\frac{1}{\Omega}\int \varepsilon \frac{\partial f}{\partial t} dp = \frac{\partial W}{\partial t} \tag{1.17}$$

$$\frac{1}{\Omega}\int \varepsilon \vec{v} \cdot \vec{\nabla_r} f dp = \nabla_r \cdot \frac{1}{\Omega}\int \varepsilon \vec{v} f dp \tag{1.18}$$

$$-q\frac{1}{\Omega}\int \varepsilon \vec{E} \cdot \vec{\nabla_p} f dp = -\frac{qE\varepsilon f}{\Omega} - \frac{q\vec{E}\vec{P}}{m^*} \tag{1.19}$$

$$\frac{1}{\Omega}\int \varepsilon s(r,p,t) dp = G_\varepsilon - R_\varepsilon \tag{1.20}$$

$$\frac{1}{\Omega}\int \varepsilon \left.\frac{\partial f}{\partial t}\right|_{\text{coll}} dp = -\frac{\Delta W}{\langle \tau_\varepsilon \rangle} \tag{1.21}$$

Here, change in average energy in time, Eqn. 1.17, is related to the flow of energy, Eqn. 1.18, self-heating, Eqn. 1.19, energy exchange due to scattering leading to recombination and generation, Eqn. 1.20, and energy exchange between lattice and carriers, Eqn. 1.21. In Eqn. 1.19, the first term on the right-hand-side is approximately zero due to fast decay of f for large p, and the second term is equal to $\vec{J} \cdot \vec{E}$ noting that $\frac{q\vec{P}}{m^*} = \vec{J}$, where $\vec{J}$ is the current density. Also, $\vec{J} \cdot \vec{E}$ can be recognized as Joule Heating. Therefore, the energy balance equation can be written in a concise form, as shown below:

$$\frac{\partial W}{\partial t} = -\nabla_r \cdot \frac{1}{\Omega}\int \varepsilon \vec{v} f dp + \vec{J} \cdot \vec{E} - \frac{\Delta W}{\langle \tau_\varepsilon \rangle} + G_\varepsilon - R_\varepsilon \tag{1.22}$$

We have shown zeroth, first and second moments of the BTE in Eqns. 1.11, 1.16 and 1.22, respectively. They are the carrier balance, momentum balance and energy balance equations in the aforementioned order. We note that each balance equation requires solution of higher order balance equations. For example, in Eqn. 1.11, $\vec{v} = \vec{P}/m$; thus, to solve for n, we need to know P. Likewise, to solve for P in Eqn. 1.16, we need the spatial variation of W. Thus, it is impossible to solve the BTE moment equations unless we achieve closure using some approximations. The most commonly-used approximation is to write the drift velocity in terms of the electric field $\vec{v} = \mu\vec{E}$, where mobility μ is the proportionality constant that is determined empirically from experimental data or detailed simulations such as low-level Monte Carlos (MCs). This provides the closure for the current continuity equation. We now need a relation between the carrier densities and the electric field to have a complete set of equations. This is provided by the Poisson Equation, which relates divergence of local electric field weighted by the dielectric constant, ϵ, to the net charge density, ρ, as written below:

$$\nabla_r \cdot \epsilon\vec{E} = \rho \tag{1.23}$$

The carrier balance (current continuity) equation 1.11 along with the Poisson Equation, written above, forms the drift diffusion model. This is the most commonly-solved model to obtain device characteristics, mainly due to its simplicity and clear physical interpretation. However, ignoring the higher order moments of the BTE results in some loss of physical details. This includes loss of distribution details of carriers in energy space or temperature, which may be important

to characterize effects involving hot-electrons. It also includes the assumption that carriers reach equilibrium with the lattice through scattering. This enables the use of mobility and diffusion coefficient concepts. For not-so-high electric fields, hot electron effects are negligible; therefore, we can tolerate the loss of those data. In terms of scattering dominated electron current, unless the system is purely quantum mechanical such as those that can be found in ultra small devices, we can still develop a mobility model that accounts for non-equilibrium conditions by adjusting the mobility through experiments, published data or MC simulations. To improve the drift-diffusion model, researchers sometimes solve for the energy balance equation in addition to the carrier balance equation. In this case, the model is called the hydrodynamic model. Even though this model may successfully predict hot-electron effects, it may also produce erroneous data showing energy peaks near the MOSFET channel-source junction [40]. Therefore, its results should be carefully interpreted.

So far, we have described classical models, which are essentially the drift-diffusion and hydrodynamic models. Next comes the semiclassical models in the list of physical models. The main characteristics of semiclassical models are their use of scattering rates at the microscopic level (instead of using a mobility concept as in the classical model), and slope of the energy dispersion curve to determine the electron velocity.

Obtaining device characteristics using a semiclassical model can be done mainly in two ways. First is the single particle approach or the Monte Carlo (MC) method. This method includes statistical means to obtain macroscopic characteristics such as terminal currents. In this method, carriers are randomly picked from an energy-

momentum pool using carrier distribution details such as Fermi-Dirac statistics in conjunction with the energy dispersion curves. They then are accelerated under the macroscopic field with an effective mass m^* for a predetermined time that is long enough for acceleration but too short for any scattering event to take place. At the end of the flight, depending on a probabilistically dictated selection criteria, either the carrier continues its free flight or a randomly determined microscopic scattering event that generally depends on the carrier's energy and momentum occurs. Repeating this drift-scattering combination a significantly large enough number of times allows determination of average macroscopic field dependent parameters such as drift velocity for the carrier with the preselected initial energy-momentum combination. For different energy-momentum combinations in the energy-momentum pool, we can obtain similar average macroscopic field dependent parameters. Lastly, scaling these average macroscopic field dependent parameters in conjunction with the associated initial selection statistics would give the average macroscopic field dependent parameters for the carrier ensemble.

The second way of calculating device characteristics using a semiclassical model is to solve the BTE that includes scattering details at the microscopic level and field-propelled drift of a particle with an effective mass m^* at the macroscopic level. One common practice is to solve for the space, time and momentum distributions. This forms a seven dimensional problem: 3 in space, 3 in momentum and 1 in time. It is a challenging problem that puts too much burden on the CPU. However, the total number of dimensions can be reduced using spherical harmonic basis functions [40]. This would make the problem more manageable for the CPU,

by putting some of the analytical burden on the programmer's side.

Semiclassical models include more physical details than classical models. They involve fewer approximations and more elemental properties. Thus, one can simulate devices with different geometries without the need to have fitting parameters. In addition, semiclassical models can provide energy distribution resolving phenomena such hot-electron effects.

One important deficiency of the BTE model is that the carriers are treated as particles; therefore, they obey classical or Newtonian mechanics. However, the particle approach can not resolve transport in ultra-small devices, and quantization in the MOSFET channel near the silicon/silicon dioxide (Si-SiO_2) interface, where a potential well depending on bias conditions may form. This is related to the failure of the particle approach in the characterization of phenomena such as quantization and tunneling. To resolve these effects, a quantum model needs to be employed. One common practice is to solve the single-particle Schröndinger equation. Using a general form, the time-independent Schröndinger equation can be written as follows:

$$-\frac{\hbar^2}{2m_o}\nabla_r^2\psi + \left[E_C(r) + U_C(r) + U_S(r)\right]\psi = \varepsilon\psi \tag{1.24}$$

Above, ψ is the wave function, from which the probability of finding an electron in a volume ($V = r^3$) can be determined using the definite integral $\int_V \psi^*(r)\psi(r)dr^3$. The actual form of ψ depends on the slowly varying potential $E_C(r)$ due to applied field and built-in potentials, quickly varying $U_C(r)$ due to crystal or lattice potential, and scattering potential $U_S(r)$ due to phonons (lattice vibrations), ionized impurities, etc. Also, ψ is the eigenfunction and ε is the eigenvalue of Eqn. 1.24.

In traditional device and material analyses, generally the three potential terms on the left-hand-side of Eqn. 1.24 are treated separately, assuming that impurity or defect densities are low compared to the actual material density, lattice vibrations do not significantly alter the relative positions of the neighboring atoms, applied potential varies slowly compared to the crystal potential, etc.

A common practice is first to solve the following equation to obtain energy dispersion curves.

$$\left[-\frac{\hbar^2}{2m_o}\nabla_r^2 + U_C(r)\right]\psi_1 = \varepsilon_1\psi_1 \tag{1.25}$$

Crystal potential $U_C(r)$ has the periodicity of the lattice due to the periodic assembly of the atoms in the material. If the sample is long enough that fringe effects can be ignored, we can assume that the wavefunctions have the same spatial periodicity. This enables us to write eigenfunctions (ψs) of Eqn. 1.25 in terms of *Bloch waves*. A *Bloch wave* is a product of two terms. The first term is a periodic function that has same periodicity with the underlying structure ($u_k(\vec{r}) = u_k(\vec{r}+\vec{T})$, where T is the lattice translational vector in space that is equal to lattice separation), and the second term is a plane wave ($e^{i\vec{k}\cdot\vec{r}}$, where lattice periodicity dictates $k = k + \frac{2\pi l}{T}$, $l = 0, \pm1, \pm2\ldots$). Mathematically, it is defined as follows:

$$\Psi_k = u_k(r)e^{i\vec{k}\cdot\vec{r}} \tag{1.26}$$

Substituting $\psi_1 = \sum \Psi_k$ in Eqn. 1.25, and canceling common terms on both sides give:

$$\left[-\frac{\hbar^2}{2m_o}\left(\nabla_r^2 + 2ik - k^2\right) + U_C(r)\right]u_k = \varepsilon(k)u_k \tag{1.27}$$

Here, we also dropped the sum written above, because solutions for each k

form a linearly independent set for solutions of the sum. This can be verified by multiplying the original equation by $e^{-i\vec{k}\cdot\vec{r}}$, and integrating it over k. Additionally, we can solve for the eigenenergies $\varepsilon(k)$ for a given k, since we know the terms in brackets on the left-hand-side of Eqn. 1.27. Tracing over k values and finding the associated allowable energies give the energy dispersion curves including conduction and valence bands for that material. Using these dispersion curves, we can extract pertinent parameters to be used in transport calculations such as effective mass m^* and bandgap ε_g.

Next, we add the scattering potential related details in addition to crystal potential related dynamics to our quantum system by solving:

$$\left[-\frac{\hbar^2}{2m^*}\nabla_r^2 + U_S(r)\right]\psi_2 = \varepsilon_2\psi_2 \tag{1.28}$$

Generally, the scattering potential is time-dependent; therefore, it is more appropriate to write the above equation as follows:

$$\left[-\frac{\hbar^2}{2m^*}\nabla_r^2 + U_S(r,t)\right]\psi_2 = i\hbar\frac{\partial\psi_2}{\partial t} \tag{1.29}$$

Equation 1.29 includes perturbations to the crystal potential. Assuming that these perturbations are small, we can use perturbation theory to calculate scattering rates. This leads to the basic result of scattering theory that is also known as *Fermi's Golden Rule*, which defines scattering rate from momentum k to k' as written below:

$$S(k,k') = \frac{2\pi}{\hbar}|H^a_{k'k}|\delta(E(k) + \hbar\omega - E(k')) + \frac{2\pi}{\hbar}|H^e_{k'k}|\delta(E(k) - \hbar\omega - E(k')) \tag{1.30}$$

Here, $H^{a,e}_{k'k}$ is the time-independent scattering potential matrix element between states k and k'. In addition, we have conservation of energy as explicitly

written in the arguments of the Dirac-δ functions for phonon absorption (^{a}) and emission (^{e}). Moreover, we also have conservation of momentum that is not explicitly shown above.

To find scattering rates from k to k', we first need to determine matrix elements either in exact form or in approximate form using the deformation potential approximation [42]. Second, we need to obtain phonon dispersion curves to calculate phonon energy $\hbar\omega$ and momentum q. This can be done in various ways including the tight binding approximation [16]. Once we calculate scattering rates, we can use Monte Carlo simulations to extract parameters such as drift velocity and mobility to be used in device simulators.

Lastly, to complete the full quantum treatment, we resolve the contribution from the last potential term in Eqn. 1.24, which is the combined term for applied and built-in potentials.

$$\left[-\frac{\hbar^2}{2m^*}\nabla_r^2 + E_C(r)\right]\psi_3 = \varepsilon_3\psi_3 \tag{1.31}$$

From the dispersion curves (solution of Eqn. 1.25), we obtain the effective mass m^*. Next, using the scattering rate data, we obtain transport details that can be used in balance equations. (For ballistic transport, the scattering potential is approximately zero resulting in undisturbed flight between two device terminals.) Now, we include the last quantum effect that provides the link between quantum potential and transport. Here, $E_C(r)$ accounts for the externally applied field, and the built-in fields due to doping, difference in workfunctions and bandgaps, etc. The details of solving this equation in conjunction with scattering data will be given later

in the following chapters.

So far, we have discussed three main physical device models: Classical, semiclassical and quantum. We here make use of modified drift-diffusion equations that combine details from the classical and quantum models, and a Monte Carlo scheme using the semiclassical model. More specifically, we solve the modified drift-diffusion equations to obtain comparative device performance figures. Our modifications are for resolving quantum, heterostructure and thermal effects. To resolve these effects using the drift-diffusion model, we develop methodologies. In the following chapters, we explain how we incorporate those effects in detail. Furthermore, to obtain electrical characteristics of CNTs, we make use of a semiclassical model, which is a Monte Carlo simulator.

1.3 Carbon Nanotube Devices

As we approach the end of the semiconductor roadmap, investigators are exploring new paradigms for electronic devices. Carbon nanotubes (CNTs) are being explored as a structure that may play a leading role in future electronic systems [12]-[15] due to their favorable electrical characteristics and angstrom scale dimensions. Therefore, we develop methodologies to obtain their electrical properties and possible relative gains associated with their use in devices.

To investigate CNT electronics, we first determine CNT electrical characteristics. Since, there has not been enough research done on CNT electrical properties, we start from low level physics, and extract pertinent CNT electrical characteristics.

Low-level CNT modeling requires solution of the single-particle Schröndinger equation. Therefore, we first determine the energy dispersion curves using details of its crystal structure. Noting that CNTs are graphite sheets (graphene) rolled into tubes, we use well-known details of the graphene band structure with modifications. Mainly, the modifications take into consideration that CNTs exhibit folding effects, which lead to bound states around the circumference. Details of these states are determined by the diameter of the tube and the wrapping angle using the zone-folding method. Moreover, restrictions due to finite lengths of the tubes are also considered for some applications. In summary, instead of solving for the CNT band diagram from scratch, well-known details of the graphene band structure are used to approximate energy-momentum curves of CNTs. Likewise, the phonon dispersion curves of CNTs are determined from those of graphene.

Next, our quantum modeling includes extraction of macroscopic electrical properties that are relevant for device performance calculations such as drift velocity and mobility. This requires resolving scattering effects on CNT electron transport using Monte Carlo (MC) simulations. In these MC simulations, scattering rate calculations are facilitated using the deformation potential approximation.

Essentially, MC simulators use numerical techniques that rely on a theorem called the weak law of large numbers [43]. According to this theorem, the average values associated with independent and identically distributed random sequences will converge to constant values as the size of the sequence, or the number of samples, go to infinity. In practical terms, average values of independent and identically distributed random sequences can be obtained if a large sequence of the correspond-

ing event is simulated, or recorded experimentally.

Once low-level effects are resolved, we obtain electrical CNT performance details, and do comparative device analyses of various CNT-embedded and traditional devices. We will give details of the underlying physics and employed modeling techniques in the following chapters.

1.4 Integrated Circuit Modeling

As industry reduces the size of devices further to increase speed and functionality of Integrated Circuits (ICs), two main challenges in IC operation have emerged: self-heating effects [1]-[11], and interconnect and input/output (I/O) delays. Here, we concentrate on the self-heating effects. (The I/O delays can partially be reduced using three-dimensional ICs or Systems-on-Chips (SoC). However, this further exacerbates the self-heating problem.)

To characterize ICs, we start from individual devices that together form the IC. This first requires the characterization of device performances, and next the resolution of their effects on the overall IC performance. Considering device and IC levels, we develop methodologies to find thermal maps of planar and three-dimensional ICs at the resolution of a single device. Resolving IC thermal effects at the device level has been done for the first time by us. Previously, it had been done using several thermal nodes and not self-consistently for the entire IC, instead for hundreds of million nodes like we do. Also, we self-consistently resolve IC self-heating effects in conjunction with device operation, IC applications and layout

details for the first time.

We first obtain device performance details for a representative device utilized for the given technology. Once we determine device performance as a function of temperature, we go to the IC level, which has dimensions millions of times bigger than those of a single device. To facilitate quick computation of the IC temperature map, we develop lumped thermal models for the whole chip. This gives a resistive-capacitive thermal network with nodes that correspond to individual devices. Also, thermal resistances and capacitances between each device is determined by the chip layout. Then, we extend the device performance details calculated for a single device to the entire chip volume using IC-wide operation details such as clock frequency, and software application details such as how frequently accesses are granted for the cache, how often arithmetic manipulations are executed at the arithmetic logic unit, etc. We do this using a Monte Carlo type methodology that determines relative power density for each device on the IC. Lastly, we iterate between the IC and the device levels until we obtain the IC thermal map. We explain the coupled algorithm in detail in the following chapters.

1.5 Thesis Overview

In the following chapter, we explain our modified drift-diffusion models. We show the methodologies we developed to resolve quantum, heterostructure and self-heating effects within the drift-diffusion model that includes the Poisson, and the electron and hole current continuity equations. We first describe the drift-diffusion

equations and the auxiliary parameters such as mobilities and current densities used in conjunction with the state variables; electrostatic potential, and electron and hole concentrations. Next, we briefly talk about the discretization method we employ. This is followed by a derivation of the methodology that shows how we incorporate the aforementioned effects. Specifically, we start from the fundamental electron current density equation, and write electron concentration in terms of the electrostatic potential and the electron Fermi potential. Next, we transform this expression in such a way that the aforementioned effects can be resolved within our discretization scheme using space dependent effective potential terms in addition to the electrostatic potential we calculate. Details concerning the derivation and utilization of the modified equations to obtain device performance details will be explained later.

In Chapter 3, we develop methodologies to obtain CNT electrical properties including mobility and intrinsic carrier concentration. To obtain an expression for the mobility, we first calculate the CNT energy dispersion curve from that of graphene using zone-folding effects. Next, we import these energy dispersion curves to the Monte Carlo (MC) simulator we developed. Briefly, we explain how we determine scattering rates and CNT electron transport. Also, we include length effects on electron transport by modifying the energy dispersion curves. Next, using our MC simulations, we obtain drift-velocity versus field curves. In addition, we report position-dependent velocity oscillations and length effects in semiconducting single-walled zig-zag carbon nanotubes for the first time. Our calculated results indicate velocity oscillations with Terahertz frequencies approaching phonon fre-

quencies. This may facilitate new high frequency RF device and circuit designs, opening new paradigms in communication networks. Furthermore, to obtain comparative device performance figures between traditional devices and the ones that embed CNTs in their channels, we extract pertinent CNT electrical parameters such as intrinsic carrier concentrations and electron affinities to be used in our semiconductor equations.

In Chapter 4, we first show how incorporating CNTs within devices modifies the governing semiconductor equations. After we import CNT parameters from the previous chapter into our device simulator, we determine interactions between the CNT and Si. More specifically, we resolve quantization and transport effects on the tube and Si. Also, we resolve CNT-Si barrier effects using a density gradient quantum treatment. Resolving the effects of CNTs on quantization and transport within devices such as a MOSFET using a modified drift-diffusion model is first done by us. Next, we compare the performances of a conventional MOSFET device and our hypothetical CNT embedded MOSFETs. We predict that the CNT-MOSFET yields a better performance than the traditional MOSFET. Especially, the CNT-MOSFETs employing lower diameter tubes exhibit improved performance capabilities. Then, we do similar analyses for CNT embedded SOI-MOSFETs for the first time for the given layouts. We find that among devices that have constant film thickness, small diameter-CNT devices yield higher transconductance. On the other hand, for devices with one layer of CNTs and film thickness equal to the CNT diameter, large diameter-CNT devices show higher transconductance.

Lastly in Chapter 5, we report novel methods for predicting the thermal profile

of complex integrated circuits (ICs) at the resolution of a single device. We explain how our technique resolves device and IC temperatures self-consistently with device performances using IC layout and running application details. We iterate between the device level, which we use to calculate performance and generated heat details, and the IC level, which we solve for the temperature. Since the IC level is millions of times bigger than the device level in terms of dimensions, we develop a lumped thermal model for the IC. This facilitates finding self-consistent IC temperature and device performance figures, which have been reported for the first time. In Chapter 5, we explain our mixed-mode algorithm for two-dimensional and three-dimensional ICs. Additionally, we show chips we had designed and fabricated through MOSIS for experimental investigations.

Chapter 2

Device Modeling

We developed simulators to obtain performance figures for devices such as silicon-MOSFETs, Silicon-On-Insulator (SOI) MOSFETs, and Carbon Nanotube (CNT) embedded MOSFETs and SOI-MOSFETs, shown in Figs. 2.1, 2.4, 2.3 and 2.2, respectively. To accurately model these devices, we introduced corrections due to confinement, barrier effects and self-heating. We developed models and numerical simulators that self-consistently solve the drift-diffusion equations, including the aforementioned effects. We introduced quantum corrections using two different methods. The first method is to solve the Schrödinger equation in addition to the drift-diffusion equations to obtain eigenwaves. Once eigenwaves are calculated, a quantum electron density equation is employed for closure and feedback to the drift-diffusion model. The second method is to utilize the density gradient formalism in the drift-diffusion framework. To add the calculated quantum corrections, numerical methods were developed and successfully implemented. Also, we developed models and novel numerical techniques that resolve heterostructure and self-heating effects. In the following sections, we explain how we achieve the embedding of these corrections.

To determine operational details of devices such as MOSFETs and SOI MOSFETs, we solve the coupled semiconductor performance equations. Our semicon-

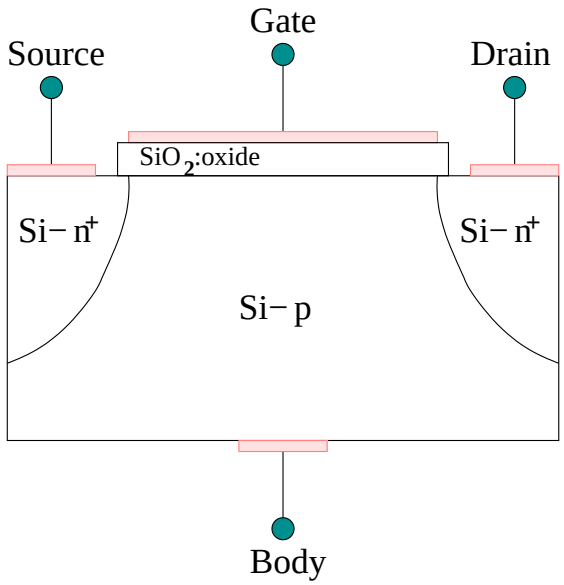

Figure 2.1: A silicon n-MOSFET.

ductor equations include those of the drift-diffusion model: Poisson equation 1.23 and carrier continuity equation 1.11 for electrons and holes. In this model, a change in carrier concentration in time is caused by a spatial change in carrier flux (From now on, all gradients are in space unless indicated otherwise by a subscript) and generation-recombination. Details related to the drift-diffusion equations and numerical algorithms to solve them self-consistently will be explained in the next section.

Also, we resolve quantum effects that greatly affect submicron device characteristics. Depending on gate bias voltages, a potential well forms at the silicon-silicon dioxide (Si-SiO_2) interface in the channels of submicron MOSFETs shown in Fig. 2.1. This well pushes electrons away from the interface, contradicting the classical assumption that electron concentration peaks where the field is highest. Since this can greatly affect device performance, we include quantum effects by ei-

ther adding the single particle Schrödinger equation 1.31, which resolves effects of applied and built-in fields on electron concentration, to our list of equations, or by making use of the density gradient formalism. Both will also be explained more in the following sections.

2.1 Drift-Diffusion Model

2.1.1 Drift-Diffusion Equations

We list below the drift-diffusion equations starting with the Poisson equation followed by the electron current continuity and hole current continuity equations.

$$\nabla^2\phi = -\frac{\rho}{\epsilon} \tag{2.1}$$

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla \cdot \vec{J}_n + G_n - R_n \tag{2.2}$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\nabla \cdot \vec{J}_p + G_p - R_p \tag{2.3}$$

Here, we have three equations with three unknowns: Electrostatic potential ϕ, electron concentration n and hole concentration p. (The other parameters in Eqns. 2.1-2.3, which are ρ, J_n, J_p, G_n and G_p, can be written in terms of the state variables.) The Poisson equation given in Eqn. 2.1 relates the net charge density ρ scaled by the dielectric constant ϵ (Generally, we use the frequency-independent dielectric constant of our medium silicon, which is 11.7 times the dielectric constant or permittivity of vacuum ϵ_o) to the gradient of the electric field ($\nabla \cdot \vec{E} = -\nabla^2\phi$). Next, the electron current continuity equation given in Eqn. 2.2 shows that a change in electron concentration n in time t in a given volume is equal to the divergence

of electron flux (J_n) and the net electron generation-recombination ("sources" and "sinks") through traps and impurities in that volume. Likewise, the hole current continuity equation given in Eqn. 2.3 shows that a change in hole concentration p in time t in a given volume is equal to the negative divergence (due to positive charge associated with holes) of hole flux (J_p) and the net hole generation-recombination through traps and impurities in that volume.

In addition to Eqns. 2.1-2.3, we need expressions for the electron current density J_n, hole current density J_p, net charge density ρ and the net generation-recombination $G-R$ in terms of the electrostatic potential ϕ, electron concentration n and hole concentration p, to have coupling between equations, and closure.

First, net charge concentration ρ can be written in terms of the electron and hole concentrations as follows:

$$\rho = q(p - n + D + T) \tag{2.4}$$

Above, q is the electronic charge, D is the net ionized donor/acceptor concentration ($D = N_D^+ - N_A^-$), and T is the net trap density. At room temperature, we can assume that all dopants are ionized; therefore, $D = N_{D_o}^+ - N_{A_o}^-$, which can be calculated from process details that give $N_{D_o}^+$ and $N_{A_o}^-$. More specifically, spatial variation of dopant concentrations can be determined experimentally or using simulation tools like Tsupreme [44]. This assumption holds unless we go to cryogenic temperatures, where ionization is adversely affected. Furthermore, we might have trapped charges T in the oxide or Si-SiO_2 interface. However, in this study, we do not consider any trapped charges ($T = 0$).

Next, we introduce electron and hole current densities, respectively.

$$\vec{J_n} = -q\mu_n n\vec{\nabla}\phi + qD_n\vec{\nabla}n \tag{2.5}$$

$$\vec{J_p} = -q\mu_p p\vec{\nabla}\phi - qD_p\vec{\nabla}p \tag{2.6}$$

The newly introduced variables are electron (hole) mobility μ_n (μ_p) and electron (hole) diffusion constant D_n (D_p). Moreover, the first terms on the right-hand-sides of Eqns. 2.5 and 2.6 are the drift components of electron and hole currents, respectively. They account for the carrier acceleration under applied and built-in fields $\vec{E} = -\vec{\nabla}\phi$. Comparing this component to the drift component of the current or particle flux derived from the zeroth moment of the BTE given in Eqn. 1.11, the following definition for mobility that relates it to the applied field and drift velocity can be derived.

$$\vec{v_n} = \mu_n\vec{E} \tag{2.7}$$

$$\vec{v_p} = \mu_p\vec{E} \tag{2.8}$$

Additionally, the second terms on the right-hand-sides of Eqns. 2.5 and 2.6 are the diffusion components of electron and hole currents, respectively. They have opposite signs due to negative and positive charges associated with electrons and holes, respectively. Furthermore, an approximate definition for the carrier diffusion constant D is given in Eqn. 1.10, which relates it to the average ensemble carrier transit length $\langle \text{Ł}_n \rangle$ and time $\langle \tau_n \rangle$, as shown below.

$$D_n \cong \frac{\langle \text{Ł}_n \rangle^2}{\langle \tau_n \rangle} \tag{2.9}$$

$$D_p \cong \frac{\langle \text{Ł}_p \rangle^2}{\langle \tau_p \rangle} \tag{2.10}$$

Lastly, we define the net generation-recombination rate $G-R$ in Eqns. 2.2-2.3, using Shockley-Hall-Read (SHR) recombination. (At the output terminal boundaries, we have infinite supply of carriers and instantaneous carrier recombinations.)

$$G - R = \frac{pn - n_o^2}{\tau_n(p + n_o) + \tau_p(n + n_o)} \tag{2.11}$$

SHR recombination accounts for the generation and recombination through localized states and centers such as ionized dopant traps within the bandgap. This recombination process accounts for the spontaneous electron and hole capture and emission. Electrons from the conduction and holes from the valence band are captured or emitted by traps such as acceptor-like recombination centers of ionized donor dopants and donor-like recombination centers of ionized acceptor dopants, respectively. Under low-level injection, the net electron emission-capture rate is equal to the net hole emission-capture rate, resulting in a detailed version of the above equation, which simplifies to Eqn. 2.11 knowing that the biggest contribution of trap assisted recombination will come from the traps near the mid-bandgap.

Now, we have the drift-diffusion equations 2.1-2.3 along with the auxiliary relations 2.4-2.11 that provide coupling and closure. In addition, we have supplementary relations that give electron and hole concentrations in terms of the electrostatic potential and Fermi levels. To derive those relations, we start from the Fermi-Dirac distribution. Basically, it gives the probability of a state being occupied by an electron that is subject to non-degenerate statistics and the Pauli exclusion principle, as defined below:

$$f(E) = \frac{1}{1 + e^{\left(\frac{E - E_f}{kT}\right)}} \theta(E_g) \tag{2.12}$$

Above, E is the electron energy ($= -q\phi$), and E_f is called the Fermi level or the Fermi energy. Moreover, $f(E)$ is zero within the bandgap through the step function $\theta(E_g)$, which is 1 within the conduction and valence bands, and 0 in between. Additionally, $f(E)$ is related to temperature T and Boltzmann's constant k.

Using non-degenerate statistics, electron (hole) concentration in the conduction (valence) band can be calculated using the probability of having an electron (hole) in a state $f(E)$ $(1 - f(E))$ and the density-of-states $g(E)$ in the conduction (valence) band, as follows:

$$n = \int_{E_c}^{\infty} f(E)g(E)dE \tag{2.13}$$

Taking $f(E) \cong e^{\left(-\frac{E-E_f}{kT}\right)}$ (Maxwell-Boltzmann distribution) and $g(E) = 4\pi\left(\frac{2m^*}{h^2}\right)^{(3/2)}$ (considering three-dimensional space, parabolic energy dispersion curves, double degeneracy and the number of available states in k space in three-dimensions), we find that:

$$n = N_c e^{\left(-\frac{(E_c-E_f)}{kT}\right)} \tag{2.14}$$

$$p = N_v e^{\left(-\frac{(E_f-E_v)}{kT}\right)}, \tag{2.15}$$

where

$$N_c = 2\left(\frac{2\pi m_n^* kT}{h^2}\right)^{3/2} \tag{2.16}$$

$$N_v = 2\left(\frac{2\pi m_p^* kT}{h^2}\right)^{3/2}. \tag{2.17}$$

In thermal equilibrium for undoped materials, the Fermi level is at mid-bandgap ($E_f = E_i$) resulting in hole and electron concentrations that both are equal to n_o, which is called the intrinsic carrier concentration. Using n_o and assum-

ing that there are two different Fermi levels at nonequilibrium (E_{f_n} and E_{f_p} are for electrons and holes, respectively), carrier concentrations can be written as shown below:

$$n = n_o e^{\left(\frac{E_{f_n} - E_i}{kT}\right)} \tag{2.18}$$

$$p = n_o e^{\left(\frac{E_i - E_{f_p}}{kT}\right)} \tag{2.19}$$

Since the electrostatic potential is with respect to a reference level, we take it as $\phi = -\frac{E_i}{q}$. Therefore, we can rewrite the carrier concentrations in terms of ϕ (in reference to mid-bandgap), quasi-electron Fermi potential $\phi_n = -\frac{E_{f_n}}{q}$, quasi-hole Fermi potential $\phi_p = -\frac{E_{f_p}}{q}$ and thermal voltage $V_{\mathrm{TH}} = \frac{kT}{q}$, as follows:

$$n = n_o e^{\left(\frac{\phi - \phi_n}{V_{\mathrm{TH}}}\right)} \tag{2.20}$$

$$p = n_o e^{\left(\frac{\phi_p - \phi}{V_{\mathrm{TH}}}\right)} \tag{2.21}$$

Next, using the Einstein relation $\frac{D}{\mu} = \frac{kT}{q}$, we can rewrite current densities in a compact form:

$$\vec{J_n} = -q\mu_n n \nabla\vec{\phi}_n \tag{2.22}$$

$$\vec{J_p} = -q\mu_p p \nabla\vec{\phi}_p \tag{2.23}$$

In the following sections, we elaborate on our numerical techniques used to solve the drift-diffusion equations. Later, we derive our modified drift-diffusion equations. During that derivation, we will make frequent use of the above equations as our fundamental current density equations.

2.1.2 Discretized Drift-Diffusion Equations

In this work, we use the finite difference method, in which the solution domain is discretized using rectangles, and derivatives are approximated by differences. Our equations 2.1-2.3 include first and second order derivatives that can be discretized in 1-D with the spacing $x_{i+1} - x_i$, ignoring third and higher order expansion terms:

$$\left.\frac{\partial^2 \zeta}{\partial x^2}\right|_{x_i} = \frac{\left.\frac{\partial \zeta}{\partial x}\right|_{x_{i+\frac{1}{2}}} - \left.\frac{\partial \zeta}{\partial x}\right|_{x_{i-\frac{1}{2}}}}{\frac{(x_{i+1}-x_{i-1})}{2}} \tag{2.24}$$

$$= \frac{2}{(x_{i+1}-x_{i-1})}\left[\frac{\zeta(x_{i+1})-\zeta(x_i)}{(x_{i+1}-x_i)} - \frac{\zeta(x_i)-\zeta(x_{i-1})}{(x_i-x_{i-1})}\right] \tag{2.25}$$

$$\left.\frac{\partial \zeta}{\partial x}\right|_{x_i} = \frac{(x_i - x_{i-1})\left.\frac{\partial \zeta}{\partial x}\right|_{x_{i+\frac{1}{2}}} + (x_{i+1}-x_i)\left.\frac{\partial \zeta}{\partial x}\right|_{x_{i-\frac{1}{2}}}}{(x_{i+1}-x_{i-1})} \tag{2.26}$$

$$= \frac{1}{(x_{i+1}-x_{i-1})}\left[(\zeta(x_{i+1})-\zeta(x_i))\frac{(x_i-x_{i-1})}{(x_{i+1}-x_i)} + (\zeta(x_i)-\zeta(x_{i-1}))\frac{(x_{i+1}-x_i)}{(x_i-x_{i-1})}\right] \tag{2.27}$$

A simplified version of the first derivative can also be used instead of the discretized form given above.

$$\left.\frac{\partial \zeta}{\partial x}\right|_{x_i} = \frac{\zeta(x_{i+1})-\zeta(x_{i-1})}{x_{i+1}-x_{i-1}} \tag{2.28}$$

Using the aforementioned approximation for the second derivative, Eqn. 2.4 for the net charge density, and Eqns. 2.20-2.21 for the electron and hole concentrations, we discretize the Poisson equation 2.1 in two-dimensions, as written below.

$$\left(\frac{2}{h_i + h_{i-1}}\right)\left[\frac{\phi_{i+1,j}-\phi_{i,j}}{h_i} + \frac{\phi_{i-1,j}-\phi_{i,j}}{h_{i-1}}\right] + \left(\frac{2}{k_j + k_{j-1}}\right)\left[\frac{\phi_{i,j+1}-\phi_{i,j}}{k_j} + \frac{\phi_{i,j-1}-\phi_{i,j}}{k_{j-1}}\right] = \frac{q}{\epsilon_{\mathrm{Si}}}\left[n_0\left(e^{\left(\frac{\phi_{i,j}-\phi_{n_{i,j}}}{V_{\mathrm{TH}}}\right)} - e^{-\left(\frac{\phi_{i,j}-\phi_{p_{i,j}}}{V_{\mathrm{TH}}}\right)}\right) - D_{i,j} - T_{i,j}\right] \tag{2.29}$$

Here, subscripts i and j are for x and y coordinates, respectively. We denote spacing in x and y directions by $h_i = x_{i+1} - x_i$ and $k_j = y_{j+1} - y_j$, respectively.

Direct discretization of the electron and hole current continuity equations 2.2 and 2.3 leads to ill-conditioned Jacobian matrices, and ignores the fact that n and p vary exponentially with ϕ between mesh points. Hence, we use the Scharfetter-Gummel discretization scheme that overcomes these problems, by approximating the current densities at half grid points using Bernoulli functions ($B(x) = \frac{x}{e^x - 1}$), as follows [45].

$$\frac{\vec{J}_{n_{i+\frac{1}{2}}}}{q\mu_{n_{i+\frac{1}{2}}} V_{\mathrm{TH}}} = \left[B\left(\frac{\vec{E}_{i+\frac{1}{2}}(x_i - x_{i+1})}{V_{\mathrm{TH}}} \right) \left(\frac{n_{i+1}}{x_{i+1} - x_i} \right) - B\left(\frac{\vec{E}_{i+\frac{1}{2}}(x_{i+1} - x_i)}{V_{\mathrm{TH}}} \right) \left(\frac{n_i}{x_{i+1} - x_i} \right) \right] \tag{2.30}$$

$$\frac{\vec{J}_{n_{i-\frac{1}{2}}}}{q\mu_{n_{i-\frac{1}{2}}} V_{\mathrm{TH}}} = \left[B\left(\frac{\vec{E}_{i-\frac{1}{2}}(x_{i-1} - x_i)}{V_{\mathrm{TH}}} \right) \left(\frac{n_i}{x_i - x_{i-1}} \right) - B\left(\frac{\vec{E}_{i-\frac{1}{2}}(x_i - x_{i-1})}{V_{\mathrm{TH}}} \right) \left(\frac{n_{i-1}}{x_i - x_{i-1}} \right) \right] \tag{2.31}$$

Therefore, electron and hole current continuity equations 2.2 and 2.3 can be written in discretized form as shown next.

$$\begin{aligned} \frac{\partial n_{i,j}}{\partial t} &= -R_{n_{i,j}} + G_{n_{i,j}} \\ &+ \left(\frac{2V_{\mathrm{TH}}}{h_i + h_{i-1}} \right) \left[\left(n_{i+1,j} \mu_{n_{i+\frac{1}{2},j}} B\left(-\frac{\phi_{i+1,j} - \phi_{i,j}}{V_{\mathrm{TH}}} \right) \right) + \left(n_{i-1,j} \mu_{n_{i-\frac{1}{2},j}} B\left(-\frac{\phi_{i,j} - \phi_{i-1,j}}{V_{\mathrm{TH}}} \right) \right) \right] \\ &- \left(\frac{2n_{i,j} V_{\mathrm{TH}}}{h_i + h_{i-1}} \right) \left[\left(\mu_{n_{i+\frac{1}{2},j}} B\left(\frac{\phi_{i+1,j} - \phi_{i,j}}{V_{\mathrm{TH}}} \right) \right) + \left(\mu_{n_{i-\frac{1}{2},j}} B\left(\frac{\phi_{i,j} - \phi_{i-1,j}}{V_{\mathrm{TH}}} \right) \right) \right] \\ &+ \left(\frac{2V_{\mathrm{TH}}}{k_j + k_{j-1}} \right) \left[\left(n_{i,j+1} \mu_{n_{i,j+\frac{1}{2}}} B\left(-\frac{\phi_{i,j+1} - \phi_{i,j}}{V_{\mathrm{TH}}} \right) \right) + \left(n_{i,j-1} \mu_{n_{i,j-\frac{1}{2}}} B\left(-\frac{\phi_{i,j} - \phi_{i,j-1}}{V_{\mathrm{TH}}} \right) \right) \right] \\ &- \left(\frac{2n_{i,j} V_{\mathrm{TH}}}{k_j + k_{j-1}} \right) \left[\left(\mu_{n_{i,j+\frac{1}{2}}} B\left(\frac{\phi_{i,j+1} - \phi_{i,j}}{V_{\mathrm{TH}}} \right) \right) + \left(\mu_{n_{i,j-\frac{1}{2}}} B\left(\frac{\phi_{i,j} - \phi_{i,j-1}}{V_{\mathrm{TH}}} \right) \right) \right] \end{aligned} \tag{2.32}$$

$$
\begin{aligned}
\frac{\partial p_{i,j}}{\partial t} &= -R_{p_{i,j}} + G_{p_{i,j}} \\
&+ \left(\tfrac{2V_{\mathrm{TH}}}{h_i+h_{i-1}}\right)\left[\left(p_{i+1,j}\mu_{p_{i+\frac{1}{2},j}} B\left(-\tfrac{\phi_{i+1,j}-\phi_{i,j}}{V_{\mathrm{TH}}}\right)\right) + \left(p_{i-1,j}\mu_{p_{i-\frac{1}{2},j}} B\left(-\tfrac{\phi_{i,j}-\phi_{i-1,j}}{V_{\mathrm{TH}}}\right)\right)\right] \\
&\quad - \left(\tfrac{2p_{i,j}V_{\mathrm{TH}}}{h_i+h_{i-1}}\right)\left[\left(\mu_{p_{i+\frac{1}{2},j}} B\left(\tfrac{\phi_{i+1,j}-\phi_{i,j}}{V_{\mathrm{TH}}}\right)\right) + \left(\mu_{p_{i-\frac{1}{2},j}} B\left(\tfrac{\phi_{i,j}-\phi_{i-1,j}}{V_{\mathrm{TH}}}\right)\right)\right] \\
&+ \left(\tfrac{2V_{\mathrm{TH}}}{k_j+k_{j-1}}\right)\left[\left(p_{i,j+1}\mu_{p_{i,j+\frac{1}{2}}} B\left(-\tfrac{\phi_{i,j+1}-\phi_{i,j}}{V_{\mathrm{TH}}}\right)\right) + \left(p_{i,j-1}\mu_{p_{i,j-\frac{1}{2}}} B\left(-\tfrac{\phi_{i,j}-\phi_{i,j-1}}{V_{\mathrm{TH}}}\right)\right)\right] \\
&\quad - \left(\tfrac{2p_{i,j}V_{\mathrm{TH}}}{k_j+k_{j-1}}\right)\left[\left(\mu_{p_{i,j+\frac{1}{2}}} B\left(\tfrac{\phi_{i,j+1}-\phi_{i,j}}{V_{\mathrm{TH}}}\right)\right) + \left(\mu_{p_{i,j-\frac{1}{2}}} B\left(\tfrac{\phi_{i,j}-\phi_{i,j-1}}{V_{\mathrm{TH}}}\right)\right)\right] \qquad (2.33)
\end{aligned}
$$

We first generate a rectangular 2-D mesh on a device such as the one shown in Fig. 2.1. We then solve the coupled drift-diffusion equations on this mesh, using the Gauss-Seidel method, where the value of a given variable is explicitly written using the current values of the neighbors. We use this method to get a good guess for the real solution without diverging between iterations. Next, we setup the Jacobian matrix for the whole system, and then update the values of our variables adding corrections calculated using the Newton-Raphson method. This method works well once we have a good initial estimate; otherwise, our solution diverges. Finally, we obtain a self-consistent solution for the electrostatic potential ϕ, electron carrier concentration n and hole carrier concentration p that satisfies all our equations.

2.2 Quantum Corrected Drift-Diffusion Model

2.2.1 Solving for the Single Particle Schröndinger Equation

We developed a device simulator that is capable of solving the coupled quantum and semiconductor equations. To add quantum corrections to the calculation of the electron density in MOSFET and SOI-MOSFET channels, we solve the

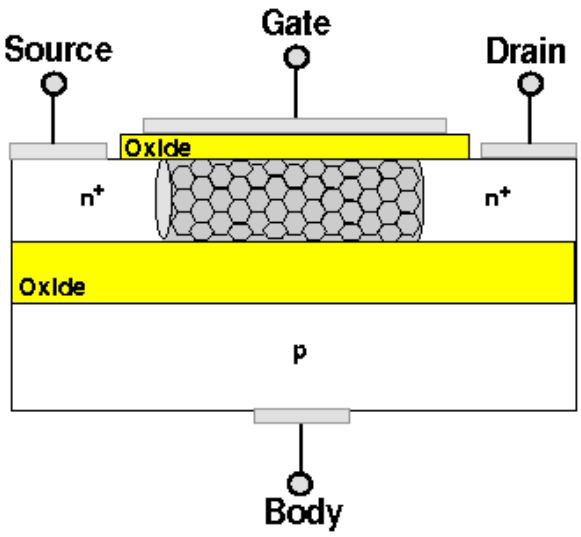

Figure 2.2: A CNT embedded SOI-MOSFET.

