CHAPTER 11

MONTE CARLO TRANSPORT AND HEAT GENERATION IN SEMICONDUCTORS

Eric Pop

Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA; E-mail: epop@stanford.edu

Understanding transport and energy use in nanoscale devices is essential for the design of low-power electronics and efficient thermoelectrics. This chapter examines transport physics and the electron-phonon interaction in the context of Monte Carlo simulations, which treat electrons and phonons with comparable attention to detail. The Monte Carlo method is described in depth, including scattering physics, electron energy band and phonon dispersions, Poisson solution, and contacts within realistic devices. This approach uncovers, for instance, that Joule heating in silicon devices is distributed between the (slow) optical and (fast) acoustic phonon modes by a ratio of two to one. In nanoscale transistors, nonequilibrium transport affects heat generation near strongly peaked electric fields, and Joule heating occurs almost entirely in the drain of short, quasi-ballistic devices. Evidence is also uncovered for thermionic cooling at the source terminal of transistors, and the physics of this phenomenon are described. Although the discussion is often with respect to silicon for specificity, key methods can be broadly applied to many semiconductor devices and structures. Such aspects are only expected to increase in importance as nanoscale devices are reduced to dimensions comparable to or smaller than the electron and phonon mean free paths (10–100 nm).

KEY WORDS: *Monte Carlo, transport, transistors, nanoscale, ballistic, energy dissipation*

1. INTRODUCTION

The power generated in nanoscale transistors and semiconductors is the fundamental source of the heat dissipated within circuits and processors, and in applications ranging from mobile devices ($\sim 10^{-3}$ W) to data centers ($\sim 10^{9}$ W), all primarily based on silicon nanotechnology today. At the individual level of the central processing unit (CPU) or microprocessor, the dissipated power has virtually stopped the race to increase operating frequency beyond a few GHz; for example, desktop computer CPUs cannot dissipate more than $\sim$100 W/cm^2 due to heat removal challenges,[1] whereas mobile CPUs are limited to $<$1 W/cm^2 in part due to passive cooling restrictions, and in part due to limited battery lifetime. At a much larger scale, data centers installed in the United States in 2011 had a total power consumption of $\sim$10 GW, equivalent to $\sim$2.5% of the national electricity budget and to the output of 10 large nuclear power plants.[1] Worldwide, if "cloud computing" were a country,* it would be among the six most electricity consuming countries

*Greenpeace International[2] estimated the worldwide cloud computing electricity use was 623 TWh per year in 2007 ($\sim$70 GW). This estimate may be high by a factor of two for data centers alone ($\sim$30 GW worldwide)[3] but becomes more accurate if *all* cloud-connected electronics (not just data centers) are considered. By comparison, China uses an estimated yearly 4700 TWh (2011) and the United States 3800 TWh (2009).[4]

ISSN: 1049–0787; ISBN: 1–978–56700–295–9/14/$35.00 + $00.00

© 2014 by Begell House, Inc.

NOMENCLATURE

Symbol	Description
a	lattice constant, m
D_A	acoustic deformation potential, eV
E_G	band gap energy, eV or J
E_k	kinetic energy, eV or J
$\mathbf{F}$	electric field, V/m
g_d	electron density of states, $cm^{-3}eV^{-1}$
g_p	phonon density of states, $cm^{-3}eV^{-1}$
G_n	generation rate, $cm^{-3}s^{-1}$
$\mathbf{G}$	reciprocal lattice vector, m^{-1}
I	electrical current, A
J	electrical current density, A/cm^2
k_B	Boltzmann constant, eV/K or J/K
$\mathbf{k}$	electron momentum, 1/m
L_D	Debye length, m
m^*	conduction effective mass, kg
m_d	density of states effective mass, kg
n	electron density, cm^{-3}
N_q	phonon occupation
p	hole density, cm^{-3}
P	power, W
P'''	power density, W/cm^3
q	elementary charge, C
$\mathbf{q}$	phonon momentum, 1/m
R	electrical resistance, Ω
R_C	electrical contact resistance, Ω
R_n	recombination rate, $cm^{-3}s^{-1}$
t	time, s
T	absolute temperature, K

Greek Symbols

Symbol	Description
α	energy band coefficient, eV^{-1}
Δ_{if}	intervalley deformation potential, eV/cm
ε_s	semiconductor dielectric constant, F/m
Γ_0	total scattering rate, 1/s
$\hbar\omega$	phonon energy, eV
λ	mean free path, m
ϕ	angle between $\mathbf{k}$ and $\mathbf{q}$, rad
Φ	voltage potential, V
ρ	mass density, g/cm^3
ρ_c	charge density, C/cm^3
τ	drift time, s
Ξ_u	shear deformation potential, eV
Ξ_d	dilation deformation potential, eV

in the world.[2–4] In addition, the installed data center capacity in the world has been increasing at 12% per year,[2] a trend not yet expected to slow down.

This chapter is concerned with the fundamental aspects of electron-phonon scattering within semiconductors, which give rise to the power dissipated in electronic circuits. Section 2 provides a brief review of Joule heating in devices. Section 3 describes some historical and contextual aspects of the Monte Carlo simulation approach. Section 4 presents an in-depth discussion of a Monte Carlo implementation, and Section 5 applies this approach to a description of transport physics in bulk silicon, and in silicon devices. Finally, Section 6 extends the Monte Carlo method to the study of energy dissipation in silicon, particularly highlighting sub-continuum aspects that come into play as device dimensions are reduced to the $\sim$10 nm scale.

2. BRIEF REVIEW OF JOULE HEATING IN DEVICES

Power generation and heat dissipation within semiconductor devices like transistors begins with the interaction between charge carriers (electrons or holes) and lattice vibrations

(phonons), as shown in Fig. 1. Applied voltages create electric fields, which accelerate charge carriers until they reach sufficient energy to emit a net positive flux of phonons. (In thermal equilibrium, the rates of phonon emission and phonon absorption by charge carriers are equal.) The simplest approach to estimate the power generated in an electronic device is derived from Ohm's law as

$$P = I^2(R - R_C) \tag{1}$$

where I is the current flowing through the device and $R - R_C$ is the intrinsic device resistance, excluding any contact resistance R_C. This is the classical expression for a device much larger than the electron and phonon mean free paths (10–100 nm in typical semiconductors operating near room temperature). However, this expression will overestimate the total power dissipated in a quasi-ballistic device,[5,6] i.e., one with dimensions comparable to or shorter than the inelastic scattering length, where the heat generated is due to only one or two discrete phonon emission events, as illustrated in Fig. 2(a).[1] In addition, the simple lumped expression in Eq. (1) does not describe the spatial distribution of heat dissipated within a semiconductor device.

2.1 Drift-Diffusion Model for Energy Dissipation in Transistors

In order to compute the spatial distribution of power dissipation, the approach most commonly used is given by drift-diffusion-based device simulations,[7–10]

$$P^{'''} = \mathbf{J} \cdot \mathbf{F} + (R_n - G_n)(E_G + 3k_BT) \tag{2}$$

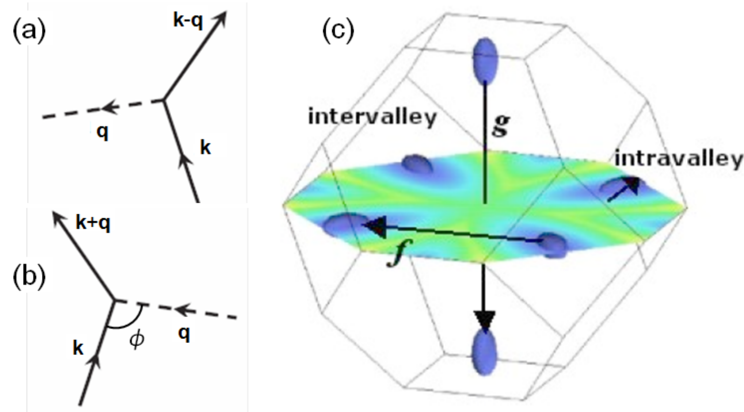

FIG. 1: Feynman diagrams of (a) phonon emission and (b) phonon absorption. Here **k** and **q** are momenta of the electron and phonon, respectively. (c) Schematic of allowed electron-phonon interactions in silicon, shown across the Brillouin zone in three-dimensional (3D) **k**-space. Intravalley transitions are those *within* one of the six conduction band minima. Intervalley transitions occur *between* two of the six equivalent conduction band minima. The f and g phonons are labeled on the phonon dispersion in Fig. 5(b). (Image courtesy C. Jungemann.)

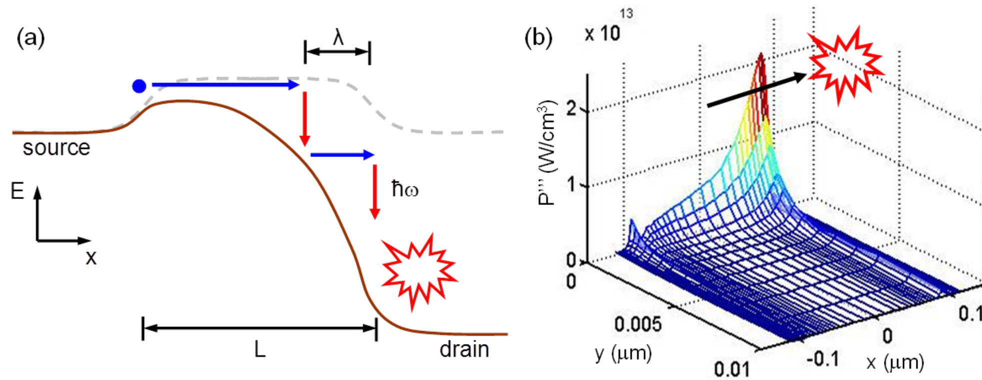

FIG. 2: Energy band diagram of electron transport, phonon emission, and heat dissipation in a quasi-ballistic transistor (L comparable to the mean free path λ). Here heat generation can be due to just a few discrete phonon emission events. (b) Drift-diffusion simulation of heat generation in a 0.18 μm long MOSFET, where heating is computed using Eq. (2) in the text. The arrow indicates the direction of electron flow from source to drain, past the location of peak electric field. The vertical plane at $y = 0$ marks the SiO_2/Si interface.

where **J** is the current density, **F** is the electric field, $(R_n - G_n)$ is the net nonradiative recombination rate (recombination minus generation), E_G is the band gap, and T is the lattice temperature. Equation (2) is typically implemented as a finite element simulation on a device grid, as shown in Fig. 2(b). Note the notation of $P^{'''}$ (power density per unit volume, e.g., in units of W/cm^3) versus Eq. (1) (total power in Watts). The total power P can be recovered by integrating Eq. (2) over the device volume V, although in small devices where hot carriers can easily escape through the contacts the actual power dissipated internally is typically $P < \int P^{'''} \mathrm{d}V$.[1]

The dot product in Eq. (2) represents the well-known Joule term, which is typically positive (power generation) as electrons drift down the band structure slope under the influence of the electric field, and gradually lose energy through net phonon emission. The Joule term can also be negative (power consumption) when electrons diffuse *against* an energy barrier,[†] and the energy required to move up the conduction band slope is extracted from the lattice through net phonon absorption.[9,11] The second term of Eq. (2) is the net heat generation rate due to non-radiative electron and hole generation and recombination processes. When an electron and a hole, both with an average energy (3/2)$k_B T$ recombine, the excitation energy $E_G + 3k_B T$ is given off either directly to the lattice, or to another charge carrier (Auger transition). In the latter case, the excited particle eventually gives off the energy to the lattice by phonon emission as well. Electron scattering with defects or impurities typically does not contribute directly to lattice heating, but can contribute indirectly by affecting the electron momentum distribution function. Equation (2) may include other terms, e.g., for electron drift along a temperature gradient (Thomson effect),[12] across a band discontinuity between two different materials (Peltier effect),[13] or at a heterojunction

[†]Such as in a forward-biased pn junction, or near the energy barrier at the injection point from the source into the channel of a MOSFET, later discussed in Fig. 13.

like in a semiconductor laser.[9,11] In direct band gap materials, optical recombination power can also play a role, radiatively cooling the device through a negative term $-\eta_{\text{opt}} J E_G / q$, which could be included in Eq. (2), where η_{opt} is the optical quantum efficiency.[13−15]

Figure 2(b) shows the heat generation rate computed in a 0.18 μm gate length silicon transistor with the approach described above, as implemented in the commercial simulator Medici. Unfortunately, this field-dependent method does not account for the microscopic nature of heat generation near a strongly peaked electric field region, such as in the drain of the transistor. Although electrons gain most of their energy at the location of the peak field, they travel several mean free paths before releasing it to the lattice, in decrements of (at most) the optical phonon energy. In silicon, the optical phonon energy is $h\omega_{OP} \approx 60$ meV and in carbon nanotubes or graphene it is approximately three times greater (160–200 meV). Typical inelastic scattering mean free paths are of the order $\lambda_{OP} \approx 10$ nm.[16] The full electron energy relaxation length is thus even longer, i.e., several inelastic mean free paths. While such a discrepancy may be neglected on length scales of microns, or even tenths of a micron, it must be taken into account when simulating transport on length scales of 10 nm, as in nanoscale transistors.[17] The highly localized electric field in such devices leads to the formation of a nanometer-sized hot spot in the drain region, which is spatially displaced (by several mean free paths) from this drift-diffusion prediction. This scenario is illustrated in Fig. 2(a) for a few discrete phonon generation events in a quasi-ballistic transistor channel. In such a situation, the $\mathbf{J} \cdot \mathbf{F}$ drift-diffusion approach cannot capture the delocalized nature of the power dissipation region.