Schrödinger equation to determine band splitting. (Confinement effects are especially strong in CNT embedded SOI-MOSFETs shown in Fig. 2.2.) Below is a set of our quantum semiconductor equations in the order of the Schrödinger, density, Poisson, electron current continuity and hole current continuity equations [7], [8], [46]-[49].

$$E\psi(y) = -\frac{\hbar^2}{2m^*}\frac{d^2}{dy^2}\psi(y) - q\phi(x,y)\psi(y) \tag{2.34}$$

$$n = \frac{m^* kT}{\pi\hbar^2}\sum_i |\psi_i|^2 \ln\left(1 + e^{\frac{E_f - E_i}{kT}}\right) \tag{2.35}$$

$$\nabla^2\phi = -\frac{\rho}{\epsilon} \tag{2.36}$$

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla\cdot\vec{J}_n + G_n - R_n \tag{2.37}$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\nabla\cdot\vec{J}_p + G_p - R_p \tag{2.38}$$

In addition to the previously introduced parameters, E_i, E_f and ψ_i are the sub-band energies, Fermi level and wave functions, respectively. Furthermore, we solve for five unknows; $|\psi_i|$ (or E_i, which are linearly dependent on each other), E_f,

ϕ, n and p using Eqns. 2.34-2.38 in the order they are listed above.

Our list of quantum-corrected drift-diffusion equations includes the Schrödinger equation 1.31, which we present below again as a reminder.

$$\left[-\frac{\hbar^2}{2m^*}\nabla_r^2 + E_C(r)\right]\psi_3 = \varepsilon_3\psi_3 \tag{2.39}$$

Here $E_C(r)$ accounts for the external and built-in potentials. Since electrostatic potential ϕ resolves effects of those fields, we replace $E_C(r)$ by $-q\phi(x,y)$. Additionally, we solve one-dimensional Schrödinger equations along the MOSFET channel starting from the Si-SiO_2 interface and going down the substrate. Therefore, we replace ∇_r by $\frac{\partial}{\partial y}$. Moreover, to ascertain average characteristics due to transport in different energy valleys of silicon, we use the harmonic mean of the effective masses in different valleys, resulting in $\frac{1}{m^*} = \frac{1}{3}\left(\frac{1}{m_t} + \frac{1}{m_t} + \frac{1}{m_l}\right)$ to be used in Eqn. 2.34. Subscripts t and l refer to the transverse and longitudinal directions, respectively. In addition, m_t=0.19m_o and m_l=0.9m_o for the Si, where m_o is the free electron mass.

We calculate the quantum electron concentration using the density (or what we also call the population) equation 2.35. We derive this equation using the two-dimensional density of states associated with the parabolic bands, and the Fermi-Dirac statistics, as shown below.

$$n = \int_{E_i}^{\infty}\sum_i \frac{m^*}{\pi\hbar^2}|\psi_i|^2\left(\frac{1}{1+e^{\frac{E-E_f}{kT}}}\right)dE \tag{2.40}$$

Since the sum does not affect the argument of the integral, we can first inte-

grate and then do the sum, as follows.

$$n = \frac{m^*}{\pi\hbar^2} \sum_i |\psi_i|^2 \int_{E_i}^{\infty} \left(\frac{1}{1 + e^{\frac{E - E_f}{kT}}} \right) dE \tag{2.41}$$

This gives the density equation 2.35 once the above integral is evaluated.

$$n = \frac{m^* kT}{\pi\hbar^2} \sum_i |\psi_i|^2 \ln \left(1 + e^{\frac{E_i - E_f}{kT}} \right) \tag{2.42}$$

We first solve the system classically using Eqns. 2.36-2.38. At this stage, we do not include the effects of quantum confinement in the channel. The carrier concentrations are determined by the solution of the continuity equations through the use of the Scharfetter-Gummel discretization scheme. Once a classical solution for the device is obtained, we add the effects of quantum confinement in the MOSFET channel by solving Eqns. 2.34-2.35 in addition to Eqns. 2.36-2.38. The potential well introduced at the Si-SiO_2 interface would change the current drive and carrier concentration considerably, especially for submicron devices. Thus, we incorporate the effects of quantum confinement in the channel. The carrier concentration lowering due to quantum confinement can be thought as bandgap broadening. This fact in turn modifies the local intrinsic carrier concentration. Thus, it introduces an additional spatial dependency on the intrinsic carrier concentration, which can be treated using an effective potential in current densities. This can be better understood by examining the electron current equation written below (the same is also true for the hole current continuity equation):

$$J_n = -qn\mu_n \nabla \phi_n \tag{2.43}$$

Here, ϕ_n is the electron quasi-fermi level. It can be expanded in terms of the

electrostatic potential, ϕ, electron carrier concentration, n, intrinsic carrier concentration, n_o, and the thermal voltage, V_{TH}, as follows:

$$\phi_n = \phi - V_{\mathrm{TH}} \ln \frac{n}{n_o} \tag{2.44}$$

For a homostructure without confinement, n_o is constant; however, n_o differs due to confinement. Multiplying both numerator and denominator of the logarithm in Eqn. 2.44 by the bulk n_o, which is the silicon's n_o (n_o^{Si}), we get the following expression for the ϕ_n:

$$\begin{aligned} \phi_n &= \phi - V_{\mathrm{TH}} \ln \left(\frac{n_o^{\mathrm{Si}}}{n_o^{\mathrm{Si}}} \frac{n}{n_o} \right) && (2.45) \\ &= \phi + V_{\mathrm{TH}} \ln \frac{n_o}{n_o^{\mathrm{Si}}} - V_{\mathrm{TH}} \ln \frac{n}{n_o^{\mathrm{Si}}} && (2.46) \end{aligned}$$

We then substitute ϕ_n back into Eqn. 2.43. This introduces an offset next to the intrinsic potential ϕ inside the gradient term, which is equal to zero outside the MOSFET channel. We below show these steps mathematically.

$$\begin{aligned} J_n &= -qn\mu_n \nabla \left(\phi + V_{\mathrm{TH}} \ln \frac{n_o}{n_o^{\mathrm{Si}}} - V_{\mathrm{TH}} \ln \frac{n}{n_o^{\mathrm{Si}}} \right) && (2.47) \\ &= -qn\mu_n \nabla \left(\phi + V_{\mathrm{TH}} \ln \frac{n_o}{n_o^{\mathrm{Si}}} \right) + qn\mu_n V_{\mathrm{TH}} \frac{\nabla n / n_o^{\mathrm{Si}}}{n / n_o^{\mathrm{Si}}} && (2.48) \end{aligned}$$

Next, we write the resulting current equation as follows:

$$J_n = -qn\mu_n \nabla \left(\phi + V_{\mathrm{TH}} \ln \frac{n_{\mathrm{QM}}}{n_{\mathrm{CL}}} \right) + q\mu_n V_{\mathrm{TH}} \nabla n \tag{2.49}$$

Here, n_{QM} is the quantum corrected electron concentration calculated using the discrete energies and corresponding occupation probabilities for the channel. On the other hand, the reference level n_{CL} is the classical solution simulated before.

Their ratio is equal to the ratio of the space-dependent (due to confinement) and the bulk intrinsic carrier concentrations. Thus, this offset takes care of the spatial dependency of the intrinsic carrier concentration and resolves the confinement effects using the drift-diffusion model.

2.2.2 Resolving Quantum Effects Using Density Gradient Formalism

We developed a device simulator that is capable of solving the semiconductor equations with quantum corrections. To add quantum corrections to the calculation of the electron density in the MOSFET channel, we use a quantum effective potential in addition to the electrostatic potential. Next, we show our set of our quantum semiconductor equations in the order of the Poisson, electron current continuity and hole current continuity equations [7], which is followed by the quantum effective potential term [46].

$$\nabla^2\phi = -\frac{\rho}{\epsilon} \tag{2.50}$$

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla \cdot \vec{J_n} + G_n - R_n \tag{2.51}$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\nabla \cdot \vec{J_p} + G_p - R_p \tag{2.52}$$

$$\phi_{\mathrm{QM}} = \frac{2\hbar^2}{12q\sqrt{n}}\left[\frac{1}{m_{||}}\frac{\partial^2\sqrt{n}}{\partial x^2} + \frac{1}{m_{\perp}}\frac{\partial^2\sqrt{n}}{\partial y^2}\right] \tag{2.53}$$

Investigations show that carrier confinement at the Si-SiO_2 interface can significantly reduce the carrier concentration adjacent to the interface [50]-[55]. In addition, there can be band-to-band and source-to-drain tunneling effects in ultra short channel devices. To incorporate these quantum effects in our device model, we use a density gradient formalism. The density gradient theory is based on an

approximate many-body quantum theory [51]. It has been shown that the density gradient theory resolves the effects of the MOSFET channel confinement [52, 53], band-to-band and source-to-drain tunneling [53, 55]. In this formalism, quantum effects are included by the introduction of an effective potential term that is proportional to the second derivative of the square root of the electron density normalized by the square root of the electron density, as shown in Eqn. 2.53 [51]-[59]. Here x is parallel to the MOSFET channel, and y is normal to x. We also use direction and location dependent effective masses if data are provided.

As described in the previous section, we treat the quantum induced effects in a manner that is analogous to the formation of position dependent heterostructures in the quantum well. Thus, we sum quantum effects in the carrier concentration term, where quantum confinement is reflected as bandgap broadening or lowering. This methodology introduces an extra potential term as follows:

$$\frac{kT}{q} \ln \frac{n_{\mathrm{QM}}}{n_{\mathrm{CL}}}, \tag{2.54}$$

where

$$\frac{n_{\mathrm{QM}}}{n_{\mathrm{CL}}} = \exp\left[\frac{\hbar^2}{6kT\sqrt{n}}\left(\frac{1}{m_{||}}\frac{\partial^2\sqrt{n}}{\partial x^2} + \frac{1}{m_{\perp}}\frac{\partial^2\sqrt{n}}{\partial y^2}\right)\right]. \tag{2.55}$$

Here subscripts refer to quantum (QM) and classical (CL) solutions. We next incorporate the above term into the current equation to account for the quantization effects on transport.

$$J_n = -qn\mu_n\nabla(\phi + \phi_{\mathrm{QM}}) + q\mu_n V_{\mathrm{TH}}\nabla n \tag{2.56}$$

As before, we first solve Eqns. 2.50-2.52 for the classical case, where $\phi_{\mathrm{QM}} = 0$.

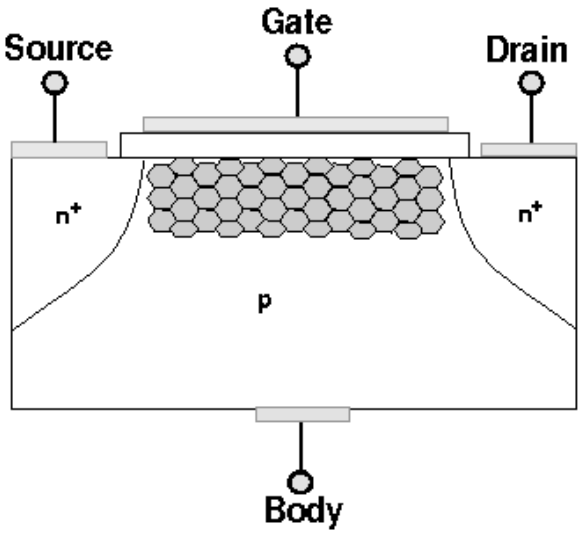

Figure 2.3: A CNT embedded silicon MOSFET.

After we obtain a solution for the classical case, we include corrections to Eqns. 2.50-2.52 calculating values for ϕ_{QM} using Eqn. 2.53.

2.3 Heterostructure Corrected Drift-Diffusion Model

We developed a device simulator to obtain performance details in devices that have other materials in addition to silicon in their active regions. To characterize the performances of these devices, we resolve barrier effects at the junctions between these materials and silicon. In addition, we resolve transport details in these materials and at the barriers. More specifically, we determine the interaction between the Si and the CNT by resolving transport and field effects self-consistently, in a CNT embedded MOSFET like the one shown in Fig. 2.3.

To find the effects of heterostructure barrier formations on transport in the MOSFET channel, we use a modified set of drift-diffusion equations. Our modifica-

tions take into account differences in bandgaps and affinities associated with the Si and the CNT. We incorporate these effects into our current equations using effective potentials similar to the ones employed to account for the quantum effects. The governing equations are listed below in order the Poisson, CNT-Si electron current continuity, and CNT-Si hole current continuity equations.

$$\nabla^2\phi = -\frac{\rho}{\epsilon} \tag{2.57}$$

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla \cdot \vec{J_n} + G_n - R_n \tag{2.58}$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\nabla \cdot \vec{J_p} + G_p - R_p \tag{2.59}$$

We next define electron and hole current densities J_n and J_p, as follows:

$$J_n = -qn\mu_n\nabla(\phi + \phi^n_{\mathrm{HS}}) + \mu_n kT\nabla n \tag{2.60}$$

$$J_p = -qp\mu_p\nabla(\phi - \phi^p_{\mathrm{HS}}) - \mu_p kT\nabla p \tag{2.61}$$

Here, ϕ_{HS} accounts for the CNT-Si barrier effects, which is composed of two parts. The first part is for resolving effects associated with the differences in intrinsic carrier concentrations due to the differences in bandgaps. The second part is for resolving the discontinuities at the conduction and valence bands due to the differences in electron affinities and bandgaps.

To obtain a form for the effective potential that accounts for the differences in intrinsic carrier concentrations, we start with the standard expression for the current as the gradient of the quasi-Fermi potential:

$$J_n = -qn\mu_n\nabla\phi_n \tag{2.62}$$

Next, we introduce the familiar relationship between the quasi-Fermi potential,

electrostatic potential, and the intrinsic carrier concentration. However, for the CNT-Si structure, the intrinsic carrier concentration, n_o, has spatial dependence.

$$\phi_n = \phi - \frac{kT}{q} \ln \frac{n}{n_o} \tag{2.63}$$

We now multiply the numerator and the denominator of the argument of the logarithm by a constant, which is equal to the intrinsic silicon carrier concentration to obtain the following expression for ϕ_n:

$$\phi_n = \left(\phi + \frac{kT}{q} \ln \frac{n_o}{n_o^{\mathrm{Si}}} \right) - \frac{kT}{q} \ln \frac{n}{n_o^{\mathrm{Si}}} \tag{2.64}$$

Substituting Eqn. 2.64 into Eqn. 2.62, we obtain the revised expression for the electron current density:

$$J_n = -qn\mu_n \nabla \left(\phi + \frac{kT}{q} \ln \frac{n_o}{n_o^{\mathrm{Si}}} \right) + \mu_n kT \nabla n \tag{2.65}$$

Here, n_o is the intrinsic carrier concentration at a grid point on our device, and n_o^{Si} is the intrinsic carrier concentration of silicon. We note that n_o takes on the intrinsic carrier concentration of either the CNT or the Si, depending on the location within the CNT-Si device. The potential within the gradient of Eqn. 2.65 includes the electrostatic potential and an effective potential due to the bandgap variations in the CNT-Si structure [50, 60]. However, CNT and Si also have different electron affinities that would change the potential barrier between these two materials. Therefore, we next introduce an additional effective potential term that arises due to the different electron affinities on both sides of the CNT-Si barrier. In Eqn. 2.60, we sum the effects of the variations of the bandstructure and the electron

affinities in $\phi^n_{\rm HS}$, which for electrons is defined as follows [60]:

$$\phi^n_{\rm HS} = \frac{1}{q}(\chi - \chi^{\rm Si}) + \frac{kT}{q}\ln\frac{n_o}{n_o^{\rm Si}} \tag{2.66}$$

Here, χ is the electron affinity at a grid point on our device and is either equal to $\chi^{\rm Si}$ or χ^{CNT}. We subtract $\chi^{\rm Si}$ from χ because our reference material is the Si, as pointed out in Eqn. 2.64. With addition of the new term that incorporates the difference in electron affinities, the electron current density becomes:

$$J_n = -qn\mu_n\nabla\left[\phi + \left(\frac{1}{q}(\chi - \chi^{\rm Si}) + \frac{kT}{q}\ln\frac{n_o}{n_o^{\rm Si}}\right)\right] + \mu_n kT\nabla n \tag{2.67}$$

We can also apply the same arguments to holes. Thus, one would find the corresponding effective potential expression for holes as written below:

$$\phi^p_{\rm HS} = -\frac{1}{q}(\chi + E_G - \chi^{\rm Si} - E_G^{\rm Si}) - \frac{kT}{q}\ln\frac{n_o}{n_o^{\rm Si}} \tag{2.68}$$

Here, bandgap E_G, like χ and n_o, refers to the same material in space. It takes on the bandgap value of either the CNT or the Si depending on the location within the CNT-Si device. We note that this formalism does not account for atomic bonding details, which could give rise to interface states and complicated junctions. These effects would likely be accounted for in the present model through the Poisson and transport equations, the Fermi level and the mobility.

2.4 Thermal Effects Included in a Drift-Diffusion Model

To incorporate self-heating effects on device performance, especially of Silicon-On-Insulator (SOI) MOSFETs shown in Fig. 2.4, we developed an efficient self-consistent method for the inclusion of self-heating effects into the drift-diffusion

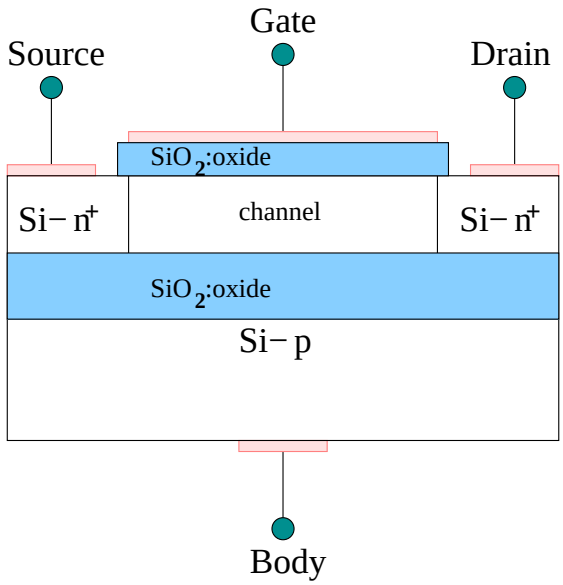

Figure 2.4: An SOI-MOSFET.

model. To resolve self-heating, we solve modified semiconductor equations. Our technique enables use of the traditional solvers for the semiconductor equations by resolving temperature effects using quasi-self-heating potentials, which can be treated similarly to those that arise due to differences in bandgap variations and work functions.

We build on our device simulator [61, 62] to solve the coupled semiconductor equations for predicting ultra-small channel thickness SOI device performances. We list our semiconductor equations from 2.69 to 2.72 in the order of the Poisson, electron current continuity, hole current continuity, and the heat flow equations.

$$\nabla^2 \phi = -\frac{\rho}{\epsilon} \tag{2.69}$$

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla \cdot \vec{J}_n + G_n - R_n \tag{2.70}$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\nabla \cdot \vec{J}_p + G_p - R_p \tag{2.71}$$

$$C\frac{\partial T}{\partial t} = \nabla \cdot \kappa \vec{\nabla} T + H \tag{2.72}$$

To achieve coupling between state variables (ϕ , n, p, T), we use the auxiliary relations 2.73, 2.74 and 2.75 for electron and hole current densities, and Joule heating (heat generated), respectively.

$$J_n = -qn\mu_n \nabla(\phi + \phi^n_{\mathrm{SH}}) + \mu_n kT\nabla n \tag{2.73}$$

$$J_p = -qp\mu_p \nabla(\phi - \phi^p_{\mathrm{SH}}) - \mu_p kT\nabla p \tag{2.74}$$

$$H = -(J_n + J_p) \cdot \nabla\phi \tag{2.75}$$

Here, we make use of the quasi-self-heating potentials ϕ^n_{SH} and ϕ^p_{SH}, which are employed as additional terms to the electrostatic potential. Next, we explain the origins of these self-heating potentials. Moreover, these quasi-potentials enable us to treat lattice heating as a mechanism that causes bandgap variations and band discontinuities.

To obtain an expression for the electron quasi-self-heating potential, we start with the familiar expression for the electron current, which states that the current is proportional to the gradient of the quasi-Fermi potential, as given below.

$$J_n = -qn\mu_n \nabla\phi_n \tag{2.76}$$

We can also express the quasi-Fermi potential in terms of the state variables ϕ and n as in Eqn. 2.77, where temperatures and intrinsic carrier concentrations are functions of space.

$$\phi_n = \phi - \frac{kT}{q} \ln \frac{n}{n_o(T)} \tag{2.77}$$

To remove the space dependent temperature term in front of the logarithm, we employ the equivalent form given in Eqn. 2.82 for the built-in potential.

$$\begin{aligned}
\frac{kT}{q}\ln\frac{n}{n_o(T)} &= \frac{kT_o}{q}\ln\left(\frac{n}{n_o(T)}\right)^{\frac{T}{T_o}} && (2.78)\\
&= \frac{kT_o}{q}\ln\left[\frac{n}{n_o(T)}\left(\frac{n}{n_o(T)}\right)^{\frac{T-T_o}{T_o}}\right] && (2.79)\\
&= \frac{kT_o}{q}\ln\frac{n}{n_o(T)} + \frac{(T-T_o)}{T_o}\left(\frac{kT_o}{q}\ln\frac{n}{n_o(T)}\right) && (2.80)\\
&= \frac{kT_o}{q}\ln\frac{n}{n_o(T)} + \frac{(T-T_o)}{T}\left(\frac{kT}{q}\ln\frac{n}{n_o(T)}\right) && (2.81)\\
&= \frac{kT_o}{q}\ln\frac{n}{n_o(T)} + \frac{(T-T_o)}{T}(\phi-\phi_n) && (2.82)
\end{aligned}$$

We next multiply the numerator and the denominator of the term inside the logarithm in Eqn. 2.82 with the intrinsic carrier concentration evaluated at room temperature.

$$\begin{aligned}
\frac{kT_o}{q}\ln\frac{n}{n_o(T)} &= \frac{kT_o}{q}\ln\left(\frac{n_o(T_o)}{n_o(T_o)}\frac{n}{n_o(T)}\right) && (2.83)\\
&= \frac{kT_o}{q}\ln\frac{n}{n_o(T_o)} + \frac{kT_o}{q}\ln\frac{n_o(T_o)}{n_o(T)} && (2.84)
\end{aligned}$$

Thus, the net built-in potential term in Eqn. 2.77 takes the form given in Eqn. 2.85.

$$\frac{kT}{q}\ln\frac{n}{n_o(T)} = \frac{kT_o}{q}\ln\frac{n}{n_o(T_o)} + \frac{kT_o}{q}\ln\frac{n_o(T_o)}{n_o(T)} + \frac{(T-T_o)}{T}(\phi-\phi_n) \qquad (2.85)$$

The first term on the right-hand-side of Eqn. 2.85 is the built-in potential without considering self-heating effects. Thus, the resulting quasi-self-heating potentials for electrons and holes are shown in Eqns. 2.86 and 2.87, respectively.

$$\phi_{\mathrm{SH}}^{n} = \frac{kT_o}{q}\ln\frac{n_o(T)}{n_o(T_o)} + \frac{(T-T_o)}{T}(\phi_n-\phi) \qquad (2.86)$$

$$\phi_{\mathrm{SH}}^{p} = \frac{kT_o}{q}\ln\frac{n_o(T)}{n_o(T_o)} + \frac{(T-T_o)}{T}(\phi-\phi_p) \qquad (2.87)$$

Comparing these expressions to those used for heterostructures, we can say that the first terms on the left-hand-side of Eqns. 2.86 and 2.87 account for the changes in built-in potential due to bandgap variations by temperature. Consequently, the second terms in those equations are analogous to terms used in heterostructures to resolve band discontinuities or differences in electron affinities.

2.5 Chapter Summary

In this chapter, we described the drift-diffusion model that includes the Poisson, and the electron and hole current continuity equations. We also gave forms for the auxiliary equations such as current densities and net charge in terms of the state variables ϕ, n and p. This provides coupling and closure for the three equations and three unknows.

Next, we introduced corrections that are necessary to resolve quantum effects. We showed two different methods to obtain corrections; first by solving the single particle Schrödinger equation, and second by using the density gradient formalism. When we employ the former method, we solve one-dimensional Schrödinger equations along the MOSFET channel direction starting from the Si-SiO_2 interface and going down the substrate enough distance away from the potential well formed at the interface. We then self-consistently solve the coupled semiconductor equations using an effective potential approach, where a quantum effective potential is added to the electrostatic potential in current densities. Likewise, the latter method also uses an effective potential term that is calculated from the classically calculated elec-

tron concentration. The expression for this effective potential term can be derived using the one particle Wigner function, and expanding it in $\hbar$ [63, 64].

Using the same effective potential approach, we derived correction terms that enable us to resolve heterostructure and self-heating effects. In the following chapters, we use our modified equations including one or more of these effects to determine device performance figures.

Chapter 3

Carbon Nanotube Modeling

We report Monte Carlo (MC) simulation results that show position-dependent velocity oscillations and length effects in semiconducting single-walled zig-zag carbon nanotubes, as shown in Fig. 3.1. The simulations show velocity oscillations at Terahertz frequencies, which approach phonon frequencies, and velocity values reaching 7×10^7cm/s. Also, our investigations on length effects show that average velocity first overshoots, then rolls off as the tube length increases, and finally reaches its steady state value. In addition, we include quantum effects due to finite lengths of the tubes, as well as their circumference.

In recent years, carbon nanotubes have prompted researchers' interest as potential candidates for use in nanoscale electronics [15, 34, 46, 65]. This is due to their favorable structural and, especially, electrical characteristics such as high electron velocities approaching 1×10^8cm/s at high fields and 1×10^7cm/s at low fields, resulting in low-field electron mobilities as much as ten times higher than that of

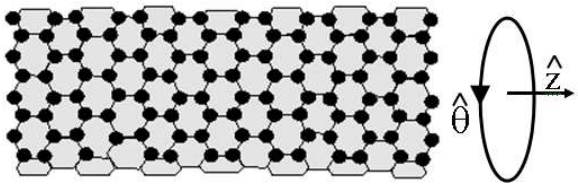

Figure 3.1: A single wall zig-zag carbon nanotube, with fundamental indices n and $m = 0$, and length L.

Si. Additionally, they exhibit negative differential velocities (NDV), similar in that respect to GaAs, opening possibilities for their use in oscillators and, therefore, communication networks. When used in circuits, our simulations indicate that they may oscillate at very high frequencies in the terahertz range, enabling data rates approaching terabits per second. Thus, to fully investigate potential gains due to CNT usage in nanoelectronics, we developed an MC CNT simulator to extract pertinent electrical CNT parameters.

CNTs are hollow tubes rolled up from planar graphite sheets (graphene). Single-walled CNTs have nanometer-scale diameters ranging from several to a few hundred angstroms. A CNT can be uniquely identified by its fundamental indices n and m, which are the coefficients of the unit vectors of the hexagonal graphite lattice used to specify the wrapping angle and the diameter. We can also relate the electrical properties of a CNT to its fundamental indices n and m such that they are metallic if $n - m$ is a multiple of three [16], or else, semiconducting with a bandgap inversely proportional to $\sqrt{n^2 + m^2 + nm}$ [17, 18]. (In this chapter, n does not refer to electron concentration unless stated otherwise as in intrinsic carrier concentration n_o, which will be mentioned toward the end of this chapter.)

Here, we concentrate on the most studied single-walled semiconducting CNT topology, which is the zig-zag (n,m = 0). From now on, "CNT" means semiconducting single-walled zig-zag carbon nanotubes. We have developed a Monte Carlo simulator for CNTs, and have used it to investigate average electron velocity as a function of position. We have also calculated average electron velocity as a function of tube length, for tubes of various indices. Next, using average velocity versus ap-

plied field curves, we have derived mobility models to be used in device simulators. To fully characterize the CNTs in our device simulators, we have also extracted additional CNT electrical parameters such as intrinsic carrier concentration and electron affinity. We first briefly describe our Monte Carlo simulator, and then show our calculated results, which are in agreement with theoretical [17, 18] and experimental [20, 21] data.

3.1 Energy Dispersion Relations

We first employ a Monte Carlo (MC) simulator [17, 18] to characterize fundamental transport properties of CNTs. Then, we incorporate these properties into our device simulators. These properties include electron drift velocity versus electric field curves, as well as CNT mobilities for zig-zag single wall CNTs. (The zig-zag CNT is probably the most studied semiconducting nanotube topology. Semiconducting zig-zag CNTs have fundamental indices $(n, 0)$, where n takes on integer values other than multiples of three.) To obtain these properties, we begin with the physical CNT system, where electrons are confined around the circumference, and move relatively freely along the tube in the direction of the longitudinal axis. Therefore, one can write the appropriate plane wave solutions that satisfy periodic boundary conditions, distinguished by the quantum number β, for the given CNT circumference. However, along the tube, electrons are not confined for long tubes. Thus, the wavevector can be written as follows:

$$\vec{k} = k_z \hat{z} + \frac{2\beta}{d}\hat{\theta} \tag{3.1}$$

Here $\hat{z}$ is parallel to the tube axis, and $\hat{\theta}$ is the unit vector along the circumference. Discrete values around the circumference, β, are bounded by the fundamental tube index $\pm n$, to take advantage of the symmetry lines in the CNT Brillouin zone.

In the simulator, electrons drift due to an external field along the length of the tube in real space, and in one of the first three lowest CNT energy subbands in the energy-momentum space until they probabilistically scatter with acoustic or optical phonons. They then start traveling again on the tube with their new energy and momentum that are calculated using the energy and momentum conservation laws. This process is repeated until the electron exits the tube from one end. Here, we consider scattering by optical and acoustic phonons, causing inter and intra subband, and inter-valley and intra-valley transitions. Both scattering mechanisms are treated within the deformation potential method using the Fermi's Golden Rule [17, 18].

3.1.1 Monte Carlo for Long Tubes: The Continuum Model

Due to confinement around the circumference, the bandstructure splits into a system of subbands when graphene is wrapped into a CNT. Each of the subbands has a characteristic effective mass, mobility and band energy minima. We determine the energy levels of CNTs by applying zone-folding methods to graphene [16]. From the two-dimensional graphene band diagram, we cut one-dimensional slices, whose numbers and locations are set by the fundamental tube indices $(n, 0)$. The resulting CNT $(n, 0)$ energy dispersion relation, which is determined by applying zone-folding

methods to the graphene energy dispersion relation that is calculated using the tight binding model [16], is shown below (In this chapter, E is used for the energy and F is used for the field):

$$E(k_z, \beta) = \pm\gamma\sqrt{1 + 4\cos\left(\frac{Tk_z}{2}\right)\cos\left(\frac{\pi\beta}{n}\right) + 4\cos^2\left(\frac{\pi\beta}{n}\right)} \qquad \text{(eV)} \qquad (3.2)$$

Here, T is the length of the translational vector, which is equal to 4.26Å for the zig-zag tubes. Also, we use 3eV for the nearest-neighbor π-hopping integral γ [16]. Now, we determine the lowest three subbands, using the above expression. We first take the derivative of Eqn. 3.2 with respect to k_z, or k $(= k_z)$.

$$\left|\frac{dE}{dk}\right| = 9\left|\frac{T\sin\left(\frac{Tk}{2}\right)}{E}\right| \qquad (3.3)$$

For the k values that give subband minima, the derivative of Eqn. 3.2 is zero. Since $\sin\left(\frac{Tk}{2}\right)$ is zero when k is zero, we have an energy minimum of each subband at $k = 0$. Moreover, we also need to check the boundaries. However, in this case, energy values at the boundaries $(\pm\frac{\pi}{T})$ are higher than energy values at $k = 0$.

Next, we determine wavevector indices, β, for the lowest three subbands by searching for the integers from zero to n that give the lowest three values for $E(0, \beta)$. To shorten our search time, we take the derivative of Eqn. 3.2 with respect to β at $k = 0$.

$$\left|\frac{dE}{d\beta}\right|_{k=0} = 6\frac{\pi}{n}\left|\sin\left(\frac{\pi\beta}{n}\right)\left(1 + 2\cos\left(\frac{\pi\beta}{n}\right)\right)\right| \qquad (3.4)$$

The above derivative tells us that the lower subbands are either around $\beta = 0$ or $\beta = \frac{2n}{3}$. From $E(0, \beta) = \pm 3\left|1 + 2\cos\left(\frac{\pi\beta}{n}\right)\right|$, we can show that they are around $\beta = \frac{2n}{3}$. Since β takes on integer values, β for the lowest subband is $\frac{2n}{3}$ rounded to

the nearest integer. Furthermore, we can also find the β values using the Brillouin zone. More specifically, from geometrical considerations in the Brillouin zone, we also found β, for the lowest subband, equal to $\frac{2n}{3}$ rounded to the nearest integer. Additionally, since cosine is an even function, $-\beta$ and β give the same energy. Therefore, we here have two identical valleys with each having three subbands.

3.1.2 Finite CNT Length: Incorporating Quantization Effects

We consider effects due to finite length of the tubes, which lead to discretization in energy dispersion curves, as shown in Fig. 3.2. For a zig-zag tube, the length of the translational vector is roughly 4.26Å; therefore, maximum electron momentum, which is equal to pi (π) over this value, is approximately 0.74Å^{-1}. Furthermore, minimum momentum step is related to the length of the tube, which is $2\pi/\text{L}$. Figure 2 shows the steps we have for a 5nm long tube. Also, for the longest tube we simulate, which is 100nm long, we have about twenty times more steps. Using this information, we include the finite contribution of longitudinal quantization on electron transport during our simulations. We calculate scattering rates using the continuous band. We have modified our MC simulator to account for this quantization. In our modified MC, the electron drifts along the tube until it hits an energy step that needs to be overcome to achieve higher momentum values (This is true only for positive momentum values). At this point, we determine reflection and transmission probabilities for this barrier. We below show the backward (reflection) scattering rate for an electron with a momentum k at the edge of a step, which is

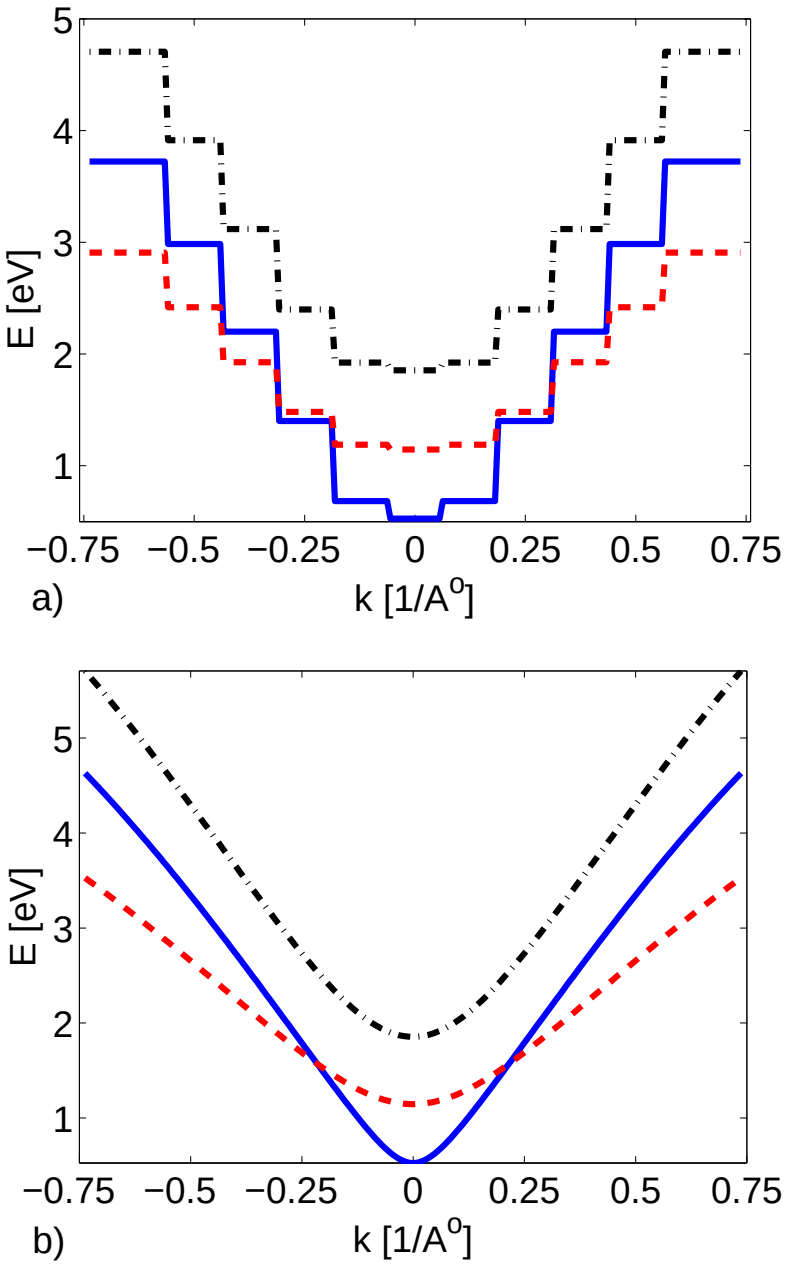

Figure 3.2: a) Discretization of the energy dispersion curves of a 5nm long n=10 CNT (T=0.46nm). b) Energy dispersion relations for the first three subbands of an infinitely long n=10 CNT.

$2\pi/L$ (=Δk) wide.

$$\Gamma_{\text{ref}} = \left(\frac{\Delta k}{2k}\right)^2 \tag{3.5}$$

When an electron, with momentum k_1 and energy $E(k_1)$, hits an energy barrier ΔE, upon successful transmission, it has a new momentum k_2 that satisfies the energy conservation written below:

$$E(k_2) = E(k_1) + \Delta E \tag{3.6}$$

For such a system, the transmitted and reflected power ratios [66] are:

$$R = \left(\frac{k_2 - k_1}{k_2 + k_1}\right)^2 \tag{3.7}$$

$$T = 1 - R. \tag{3.8}$$

In our case, $k_2 - k_1$ is $2\pi/L$. Since the electron keeps gaining energy due to the applied field according to the continuum model, to retain consistency, it does not suddenly gain energy if transmitted. Therefore, $k_2 + k_1$ becomes $2k_1$. Depending on the likelihood of transmission, the electron either continues gaining momentum until it hits the next step or reflects back to negative momentum values ($-k_1$). The longer the CNT ($\Delta k \to 0$), the smaller the barriers become, with reflection coefficients approaching zero and the continuum approximation for long tubes.

3.1.3 Phonon Energy Dispersion Relations

To obtain CNT phonon energy spectra, we start from the phonon dispersion curves of the graphene. We first calculate the graphene phonon spectra using the forth nearest neighbor force constant model, where force —equivalently, spring—

constants determine the inter-atomic interactions. We derive this model from the equation of motion, as follows [16].

$$M_i \frac{\partial^2 u_i}{\partial t^2} = \sum_j K_{ij}(u_j - u_i) \qquad (i, j = 1, 2, 3...) \tag{3.9}$$

Above, i and j represent one of the N atoms in the unit cell. In addition, M_i and u_i are the mass and the location of the ith atom, and the force constant between the i^{th} and the j^{th} atoms is K_{ij}.

To obtain phonon spectra, we first apply a Fourier transform, and substitute u_i with $\frac{1}{\sqrt{N}} \int e^{-i(\vec{k}\cdot\vec{r_i}-\omega t)} u_{k_i} dk$. In the exponential, the coefficient i is the complex number $\sqrt{-1}$. Moreover, this gives the following equation of motion

$$\begin{aligned} -\omega^2 M_i \frac{1}{\sqrt{N}} \int e^{-i(\vec{k}\cdot\vec{r_i}-\omega t)} u_{k_i} dk = \\ \sum_j K_{ij} \left(\frac{1}{\sqrt{N}} \int e^{-i(\vec{k}\cdot\vec{r_j}-\omega t)} u_{k_j} dk - \frac{1}{\sqrt{N}} \int e^{-i(\vec{k}\cdot\vec{r_i}-\omega t)} u_{k_i} dk \right) \end{aligned} \tag{3.10}$$

Canceling out common terms from both sides, and using the orthogonality condition, we get a matrix equation of the form $\overline{\overline{A}}\overline{u}_k = 0$, where $\overline{u}_k = [u_{k_1}..u_{k_i}..u_{k_N}]^T$. The diagonal elements of $\overline{\overline{A}}$ are $\left(\sum_j K_{ij} - M_i\omega^2\right) - K_{ii}e^{i\vec{k}.(\vec{r_i}-\vec{r_i})}$. Additionally, the off-diagonal elements of $\overline{\overline{A}}$ are $-K_{ij}e^{i\vec{k}.(\vec{r_i}-\vec{r_j})}$. Furthermore, for different k values, we find the corresponding eigenvalues of $\overline{\overline{A}}$. Tracing over all ks gives the dispersion curve of the material.

Graphene has two atoms in its unit cell, as shown with the closest pairs of AB along the x axis in Fig. 3.3. We use the force constants to find K_{ij} between the given atom (A or B), and its neighbors. Below, we write in Table 3.1 the force constants in the x (radial), y (transverse in-plane) and z (transverse out-of-plane) directions

Table 3.1: Spring constants, in N (kg·m/s^2), of the graphene in x (radial), y (transverse in-plane) and z (transverse out-of-plane) directions, shown in Fig. 3.3, for the first to the forth nearest neighbors [16].

$K_{x_1} = 3.65$	$K_{y_1} = 2.45$	$K_{z_1} = 0.982$
$K_{x_2} = 0.88$	$K_{y_2} = -0.323$	$K_{z_2} = -0.04$
$K_{x_3} = 0.3$	$K_{y_3} = -0.525$	$K_{z_3} = 0.015$
$K_{x_4} = -0.192$	$K_{y_4} = 0.229$	$K_{z_4} = -0.058$

[16]. To find the force constants in any other direction, we rotate the force constants' matrix by θ using the rotation matrix U and transformation $K^{'} = U^{-1}KU$.

$$U = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3.11}$$

We next show our calculated phonon dispersion curves in Fig. 3.4, using the prescription described before. We take the mass of a carbon atom in the graphene as about $12m_o$, where m_o is the free electron mass. Additionally, bond lengths in Fig. 3.3 are 2.49Å. Furthermore, to find the dispersion curves of the CNT, we approximate the dispersion curves of graphene around the Γ and K points, which are important for transport. We then calculate the phonon energy spectrum by applying zone-folding methods to graphene. Our calculated energy dispersion relations for acoustic and optical phonons can be found in [18]. We give a generalized formula for our dispersion curves below:

$$E_p(q,\eta) = E_{p_o}(\eta) + \hbar\upsilon_s|q|\theta\left(|q| - \lambda|\frac{\eta}{d}|\right) \tag{3.12}$$

Here q, η, θ and λ are respectively the phonon wavevector along the length of

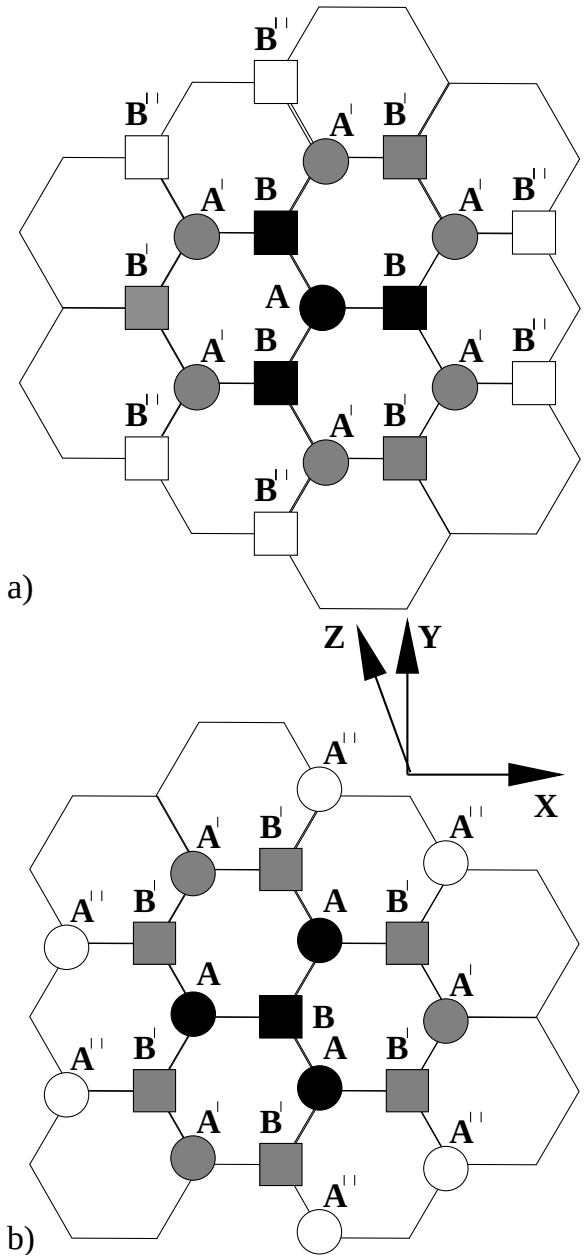

Figure 3.3: Four nearest neighbors of the two atoms, solid circle A in a) and solid square B in b), in the graphene unit cell [16].

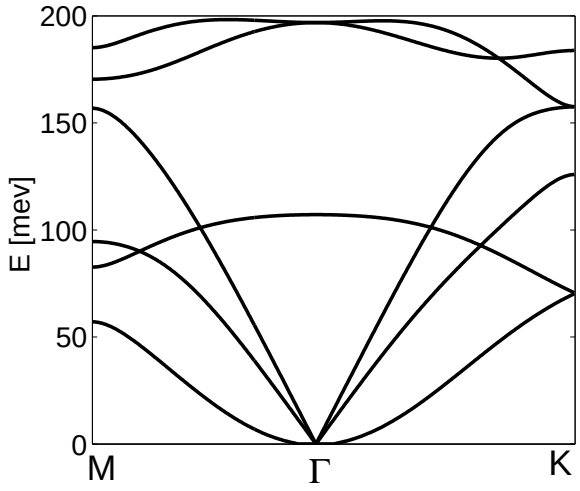

Figure 3.4: The graphene phonon dispersion curves along the symmetry lines.

the tube, phonon wavevector index around the circumference of the tube, a dispersion coefficient and a kink factor that is zero for optical phonons and one for acoustic phonons. Additionally, d is the diameter of the tube and υ_s is the longitudinal sound velocity in graphene, which is 200Å/ps.

3.2 Scattering Rates

To determine the electron-phonon scattering rates, we employ the deformation potential approximation and Fermi's Golden Rule [18]. In this scheme, the total scattering rate ($\Gamma_i(k)$) for an electron in subband i with a wavevector k (and β_i) to any other subband j by absorbing or emitting an intra-valley ($q, \eta = \beta_i - \beta_j$) or inter-valley ($q, \eta = \beta_i - \beta_j \pm (2n)$) phonon can be written as follows:

$$\Gamma_i(k) = \sum_q \frac{\hbar D^2 Q^2 DOS_j(E(k+q, \beta_j))}{2\rho E_p(q, \eta)} \left[N(q, \eta) \pm \frac{1}{2} \right] \tag{3.13}$$

Here, D is the deformation potential taken to be 9eV, Q is a wavevector for

optical and acoustic phonons, DOS is the density of states calculated by the inverse slope of Eqn. 3.2, ρ is the linear mass density, and N is the Bose-Einstein phonon occupation number at equilibrium. Additionally, the above sum has non-vanishing values for phonon wavevectors that satisfy energy and momentum conservation laws: $\delta\left(\mathrm{E(k+q,}\beta_j) - E(k,\beta_i) - E_p(q,\beta_i - \beta_j \pm (2n)\right)$.

We below show the density of states. It has singularities near the band minima, where k and $\sin(k)$ are zero. To avoid numerical problems, we add an epsilon to k when it is exactly zero. For proper handling of this, we need to use the collision broadening concept. However, our investigations show that they give the same results for this problem, enabling us to use the aforementioned truncation for fast computation.

$$DOS(k,\beta) = \left| \frac{E(k,\beta)}{9T\sin\left(\frac{Tk}{2}\right)} \right| \tag{3.14}$$

In addition, the zig-zag CNTs have a number $2n$ of hexagons in their unit cells, with each hexagon weighing M $(12m_o)$. Therefore, the linear mass density is $\frac{2nM}{\mathrm{AT}}$, where T is the length of the translational vector, and A is the Avogadro's number.

Figure 3.5 shows the scattering rates calculated using Eqn. 3.13 for the lowest three subbands of n=10 and n=22 CNTs for energies lower than the energy minima of the fourth subbands. We attribute the peaks in Fig. 3.5 to the singularities associated with the density of states at band minima. In addition, we observe oscillatory behavior in the scattering rate curves, which are visible in Fig. 3.5(a) and 3.5(b) between the first two peaks. We associate this with the slightly different energies required to emit or absorb a phonon.

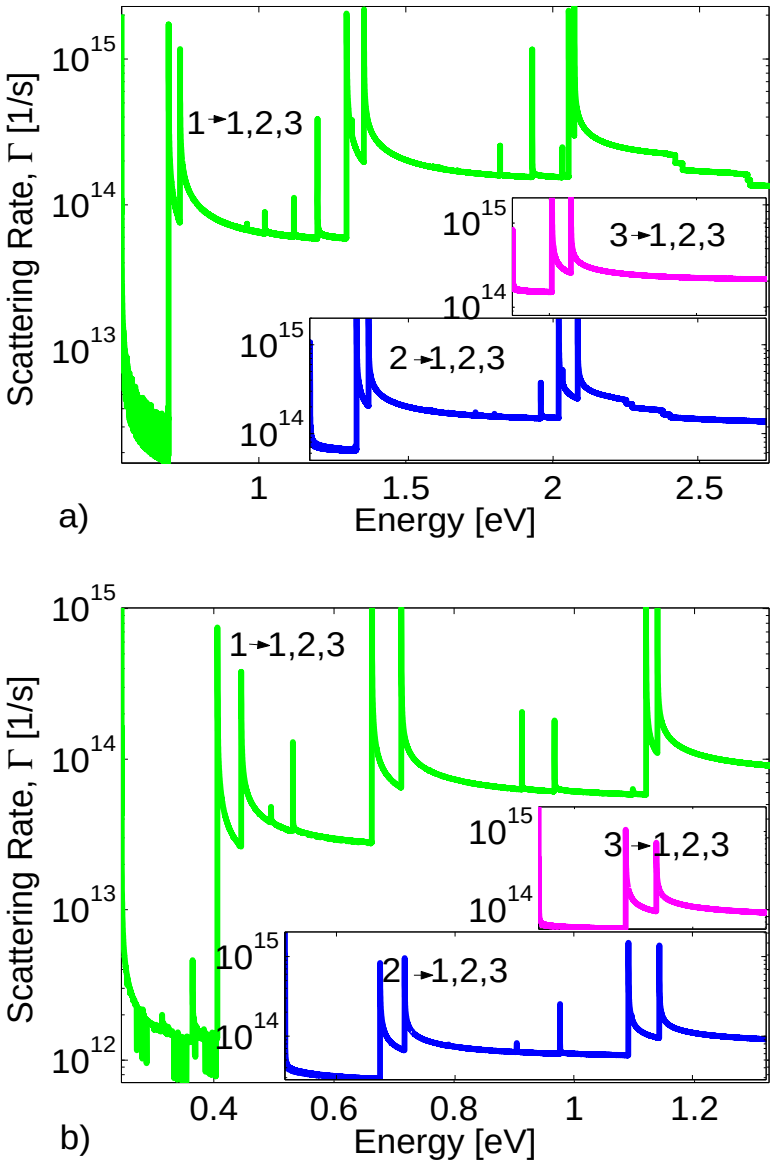

Figure 3.5: Scattering rates from the first, second (lower left corner) and third (on top of the lower left corner plot) subbands to the lowest three subbands of CNTs with indices of a) 10 and b) 22. Insets share the same abscissa with the mother plot.

3.3 Velocity Curves

3.3.1 Position-Dependent Velocity Oscillations

Using our Monte Carlo simulator, we first investigate how local CNT electron velocities change by varying the applied field. To obtain average local electron velocities as a function of position, we inject electrons, which are picked from a Fermi-Dirac distribution, from both sides of the tube. We then keep track of their position, average energy and momentum. Our calculated average electron velocities on 100nm-long CNTs with indices of 10 (diameter, d=0.8nm) and 22 (d=1.7nm) are shown in Figs. 3.6(a) and 3.6(b), respectively. From $\frac{\Delta E}{\hbar \Delta k}$, which is also equal to $\frac{l}{\tau}$, we calculate average velocities. The newly introduced variables ΔE, Δk, l and τ are change in total energy, change in total momentum, length and time spent around the vicinity of a given location, respectively. The two aforementioned methods to calculate average velocities give the same answer because of the following reasons. The $\hbar \Delta k$ term is equal to change in momentum Δp, which is also equal to the product of the elapsed time Δt ($= \tau$) and the electric force (qF) due to the applied field F. In addition, change in energy due to drift can be calculated from force, due to the electric field, (qF) times distance l. Therefore, $\frac{\Delta E}{\hbar \Delta k} = \frac{qFl}{qF\tau}$, resulting in $\frac{l}{\tau}$ that was shown before.

Simulations predict velocity oscillations at Terahertz frequencies, with a highest frequency of approximately 30THz among the simulated cases. From the velocity versus location curve of the n=10 tube under 100kV/cm, shown in Fig. 3.6a, we take the average wavelength and velocity of the oscillations roughly as 15nm

and 4.5×10^7cm/s. This results in $f = \frac{4.5\times10^7\text{cm/s}}{15\times10^{-7}\text{cm}} = 30$THz. (Here, we have velocity oscillations in space, which might induce dipole formations within the material. These dipoles are likely to travel on the CNT, resulting in velocity oscillations in time. To observe if this phenomenon does indeed occur, transient simulations need to be performed.)

We associate such high oscillation frequencies with the phonon spectrum and the one-dimensional nature of the system, which results in the average scattering rates and momenta that are shown in Fig. 3.6(c). More specifically, Figure 3.6(c) shows that average scattering rate has oscillations with a period of 16nm (first maximum) and 20nm (second maximum) for the first few cycles and at their harmonics thereafter. Under an applied field of 100kV/cm, an electron will gain 160meV and 200meV after a free flight of 16nm and 20nm, respectively. These 160meV and 200meV energies above the energy band minima correspond to energy differences sufficient enough to have inter-valley acoustic and optical, and intra-valley and inter-valley optical phonon emissions in addition to all the other scattering mechanisms. When this happens, electrons are much more likely to scatter to lower momentum values where densities of states (equivalently scattering rates to those states in our one dimensional system) are much higher. This is in agreement with average local momentum curve shown in Fig. 3.6(c). In addition, we observe that all except a negligible portion of the electrons travel in the first subband, thus eliminating the possibility to have the velocity oscillations due to transfer of electrons from the first to the second subband, and vice versa.

In summary, we theoretically show that one-dimensional CNT system has

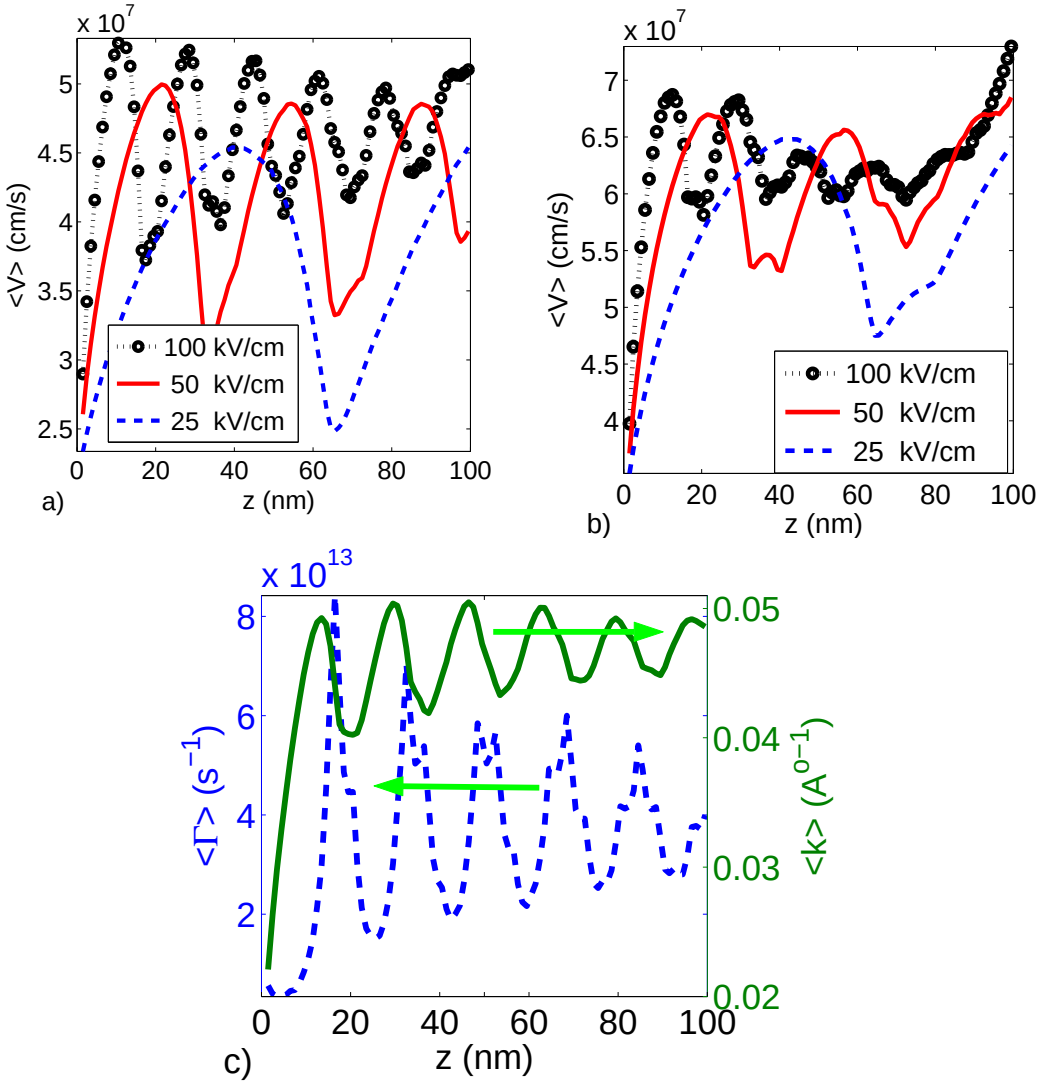

Figure 3.6: Average local electron velocities on 100nm-long CNTs with indices of a) 10 and b) 22. c) Average local scattering rate and momentum for the n=10 tube under F=100kV/cm.

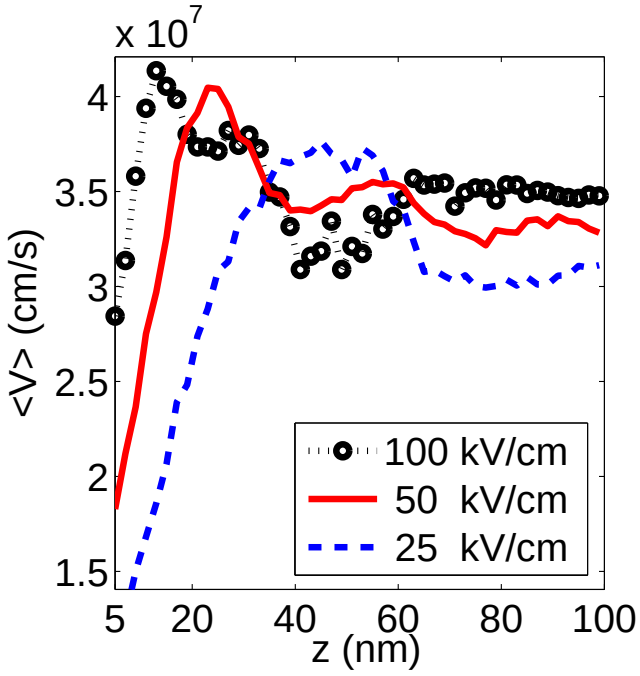

Figure 3.7: Average velocity of an electron on various length n=10 CNTs

velocity oscillations with Terahertz frequencies, approaching to those of phonons. This may facilitate very high frequency oscillators similar to Gunn diodes, opening new paradigms for Terahertz RF electronics.

3.3.2 Length-Dependent Velocity Overshoots

We next show in Fig. 3.7 our calculated average velocity as a function of CNT length. It shows how the forward and backward currents cancel each other out yielding an increase in electron velocity for an increase in length. As we increase tube length, backward current decreases exponentially due to an exponential decrease in the probability of electrons succesfully travelling the length of the tube in the reverse field direction, therefore contributing less to the net reverse current. We also have overshoots from a combination of the previously mentioned scattering mechanisms.

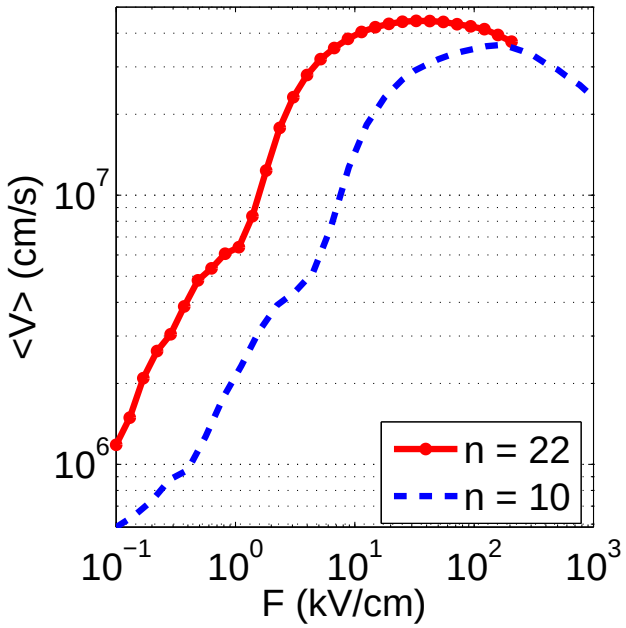

Figure 3.8: Average electron velocities as a function of applied field on infinitely long CNTs with indices of 10 and 22.

3.3.3 Continuum Model: Velocity Curves

We last show in Fig. 3.8 our calculated average ensemble electron velocities as a function of applied field. It shows that velocity first increases linearly with the applied field, reaches a peak, and then rolls off. This negative differential velocity is caused by the transfer of electrons, as applied field increases, from the first subband to the second subband where effective electron mass and velocity is higher and lower, respectively, than that of the other.

3.4 Mobility Models

Electron mobility can be very high in nanotubes with negligible defect densities. The high mobility is due to small effective masses and low scattering rates

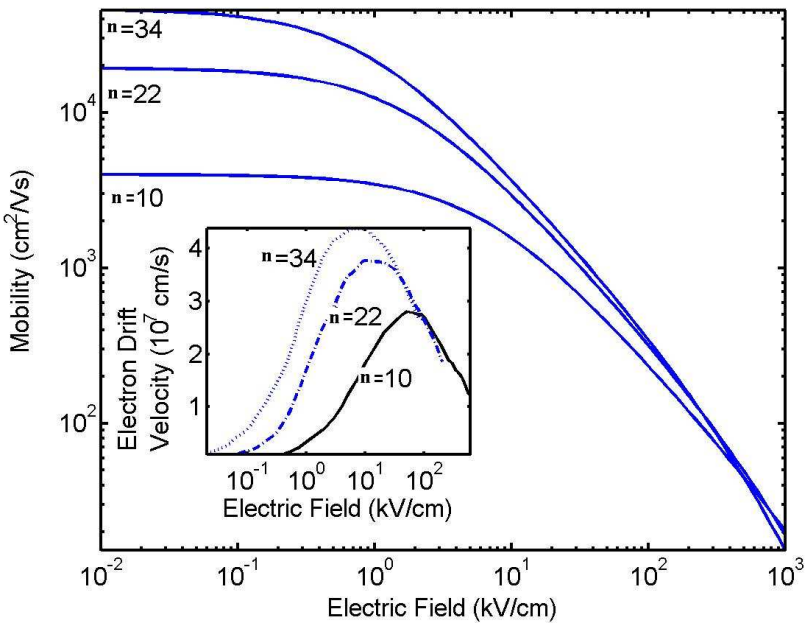

Figure 3.9: Average electron velocities as a function of applied field on infinitely long CNTs with indices of 10 and 22.

resulting from the quasi-one-dimensional transport. As electric fields increase, the scattering rate increases, and mobility decreases. Low-field mobility has been measured and calculated theoretically to be greater than $10^5 \mathrm{cm}^2/\mathrm{Vs}$ [17]-[21]. We derived a mobility model based on our MC simulation results of drift velocity versus electric field curves. These velocity versus field curves are plotted again in the Fig. 3.9 inset for CNTs with tube indices ranging from 10 to 34, which correspond to diameters of 8Å to 27Å. Simulations indicate that electron drift velocity first increases linearly with the applied field, reaches a maximum, and then rolls off, showing a negative differential mobility (NDM). We find that peak electron velocities are as much as five times higher than what they are in silicon. Electrons reach velocities as high as 4.5×10^7cm/s in large diameter CNTs ($n = 34$). The maxi-

mum velocity drops for smaller diameter tubes, which is approximately 3×10^7cm/s for n=10. However, the peak velocities in narrow tubes are still larger than the corresponding velocities in other semiconductors. Calculated results also show that the critical field, where we have the peak drift velocity, increases from 1kV/cm to 10kV/cm as we reduce the tube diameter from 27Å to 8Å.

Figure 3.9 inset shows three main characteristics of the CNTs. First, CNTs attain drift velocities larger than other semiconductors. Investigations show that this is due to small effective masses and decreased scattering rates, which is a result of the quasi-one-dimensional system. Second, electrons in large diameter tubes have higher velocities than the ones in small diameter tubes for a given applied field, unless the applied field is too large. This leads to higher low-field mobilities for larger diameter tubes. (Low scattering rates on bigger diameter tubes are due to their higher linear mass densities, which is inversely proportional to the scattering rate.) Analysis shows that this is due to lower effective masses in the larger diameter tubes. Third, all CNTs show NDM. They are similar to GaAs in that respect, where conduction band velocity of the first subband is larger than that of the second.

3.4.1 Field and Index Dependent CNT Mobility

We develop an analytical mobility model for small diameter tubes, considering the two lowest subbands which dominate the conduction. By that means, we embed the effects of NDM in our mobility model. We then express the final mobility using

Mathiessen's rule, as follows:

$$\frac{1}{\mu(n,F)} = \frac{1}{\mu_1(n,F)} + \frac{1}{\mu_2(n,F)} \tag{3.15}$$

Here, $\mu_1(n,F)$ and $\mu_2(n,F)$ refer to the mobilities in the first and second subbands, respectively. The mobilities are functions of the fundamental tube index, n, and the electric field, F. The mobility of the first subband is:

$$\mu_1(n,F) = \frac{\mu_o(n)}{1+\frac{F}{F_c(n)}} \tag{3.16}$$

Above, $\mu_o(n)$ is the low-field mobility, and $F_c(n)$ is the critical electric field. The critical electric field corresponds to the peak electron drift velocity. We have empirically determined the following expressions for the low-field mobility and the critical field in terms of the tube index n:

$$\mu_o(n) = 40n^2\left(1+\frac{\rho}{n^{2/3}}\right) \quad (\mathrm{cm}^2/\mathrm{Vs}) \tag{3.17}$$

$$F_c(n) = \frac{1}{n^{3/2}}\left(1+\frac{64\rho}{n^2}\right)10^6 \quad (\mathrm{V/cm}) \tag{3.18}$$

Here, ρ=1-gcd(n+1,3) (= 0, -2), where gcd(n+1,3) is the greatest common divisor of n+1 and 3. The expression for the low field mobility can be obtained from the familiar expression $\mu_o = q\tau/m^*$. Results from our previous work on small diameter tubes, with tube index n less than 37, indicate that τ is proportional to n, and m^* is inversely proportional to n [18, 30] thereby giving the quadratic-type form of Eqn. 3.17. We empirically write the mobility of the second subband as follows:

$$\mu_2(n,F) = \frac{V_{\max}(n)}{F\left(1+\lambda\frac{F}{F_c(n)}\right)} \tag{3.19}$$

Here, λ is an empirical parameter which we find to have the value of 0.01. $V_{\max}(n)$ is the maximum drift velocity of the electrons, shown in Fig. 3.9 inset. We find it to be the following function of n:

$$V_{\max}(n) = 1.5n^{1/3}\left(1+\frac{\rho}{2n}\right)10^7 \qquad \text{(cm/s)} \tag{3.20}$$

In Fig. 3.9, we show our calculated mobility versus field curves, using Eqns. 3.15-3.20. For n=34, low-field mobility is as high as 4×10^4cm^2/Vs, while for n=10 it is approximately 4×10^3cm^2/Vs. Such high mobilities indicate that incorporating CNTs into MOSFETs may yield high drive currents and transconductances.

We also found that high scattering rates help to validate the use of a mobility model for the CNTs we simulated, that are about $0.14\mu m$ long and ranging in diameter from 8Å to 17Å. Furthermore, mean free length versus electric field curves are concave down, like drift velocity versus electric field curves. In addition, the mean free paths (mfp) range from approximately 10nm to a maximum of 100nm for the CNTs we use here. For smaller diameter tubes (8Å) the mfp has a narrow peak value of approximately 30nm [18, 67].

3.4.2 Field and Diameter Dependent CNT Mobility

We convert the field and index dependent CNT mobility into a field and diameter dependent CNT mobility using the following transformation for the single-walled zig-zag tubes.

$$n = \frac{d\pi}{a} \tag{3.21}$$

Above, d is the diameter of the tube in angstroms, and a is the length of the

graphene unit vector, which is 2.49Å. Next, using the dimensionless $\hat{d}$ (=d(Å)/1Å) in Eqns. 3.15, 3.16 and 3.19, we replace n by $\hat{d}\pi/2.49$.

$$\frac{1}{\mu(\hat{d},F)} = \frac{1}{\mu_1(\hat{d},F)} + \frac{1}{\mu_2(\hat{d},F)} \tag{3.22}$$

$$\mu_1(\hat{d},F) = \frac{\mu_o(\hat{d})}{1+\frac{F}{F_c(\hat{d})}} \tag{3.23}$$

$$\mu_2(\hat{d},F) = \frac{V_{\max}(\hat{d})}{F\left(1+\lambda\frac{F}{F_c(\hat{d})}\right)} \tag{3.24}$$

Now, we have the following diameter dependent parameters to be used in the above equations.

$$\mu_o(\hat{d}) = 63.5\hat{d}^2\left(1+\frac{\rho}{1.17\hat{d}^{0.67}}\right) \quad (\mathrm{cm}^2/\mathrm{Vs}) \tag{3.25}$$

$$F_c(\hat{d}) = \frac{1}{1.41\hat{d}^{1.5}}\left(1+\frac{40.3\rho}{\hat{d}^2}\right)10^6 \quad (\mathrm{V/cm}) \tag{3.26}$$

$$V_{\max}(\hat{d}) = 1.62\hat{d}^{0.33}\left(1+\frac{\rho}{2.52\hat{d}}\right)10^7 \quad (\mathrm{cm/s}) \tag{3.27}$$

3.4.3 Temperature Dependent CNT Mobility

We have presented our room temperature mobility model. We next include temperature dependencies in the overall mobility. We use our MC simulation results to obtain the temperature dependent velocity curves shown in Fig. 3.10. Our analyses have shown that CNT mobilities are phonon scattering limited down to 100°K [67], which agrees quite well with a recent experiment [20, 21, 68]. Thus our MC simulations that take into account longitudinal acoustic and optical phonons with intra- and inter-valley scatterings would suffice to describe the temperature dependent CNT behavior.

To formulate the temperature dependency, we follow the methods used for

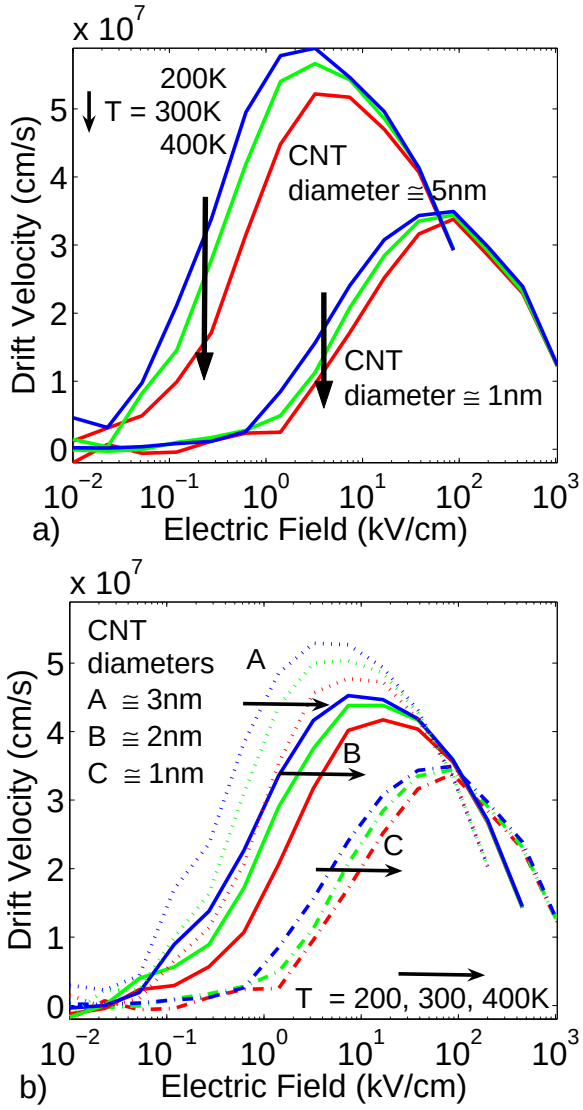

Figure 3.10: Electron drift velocities as a function of the applied electric field for different CNTs varying in diameter and temperature.

silicon, where a power-law relationship relates the mobility, drift velocity and critical field values at different temperatures to that of the room temperature [69, 70] as follows:

$$\mu = \mu(T_o)\left(\frac{T}{T_o}\right)^{\alpha} \tag{3.28}$$

$$V = V(T_o)\left(\frac{T}{T_o}\right)^{\beta} \tag{3.29}$$

$$F_c = F_c(T_o)\left(\frac{T}{T_o}\right)^{\gamma} \tag{3.30}$$

Here α, β and γ are parameters that need to be determined empirically. For silicon, α ranges from -1.4 to -2.5, and β and γ take on values of -0.87 and 1.55, respectively [69, 70]. Our calculations for the CNTs yield -0.5, $-0.05\hat{d}^{3/4}$ and 1.18 for the powers of low-field mobility, peak drift velocity and critical field, respectively.

According to Fermi's Golden Rule, temperature dependency of the scattering rate for the low-field (correspondingly, also low in energy) region is affected mostly by the density of states, which is inversely proportional to $\sqrt{E}$ in a one-dimensional system. However, integration over possible states yields $\sqrt{E}$, which is proportional to $\sqrt{T}$. Since low-field mobility is inversely proportional to the scattering rate, α takes the value of -0.5, which is also the value found from simulations. In addition, at low temperatures, low-field electron transport is affected by the Bose-Einstein phonon occupation number, where low-field mobility is inversely proportional to T, due to expansion of $e^{\frac{\hbar\omega}{kT}} - 1 = \frac{\hbar\omega}{kT}$.

We next investigate the temperature dependency of the critical field and the maximum drift velocity. Critical field increases as temperature rises. We attribute this to the increase in Bose-Einstein phonon occupation number as temperature

increases, resulting in higher scattering rates. This reduces average drift velocity for a given field as temperature increases. Also, due to higher scattering rates, electrons spend more time in the first subband causing an increase in the critical electric field as temperature rises. Additionally, we attribute weak temperature dependencies of maximum drift velocities to the one-dimesional nature of the CNTs. The peak velocities are affected by lower mobilities and higher critical fields as temperature increases.

3.4.4 Length Dependent CNT Mobility

So far, we have included the index — equivalently, the diameter—, field and temperature dependencies into our CNT mobility model. To accurately model the CNT mobility in short length tubes, we additionally need to consider the length effects on average CNT electron velocities. To derive an analytical formula that scales the CNT electron mobility depending on the tube length, we fit the curves shown in Fig. 3.7 to an analytical expression, using the following relation.

$$V(z,F) = V_{\infty}\left[1 - e^{-A(F)z}\cos(B(F)z)\right] \tag{3.31}$$

Here, z is the length of the CNT. $V(z,F)$ is the average velocity on the tube as a function of tube length and applied field. V_{∞} is the average CNT velocity on sufficiently long tubes for an applied field. Furthermore, $A(F)$ and $B(F)$ are field dependent coefficients.

We use the above expression due to the average velocity versus length curves' resemblance to a unit step response of a second order differential system for the

damped case [71], as shown in Fig. 3.11(a). The plot in 3.11(a) can be described using an expression like the one above in Eqn. 3.31.

$$y(x) = y_\infty \left[1 - e^{-Ax}\cos(Bx)\right] \tag{3.32}$$

Here, B is $\frac{2\pi}{T}$, and A is the damping factor [71].

Next, we empirically determine the oscillation period T for the curves in Fig. 3.7, also shown in Figs. 3.11(b)-3.11(d), as 36nm, 60nm and 100nm for the applied fields of 100kV/cm, 50kV/cm and 25kV/cm, respectively. These values for the periods make $B(F)$ in Eqn. 3.31 equal to $\left(\frac{\pi}{18\text{nm}}\right)\left(\frac{3}{5}\right)^r$, where r is 0, 1, 2 for the external fields of 100kV/cm, 50kV/cm and 25kV/cm, respectively. Likewise, we find that $A(F) = 0.1\left(\frac{3}{5}\right)^r$, where it is in nm^{-1}. Therefore, analytical expressions for the fit curves plotted in Figs. 3.11(b)-3.11(d) corresponding to an n=10 tube for the applied fields of 100kV/cm, 50kV/cm and 25kV/cm are written below:

$$V(z, 100\text{kV/cm}) = 3.5\times10^7\left[1 - e^{-0.1z}\cos\left(\frac{\pi z}{18\text{nm}}\right)\right] \quad \text{(cm/s)} \tag{3.33}$$

$$V(z, 50\text{kV/cm}) = 3.4\times10^7\left[1 - e^{-0.06z}\cos\left(\frac{\pi z}{30\text{nm}}\right)\right] \quad \text{(cm/s)} \tag{3.34}$$

$$V(z, 25\text{kV/cm}) = 3.1\times10^7\left[1 - e^{-0.036z}\cos\left(\frac{\pi z}{50\text{nm}}\right)\right] \quad \text{(cm/s)} \tag{3.35}$$

Next, we find the corresponding length dependency of the CNT mobility. Since $\mu = \frac{V_\infty}{F}$, where V_∞ is the average electron velocity on long tube, as mentioned before, the length dependent mobility for n=10 tube is written as follows (we now denote length by L):

$$\mu(L, F) = \mu(L_\infty)\left[1 - e^{-\left[0.1\left(\frac{3}{5}\right)^r\right]L}\cos\left(\frac{\pi L}{18\text{nm}}\left(\frac{3}{5}\right)^r\right)\right] \tag{3.36}$$

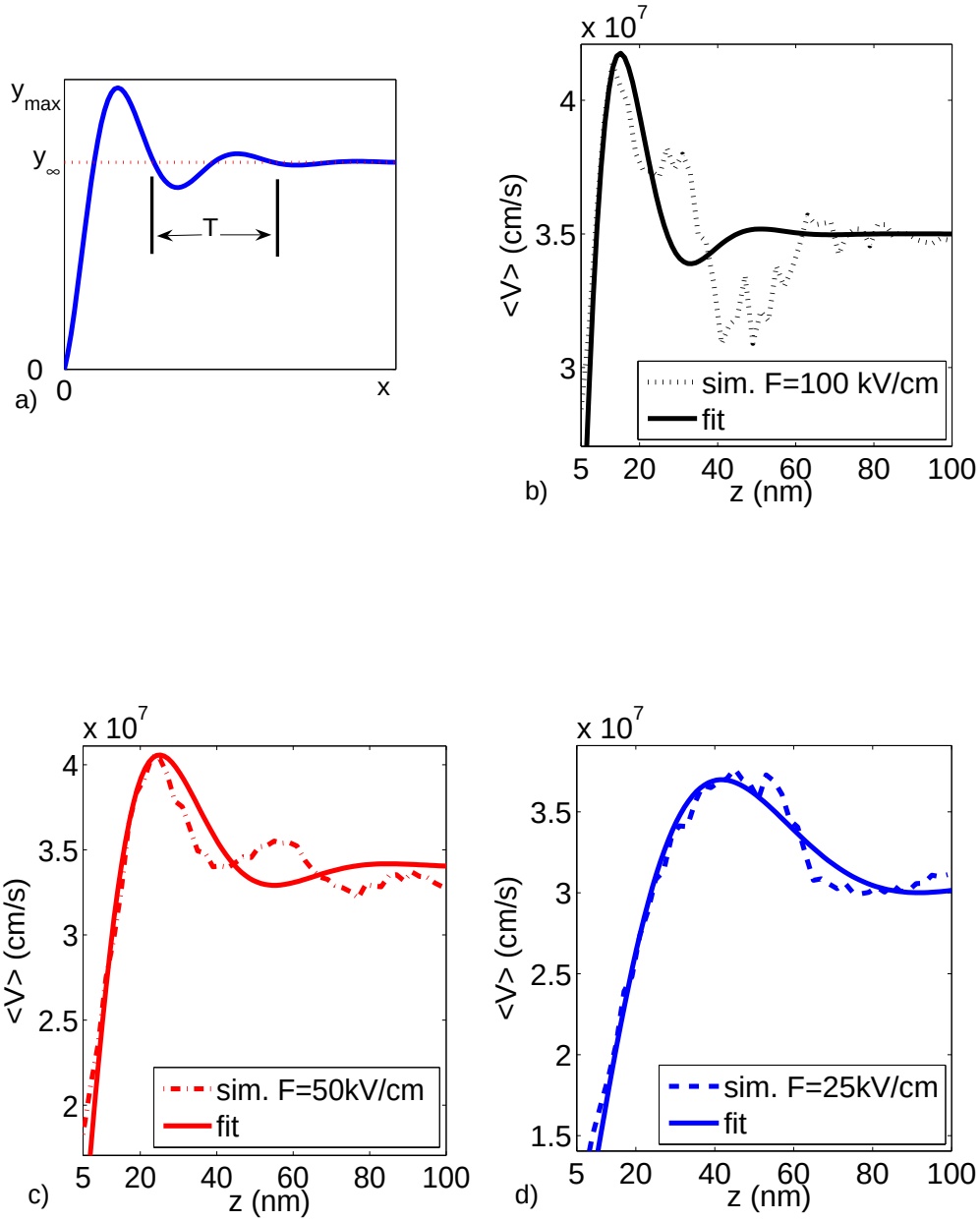

Figure 3.11: a) Unit response of a second order differential system (damped case). b), c), d) Average velocity curves of an electron on various length n=10 CNTs for different applied fields are fitted to an analytical expression given in Eqn. 3.31.

As previously mentioned, r is 0, 1 or 2 depending on the applied fields of 100kV/cm, 50kV/cm and 25kV/cm, respectively.

3.5 CNT Intrinsic Carrier Concentration

To investigate the effects of embedding a CNT into a MOSFET, we developed a novel device simulator. One of the fundamental quantities required by our CNT-MOSFET solver is the CNT intrinsic carrier concentration. Therefore, we develop a methodology to obtain the intrinsic carrier concentrations of different tubes. We start from the parabolic energy dispersion approximation. (We approximate bands like the ones shown in Fig. 3.2 using parabolic energy dispersion relations.) The density of states for each subband is zero for energies less than the energy minimum of that particular subband, and becomes the following for energies greater than the subband energy minimum:

$$DOS(n,\beta) = \sqrt{\frac{m^*_{n,\beta}}{2\hbar^2(E - E^n_\beta)}} \tag{3.37}$$

Using nondegenerate statistics and the zero energy point at the midgap, we get the following expression for the intrinsic carrier concentration n_o as a function of the fundamental tube index n:

$$n_o(n) = 2\frac{1}{2\pi}\sum_\beta \int_{E_\beta}^{\infty} DOS(n,\beta)e^{-E/kT}dE \tag{3.38}$$

We then change the variable in the integral from E to $t = \frac{E-E_\beta}{kT}$:

$$n_o(n) = \sum_\beta \sqrt{\frac{kTm^*_{n,\beta}}{2\pi^2\hbar^2}}e^{-E_\beta/kT}\int_0^{\infty} t^{-1}e^{-t}dt \tag{3.39}$$

The integral in Eqn. 3.39 can be recognized as the gamma function with an argument equal to $\frac{1}{2}$, which makes the integral equal to $\sqrt{\pi}$. The expressions in Eqns. 3.38-3.39 give the one-dimensional carrier concentration. To obtain the intrinsic carrier concentration per unit volume, we calculate the concentration that would arise by stacking quasi-2-dimensional sheets of CNTs directly on top of each other to form a 3-dimensional volume filled with nanotubes. Finally, we arrive at the following formula for the intrinsic carrier concentration, which is a function of the fundamental tube index n:

$$n_o(n) = \sum_{\beta} \frac{1}{(2.49n/\pi)^2} \sqrt{\frac{kTm^*{}_{n,\beta}}{2\pi\hbar^2}} e^{-E_\beta/kT} \tag{3.40}$$

3.6 CNT Electron Affinity

We also need the electron affinities of different size CNTs in addition to the intrinsic carrier concentrations to incorporate the effects of the CNT-Si barrier into the carrier continuity equations. We use the bandgap of the CNT and the electron affinity of the graphite to obtain the electron affinities of the CNTs. We then calculate CNT affinities by subtracting half the bandgap value of the lowest subband of the CNT from the electron affinity of graphite, which is 4.4eV [72]. This results in 3.87eV and 4.16eV for the electron affinities of the n=10 and n=22 CNTs, respectively. Comparing these electron affinities to that of the Si (4.05eV), these CNTs form barriers with opposite signs when they have a junction with the Si.

3.7 Chapter Summary

In this chapter, we described how we obtain the CNT electrical parameters, using an MC simulator. We also gave empirically determined analytical expressions for these parameters such as field, diameter and length dependent electron mobilities, and electron affinities.

The CNT low-dimensional system results in many interesting transport characteristics. First, we have very high densities of states at band energy minima, causing spikes in the scattering rates. However, the overall scattering rate still results in very high electron velocities, approaching 10^8cm/s. In addition, our calculated mobilities are as much as five-to-ten times higher than that of the Si. Thus, embedding CNTs in active device regions may facilitate devices with very high transconductances and drive currents.

Our simulations also show velocity oscillations on the tubes, reaching tens of Terahertz. We attribute this to scattering due to phonons with energies of 160meV and 200meV, leading to voltage controlled, very high frequency oscillators. We believe that if these can be used as high frequency oscillators, like Gunn diodes, they would revolutionize future high frequency RF designs. In addition, we also investigated length effects on average velocities, resolving the quantization effects due to finite lengths of the tubes. Our calculated results show that average velocity first overshoots, and then reaches its steady state value.

In the following chapter, we investigate whether the usage of the CNTs in devices can lead to better device performances. We achieve this by solving for

the semiconductor equations, including the Si and the CNT barrier, transport and quantization effects.

Chapter 4

Carbon Nanotube Embedded Device Modeling

As we approach the end of the semiconductor roadmap, investigators are exploring new paradigms for electronic devices. Carbon nanotubes (CNTs) are being explored as a structure that may play a leading role in future electronic systems [12]-[15]. CNTs are planar graphite sheets (graphene) that are seamlessly wrapped into tubes. CNTs possess favorable electrical characteristics, and can be fabricated in dimensions as small as 8Å in diameter. The electrical characteristics of CNTs vary with the diameter and the wrapping angle of the graphene [16]. Both the diameter and the wrapping angle can be described by the tube's fundamental indices (l,m) (Standard notation uses (n,m); however, l is used here instead of n to avoid confusion with electron concentration). Theory indicates that CNTs can be metallic or semiconducting depending on the fundamental tube indices (l,m), with bandgap of the semiconducting tube inversely proportional to the CNT diameter. Experimental and theoretical analyses show semiconducting CNTs having electron mobilities even higher than $10^5 \text{cm}^2/\text{Vs}$, with peak drift electron velocities that can be as much as five times higher than that of silicon [17]-[21]. It has also been shown that tubes can be doped by donors and acceptors [22]-[24], and low resistance contacts can be made to tubes [25]-[29]. Experiments and calculations also indicate that CNTs may facilitate devices with large transconductances and high drive currents

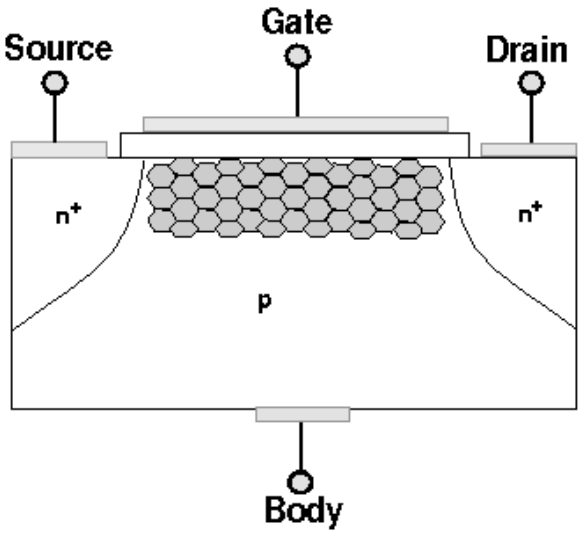

Figure 4.1: Simulated CNT-MOSFET device.

[20]-[40]. Experiments also have demonstrated the viability of CNT-based FETs [34, 35], and CNT-SOI type MOSFETs [36, 37]. Furthermore, preliminary research has been done to model and design CNT embedded bulk MOSFETs [30, 31].

In this chapter, we investigate several hypothetical CNT-MOSFET devices, similar to the one shown in Fig. 4.1. Our calculations indicate that if successfully fabricated, CNT-MOSFETs can have improved device performance over conventional MOSFETs [30, 31]. To investigate the potential attributes of the new design, we developed a methodology for modeling nanoscale CNT-MOSFETs. It includes determination of the electrical characteristics of single wall zig-zag CNTs, and the merging of the CNT results into our quantum device solver. To electrically characterize the CNT, we developed a Monte Carlo (MC) simulator for CNTs. Using the MC simulator described in the previous chapter, we first calculate electron and phonon dispersion relations for single wall zig-zag CNTs with different tube indices l. We then derive the selection rules and the scattering matrix elements, using the

Fermi's Golden Rule. Using the MC results, we derive analytical models for CNT parameters such as mobility and density of states. Once we obtain CNT parameters, we import them to our quantum device solver. Our device solver is based on the semiconductor equations, modified to account for the CNT-silicon (CNT-Si) barrier [38, 39] and quantum effects. We solve these coupled equations on a mesh within our CNT-MOSFET device. The solution gives results, which include CNT-MOSFET current voltage curves, and the electron concentration profile in both the bulk MOSFET and the CNT enhanced channel. In addition, we also do similar analyses for CNT embedded SOI-MOSFETs.