2.2 Hydrodynamic Model for Energy Dissipation in Transistors

An improvement over the drift-diffusion approach is provided by the hydrodynamic model,[9,19] which introduces the electron and hole temperature ($T_{n,p}$) and an average electron and hole energy relaxation time (τ_{nL}, τ_{pL}) to compute the power dissipation as[20]

$$\begin{aligned} P''' = {} & \frac{3}{2} k_B \left[\frac{n(T_n - T_L)}{\tau_{nL}} + \frac{p(T_p - T_L)}{\tau_{pL}} \right] \\ & + q(G-R)_{bb} \left[T_p \left(\frac{\partial \phi_p}{\partial T_p} \right) - T_n \left(\frac{\partial \phi_n}{\partial T_n} \right) + \phi_n - \phi_p \right] \end{aligned} \tag{3}$$

Here the subscript L refers to the lattice, *bb* refers to band-to-band processes,[20] and other quantities are as defined earlier. ϕ_n and ϕ_p are the electron and hole quasi-Fermi levels, respectively. Unlike the drift-diffusion model, this approach is better suited for capturing transport near highly peaked electric fields. However, the hydrodynamic model suffers from the simplification of a single averaged carrier temperature and relaxation time, as scattering rates are strongly energy dependent.[21] In addition, in practical device simulations the hydrodynamic method often offers challenges in achieving convergence. Neither of the two methods summarized above gives information regarding the frequencies and wave vectors of phonons emitted. Such details are important because the emitted phonons have different velocities, different scattering rates and lifetimes, and widely varying contributions to heat transport[22−25] and device heating.[26,27]

2.3 Monte Carlo Model for Energy Dissipation in Transistors

The mechanism by which lattice self-heating occurs is that of electron scattering with phonons, and therefore a model that deliberately incorporates all scattering events will also capture such energy dissipation details. Thus, the Monte Carlo (MC) method,[28] originally developed for studying hot electron effects,[29] is also well suited for computing a detailed picture of energy dissipation. This was the approach adopted in Refs. 30–35, where power dissipation was computed as a sum of all phonon emission minus all phonon absorption events,

$$P''' = \frac{n}{N_{\text{sim}}\Delta t} \sum (\hbar\omega_{\text{ems}} - \hbar\omega_{\text{abs}}) \tag{4}$$

where n is the real-space carrier density, N_{sim} is the number of simulated particles (e.g., 10,000 simulated particles could be used to describe 10^{19} cm^{-3} real-space concentration), and Δt is the time. This approach has been used to investigate phonon emission as a function of phonon frequency and mode in silicon, as well as to study heat generation near a strongly peaked electric field in a realistic device geometry. In the remainder of this chapter we describe in greater detail the MC model for heat generation in semiconductors, with particular attention to the electron-phonon interaction where the lattice heating processes begin.

3. MONTE CARLO METHOD FOR TRANSPORT IN SEMICONDUCTORS

The Monte Carlo method is regarded as the most comprehensive approach for simulating charge transport in semiconductors. An early standard was set by the work of Canali et al.[36] and that of Jacoboni and Reggiani[28] using analytic, ellipsoidal descriptions of the energy band structure of silicon. Over the past three decades, the research community has added numerous enhancements, including more comprehensive physical models, more efficient computer algorithms, new scattering mechanisms, boundary conditions, electrostatic self-consistency in device simulations, etc. A significant enhancement of the physical models was the introduction of full electron energy bands from empirical pseudopotential calculations.[29,37] The full-band MC method has been very useful with high-field and high-energy transport simulations, including impact ionization,[37,38] where details of the full band structure are essential.

3.1 Historical Overview

Figure 3 shows a brief historical overview of various MC simulation methods for charge transport in silicon. Canali et al.[36] introduced the first multivalley model with parabolic, ellipsoidal bands and phonon scattering with a single dispersionless longitudinal acoustic (LA) mode and six fixed-energy intervalley phonons. Jacoboni et al.[39] accounted for analytic band non-parabolicity and slightly altered Canali's set of phonon deformation potentials. A few years later, Brunetti et al.[40] introduced a new set of deformation potentials, more closely matching available data on the anisotropy of electron diffusion in silicon. This phonon model was used by Jacoboni and Reggiani[28] in an excellent and frequently

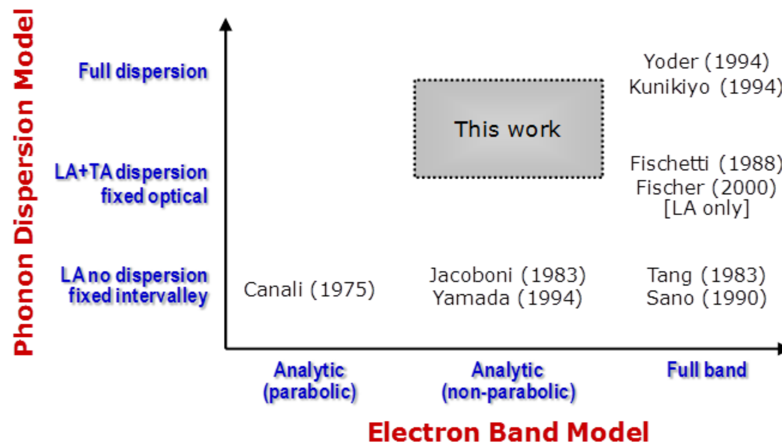

FIG. 3: Historical context of various Monte Carlo models for electron transport in silicon. The computational burden increases for full-band (and full dispersion) simulations.

referenced review of the MC method, and it subsequently became the set of phonon energies and deformation potentials most often employed in the literature over the past two decades. Other workers[41] also introduced scattering with first-order intervalley phonons. Tang and Hess[37] were the first to incorporate the full band structure of silicon (computed from empirical pseudopotentials) for MC transport. However, they used the simple phonon model of Canali and Brunetti (dispersionless LA phonons, six fixed intervalley phonons), and the deformation potentials of Brunetti et al.[40] Sano et al. introduced wave vector dependent impact ionization rates in a full-band MC formulation,[38] but computed phonon scattering rates with the multivalley deformation potentials of Canali et al.[36]

Realistic device simulations using electrostatically self-consistent full-band MC were first performed by Fischetti and Laux.[29] They were also the first to make the distinction between longitudinal (LA) and transverse acoustic (TA) intravalley scattering, using a simple analytic dispersion for both modes. Fischer and Hofmann[42] pointed out the poor definition of energy "valleys" in the context of full-band models, and used only two averaged deformation potentials: one for fixed-energy optical phonons and another for acoustic phonons (LA, but not TA), including their dispersion. The most sophisticated MC models for charge transport in silicon were developed by Yoder and Hess[43] and Kunikiyo et al.[44] They employed the full band structure computed from empirical pseudopotentials and the full (anisotropic) phonon dispersion obtained from an adiabatic bond-charge model. The electron-phonon scattering rates were calculated as a function of energy and wave vector, consistently with the band structure and phonon dispersion. In the absence of any adjustable parameters, mobilities computed with these ab initio models are typically less accurate than those computed using more empirical simulators. Such codes also present formidable computational burdens, rendering them impractical for simulations of realistic devices. Their only applications have been for very detailed bulk transport calculations.

Most MC codes found in practice today employ a sophisticated, full description of the electron energy bands (often including quantum effects),[45] yet scattering rates and energy

exchange with the lattice are only computed with a simplified phonon dispersion.[46,47] The phonon energies and deformation potentials in use most often are those originally introduced by Brunetti et al.[40] Optical phonon dispersion is ignored and often only one acoustic branch (LA) is considered for intravalley scattering. Such models can lead to unphysical thresholds in the electron distribution function[42] and cannot be used to compute phonon generation rates for detailed phonon dynamics simulations (e.g., phonon Boltzmann transport or molecular dynamics). In a realistic electron device, a full phonon dispersion is essential for extracting the correct phonon generation spectrum from Joule heating.[30,48] Use of the full phonon dispersion is also important in strained or confined materials and devices, where the dispersion relationship is altered from its bulk form. In the discussion below, we describe a MC code that uses analytic descriptions for both the electron bands and the phonon dispersion. This computationally efficient method is suitable for simulating low-voltage nanodevices, while treating the electron bands and phonon dispersion with equal attention.

3.2 General Monte Carlo Aspects

The general aspects of the Monte Carlo method for charge transport in semiconductors have been well described before.[28,49,50] This section provides but a brief overview of the MC algorithm, summarized with the diagram in Fig. 4. The ensemble MC approach used in this work preselects several tens of thousands "super-particles" to represent the mobile charge inside the semiconductor. This number is limited by computational (and to a lesser extent, today, by memory) constraints, but good statistics can be obtained if the simulation is run for an adequately long time. The particles are initialized with thermal energy distributions (average energy $3k_BT/2$) and with randomly oriented momenta. Spatially, in the case of a realistic device simulation (as opposed to modeling the transport properties of bulk silicon), the particles are initially distributed following the device doping profile or based on initial conditions read from, for example, a drift-diffusion device simulator. Once the simulation is started, the particles are allowed to drift for short periods of time (τ shorter than the average time between collisions), then a scattering process (if any) is selected. A fictive "self-scattering" rate can be chosen such that the sum of all scattering rates is constant (Γ_0) and independent of the carrier energy. The distribution of each particle's free flight time intervals τ is then directly related to this total scattering rate as[49]

$$\tau = -\frac{1}{\Gamma_0} \ln(r_1) \tag{5}$$

where r_1 is a random number uniformly distributed between 0 and 1. During its free flight, the carrier is allowed to drift under the influence of the electric fields, as dictated by Newton's laws of motion with an effective mass (as opposed to the free electron mass), which represents the collective influence of the lattice. Then another random number r_2 between 0 and 1 is drawn[‡] and $r_2\Gamma_0$ is compared with cumulative probabilities of scattering, which

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

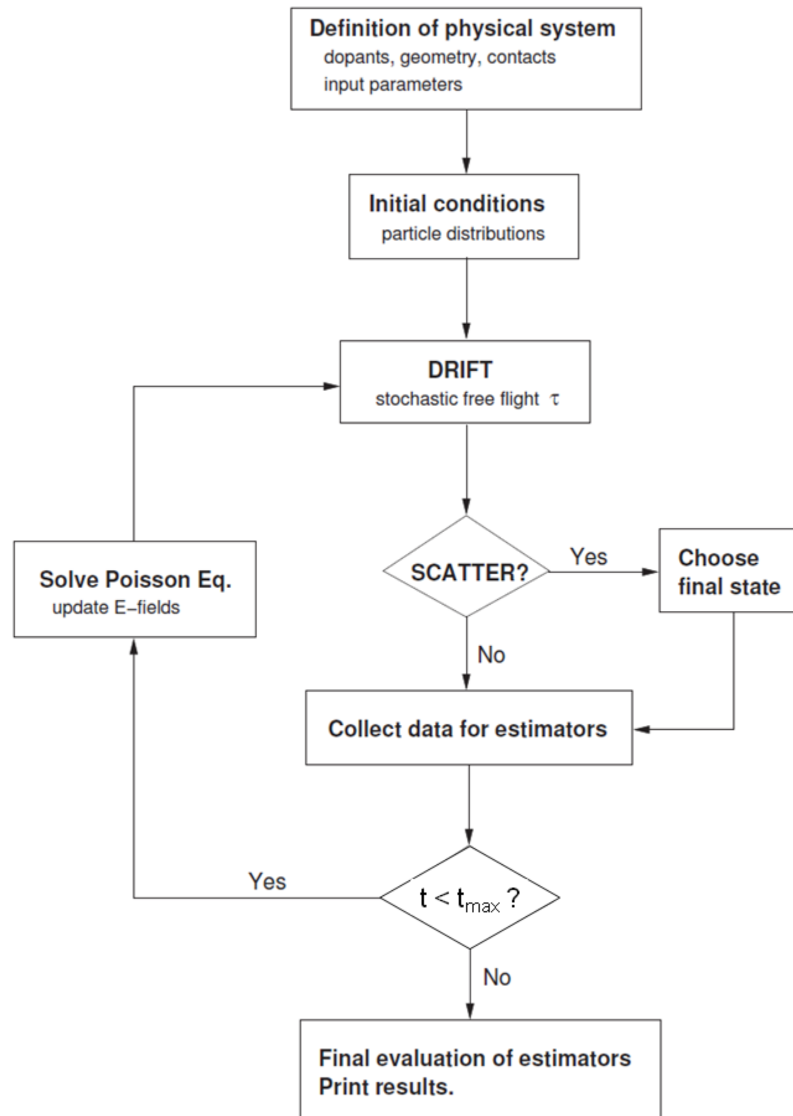

FIG. 4: Basic Monte Carlo algorithm flowchart, after Ref. 51.

have been pre-computed at the beginning of the simulation as a function of energy. A scattering mechanism (e.g., with impurities, acoustic or optical phonons) is selected in proportion to the strength of each process. If self-scattering is selected, the particle continues its free flight unimpeded. If a real scattering process is selected, the particle's state after scattering is stochastically chosen taking into account both energy and momentum conservation, then another random time of flight is drawn. This procedure then repeats for all particles.

In the case of a realistic device simulation, the Poisson equation must be solved at every time step, to self-consistently update the electric fields as the mobile charge carriers

move inside the device. The Monte Carlo simulation can also be run without the Poisson equation, in postprocessor mode on the fixed ("frozen") fields initially read from a drift-diffusion simulator, although extensive work has shown[52] that the results are less accurate and predictive, particularly for noise simulations. The super-particles are treated as single carriers during their free flights, and as charge clouds when the Poisson equation is solved. The cloud-in-cell method[50] is most often employed for assigning the super-particle charge to the grid nodes before Poisson's equation is solved. The charge on each super-particle is

$$Q = q\frac{N}{N_{\text{sim}}} \tag{6}$$

where q is the elementary charge, N is the total number of mobile charges in the device, and N_{sim} is the number of super-particles used in the simulation. It should be noted that the coupled solution to Poisson's equation yields a much more stringent requirement on the simulation time step, necessary to avoid charge imbalance due to plasma oscillations.[49] The Poisson equation therefore ought to be solved every

$$\Delta t < \frac{1}{2}\sqrt{\frac{\varepsilon_s m^*}{q^2 n}} \tag{7}$$

where ε_s is the dielectric constant of the semiconductor, m^* is the lighter effective mass of the carrier in the material (e.g., the transverse mass m_t for electrons in silicon), and n is the mobile charge density. In the heavily doped contact regions of a device, where $n \approx 10^{20}$ cm^{-3}, very short (and therefore time-consuming) time steps of <1 fs are necessary. The charge density at the device contacts must also be updated at the end of each time step. This is done by injecting (or deleting) thermal electrons at the grid nodes adjacent to the contacts, to maintain charge neutrality there. Ensemble averages are updated every time step, and statistics are gathered by sampling the super-particle system at regular time intervals, until reaching a targeted accuracy. The error margins are inversely proportional to the square root of the number of particles being simulated, $(N_{\text{sim}})^{-1/2}$. The run of the algorithm ends when the total time allotted for the simulation ends (typically, on the order of tens or hundreds of picoseconds), or when enough statistics have been gathered and the error margins of the sought-after ensemble averages are deemed appropriate. It should be noted that Monte Carlo simulations are not well suited for low-field transport, where other, simpler but much faster methods may be preferred (e.g., drift diffusion). However, the method represents the most physically comprehensive simulation approach for charge transport in semiconductors, and is usually the standard against which all other methods are judged. Several reference works have been dedicated to thorough reviews of the Monte Carlo method[28,49,50] and additional information can be gathered therein.

4. MONTE CARLO IMPLEMENTATION

This section describes the implementation of a Monte Carlo model for electron transport, specifically developed to compute heat (phonon) generation rates in bulk and strained

silicon, as well as in simple nanoscale device geometries. The model uses analytic, non-parabolic electron energy bands and an isotropic, analytic phonon dispersion model, which distinguishes between the optical/acoustic and longitudinal/transverse phonon branches. A unified set of deformation potentials for electron-phonon scattering is used to yield accurate transport simulations (versus the available data) in bulk and strained silicon across a wide range of electric fields and temperatures. The Monte Carlo model is then applied in the context of transport in 1D (self-consistent with the Poisson equation) and 2D device geometries.