Next, we show the methodology developed to obtain device performance details of CNT-MOSFETs. We first show our algorithm to resolve the quantum and the CNT-Si barrier effects. After we give an insight to our CNT-Si device simulator, we present our calculated results for the CNT-MOSFETs. We then apply the same methodology to CNT embedded SOI-MOSFETs, and show our calculated performance details for these devices.

4.1 Quantum Modeling and Proposed Designs of Carbon Nanotube (CNT) Embedded Nanoscale MOSFETs

We propose a novel MOSFET design that embodies single wall zig-zag semiconducting Carbon Nanotubes (CNTs) in the channel. Investigations show that CNTs have high low-field mobilities, which can be as great as $4\times10^4 \mathrm{cm}^2/\mathrm{Vs}$. Thus, we expect that MOSFET performance can be improved by embedding CNTs in

the channel. To investigate the performance of a newly proposed CNT-MOSFET device, we develop a methodology that connects CNT modeling to MOSFET simulations. Our calculations indicate that by forming high mobility regions in the channel, MOSFET performance can be boosted. However, barriers formed between the CNT and the Si due to the variations of the bandgaps and the electron affinities can degrade MOSFET performance improvements. Our calculations were obtained by building on our existing CNT Monte Carlo (MC) simulator [17, 18] and quantum based device solver [30, 31].

4.1.1 Quantum CNT-Silicon Device Simulator

We develop a two-dimensional quantum device solver based on the Poisson equation and the modified semiconductor equations. We here take the invariance in the width direction as retained by the introduction of tubes in the channel. Since CNTs in our simulations have small diameters, the bending of the field around the tube is limited [73] and the associated dielectric relaxation lengths are high enough to ensure smooth field curves. The governing equations are listed below in the order of Poisson, quantum/CNT-Si electron current continuity, and quantum/CNT-Si hole current continuity equations.

$$\nabla^2\phi = -\frac{q}{\varepsilon}(p - n + D) \tag{4.1}$$

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla \cdot J_n + GR_n \tag{4.2}$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\nabla \cdot J_p + GR_p \tag{4.3}$$

Here, the variables n (p), J_n (J_p), D and GR_n (GR_p) are electrostatic po-

tential, electron (hole) concentrations, electron (hole) current densities, net dopant concentration, and electron (hole) Shockley-Hall-Read net generation-recombination rates, respectively. We next define electron and hole current densities J_n and J_p as follows:

$$J_n = -qn\mu_n\nabla\left(\phi + \phi_{\mathrm{QM}} + \phi^n_{\mathrm{HS}}\right) + \mu_n kT\nabla n \tag{4.4}$$

$$J_p = -qp\mu_p\nabla\left(\phi - \phi_{\mathrm{QM}} - \phi^p_{\mathrm{HS}}\right) - \mu_p kT\nabla p \tag{4.5}$$

We here symbolize electron and hole mobilities by μ_n and μ_p, respectively. We also introduce two additional effective potential terms ϕ_{QM} and ϕ_{HS} to account for the quantum and the CNT-Si barrier effects, respectively. We next will discuss how these two phenomena are taken care of by the effective potential terms, beginning with the CNT-Si barrier effects.

Solution of the CNT-MOSFET system requires proper handling of two phenomena. The first is the effect of the quantum well formed at the Si-SiO_2 interface that causes band splitting, thus lowering the carrier concentration. Second one is the influence of the barrier formed at the CNT-Si interface that results from the difference in bandstructures and electron affinities of the CNT and the Si. A quantum well may also form at the CNT-Si junction due to the band-offsets.

As an initial guess, we first solve our system without considering quantum confinement effects. This translates to the coupled solution of Eqns. 4.1-4.5. At this stage, we resolve the effects of CNT-Si barrier through the use of revised current equations, given in Eqns. 4.4 and 4.5, with the following effective potential terms:

$$\phi^n_{\mathrm{HS}} = \frac{1}{q}\left(\chi - \chi^{\mathrm{Si}}\right) + \frac{kT}{q}\ln\frac{n_o}{n_o^{\mathrm{Si}}} \tag{4.6}$$

$$\phi^{p}_{\mathrm{HS}} = -\frac{1}{q}\left(\chi + E_G - \chi^{\mathrm{Si}} - E_G^{\mathrm{Si}}\right) - \frac{kT}{q}\ln\frac{n_o}{n_o^{\mathrm{Si}}} \tag{4.7}$$

$$\phi_{\mathrm{QM}} = 0 \tag{4.8}$$

Here, n_o is the intrinsic carrier concentration at a grid point on our device, and n_o^{Si} is the intrinsic carrier concentration of silicon. We note that n_o takes on either the intrinsic carrier concentration of the CNT or the Si, depending on the location within the CNT-MOSFET. Also, χ is the electron affinity at a grid point on our device and is either equal to χ^{Si} or χ^{CNT}. We subtract χ^{Si} from χ, because our reference material is the Si. In addition, E_G, like χ and n_o, refers to the same material in space. It takes on the bandgap value of either the CNT or the Si depending on the location inside our CNT-MOSFET. Furthermore, we note that this formalism does not account for atomistic bonding details, which could give rise to interface states and complicated junctions. These effects would likely be accounted for in the present model through the Poisson and transport equations, the Fermi level and the mobility.

Investigations show that carrier confinement at the Si-SiO_2 interface and the CNT-Si barrier can significantly reduce the carrier concentration adjacent to these interfaces [50]-[57]. In addition, the potential well formed at the band discontinuities between the CNT and the Si can result in confinement and band-to-band tunneling effects. To incorporate these quantum effects in our device model, we use the density gradient formalism. The density gradient theory is based on an approximate many-body quantum theory [51]. It has been shown that the density gradient theory resolves the effects of the MOSFET channel confinement [52, 53], band-to-band and

source-to-drain tunneling [53]-[56]. In this formalism, quantum effects are included by the introduction of an effective potential term that is proportional to the gradient of the electron density. Using that model, we resolve quantum effects by using non-zero quantum effective potentials in revised current equations 4.4 and 4.4 [51]-[59]. Here, we treat the quantum induced effects in a manner that is analogous to the formation of position dependent heterostructures in the quantum well, using the following effective potential term.

$$\phi_{\mathrm{QM}} = \frac{2\hbar^2}{12q\sqrt{n}}\left[\frac{1}{m_{\parallel}}\frac{\partial^2\sqrt{n}}{\partial x^2} + \frac{1}{m_{\perp}}\frac{\partial^2\sqrt{n}}{\partial y^2}\right] \tag{4.9}$$

Here, x is parallel to the MOSFET channel and tube axis, and y is normal to x. Also, we use the effective mass of the Si or the CNT depending on the direction and location.

The effective potential term can be derived using either the one particle Wigner function or the single particle Schrödinger equation. Furthermore, the one particle Wigner function is the BTE with corrections due to non-local driving potentials [63, 64], as shown below:

$$\frac{\partial f}{\partial t} + \vec{v}\cdot\vec{\nabla_r} f - \frac{2}{\hbar}V(r)\sin\left[\frac{\hbar\overleftarrow{\nabla_r}\vec{\nabla_k}}{2}\right] f = s(k,p,t) + \left.\frac{\partial f}{\partial t}\right|_{\mathrm{coll}} \tag{4.10}$$

Above, $f(k,r,t)$ is the distribution function. We expand the sine, assuming that the argument is small. Next, we multiply the entire equation by $\frac{1}{\Omega}\int w(k)dk$, as described in Section 1.2, to find the moments of the above equation. If we only include the first order term, it gives the BTE. Using the second order term, we obtain the correction factor, shown in Eqn. 4.9, to the electrostatic potential. Additionally,

we can ignore higher order terms since their contributions are very small because of having $\hbar^r$, where r is 4, 6, 8, ..., as coefficients.

We here derive the effective potential in Eqn. 4.9 using the single particle Schrödinger equation shown below:

$$\left[-\frac{\hbar^2}{2m^*}\nabla^2 - q\phi(r)\right]\psi = i\hbar\frac{\partial\psi}{\partial t} \tag{4.11}$$

For a stationary case, the wavefunction ψ can be expressed using the complex expression $|\psi|e^{-i^{\varepsilon}/\hbar t}$. Substituting this form for ψ above, and equating the real parts, we obtain the following expression:

$$\varepsilon = -q\phi - \frac{\hbar^2}{2m^*}\frac{\nabla^2|\psi|}{|\psi|} \tag{4.12}$$

We first note that $|\psi|^2$ is the probability density that gives the electron concentration. We then replace the energy ε by a quantum potential $-q\phi_{\mathrm{QM}}$. These substitutions give the density gradient effective potential for one band, as follows:

$$-q\phi_{\mathrm{QM}} = -q\phi - \frac{\hbar^2}{2m^*}\frac{\nabla^2\sqrt{|\psi|^2}}{\sqrt{|\psi|^2}} \tag{4.13}$$

$$\phi_{\mathrm{QM}} = \phi + \frac{\hbar^2}{2qm^*}\frac{\nabla^2\sqrt{n}}{\sqrt{n}} \tag{4.14}$$

The first equation above is called the Schrödinger-Bohm equation. It gives the density gradient equation 4.14 for a pure state [59].

Using a combination of numerical methods, we finally solve our coupled quantum semiconductor Eqns. 4.1-4.5 along with Eqns. 4.6-4.7, for the electrostatic potential, quantum/CNT-Si electron concentration, and quantum/CNT-Si hole concentration for the CNT-MOSFET. More specifically, at each grid point on our mesh,

we first calculate values for the effective heterostructure potentials for the electrons and holes. We then add these effective potentials to the electrostatic potentials ($\phi_{i,j}$) at each grid point; i, j. Next, we use these newly calculated potentials in the Bernoulli functions of the Scharfetter-Gummel discretization scheme, as described in Section 2.1.2. We apply the same method of calculating potentials to find the electron and hole concentrations to be used in the discretized Poisson equation. Since we take the reference as the Si, we use intrinsic carrier concentration n_o of the Si in the semiconductor equations wherever an intrinsic carrier concentration is needed, except for the calculation of the aforementioned heterostructure effective potentials. Next, we solve for the electrostatic potential, and the electron and hole concentrations. To solve for the state variables, we first use the Gauss-Seidel method, and then simultaneously find corrections to all the state variables using the Newton-Raphson method. We obtain the classical solution once the corrections are insignificantly small. At this point, we calculate the quantum effective potentials at each point in the channel of our device. We then add these quantum effective potentials to the electrostatic potentials and the heterostructure effective potentials, and then use the new potential terms to calculate the drift components of carrier's current densities. As before, we first use the Gauss-Seidel method to get an estimate for the solution. For the final tune-up, we use a matrix solver to calculate corrections for the state variables using the Newton-Raphson method. Once the aforementioned variables are determined, we use them to calculate the current-voltage characteristics of the CNT-MOSFET. In summary, we solve the system numerically using the overall algorithm given in Fig. 4.2.

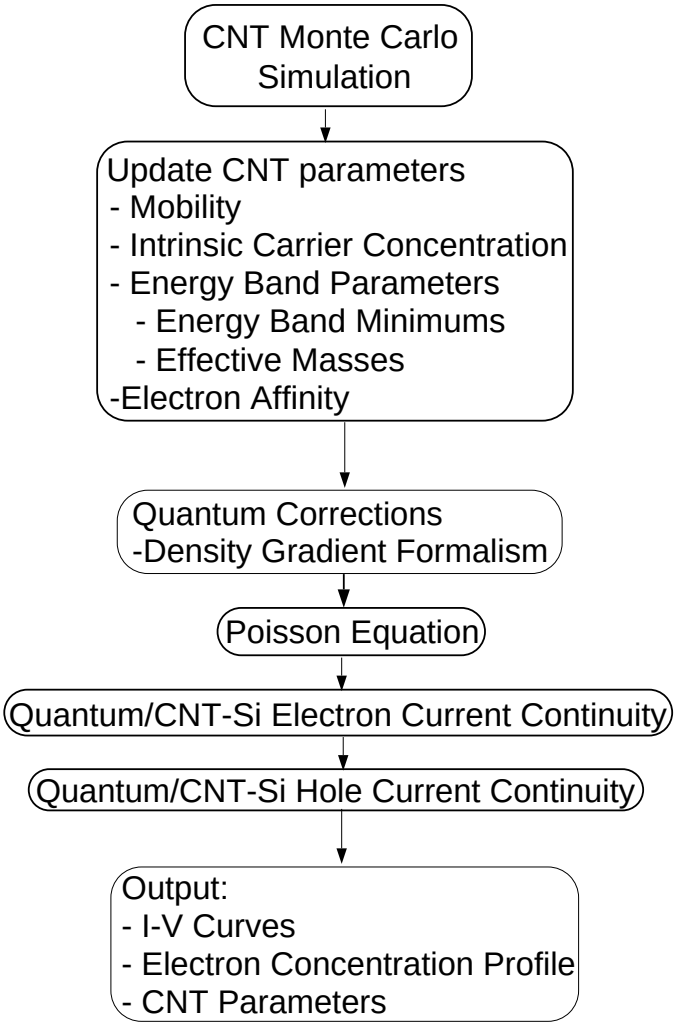

Figure 4.2: Coupled algorithm flowchart.

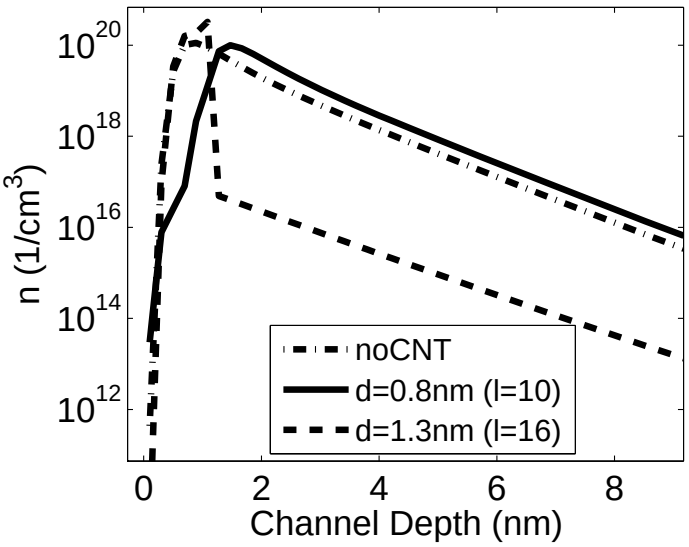

Figure 4.3: Calculated electron concentration profile in the middle of the CNT-MOSFET channel, for different diameter CNTs and V_G=1.5V (V_D and V_S are 0V), starting from the Si-SiO_2 interface and going down about 9nm.

4.1.2 Simulation Results

We applied our modeling methodology to simulate a 0.15μm well-tempered (having a good on/off current ratio) CNT-MOSFET [74]. We first simulated CNT-MOSFETs with a single layer of CNT in the MOSFET channel parallel to the interface as illustrated in Fig. 4.1. The parameter we investigate in these simulations is the effect of different diameter tubes. We next study how incorporating additional layers of 8Å-diameter tubes affects the device characteristics.

In Fig. 4.3, we show our calculated electron concentration in the vertical direction of the MOSFET channel, starting from the Si-SiO_2 interface. We applied 1.5V to the gate terminal, and grounded others. CNT-MOSFET contains one layer of tube. The device with the medium diameter tubes (d=13Å) shows high concentrations in the channel. The abrupt change in the carrier concentration can be

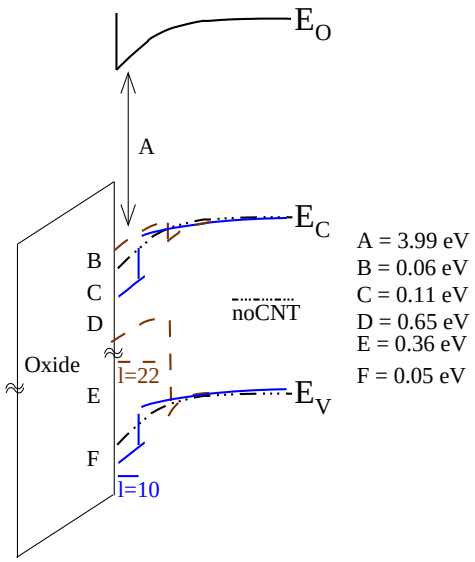

Figure 4.4: Energy-band diagrams of CNT-MOSFETs, with diameters of 0.8nm and 1.3nm, and a MOSFET in the vertical channel direction. Dashed line is the band diagram of a CNT-MOSFET that has l=22 (d=1.3nm) CNTs in its channel. Solid line is the band diagram of a CNT-MOSFET that has l=10 (d=0.8nm) CNTs in its channel. Dot-dash line is the band diagram of the silicon in the vertical MOSFET channel direction.

attributed to the differences in the conduction band offset between the CNT and the Si, as shown in Fig. 4.4. We associate this with the high intrinsic carrier concentration and lower work function (compared to the Si) of the larger diameter tubes, which attract electrons even in the absence of a gate field. On the other hand, the intrinsic carrier concentration of the d=8Å CNT is close to that of the Si, and the CNT has a higher work function. Thus a potential well is formed on the tube which in turn pushes electrons away from the channel of this CNT-MOSFET. Thus, the larger diameter CNTs appear to be likely to sustain large transconductances.

We next investigate whether the band-offsets between the wider tubes and silicon appear to negate the potential improvement of higher electron concentration in the channel of the larger diameter tube CNT-MOSFETs. Therefore, we obtain the

current-voltage characteristics of the 0.15μm CNT-MOSFETs in the subthreshold, linear and saturation regions. In Fig. 4.5(a), we compare the drain current density versus applied drain voltage curves for four MOSFET configurations. One set of curves is for the conventional MOSFET without any CNTs in the channel. The other three sets of curves are for the single layer CNT-MOSFETs with small (8Å), medium (13Å) and large (17Å) diameter CNTs in the channel, just below the SiO_2. We find that for high bias conditions, CNT-MOSFETs utilizing larger diameter tubes attain higher drive currents than the ones having the small diameter tubes, followed by the conventional MOSFET.

One of the main differences in performance between CNT-MOSFETs can be attributed to the height of the barrier formed at the CNT-Si junction. The smaller diameter tubes have less barrier height offset since their intrinsic carrier concentration is closer to that of the silicon. However, the small diameter tubes form a potential well at the channel, unlike the larger diameter tubes that attract more electrons as the diameter gets bigger. The CNT-MOSFETs have improved drive current characteristics over the conventional MOSFET. We attribute these higher currents to larger channel electron concentrations, as shown in Fig. 4.3, and larger mobility values in the CNTs. However, the large diameter tube CNT-MOSFET behaves more like a resistor with a low output resistance due to its band-offset and high mobility. In addition, the small diameter tube CNT-MOSFET has a jump in its current drive around V_{DS}=0.6V, where the electron concentration on the tube suddenly jumps from the levels shown in Fig. 4.3 ($10^{16}cm^{-3}$) to higher values ($10^{18}cm^{-3}$) indicating that new subbands are populated on the tube as we increase the drain bias.

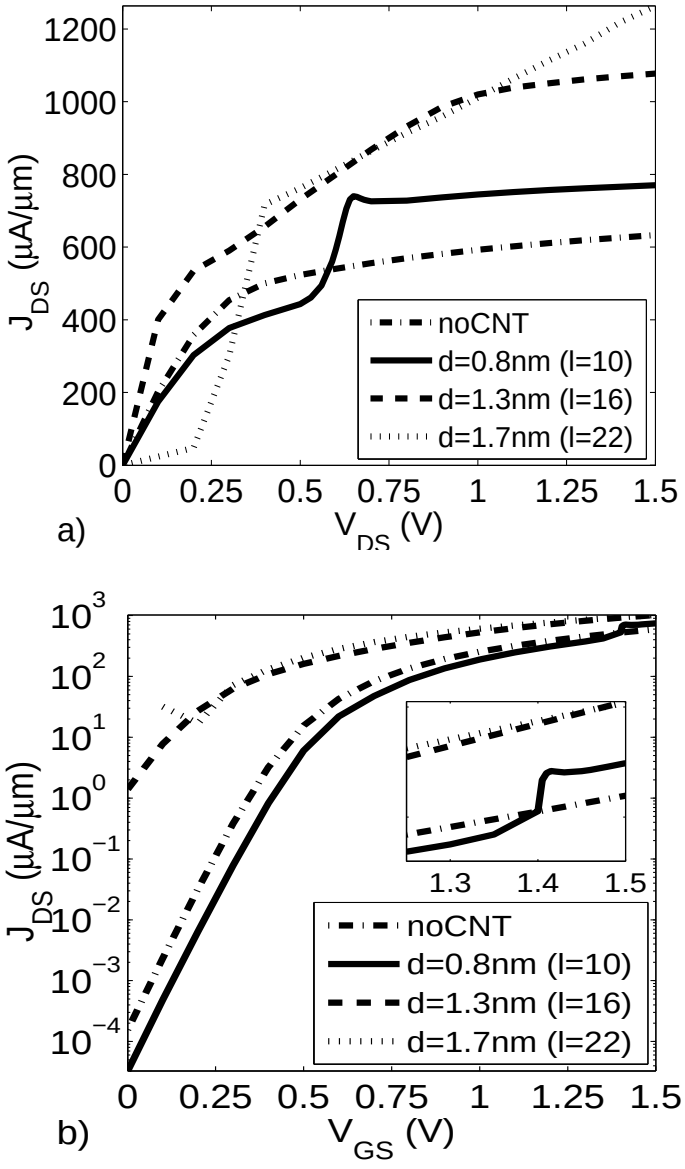

Figure 4.5: Current-voltage curves for CNT-MOSFETs with different diameter CNTs. Calculated currents are for a) V_{GS}=1.5V and b) V_{DS}=1.0V (Inset shows the local maximum point for the d=0.8nm tube CNT-MOSFET around V_{GS}=1.4V.).

We show the subthreshold characteristics of the aforementioned CNT-MOSFETs in Fig. 4.5(b). The small diameter tube CNT-MOSFET has a steep subthreshold slope (like the conventional device) with a lower leakage level and higher drive current when compared to the conventional device at high gate biases. We attribute this to the band-offset and high mobility associated with the small diameter CNTs. Additionally, the small diameter CNT-MOSFET shows negative differential transconductance. We associate this with the occupation of new subbands on the tube as the gate bias increases. For the same bias range, larger diameter tube CNT-MOSFETs have a much higher leakage level which gets worse as the drain bias increases. However the on/off current ratio is still on the order of a thousand, which should enable their use as FETs but may limit their low power applications. We attribute this to the band-offsets and high mobility of the larger diameter tubes.

We next investigate ways to increase the electron concentration in the channel of the small diameter tube CNT-MOSFET to achieve even higher current drives. The small diameter tube device already has improved subthreshold characteristics, which are mainly controlled by the band-offsets at the drain and source sides. However, drive current is controlled by the gate via the electron channel formed in the CNT-MOSFET. Since electron concentration is low on the tube due to confinement, we add extra layers of CNTs in the vertical channel direction to increase the physical size of the well. (The length of the tube is still in the direction of the channel.) Therefore, more electrons can fit in the well. In Fig. 4.6, we show the electron concentration in the channel of the small diameter tube CNT-MOSFET for various numbers of vertically stacked CNT layers. We observe that the confinement effects

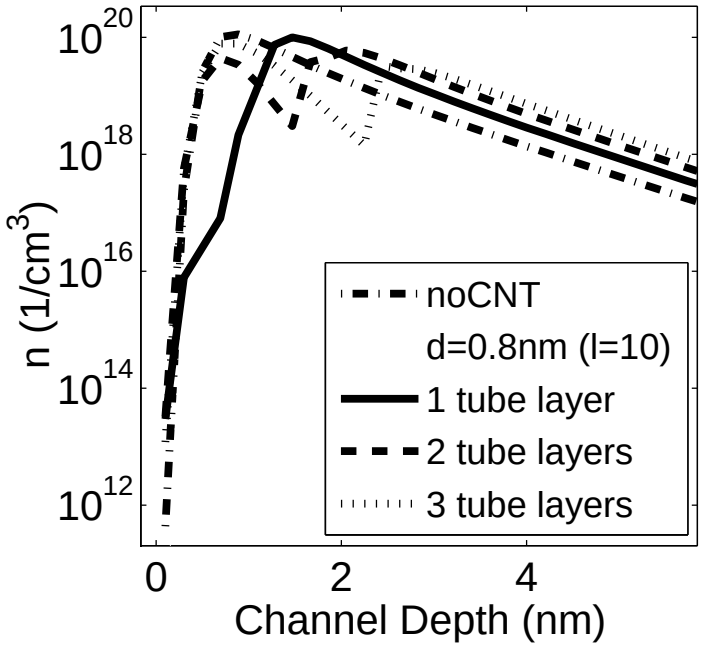

Figure 4.6: Electron concentration profile in the middle of the CNT-MOSFET channel, for different number of CNT layers in the vertical channel direction and V_G=1.5V (V_D and V_S are 0V), starting from the Si-SiO_2 interface and going down about 6nm.

are less pronounced as the number of layers increases from one to three. This enables the peak electron concentration to be on the CNTs, with a highest level reached for the three layered device. Therefore, we expect this to be mirrored in the drive current capabilities. We show the current curves for high gate bias in Fig. 4.7(a), where the highest current is supplied by the three layered CNT-MOSFET. Additionally, the jump in the current drive of the one layered device becomes less pronounced as the number of layers increases. We associate this with less confinement in a well with bigger dimensions, where most of the states are already occupied. In Fig. 4.7(b), we show the subthreshold characteristics of these CNT-MOSFETs. Our calculated currents show performance improvements as the number of layers increases.

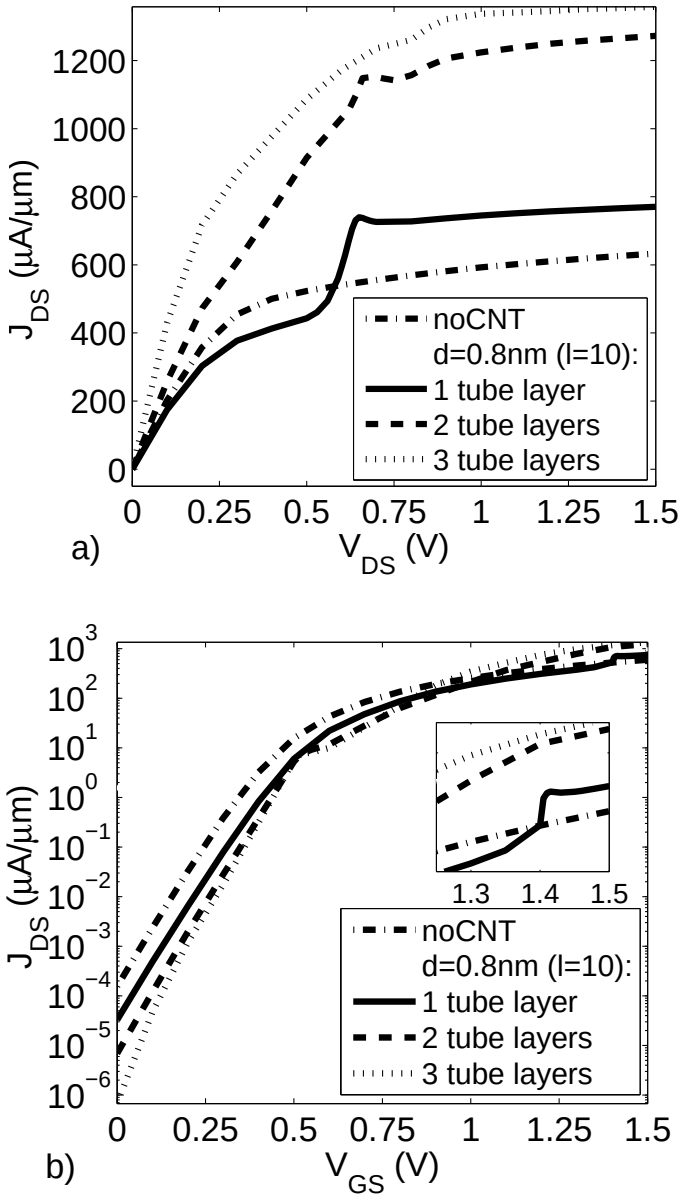

Figure 4.7: Current-voltage curves for CNT-MOSFETs with CNTs of 0.8nm in diameter and varying number of tube layers (planar CNT sheets) in the vertical channel direction. Calculated currents are for a) V_{GS}=1.5V and b) V_{DS}=1.0V (Inset shows the local maximum point for the one layered CNT-MOSFET around V_{GS}=1.4V. Two and three layered CNT-MOSFETs show a weaker local maxima around V_{GS}=0.5V.).

4.1.3 Section Summary

We propose and investigate a novel device structure that combines MOSFET technology with CNT nanostructures. We report that the CNT-MOSFET device appears to yield better performance than the conventional MOSFET. To analyze the new design, we develop a methodology for modeling CNT-MOSFETs. We first employ MC techniques to electrically characterize single wall zig-zag CNTs. We then derive analytical models for important CNT parameters, including mobility and intrinsic carrier concentration. We next develop a methodology for incorporating these CNT characteristics into a quantum device solver. We use the solver to calculate the current-voltage characteristics of CNT-MOSFETs, as well as internal dynamic variables such as quantum/CNT-Si electron concentration, and electrostatic potential. Our new CNT-MOSFET simulator predicts that the drive current of CNT-MOSFETs is higher than that of conventional MOSFETs. Likewise, in the subthreshold region, the narrow diameter tube CNT-MOSFET shows similar performance compared to the conventional device. Therefore, CNT-MOSFETs employing smaller diameter carbon nanotubes outperform other devices.

4.2 Device Behavior Modeling for Carbon Nanotube Silicon-On-Insulator MOSFETs

We offer a methodology for the numerical analysis of carbon nanotube (CNT) embedded silicon-on-insulator (SOI) MOSFETs. We examine CNT-SOI-MOSFETs that have a planar sheet of single-walled zig-zag semiconducting CNTs embedded

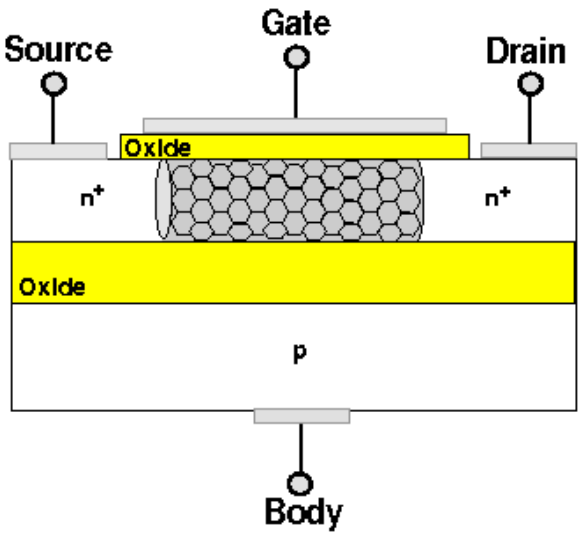

Figure 4.8: Simulated design of CNT-SOI-MOSFET.

along the channel, as shown in Fig. 4.8. To obtain device performance details including current-voltage characteristics, we employ a quantum based device solver [46] along with a Monte Carlo simulator [18]. Our calculated results show that replacing the silicon with CNTs in the channel may significantly improve device performance. The CNT-SOI-MOSFET with the smallest diameter tube may surpass other configurations of CNT-SOI-MOSFETs and conventional SOI-MOSFET in performance if fabricated successfully with the same channel thickness. In addition, under certain conditions, the CNT-SOI-MOSFETs show negative differential resistance.

Here, we first discuss the energy band diagram of CNTs. We then show how we integrate the details of CNT energy dispersion curves into our device solver. We next discuss our methodology, and show our calculated current-voltage characteristics.

4.2.1 Carbon Nanotube Model

To account for the CNT related quantum effects, we need to determine the band-structure of the CNTs. Due to confinement introduced around the circumference when graphene is wrapped into a CNT, the bandstructure splits into a system of subbands. Each of the subbands has a characteristic effective mass, mobility and band energy minima. We determine the energy levels of CNTs by applying zone-folding methods to graphene. The following formula gives the energy dispersion for a zig-zag CNT, which has fundamental tube indices (l,0), as a function of electron momentum along the tube, k_x, and subband index, β, (a, 2.46Å, is the lattice constant of two dimensional graphite.) [16]:

$$E(k_x,\beta) = \pm 3\sqrt{1 + 4\cos\left(\frac{Tk_x}{2}\right)\cos\left(\frac{\pi\beta}{n}\right) + 4\cos^2\left(\frac{\pi\beta}{n}\right)} \qquad \text{(eV)} \qquad (4.15)$$

To extract pertinent information that can be easily integrated into our device simulator, we approximate Eqn. 4.15 by a quadratic energy dispersion relation. Conduction band minimum, effective mass and non-parabolicity factor for the quadratic energy dispersion relation can be calculated using Eqn. 4.15 for different subbands β. For a zig-zag CNT, the total number of subbands are $2l$. In accordance with this, we set the prime values of β to integers from -l to l excluding one of the boundaries. For each subband, conduction band minimum and effective mass can be found by setting k_x to zero and finding the curvature around k_x=0, respectively:

$$E_\beta^l = \left(3\left|1 + 2\cos\left(\frac{\pi\beta}{l}\right)\right|\right) \qquad \text{(eV)} \qquad (4.16)$$

$$\frac{m_\beta^{*l}}{m_o} \cong 0.0910\frac{\left(E_\beta^l/1\text{eV}\right)}{\left|\cos\left(\frac{\pi\beta}{l}\right)\right|} \qquad (4.17)$$

Table 4.1: CNT parameters.

	m^*/m_o	$E_{\min}$ (eV)	$\pm\beta$
l=10			
	0.082	0.53	7
	0.339	1.15	6
	0.208	1.85	8
l=22			
	0.040	0.24	15
	0.112	0.52	14
	0.129	0.93	16

Specifically, we include the statistics of the lowest six CNT subbands, where all the electron transport takes place in our simulations. Among these six subbands, pairs of two subbands have the same energy dispersion curves because -β and β give the same cosine value. We list in Table 4.1 the energy band minima and the effective masses of the lowest three subbands for l=10 and l=22 tubes.

Using an MC simulator similar to the one described in the previous chapter, we obtain velocity versus electric field curves. Using these curves, we derive a diameter and field dependent mobility model [46]. Our MC calculations indicate that the low field electron mobility of l=10 tube is as much as five times higher than that of the silicon. The low field electron mobility of l=22 tube is even higher; it approaches a value ten times higher than that of the silicon.

We next obtain momentum relaxation length versus field curves of the CNTs. Our calculations show that these curves and velocity versus field curves show similar characteristics. Momentum relaxation length versus electric field curves first increase with applied field, reach a peak and then roll off [67]. The peak values of l=10 and

l=22 tubes are approximately 40nm and 100nm, respectively. To avoid ballistic transport, we here simulate sufficiently long CNTs. Therefore, we ensure being in the scattering limited solution domain.

After we obtain CNT characteristics, we import them into our device simulator. We treat the CNT in the device as a material with different bandstructure, intrinsic carrier concentration, electron affinity, electron mobility, etc.

4.2.2 Quantum CNT-SOI-MOSFET Model

We develop a two-dimensional quantum SOI-MOSFET simulator by modifying our quantum bulk device solver [46]. Our simulator is capable of obtaining a coupled solution to the Poisson equation along with the quantum semiconductor CNT/Si electron and hole current continuity equations. We list these equations in the aforementioned order:

$$\nabla^2\phi = -\frac{q}{\varepsilon}\left(p_{\mathrm{QM}} - n_{\mathrm{QM}} + D\right) \tag{4.18}$$

$$\frac{\partial n_{\mathrm{QM}}}{\partial t} = \frac{1}{q}\nabla \cdot J_{n_{\mathrm{QM}}} + GR_n \tag{4.19}$$

$$\frac{\partial p_{\mathrm{QM}}}{\partial t} = -\frac{1}{q}\nabla \cdot J_{p_{\mathrm{QM}}} + GR_p, \tag{4.20}$$

where

$$\begin{aligned}\frac{J_{n_{\mathrm{QM}}}}{q} &= \mu_n \frac{kT}{q}\nabla n_{\mathrm{QM}} \\ &\quad -n_{\mathrm{QM}}\mu_n \nabla\left(\phi + \frac{1}{q}\left(\chi - \chi^{\mathrm{Si}}\right) + \frac{kT}{q}\ln\frac{n_o}{n_o^{\mathrm{Si}}} + \phi_{\mathrm{QM}}\right)\end{aligned} \tag{4.21}$$

$$\begin{aligned}\frac{J_{p_{\mathrm{QM}}}}{q} &= \mu_p \frac{kT}{q}\nabla p_{\mathrm{QM}} \\ &\quad +p_{\mathrm{QM}}\mu_p \nabla\left(\phi + \frac{1}{q}\left(\chi + E_G - \chi^{\mathrm{Si}} - E_G^{\mathrm{Si}}\right) - \frac{kT}{q}\ln\frac{n_o}{n_o^{\mathrm{Si}}} - \phi_{\mathrm{QM}}\right).\end{aligned} \tag{4.22}$$

The Poisson equation 4.18 solves for the electrostatic potential, ϕ, in conjunction with the quantum CNT/Si electron, $n_{\rm QM}$, hole, $p_{\rm QM}$, and net dopant, D, concentrations. In addition, we introduce CNT-Si electron (hole) mobilities, μ_n (μ_p), intrinsic carrier concentration, n_o, electron (hole) Shockley-Hall-Read net generation-recombination rates, GR_n (GR_p), electron affinity, χ, bandgap, E_G, and temperature, T, along with the familiar constants.

To obtain CNT-SOI-MOSFET performance details, we first solve Eqns. 4.18-4.20 (we solve only Eqn. 4.18 within the oxide) in conjunction with Eqns. 4.21-4.22. At this point, we ignore the quantum effects. This gives a modified version of Eqns. 4.21-4.22, which can be obtained by setting $\phi_{\rm QM}$ to zero in the CNT/Si electron and hole current continuity equations, and replacing the subscript QM for quantum by CL for classical. Solving for the classical set of equations, we resolve CNT-Si heterostructure effects including intrinsic variations of CNT/Si bandgaps and workfunctions.

We then include quantum effects to resolve carrier confinement between the gate and buried oxides. Additionally, potential wells at CNT-Si band discontinuities can significantly affect carrier transport phenomena due to confinement and band-to-band tunneling. To resolve quantum effects, we employ the density gradient theory [51]-[59].

We next use a combination of numerical methods to solve Eqns. 4.18-4.20, using the calculated $\phi_{\rm QM}$ values, to obtain CNT-SOI-MOSFET device performance including current-voltage characteristics and carrier concentrations. Moreover, our numerical method is similar to what we use to solve the CNT-MOSFET system.

Using our methodology, we also resolve the additional confinement in the substrate direction between the the gate oxide and the buried oxide.

4.2.3 Simulation Results

We simulated a 0.15μm SOI-MOSFET with a roughly 0.1μm thick buried oxide. We first investigate the effects of a single planar layer of CNT sheet embedded under the gate to fully fill the channel between the two oxide layers, as shown in Fig. 4.8. In this case, device performance is affected by different size channel cavities in the normal direction in addition to different CNTs in the channel with varying electrical parameters. To equate the effects of channel cavity thickness on electron transport, we next embed planar sheets of different diameter CNTs into a channel with a fixed channel thickness. We decide on the channel thickness such that one layer of the biggest diameter tube can fit. Therefore, we obtain comparative analyses of the electrical parameters of different size tubes on electron transport.

In Fig. 4.9(a) and 4.9(b), we show our calculated device performance for the current-voltage and subthreshold characteristics of CNT-SOI-MOSFETs employing various size CNTs. Each CNT-SOI-MOSFET has an associated channel thickness equal to the diameter of the tube used. Among those CNT-SOI-MOSFETs, the one that incorporates the biggest diameter CNT (d=1.76nm, CNT fundamental index l=22) outperforms other configurations by supplying more drive currents in the linear and saturation regions for the two different gate biases (V_{GS}=1.0V, 1.5V). It also has good subthreshold characteristics. We attribute the best device performance

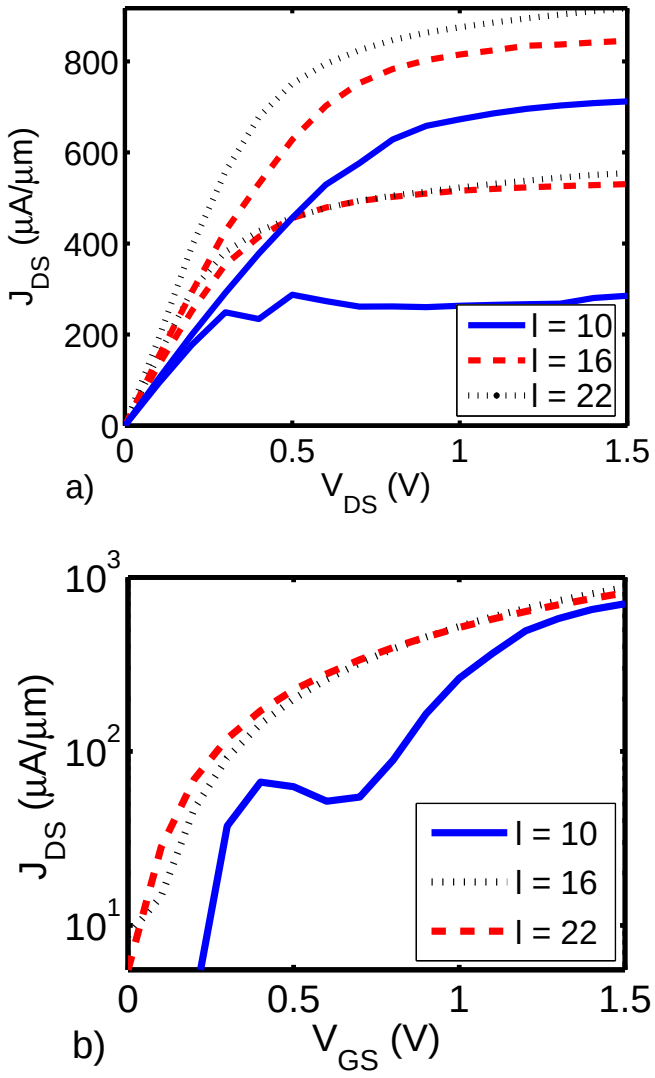

Figure 4.9: a) Current-voltage (V_{GS}=1.0V, 1.5V) and b) subthreshold (V_{DS}=1.0V) characteristics for CNT-SOI-MOSFETs with channel thicknesses equal to the diameter of the tube embedded. (Nanometer scale diameters of l= 10, 16 and 22 tubes are 0.8, 1.28 and 1.76, respectively.)

of l=22 tube embedded CNT-SOI-MOSFET to higher low-field mobilities associated with bigger diameter tubes. (Low-field electron mobility of l=22 CNT is about twice as large as that of the l=10 CNT.) In addition, the lowest diameter CNT (d=0.8nm, l=10), when embedded in an SOI-MOSFET, shows negative differential resistance (NDR). We relate this NDR to high mobilities, band discontinuities between the CNT and the Si, and the smallest cavity formed between the buried oxides.

We then investigate the effects of CNTs on device performance for the same film dimensions, thereby eliminating channel film thickness as a variable on device performance. So, we set the film thickness equal to the diameter of the biggest CNT; therefore, the devices with the largest tubes only have one layer, whereas the l=10 and l=16 devices have film thickness composed of multiple CNT layers. We also simulate one conventional SOI-MOSFET with a silicon film in the channel. In Fig. 4.10(a), our calculated current-voltage curves show that smaller the CNT diameter, the higher the supplied current, with the conventional Si-SOI-MOSFET outperformed by others. We attribute the difference between the SOI-MOSFETs having the Si channel and the ones with CNTs in the channel, to higher mobilities associated with the CNTs, and band discontinuities between the CNT and the Si. Additionally, we relate the difference in the performance of SOI-MOSFETs employing CNTs mainly to the amplitude of the band discontinuities between the utilized CNT and the heavily doped Si terminals.

In Fig. 4.10(b), our calculated subthreshold curves for the devices in 4.10(a) indicate that the CNT-SOI-MOSFET with the lowest diameter tube outperforms other SOI-MOSFETs. As in 4.9(b), it also shows NDR. This is related to low band

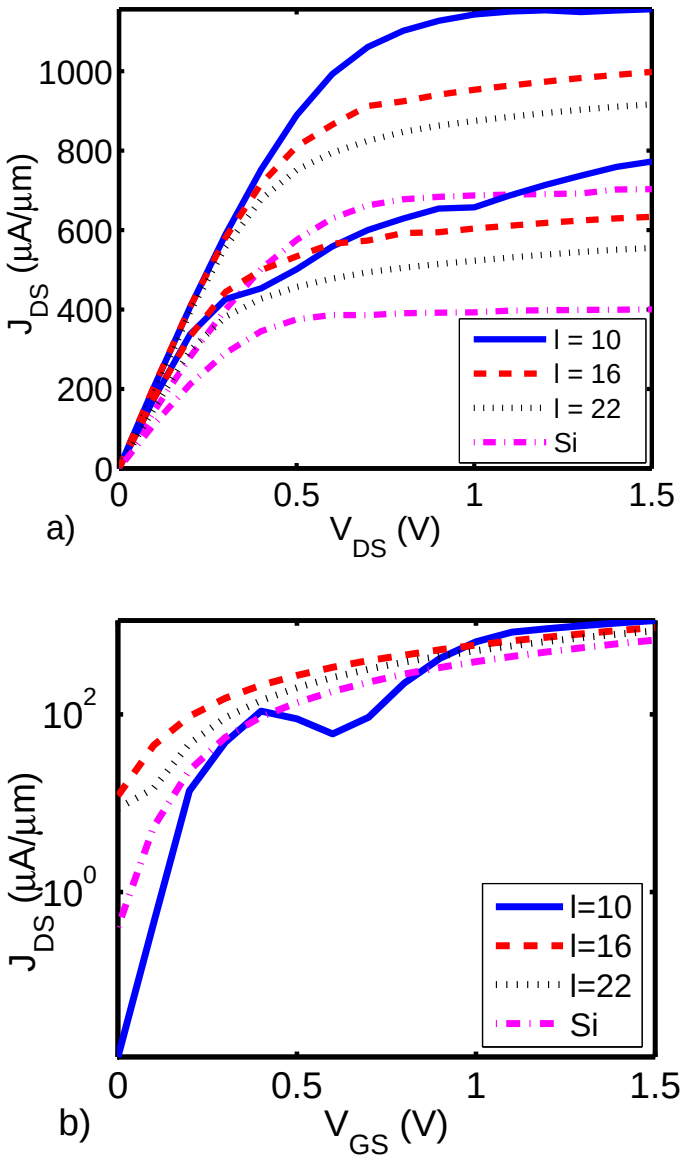

Figure 4.10: a) Current-voltage (V_{GS}=1.0V, 1.5V) and b) subthreshold (V_{DS}=1.0V) characteristics for CNT-SOI-MOSFETs with channel thicknesses equal to 1.76nm, which is the diameter of the biggest tube. (Nanometer scale diameters of l= 10, 16 and 22 tubes are 0.8, 1.28 and 1.76, respectively.)

discontinuity between the l=10 tube and the Si, and high electron mobility on the l=10 tube.

In summary, we have developed a device simulator for modeling CNT-SOI MOSFETs. We find that among devices that have constant film thickness, the small diameter-CNT device yields higher transconductance. On the other hand, devices with one layer of CNTs, with a film thickness equal to the CNT diameter, show that larger diameter-CNT devices have higher transconductance.

4.3 Chapter Summary

In this chapter, we analyzed novel MOSFET designs that include CNTs in their active channels. We suggested these structures because our Monte Carlo (MC) transport simulations of CNTs indicate that they exhibit very high mobilities. Therefore, their usage in the active regions of MOSFET devices may facilitate high current densities, leading to higher transconductances and switching speeds.

We developed novel methodologies to obtain their device performances. To compare them with each other and the traditional all silicon channel devices, we first determine the CNT electrical parameters using an MC simulator. We then analytically or empirically obtain relations for their electrical parameters in terms of external variables and physical properties such as applied field, diameter, or fundamental index. Once the electrical parameters are determined, they are used in the device simulator to resolve interactions between the Si and the CNT. Thus, we determine transport on the tube and in the Si along the channel direction, and

quantization on the tube and in the Si normal to the channel direction. Also, the method we use to resolve quantum effects also resolves the CNT-Si barrier effects and tunneling from source-to-drain.

Our calculated CNT-MOSFET performance figures show that CNT-MOSFETs employing lower diameter tubes outperform the conventional MOSFET, and the CNT-MOSFETs that have bigger diameter tubes in their channels. However, more than one layer of CNT sheets needs to be utilized to achieve such gains. Otherwise, small dimensions of the lower diameter tubes, which are adjacent to the Si-SiO_2 interface, prevents the peak electron concentration being on the tube, resulting in a current flow on the tube that is a small percentage of the total current. Furthermore, we also obtain similar performance results for the SOI-MOSFETs. CNT-SOI-MOSFETs outperform traditional SOI-MOSFETs that have silicon channels. When a single layer of CNT sheet is employed, lower diameter tubes suffer more from smaller thickness quantum wells formed between the gate and the buried oxides. To reduce this confinement effect, more than one layer of smaller diameter tube sheets are used in the channel. This gives the best device performance figures compared to the other SOI-CNT-MOSFETs and the traditional SOI-MOSFETs for the same channel thickness.

In summary, we conclude that CNT-MOSFETs and CNT-SOI-MOSFETs employing lower diameter carbon nanotubes appear to exhibit improved capabilities and, therefore, may represent a new paradigm for devices in the 21^{st} century.

Chapter 5

Integrated Circuit Modeling: Heating Effects

As integrated circuits (ICs) become more densely packed with transistors, manufacturers are facing several important problems threatening chip performance [1]-[11]. One especially important difficulty is chip heating. Investigators have pointed out that toward the end of the semiconductor roadmap, there will be more devices per unit area due to scaling of physical device dimensions. This real estate crowding induces high temperatures, since power density can not be kept in line with the well-known scaling algorithm that guarantees constant power densities between different generations. On the contrary, high device densities cause elevated power densities. According to the traditional device scaling, when device dimensions are scaled downward by a factor of S, all other parameters are scaled by the same factor, either downward (physical features, supply voltage...) or upward (frequency and capacitance per area...), in order to maintain a fixed power density per unit area. However, as dimensions become smaller, manufacturers must deviate from this, and especially from voltage scaling, because of the intrinsic limitations of the silicon bandgap and built-in voltages [1]-[6]. The result is higher power densities because of higher clock frequencies and supply voltages. Additionally, isolation between supply rails gets smaller in nano-devices, leading to higher leakage levels. The chip is also likely to overheat faster than conventional cooling methods can

account for. Thus, power density per unit area keeps increasing exponentially for future electronic devices, making full-chip heating substantially influential in the performance of next generation ICs. Hence, chip heating is considered as one of the major obstacles to be overcome for future IC designs [1]-[11]. (This also depends on the usage of the silicon CMOS technology for future electronics. In this respect, CNT embedded devices may offer faster and cooler alternatives, considering fast electron and heat transport on the tubes.)

To fully understand the chip-heating problem, researchers need modeling tools to simulate and examine the phenomenon. These tools can also be used to relieve heating problems by offering new design approaches to chip layout. Preliminary research has been done to estimate the temperature profile for given chips [6]-[9]. Here, we address the need for a tool that establishes the necessary link between single device operation and the full-chip heating. We present a new methodology for predicting full-chip heating at the resolution of a single device. On the device level, we first obtain electrical characteristics of an n-MOSFET for the given voltage and temperature boundary conditions by self-consistently solving the coupled quantum and semiconductor equations. We then solve the system on the chip level, where the thermal coupling between devices is modeled by a lumped circuit-type thermal network. We obtain the model for the thermal network comprised of passive thermal elements like thermal resistances and capacitances, and heating sources. From the layout design and spatial considerations, we calculate values for the thermal resistances and capacitances between individual devices, and a single device and ground. To determine the strength of each heating source (driving force

in the thermal network corresponding to a single device), we extend the results of the individual MOSFET operation to the entire chip by a Monte Carlo type algorithm. Thus, we account for application and location specific effects of full-chip heating, while achieving the coupling between individual devices and their collective operation. Using our modeling technique, we obtain the effects of power density on full-chip heating and single device performance. To achieve efficient chip designs, we also offer solutions for removing heat from the hottest regions of the chip using thermal contacts.

5.1 Planar Integrated Circuits (ICs): Two-Dimensional (2D)

We report on a novel method for predicting the temperature profile of complex integrated circuits at the resolution of a single device. The proposed new modeling method establishes the necessary link between full-chip heating and non-isothermal device operation for resolving effects of the individual devices on the overall full-chip heating. The technique accounts for the application specific activity levels and the layout placements of individual devices. We use a lumped full-chip heating model that has thermal resistances and capacitances determined by the layout design, and heat sources that are set according to the operational statistics of devices on the chip. To embed the effects of operational statistics for a given application, we use a Monte Carlo type methodology. We analyzed a Pentium III [1] chip considering a realistic layout geometry and averaged activity statistics. Our analysis shows forty three and thirty three degrees Kelvin increases above the ambient for the peak and

median temperatures, respectively.

In Fig. 5.1, we show our device and chip levels, and their interaction. To obtain performance figures at the device and IC levels, we solve coupled device performance equations along with the full-chip heating model. To obtain device performance for the given boundary conditions, we solve the semiconductor equations along with the Schrödinger equation. We next solve the lumped thermal network for the full-chip. Here, we first elaborate on the device model and later on the thermal network.

5.1.1 Device Performance Model

We develop a quantum device solver based on the quantum and semiconductor equations. We list these device equations below starting from the Schrödinger equation, and followed by the Poisson, electron current continuity, hole current continuity, and the lattice heat flow equations. In addition, we have one more equation, which we call the population equation, that gives the density of electrons in the channel by summing contributions from different subbands.

$$E_i\psi_i(y) = -\frac{\hbar^2}{2m^*}\frac{d^2\psi_i(y)}{dy^2} - q\phi(x,y)\psi_i(y) \tag{5.1}$$

$$\nabla^2\phi = -\frac{q}{\varepsilon}\left(p - n + D\right) \tag{5.2}$$

$$\frac{\partial n}{\partial t} = \nabla\cdot\left(-n\mu_n\nabla\phi + \mu_n V_{\mathrm{TH}}\nabla n\right) + GR_n \tag{5.3}$$

$$\frac{\partial p}{\partial t} = \nabla\cdot\left(p\mu_p\nabla\phi + \mu_p V_{\mathrm{TH}}\nabla p\right) + GR_p \tag{5.4}$$

$$C\frac{\partial T}{\partial t} = \nabla\cdot\left(\kappa\nabla T\right) + H \tag{5.5}$$

$$n = \frac{m^*kT}{\pi\hbar^2}\sum_i |\psi_i|^2 \ln\left(1 + e^{(E_F - E_i)/kT}\right) \tag{5.6}$$

The heat flow equation 5.5 provides the coupling between the lattice temper-

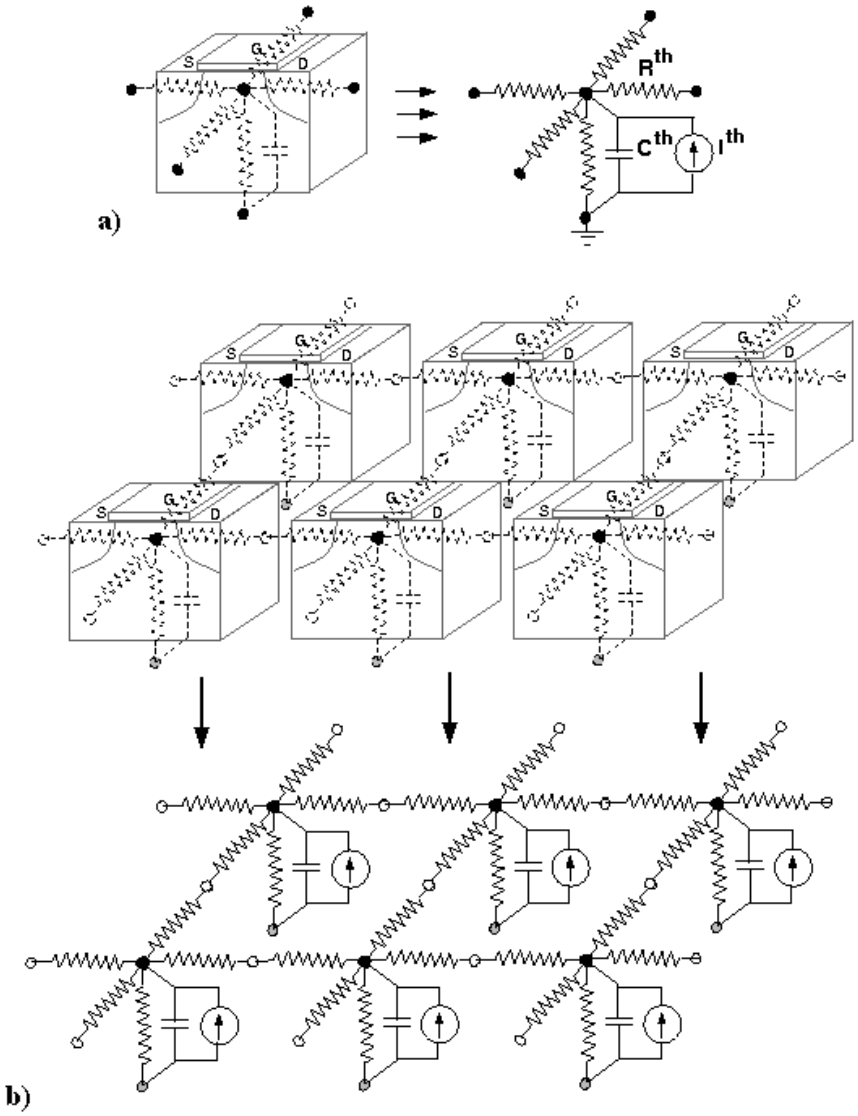

Figure 5.1: a) Each MOSFET device is modeled by a lumped circuit for chip thermal analysis. b) Devices and their interaction are shown. Heat flow between devices causes thermal coupling.

ature and the state variables such as current density and electric field. Here, the major source of heat is Joule heating ($H = -\vec{J} \cdot \vec{\nabla}\phi$, where $\vec{J}$ is the total current density). To help determine the effects of temperature variations within the device on its performance, we explicitly include the temperature dependence on the following parameters [69, 75, 76]: thermal voltage, $V_{\mathrm{TH}}(T)$, intrinsic carrier concentration, $n_o(T)$, electron and hole mobility, $\mu(T)$, electron and hole saturation velocity, $\upsilon_{\mathrm{sat}}(T)$, built-in potentials, $\phi_{\mathrm{built-in}}(T)$, bandgap of silicon, $E_g(T)$, and the thermal diffusion constant , $\kappa(T)$.

$$V_{\mathrm{TH}}(T) = V_{\mathrm{TH}}(T_o)\left(\frac{T}{T_o}\right) \tag{5.7}$$

$$n_o(T) = n_o(T_o)\left(\frac{T}{T_o}\right)^{1.5} e^{\left(-E_g(T)/2kT\right)\left(1-\left(\frac{T}{T_o}\right)\frac{E_g(T_o)}{E_g(T)}\right)} \tag{5.8}$$

$$\mu(T) = \mu(T_o)\left(\frac{T}{T_o}\right)^{-2.5} \tag{5.9}$$

$$\upsilon_{\mathrm{sat}}(T) = \upsilon_{\mathrm{sat}}(T_o)\left(\frac{1+e^{-T/2T_o}}{1+e^{-1/2}}\right) \tag{5.10}$$

$$\phi_{\mathrm{built-in}}(T) = V_{\mathrm{TH}}(T)\ln\frac{n}{n_o(T)} \tag{5.11}$$

$$E_g(T) = E_g(T_o)\left(1 - 2.4\times10^{-4}(T - T_o)\right) \tag{5.12}$$

$$\kappa(T) = \frac{\kappa(T_o)}{\left(1+\frac{D}{2.8\times10^{19}}\right)}\left(\frac{T}{T_o}\right)^{-4/3} \tag{5.13}$$

T_o is the ambient temperature, taken to be 300°K for this work.

We solve device equations 5.1-5.6, with the aid of temperature relations 5.7-5.13, to obtain the non-isothermal device characteristics. More specifically, we first solve equations 5.2 through 5.5 to obtain the semiclassical values of ϕ, n, p, and T throughout the device. Then we include the quantum effects by solving equations 5.1 and 5.6, in addition to equations 5.2 through 5.5, while using the semiclassi-

cal solution as the initial guess. We self-consistently solve these equations for the quantum corrected values of the state variables: ϕ, n, p, T, E_F, and ψ_i [46, 47, 49].

We use the discretization scheme described in section 2.1.2 for the Poisson equation and the continuity equations. To obtain a discretized form for the heatflow equation, we write it as follows:

$$C\frac{\partial T}{\partial t} = \nabla\kappa \cdot \nabla T + \kappa\nabla^2 T + H \tag{5.14}$$

Since we do not consider the transient case for the differential heatflow equation, the left-hand-side of the above equation is zero. We discretize the $\kappa\nabla^2 T$ term as in Eqn. 5.15, and the $\nabla\kappa \cdot \nabla T$ term as in Eqn. 5.16.

$$\kappa\nabla^2 T|_{i,j} =$$
$$2\kappa_{i,j}\left[\frac{(T_{i+1,j}-T_{i,j})}{h_i(h_i+h_{i-1})} + \frac{(T_{i-1,j}-T_{i,j})}{h_{i-1}(h_i+h_{i-1})} + \frac{(T_{i,j+1}-T_{i,j})}{k_j(k_j+k_{j-1})} + \frac{(T_{i,j-1}-T_{i,j})}{k_{j-1}(k_j+k_{j-1})}\right] \tag{5.15}$$

$$\nabla\kappa \cdot \nabla T|_{i,j} =$$
$$\left[\frac{h_{i-1}(\kappa_{i+1,j}-\kappa_{i,j})}{h_i(h_{i-1}+h_i)} + \frac{h_i(\kappa_{i,j}-\kappa_{_{i-1,j)}})}{h_{i-1}(h_{i-1}+h_i)}\right]\left[\frac{h_{i-1}(T_{i+1,j}-T_{i,j})}{h_i(h_{i-1}+h_i)} + \frac{h_i(T_{i,j}-T_{i-1,j})}{h_{i-1}(h_{i-1}+h_i)}\right] +$$
$$\left[\frac{k_{j-1}(\kappa_{i,j+1}-\kappa_{i,j})}{k_j(k_{j-1}+k_j)} + \frac{k_j(\kappa_{i,j}-\kappa_{_{i,j-1)}})}{k_{j-1}(k_{j-1}+k_j)}\right]\left[\frac{k_{j-1}(T_{i,j+1}-T_{i,j})}{k_j(k_{j-1}+k_j)} + \frac{k_j(T_{i,j}-T_{i,j-1})}{k_{j-1}(k_{j-1}+k_j)}\right] \tag{5.16}$$

Above, the spacing in x and y directions are $h_i = x_{i+1} - x_i$ and $k_j = y_{j+1} - y_j$. Moreover, the source term, which is the heat generated, is $H = -\vec{J} \cdot \vec{\nabla}\phi$, and it is discretized as shown below, denoting the current densities in x and y directions by J_x and J_y:

$$\begin{aligned} H_{i,j} = &-(J_{n_{x_{i,j}}} + J_{p_{x_{i,j}}})\left[\frac{h_{i-1}(\phi_{i+1,j}-\phi_{i,j})}{h_i(h_{i-1}+h_i)} + \frac{h_i(\phi_{i,j}-\phi_{i-1,j})}{h_{i-1}(h_{i-1}+h_i)}\right] \\ &-(J_{n_{y_{i,j}}} + J_{p_{y_{i,j}}})\left[\frac{k_{j-1}(\phi_{i,j+1}-\phi_{i,j})}{k_j(k_{j-1}+k_j)} + \frac{k_j(\phi_{i,j}-\phi_{i,j-1})}{k_{j-1}(k_{j-1}+k_j)}\right] \end{aligned} \tag{5.17}$$

To obtain a solution for the Schrödinger equation at each grid point along the channel direction starting from the Si-SiO_2 interface and going down the substrate, we solve the following matrix [77] for its eigenvalues ψ and eigenenergies E that correspond to different subbands i (Below, subscripts refer to locations, except for the one used for E):

$$\begin{pmatrix} (2t-V_1) & -t & \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \ddots & \ddots & \ddots & \cdots & \cdots & \cdots \\ \vdots & \vdots & -t & (2t-V_l) & -t & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & -t & (2t-V_N) \end{pmatrix} \begin{pmatrix} \psi_1 \\ \vdots \\ \vdots \\ \vdots \\ \psi_N \end{pmatrix} = E_i \begin{pmatrix} \psi_1 \\ \vdots \\ \vdots \\ \vdots \\ \psi_N \end{pmatrix} \tag{5.18}$$

Here, t is $\frac{\hbar^2}{2m^*\Delta^2}$, where $\Delta(\cong 1.5\text{Å or } 2\text{Å})$ is the uniform spacing. Also, potential energy V_l at grid point l, which is one of N (200 or 150) points, is $E_{c_l} - q\phi_l$ in terms of the electrostatic potential ϕ and the subband energy minimum E_c. Since we only have silicon in the channel, we set E_c to zero for all l.

To determine a Fermi level for each of the one-dimensional lines along the channel that we solve the above matrix on, we use the population equation 5.6. Since there is no current flow in the vertical channel direction, there is only one Fermi potential for all points on a line along that direction. Thus, we only solve the population equation for the Fermi level just below the Si-SiO_2 interface at y_1, using the Newton-Raphson method, and the function f and its derivative.

$$n(y_1) = \frac{m^* kT}{\pi\hbar^2} \sum_i |\psi_i(y_1)|^2 \ln\left(1 + e^{\left(E_F^k - E_i\right)/kT}\right) \tag{5.19}$$

$$f = n(E_{Fo}) - n(E_F^k) \tag{5.20}$$

$$\frac{\partial f}{\partial E_F^k} = -\frac{m^*}{\pi\hbar^2} \sum_i |\psi_i|^2 \frac{1}{1 + e^{(E_i - E_F)/kT}} \tag{5.21}$$

We add the correction terms $-f/\left(\frac{\partial f}{\partial E_F^k}\right)$ to the previous value of the Fermi potential E_F^k to find the updated value E_F^{k+1}, until the Fermi potential gives the electron concentration calculated from the continuity equation, using the current values of ψ_i and E_i.

Furthermore, our analyses show that non-isothermal MOSFET operation is affected mostly through carrier mobility, saturation velocity and built-in boundary potentials. As temperature increases, current decreases due to mobility reduction, and decreases slightly due to carrier saturation velocity and built-in boundary potentials (increasing temperature effectively lowers the threshold voltage). Thus, as temperature increases, current decreases for moderate temperatures, which are in the operating range of most of today's devices. However, for high temperatures such as 100° above the ambient, the effects of intrinsic carrier concentration may play a leading role and the MOSFET might run into a condition much like thermal runaway in *pn* junctions. (Intrinsic carrier concentration has an exponential dependency on temperature. Thus it appears that it is likely to have the strongest influence on the device performance. However, investigations have shown that although that might be the case in *pn* junctions, it is not the case in MOSFET devices unless temperature increases to such high levels where the control of the gate over the channel is lost due to an abundance of intrinsic carriers for transport.)

5.1.2 Full-Chip Heating Model

We obtain the temperature map of the full-chip by solving the heat flow equation, using the heat produced by the chip transistors as input. We transform the differential heat flow equation given in Eqn. 5.5 to a lumped heat flow equation [7, 78]. We do this to overcome the difficulties introduced by finite differences in the scales of a single device and the full-chip, where the dimensions of the full-chip are thousands of times larger than the corresponding dimensions in a MOSFET. Using the differential heat flow equation for the full-chip requires too many mesh points and is not practical for our application.

To adapt heat flow equation 5.5 into a form that is suitable for the entire chip, it is beneficial to employ the following Kirchoff's transformation [78].

$$\overline{T} = T_o + \frac{1}{\kappa(T_o)} \int_{T_o}^{T} \kappa(\tau) d\tau \tag{5.22}$$

Substituting $\kappa(T)$ with $\kappa(T_o)\left(\frac{T}{T_o}\right)^{-4/3}$, evaluation of the above integral gives:

$$\overline{T} = T_o \left[4 - 3\left(\frac{T}{T_o}\right)^{-1/3} \right] \tag{5.23}$$

Accordingly, temperature T can be written in terms of $\overline{T}$ as shown below:

$$T = T_o \left(1 - \frac{\left(\overline{T} - T_o\right)}{3T_o} \right)^{-3} \tag{5.24}$$

Furthermore, using the above relation, the derivative of T is related to the derivative of $\overline{T}$ as follows:

$$\partial T = -3T_o \left(1 - \frac{\left(\overline{T} - T_o\right)}{3T_o} \right)^{-4} \frac{\partial \overline{T}}{-3T_o} \tag{5.25}$$

$$= \left(\frac{T}{T_o}\right)^{4/3} \partial \overline{T} \tag{5.26}$$

Next, we substitute ∂T with $\left(\frac{T}{T_o}\right)^{4/3}\partial\overline{T}$, and $\kappa(T)$ with $\kappa(T_o)\left(\frac{T}{T_o}\right)^{-4/3}$ in the differential heatflow equation.

$$C\left(\frac{T}{T_o}\right)^{4/3}\frac{\partial\overline{T}}{\partial t} = \nabla\cdot\kappa(T_o)\left(\frac{T}{T_o}\right)^{-4/3}\left(\frac{T}{T_o}\right)^{4/3}\nabla\overline{T} + H(T) \tag{5.27}$$

This leads to the modified differential heat flow equation in terms of the new temperature variable $\overline{T}$:

$$\overline{C}\frac{\partial\overline{T}}{\partial t} = \kappa(T_o)\nabla^2\overline{T} + H \tag{5.28}$$

The benefits of the applied transformation can be seen in Eqn. 5.28, where $\kappa(T_o)$ no longer varies with temperature, but is only evaluated at the ambient temperature of T_o=300°K. (Here, we mostly concentrate on the steady-state case, where the time derivative is zero. For the transient case, heat capacity is the specific heat, which is 0.7J/gK for the silicon, with the added temperature coefficient if Eqn. 5.28 is employed.)

We then integrate Eqn. 5.28 around our unit device, a single MOSFET, assuming that the thermal diffusion constant does not change much within the volume of interest:

$$\overline{C}\int_V\frac{\partial\overline{T}}{\partial t}dV = \kappa_o\int_V\nabla^2\overline{T}dS + \int_V HdV \tag{5.29}$$

Using Stoke's theorem, the first volume integral on the left-hand-side can be written in terms of a surface integral:

$$\overline{C}\int_V\frac{\partial\overline{T}}{\partial t}dV = \kappa_o\int_S\nabla\overline{T}dS + \int_V HdV \tag{5.30}$$

We enclose the MOSFET by a rectangular prism. Here, V and S are the volume and the six faces of that prism shown in Fig. 5.2, respectively. Each MOSFET

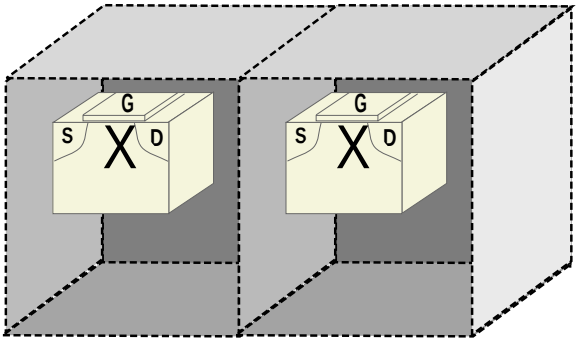

Figure 5.2: We enclose each MOSFET by a rectangular prism to derive the lumped model. Here, the two enclosing prisms for two adjacent MOSFETs are shown, with X showing their centers of heat generation.

is then represented by a single thermal node at its center of heat generation, which is in the MOSFET channel and closer to the drain junction where Joule heating peaks, as shown by an X in Fig. 5.2. Next, we assume that temperature changes linearly between MOSFET centers of heat generation. Also, we note that heat flows in the direction of decreasing temperature, thus $-\kappa\nabla\overline{T}$ represents the heat flux. We take time and space derivatives of temperature as constant in the volume and on the given face, respectively. Taking the integrals in Eqn. 5.30, we obtain:

$$\overline{C}V\frac{\Delta\overline{T}}{\Delta t}+\sum_{f=1}^{6}\frac{\kappa_o\Delta\overline{T}_f S_f}{\Delta l_f}=\int_V H dV \tag{5.31}$$

Here l_f and $\Delta\overline{T}_f$ are the distance and temperature difference between the centers of adjacent prisms going normal to one of the six faces S_f. $\Delta\overline{T}$ shows the transformed temperature variation at the mid-point of that prism. The expression in Eqn. 5.31 is analogous to a KCL type nodal equation, where terms on the left hand side are capacitive and resistive components of the network, while the right side is the source term like a current source in the KCL network. Thus taking $\overline{T}$ analogous to voltage, we can write equivalent thermal resistances, capacitances and

current sources, as follows:

$$C^{\text{th}} = \overline{C}V \tag{5.32}$$

$$R_f^{\text{th}} = \frac{\Delta l_f}{\kappa_o S_f} \tag{5.33}$$

$$I = \int_V H dV \tag{5.34}$$

One capacitive and six resistive components connect the device to other devices and ground. We calculate values for thermal resistances and capacitances from the layout design and the geometrical considerations. In addition to the geometrical considerations, we use values of 1.5K/Wcm and 0.015K/Wcm for the room temperature thermal diffusion constants of silicon and SiO_2, respectively. We then use Eqn. 5.31 to find the temperature for the calculated resistances, capacitances and source term. We obtain the Joule heating from MOSFET simulations using the actual temperature as described in the previous section. As a reminder, we obtain MOSFET performance for given boundary conditions and then extend these results to the chip surface. We get Joule heating for each device by a Monte Carlo type methodology.

Once we have the values of resistances and sources for all the nodes, we obtain the temperature that corresponds to each node or device by solving a KCL-type equation for each node (i,j), where the number of nodes typically equals to the number of transistors of the chip.

$$C_{i,j}^{\text{th}} \frac{(\overline{T}_{i,j}^{k} - \overline{T}_{i,j}^{k-1})}{\Delta t} + \frac{\overline{T}_{i,j}^{k}}{R_{i,j}^{\text{th}}} + \frac{(\overline{T}_{i,j}^{k} - \overline{T}_{i\pm1,j}^{k})}{R_{i\pm 1/2,j}^{\text{th}}} + \frac{(\overline{T}_{i,j}^{k} - \overline{T}_{i,j\pm1}^{k})}{R_{i,j\pm 1/2}^{\text{th}}} = I_{i,j}^{k}(T_{i,j}^{k-1}) \tag{5.35}$$

Here, $R_{i+1/2,j}^{\text{th}}$ is the resistance between nodes (i,j) and $(i+1,j)$.

5.1.3 Coupled Device and Full-Chip Heating Model: Methodology

To obtain the temperature profile of the chip, and its effect on device current-voltage characteristics, we self-consistently solve the device equations along with full-chip heating equations. The solution necessitates convergence at the device level and the chip level. To achieve convergence, we employ the following algorithm for a given digital chip:

Set up Chip Geometry

For a given digital chip, we first decide on spatial chip resolution. We then define the $R^{\text{th}}C^{\text{th}}$ thermal network in conjunction with the chip layout and the device geometry. This includes calculation of thermal resistances and capacitances shown in Fig. 5.1, where each node represents a device. The power supplied to the thermal network is the heat produced by each transistor. The heat sources are represented by the current sources at each node in the network. We then divide our chip into functional blocks (cache, floating point unit, execution unit, clock, etc.).

Determine normalized power per area generated in each functional block

We obtain the percentage of total power consumed in each block. We later normalize each block's power percentage by its corresponding area percentage. Thus we obtain an estimate of the likelihood of finding an active device in that block relative to others. We next renormalize the power per area for each block by the maximum power per area calculated

for a block. Therefore, we determine the comparative activity levels.

Statistically determine normalized power for each transistor on the chip

To determine normalized power for each transistor on the chip, we use a statistical Monte Carlo type methodology. In each functional block of a working IC, some devices will be turned on, and others will be turned off. To determine the number of devices that are on, we use a probability density function. We divide the probability density function into two parts: one gives the on-probability, the other gives the off-probability. We then weight the on-probability density function by the normalized power per area of the particular functional block we are concerned with. We weight the off-probability by the complement of the normalized power per area (1 - normalized power per area). This gives statistically the relative power consumed by each device on the chip.

Calculate the unit response temperature for the entire chip

We next find the unit response by taking all power "current" sources to have unit strength multiplied by the probability weighting factor described in the paragraph above. We then solve for the nodal temperatures of the $R^{\text{th}}C^{\text{th}}$ thermal network. Lastly, we obtain the value of the median temperature corresponding to the unit input.

Find initial value for heat produced by representative device

For our initial device conditions, we take the temperature to be equal

to 300°K (room temperature). We solve device equations 5.1-5.6 (using 5.7-5.13), and calculate the Joule heating for the room temperature boundary condition, as well as the current-voltage characteristics of the device. We next weight the calculated Joule heating by the percentage on-time during switching, to adjust total Joule heating for one clock cycle.

Calculate temperature profile of the entire chip

We next calculate the median temperature of the chip using the calculated Joule heating. To obtain the median temperature, we make use of the linearity of $R^{\text{th}}C^{\text{th}}$ thermal network equations. The linearity allows us to multiply the chip temperature obtained from the unit heat input by the calculated Joule heating of the single device. This scales all the distributed heat sources for the entire chip by the average Joule heating of the distribution. This gives us the temperature as a function of position on the chip.

Mixed-mode solution

We then update the temperature boundary condition of the representative device, and perform device simulation to find the Joule heating that satisfies the device equations for the new temperature boundary. We next calculate the temperature that satisfies thermal network equations for the updated Joule heating (by multiplying the unit response by the calculated Joule heating). To get a self-consistent solution, we iterate

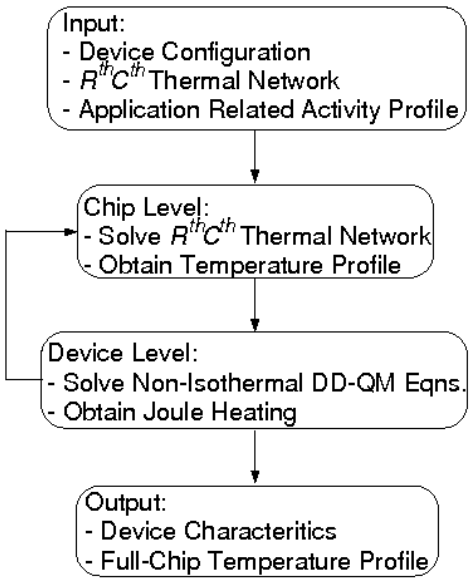

Figure 5.3: Coupled algorithm flowchart.

between the device and IC levels until the heating and temperature figures are consistent. We obtain the final temperature profile of the entire chip by scaling all device temperatures by the temperature ratio found for the representative device.

We summarize our algorithm in Fig. 5.3.

5.1.4 Coupled Device and Full-Chip Heating Model: Application and Results

To test our technique, we apply it to an integrated circuit that is modeled after a Pentium III processor. The block diagram of the example chip is given in Fig. 5.6(a). We use a 0.13μm well-tempered MOSFET given by [74] as our fundamental transistor unit. We first set up our thermal network. We roughly

Table 5.1: Percentage areas and powers of functional blocks in a Pentium III chip [79, 80].

Pentium III Unit	**Percentage Area**	**Percentage Power**	**Normalized Power/Area**
Clock (CLK)	1.0	5.2*	1.0
Issue Logic (ISL)	9.5	14.1	0.71
Memory Order Buffer (MOB)	3.3	4.7	0.68
Register Alias Table (RAT)	3.3	4.7	0.68
Bus Interface Unit (BIU)	4.3	5.9	0.66
Execution Unit (EU)	9.5	13.0	0.66
Fetch	12.5	16.9	0.65
Decode Unit (DU)	14.6	17.2	0.57
L1 Data Cache (L1C)	12.5	9.8	0.38
L2 Data Cache (L1C)	29.8	8.5	0.14

* 40% consumed in the clock block, 60% consumed in the clock network throughout the chip.

estimate that there are forty million devices in an area of one square centimeter. Our geometry yields values of 70 Degree-Kelvin/Watt (K/W) for the mutual thermal resistances, and 5×10^5K/W for the thermal resistance connected to the ground, including package resistance. We take devices to be uniformly distributed on the surface. Percentage areas and powers of each block in Fig. 5.6(a) are written in Table 5.1.

For a single MOSFET occupying an area of approximately $4\mu\text{m}^2$, the $R^{\text{th}}C^{\text{th}}$ thermal network translates into a system of forty million KCL equations (corresponding to forty million MOSFETs, with the generated heat for each modeled by a current source). This is a very large numerical problem. To solve the KCL equations, we first reduce the size of the system [81] using a Norton equivalent circuit on a sub-block of twelve by twelve nodes. At each side of the block, we introduce

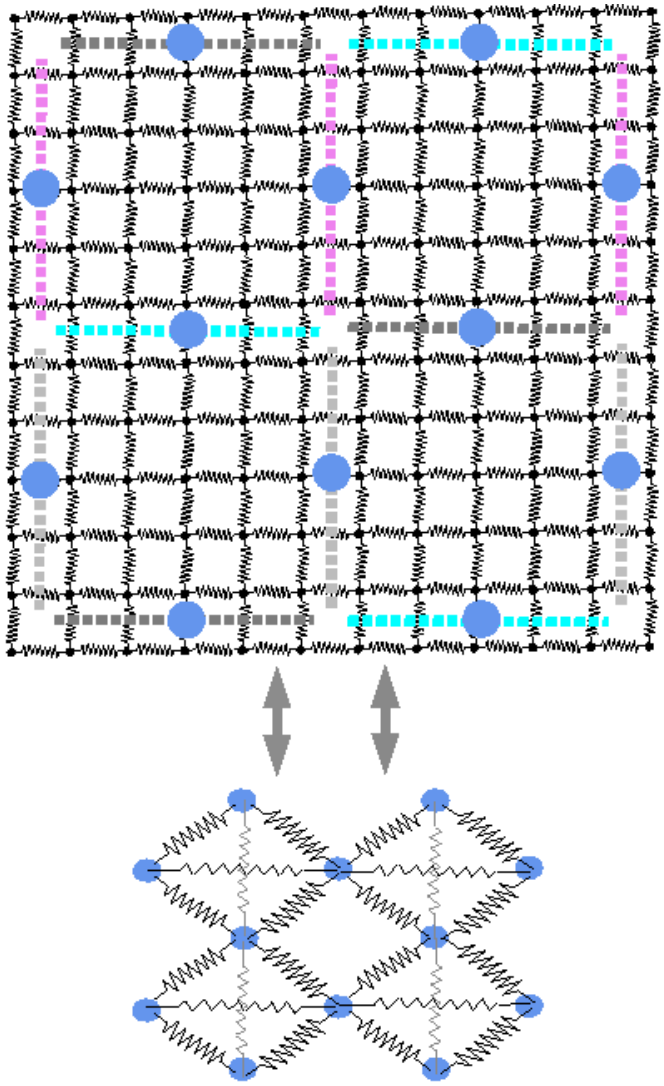

Figure 5.4: Size reduction methods are applied on a subblock of five by five. We obtain four-port Norton representation of each block and use that representation instead, as shown at the bottom of the figure.

new nodes that are half the resistance away from the boundary nodes. We then separately short the new nodes introduced on each of the four sides. Size reduction and the formation of new nodes are shown in Fig. 5.4. With the addition of new nodes, we relate voltages at each node of the entire block and the boundaries to currents using the Kirchoff's Current Law (KCL), as follows:

$$\begin{pmatrix} G_1 & G_2 \\ G_2^T & G_3 \end{pmatrix} \begin{pmatrix} V_{\text{boundary}} \\ V_{\text{inside}} \end{pmatrix} = \begin{pmatrix} I_{\text{boundary}} \\ I_{\text{inside}} \end{pmatrix} \tag{5.36}$$

Here, G_1 is a 4×4 diagonal matrix, which for each diagonal entry has one over the sum of all half resistances that are between the boundary node represented by that row and its nearest neighbors. Consequently, each row of G_2 has nonzero entries equal to minus one over half the resistance, if that column corresponds to a closest neighbor to the boundary node represented by that row. Moreover, G_3 represents the KCL equations, written for the inner nodes, in matrix form. It includes the conductances for the inner nodes to their four neighbors and the ground; therefore, each row has five nonzero entries, if it is not one of the closest neighbors to a boundary. Moreover, we know the values of the conductances from the layout, and the inner current sources, which we assume that all have the same strength equal to unity. Next, using the KCL matrix written above, we calculate the impedance matrix Z ($=G^{-1}$):

$$\begin{pmatrix} Z_1 & Z_2 \\ Z_3 & Z_4 \end{pmatrix} \begin{pmatrix} I_{\text{boundary}} \\ I_{\text{inside}} \end{pmatrix} = \begin{pmatrix} V_{\text{boundary}} \\ V_{\text{inside}} \end{pmatrix} \tag{5.37}$$

To obtain the Norton equivalent circuit seen from the boundaries, we first write V_{boundary} in terms of I_{boundary} and I_{inside}, which is a vector of ones —later, this

is scaled by the heat generated of a device.

$$V_{\text{boundary}} = Z_1 I_{\text{boundary}} + Z_2 I_{\text{inside}} \tag{5.38}$$

This equation gives the Thevenin equivalent circuit. Then, we write I_{boundary} in terms of V_{boundary} to determine the Norton equivalent circuit. (Since we calculate strengths of the current sources from device simulations, and take I_{inside} as a vector of ones, the strengths of the independent current sources in the Norton network are proportional to the calculated heat generated of devices, with a proportionality constant equal to one.) The conductance matrix that determines the Norton equivalent conductances between the four boundaries, and a boundary and the ground is Z_1^{-1}. Also, the strengths of the Norton equivalent current sources at the boundaries are determined from $Z_1^{-1} Z_2 I_{\text{inside}}$, which for a square block is a vector with all entries equal to each other due to symmetry. We use the Norton equivalent thermal resistances between the four nodes. To find the $R^{\text{th}}C^{\text{th}}$ thermal network resistance from a node to the ground, we divide the square subblock's calculated Norton equivalent resistance to the ground by two since the same node is shared by a corresponding node of an adjacent block resulting in two parallel ground resistors. Also, the strength of the Norton equivalent current source doubles due to the two parallel Norton equivalent current sources connected to a node in the reduced system. Furthermore, this method reduces the number of equations that needs to be solved from forty million to approximately one million, using a block of 12×12 nodes and replacing it with four new nodes. Once we reduce the number of equations, we solve the KCL system by a bilateral conjugate gradient method for nodal temperatures.

Next, we statistically determine normalized power for each transistor on the chip. As mentioned in the overall algorithm, we first divide the chip into functional blocks. We then determine normalized power per area generated in each functional block. Later, we associate these normalized power per areas with probability density functions that are different for each functional area. Using these probability density functions, we find normalized powers for each transistor on the chip, with one and zero, for the normalized power, meaning that it is always "on" or "off", respectively. More specifically, we attribute a normalized power from 0.5 to 1 to a device that is mostly "on". Likewise, a device that is mostly "off" is associated with a normalized power from 0 to 0.5. Since we consider a uniformly likely "on" or "off" probability, between 0.5 and 1, and 0 and 0.5, our probability density function is comprised of two steps from 0 to 1, having a jump at 0.5, as shown in Fig. 5.5. Next, we determine the magnitudes of these two steps. We first assume that the functional block with the highest normalized power per area has devices that are mostly "on". Therefore, all devices in that block are associated with a power weighting coefficient from 0.5 to 1.0. In Fig. 5.5, the probability density function for that block is denoted by $P_{\max}$. We take the "on" state probability level for that block as 2, to make the normalization factor, or the area enclosed by the function, equal to 1. We next take that maximum normalized power per area as a reference for the "on" states of the other functional blocks. For example, if the ratio of normalized power per area of a functional block to the one with the highest normalized power per area is R, then the "on" state probability level of that functional block is $2R$. Since the total area enclosed by the probability density function is 1, the "off" state probability level is

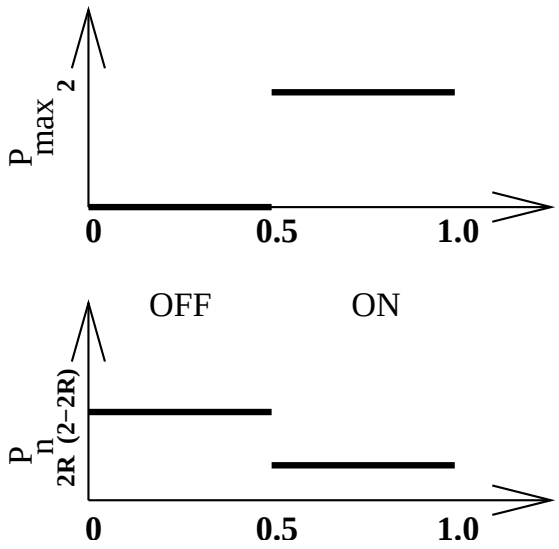

Figure 5.5: Probability density functions for calculating the heat generated of devices in different functional blocks. Top is for a functional block, which has devices that are always mostly "on", Bottom is for any other functional block that has devices in "on" and "off" states.

(2-2R), making the enclosed area equal to 1 ($= 0.5 \times (2 - 2R) + 0.5 \times (2R)$). Next, using the normalized power per areas given in Table 5.1, we calculate the probability density functions for each functional block, where R for each functional block is given in the third column of that table under the title "normalized power/area".

To calculate weighting coefficients probabilistically, using those distributions, we map them to a uniform random distribution function. This is necessary, noting that the built-in random number generators of compilers only output uniformly distributed random numbers from 0 to 1. We achieve the mapping using the following transformation:

$$\int_0^x P_n(\tau)d\tau = R_u \tag{5.39}$$

Here, R_u is a random number generated by a uniform random number generator P_u. Solving for x, upper limit of the integral, we obtain the corresponding

random number for the probability density function P_n. Since we already normalized our probability density functions by making their enclosed area equal to 1, we do not have a normalization factor in front of the integral. Moreover, for the probability density function P_n given in Fig. 5.5, an analytical expression for x can be written as follows:

$$x = \begin{cases} R_u/(2-2R), & R_u < 1-R \\ (R_u - 1 + 2R)/2R, & \text{otherwise} \end{cases} \tag{5.40}$$

Once we determine activity levels of each of the forty million transistors on the chip using the above prescription, we multiply those activity levels with the full heat generated by a device. This heat generated is first determined for the steady-state case, and then weighted for the digital operation, where we assume full power is consumed ten percent of the time. Finally, the calculated heat generated values become the current source strengths in the lumped $R^{\text{th}}C^{\text{th}}$ thermal network.