4.1 Electron Energy Band Model

This work models the electron energy bands analytically, following Jacoboni and Reggiani,[28] and including the non-parabolicity parameter α (= 0.5 eV^{-1} at room temperature). With $\alpha = 0$, the kinetic energy is purely parabolic and Canali's original model[36] is recovered. All six ellipsoidal, energetically equivalent conduction band valleys of silicon are explicitly included, as in Fig. 1(c). Figure 5(a) shows a comparison between the total conduction band density of states (DOS) computed in the non-parabolic band approximation and the full-band DOS. From the point of view of the DOS, which determines the scattering rates (described in Section 4.3), the analytic band approximation is sufficient up to 1.5 eV in electron energy. These energies are sufficient for future low-voltage nanotechnologies, where impact ionization and high energy transport are not expected to play a significant

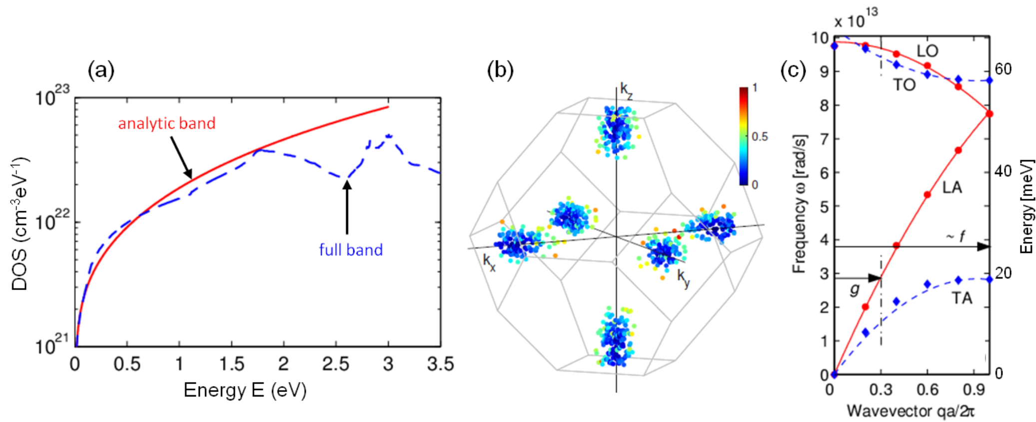

FIG. 5: (a) Conduction band density of states (DOS) in silicon from full-band calculation (courtesy C. Jungemann) compared to the DOS computed with the non-parabolic analytic approximation from Eq. (8). (b) Electron distribution in momentum space, for an electric field 50 kV/cm in the $\langle 111 \rangle$ direction, at 300 K. The color bar represents the electron energy, in electron volts. (c) Phonon dispersion in silicon along the $\langle 100 \rangle$ direction, from neutron scattering data (symbols).[53] The lines represent the quadratic approximation introduced in Ref. 21. The f- and g-phonons participate in the intervalley scattering of electrons,[54] as labeled on Fig. 1(c). (Reprinted with permission from American Institute of Physics Publishing LLC, Copyright 2015.)[30]

role. The analytic, non-parabolic relationship between electron energy E_k and the wave vectors k_i (i = 1, 2, or 3, for the three Cartesian axes) is

$$E_k\,(1+\alpha E_k) = \frac{\hbar^2}{2}\sum_{i=1}^{3}\frac{(k_i-\kappa_{vi})^2}{m_i} \tag{8}$$

where m_i is the component of the electron mass tensor along the ith direction and κ_{vi} represents the coordinates of the respective conduction band minimum. Silicon has six equivalent conduction band minima near the X symmetry points, located at ±85% of the way to the edge of the Brillouin zone along the three $\langle 100\rangle$ axes (the Δ lines), as shown in Figs. 1(c) and 5(b). For example, the X-valley (sometimes also called the Δ-valley) along the $\langle 100\rangle$ direction is centered at (0.85, 0,0)G where $|\mathbf{G}| = 2\pi/a$ is the reciprocal lattice vector and a = 5.431 Å is the silicon lattice constant. The mass tensor components are the longitudinal mass m_l/m_o = 0.916 and the transverse mass m_t/m_o = 0.196 at room temperature, where m_o is the free electron mass. The temperature dependence of the band gap $E_G(T)$ is also included analytically, following Ref. 55,

$$E_G(T) = 1.1756 - 8.8131\times 10^{-5}T - 2.6814\times 10^{-7}T^2 \tag{9}$$

where T is the absolute temperature in Kelvin. This dictates a slight temperature dependence of the transverse mass as $m_t/m_o = 0.196E_{G0}/E_G(T)$ and of the non-parabolicity parameter as $\alpha = 0.5E_{G0}/E_G(T)$ eV^{-1}, where E_{G0} is the silicon band gap at room temperature.[55] Figure 5(b) shows a typical "snapshot" of the electron distribution in momentum space, as computed here.

4.2 Phonon Dispersion Model

The present work treats all phonon scattering events inelastically, hence the electrons exchange the correct amount of energy (corresponding to the absorption or emission of a phonon) with each scattering event. Particular attention is paid to the treatment of inelastic acoustic phonon scattering, to properly account for energy dissipation at low temperatures and low electric fields. Treating the acoustic phonons inelastically is also important for heat generation calculations, as shown in Section 5 and Ref. 30. Figures 1(c) and 5(b) illustrate the ellipsoidal conduction band valleys and the allowed phonon scattering transitions. As in the traditional analytic-band approach,[28] scattering with six types of intervalley phonons is incorporated. Intervalley scattering can be of g-type, when electrons scatter between valleys on the same axis, e.g., from $\langle 100\rangle$ to $\langle -100\rangle$, or of f-type when the scattering occurs between valleys on perpendicular axes, e.g., from $\langle 100\rangle$ to $\langle 010\rangle$. The phonons involved in these scattering transitions (three of f-type and three of g-type) can be determined from geometrical arguments[54] and are labeled in Fig. 6(a).[53] Intravalley scattering refers to scattering within the same conduction band valley and usually involves only acoustic phonons.[56]

Most typical MC implementations,[28,38–42] both analytic- and full-band, have treated intravalley scattering with a single kind of acoustic phonon. This simplification is accomplished by grouping the longitudinal acoustic (LA) and transverse acoustic (TA) branches

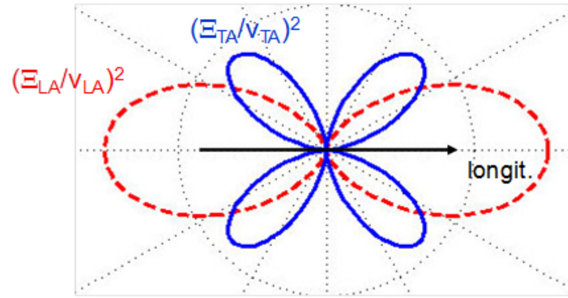

FIG. 6: Polar plot showing the angular dependence of the deformation potentials for electron intravalley scattering with LA and TA phonons in silicon [Eqs. (20) and (21)]. The angle θ is between the phonon wave vector and the longitudinal axis of the conduction band valley [Fig. 1(c)]. The isotropic, angle-averaged integrals lead to the expressions in Eqs. (22) and (23).

into a dispersionless mode with a single velocity and a single deformation potential. Historically, TA modes have been neglected because their matrix element is zero for intravalley scattering within a band located at the center of the Brillouin zone.[28,56] This is not the case for silicon, hence in a more comprehensive approach (where scattering with *all* phonon modes matters), intravalley scattering with TA modes should be considered. Unlike the traditional approach, this work considers scattering with LA and TA modes separately. Each phonon dispersion branch from Fig. 5(c), including the optical modes, is treated with the isotropic approximation

$$\omega_q = \omega_0 + v_s q + c q^2 \tag{10}$$

where ω_q is the phonon frequency and q the wave vector. For the acoustic phonons, the parameters v_s and c can be chosen to capture the slope of the dispersion near the Brillouin zone center and the maximum frequency at the zone edge, similar to Ref. 42. The choice of parameters for longitudinal optical (LO) phonons insures that they meet the zone edge LA frequency. For both TA and transverse optical (TO) phonons, the zone edge slope, i.e., their group velocity, is fit to zero. The continuous (longitudinal) and dashed (transverse) lines in Fig. 5(c) represent these quadratic approximations, and the fitting coefficients are listed in Table 1.[21,30] Quartic polynomials would offer a better fit in the $\langle 100 \rangle$ crystal direction but no advantage in the other directions, hence the quadratics are entirely sufficient for this isotropic approximation. They track the phonon dispersion data closely, especially in the regions relevant to electron-phonon scattering in silicon: near the Brillouin zone center for long wavelength intravalley acoustic phonons, and near the frequencies corresponding to intervalley f- and g-type phonons. The quadratics are also easy to invert and, where needed, to extract the phonon wave vector as a function of frequency.

The same approach can be used to extend this phonon dispersion model to other materials or confined dimensions, and a similar example for graphene is given in Ref. 57. Changes in the phonon dispersion due to strain or confinement (e.g., in nanostructures) can be easily included. The challenge in this case lies chiefly in determining the correct modified phonon dispersion to use in such circumstances. The electron-phonon scattering

TABLE 1: Quadratic dispersion coefficients for each branch of the phonon spectrum, see Eq. (10) and Fig. 5(c). Longitudinal acoustic (LA), transverse acoustic (TA), longitudinal optical (LO), and transverse optical (TO). After Refs. 21 and 30.

Phonon Mode	ω_0 (10^{13} rad/s)	v_s (10^5 cm/s)	c (10^{-3} cm^2/s)
LA	0.00	9.01	–2.00
TA	0.00	5.23	–2.26
LO	9.88	0.00	–1.60
TO	10.20	–2.57	1.12

rates need to be numerically recomputed with the modified phonon description (as outlined below), which can be done efficiently if the dispersion is written as a set of analytic functions, like the polynomials in this work.

4.3 Electron-Phonon Scattering

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

$$M(\mathbf{k}, \mathbf{k}') = \langle \mathbf{k}' | U | \mathbf{k} \rangle \tag{11}$$

between the initial state $\mathbf{k}$ and the final state $\mathbf{k}'$.[49,58] This matrix element contains the momentum conservation condition, $\mathbf{k}' = \mathbf{k} \pm \mathbf{q} + \mathbf{G}$, where $\mathbf{q}$ is the phonon wave vector, $\mathbf{G}$ is a reciprocal lattice vector, and the upper and lower signs correspond to the absorption and emission of a phonon; also see Fig. 1(a) and 1(b). The electronic wave functions are typically taken to be Bloch functions that exhibit the periodicity of the lattice. The electron-phonon scattering rate is based on Fermi's golden rule, which is derived from first-order time-dependent perturbation theory[49,50] and gives the transition probability between the two eigenstates,

$$P(\mathbf{k}, \mathbf{k}') = \frac{2\pi}{\hbar} \left| M(\mathbf{k}, \mathbf{k}') \right|^2 \delta(E_{\mathrm{k}} - E_{\mathrm{k}'} \pm \hbar\omega_q) \tag{12}$$

where the upper and lower signs have the same meaning as above. It is assumed that the scattering potential is weak, such that it can be treated as a perturbation of the well-defined energy bands, and the δ-function ensures that two collisions do not "overlap" in space or in time, i.e., they are infrequent, or that the scattering time is much shorter than the time between collisions. The total scattering rate out of state $\mathbf{k}$ is obtained by integrating over all final states $\mathbf{k}'$ the electron can scatter into. Mathematically, this integration can be carried out over $\mathbf{k}'$ or $\mathbf{q}$ with the same result.[58] In those cases in which the matrix element

is independent of the phonon wave vector, the matrix element can be removed from the integral, which leaves a total scattering rate directly dependent on the density of states $g_d(E)$,

$$\Gamma(\mathbf{k}) = \frac{2\pi}{\hbar} |M(\mathbf{k})|^2 g_d(E_k \pm \hbar\omega_q) \tag{13}$$

where $M(\mathbf{k})$ includes the dependence on the phonon occupation of states, on the wave function overlap integral, and on the deformation potential characteristic of the particular phonon involved.[§] The dependence of the total scattering rate on the density of final states has a satisfying interpretation,[58] since it gives us a means for comparing scattering rates in 1-, 2-, or 3D systems. In three dimensions, the electron-phonon scattering rate increases roughly as the square root of the electron energy, just like the density of states (DOS),[¶]

$$g_d(E_k) = \frac{(2m_d)^{3/2}}{2\pi^2\hbar^3}\sqrt{E_k(1+\alpha E_k)}(1+2\alpha E_k) \tag{14}$$

written here in the non-parabolic, analytic band approximation [Eq. (8)] adopted in this work, where $m_d = (m_t^2 m_l)^{1/3}$ is the electron density of states effective mass.

4.3.1 Intravalley Scattering

Intravalley scattering refers to scattering within the same conduction band valley and it usually involves only acoustic phonons.[56] In this work, the total intravalley scattering rate is calculated separately with LA and TA phonons, as a function of the initial electron energy E_k,

$$\Gamma_i(E_k) = \frac{D_A^2 m_d}{4\pi\rho\hbar^2 k_s}\int_q \frac{1}{\omega_q}\left(N_q + \frac{1}{2} \mp \frac{1}{2}\right)\mathcal{I}_q^2 q^3 dq \tag{15}$$

where D_A is the respective deformation potential (D_{LA} or D_{TA}) and ρ is the mass density of silicon. The top and bottom signs refer to phonon absorption and emission, respectively. The electron wave vector is transformed to spherical Herring-Vogt[28,59] space as

$$k_s = \frac{\sqrt{2m_d E_k(1+\alpha E_k)}}{\hbar} \tag{16}$$

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

$$N_q = \frac{1}{\exp(\hbar\omega_q/k_B T) - 1} \tag{17}$$

without resorting to the equipartition or Joyce-Dixon approximations normally used.[28] The wave function overlap integral is included in the rigid ion approximation,[60]

§As will be shown below, deformation potentials are typically extracted empirically from comparison to electrical transport data.
¶This is the DOS per energy ellipsoid in silicon, including the factor of 2 for spin. Note this must be multiplied by a factor of 6 for all conduction band ellipsoids in silicon. See Figs. 1(c) and 5(b).

$$\mathcal{I}_q = \frac{3}{(qR_s)^3}\left[\sin(qR_s) - qR_s\cos(qR_s)\right] \tag{18}$$

where $R_s = a[3/(16\pi)]^{1/3}$ is the radius of the spherical Wigner-Seitz cell, R_s = 2.122 Å for silicon. All quantities are numerically evaluated using the corresponding phonon dispersion. The scattering rate integral in Eq. (15) is carried out over all phonon wave vectors **q** that conserve both energy ($E'_k = E_k \pm \hbar\omega_q$) and momentum (**k′** = **k** ± **q**). These arguments can be used to establish the range of q, as required by $|\cos(\phi)| \leq 1$, where

$$\cos(\phi) = \mp\frac{q}{2k_\mathrm{s}} + \frac{m_\mathrm{d}\omega_q}{\hbar q k_\mathrm{s}}\left[1 + \alpha(2E_\mathrm{k} \pm \hbar\omega_q)\right] \tag{19}$$

and ϕ is the angle between the phonon and the initial electron wave vector, see Fig. 1(b). As in the rest of this chapter, the top and bottom signs refer to phonon absorption and emission, respectively. The intravalley scattering rate typically used in the literature[28] can be recovered by substituting the simple, dispersionless phonon frequency $\omega_q = v_s q$ (typically for LA phonons only), $\mathcal{I}_q = 1$, and using an approximation for N_q, which allows Eq. (15) to be integrated analytically.