For our simulations, we take the supply voltages to be 1.5V, and apply the algorithm described in the previous section to obtain the temperature map for the chip, as well as the device temperature-dependent current-voltage characteristics.

In Fig. 5.6(b), we show the calculated temperature profile for the chip. The figure shows that temperature reaches peak at forty three degrees above ambient, while the median and lowest temperatures are thirty three and twenty degrees above the outside temperature. The clock and L2 cache have the highest and lowest temperatures, respectively, because the clock has the highest normalized power and L2 cache has the least. A device in the clock unit operates much more frequently, and thus generates more heat than a device in the L2 cache. This thermal behavior

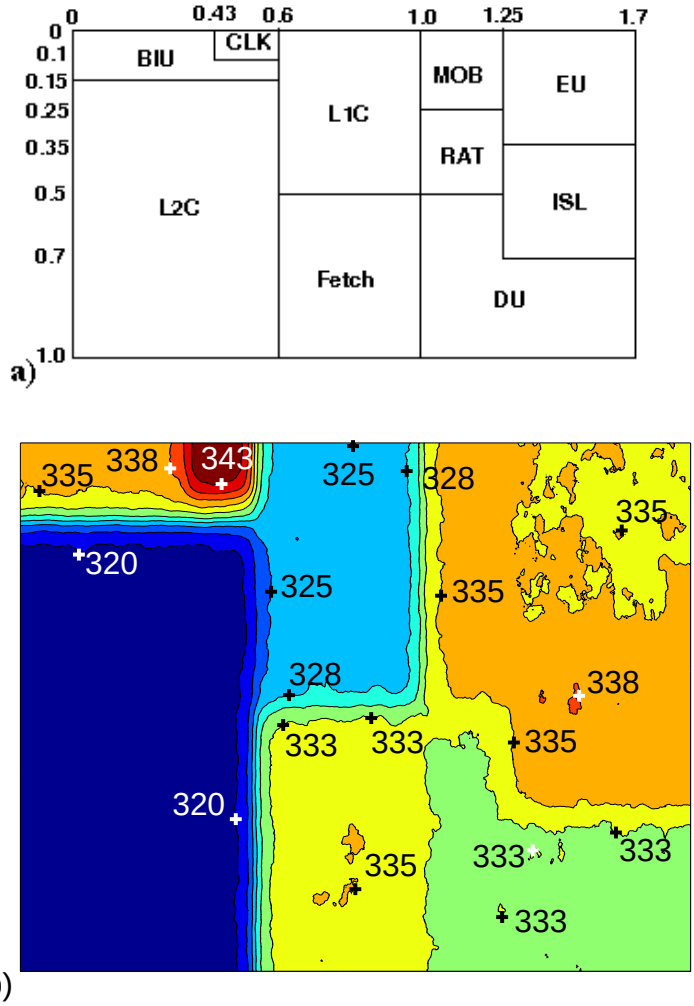

Figure 5.6: a) Functional blocks of the Pentium III chip: Clock has the smallest area but the largest normalized power. Unlike L2 Cache that has the largest area but smallest normalized power as pointed out in Table 5.1. b) Our calculated temperature map for Pentium III reaches a peak in the clock block (forty three degrees above the ambient) and has the lowest temperature plateau in L2 cache (twenty degrees above the ambient). Ambient temperature is 300 degrees Kelvin.

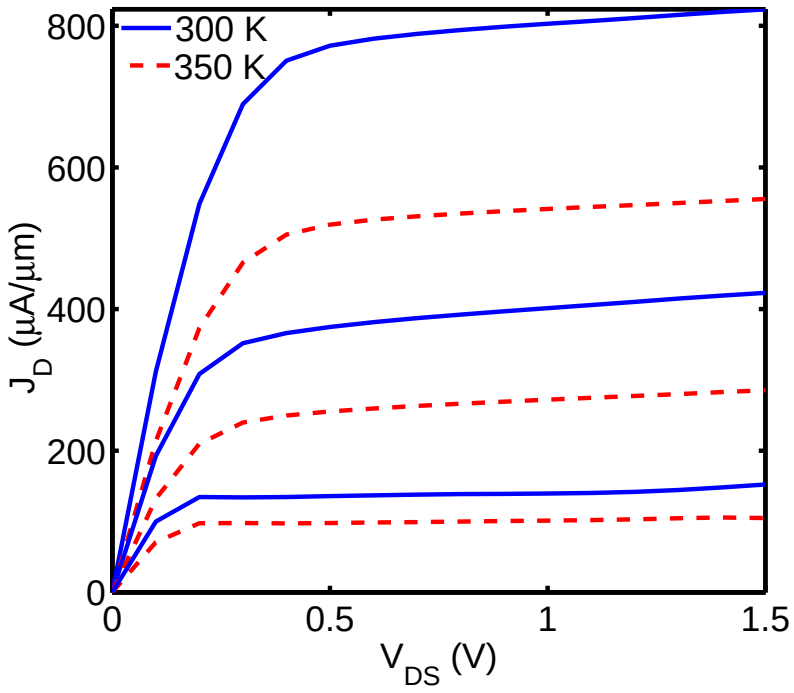

Figure 5.7: Temperature dependent current-voltage characteristics of a 0.13μm n-MOSFET for V_{GS}=0.7V, 1.0V, 1.5V. As temperature increases, current decreases.

is consistent with references [1, 5] for Pentium III processors. Furthermore, the temperature profile can be used to relieve problems related to hot spots on the chip by offering ways for rearranging the spatial distribution of functional units, and utilizing thermal contacts with direct connections to the problematic areas.

In Fig. 5.7, we show temperature dependent device performance characteristics for our n-MOSFET. For an n-MOSFET, as temperature increases, current decreases in the linear and saturation regions. However, current-voltage characteristics differ from that under high temperature conditions (one hundred fifty degrees Celsius and higher) where as temperature increases, current also increases. This can result in a positive feedback and thermal instability.

5.1.5 Section Summary

We present a novel method for obtaining the temperature profiles of ICs with the resolution of a single device. We start from single device simulations and calculate Joule heating for given temperature boundary conditions. We then use a Monte Carlo type methodology to extend our results to the chip surface. We achieve this by assigning different activity levels to the chip's devices, which are then used to calculate Joule heating and temperature distribution of the entire integrated circuit. The method also provides the change in I-V characteristics of the individual transistors as a function of chip heating. Our methodology can be applied to different IC configurations with different running applications. Thus we offer new paradigms to researchers for designing robust designs. Our method can also be easily integrated to a computer-aided-design software and facilitate novel layout designs.

5.2 Stacked Integrated Circuits (ICs): Three-Dimensional (3D)

We present a new method for finding the temperature profile of vertically stacked three-dimensional (3D) digital integrated circuits (ICs), as a three layer 3D IC shown in Fig. 5.8(a). Using our model, we achieve spatial thermal resolution at the desired circuit level, which can be as small as a single MOSFET. To resolve heating of 3D ICs, we solve non-isothermal device equations self-consistently with lumped heat flow equations for the entire 3D IC. Our methodology accounts for operational variations due to technology nodes (hardware: device), chip floor plans (hardware: layout), operating speed (hardware: clock frequency) and running ap-

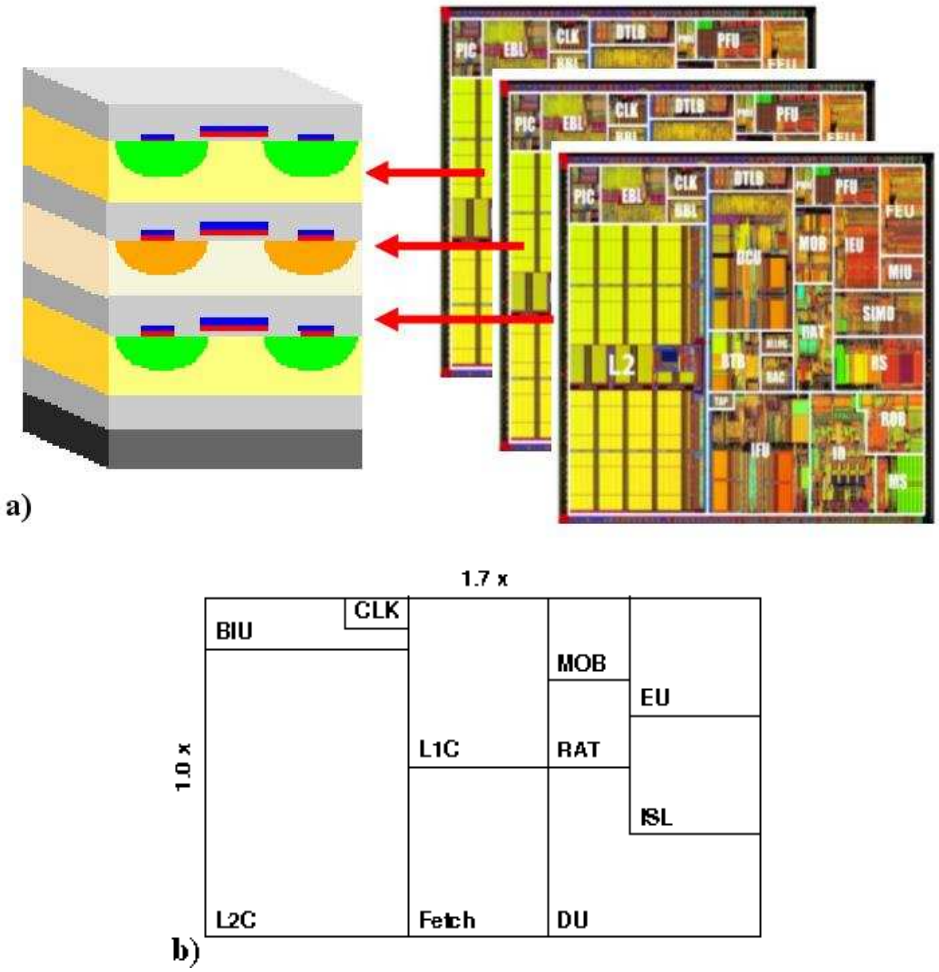

Figure 5.8: a) A vertically stacked three layer 3D IC, where each layer is modeled after a Pentium III [1]. b) Floor plan of each layer in conjunction with Table 5.1.

plications (software). To model hardware, we first decide on an appropriate device configuration. We then calculate elements of the lumped thermal network using the 3D IC layout. To include software, chip floor plan and duty cycle related performance variations, we employ a statistical Monte Carlo (MC) type algorithm. In this work, we investigate performances of vertically stacked 3D ICs, with each layer modeled after a Pentium III [1]. Our calculated results show that layers within the stacked 3D ICs, especially the ones in the middle, may suffer greatly from thermal heating.

As industry makes devices smaller to increase the speed and functionality of integrated circuits, a challenge in IC operation has emerged: interconnect and input/output (I/O) delays. This is especially evident where systems require multiple integrated circuits that communicate through printed circuit boards, I/O pads and bond wires. To alleviate the problem, manufacturers are investigating the development of 3-dimensional integrated circuits (3D ICs). 3D designs can diminish the need for many I/O pads, bond wires, package pins and PCB interconnects. Additionally, 3D designs offer substantial real estate gains. However, while chip heating has become a big problem for standard planar integrated circuits [1]-[11], it is exacerbated for 3D ICs. Silicon dioxide (SiO_2), which acts like a thermal and electrical insulator between stacked chips in a 3D IC, aggravates heating problem by greatly restricting the flow of heat generated. The main result is increased thermal resistance and power density, leading to higher chip temperatures — temperatures higher than conventional cooling methods can account for. Thus, as feature sizes shrink, the power density is increasing exponentially, demanding a focus on heating

and cooling of 3D ICs and planar chips if this barrier is to be overcome [1]-[11].

For the chip heating problem, our simulator should predict localized and overall chip heating for a given 3D IC architecture. It should also assist in developing alternate IC layouts that could help keep localized temperatures low. A foundation has already been established for estimating chip temperatures [6]-[9]. Here, we bring to light the need for a simulator that can connect individual device operations with heating of 3D ICs. Since there can be over a billion devices on a 3D integrated circuit, it is a challenge to calculate the details of device and chip heating simultaneously. Here, we present a method to achieve this connection. First, by self consistently solving coupled quantum and semiconductor equations, we find the electrical characteristics of an n-MOSFET. Next, we take each device on a 3D IC as a cell and model the thermal connections between devices using a lumped circuit type thermal network of thermal resistances, capacitances, and heating sources. From the architectural aspects of the chip layout, we determine the values of the thermal resistances and capacitances in the network. Since the heating source for each device is the driving force in the thermal network, we incorporate the results of the individual MOSFET operations into the millions of thermal elements of the IC. We do this using a Monte Carlo type algorithm, which allows us to realize the goal of connecting 3D IC heating of billions of transistors to individual device operations. Finally, we suggest chip design solutions for cooling the warmest areas of a chip. We present our device and IC levels, and their collective relation in Fig. 5.9.

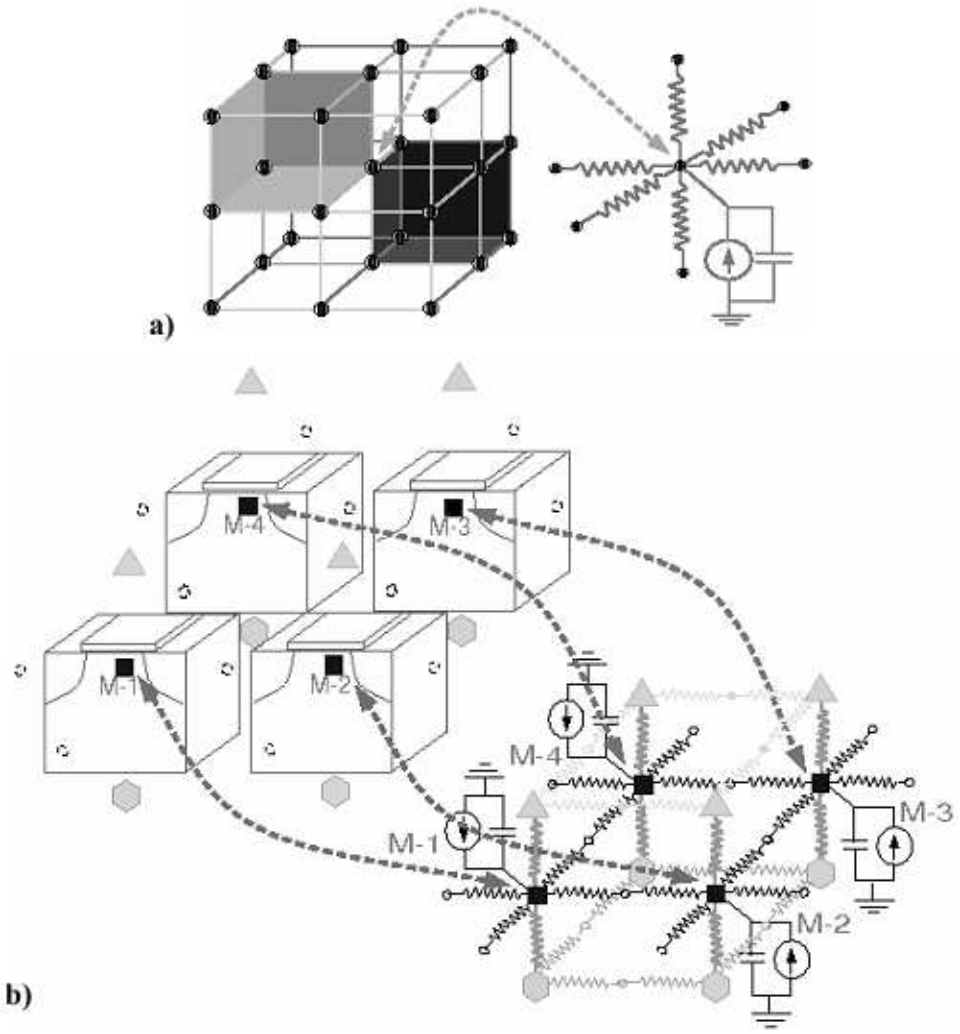

Figure 5.9: a) To analyze 3D IC heating, each MOSFET (M) device is replaced by a current source and an $R^{\text{th}}C^{\text{th}}$ circuit. b) 3D IC's transistors interact thermally with each other as a result of thermal coupling.

5.2.1 Device Performance and 3D IC Modeling

We self-consistently solve device performance and full-3D IC heating equations. We first obtain device performance at different temperatures by solving the semiconductor equations along with the Schrödinger equation. Second, we achieve heating figures of vertically stacked 3D ICs by solving a lumped thermal network in conjunction with device performance results and averaged operational statistics.

We developed a device simulator that is capable of solving the coupled quantum and semiconductor equations. The device simulator provides the electron and hole concentrations, electrostatic potential, current densities, lattice temperature and Joule heating as a function of position and temperature inside the device. Pertinent details of the device simulator were given in the previous section.

Using our device simulator, we first investigate the temperature profile within a single MOSFET. Our analyses indicate that temperature variation within a bulk MOSFET channel is small, unless it is a Silicon-On-Insulator (SOI) device. The lattice temperature inside a bulk MOSFET differs only a few percent from the value at the boundary.

We also developed a lumped thermal network model based on the differential heat flow equation to obtain the temperature profile of vertically stacked 3D ICs. In our model, we account for the 3D IC's layout and floor plan, and the chip transistors' performance details including heat generated, duty cycle and averaged operational statistics.

Large differences in the scales of an entire 3D IC and a single transistor ne-

cessitate use of a lumped thermal network model [7, 78]. We use a similar lumped thermal network derived in the last section. We first obtain thermal capacitances, resistances, and non-isothermal device performance figures, and decide on an appropriate MC methodology. Next, we determine the temperature of each transistor on the 3D IC, represented by (i,j,k), by solving KCL-type equations of the following form:

$$C^{\text{th}}_{i,j,k}\frac{(\overline{T}^{l}_{i,j,k}-\overline{T}^{l-1}_{i,j,k})}{\Delta t}+\frac{(\overline{T}^{l}_{i,j,k}-\overline{T}^{l}_{i\pm1,j,k})}{R^{\text{th}}_{i\pm 1/2,j,k}}+ \frac{(\overline{T}^{l}_{i,j,k}-\overline{T}^{l}_{i,j\pm1,k})}{R^{\text{th}}_{i,j\pm 1/2,k}}+\frac{(\overline{T}^{l}_{i,j,k}-\overline{T}^{l}_{i,j,k\pm1})}{R^{\text{th}}_{i,j,k\pm 1/2}}=I^{l}_{i,j,k}(T^{l-1}_{i,j,k}) \quad (5.41)$$

Here, $1/2$ in the subscript gives the resistance between nodes in the given direction. Furthermore, (i,j) represents a device within a layer k. The superscript l shows the iteration number for our numerical solver.

At the boundaries of the chip, we include the thermal resistances of the package, in addition to the substrate and oxide resistance. This resistor connects to a ground that represents the temperature at the ambient. For these calculations, we take the ambient to be at room temperature. The solutions to these equations give the temperature variation from the ambient.

5.2.2 Mixed-Mode Device Performance and 3D IC Heating: Coupled Algorithm

To obtain the temperature map of 3D ICs, we self-consistently solve lumped thermal network equations for the entire vertically stacked 3D IC in conjunction with

device performance details. These details include non-isothermal device performance figures including current-voltage characteristics, and operational statistics such as duty cycle and functionality. Moreover, we achieve convergence at the device level and the 3D IC level as described below in our coupled algorithm:

Obtain device performance as a function of temperature

For a given vertically stacked 3D IC, we first find the technology node used for fabrication. We determine the average dimensions of a typical transistor on the chip. (We use a MOSFET as our unit cell, but fundamental logic gates such as an inverter can also be used instead.) We then input our representative device in our device simulator. We also decide on typical bias conditions and average on-power during switching for that particular digital IC to adjust total Joule heating for one clock cycle. To obtain device performance including current-voltage characteristics and heat generated at different temperatures, we solve quantum device equations, and prepare a look-up table.

Fit device performance results to a polynomial

We obtain a heat generated, H, versus temperature, T, curve for drain-to-source and gate-to-source biases of 1.5V, which is equal to the on-state bias. Since our KCL-type equations for the lumped thermal network are derived after we apply Kirchoff's transformation to the differential heat flow equation as in Eqn. 5.22, we also produce a heat generated, H, versus transformed temperature, $\overline{T}$, curve. We then fit the H vs. $\overline{T}$

curve to a second-order polynomial and obtain an analytical expression for their relationship.

Set spatial resolution for the 3D IC

We next focus on the geometry of the 3D IC. We first set the spatial resolution in accordance with the average size of the 3D IC's transistors. We then determine the thermal link between devices by defining the thermal resistances, R^{th}, and thermal capacitances, C^{th}, in conjunction with the 3D IC's layout and device architecture. Thus, we obtain values for all the lumped thermal elements except the current sources shown in Fig. 5.9. The strengths of the current sources are related to the heat generated by each transistor on the 3D IC. Therefore, we find their actual values along with the temperature of each device at the end of our mixed-mode simulation.

Obtain effects of 3D IC's floor plan, and software application on performance

To embed effects of 3D IC's floor plan on performance, we group transistors in each layer into a few functional blocks such as cache, floating point unit, execution unit, clock, etc., as shown in Fig. 5.8(b). Next, to embed the effects of the typical software applications on IC performance, we determine consumed percentage power for each functional block in that layer. Then, to obtain the activity level of a transistor within a functional block relative to one within another functional block, we normalize these percentage powers by the corresponding areas of each block.

We then renormalize these percentage powers per area by the maximum for that particular layer.

Statistically extend effects of operational device variations to the entire 3D IC

To extend the effects of operational device variations to the entire 3D IC, we employ a statistical Monte Carlo-type methodology. We first generate a random number for each transistor as a function of the calculated normalized percentage power per area corresponding to that device. We then assign this calculated random number to the corresponding 3D IC's transistor as an indicator of the likelihood of the full power that the particular device is consuming on average. This procedure is applied to each transistor in the 3D chip. In essence, we statistically determine the relative power consumed by each transistor in the 3D IC.

Compilation of data

At this point, we know the following:

- Device performance details including heat generated (H) versus transformed temperature ($\overline{T}$) curve for a single transistor, as well as an analytical expression for a second order polynomial fit,
- 3D IC geometry and layout dependent thermal resistances and capacitances between the 3D IC's transistors, and devices and ambient,
- Statistically determined normalized powers for each transistor that

are obtained using the given 3D IC floor plan and the typical application running on that 3D IC.

Mixed-mode solution

We now can solve the KCL-type lumped thermal network equations given in Eqn. 5.41. From the layout, we know the coefficients of the temperature $\overline{T}$ on the left-hand-side of Eqn. 5.41. We also know the heat generated as a function of temperature. In addition, we know the percentage of the heat generated by each transistor. We have as many equations as the number of transistors on the 3D IC. Each equation is non-linear due to the square law dependency of heat generated on temperature. To solve, we first assign the heat generated at room temperature to all nodes (devices) as an initial guess. We then use a preconditioned bi-conjugate gradient solver to obtain nodal temperatures. We next update the heat generated of each transistor in conjunction with its calculated temperature value. During each iteration, we update temperature and heat generated of each node alternately. Finally, the solution gives the temperature map of the 3D IC as well as the heat generated of each device.

For easy reference, we summarize our algorithm in Fig. 5.10.

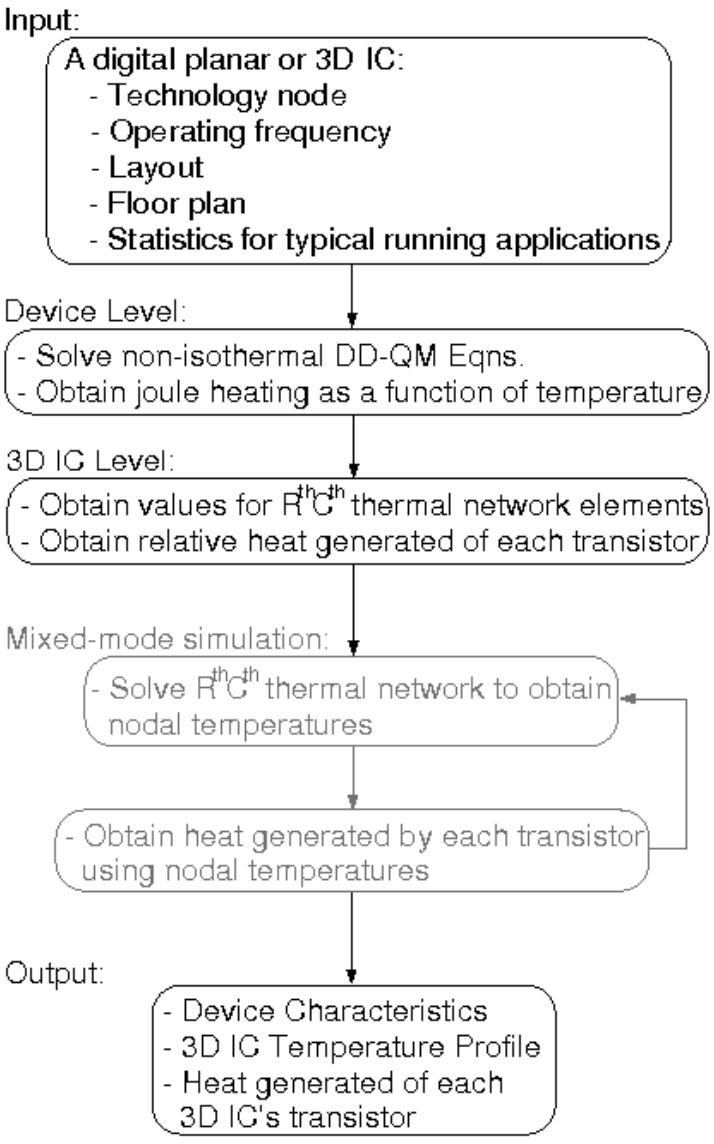

Figure 5.10: Coupled algorithm flowchart.

5.2.3 Mixed-mode Device Performance and 3D IC Heating: Application and Results

After establishing our methodology, we test it on hypothetical digital 3D ICs that have layers modeled after a Pentium III, as shown in Fig. 5.8. We take 0.13μm as the technology node for that chip, and model a device after [74]. We then obtain device performance and heat generated as a function of temperature. We next determine the thermal network associated with this 3D IC, representing a single transistor by a thermal node. We last obtain nodal temperatures (temperature of each transistor) on the 3D IC.

To obtain device performance as a function of temperature, we simulate a 0.13μm N-MOSFET with drain-to-source and gate-to-source biases of 1.5V, at different temperatures, by solving the semiconductor device equations 5.2-5.5 along with the Schrödinger equation 5.1. We then fit the device performance results to a polynomial function. We also weight the calculated steady state powers by the percentage of the on-power during switching.

We next set spatial resolution for our 3D IC by taking a single transistor as our unit cell. Consequently, we have roughly forty million devices in each layer of about 1.6cm^2, with each device occupying approximately an area of 4μm^2. If we have a five layer 3D IC, this yields a very large coupled system of two hundred million nodes. To simplify the problem, we take the 3D IC's transistors to be laid out uniformly in each layer. Each layer is separated by a substrate and an insulating layer, with thicknesses of 250μm and 0.5μm, respectively. Additionally, for the top

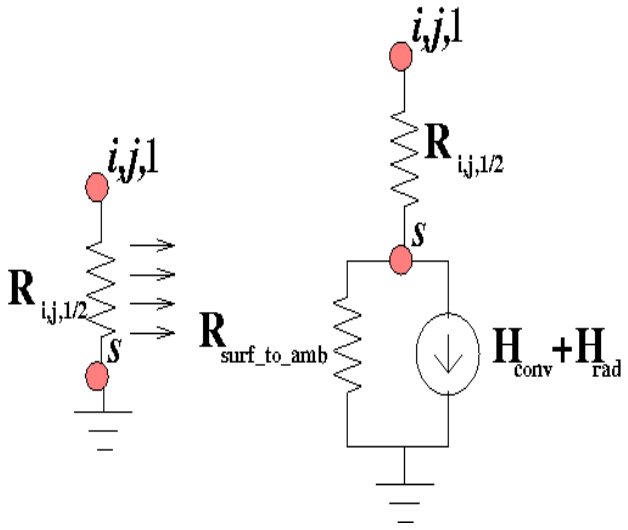

Figure 5.11: To include surface heat transfer due to convection and radiation, we replace the ground resistor connected to the chip's surface, s, shown on the left with the circuit shown on the right. The figure shows the boundary for the bottom layer, k=1, in the vertical direction.

and bottom layers, we use the configuration given in [82]. The bottom and top layers have connections to extra layers such as package and supporting substrate. Using the given 3D IC's layout and package details, we calculate thermal resistance between nodes in the vertical direction between layers, as follows:

$$R_f^{\text{th}} = \frac{250\mu\text{m}}{1.5^{\text{K}}/_{\text{Wcm}} \times 2\mu\text{m} \times 2\mu\text{m}} + \frac{0.5\mu\text{m}}{0.015^{\text{K}}/_{\text{Wcm}} \times 2\mu\text{m} \times 2\mu\text{m}} \simeq 5 \times 10^5 {}^{\text{W}}/_{\text{K}} \quad (5.42)$$

We use values of 1.5K/Wcm, 0.015K/Wcm and 3K/Wcm for the thermal conductivities of silicon, silicon dioxide and metal interconnects, respectively. (Dimension details used to calculate thermal resistances for metals are taken from [83] for 0.1 micron technology.) We have revised the model to include heat losses at the surface due to convection and radiation, and packaging as shown in Fig. 5.11. To account for packaging, the user can adjust the values of the thermal resistors shown in Fig. 5.9, which are at the chip boundary. (More information can be obtained

about thermal packages in [84]-[86].) For this work, we take the thermal resistance between the surface and the ambient to be zero. This has the effect of approximating the surface temperature to be the same as the ambient. To include detailed heat transfer at the surface, formulas given in [87, 88] for the heat transfer coefficients can be used. We then can determine a value for a heat sink (or current sink) at a node on the surface of our network. We next can substitute the sub-network given in Fig. 5.11 in place of our regular ground node.

To calculate thermal resistance between nodes in the same layer, we evaluate the expression written below:

$$R_f^{\text{th}} = \frac{2\mu\text{m}}{1.5^{\text{K}}/_{\text{Wcm}} \times 2\mu\text{m} \times 250\mu\text{m}} \simeq 27^{\text{W}}/_{\text{K}} \tag{5.43}$$

We round this value down to 25W/K. (Thermal resistances of the parallel interconnect lines are much larger than this value.)

We next work on the solution of this thermal network, which consists of forty million nodes (corresponding to all transistors) in each layer and up to five layers. To make the problem tractable, we reduce the associated number of equations while increasing the bandwidth of the connectivity matrix that defines the connections between nodes [81]. To achieve this, we replace sub-blocks in each layer by their Norton equivalent circuits, reducing the size of the system of equations. We enclose a sub-block of N×N nodes in each layer, where N is greater than 2, for size reduction. In Fig. 5.12(a), we show the resulting graph as an example for a 2D planar chip layout with 100 thermal nodes inside. In the figure, we reduce those 100 nodes, as shown on the left network, to the network on the right with 12 nodes, using 4

blocks that are 5×5 each. Each reduced node has an explicit connection to six other reduced nodes, as opposed to four other nodes in the original rectangular grid.

In 3D, we have an analogous situation. We use a three dimensional sub-block consisting on one layer and extending half way above and below the layer (N×N×1). Using Norton's theorem, we reduce these N×N nodes to six nodes, each having ten connections. In Fig. 5.12(b), we show the reduced thermal network for 3D structures. In summary, the resulting graph for 3D has tetrahedral shape unit cells, where each node has explicit connection to ten other nodes as opposed to six other nodes in the rectangular grid. For our calculations, we chose our 3D sub-block as 22×22×1. This reduces the number of simultaneous equations that we solve from approximately 200 million to a more tractable 3 million for the CPU. While there is nothing unique about our choice of 22 nodes, it allows us to obtain a solution for the thermal network in five-to-ten minutes on a standard Pentium4 PC.

The boundary of each sub-block is a half resistor away from the nodes that are closest to the enclosing surface in all six directions. We then short the sub-block borders on each side, yielding six new nodes (four in the same layer, and two in the top and bottom of that layer). We next obtain the six-port Norton equivalent circuit seen from these nodes, with equivalent thermal resistances, a capacitance and a heat source attached to each. To obtain the Norton equivalent circuit, we write the impedance matrix for the N^2 (N=22) nodes inside the cube and the six nodes on the sides using the KCL analysis. We then divide our calculated impedance matrix up into four sub-matrices, and multiply two of these sub-matrices with the unknown currents of the six outer nodes and the known currents of the N^2 inside

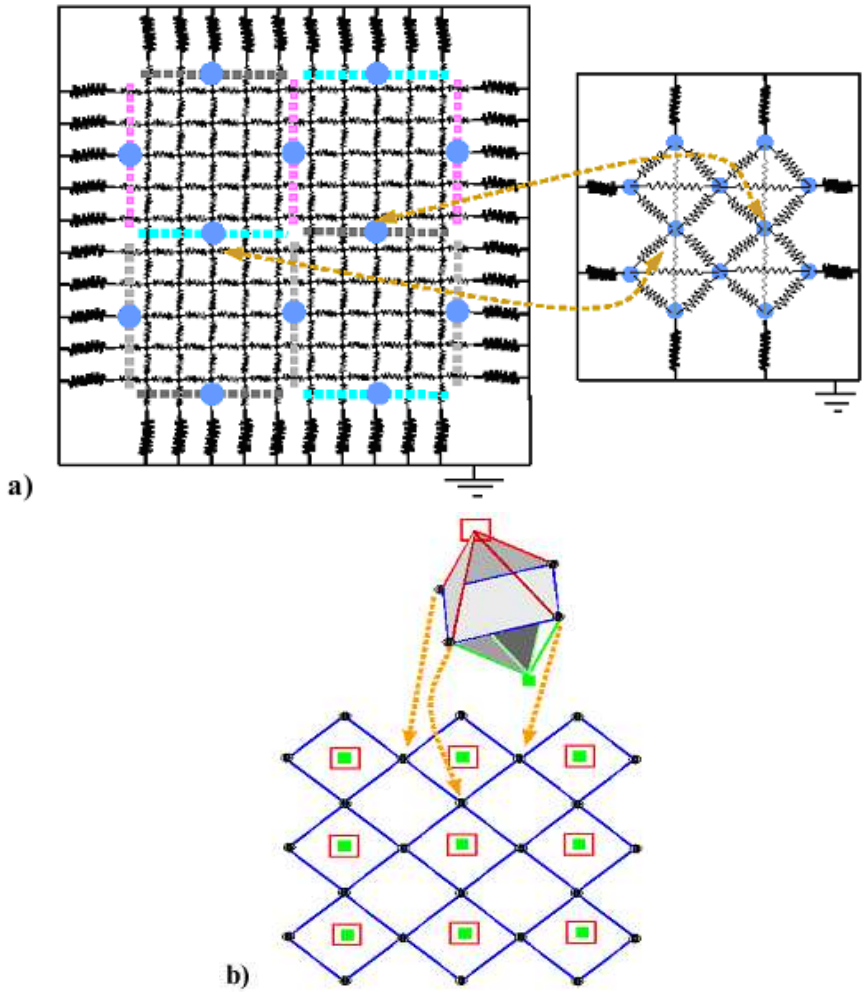

Figure 5.12: a) We apply size reduction methods to a planar chip with one hundred mesh points. We divide it up into four blocks. We then replace the original mesh with twelve nodes corresponding to four-port Norton representations of each block. (Bold resistors are for package.) b) In 3D, we have six-port tetrahedral shape Norton representations for cubes of grid points like the one shown in Fig. 5.9(a). Coupling to layers above and below is through nodes at the top and bottom of each tetrahedral shape, respectively.

nodes. This gives the unknown voltages of the six outer nodes. Multiplication of the sub-impedance matrix with the known current sources gives the Thevenin equivalent voltage sources. The sub-impedance matrix, which is the current coefficient matrix of the outer nodes, is the Thevenin equivalent impedance matrix. Next, we transform the Thevenin equivalent circuit to the Norton equivalent circuit.

We then extend our calculated heat generated results to the 3D IC volume using a Monte Carlo (MC) type methodology. We use an MC algorithm to statistically determine each equivalent node's source strength. Our MC algorithm makes use of the floor plan shown in Fig. 5.8(b) with percentage powers and areas given in Table 5.1. After we set up our thermal network including the source components, we solve the reduced system of equations for nodal temperatures using a bilateral conjugate gradient method.

In Fig. 5.13, we show steady state device performance figures including current-voltage characteristics and heat generated as a function of temperature. Figure 5.13(a) indicates that as temperature increases, current decreases both in the linear and saturation regions. This is in accordance with the downward slope of the heat generated versus temperature curve, as shown in Fig. 5.13(b). (We note that temperatures calculated have not gone beyond device operating limits where intrinsic carrier concentration approaches that of the doping.)

In Fig. 5.14(a), we show a five layered vertically stacked 3D IC with a Pentium III type chip in each layer. Our calculated temperature maps for the middle, second and bottom layers of that 3D IC are shown in Figs. 5.14(b)-(d), respectively. We note the dramatic increase for the peak temperature value from the bottom, 365°K,

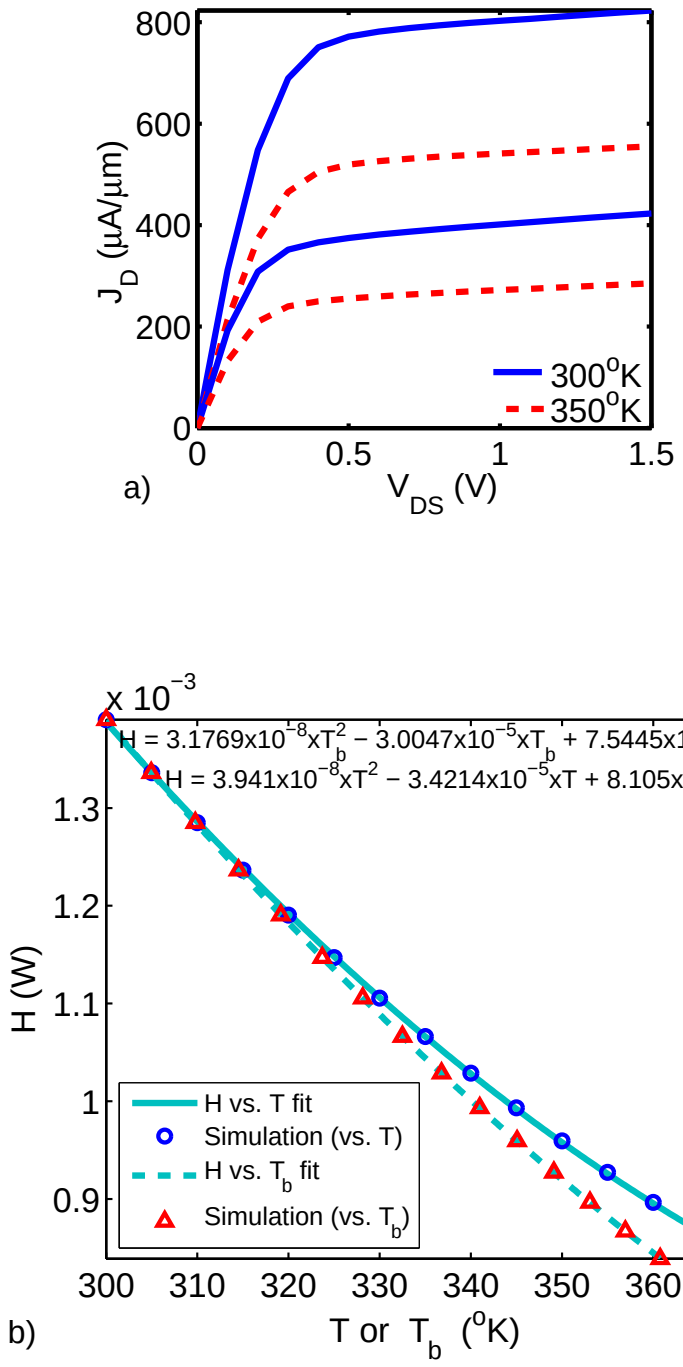

Figure 5.13: a) Temperature dependent current-voltage characteristics of a 0.13μm N-MOSFET for V_{GS}=1.0V, 1.5V. a) Steady-state heat generated (V_{GS} = V_{DS} = 1.5V) as a function of temperature (T) and $\overline{T}$ (T_b). Conversion from T to $\overline{T}$ is given in Eqn. 5.22.

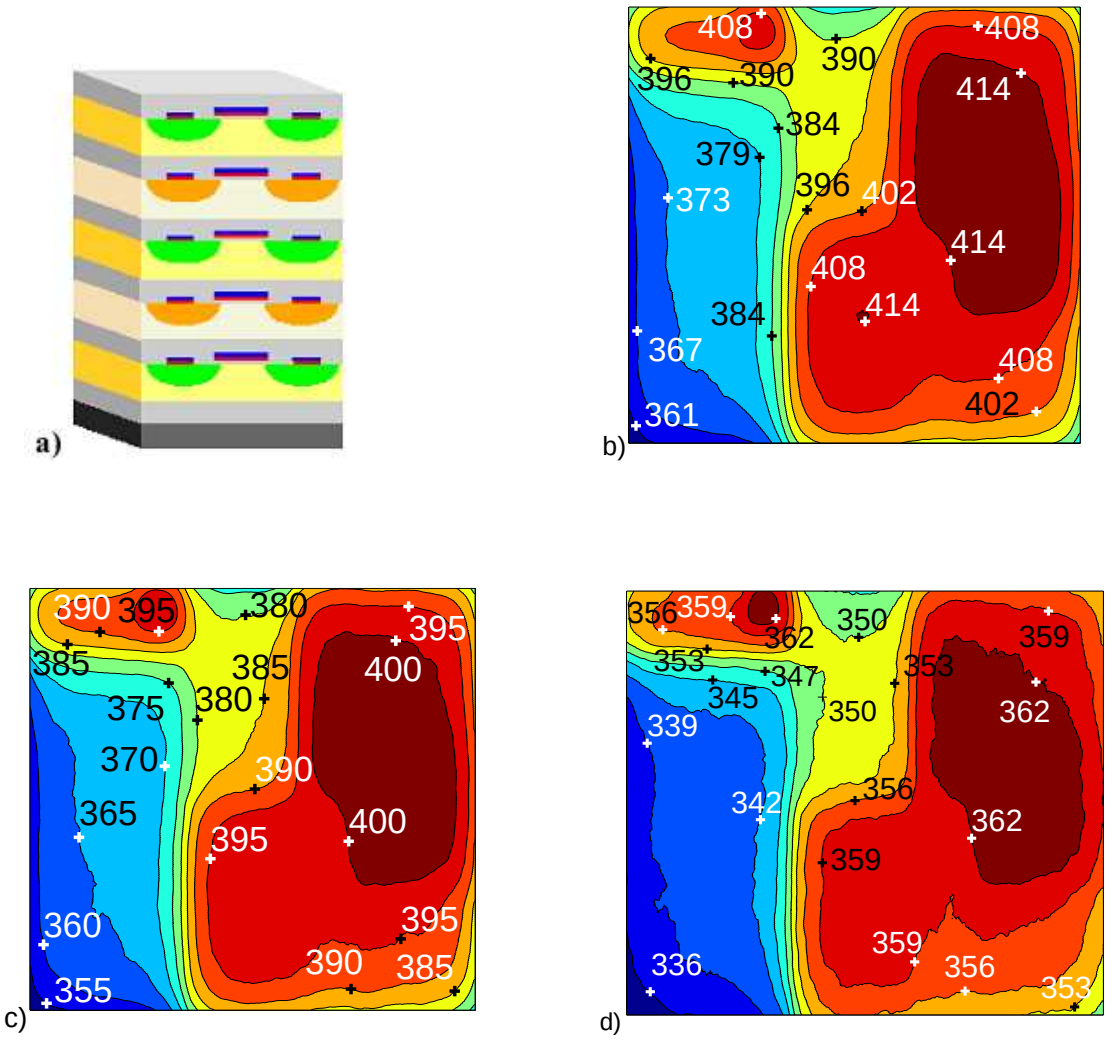

Figure 5.14: a) A 3D IC with five layers of stacked Pentium III chips. Our calculated temperature maps corresponding to the b) middle, c) second and d) bottom layers shown in a). Here, ambient is at room temperature (300°K).

to the middle layer, 420°K. We attribute this to the low thermal diffusion constant of the SiO_2, which traps heat in sandwiched layers. In addition, we also note that the location of the peak temperature moves from the clock block in the bottom layer (and one-to-three layered 3D ICs) to the issue unit in the middle layer, as shown in Fig. 5.14 (in relation to the layout shown in Fig. 5.8(b)). We associate this with the increase of equivalent thermal resistance with stacking for each node. This also causes an increase in maximum 3D IC temperature, as well as the peak temperature of the bottom layer, as we utilize more layers in a stacked 3D IC configuration, as can be seen in Fig. 5.15(a). This shows the marked heating problem in 3D ICs. Moreover, high temperature variations on a 3D IC are likely to have detrimental effects on device and circuit operations. For example, temperature related phase delays may result in the failure of synchronous circuit operation. In Fig. 5.15(b), we show the oscillation frequency of a thirty one stage ring oscillator as a function of temperature. This shows that if such a circuit is used as a clock generator for each layer, the speed of each layer will deviate from the others even though they all have the same room temperature operating frequency when the 3D IC is first turned on.

The temperature map of a 3D IC can also be used in conjunction with computer aided design (CAD) tools to relieve problems related to hot spots and high temperature gradients on the chip. To achieve this, chip floor plans can be rearranged to distribute active units over the whole volume. Additionally, thermal contacts can be utilized to pull high temperatures to low at problematic regions. We test the effects of perfect vertical thermal contacts (shorts to ambient) on a layer that has the temperature profile given in Fig. 5.14. Utilization of one thermal

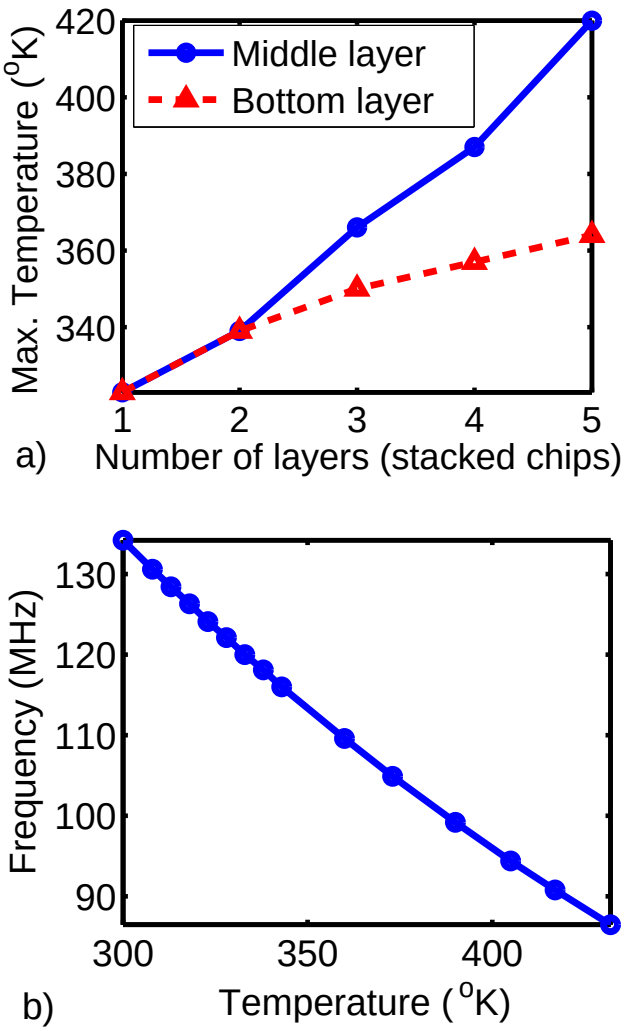

Figure 5.15: a) Maximum temperature of the middle (also the maximum of the entire 3D IC) and bottom layers as a function of number of layers. b) Oscillation frequency of a thirty one stage ring oscillator calculated by Cadence [89] decreases as temperature increases. Here, ambient is at room temperature (300°K).

contact near the peak temperature location of the middle layer pulls the maximum temperature couple of degrees down; however, an array of ten by ten thermal contacts pulls the peak temperature down about fifty degrees.

To help verify our new algorithm for mixed mode device-chip temperature modeling, we applied our method to a chip whose temperature profile was recently provided in the literature [90, 91]. To compare our approach with the published chip temperature results, we use the layout given in [90, 91] to reproduce the temperature map given for that chip. Using the layout in [90, 91], we grouped some of the functional blocks, given in that paper, together and assigned a single power density to each new block. In Fig. 5.16(a), we show our extracted layout, and power consumed in each block over enclosed number of original functional blocks for the chip. We use the layout, geometry and power profile given in Fig. 5.16(a) in conjunction with vertical (including package) and lateral resistances whose values are proportional to those in Eqns. 5.42 and 5.43: 1.5×10^4W/K and 0.75W/K, respectively. This corresponds to approximately 55.5 million grid points. Using our simulator, we obtain the temperature map shown in Fig. 5.16(b), which is quite similar to the thermal map given in [90, 91]. In addition, our simulation takes only about one minute of computing time.

5.2.4 Effects of Different Layer Thicknesses on 3D IC Heating

So far, we have used the bulk MOSFET heating figures as inputs to the current sources of the $R^{\text{th}}C^{\text{th}}$ thermal network. To obtain effects of silicon substrate

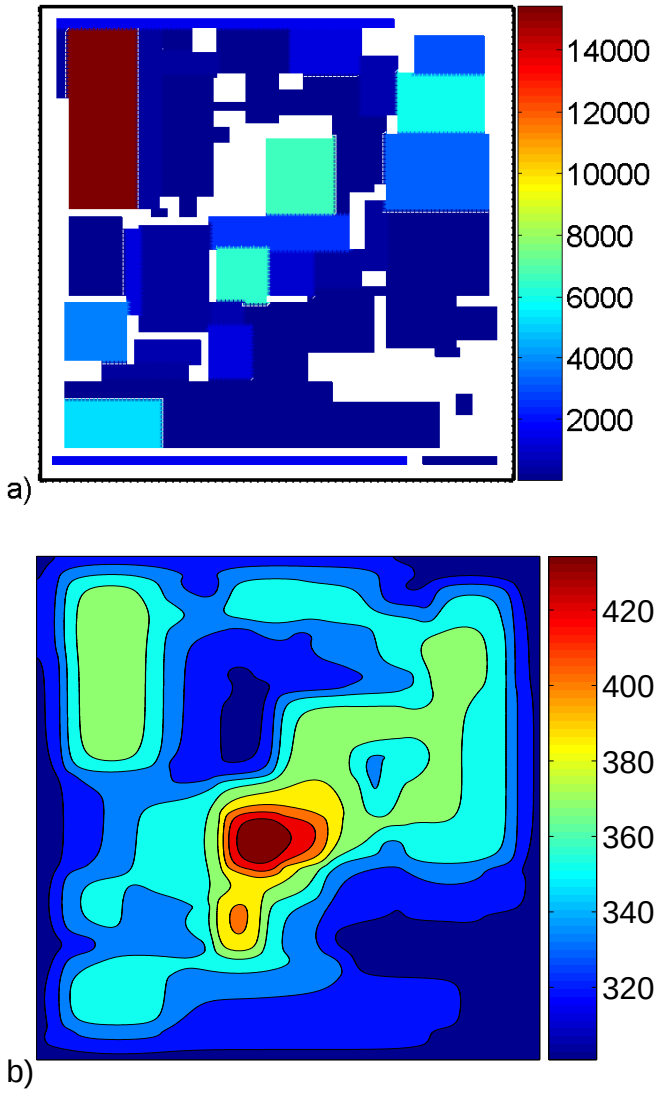

Figure 5.16: a) Maximum temperature of the middle (also the maximum of the entire 3D IC) and bottom layers as a function of number of layers. b) Oscillation frequency of a thirty one stage ring oscillator calculated by Cadence [89] decreases as temperature increases. Here, ambient is at room temperature (300°K).

thinning on the channel and device temperatures, we solve the semiconductor equations along with the quantum corrections for a bulk MOSFET and two different SOI-MOSFETs, which have channel thicknesses ($t_{\rm ch}$) of 1nm or 20nm. To include quantum corrections, we make use of the density gradient effective potential term, which was discussed in the previous chapter. We also include the thermal effects in the drift-diffusion model, as described in Section 2.4.

In Fig. 5.17, we show our calculated current and heat generation plots for the bulk and the SOI-MOSFETs, with channel thicknesses of 1nm or 20nm. As before, we use a 0.13μm channel length device. Our numerical results show that the bulk and the 20nm channel thickness SOI-MOSFET have similar heat generation, whereas the 1nm SOI-MOSFET has about twenty percent less heat generation. (We note that the heat generated curve of the bulk MOSFET is lower than the one shown in Fig. 5.13(b). We associate this with different current levels obtained by solving the Schrödinger equation or by using the density gradient formalism. To ascertain similar performance figures, both simulators need to be calibrated.) We attribute this to confinement effects, and excessive channel temperatures in small channel thickness SOI-MOSFETs. To show how peak channel temperatures differ between these three device configurations, in Fig. 5.18(a), we show the channel maximum temperature as a function of boundary temperature, which is attributed to the device terminals. In Fig. 5.18(b), we present how much the channel heats up in excess of the boundary. It shows that the smaller the SOI channel, the higher the channel temperatures are, with the highest difference between the channel and the device boundary being 100°K for the 1nm channel SOI-MOSFET among the

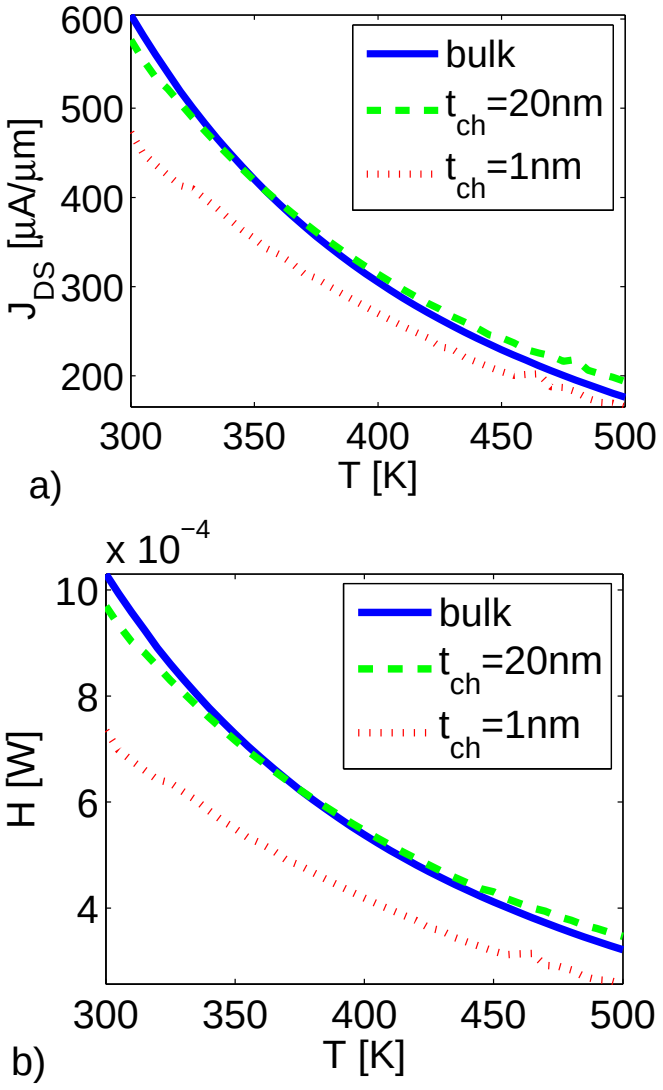

Figure 5.17: Calculated a) current and b) heating figures of bulk and SOI-MOSFETs (0.13μm).

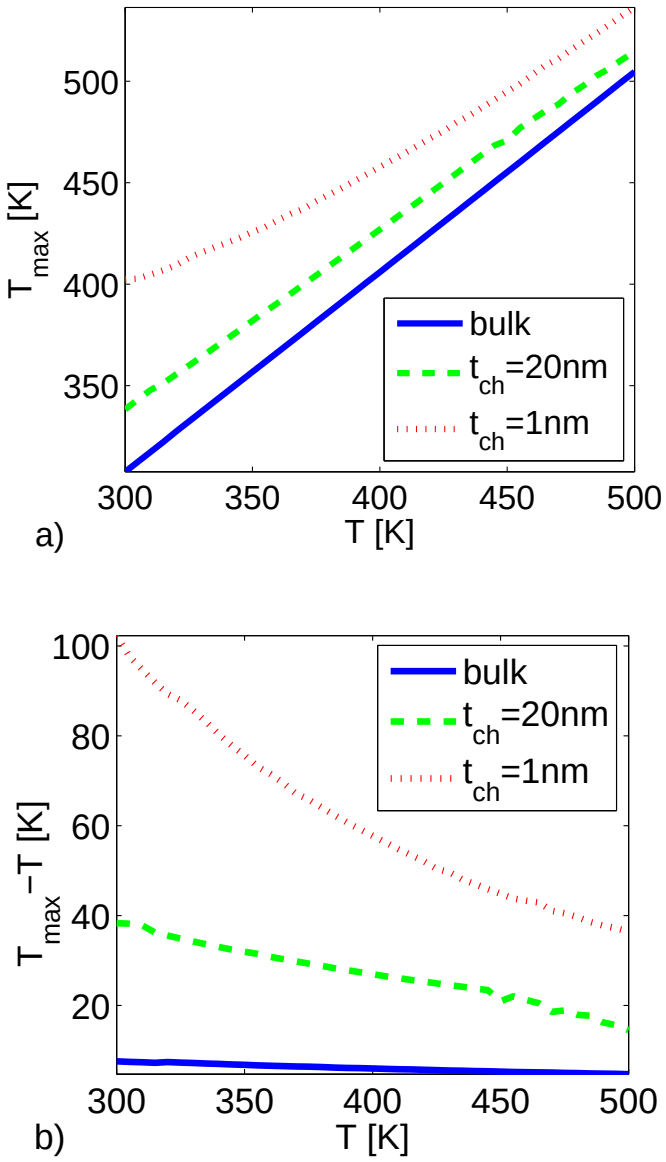

Figure 5.18: Calculated a) current and b) heating figures of bulk and SOI-MOSFETs (0.13μm).

simulated devices.

To obtain IC temperature maps using the heat generated figures of these three devices, we fit the heat generated curves to polynomials. For the bulk MOSFET, and 1nm and 20nm channel thickness SOI-MOSFETs, we have the following relations between temperatures and heat generation (H is in Watts and T is in Kelvin):

$$H_{\mathrm{b}} = -5.0037\times10^{-11} \times T^3 + 7.3639\times10^{-8} \times T^2 - 3.7933\times10^{-5} \times T + 7.1303\times10^{-3} \tag{5.44}$$

$$= -2.8947\times10^{-11}\times\overline{T}^3 + 4.8727\times10^{-8}\times\overline{T}^2 - 2.9061\times10^{-5}\times\overline{T} + 6.1439\times10^{-3} \tag{5.45}$$

$$H_{20\mathrm{nm}} = -2.649\times10^{-11} \times T^3 + 4.2773\times10^{-8} \times T^2 - 2.4325\times10^{-5} \times T + 5.127\times10^{-3} \tag{5.46}$$

$$= 1.2769\times10^{-11}\times\overline{T}^3 - 2.1894\times10^{-9}\times\overline{T}^2 - 8.0769\times10^{-6}\times\overline{T} + 3.2402\times10^{-3} \tag{5.47}$$

$$H_{1\mathrm{nm}} = -6.8594\times10^{-12} \times T^3 + 1.5415\times10^{-8} \times T^2 - 1.1297\times10^{-5} \times T + 2.9098\times10^{-3} \tag{5.48}$$

$$= 3.4541\times10^{-11}\times\overline{T}^3 - 3.204\times10^{-8}\times\overline{T}^2 + 6.0457\times10^{-6}\times\overline{T} + 8.5987\times10^{-4} \tag{5.49}$$

Solving the $R^{\mathrm{th}}C^{\mathrm{th}}$ thermal network in the previous section for a five-layered IC using the above heat generated temperature relations, we obtain the comparative temperature values presented in Table 5.2.

Table 5.2 indicates that for future generation 3D ICs with smaller layer thick-

Table 5.2: Comparison of peak boundary and channel temperatures

Peak Channel Temperature		
	Bottom Layer	Middle Layer
Bulk MOSFET	360°K	405°K
SOI-MOSFET (t_{ch}=20nm)	385°K	427°K
SOI-MOSFET (t_{ch}=1nm)	422°K	445°K
Peak Boundary Temperature		
	Bottom Layer	Middle Layer
Bulk MOSFET	353°K	400°K
SOI-MOSFET (t_{ch}=20nm)	353°K	400°K
SOI-MOSFET (t_{ch}=1nm)	343°K	343°K

nesses, the overall heating at the device terminals will be lower at the expense of much higher channel temperatures.

5.2.5 Section Summary

We present a new method for determining the temperature profile of complex digital 3D ICs. Using the new methodology, we achieve spatial resolution of a single device. We first obtain device performance figures such as heat generated as a function of temperature. We then calculate values for thermal lumped elements using the 3D IC geometry. After extending our device results to each transistor on the 3D IC using an MC type algorithm, we iteratively solve for nodal temperatures to obtain the thermal map of the 3D IC in conjunction with each transistor's performance. Details of our algorithm can easily be modified for other planar (2D) or 3D

ICs with different designs and operating conditions. Knowing potential hot spots can facilitate new design strategies for 3D ICs that are less susceptible to thermal damage. It can also suggest new floor plans and ways to monitor effects of thermal contacts.

5.3 Methods for Cooling ICs

Our calculated 3D IC heating figures indicate that planar (2D) and 3D ICs suffer extensively from heating. This heating problem is more pronounced in 3D ICs, where it imposes upper limits on the number of layers and device densities for safe device and IC operations. Since success of future electronics relies heavily on device scaling, to relieve the heating problem, instead of bringing down the power density dramatically using bigger devices or devices set apart, we seek other solutions. One such solution is use of alternative cooling methods for the IC. We propose use of thermal vias [92] or metal lines to remove the heat from the hot areas. This method enables use of current assembly lines; thus, it obviates the need for constructing a prohibitively expensive new fabrication facility, which might be the case if unconventional electrically insulating high thermal conductivity materials are employed [93].

Our previous investigations showed that the 3D IC employing 1nm channel thickness SOI-MOSFETs reaches the highest temperatures (in terms of channel temperatures, but not device terminal, boundary, temperatures). To relieve extreme channel temperatures, we propose the use of an array of thermal vias. To

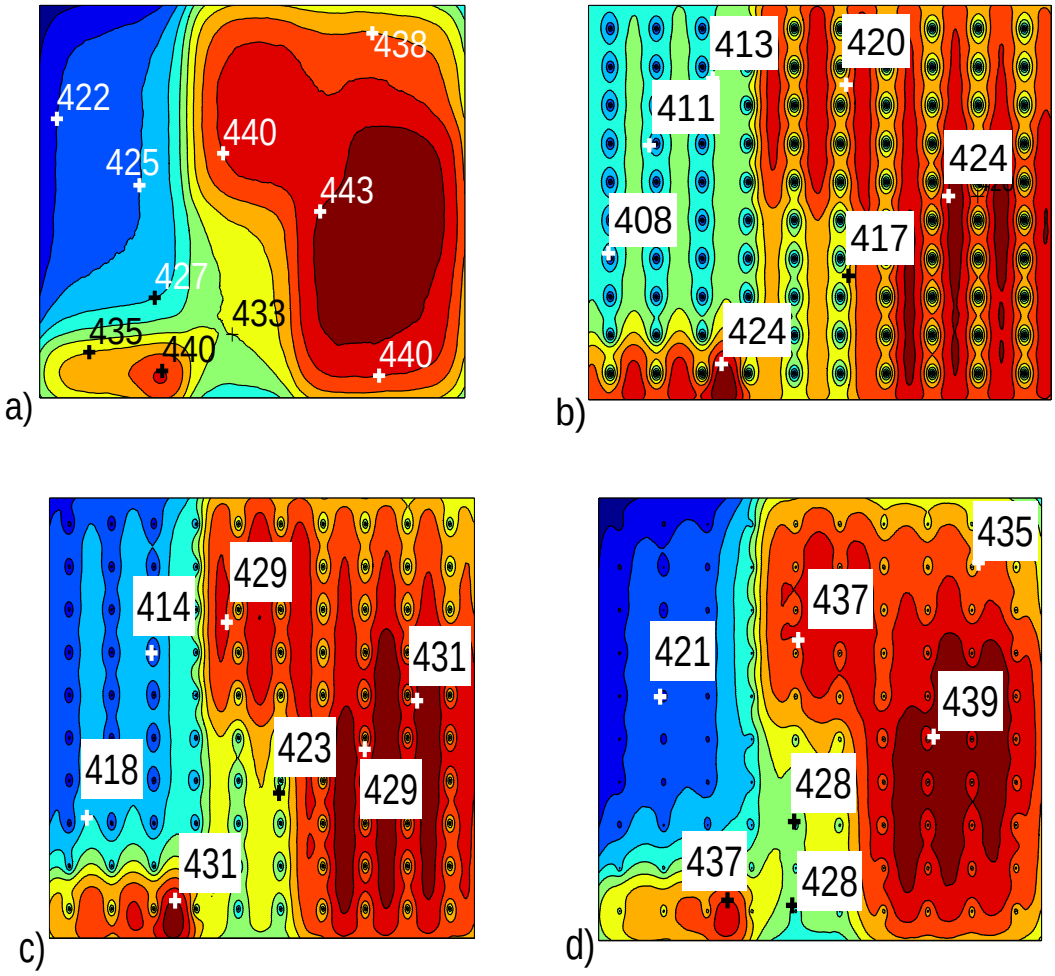

Figure 5.19: Thermal maps for a 3D IC employing 1nm channel thickness SOI-MOSFETs and an array of 10 x 10 vertical vias. Thermal maps of peak channel temperatures are shown for the middle layer of a five layered 3D IC that employs thermal vertical vias, between the layers (R_l), and the top or bottom layer and the ambient (R_b). a) No vertical vias, where $T_{\max}$=445°K and T_{ave}=436°K. b) R_l=0.01K/W and R_b=0.04K/W, where $T_{\max}$=426°K and T_{ave}=417°K. c) R_l=10K/W and R_b=0.04K/W, where $T_{\max}$=432°K and T_{ave}=426°K. d) R_l=10K/W and R_b=40K/W, where $T_{\max}$=441°K and T_{ave}=433°K.

investigate the effects of vertical vias that extend from the top to the bottom of the 3D IC, we use a 10×10 array of vertical vias uniformly distributed over the chip surface. In Fig. 5.19, we show our calculated thermal maps, using the peak channel temperatures, for the middle layer of a five layer 3D IC. Figure 5.19(a) is for the one without any thermal vias, while Fig. 5.19(b) is for the one that employs thermal vertical vias with the smallest thermal resistances of 0.01K/W between the layers, and 0.004K/W between the top or bottom layer and the ambient. These thermal resistances are not close to ideal thermal contacts, which are shorts to ambient that pull the temperature around them down to the ambient value of 300°K. They pull the minimum temperature down to 401°K from the 417°K minimum of the one without any vias. Also, the average and maximum temperatures drop from 436°K to 417°K, and 445°K to 426°K. When the thermal resistances of the thermal vias increase between the layers from 0.01K/W to 10K/W, maximum, average and minimum temperatures increase respectively from 426°K, 417°K and 401°K to 432°K, 426°K and 410°K. If we also increase the thermal resistance of the boundary thermal via from 0.04K/W to 40K/W, maximum, average and minimum temperatures further rise to 441°K, 433°K and 416°K, closing in to those values of the one without any vias. We attribute the temperature variations to changes in the equivalent Norton resistances seen from each node, where the percentage change rapidly gets smaller as thermal resistances of the vias rise. In addition, thermal vias can sink limited amount of heat due to their finite resistances. Therefore, to lower the overall temperature, we need to utilize thermal vias with very low thermal resistances. To remove heat from the hottest region of the chip, we can strategically place thermal

vias around the hot region. This can facilitate effective heat removal from that region.

In addition to thermal vias, we also investigate the effects of horizontal heat sinks that extend from one side of the chip to the other. Such heat sinks can be made from metal lines that are on top of the active device regions. In Fig. 5.20, we show thermal maps for the middle layer of a five layer 3D IC, corresponding to the same 3D IC as in Fig. 5.19. We employ ten lines of heat sinks that are uniformly distributed over the side. When we use thermal heat sink resistances of 0.01K/W within the layer, and 0.04K/W at the boundaries, we pull the maximum, average and minimum temperatures from 445°K, 436°K, 417°K down to 424°K, 414°K, 401°K, as shown in Fig. 5.20(a). However, a rise in the thermal heat sink resistance within the layer from 0.01K/W to 10K/W diminishes the cooling effect, and pulls the maximum, average and minimum temperatures up to 445°K, 434°K and 401°K, approaching the values for the one without any vias or lateral heat sinks.

To make the temperature variations small within a layer, we can also change the chip's layout. To examine how such a change would affect the thermal profile, we simulated a five layer 3D IC. For the bottom layer, we use the layout given in 5.8(b), as has been used for all layers to obtain results shown in Figs. 5.19 and 5.20. For each consecutive layer above, we rotate this layout ninety degrees clockwise. The resulting thermal map of the middle layer before and after the rotations are shown in Figs. 5.21(a) and 5.21(b). This reduces maximum, average, minimum temperatures from 445°K, 436°K, 417°K to 438°K, 435°K, 425°K.

As a summary, to relieve the heating problem, we offer solutions such as chang-

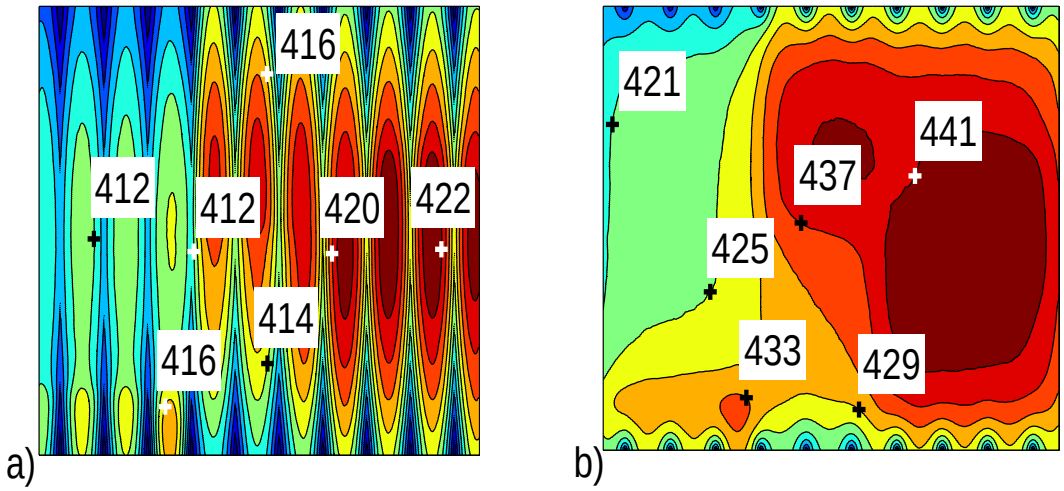

Figure 5.20: Thermal maps for a 3D IC employing 1nm channel thickness SOI-MOSFETs and an array of 10 lateral vias. Thermal maps of peak channel temperatures are shown for the middle layer of a five layered 3D IC that employs lateral heat sinks, with resistances of R_l within the layer, and R_b at the boundaries. a) R_l=0.01K/W and R_b=0.04K/W, where $T_{\max}$=424°K and T_{ave}=414°K. b) R_l=10K/W and R_b=0.04K/W, where $T_{\max}$=445°K and T_{ave}=434°K.

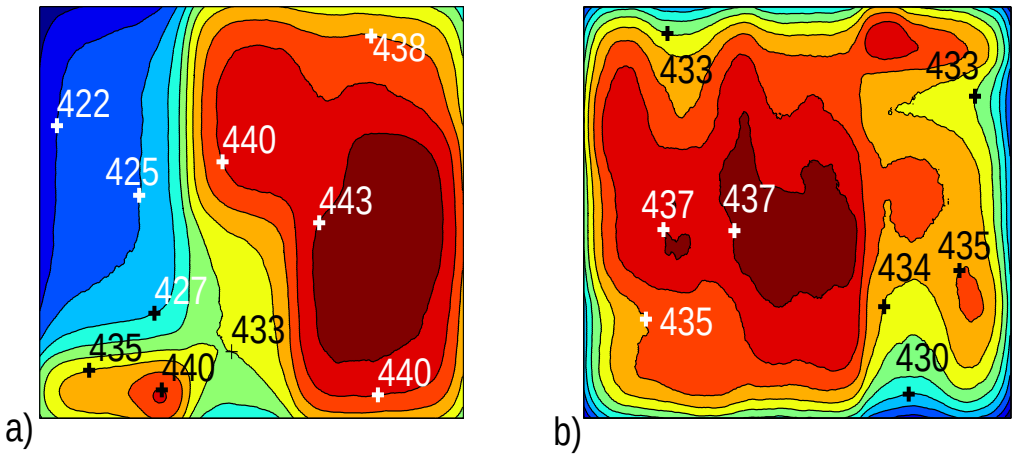

Figure 5.21: Thermal maps for a 3D IC employing 1nm channel thickness SOI-MOSFETs. Thermal maps of peak channel temperatures are shown for the middle layer of a layered 3D IC using a) the same layout for each layer ($T_{\max}$=445°K and T_{ave}=436°K), or b) the ninety degrees rotated version for each consecutive layer ($T_{\max}$=438°K and T_{ave}=435°K).

ing the chip's layout, and the use of thermal contacts in the vertical chip direction with vias or in the lateral chip direction with metal lines. Our investigations show that for effective heat removal, we need low resistance thermal contacts. Such low thermal resistances can pull the overall chip temperatures considerably down. However, any unintentional fabrication of high resistance thermal contacts may significantly diminish cooling effects, and result in higher temperatures than predicted. Instead of uniformly lowering all temperatures, thermal contacts can be utilized more intensely around hot regions. In that case, they are likely to relieve the heating problem locally. In addition to thermal contacts, we also offer methodologies for novel chip layout designs. Even a simple ninety degrees rotation between the layers may move the mid-layer's temperatures to safer operational limits.

5.4 Experimental Investigations

To experimentally investigate the chip heating effects, we fabricated chips with thermal sensors and heaters. As temperature sensors, we use *pn* junction diodes, like the one shown in Fig. 5.22, because of their currents' high sensitivity to varying temperatures. To specify the percentage change in diode current with temperature, we write diode current in terms of its geometrical and electrical parameters, as follows:

$$I = qA\left(\frac{D_p}{N_d L_p} + \frac{D_n}{N_a L_n}\right) n_o^2 \left(e^{\frac{V_A}{V_{TH}}} - 1\right) \tag{5.50}$$

Above, D_n (D_p) is the electron (hole) diffusion constant, and L_n (L_p) is the corresponding electron (hole) diffusion length, as described in Chapter 2. (For short

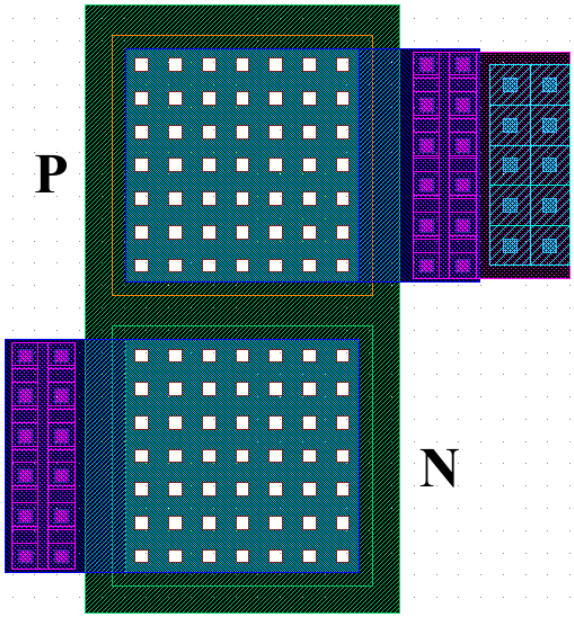

Figure 5.22: We use a *pn* junction diode as a temperature sensor. This 10 x 10μm^2 diode was laid out using the Cadence Virtuoso tool [89].

channel *pn* junction diodes ($W < L_p, L_n$), we replace L_p and L_n with the diode's physical length W.) Moreover, N_d and N_a are the donor and acceptor levels in the n and p regions of the *pn* junction diode, respectively. Furthermore, the change in diode current with temperature is mostly determined by the temperature dependency of the intrinsic carrier concentration n_o and the exponential term within the parenthesis that has the ratio of the applied bias $V_{\rm A}$ and the thermal voltage $V_{\rm TH}$ as the exponent. The thermal voltage and the square of the intrinsic carrier concentration can be written as functions of temperature as shown below:

$$V_{\rm TH} = \frac{kT}{q} \tag{5.51}$$

$$= \alpha T \tag{5.52}$$

$$n_o^2 = 4\left(m_n^* m_p^*\right)^{3/2}\left(\frac{2\pi kT}{h^2}\right)^3 e^{\frac{-E_g}{kT}} \tag{5.53}$$

$$= \beta T^3 e^{\frac{-qE_g}{\alpha T}} \tag{5.54}$$

Here, α, β and bandgap E_g are not functions of temperature. Therefore, the ratio of the diode currents at two different temperatures can be written as follows:

$$\frac{I(T_1)}{I(T_2)} = \left(\frac{T_1}{T_2}\right)^3 e^{\left(\frac{V_A - qE_g}{\alpha}\right)\left(\frac{1}{T_1} - \frac{1}{T_2}\right)} \tag{5.55}$$

Since diode currents change exponentially with temperature, it makes them an attractive candidate for use in thermal sensors. That is why we picked a diode as our fundamental thermal sensor block. Once we decided on our thermal sensor configuration, which is a $10\times10\mu\text{m}^2$ diode, we laid out a chip that has an array of 10×10 thermal diodes uniformly distributed over the chip's surface, as shown in Fig. 5.23.

Here, our goal was to fill up the remaining space with circuits that function as microheaters, and then to measure the chip temperatures using the diode currents. As our fundamental circuit and microheater, we employed an NMOS block that was comprised of hundreds of smallest size NMOS devices with their gates, sources and drains shorted together to enable maximum heat generation, as shown in Fig. 5.24. We next shorted their sources and drains to the chip ground and V_{DD}, respectively. To be able to control the amount of heat generated, we connected the gate to an output pin so that the total current can be set to a desired level. We used a 4×4 array of NMOS blocks. The chip layout with the 10×10 diode sensor array and the 4×4 NMOS heater blocks is shown in Fig. 5.25. Even though we successfully observed diode current change with temperature, due to a biasing problem, we were not be able to simultaneously heat up the chip and measure the temperature.

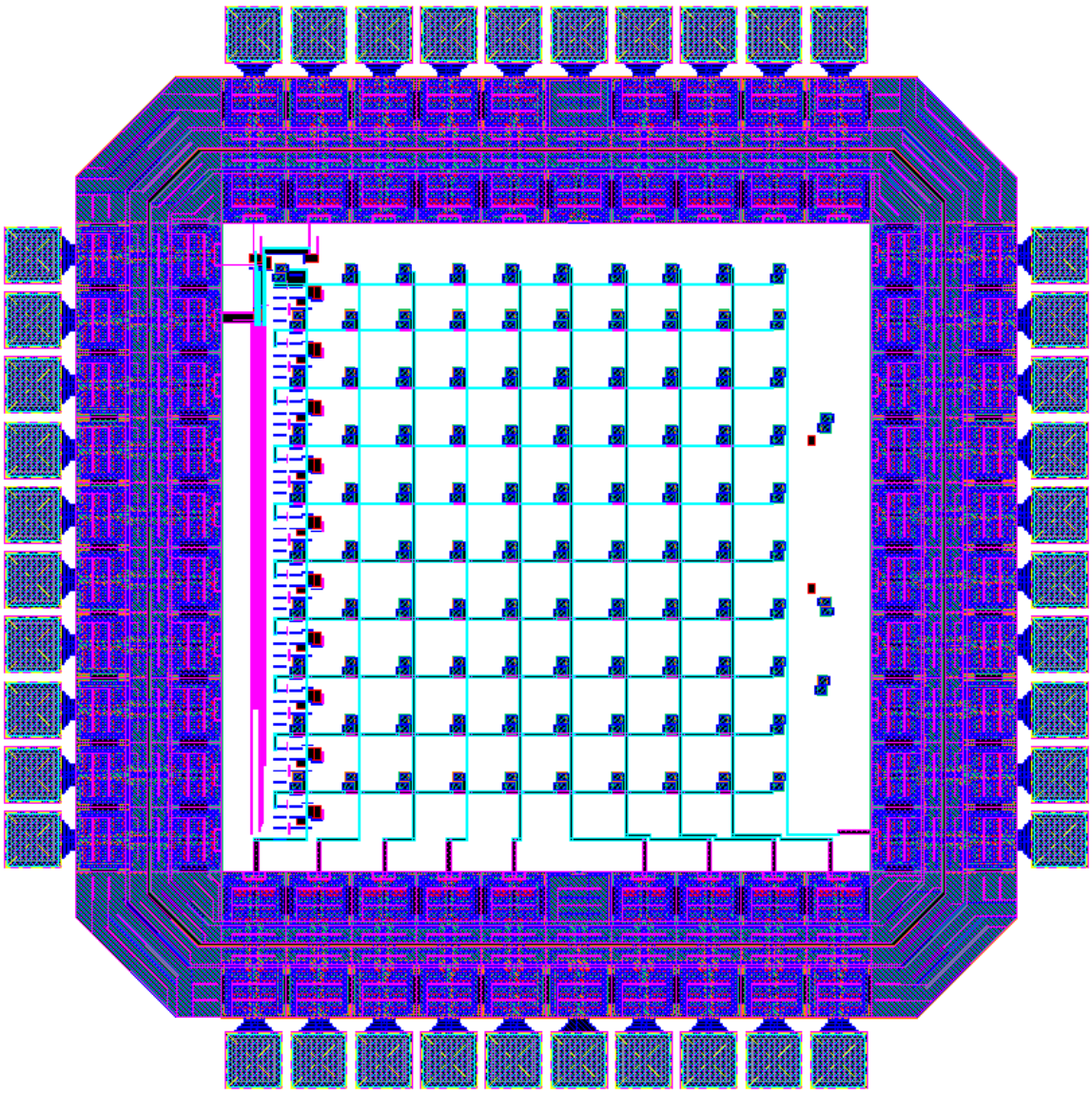

Figure 5.23: A 10×10 diode array is laid out to locally measure temperatures on the chip. To facilitate readout, we included a multiplexer on the left to selectively enable different rows. The chip was laid out using the Cadence Virtuoso tool [89], and was fabricated through MOSIS [94].

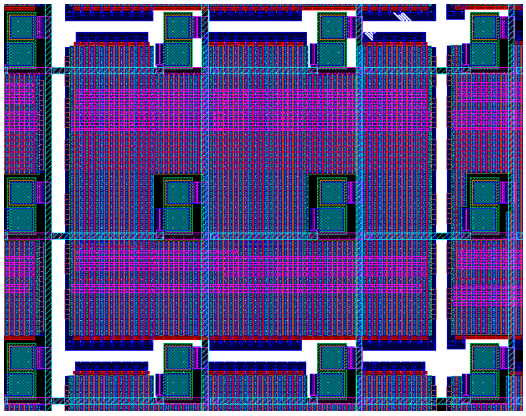

Figure 5.24: A rectangular NMOS microheater block is shown. The NMOS block is comprised of hundreds of smallest size NMOS devices with their gates, sources and drains shorted together to enable maximum heat generation.

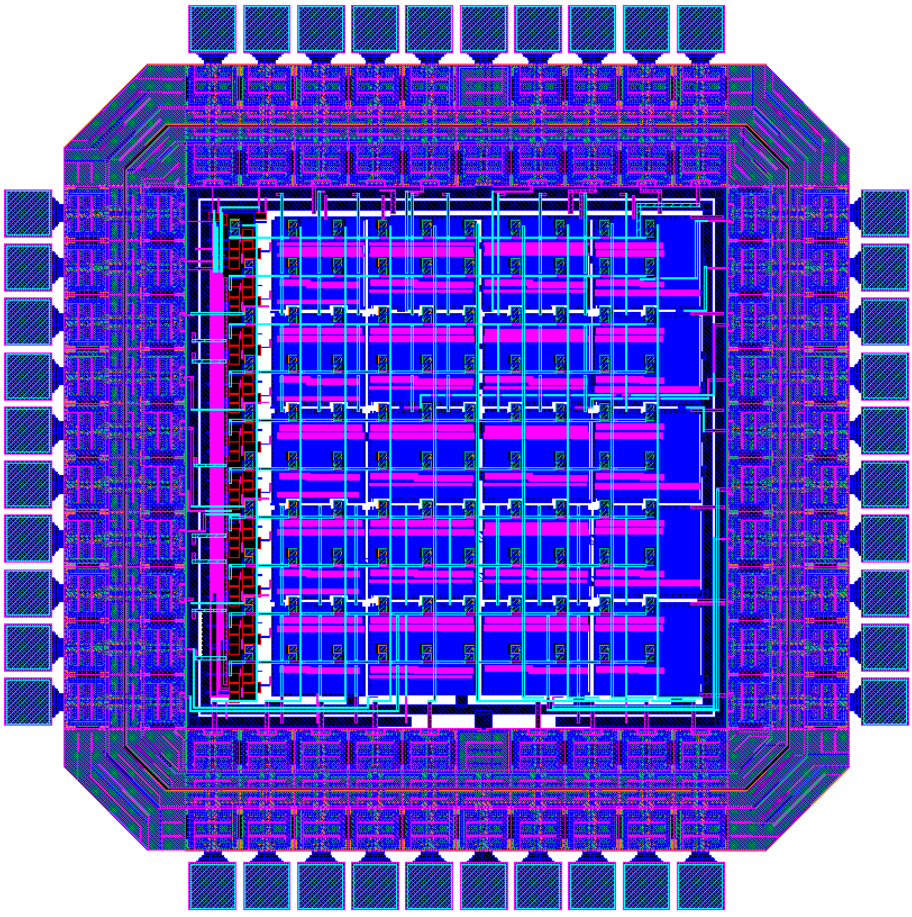

Figure 5.25: An array of 4×4 NMOS heater blocks was superimposed onto the temperature sensing diode array network shown in Fig. 5.23. The chip was laid out using the Cadence Virtuoso tool [89], and was fabricated through MOSIS [94].

Next, instead of using NMOS blocks as microheaters, we employed poly silicon resistors as our heat generators [95]. In addition, we also made the diode array denser by making the diodes smaller to ascertain higher spatial resolution. We show our fabricated chip with the 15×15 diode sensor array and the 4×4 poly silicon resistor heater array in Fig. 5.26. By enabling controlled current flow on different poly silicon resistor blocks, we induced temperature gradients on the surface. To measure the exact temperatures, we prepared a look-up table for the diode currents as functions of temperature. In Fig. 5.27, we show how the measured diode current increases with temperature. Next, we turned on different resistor blocks on the chip, and measured the local temperatures on the surface of the chip using the diode currents. To determine the diode temperature, we compared the observed currents to the ones in their measured current versus temperature look-up tables. In Fig. 5.28, we show our calculated temperature map using the measured diode currents for our chip when the third row-first column poly silicon resistor block is on [95].

In conclusion, we designed thermal sensors and microheaters. We successfully observed the change in diode current as a function of temperature. Moreover, our comparison of measured temperatures with the calculated ones are in accordance. We use the experimental data to calibrate our thermal resistance values for the chip.

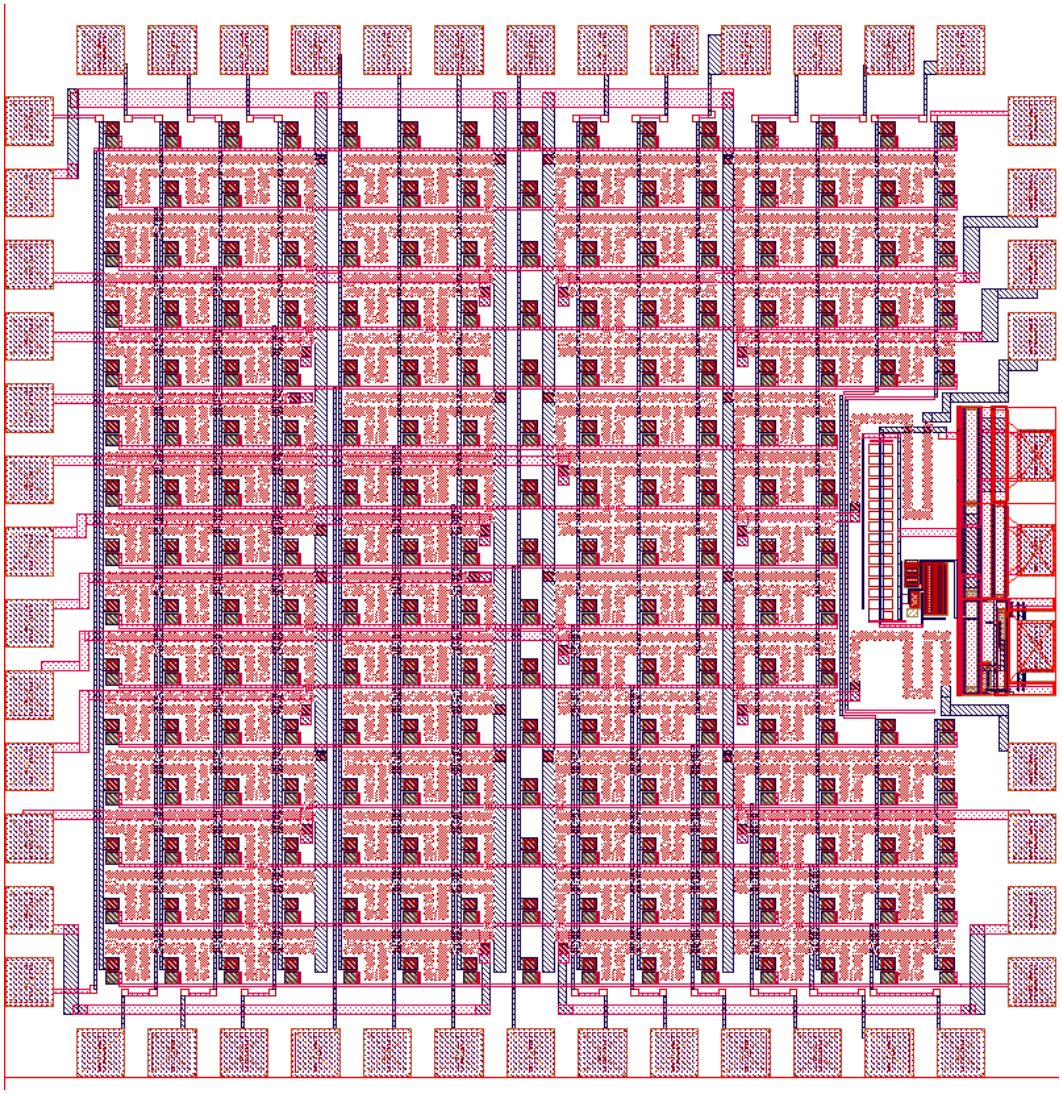

Figure 5.26: Our fabricated chip with the 4×4 poly silicon differential microheater blocks superimposed onto the diode array sensor [95].