The final state of the electron after scattering $|E'_k, \mathbf{k}'\rangle$ reflects both the energy and momentum exchange with the phonon, as follows. First the magnitude of the phonon wave vector **q** is selected within the allowed range using a rejection algorithm[28] applied to the integrand in Eq. (15), which includes the overlap integral. Then the magnitude of the electron wave vector **k′** after scattering is found by energy conservation, while the angle between **k′** and **k** is obtained by momentum conservation. The final electron state is only accepted if it falls within the first Brillouin zone, otherwise the rejection algorithm is repeated.

The intravalley deformation potentials have a general angular dependence that can be written as[59] (and is plotted in Fig. 6)

$$\Xi_{LA} = \Xi_d + \Xi_u \cos^2\theta \tag{20}$$

$$\Xi_{TA} = \Xi_u \sin\theta\cos\theta \tag{21}$$

where θ is the angle between the phonon wave vector and the longitudinal axis of the conduction band valley, Ξ_u is the shear, and Ξ_d is the dilation deformation potential. Detailed calculations have shown that the influence of this angular dependence on the electron transport is relatively small.[61] Hence, the intravalley deformation potentials can be averaged over the angle θ, consistently with the general isotropic approach adopted in this work. The isotropically averaged deformation potentials become

$$D_{LA} = \sqrt{\langle\Xi_{LA}^2\rangle|_\theta} = \left[\frac{\pi}{2}\left(\Xi_\mathrm{d}^2 + \Xi_d\Xi_u + \frac{3}{8}\Xi_u^2\right)\right]^{1/2} \tag{22}$$

$$D_{TA} = \sqrt{\langle\Xi_{TA}^2\rangle|_\theta} = \frac{\sqrt{\pi}}{4}\Xi_u \tag{23}$$

which are used for computing the intravalley scattering rates in Eq. (15). There is considerable variation in the values of the shear (Ξ_u) and dilation (Ξ_d) deformation potentials

reported in the literature over the years. A good summary of these values can be found in Ref. 62: various theoretical and empirical studies have estimated Ξ_u in the range of 7.3–10.5 eV, while Ξ_d has been previously cited both as –11.7 eV (Ref. 63) and near 1.1 eV (Ref. 62). Although, perhaps surprisingly, both values can be used to describe electron mobility (hence the original confusion over the correct choice), it was shown that only the latter (Ξ_d = 1.1 eV) yields the correct mobilities for both electrons and holes.[62] This is the value adopted in the current study. Then Ξ_u is used as a fitting parameter while calculating the low-field, low-temperature (T = 77 K) electron mobility, a regime dominated by scattering with intravalley phonons. An empirical best-fit value of Ξ_u = 6.8 eV is found, in reasonable agreement with previous work. With these values of Ξ_d and Ξ_u, the isotropically averaged deformation potentials are D_{LA} = 6.39 eV and D_{TA} = 3.01 eV. These are comparable to the value of 9 eV typically cited in the literature for MC models where scattering is only taken into account with the LA modes.[28]

4.3.2 Intervalley Scattering

As outlined in Section 4.2 and in Fig. 1(c), intervalley scattering in silicon can take electrons between equivalent (g-type) and nonequivalent (f-type) valleys. Based on geometrical arguments,[54] both f- and g-type scattering are Umklapp processes, involving a reciprocal lattice vector $|\mathbf{G}| = 2\pi/a$. Since the X-valley minima are located at 0.85 from the center to the edge of the Brillouin zone, the change required in electron momentum is (0, 0.85, 0.85)G for f-type scattering and (1.7, 0, 0)G for g-type scattering. Reduced to the first Brillouin zone, the phonons involved are (1, 0.15, 0.15)G and (0.3, 0, 0)G, respectively.[54,64] The f-phonon is just 11 deg off the $\langle 100\rangle$ direction, while the g-phonon is along $\langle 100\rangle$, at 0.3G. These phonons are schematically drawn on the dispersion relation in Fig. 5(c). The g-phonon frequencies can be directly read off the $\langle 100\rangle$ dispersion, while the f-phonons are typically assumed to be those at the edge of the Brillouin zone. In this work, ω_q is computed from the analytic phonon dispersion, and the intervalley scattering rate between the initial (i) and final (f) valley can be written as[28,49]

$$\Gamma_{if}(E_k) = \frac{\pi \Delta_{if}^2 Z_f}{2\rho\omega_q}\left(N_q + \frac{1}{2} \mp \frac{1}{2}\right) g_{df}\left(E_k \pm \hbar\omega_q\right) \tag{24}$$

where Z_f is the number of available final valleys (four for f-type and 1 for g-type scattering), $g_{df}(E_k)$ is the density of states in the final valley [Eq. (14)], and other symbols are the same as previously defined. Intervalley scattering can also include an overlap factor, but its value is typically incorporated into the scattering constant Δ_{if}. The six phonons involved in intervalley scattering, along with their approximate energies, equivalent temperatures (as $T = \hbar\omega_q/k_B$), and deformation potential scattering constants, are listed in Table 2.

Traditional MC models (apart from the ab initio approaches of Refs. 43 and 44) assume the phonon energies involved in intervalley scattering are fixed at the values determined by transitions between the X-valley minima. Also, the state of the electron in the final valley is computed isotropically.[28] These geometrical arguments only hold strictly for the lowest energy electrons at the bottom of the bands. This work takes into account the phonon

TABLE 2: Summary of phonon energies and deformation potentials Δ_{if} for intervalley electron-phonon scattering in silicon

Phonon Type	*E* (meV)	*T* (K)	Δ_{if} old model[a]	Δ_{if} new model[b]
			($\times$ 10^8 eV/cm)	
f-TA	19	220	0.3	0.5
f-LA/LO	51	550	2	3.5†
f-TO	57	685	2	1.5
g-TA	10	140	0.5	0.3
g-LA	19	215	0.8	1.5[c]
g-LO	63	720	11	6†

[a]Old model refers to the work of Jacoboni and Reggiani,[28] which was calibrated against bulk silicon mobility data.
[b]New model refers to the work of Pop et al.[21,30] which was calibrated against bulk *and* strained silicon transport data.
[c]Values marked with a dagger are also consistent with recent ab initio calculations.[44,61]

dispersion for scattering with both optical and acoustic phonons when calculating the final state of the electron. After the type of intervalley scattering mechanism is determined, the state of the electron in the final valley is first chosen isotropically, as in the traditional approach. The phonon wave vector necessary for this transition can be calculated as $\mathbf{q} = \mathbf{k}' - \mathbf{k}$ because the initial state of the electron is known. The phonon is then reduced to the first Brillouin zone and its energy is obtained using the phonon dispersion described earlier. This procedure is applied to both acoustic and optical phonons. The phonons that do not satisfy both energy and momentum conservation within a certain tolerance are discarded with a rejection algorithm. This is a relatively inexpensive search that ends when a suitable phonon is found. The effect of this algorithm is to smear out any "hard" thresholds associated with intervalley phonon energies in the electron distribution, as was shown in Ref. 21. The present model removes such unphysical thresholds in a computationally inexpensive way, while satisfying energy and momentum conservation for all scattering events.

Despite the added complexity of the full phonon dispersion, the analytic band code is more than an order of magnitude faster when compared to typical full-band programs (using a simpler phonon description) doing the same velocity-field curve calculations.[21] This work also incorporates the phonon dispersion in an efficient way, giving significantly more physical insight than the typical analytic band code for very little computational overhead, while still being more than an order of magnitude faster than a typical full-band code. The analytic phonon dispersion and analytic electron bands significantly speed up the calculations of the final electron state after scattering, compared to the look-up tables and interpolation schemes found in full-band codes. Further speed improvements could be obtained by including an energy-dependent total scattering rate,[63] which would significantly reduce the number of self-scattering events.

4.3.3 Electron-Ionized Impurity Scattering

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

$$U(r) = \frac{Zq^2}{4\pi\varepsilon_s r} \exp\left(-\frac{r}{L_D}\right) \tag{25}$$

depending on how many free charge carriers are present. Here Zq is the net extra charge on the impurity atom (for example, $Z = 1$ for n-type dopants from group V, like As or P), ε_s is the dielectric constant of the semiconductor, r is the distance from the scattering center, and $L_D = [\varepsilon_s k_B T/(q^2 n)]^{1/2}$ is the Debye length, where n is the free charge carrier (e.g., electron) density responsible for screening the potential in Eq. (25). Impurity scattering is a highly anisotropic process,[49,67] showing a strong preference for small scattering angles. Although physically sound, a direct implementation of this approach in a Monte Carlo technique would yield several problems. Many small-angle scattering events would have to be processed, consuming computational time. Also, many short free-flight times would be obtained, further degrading the efficiency of the procedure. The scattering model proposed by Kosina[68,69] avoids such pitfalls by reformulating impurity scattering as an isotropic process with the same momentum relaxation time as the anisotropic process. This work implements Kosina's model, including the screening function from Ref. 69. The model has been shown to be adequate for doping concentrations up to 10^{20} cm^{-3}, with particularly notable improvements in efficiency at lower (less than 10^{17} cm^{-3}) doping levels. The dashed line and solid symbols in Fig. 7(a) show a comparison between velocity-field data obtained in 10^{17} cm^{-3} doped bulk silicon and Monte Carlo simulations using the isotropic scattering model. Good agreement is found over a wide range of electric fields. Similarly, the low-field mobility was computed over a wide range of doping densities and good agreement was found with available experimental data. Figure 7(b) shows the results of transport simulations using the updated set of deformation potentials listed in Table 2. Note the wide range of electric fields and temperatures (from 30 to 600 K) covered by the simulations and their comparison with the transport data in Fig. 7, including that for strained Si.

5. APPLICATIONS TO TRANSPORT

5.1 One-Dimensional Device Applications

In the ensemble Monte Carlo method for device simulation, several things must be taken into account in addition to the ensemble Monte Carlo method for bulk semiconductors (described in the previous sections). One is that the motion of the particles is spatially restricted to the device domain, hence suitable boundary conditions must be set up. Another is that the impurity concentration, and hence the impurity scattering rate, is dependent

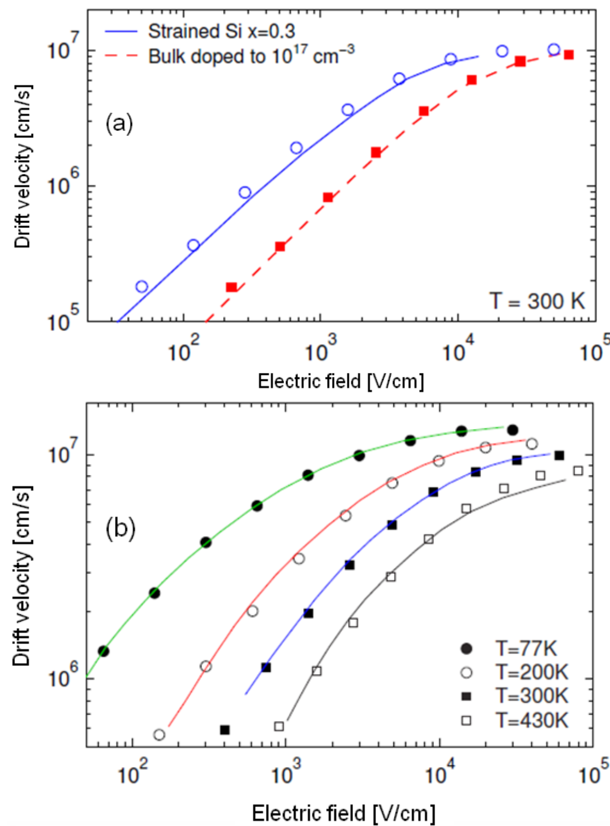

FIG. 7: (a) Electron velocity field relationship in doped bulk and strained silicon. The dashed lines represent data for 1017 cm^{-3} doped bulk silicon, the solid lines are data for strained silicon on $x = 0.3$ substrate Ge fraction.[65] The symbols are our simulation results for the two respective cases. Reprinted with permission from Ref. 30. Copyright 2005, AIP Publishing LLC. (b) Electron drift velocity versus electric field in unstrained bulk silicon over a wide range of temperatures. Symbols are the Monte Carlo simulations of this work. The lines represent the time of flight experimental data of Canali et al.[36] All data sets for bulk and strained silicon are fitted with the new set of new deformation potentials listed in Table 2.

on position, i.e., on the doping profile. Finally, the electric fields must be updated self-consistently with the motion of the charged particles, through repeated solutions of the Poisson equation (at every time step) with appropriate boundary conditions, which are consistent with the boundary conditions applied to the carrier dynamics.