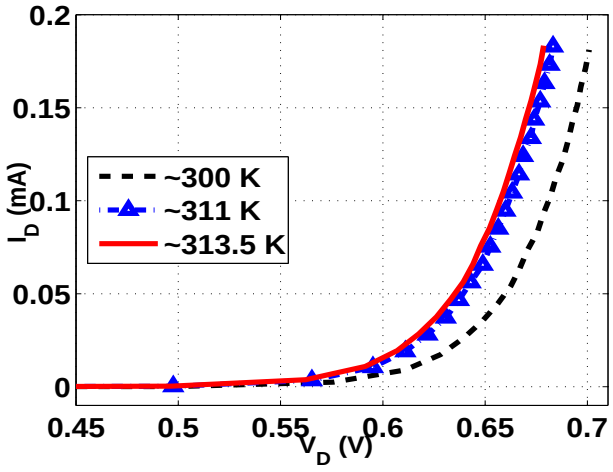

Figure 5.27: Measured current-voltage characteristics of a diode used in the diode array as a function of temperature.

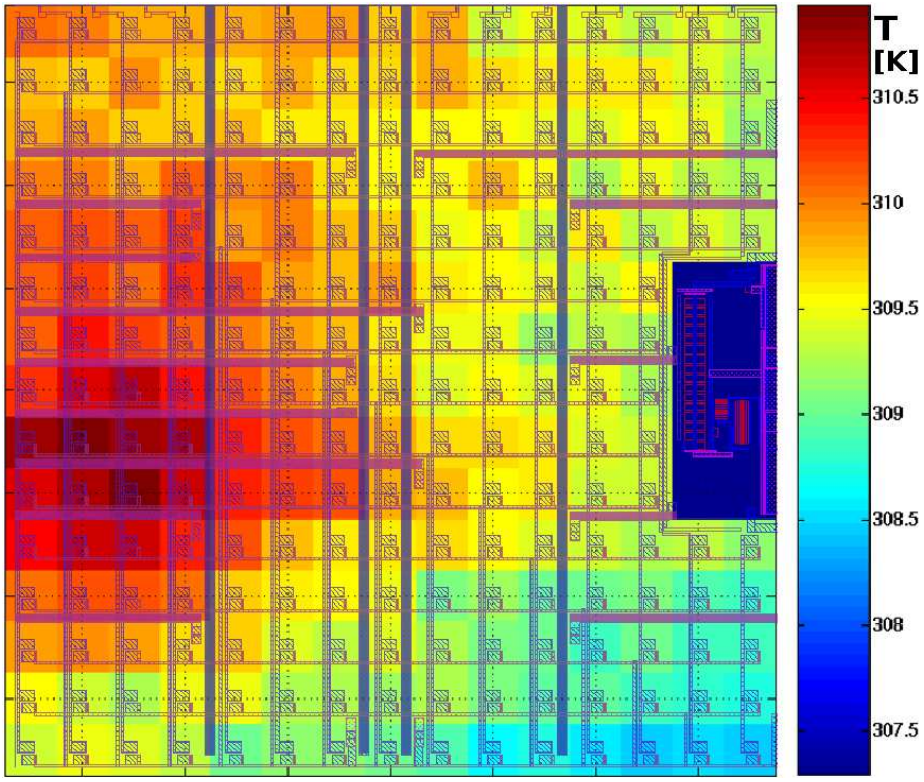

Figure 5.28: Measured temperatures after turning the third row-first column poly resistor block on. Peak temperatures, reaching 10 degrees above the ambient, are induced around this block, as shown on the left of the figure.

5.5 Self-Heating Effects at Cryogenic Temperatures

5.5.1 Device and Chip Model

As use of orbiting satellites steadily increases, the modeling of electronics that efficiently work under extreme space conditions has been gaining great importance. To aid device and chip designs that can work in such low temperatures, we present a methodology for determining device performance details at cryogenic temperatures in conjunction with chip and package details. Using our technique that takes into account package thermal resistances and generated heat, we obtain possible temperature operating conditions for a device used in space applications. Moreover, to enable device operation at higher temperatures that would result in higher transconductances and operating speeds, we also offer methods for initial temperature boosting using heat kick-start circuits.

To obtain device performance details, we solve the coupled semiconductor equations [61, 62] including the Poisson equation, electron and hole current equations, and the differential heat flow equation.

In addition, we explicitly include the temperature dependence on the following parameters [69, 75, 76]: thermal voltage, $V_{\mathrm{TH}}(T)$, intrinsic carrier concentration, $n_o(T)$, electron and hole mobility, $\mu(T)$, electron and hole saturation velocity, $\upsilon_{\mathrm{sat}}(T)$, built-in potentials, $\phi_{\mathrm{built-in}}(T)$, bandgap of silicon, $E_g(T)$, and the thermal diffusion constant , $\kappa(T)$, given in Eqns. 5.7-5.13. Moreover, at cryogenic tempera-

tures, we also consider incomplete ionization effects [75]:

$$D(T) = \frac{N_d^+}{1 + \frac{g_d n(x,y,T)}{N_C e^{-\frac{E_D}{kT}}}} - \frac{N_a^-}{1 + \frac{g_a p(x,y,T)}{N_V e^{-\frac{E_A}{kT}}}} \tag{5.56}$$

Here, N_C and N_V are the effective densities of state at the conduction and valence band edges. Also, E_A and E_D are the energy differences between the acceptor and donor levels, and the valence and conduction bands, respectively. In our study, we take them as both equal to 45meV. Moreover, the net ionized dopant concentration is related to the electron $n(x, y, T)$ and hole $p(x, y, T)$ concentrations, and the net acceptor N_a^- and donor N_d^+ levels (ionized and unionized together). Above, g_d and g_a, which are fitting parameters for the donors and acceptors, are 2 and 4, respectively.

Our analyses indicate that around room temperature down to approximately 100 degrees Kelvin, device performance is mostly affected through changes in carrier mobility, saturation velocity and built-in boundary potentials. As temperature increases from 100°K, mobility and saturation velocity decrease, resulting in lower current values. However, change in built-in boundary potentials results in effective threshold voltage lowering as temperature rises, which increases current. At cryogenic temperatures from 100°K down to 20°K, device performance degrades due to incomplete ionization and low intrinsic carrier concentration. Freeze-out of dopants especially at the source and drain terminals adversely affects drive currents, which leads to dramatic current drops as temperature decreases, as shown in Fig. 5.29.

Our goal is to find out how much power must be added externally to achieve acceptable device performance and to see if unaided power dissipation in the circuit

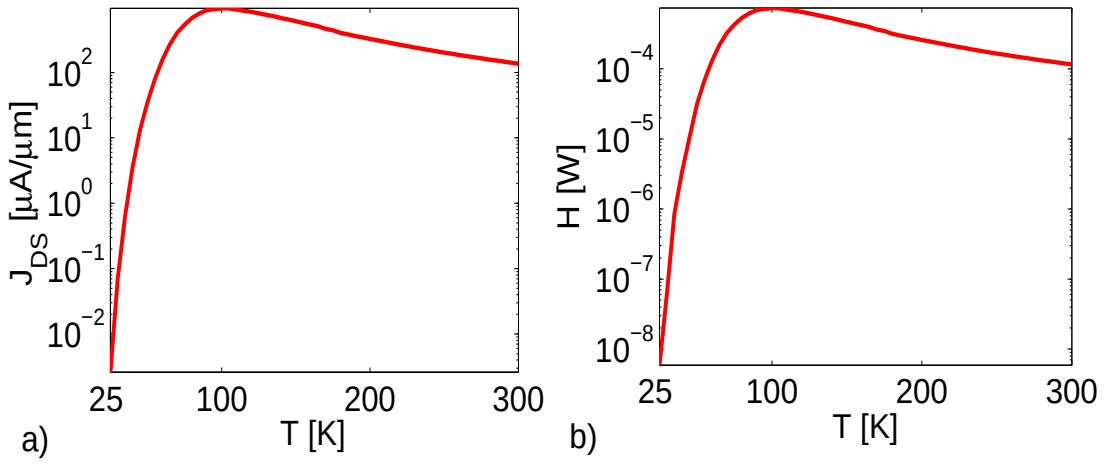

Figure 5.29: Calculated a) current density and b) heat generated of a 0.13μm N-MOSFET. (V_{GS}=V_{DS}=0.7V)

can be sufficient to create sustained device operation.

To obtain effects of chip and package on device temperature, we first calculate a value for the equivalent thermal resistance between that device and the ambient. We have shown that one appropriate value for that resistance is 4×10^{5}K/W. Then, we solve self-consistently for the generated heat and thermal current through a single resistance using the electrical analogy derived previously. We solve the below equation using graphical methods.

$$H = \frac{T - T_A}{R_C} \tag{5.57}$$

Here, H is the heat generated by the device, which is equivalent to the integrated Joule heating over the device volume; T is the device temperature; T_A is the ambient temperature; R_C is the Norton equivalent thermal resistance seen from that node including chip and package. When heat generated, or the source term, is equal to the resistive heat flow, we have an operating temperature point.

5.5.2 Simulation Results

We first find temperature versus heat generated curve of a 0.13μm N-MOSFET at V_{GS}=V_{DS}=0.7V to be used in Eqn. 5.57. We next calculate the heat flow on R_C=4×10^{5}K/W for T_A=40°K, which is the approximate ambient temperature for our satellite application. Then, we determine temperature operating conditions graphically, as shown in Fig. 5.30(a). It shows that we have three intersections at about 43°K, 46°K and 175°K. Moreover, 43°K and 175°K are stable operating temperature points because generated heat is high when device temperature is lower than these values, and vice versa if it is low. Therefore, if we want the circuit to operate at 175°K, a temperature boosting circuit to push initial device temperature higher than the unstable operating point of 46°K is needed.

In Fig. 5.30(b), we graphically solve Eqn. 5.57; however, R_C is 1×10^{6}K/W compared to the previous case. It shows that for a high Norton equivalent resistance, we do not need a heat kick-start circuit. Likewise, for a low Norton equivalent resistance, we have one operating point, which is close to the ambient.

In Fig. 5.31(a), we show the differential microheater and temperature sensors we had fabricated through MOSIS. It contains sixteen heater blocks shown by dashed lines. By turning a specific block on, we provide temperature boosting for a device in a specific location. Figure 5.31(b) shows the effects of such differential heating. (All temperatures were recorded and mapped except the dark area on the right.)

In summary, we provide means to calculate device performance at cryogenic temperatures. Also, we determine package induced self-heating effects for that

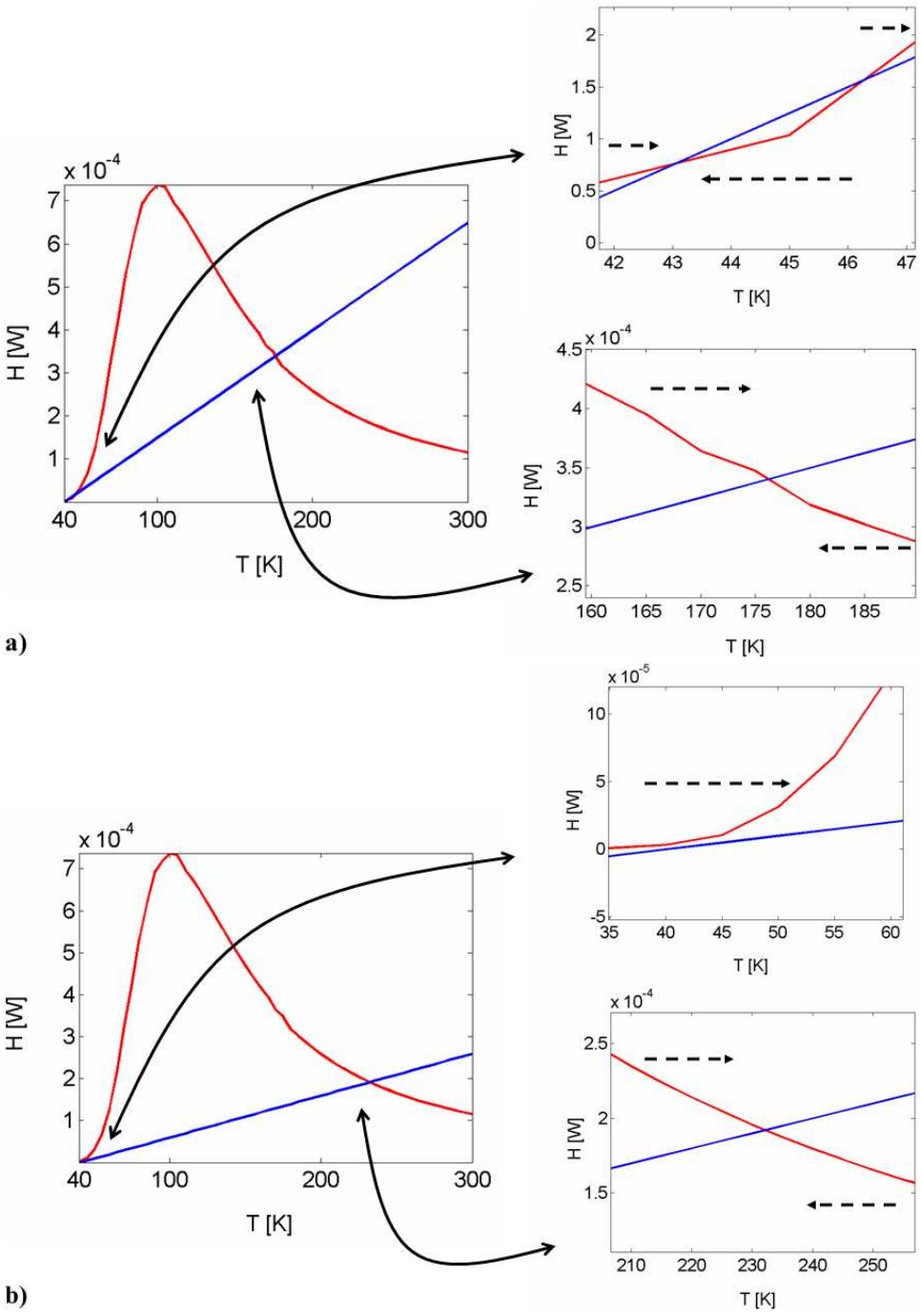

Figure 5.30: Heat generated by a device and the resistive linear thermal current. Intersections (zoomed in on the right) are operating temperature conditions. a) T_{A}=40°K, R_C=4×10^5 K/W b) T_{A}=40°K, R_C=1×10^{6}K/W.

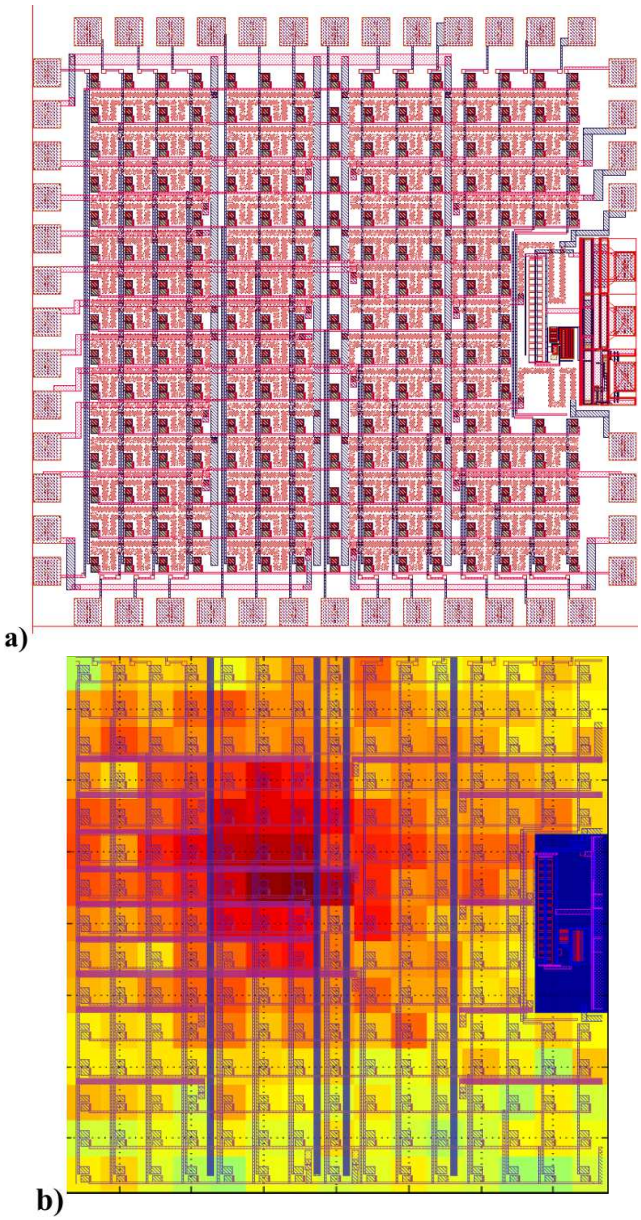

Figure 5.31: a) Our fabricated chip with uniformly distributed 4×4 differential microheater blocks and 15×15 thermal diode sensors. b) Induced temperatures by turning the second row-second column resistor block on. Darker middle region is about seven degrees warmer than the lighter regions.

device. Additionally, we include differential microheater circuits for temperature boosting applications.

5.6 Chapter Summary

In this chapter, we proposed novel methodologies for calculating planar (2D) and 3D IC temperatures at the resolution of a single device. At the device level, we solve the coupled semiconductor equations with the inclusion of quantum effects, to determine the heat generated and current-voltage curves for given device boundary temperatures. We then use the heat generation figures supplied by the device simulator, as current sources on the lumped thermal network. Next, at the chip level, we solve for nodal temperatures that represent device temperatures, using that lumped $R^{\text{th}}C^{\text{th}}$ thermal network. Therefore, the feedback between the chip and IC levels is achieved by obtaining heat generated at the device level for use in the chip's thermal network, and calculating nodal temperatures from the chip's lumped $R^{\text{th}}C^{\text{th}}$ thermal network for use as device temperatures.

Since we have tens of millions of devices on an IC, we developed techniques to extend the device performance results to the overall chip volume. More specifically, we first calculate the device heat generation at different temperatures for given bias conditions that are plausible during switching. However, we solve the device equations for the steady state, then assume that a device is consuming a percentage of its steady state heat generated during switching for the given clock frequency. Next, we extend these device results to the chip's volume using a Monte Carlo type

methodology, where each transistor's heat generation is found by multiplying the calculated full power by a probabilistically determined weighting coefficient that takes into account relative activity levels of different transistors on the chip. To obtain probability density functions that we use to calculate devices' weighting coefficients, we first group the chip's transistors into several functional blocks such as cache, clock, arithmetic logic unit, etc. We then find the normalized power per area for each block. We assume that the block with the highest normalized power per area has devices that are partially or fully on. Using that block's normalized power per area as a benchmark for the on-probability, we calculate the on-probabilities of the other functional blocks. Moreover, using the normalization condition, which states that the integral of the probability density function is equal to one, we obtain the corresponding off-probabilities. We use these probability density functions to determine weighting coefficients for each transistor on the chip. We import the heat generated figures scaled by those coefficients to our lumped $R^{\text{th}}C^{\text{th}}$ thermal network as current sources. This provides the transition from the device to the IC level. Lastly, from the IC layout's geometrical and fabrication related details, we calculate values for the thermal resistances and capacitances. After we obtain the thermal resistances, thermal capacitances and the current sources in the thermal network, we solve for the nodal temperatures that are associated with tens of millions of devices. The solution, after transforming the temperatures back to those before applying Kirchoff's integral, gives the device temperatures, which are directly fed back to device simulations. We iteratively solve for the device temperatures and the heat generated for each device until convergence.

Our numerical predictions for the 2D and 3D ICs indicate that chip heating is an important problem to be overcome for successful device and chip operation. As devices are scaled down, power densities are increasing rapidly, with the heating figures exponentially diverging from safe operational limits. The self-heating is more pronounced in 3D ICs, where each layer is separated by an electrical insulator such as SiO_2 that is also a thermal insulator. To relieve the heating problem, we offer methods to pull the extremely high chip temperatures locally or globally to lower temperatures that are within safe operational bounds. The first novel method we offer is the use of thermal vias that span from the top to the bottom of the chip to extract heat vertically away from hot spots, especially the high temperature middle layers of the 3D ICs. Our investigations show that uniformly distributed low thermal resistance vias successfully relieve the heating problem. However, as the thermal resistance of these vias get bigger, they become less influential. Instead of trying to reduce the overall temperature using a uniform array of vias, their concentrated utilization around hot spots successfully helps heat removal from hot regions. Furthermore, we also investigate the use of metal lines or horizontal thermal contacts, which extend from one side of the chip to the other, for heat removal. This method proves to be less effective than the use of vertical thermal vias. However, they are effective for removal of heat from the hot regions close to the chip's boundaries. In addition, we also note that a rearrangement of the chip's layout can distribute the active devices uniformly over the chip's volume, making the temperature gradients on the chip smoother and the temperature values lower.

Lastly, we show how the knowledge of non-isothermal device performance de-

tails helps us build systems and chips operating at room temperature and cryogenic temperatures. We show how design considerations change at cryogenic temperatures. Unlike operation at room temperature, device and chip operation at cryogenic temperatures rely on self-heating effects. We provide methodologies to predict temperature operating points. Also, we offer design strategies, which can result in more efficient chip operation using kick-start microheater circuits.

Chapter 6

Thesis Publications

6.1 Journal Publications

A. Akturk, N. Goldsman, and G. Metze, "Self-consistent modeling of heating and mosfet performance in three-dimensional integrated circuits," IEEE Trans. on Electron Devices **52(11)**, 2395 (2005).

A. Akturk, N. Goldsman, L. Parker, and G. Metze, "Mixed-mode temperature modeling of full-chip based on individual non-isothermal device operations," Solid-State Electronics **49(7)**, 1127 (2005).

A. Akturk, G. Pennington, and N. Goldsman, "Quantum modeling and proposed designs of carbon nanotube (cnt) embedded nanoscale mosfets," IEEE Trans. on Electron Devices **52(4)**, 577 (2005).

A. Akturk, N. Goldsman, and G. Metze, "Increased cmos inverter switching speed with asymmetrical doping," Solid-State Electronics **47(2)**, 185 (2003).

6.2 Conference Publications

A. Akturk, G. Pennington, N. Goldsman and A. Wickenden, "Quantum electron transport in carbon nanotubes: length dependence and velocity oscillations," Int. Conf. on Simulation of Semiconductor Processes and Devices (SISPAD), accepted (6-8 Sept. 2006).

A. Akturk, N. Goldsman, Z. Dilli and M. Peckerar, "Device performance and package induced self-heating effects at cryogenic temperatures," Int. Conf. on Simulation of Semiconductor Processes and Devices (SISPAD), accepted (6-8 Sept. 2006).

Z. Dilli, N. Goldsman, A. Akturk and G. Metze, "A 3-d time-dependent greens function approach to modeling electromagnetic noise in on-chip interconnect networks," Int. Conf. on Simulation of Semiconductor Processes and Devices (SISPAD), accepted (6-8 Sept. 2006).

A. Akturk, N. Goldsman, and G. Metze, "An efficient inclusion of self-heating and quantum effects in soi device simulations," Int. Semiconductor Device Research Symp. (ISDRS), 99 (7-9 Dec. 2005).

A. Akturk, N. Goldsman, N. Dhar, and P. S. Wijewarnasuriya, "Modeling the temperature dependence and optical response of hgcdte diodes," Int. Semiconductor Device Research Symp. (ISDRS), 70 (7-9 Dec. 2005).

G. Pennington, A. Akturk, J. M. McGarrity, and N. Goldsman, "Transport properties of wide band gap nanotubes," Int. Semiconductor Device Research Symp. (IS-

DRS), 346 (7-9 Dec. 2005).

Z. Dilli, N. Goldsman, and A. Akturk, "An impulse-response based methodology for modeling complex interconnect networks," Int. Semiconductor Device Research Symp. (ISDRS), 64 (7-9 Dec. 2005).

A. Akturk, G. Pennington, and N. Goldsman, "Numerical device analysis of all-around gate carbon nanotube (cnt) embedded field-effect transistors (fets)," 16th European Conf. on Diamond, Diamond-Like Materials, Carbon Nanotubes and Nitrides (11-16 Sep. 2005).

G. Pennington, A. Akturk, and N. Goldsman, "Low-field electronic transport in single-walled semiconducting carbon nanotubes," 16th European Conf. on Diamond, Diamond-Like Materials, Carbon Nanotubes and Nitrides (11-16 Sep. 2005).

A. Akturk, N. Goldsman, and G. Metze, "Coupled simulation of device performance and heating of vertically stacked three-dimensional integrated circuits," Int. Conf. on Simulation of Semiconductor Processes and Devices (SISPAD), 51 (1-3 Sept. 2005).

A. Akturk, G. Pennington, and N. Goldsman, "Device behavior modeling for carbon nanotube silicon-on-insulator mosfets," Int. Conf. on Simulation of Semiconductor Processes and Devices (SISPAD), 115 (1-3 Sept. 2005).

G. Pennington, A. Akturk, and N. Goldsman, "Low-field transport model for semiconducting carbon nanotubes," Int. Conf. on Simulation of Semiconductor Processes and Devices (SISPAD), 87 (1-3 Sept. 2005).

G. Pennington, A. Akturk, and N. Goldsman, "Phonon-limited transport in carbon nanotubes using the monte carlo method," Int. Workshop on Computational Electronics (IWCE-10) (24-27 Oct. 2004).

A. Akturk, G. Pennington, and N. Goldsman, "Numerical performance analysis of carbon nanotube (cnt) embedded mosfets," Int. Conf. on Simulation of Semiconductor Processes and Devices (SISPAD), 153 (2-4 Sept. 2004).

A. Akturk, G. Pennington, and N. Goldsman, "Temperature dependent mobility model for single-walled zig-zag carbon nanotubes (cnts)," 8th Int. Conf. on Nanometer-Scale Science and Technology (NANO-8), 728[1846], (28 June - 2 July 2004).

A. Akturk, G. Pennington, and N. Goldsman, "Characterisation of nanoscale carbon nanotube (cnt) embedded cmos inverters," 8th Int. Conf. on Nanometer-Scale Science and Technology (NANO-8), 769[413], (28 June - 2 July 2004).

A. Akturk, L. Parker, N. Goldsman, and G. Metze , "Mixed-mode simulation of non-isothermal quantum device operation and full-chip heating," Int. Semiconductor Device Research Symp. (ISDRS), 508 (10-12 Dec. 2003).

G. Pennington, A. Akturk, and N. Goldsman, "Electron mobility of a semiconducting carbon nanotube," Int. Semiconductor Device Research Symp. (ISDRS), 412 (10-12 Dec. 2003).

A. Akturk, N. Goldsman, and G. Metze, "Coupled modeling of time-dependent

full-chip heating and quantum non-isothermal device operation," Int. Conf. on Simulation of Semiconductor Processes and Devices (SISPAD), 311 (3-5 Sept. 2003).

A. Akturk, G. Pennington, and N. Goldsman, "Modeling the enhancement of nanoscale mosfets by embedding carbon nanotubes in the channel," 3rd IEEE Conf. on Nanotechnology (IEEE-NANO) **1** , 24 (12-14 Aug. 2003).

A. Akturk, N. Goldsman, and G. Metze, "Faster cmos inverter switching obtained with channel engineered asymmetrical halo implanted mosfets," Int. Semiconductor Device Research Symp. (ISDRS), 118 (5-7 Dec. 2001).

BIBLIOGRAPHY

[1] http:\\www.intel.com.

[2] P. Gelsinger, "Microprocessors for the new millennium: challenges, opportunities, and new frontiers," ISSCC, 22 (2001).

[3] D.J. Frank, R.H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H. P. Wong, "Device scaling limits of si mosfets and their application dependencies," Proc. of the IEEE **89(3)**, 259 (2001).

[4] R. Ronen, A. Mendelson, K. Lai, S-L. Lu, F. Pollack, and J. P. Shen, "Coming challenges in microarchitecture and architecture," Proc. of the IEEE **89(3)**, 325 (2001).

[5] F. Pollack, "New microarchitecture challenges in the coming generations of cmos process technologies," keynote presentation at the 32nd Int. Symp. on Microarchitecture (1999).

[6] S. Im, and K. Banerjee, "Full chip thermal analysis of planar (2d) and vertically stacked integrated (3d) high performance ics," IEDM, 727 (2000).

[7] A. Akturk, N. Goldsman, and G. Metze, "Coupled modeling of time-dependent full-chip heating and quantum non-isothermal device operation," SISPAD, 311 (2003).

[8] A. Akturk, L. Parker, N. Goldsman, and G. Metze, "Mixed-mode simulation of non-isothermal quantum device operation and full-chip heating," ISDRS, 508 (2003).

[9] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, "Temperature-aware microarchitecture," ISCA, 2 (2003).

[10] E. Pop, K. Banerjee, P. Sverdrup, R. Dutton, and K. Goodson, "Localized heating effects and scaling of sub-0.18 micron cmos devices," IEDM, 31.1.1 (2001).

[11] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, "3d ics: a novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration," Proc. of the IEEE **89(5)**, 602 (2001).

[12] M. Lundstrom, "A top-down look at bottom-up electronics," Symp. on VLSI Cir., 5 (2003).

[13] H. S. P. Wong, "Field effect transistors - from silicon mosfets to carbon nanotube fets," MIEL **1**, 103 (2002).

[14] J. Guo, S. Datta, and M. Lundstrom, "Assesment of silicon mos and carbon nanotube fet performance limits using a general theory of ballistic transistors," IEDM, 29.3.1 (2002).

[15] Ph. Avouris, "Molecular electronics with carbon nanotubes," Acc. Chem. Res. **35**, 1026 (2002).

[16] R. Saito, M. S. Dresselhaus, and G. Dresselhaus, *Physical properties of carbon nanotubes* (Imperial College Press, London, 1998).

[17] G. Pennington, and N. Goldsman, "Semiclassical transport and phonon scattering on electrons in semiconducting carbon nanotubes," Phys. Rev. B **68**, 45426 (2003).

[18] G. Pennington, and N. Goldsman, "Monte carlo study of electron transport in a carbon nanotube," IEICE Trans. Electron. **E86-C**, 372 (2003).

[19] H. C. d'Honincthun, S. G.-Retailleau, J. See, and P. Dollfus, "Electron-phonon scattering and ballistic behavior in semiconducting carbon nanotubes," Appl. Phys. Lett. **87**, 172112 (2005).

[20] T. Durkop, S. A. Getty, E. Cobas, and M. S. Fuhrer, "Extraordinary mobility in semiconducting carbon nanotubes," Nano Lett. **4**, 35 (2004).

[21] T. Durkop, B. M. Kim, and M. S. Fuhrer, "Properties and applications of high-mobility semiconducting nanotubes," Jour. Phys.: Condens. Matt. **16**, R553 (2004).

[22] R. S. Lee, H. J. Kim, J. E. Fischer, A. Thess, and R. E. Smalley, "Conductivity enhancement in single-walled carbon nanotube bundles doped with k and br," Nature **388**, 255 (1997).

[23] L. Grigorian, G. U. Sumanasekera, A. L. Loper, S. Fang, J. L. Allen, and P. C. Eklund, "Transport properties of alkali-metal-doped single-wall carbon nanotubes," Phys. Rev. B **58**, R4195 (1998).

[24] V. Derycke, R. Martel, J. Appenzeller, and Ph. Avouris, "Controlling doping and carrier injection in carbon nanotube transistors," Appl. Phys. Lett. **80(15)**, 2773 (2002).

[25] R. Martel, V. Derycke, C. Lavoie, J. Appenzeller, K. K. Chan, J. Tersoff, and Ph. Avouris, "Ambipolar electrical transport in semiconducting single-wall carbon nanotube," Phys. Rev. Lett. **87(25)**, 256805 (2001).

[26] J.-O. Lee, C. Park, J.-J. Kim, J. Kim, J. W. Park, and K.-H. Yoo, "Formation of low-resistance ohmic contacts between carbon nanotube and metal electrodes by a rapid thermal annealing method," Jour. Phys. D: Appl. Phys. **33**, 1953 (2000).

[27] M. H. Yang, K. B. K. Teo, W. I. Milne, and D. G. Hasko, "Carbon nanotube schottky diode and directionally dependent field-effect transistor using asymmetrical contacts," Appl. Phys. Lett. **87**, 253116 (2005).

[28] L. Vitali, M. Burghard, P. Wahl, M. A. Schneider, and K. Kern, "Local pressure-induced metallization of a semiconducting carbon nanotube in a crossed junction," Phys. Rev. Lett. **96**, 086804 (2006).

[29] Z.-B. Zhang, S.-L. Zhang, and E. E. B. Campbell, "All-around contact for carbon nanotube field-effect transistors made by ac dielectrophoresis," Jour. Vac. Sci. Tech. B **24(1)**, 131 (2006).

[30] A. Akturk, G. Pennington, and N. Goldsman, "Modeling the enhancement of nanoscale mosfets by embedding carbon nanotubes in the channel," 3rd IEEE Conf. on Nanotech., 24 (2003).

[31] Akturk, G. Pennington, and N. Goldsman, "Numerical performance analysis of carbon nanotube (cnt) embedded mosfets," SISPAD, 153 (2004).

[32] Z. Yao, C. L. Kane, and C. Dekker, "High-field electrical transport in single-wall carbon nanotubes," Phys. Rev. Lett. **84**, 2941 (2000).

[33] J. Hone, M. Whitney, C. Piskoti, and A. Zettl, "Thermal conductivity of single-walled carbon nanotubes," Phys. Rev. B **59**, R2514 (1999).

[34] A. Bachtold, P. Hadley, T. Nakanishi, and C. Dekker, "Logic circuits with carbon nanotube transistors," Science **294**, 1317 (2001).

[35] R. Martel, V. Derycke, J. Appenzeller, S. Wind, and Ph. Avouris, "Carbon nanotube field-effect transistors and logic circuits," IEEE Design Auto. Conf., 94 (2002).

[36] S. J. Tans, A. R. M. Verschueuren, and C. Dekker, "Room temperature transistor based on a single carbon nanotube," Nature **393**, 49 (1998).

[37] S. J. Wind, J. Appenzeller, R. Martel, V. Derycke, and Ph. Avouris, "Vertical scaling of carbon nanotube field-effect transistors using top gate electrodes," Appl. Phys. Lett. **80(20)**, 3817 (2002).

[38] J. Appenzeller, J. Knoch, and Ph. Avouris, "Carbon nanotube field-effect transistors - an example of an ultra-thin body schottky barrier devices," Device Research Conf., 167 (2003).

[39] J. Appenzeller, J. Knoch, R. Martel, V. Derycke, S. Wind, and Ph. Avouris, "Short-channel like effects in schottky barrier carbon nanotube field-effect transistors," IEDM, 285 (2002).

[40] N. Goldsman, *Modeling electron transport and degradation mechanisms in semiconductor submicron devices*, Ph. D. Thesis, Cornell University, NY (1989).

[41] M. Lundstrom, *Fundamentals of carrier transport* (2^{nd} ed., Cambridge University Press, Cambridge, 2000).

[42] S. Datta, *Quantum phenomena (modular series on solid state devices, vol 8)* (Addison-Wesley, 1989).

[43] S. M. Ross, *Introduction to probability and statistics for engineers and scientists* (Wiley & Sons, 1987).

[44] TCAD Taurus Tsupreme-4, http:\\www.synopsys.com/products/tcad/taurus_tsuprem4_ds.html.

[45] A. Akturk, *Investigation of transient and dc characteristics of cmos inverters*, M.S. Thesis, University of Maryland College Park, MD (2001).

[46] A. Akturk, G. Pennington, and N. Goldsman, "Quantum modeling and proposed designs of cnt-embedded nanoscale mosfets," IEEE Trans. Electron Dev. **52(4)**, 577 (2005).

[47] C. K. Huang, and N. Goldsman, "2-d self-consistent solution of schrödinger equation, boltzmann transport equation, poisson and current-continuity equation for mosfet," SISPAD, 148, Spring-Verlag (2001).

[48] Z. Ren, R. Venugopal, S. Goasguen, S. Datta, and M. S. Lundstrom, "nanoMOS 2.5: a two-dimensional simulator for quantum transport in double-gate MOSFETs," IEEE Trans. Electron Dev. **50(9)**, 1914 (2003).

[49] P. B. M. Wolbert, G. K. M. Wachutka, and B. H. Krabbenborg, T. J. Mouthaan, "Nonisothermal device simulation using the 2d numerical process/device simulator trendy and application to soi devices," CAD of ICS **13(3)**, 293 (1994).

[50] A. S. Spinelli, A. Benvenuti, and A. Pacelli, "Self-consistent 2-d model for quantum effects in n-mos transistors," IEEE Elect. Dev. **45**, 1342 (1998).

[51] M. G. Ancona, and G. J. Iafrate, "Quantum correction to the equation of state of an electron gas in a semiconductor," Phys. Rev. B **39(13)**, 9536 (1989).

[52] M. G. Ancona, "Equations of state for silicon inversion layers," IEEE Elect. Dev. **47**, 1449 (2000).

[53] M. G. Ancona, Z. Yu, R. W. Dutton, P. J. V. Voorde, M. Cao, and D. Vook, "Density-gradient analysis of mos tunneling," IEEE Elect. Dev. **47(12)**, 2310-9 (2000).

[54] J. R. Watling, A. R. Brown, and A. Asenov, "Can the density gradient approach describe the source-drain tunneling in decanano double-gate mosfets?," Jour. Computational Elect. **1**, 289 (2002).

[55] M. G. Ancona, "Macroscopic description of quantum-mechanical tunneling," Phys. Rev. B **42(2)**, 1222 (1990).

[56] M. G. Ancona, and H. F. Tiersten, "Macroscopic physics of the silicon inversion layer," Phys. Rev. B **35(15)**, 7959 (1987).

[57] H. Tsuchiya, B. Fischer, and K. Hess, "A full-band Monte Carlo model for silicon nanoscale devices with a quantum mechanical correction of the potential," IEDM, 283 (2000).

[58] J. R. Watling, A. R. Brown, A. Asenov, A. Svizhenko, and M. P. Anantram, "Simulation of direct source-to-drain tunneling using the density gradient formalism: non-equilibrium green's function calibration," SISPAD, 267 (2002).

[59] T-W Tang, and B. Wu, "Quantum correction for the monte carlo simulation via the effective conduction-band edge equation," Semicond. Sci. Tech. **19(1)**, 54 (2004).

[60] Y. Leblebici, S. Unlu, S-M. Kang, and B. M. Onat, "Transient simulation of heterojunction photodiodes-part I: computational methods," Jour. of Lightwave Tech. **13(3)**, 396 (1995).

[61] A. Akturk, N. Goldsman, L. Parker, and G. Metze, "Mixed-mode temperature modeling of full-chip based on individual non-isothermal device operations," Solid-State Electronics **49(7)**, 1127 (2005).

[62] A. Akturk, N. Goldsman, and G. Metze, "Self-consistent modeling of heating and mosfet performance in three-dimensional integrated circuits," IEEE Trans. Elec. Dev. **52(11)**, 2395 (2005).

[63] P. Carruthers, and F. Zachariasen, "Quantum collision theory with phase-space distributions," Rev. of Modern Phys. **55(1)**, 245 (1983).

[64] Z. Han, N. Goldsman, and C-H Lin, "Incorporation of quantum corrections to semiclassical two-dimensional device modeling with the wignerboltzmann equation," Solid-State Electronics **49(2)**, 145 (2005).

[65] J. Appenzeller, "Electronic transport in semi-conducting carbon nanotube transistor devices," talk can be found at https:\\www.nanohub.org (2003).

[66] B. R. Nag, *Theory of electrical transport in semiconductors* (Pergamon Press, 1971).

[67] G. Pennington, A. Akturk, and N. Goldsman, "Phonon-limited transport in carbon nanotubes using the monte carlo method," IWCE-10 (2004).

[68] X. Zhou, J.-Y. Park, S. Huang, J. Liu, and P. L. McEuen, "Band structure, phonon scattering, and the performance limit of single-walled carbon nanotube transistors," Phys. Rev. Lett. **95**, 146805 (2005).

[69] R. S. Muller, *Device electronics for integrated circuits* (2^{nd} ed., Wiley & Sons, New York, 1986).

[70] B. G. Streetman, *Solid state electronic devices* (4^{th} ed., Prentice Hall, 1995).

[71] G. C. Goodwin, S. F. Graebe, and M. E. Salgado, *Control system design* (Prentice Hall, New Jersey, 2001).

[72] A. Charlier, R. Setton, and M-F. Charlier, "Energy component in a lattice of ions and dipoles: application to the K(THF)1,2C24 compounds," Phys. Rev. B **55(23)**, 15537 (1997).

[73] X. Wang, H-S. P. Wong, P. Oldiges, and R. J. Miller, "Electrostatic analysis of carbon nanotube arrays," SISPAD, 163 (2003).

[74] http:\\www-mtl.mit.edu\Well\.

[75] S. Powell, *Surface physics modeling and evaluation of 6H-silicon carbide metal-oxide-semiconductor field effect transistors with experimental corroboration*, Ph. D. Thesis, University of Maryland College Park, MD (2003).

[76] S. M. Sze, *Physics of semiconductor devices* (2^{nd} ed., Wiley & Sons, New York, 1981).

[77] S. Datta, "Nanoscale device modeling: the green's function method," Superlattices and Microstructures **28(4)**, 253 (2000).

[78] S. S. Lee, and D. J. Allstot, "Electrothermal simulation of integrated circuits," Solid-State Cir. **28(12)**, 1283 (1993).

[79] D. Brooks, V. Tiwari, and M. Martonosi, "Wattch: a framework for architectural-level power analysis and optimizations," Int. Conf. on Computer Architecture **28(2)**, 83 (2000).

[80] M. Martonosi, D. Brooks, and P. Bose, "Modeling and analyzing cpu power and performance: metrics, methods, and abstractions," tutorial given at SIGMETRICS (2001).

[81] E. A. Dengi, and R. A. Rohrer, "Hierarchical 2d field solution for capacitance extraction for vlsi interconnect modeling," Design Auto. Conf., 127 (1997).

[82] V. W. C. Chan, P. C. H. Chan, and M. Chan, "3d integrated circuit using large grain polysilicon film," Int. Conf. on Solid-State and Integrated-Circuit Tech. **1**, 58 (2001).

[83] S.-Y. Oh, K. Rahmat, O. S. Nakagawa, and J. Moll, "A scaling scheme and optimization methodology for deep sub-micron interconnect," ICCD, 320 (1996).

[84] O. Steffens, P. Szabo, M. Lenz, and G. Farkas, "Thermal transient characterization methodology for single-chip and stacked structures," STHERM, 313 (2005).

[85] S. P. Gurrum, S. K. Suman, Y. K. Joshi, and A. G. Fedorov, "Thermal issues in next-generation integrated circuits," IEEE Trans. on Device and Materials Reliability **4(4)**, 709 (2004).

[86] L. Zhang, N. Howard, V. Gumaste, A. Poddar, and L. Nguyen, "Thermal characterization of stacked-die packages," STHERM, 55 (2004).

[87] G. N. Ellison, "Thermal analysis of microelectric packages and printed circuit boards using an analytic solution to the heat conduction equation," Advances in Eng. Software **22(2)**, 99 (1995).

[88] J.-M. Koo, L. Jiang, L. Zhang, P. Zhou, S. S. Banerjee, T. W. Kenny, J. G. Santiago, and K. E. Goodson, "Modeling of two-phase microchannel heat sinks for vlsi chips," MEMS, 422 (2001).

[89] http:\\www.cadence.com

[90] T.-Y. Wang, and C. C.-P. Chen, "3d-thermal adi: a linear-time chip level transient thermal simulator," IEEE Trans. on CAD of ICs and Systems **21(12)**, 1434 (2002).

[91] T.-Y. Wang, Y.-M. Lee, and C. C.-P. Chen, "3d-thermal adi: an efficient chip-level transient thermal simulator," ACM/SIGDA ISPD (2003).

[92] M. Khbeis, G. Metze, N. Goldsman, and A. Akturk, "Use of thermally conductive vias to extract heat from microelectronic chips and method of manufacturing," United States Patent Application, no. 20050254215 (2005).

[93] D. D. L. Chung, "Materials for thermal conduction," Applied Thermal Eng. **21**, 1593 (2001).

[94] http://mosis.org

[95] Zeynep Dilli, private communication (2005).