The most frequently studied, realistic, 1D device in the Monte Carlo and device transport community is the n$^+$nn$^+$ (or n$^+$in$^+$) structure, sometimes referred to as a “ballistic” diode.[70,71] The energy band diagram of the ballistic diode is such that it represents a simple model for a cross section along the channel of a metal-oxide semiconductor field effect transistor (MOSFET), as shown in Fig. 8. The n$^+$nn$^+$ band diagram has similar features,

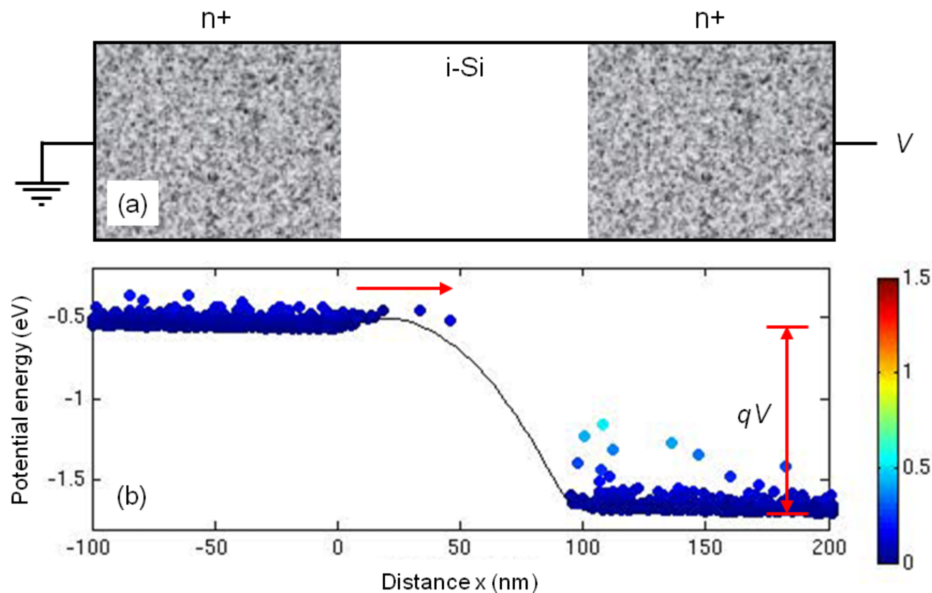

FIG. 8: Physical structure of n^+in^+ ballistic diode. (b) Energy band diagram that approximates that along the channel of a MOSFET. The colored balls represent the super-particles during MC simulation (here electrons), and their color marks the energy (measured in electron volts). See Fig. 4 for a block diagram of various key steps during the simulation.

like the voltage-controlled injection barrier at the beginning of the "channel," followed by a steep drop in potential (i.e., highly peaked lateral electric field). Charge transport may be quasi-ballistic across the channel region, provided it is short enough compared to the electron mean free path. This device structure is also ideal as a test bed for the comparison of various simulation approaches (e.g., drift-diffusion, energy balance, or Monte Carlo) since it incorporates impurity scattering, charge transport (with likely velocity overshoot), and realistic boundary conditions. On the other hand, transport in a ballistic diode is not complicated by 2D potential or quantum confinement effects (both present in the channel of a MOSFET), which allows for the other transport features mentioned above to be better isolated and understood. The program code described here has been implemented to simulate any electron device, but focus in this section will be given to the ballistic diode because of its relevance to a variety of transport problems. The code was named Monte Carlo electron transport (MONET),[51] and it is occasionally referred to as such in the remainder of this chapter.

5.2 Self-Consistent Poisson Equation

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

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

As mentioned in Section 3, the ensemble Monte Carlo method models the entire mobile charge inside the semiconductor device with a few thousand (e.g., 10,000 to 20,000) particles. These "super-particles" are treated as individual charge carriers while they drift, but as clouds of charge when the simulation is stopped and the Poisson equation is solved. The amount of charge then assigned to each super-particle is given by (from Section 3) $Q = qN/N_{\text{sim}}$, where q is the elementary charge, N is the total number of mobile charges expected in the real device, and N_{sim} is the number of super-particles used in the simulation. Charge assignment on the device grid is done with the cloud-in-cell method, with

$$w_G = \frac{X_{G+1} - x}{X_{G+1} - X_G} \tag{26}$$

$$w_{G+1} = 1 - w_G \tag{27}$$

which are weights used in a simple linear interpolation of the charge Q at position x, onto grid nodes at locations X_G and X_{G+1} (where $X_G < x < X_{G+1}$), as shown in Fig. 9. The charge assigned to the grid nodes is then given by Qw_G for grid node G, and Qw_{G+1} for grid node $G + 1$.

In order to self-consistently update the electric field as the mobile charge moves during the simulation, Poisson's equation must be solved at every time step Δt. In other words, the mobile charge is allowed to drift under the influence of the electric fields for Δts (an upper limit on this time step being given by the plasma oscillation period, as explained in Section 3), then the simulation is stopped, the mobile charge is assigned to the grid nodes, and the Poisson equation is solved in order to update the electric fields. The Poisson equation may be written as

$$\nabla^2 \Phi(x) = -\frac{\rho_c(x)}{\varepsilon_s} = -\frac{q}{\varepsilon_s}\left[p(x) - n(x) + N_D(x) - N_A(x)\right] \tag{28}$$

where Φ is the voltage potential, ρ_c is the net charge density, and ε_s is the dielectric constant of the semiconductor. The mobile charge densities (after charge assignment with the cloud-in-cell method) are given by n and p for electrons and holes, while the fixed charge is determined by N_D and N_A, the donor and acceptor doping profiles. In the simulation of an n^+nn^+ ballistic diode, the acceptor and hole densities are zero. The Poisson equation can be discretized in general as

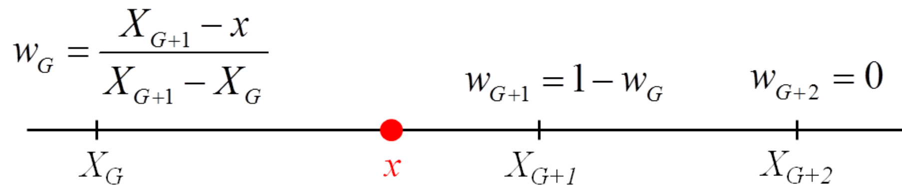

FIG. 9: Charge density assignment using the cloud-in-cell method.

$$-\frac{1}{h_i}\Phi_{i-1} + \left[\frac{1}{h_i} + \frac{1}{h_{i+1}}\right]\Phi_i - \frac{1}{h_{i+1}}\Phi_{i+1} = \frac{q}{2\varepsilon_s}\left(N_{D,i} - n_i\right)\left(h_i + h_{i+1}\right) \tag{29}$$

where $h_i = x_i - x_{i-1}$ and $h_{i+1} = x_{i+1} - x_i$ and these differences become simply Δx on a uniform grid. The discretized Poisson equation can then be written as a set of linear algebraic equations that can be easily solved through conventional means, e.g., tridiagonal elimination.[50,74] Once the potential is found, the electric field is written as its negative derivative through centered differencing,[75]

$$F_i = -\frac{d\Phi}{dx} \simeq \left[\frac{h_{i+1}/h_i}{h_i + h_{i+1}}\right]_i \Phi_{i-1} + \left[\frac{1}{h_{i+1}} - \frac{1}{h_i}\right]\Phi_i - \left[\frac{h_i/h_{i+1}}{h_i + h_{i+1}}\right]\Phi_{i+1} \tag{30}$$

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

$$F_i \simeq -\frac{\Phi_{i+1} - \Phi_{i-1}}{2\Delta x} \tag{31}$$

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

$$F_1 \simeq -\frac{3\Phi_1 + 4\Phi_2 - \Phi_3}{X_3 - X_1} \tag{32}$$

$$F_n \simeq -\frac{3\Phi_n - 4\Phi_{n-1} + \Phi_{n-2}}{X_n - X_{n-2}} \tag{33}$$

where the denominator, in both cases, is equal to $2\Delta x$ for a uniformly spaced grid. After the electric field is found, the simulation resumes and particles are allowed to drift under the influence of the new field distribution for another Δt seconds, after which this process repeats [see Fig. 4 and Eq. (7)].

5.3 Contact Boundary Conditions

In the case of 1D simulation, only two boundaries are present, which are the contacts where the voltage is applied. In general, these contacts are unions of mesh nodes where the device domain touches an ideal source/sink of carriers. In most Monte Carlo simulations, these boundaries are treated as ideal ohmic contacts, absorbing all incident electrons that actually reach them, and emitting (as necessary, and explained further below) only electrons in thermal equilibrium with the contact temperature.[71] The boundary conditions for particle transport must be consistent with those for the electric field and potential. There are two ways that are usually employed to treat the particle flux at boundaries within Monte Carlo simulation. They have both been implemented within MONET, the code developed during this dissertation, and one or the other can be selected when the code is compiled. The

simplest way to model the two contacts is to assume periodic boundary conditions, that is, particles that escape from one contact are reinjected at the other with thermal energy, and with a momentum component weighed toward the inside of the device as[49]

$$p_x = \sqrt{-2m_x k_B T \ln(r)} \tag{34}$$

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

$$I = \frac{1}{t_{\text{sim}}} Q \left(N_{\text{right}} - N_{\text{left}}\right) \tag{35}$$

where Q is the super-particle charge [Eq. (6)], t_{sim} is the simulation time, and the term in parenthesis is the difference between the number of particles that exit through the right versus the left contacts. The instantaneous current (e.g., during transients) can be similarly computed by counting particles exiting through the contacts during shorter periods of time, e.g., only a few time steps Δt [also see Eq. (7) and Section 3.2].

Another method for treating device boundaries is more frequently employed because it can be extended to devices with an arbitrary number of contacts. It involves maintaining local charge neutrality at the grid nodes adjacent to the contact, which is done as follows. At the beginning of the simulation, a target super-particle density is calculated at each contact, as consistent with local charge neutrality. During the simulation, the particles exiting through the contacts are deleted and tallied as current. Within the Monte Carlo code MONET, this is done by copying the information of the last particle in the array where particles are stored on top of the ith particle to be deleted, then shrinking the array size by one. After each time step Δt, just before the Poisson equation is solved, the program examines the super-particle count at each contact node and determines how many particles should be injected or deleted to reach the charge-neutral target initially determined. The injected particles are assumed to have thermal equilibrium energy, and a momentum component forward weighed into the device, as previously described [Eq. (34)]. This velocity weighing is essential, since it accounts for the higher probability of a "fast" particle entering the device from the conceptual thermal carrier gas considered touching the contact. Every particle injected or deleted is also tallied as current. Note that with this second method for modeling device contacts, the number of super-particles present in the device at any time during the simulation is not constant. This is also the method preferred for Monte Carlo noise simulations.[52,71]

5.4 One-Dimensional Device Simulation Results

To illustrate 1D device applications of the Monte Carlo code MONET, an n^+nn^+ ballistic diode was simulated. The results are shown in Fig. 10 for the potential, electric field, average electron velocity, and density (solid lines), and they are compared with the results of

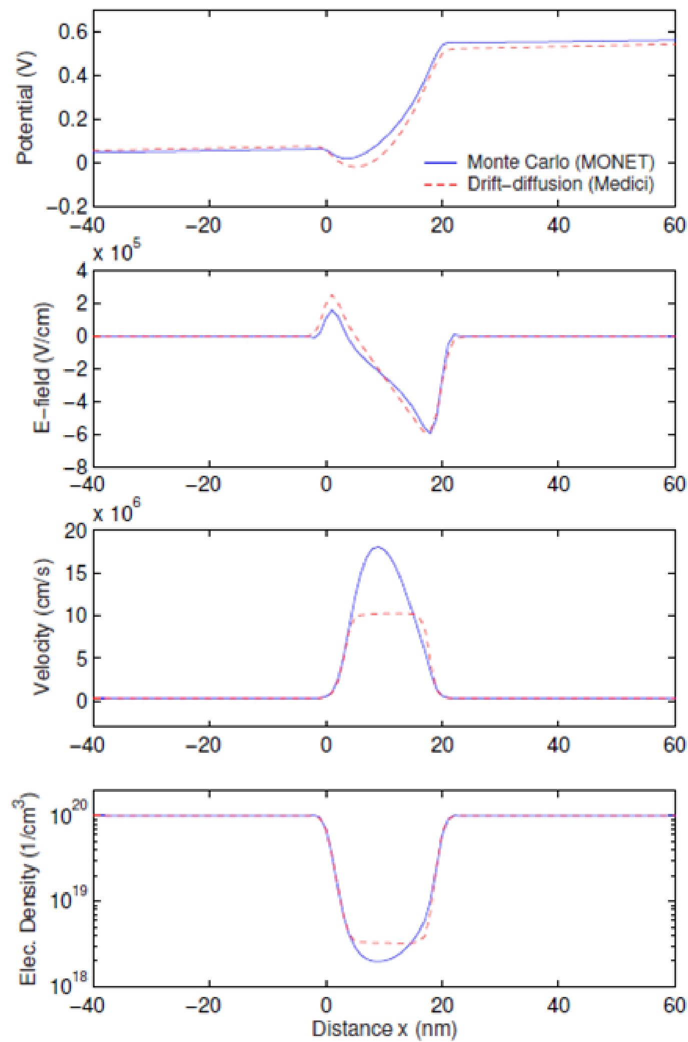

FIG. 10: One-dimensional (1D) device simulation with the Monte Carlo program described here (MONET, solid lines), compared to a commercial drift-diffusion device simulator (Medici, dashed lines). The middle "n" region is 20 nm long and the applied voltage is 0.6 V. Note that the Monte Carlo simulation indicates significant velocity overshoot, which is not captured by the drift-diffusion simulator.

the commercial drift-diffusion code Medici (dashed lines). The n^+nn^+ diode has a "channel" length of 20 nm and source and drain lengths of 100 nm (although only 40 nm of each are shown in the plots). The source/drain doping is 10^{20} cm^{-3} and the channel doping is 10^{16} cm^{-3}. The applied voltage for the simulations in the figure was 0.6 V. The 1D device structure was first "built" and simulated with the commercial code Medici, with a uniform grid spacing. The resulting grid, charge, potential, and electric field distributions were then saved and imported into MONET, where they served as the initial conditions.

The Poisson equation was self-consistently solved along with the Monte Carlo transport of charge.

Several similarities and differences can be pointed out between the drift-diffusion code and the Monte Carlo results. As can be seen from the plots, the potential and electric field distributions are very similar. The Monte Carlo code, however, predicts significant velocity overshoot in the short "channel" region, whereas the average velocity predicted with the drift-diffusion model plateaus at 10^7 cm/s, the saturation velocity in silicon. Moreover, the influence of the heavily doped drain region (which injects cool, slow electrons) is clearly seen in the velocity distribution computed by the Monte Carlo method, which is slightly skewed toward the source side. It is also clear that the average electron velocity is not at all a local function of the electric field. The differences in the particle density distributions are consistent with the differences in the average velocity between the two computational methods, since the net current density (proportional to $n \times v$) is the same, and constant through the 1D profile, as required by current continuity. This example shows the applicability of such a Monte Carlo simulator to 1D transport problems, including self-consistently computed electric field distribution, spatially varying doping profile, and realistic device contacts.

5.5 Two-Dimensional (2D) Device Simulation Results

As an example of a 2D device application, we focus on a silicon-on-insulator (SOI) MOSFET[17] with 18 nm gate length, as in Fig. 11. The 2D grid (including electric fields, doping, and device boundaries) was imported from a previous drift-diffusion simulator run (e.g., Medici). The Monte Carlo particle motion was computed on the "frozen" electric field grid imported at the beginning of the simulation. This is the so-called non-self-consistent approximation, which has limited applications, and has been shown[52] (as it might be expected) to not yield significant improvements in accuracy over the drift-diffusion approach. However, the results of such simulations can yield significant physical insight, as shown here.

Figure 11 illustrates the three-step process by which MONET can be used to perform such simulations. The mesh (top subplot) and electric field distribution (middle subplot) are imported from a drift-diffusion simulation with Medici, with voltages applied as necessary. MONET initially distributes particles in proportion with the charge density (not the doping density) imported from Medici. These super-particles are first assigned thermally distributed energies and randomly oriented momenta. Then, the particles are allowed to drift under the influence of the electric field grid, but the electric fields are not updated as the charge moves around. Boundary conditions at the source and drain electrodes are similar to those described in the previous section. Scattering with the other surfaces (e.g., between Si and SiO_2) reflects the particles back into the simulation domain, with unchanged energy, but newly oriented momenta. This scattering can be either specular (the reflection angle is the same as the incident angle) or diffuse (randomly chosen reflection angle). A specularity parameter is used to choose between the two types of surface scattering, and the ratio of diffuse to specular scattering is set at 0.15.[76] The bottom subplot in Fig. 11 shows a snapshot of such a Monte Carlo simulation with only a few hundred super-particles shown, for

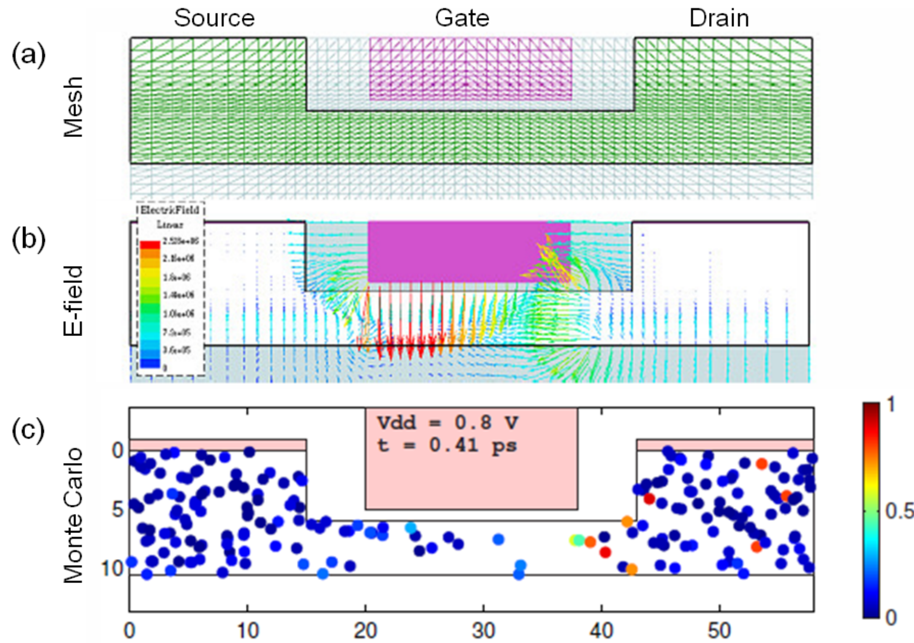

FIG. 11: (a) Simulation mesh, (b) electric fields, and (c) Monte Carlo simulation snapshot of an 18 nm gate length thin-body SOI device, with 0.8 V applied to the drain and gate. The mesh and electric field distribution are imported from a commercial drift-diffusion simulator (Medici). The Monte Carlo simulation only shows a few hundred super-particles, for clarity. The color bar is the electron energy scale (measured in electron volts), the physical axes are in nm.

clarity. The device being simulated is an 18 nm gate length thin-body SOI with 10^{20} cm^{-3} doped source and drain, undoped body, and molybdenum gate. The body thickness is 4.5 nm. The on/off current ratio predicted by Medici for this layout is 1000:1. Qualitatively, some important observations can be made based on this simulation. For example, we note the presence of hot electrons almost entirely in the drain of the device. This indicates that (i) transport across the short channel is nearly ballistic, and that (ii) energy relaxation of the carriers, and therefore Joule heating of the lattice, happens entirely in the drain region of the device. This point will be discussed in more detail in Section 6, and the exact location of the heat generation region will be analyzed with electrostatically self-consistent simulations.

6. APPLICATIONS TO DEVICE POWER DISSIPATION

One of the unique applications of the Monte Carlo approach described in this work is for heat generation simulations within functioning silicon transistors. The simulations here are particularly well suited for this task, since they incorporate realistic phonon dispersion and all electron-phonon scattering events are (correctly) taken to be inelastic, meaning that energy is exchanged. These have not always been possible within Monte Carlo simulations,

which typically simplify the phonon dispersion, and treat acoustic phonon scattering as elastic. The phonon dispersion is also used when computing the final electron state after scattering, taking into account both momentum and energy conservation. This approach allows a range of phonon wave vectors and energies *around* the six typical f- and g-type phonons to participate in scattering. This is an innovative, efficient, and physically realistic approach introduced for the first time in Refs. 21, 30, and 51. During the simulation, all phonons absorbed and emitted are tallied, and full phonon generation statistics can be computed. The total heat generation rate can be obtained from the sum of all phonon emission events minus all phonon absorption events per unit time and unit volume, as briefly discussed in Section 2.3.

6.1 Heat Generation in Bulk and Strained Silicon

In this section, we examine the details of net phonon generation as a function of phonon frequency, in order to find out exactly which branches (modes) of the phonon dispersion are excited when current flows in a constant electric field. Figure 12 shows the computed phonon generation spectrum in 10^{17} cm^{-3} doped bulk and strained silicon with both a lower (5 kV/cm) and higher (50 kV/cm) applied electric field. These electric field values were chosen from two regions of Fig. 7(a) such that the mobility enhancement in strained silicon is maintained at the lower field value, but not at the higher field. To facilitate comparison, Fig. 12(b)–12(e) subplots are drawn such that the vertical axes with energy units in 10^{-3} eV match the vertical frequency axis of the phonon dispersion in subplot 12(a), with units in rad/s, as given by $E = \hbar\omega$. Note the cutoff energies of the various emitted phonon

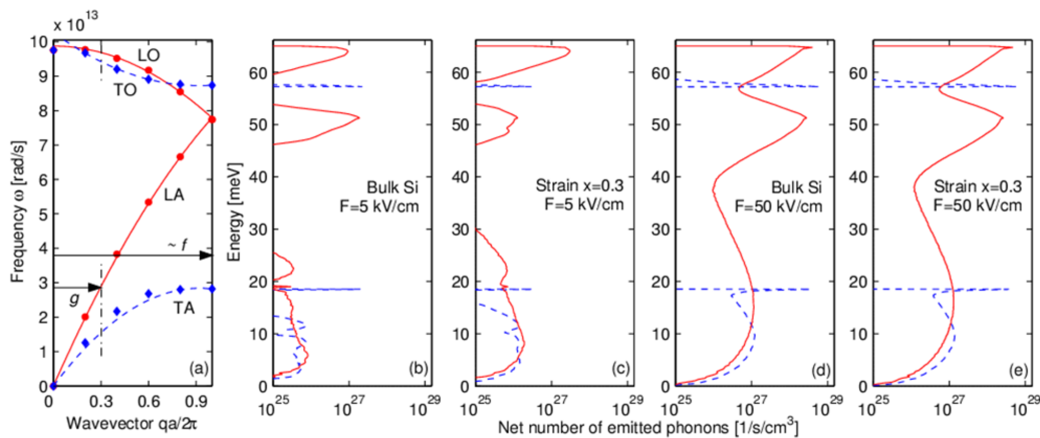

FIG. 12: Phonon dispersion in silicon (a) and computed net phonon generation rates (emission minus absorption) with low field (b,c) and high field (d,e) in strained and bulk silicon doped to 10^{17} cm^{-3}, at T = 300 K. Subplot (a) shows the dispersion data of Ref. 53 (symbols), our quadratic approximation (lines),[21] and the vector magnitude of f- and g-type intervalley phonons. Dashed lines represent transverse, while solid lines represent longitudinal phonons throughout. Reprinted with permission from Ref. 30. Copyright 2005, AIP Publishing LLC.

populations as required by their respective dispersion relation. Few acoustic phonons are generated through intravalley scattering at low energies because the 3D phonon density of states $g_p(\omega)$ vanishes near the Brillouin zone (BZ) center, where the phonon wave vector $q \to 0$, as[77]

$$g_p(\omega) = \frac{\partial N_s}{\partial \omega} = \frac{q^2}{2\pi^2}\left(\frac{dq}{d\omega}\right) \tag{36}$$

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

The sharp peaks in the phonon generation plots occur due to intervalley scattering with the three g-type (TA, LA, and LO, at 0.3 of the distance to the edge of the BZ) and three f-type (TA, LA/LO, and TO, at the edge of the BZ) phonons, see Table 2. The momenta and hence the location within the BZ of these six intervalley phonons are given by scattering selection rules.[54] The relative magnitude of their generation rates depends on the choice of scattering deformation potentials Δ_{if}, which have been carefully calibrated in Section 4.3.2 and Ref. 21. The deformation potential values determined here are the only ones currently available in the literature that reproduce the experimental mobility data for both bulk *and* strained silicon. Figures 12(b) and 12(c) highlight the difference in the phonon emission spectrum between strained and bulk silicon at low electric fields. The strain-induced band splitting suppresses f-type phonon emission between the two lower and four upper valleys.[51] However, since most conduction electrons in strained silicon are confined to the two lower valleys (of lighter mass m_t), they quickly gain energy and g-type emission between the lower valleys is enhanced. Comparing Figs. 12(d) and 12(e), it can be noted that phonon generation in strained and bulk silicon at high field is essentially identical, when electrons have enough energy to emit across the entire phonon spectrum despite the strain-induced band splitting. This is consistent with the observation of similar saturation velocity in strained and bulk silicon [Fig. 7(a)].

6.2 Heat Generation in Quasi-Ballistic Devices

This section examines heat (phonon) generation in silicon devices as they transition from the diffusive conduction regime (size $L \gg$ electron mean free path λ) to the quasi-ballistic transport regime (L comparable with λ). Three n^+nn^+ devices are considered, with channel lengths of 500, 100, and 20 nm (also see Fig. 8). The source and drain regions are assumed doped to 10^{18}, 10^{19}, and 10^{20} cm^{-3}, and the applied voltages are 2.5, 1.2, and 0.6 V, respectively. The latter are roughly equivalent to the operating voltages recommended by the International Technology Roadmap for Semiconductors guidelines[78] for complementary metal oxide semiconductor devices of similar channel lengths. The middle (channel) region is assumed doped to 10^{16} cm^{-3} throughout. Monte Carlo simulations of heat generation using the approach described here are compared to heat generation rates computed using the commercial drift-diffusion simulator Medici, with the $\mathbf{J} \cdot \mathbf{F}$ approach of Eq. (2).

In general, Monte Carlo simulation results are expected to be similar to those of the drift-diffusion calculations for "long" devices ($L \gg \lambda$), i.e., in the continuum approximation. This limit provides a check on the accuracy of the Monte Carlo simulation, and enables a study of the conditions under which the drift-diffusion heat generation calculations break down. The Monte Carlo results are expected to differ from (and be more physically accurate than) the drift-diffusion results in the limit of short channel lengths ($L \sim \lambda$), where velocity overshoot and other nonequilibrium transport effects are expected to dominate. This is the limit under which the "granularity" of charge transport and phonon emission becomes important, and the continuum approximation of the drift-diffusion method breaks down.

Figure 13 displays heat generation rates computed along the three n^+nn^+ devices of varying channel lengths. Both the drift-diffusion (Medici) and Monte Carlo (MONET) simulations are solved self-consistently with the Poisson equation, as described in Section 5.2. As expected, the two approaches give very similar results for the longest simulated device, with channel length (500 nm) much greater than the average electron-phonon scattering length (5–10 nm). This is essentially still in the continuum limit, and the drift-diffusion simulation approach is adequate. However, for the two shorter (100 and 20 nm) devices, the heat generation rates computed by the Monte Carlo approach are seen to differ significantly from the drift-diffusion results. The peak of the Monte Carlo heat generation is "displaced" from the peak of the drift-diffusion heat generation. This outcome is qualitatively expected, and an explanation for it was already suggested in Section 2.1: electrons gain most of their energy at the location of the peak electric field, yet they travel several mean free paths until they release this energy back to the lattice. Note that since the transport is 1D, the current density $J = qnv$ is constant along the length of the device, and the heat generation rate computed by the drift-diffusion ($\mathbf{J} \cdot \mathbf{F}$) reaches its peak at the location of the electric field maximum. By comparison to the channel length L, the "nonlocal" errors incurred by using the drift-diffusion versus the Monte Carlo approach when finding the location of the peak heat generation rate are $\Delta L/L = 0.10$, 0.38, and 0.82 for the three device lengths $L = 500$, 100, and 20 nm.

Another observation can be made about the "shape" of the heat generation in the drain region of the device, downstream from the E-field. Because, in reality, electrons can only release energy in discrete packets (phonons) of at most 50–60 meV (the energy range of the optical phonons in silicon), the heat generation region computed by the (physically correct) Monte Carlo approach spreads deep into the device drain, as electrons drift toward the contact. This situation is particularly noticeable for the shortest device (20 nm), where transport in the channel is nearly ballistic, and almost the *entire* heat generation occurs in the drain. Note that the Monte Carlo method also computes the integrated optical and acoustic phonon generation rates, with dotted lines in Fig. 13. It can be seen that about twice as much energy is deposited in the optical (LO and TO) compared to the acoustic (LA and TA) modes, along the length of the simulated quasi-ballistic devices. This is consistent with (and an integral of) the spectral distribution of net generated phonons in Fig. 12, for Joule heating in silicon.

Before concluding, we explore the heat generation in the 20 nm device in more detail in Fig. 14. Several voltages are considered, from 0.2 to 1.0 V, for the self-consistent

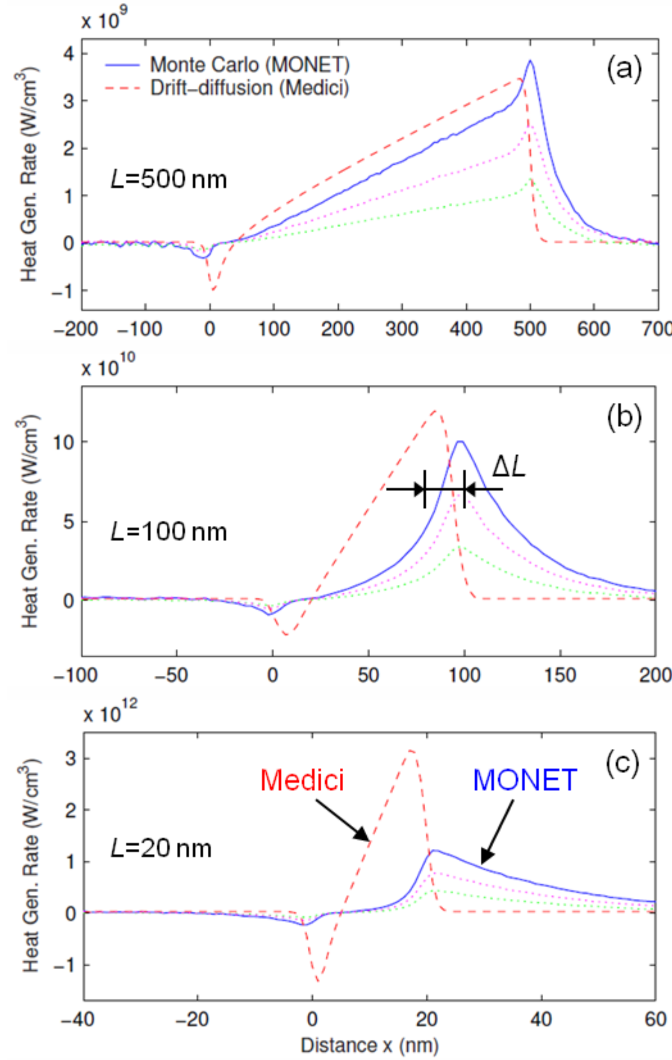

FIG. 13: Heat generation along three n+nn+ devices with middle ("channel") regions of length: (a) 500 nm, (b) 100 nm, and (c) 20 nm. Applied voltages are 2.5, 1.2, and 0.6 V, respectively. Solid lines are Monte Carlo results with MONET,[51] dashed lines are drift-diffusion calculations using the commercial simulator Medici. The dotted lines represent the optical (upper) and acoustic (lower) phonon heat generation rates, as computed by MONET.

Monte Carlo analysis. It can be easily seen that the maximum heat generation rate scales linearly with the potential drop across the channel, hence essentially with the applied voltage. The maximum average electron energy in Fig. 14(b) also scales linearly with the applied voltage V, approximately as $q \times 0.4V$, where q is the elementary charge. However, the characteristic (exponential) decay length of the heat generation region in the

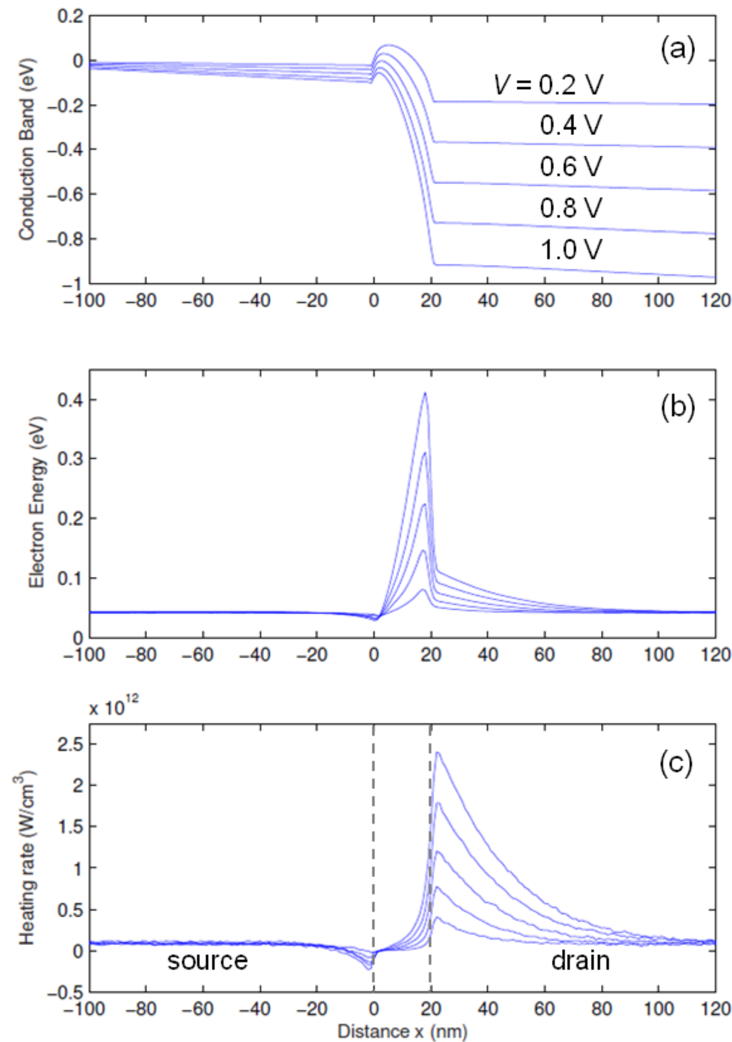

FIG. 14: Monte Carlo simulations of a quasi-ballistic device (channel length L = 20 nm) with applied V = 0.2, 0.4, 0.6, 0.8, and 1.0 V. The source and drain n^+ regions are doped 10^{20} cm^{-3}, the middle region is 10^{16} cm^{-3}. The edges of the channel are at 0 and 20 nm. (a) Conduction band, (b) average electron energy, and (c) net heat generation rate (increasing with V from bottom to top). Note that the heat generation is almost entirely "displaced" into the drain of the device.

drain is always approximately Λ_h = 20 nm, regardless of the applied voltage. This can be qualitatively understood because electrons lose $\hbar\omega$ (a phonon of) energy approximately every $v_e\tau_o$, the inelastic scattering length. Neglecting non-parabolicity, the electron velocity v_e scales as the square root of energy, while the inelastic (phonon) scattering time τ_o scales as $1/\sqrt{E}$ because the phonon scattering rate ($1/\tau_o$) scales with $\sqrt{E}$ from the density

of states [Eq. (14)]. Therefore, the inelastic scattering length is relatively independent of the electron energy and of the applied voltage.

The extent of the heat generation region in the drain can be understood in more detail as follows. The electrons present in the drain are a heterogeneous mixture of two populations, one being the "hot" electrons injected across the channel, and another made up of the many "cold" electrons already present there due to the high doping. The cold electrons have an average energy of $3k_BT/2$ (the thermal average) and they do not contribute to any net heat generation. Hence, the heat generation in Fig. 13(c) is entirely caused by the hot electrons injected quasi ballistically across the channel. While crossing the channel, these electrons acquire an amount of energy that is a significant fraction of the applied voltage, qV. This energy is then released, in discrete amounts of $\hbar\omega$ (the phonon energy) to the lattice in the drain. Assuming an average inelastic scattering time τ_o = 0.05–0.1 ps (based on the Monte Carlo scattering rates computed in this work) and an average injected electron velocity v_e = 10^7 cm/s, the inelastic scattering length is about 5–10 nm. Since an electron of energy E must release multiple phonons to relax its energy fully down to the thermal average, the total length of the heat generation in the drain can be much longer than the inelastic scattering length[33] and can be written approximately as

$$L_h \simeq \frac{E - (3/2)k_BT}{\hbar\bar{\omega}} v_e\tau_o \tag{37}$$

where $\hbar\bar{\omega}$ is the average emitted phonon energy. The average energy of the hot electrons injected across the drain scales linearly with the applied voltage and it is a significant fraction of it ($E \sim \alpha V$). Furthermore, if the electron energy is significantly larger (several tenths of an electron volt) than $3k_BT/2$ (39 meV at 300 K), the multiplying fraction in Eq. (37) can be reduced to $E/(\hbar\bar{\omega})$. If the average emitted phonon (including acoustic and optical modes) has an energy about $\hbar\bar{\omega}$ = 50 meV, the multiplying factor is approximately 10–20 at biases near 1 V. Hence, the length of the heat generation region in the drain is on the order of $L_h \approx 100$ nm, which is consistent with both our[79] and other's findings[33] from Monte Carlo simulations, as shown in Fig. 14(c). Equation (37) is a crude approximation, but it gives a good order of magnitude estimate and correctly explains the long (much longer than the channel length when quasi-ballistic transport dominates) heat generation region in the device drain. These findings are also consistent with the work of Lake and Datta,[5] implying that heat dissipation in mesoscopic devices occurs in or near the contacts rather than in the active device region, i.e., when the length of the active region is on the order of the inelastic mean free path.

6.3 Thermionic Cooling at the Source

Unlike in the drain, the electrons in the source region are very close to thermal equilibrium with the lattice temperature. However, a careful examination of both Figs. 13 and 14 reveals a small, but consistently negative heat generation region (lattice cooling) at the beginning of the channel. This is a thermionic (TI) cooling effect due to the presence of the potential barrier at the injection point from the source into the channel. The situation is similar to the

Peltier effect, but the root cause is slightly different.[33,80,81] Thermionic cooling is a non-equilibrium effect similar to evaporative cooling, in which hot electrons are selectively emitted over an energy barrier.[13,82] To understand the TI cooling effect when current flows over the potential barrier into the channel, consider the electron energy distribution just to the left of the barrier. The electrons in the source are essentially in thermal equilibrium and the distribution is a Fermi-Dirac function at temperature T. From this distribution, only the electrons with forward-oriented momenta and energies greater than the barrier height are going to travel into the channel. Since the high energy tail of the distribution is able to leave, the remaining electrons will have an average energy below the thermal average. By the principle of detailed balance, these remaining electrons will, on average, absorb more phonons than they emit, which contributes to a net effective cooling of the lattice.

The TI cooling effect as current flows over an energy barrier can also be explained from the classical drift-diffusion theory of Eq. (2) (the $\mathbf{J} \cdot \mathbf{F}$ approach) and the discussion surrounding it. The electric field and the direction of current flow are pointing in opposite directions at the beginning of the energy barrier into the channel, hence the $\mathbf{J} \cdot \mathbf{F}$ product is negative, and so is the heat generation rate. In other words, electrons diffusing *against* an energy barrier extract the energy required to move up the conduction band slope (against the electric field) from the lattice, through net phonon absorption. This phenomenon has been studied and exploited in the design of heterojunction laser diodes, where the energy barriers introduced by band structure offsets can be optimized to provide internal thermoelectric cooling near the active laser region.[11]

7. SUMMARY

The functionality, transport, and energy consumption of electronics is strongly influenced by the electron-phonon interaction. Therefore, understanding and controlling such fundamental aspects could impact a wide range of applications from mobile devices (10^{-3} W) to massive data centers (10^9 W). In this chapter, we described the electron transport and energy dissipation, particularly from the point of view of a Monte Carlo simulation approach. Various aspects of the Monte Carlo implementation, scattering physics, modeling of energy bands, and phonon dispersion were described. Applying the method to transport in silicon we uncovered, for example, that heat generation is not evenly divided among phonon modes, but that acoustic phonons receive approximately 1/3 and optical phonons 2/3 of the energy dissipated. We also found that heat dissipation in nanoscale transistors becomes highly asymmetric and nonlocal (with respect to the electric field) in quasi-ballistic devices, when the electron-phonon scattering mean free path becomes comparable to the device size. Finally, we demonstrated the existence of thermionic cooling effects within silicon devices, particularly close to the device source terminal, where charge carriers undergo energy “filtering” in the presence of a potential barrier. While the discussion typically referred to silicon for specificity, the results described can be broadly applied to many other semiconductors and nanoscale device structures. Such aspects are only to be expected to increase in importance as nanoscale devices are reduced to dimensions comparable to or smaller than the electron and phonon mean free path ($\sim$10 nm).

8. ACKNOWLEDGMENTS

I am indebted to S. Sinha, C. Jungemann, M. Fischetti, U. Ravaioli, K. Goodson, and R. Dutton for many discussions and advice. This work was in part supported by the Semiconductor Research Corporation (SRC), the Nanoelectronics Research Initiative (NRI), the ARO Presidential Early Career (PECASE) Award, and the National Science Foundation (NSF) CAREER award.

REFERENCES

1. Pop, E., Energy Dissipation and Transport in Nanoscale Devices, *Nano Res.,* vol. 3, pp. 147–169, 2010.
2. *How Dirty Is Your Data? A Look at the Energy Choices That Power Cloud Computing*, Greenpeace Intl., Amsterdam, The Netherlands, 2011; available at http://www.greenpeace.org/international/en/publications/reports/How-dirty-is-your-data/, [accessed March 1, 2013].
3. Koomey, J., *Growth in Data Center Electricity Use 2005 to 2010*, Analytics Press, Burlingame, CA, 2011; available at http://www.analyticspress.com/datacenters.html, [accessed May 1, 2013].
4. Country Comparison: Electricity Consumption, *CIA World Factbook*, Central Intelligence Agency, Washington, DC, 2011; Available at https://www.cia.gov/library/publications/the-world-factbook/rankorder/2042rank.html, [accessed April 1, 2013].
5. Lake, R. and Datta, S., Energy Balance and Heat Exchange in Mesoscopic Systems, *Phys. Rev. B,* vol. 46, pp. 4757–4763, 1992.
6. Ouyang, Y. and Guo, J., Heat Dissipation in Carbon Nanotube Transistors, *Appl. Phys. Lett.,* vol. 89, p. 183122, 2006.
7. Assad, F., Banoo, K., and Lundstrom, M., The Drift-Diffusion Equation Revisited, *Solid-State Electronics,* vol. 42, pp. 283–295, 1998.
8. Wachutka, G. K., Rigorous Thermodynamic Treatment of Heat Generation and Conduction in Semiconductor Device Modeling, *IEEE Trans. Comput.-Aided Des.,* vol. 9, pp. 1141–1149, 1990.
9. Lindefelt, U., Heat Generation in Semiconductor Devices, *J. Appl. Phys.,* vol. 75, pp. 942–957, 1994.
10. Sverdrup, P. G., Ju, Y. S., and Goodson, K. E., Sub-Continuum Simulations of Heat Conduction in Silicon-on-Insulator Transistors, *ASME J. Heat Transfer,* vol. 123, pp. 130–137, 2001.
11. Pipe, K. P., Ram, R. J., and Shakouri, A., Internal Cooling in a Semiconductor Laser Diode, *IEEE Photon. Technol. Lett.,* vol. 14, pp. 453–455, 2002.
12. Mastrangelo, C. H., Yeh, J. H.-J., and Muller, R. S., Electrical and Optical Characteristics of Vacuum-Sealed Polysilicon Microlamps, *IEEE Trans. Electron. Devices,* vol. 39, pp. 1363–1375, 1992.
13. Pipe, K. P., Ram, R. J., and Shakouri, A., Bias-Dependent Peltier Coefficient and Internal Cooling in Bipolar Devices, *Phys. Rev. B,* vol. 66, p. 125316, 2002.
14. Heikkila, O., Oksanen, J., and Tulkki, J., Ultimate Limit and Temperature Dependency of Light-Emitting Diode Efficiency, *J. Appl. Phys.*, vol. 105, p. 093119, 2009.

15. Santhanam, P., Gray, Jr., D. J., and Ram, R. J., Thermoelectrically Pumped Light-Emitting Diodes Operating above Unity Efficiency, *Phys. Rev. Lett.,* vol. 108, p. 097403, 2012.
16. Liao, A., Zhao, Y., and Pop, E., Avalanche-Induced Current Enhancement in Semiconducting Carbon Nanotubes, *Phys. Rev. Lett.,* vol. 101, p. 256804, 2008.
17. Pop, E., Chui, C. O., Sinha, S., Dutton, R., and Goodson, K., Electro-Thermal Comparison and Performance Optimization of Thin-Body SOI and GOI MOSFETs, *Proceedings of IEEE International Electron Devices Meeting*, San Francisco, IEEE, Piscataway, NJ, pp. 411–414, 2004.
18. Quade, W., Schöll, E., and Rudan, M., Impact Ionization within the Hydrodynamic Approach to Semiconductor Transport, *Solid-State Electron.,* vol. 36, pp. 1493–1505, 1993.
19. Lai, J. and Majumdar, A., Concurrent Thermal and Electrical Modeling of Sub-Micrometer Silicon Devices, *J. Appl. Phys.,* vol. 79, pp. 7353–7361, 1996.
20. Wachutka, G., Consistent Treatment of Carrier Emission and Capture Kinetics in Electrothermal and Energy Transport Models, *Microelectron. J.*, vol. 26, pp. 307–315, 1995.
21. Pop, E., Dutton, R. W., and Goodson, K. E., Analytic Band Monte Carlo Model for Electron Transport in Si Including Acoustic and Optical Phonon Dispersion, *J. Appl. Phys.,* vol. 96, pp. 4998–5005, 2004.
22. Ju, Y. S. and Goodson, K. E., Phonon Scattering in Silicon Thin Films with Thickness of Order 100 nm, *Appl. Phys. Lett.,* vol. 74, pp. 3005–3007, 1999.
23. Mazumder, S. and Majumdar, A., Monte Carlo Study of Phonon Transport in Solid Thin Films Including Dispersion and Polarization, *ASME J. Heat Transfer,* vol. 123, pp. 749–759, 2001.
24. Henry, A. S. and Chen, G., Spectral Phonon Transport Properties of Silicon Based on Molecular Dynamics Simulations and Lattice Dynamics, *J. Comput. Theoret. Nanosci.,* vol. 5, pp. 141–152, 2008.
25. Fischetti, M. V., Neumayer, D. A., and Cartier, E. A., Effective Electron Mobility in Si Inversion Layers in MOS Systems with a High-K Insulator: The Role of Remote Phonon scattering, *J. Appl. Phys.,* vol. 90, pp. 4587–4608, 2001.
26. Artaki, M. and Price, P. J., Hot Phonon Effects in Silicon Field-Effect Transistors, *J. Appl. Phys.,* vol. 65, pp. 1317–1320, 1989.
27. Lugli, P. and Goodnick, S. M., Nonequilibrium Longitudinal-Optical Phonon Effects in GaAs-AlGaAs quantum wells, *Phys. Rev. Lett.,* vol. 59, pp. 716–719, 1987.
28. Jacoboni, C. and Reggiani, L., The Monte Carlo Method for the Solution of Charge Transport in Semiconductors with Applications to Covalent Materials, *Rev. Mod. Phys.,* vol. 55, pp. 645–705, 1983.
29. Fischetti, M. V. and Laux, S. E., Monte Carlo Analysis of Electron Transport in Small Semiconductor Devices Including Band-Structure and Space-Charge Effects, *Phys. Rev. B,* vol. 38, pp. 9721–9745, 1988.
30. Pop, E., Dutton, R. W., and Goodson, K. E., Monte Carlo Simulation of Joule Heating in Bulk and Strained Silicon, *Appl. Phys. Lett.,* vol. 86, p. 082101, 2005.
31. Rowlette, J. A. and Goodson, K. E., Fully Coupled Nonequilibrium Electron-Phonon Transport in Nanometer-Scale Silicon FETs, *IEEE Trans. Electron. Devices,* vol. 55, pp. 220–232, 2008.
32. Sinha, S., Pop, E., Dutton, R. W., and Goodson, K. E., Non-Equilibrium Phonon Distributions in Sub-100 nm Silicon Transistors, *ASME J. Heat Transfer,* vol. 128, pp. 638–647, 2006.

33. Zebarjadi, M., Shakouri, A., and Esfarjani, K., Thermoelectric Transport Perpendicular to Thin-Film Heterostructures Calculated Using the Monte Carlo Technique, *Phys. Rev. B,* vol. 74, p. 195331, 2006.

34. Raleva, K., Vasileska, D., Goodnick, S. M., and Nedjalkov, M., Modeling Thermal Effects in Nanodevices, *IEEE Trans. Electron. Devices,* vol. 55, pp. 1306–1316, 2008.

35. Vasileska, D., Raleva, K., and Goodnick, S. M., Modeling Heating Effects in Nanoscale Devices: The Present and the Future, *J. Comput. Electron.,* vol. 7, pp. 66–93, 2008.

36. Canali, C., Jacoboni, C., Nava, F., Ottaviani, G., and Alberigi-Quaranta, A., Electron Drift Velocity in Silicon, *Phys. Rev. B,* vol. 12, pp. 2265–2284, 1975.

37. Tang, J. and Hess, K., Impact Ionization of Electrons in Silicon (Steady State), *J. Appl. Phys.,* vol. 54, pp. 5139–5144, 1983.

38. Sano, N., Aoki, T., Tomizawa, M., and Yoshii, A., Electron Transport and Impact Ionization in Si, *Phys. Rev. B,* vol. 41, pp. 12122–12128, 1990.

39. Jacoboni, C., Minder, R., and Majni, G., Effects of Band Non-Parabolicity on Electron Drift Velocity in Silicon above Room Temperature, *J. Phys. Chem. Solids,* vol. 36, pp. 1129–1133, 1975.

40. Brunetti, R., Jacoboni, C., Nava, F., Reggiani, L., Bosman, G., and Zijlstra, R., Diffusion Coefficient of Electrons in Silicon, *J. Appl. Phys.,* vol. 52, pp. 6713–6722, 1981.

41. Yamada, T., Zhou, J.-R., Miyata, H., and Ferry, D., In-Plane Transport Properties of Si/$Si_{1-x}Ge_x$ Structure and its FET Performance by Computer Simulation, *IEEE Trans. Electron. Devices,* vol. 41, pp. 1513–1522, 1994.

42. Fischer, B. and Hofmann, K. R., A Full-Band Monte Carlo Model for the Temperature Dependence of Electron and Hole Transport in Silicon, *Appl. Phys. Lett.,* vol. 76, pp. 583–585, 2000.

43. Yoder, P. D. and Hess, K., First-Principles Monte Carlo Simulation of Transport in Si, *Semiconduct. Sci. Technol.,* vol. 9, pp. 852–854, 1994.

44. Kunikiyo, T., Takenaka, M., Kamakura, Y., Yamaji, M., Mizuno, H., Morifuji, M., Taniguchi, K., and Hamaguchi, C., A Monte Carlo Simulation of Anisotropic Electron Transport in Silicon Including Full Band Structure and Anisotropic Impact-Ionization Model, *J. Appl. Phys.,* vol. 75, pp. 297–312, 1994.

45. Winstead, B. and Ravaioli, U., A Quantum Correction Based on Schrodinger Equation Applied to Monte Carlo Device Simulation, *IEEE Trans. Electron. Devices,* vol. 50, pp. 440–446, 2003.

46. Duncan, A., Ravaioli, U., and Jakumeit, J., Full-Band Monte Carlo Investigation of Hot Carrier Trends in the Scaling of Metal-Oxide-Semiconductor Field-Effect Transistors, *IEEE Trans. Electron. Devices,* vol. 45, pp. 867–876, 1998.

47. Bufler, F. M., Asahi, Y., Yoshimura, H., Zechner, C., Schenk, A., and Fichtner, W., Monte Carlo Simulation and Measurement of Nanoscale n-MOSFETs, *IEEE Trans. Electron. Devices,* vol. 50, pp. 418–424, 2003.

48. Pop, E., Sinha, S., and Goodson, K. E., Heat Generation and Transport in Nanometer-Scale Transistors, *Proc. IEEE,* vol. 94, pp. 1587–1601, 2006.

49. Lundstrom, M., *Fundamentals of Carrier Transport*, 2nd ed., Cambridge University Press, Cambridge, UK, 2000.

50. Tomizawa, K., *Numerical Simulation of Submicron Semiconductor Devices*, Artech House,

Boston, 1993.

51. Pop, E., *Self-Heating and Scaling of Thin-Body Transistors*, PhD, Stanford University, Stanford, 2005.
52. Jungemann, C. and Meinerzhagen, B., On the Applicability of Nonself-Consistent Monte Carlo Device Simulations, *IEEE Trans. Electron. Devices,* vol. 49, pp. 1072–1074, 2002.
53. Dolling, G., Lattice Vibrations in Crystals with the Diamond Structure, Proc of *Symposium on Inelastic Scattering of Neutrons in Solids and Liquids*, IAEA, Vienna, pp. 37–48, 1963.
54. Long, D., Scattering of conduction electrons by lattice vibrations in silicon, *Phys. Rev.,* vol. 120, pp. 2024–2032, 1960.
55. Green, M. A., Intrinsic concentration, effective densities of states, and effective mass in silicon, *J. Appl. Phys.,* vol. 67, pp. 2944–2954, 1990.
56. Hamaguchi, C., *Basic Semiconductor Physics*: Springer, 2001.
57. Pop, E., Varshney, V., and Roy, A. K., Thermal Properties of Graphene: Fundamentals and Applications, *MRS Bull.,* vol. 37, pp. 1273–1281, 2012.
58. Ferry, D. K., *Semiconductor Transport*, Taylor & Francis, New York, 2000.
59. Herring, C. and Vogt, E., Transport and Deformation-Potential Theory for Many-Valley Semiconductors with Anisotropic Scattering, *Phys. Rev.,* vol. 101, pp. 944–961, 1956.
60. Haug, A., *Theoretical Solid State Physics*, Vol. 2, Pergamon Press, New York, 1972.
61. Mizuno, H., Taniguchi, K., and Hamaguchi, C., Electron-Transport Simulation in Silicon Including Anisotropic Phonon Scattering Rate, *Phys. Rev. B,* vol. 48, pp. 1512–1516, 1993.
62. Fischetti, M. and Laux, S., Band Structure, Deformation Potentials, and Carrier Mobility in Strained Si, Ge, and SiGe Alloys, *J. Appl. Phys.,* vol. 80, pp. 2234–2252, 1996.
63. Fischetti, M. V. and Laux, S., Monte Carlo Study of Electron Transport in Silicon Inversion Layers, *Phys. Rev. B,* vol. 48, pp. 2244–2274, 1993.
64. Yu, P. Y. and Cardona, M., *Fundamentals of Semiconductors*, Springer, New York, 1996.
65. Ismail, K., Nelson, S., Chu, J., and Meyerson, B., Electron Transport Properties of Si/SiGe Heterostructures: Measurements and Device Implications, *Appl. Phys. Lett.,* vol. 63, pp. 660–662, 1993.
66. Sangiorgi, E., Ricco, B., and Venturi, F., MOS^2: An Efficient MOnte Carlo Simulator for MOS Devices, *IEEE Trans. Comput.-Aided Des. Integr. Circuits Sys.,* vol. 7, pp. 259–271, 1988.
67. Ridley, B., Reconciliation of the Conwell-Weisskopf and Brooks-Herring Formulae for Charged-Impurity Scattering in Semiconductors: Third-Body Interference, *J. Phys. C,* vol. 10, pp. 1589–1593, 1977.
68. Kosina, H., A Method to Reduce Small-Angle Scattering in Monte Carlo Device Analysis, *IEEE Trans. Electron. Devices,* vol. 46, pp. 1196–1200, 1999.
69. Kosina, H. and Kaiblinger-Grujin, G., Ionized-Impurity Scattering of Majority Electrons in Silicon, *Solid-State Electron.,* vol. 42, pp. 331–338, 1998.
70. Chen, D., Kan, E. C., Ravaioli, U., Shu, C.-W., and Dutton, R. W., An Improved Energy Transport Model Including Nonparabolicity and Non-Maxwellian Distribution Effects, *IEEE Electron. Device Lett.,* vol. 13, pp. 26–28, 1992.
71. Woolard, D., Tian, H., Littlejohn, M., and Kim, K., The Implementation of Physical Boundary Conditions in the Monte Carlo Simulation of Electron Devices, *IEEE Trans. Comput.-Aided*

Des. Integr. Circuits Syst., vol. 13, pp. 1241–1246, 1994.

72. Hess, K., *Monte Carlo Device Simulation: Full Band and Beyond*, Kluwer Academic Publishers, Boston, MA, 1991.
73. Hockney, R. W. and Eastwood, J. W., *Computer Simulation Using Particles*, IOP Publishing, 1988.
74. Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P., *Numerical reCipes in C*, 2nd ed., Cambridge University Press, Cambridge, UK, 1992.
75. Ferziger, J. H., *Numerical Methods for Engineering Applications*, 2nd ed., Wiley, Hoboken, NJ, 1998.
76. Jungemann, C., Emunds, A., and Engl, W., Simulation of Linear and Nonlinear Electron Transport in Homogeneous Silicon Inversion Layers, *Solid-State Electron.,* vol. 36, pp. 1529–1540, 1993.
77. Kittel, C., *Introduction to Solid State Physics*, 7th ed., Wiley, Hoboken, NJ, 1995.
78. *International Technology Roadmap for Semiconductors*, 2007; Available at http://public.itrs.net, [accessed Mar. 2007].
79. Pop, E., Rowlette, J., Dutton, R. W., and Goodson, K. E., Joule Heating under Quasi-Ballistic Transport Conditions in Bulk and Strained Silicon Devices, *Proceedings of International Conference on Simulation of Semiconductor Processes and Devices*, Tokyo, IEEE, Piscataway, NJ, pp. 307–310, 2005.
80. Shakouri, A. and Bowers, J. E., Heterostructure Integrated Thermionic Coolers, *Appl. Phys. Lett.,* vol. 71, pp. 1234–1236, 1997.
81. Mahan, G. D. and Woods, L. M., Multilayer Thermionic Refrigeration, *Phys. Rev. Lett.,* vol. 80, pp. 4016–4019, 1998.
82. Mahan, G. D., Sofo, J. O., and Bartkowiak, M., Multilayer Thermionic Refrigerator and Generator, *J. Appl. Phys.,* vol. 83, pp. 4683–4689, 1998.