Fully Coupled Nonequilibrium Electron–Phonon Transport in Nanometer-Scale Silicon FETs

Jeremy A. Rowlette and Kenneth E. Goodson

(Invited Paper)

***Abstract*—Heat conduction from transistors and interconnects is a critical design consideration for computing below the 20-nm milestone. This paper reviews detailed heat generation and transport mechanisms in silicon devices with a focus on the nonequilibrium behavior of electrons and phonons. Fully coupled and self-consistent ballistic phonon and electron simulations are developed in order to examine the departure from equilibrium within the phonon system and its relevance for properly simulating the electrical behavior of devices. We illustrate the manner in which nanoscale-transport phenomena are critically important for a broad variety of low-dimensional silicon-based devices, including FinFETs and depleted substrate transistors.**

***Index Terms*—CMOS, FinFET, heat, Monte Carlo, multigate, nanoscale, nanotechnology, nonequilibrium, optical phonon, phonon lifetime, phonons, power, thermal, transistor, transport.**

I. Introduction

The continued scaling of transistor dimensions and spatial density is causing major thermal management challenges on the chip. Effective removal of heat from the transistor and interconnect layers will be a growing challenge to the successful scaling of digital nanotechnologies for the foreseeable future. Today, thermal management is necessarily being addressed at all levels of design from the transistor to the circuit and microarchitecture and to the package and enclosure. While the challenges are growing at all of these levels, the electrothermal phenomena occurring within transistors are particularly challenging because of the multicarrier transport physics involved. Reducing channel lengths L_g in order to increase packing density and to reduce energy-delay product [1] has a direct impact on the departure from equilibrium of the electron and phonon systems within devices, thus increasing both the complexity and the importance of nanoscale electrothermal phenomena. Transistor-level thermal management is made more important by the move to thin-body single and multigate devices that provide an improved control of the channel electric field. The most promising multigate device is the FinFET [2] of which numerous derivatives have been proposed in recent years [3], [4]. While these devices offer superior subthreshold slopes approaching the theoretical limit of 60 mV/decade, they exhibit higher thermal resistance than bulk devices because of geometric confinement by low thermal-conductivity materials as well as enhanced phonon-boundary scattering in the active layers.

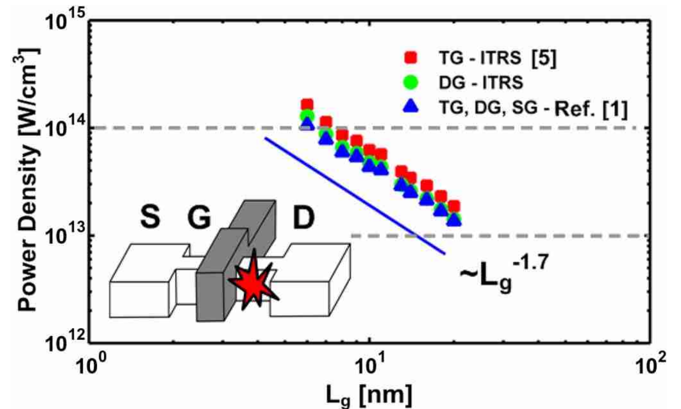

Fig. 1. Estimated steady-state volumetric power density in the drain of future thin-body TG, DG, and SG transistors as a function of physical channel length L_g.

The importance of nonequilibrium thermal-transport phenomena owes much to the high power densities per unit volume in nanodevices, which are continuing to increase despite a gradually reducing operating voltage. Based on simple volumetric scaling arguments and by considering a modest rate of reduction in voltage, it can be argued that the power density should be proportional to $L_g^{-1.7}$. This basic trend is shown in Fig. 1 for sub-20-nm single-gate (SG), double-gate (DG), and triple-gate (TG) thin-body silicon transistors. The upper two curves correspond to the International Technology Roadmap for Semiconductors [5] values for the TG and DG devices, and the lower curve represents the scaling set forth in [1] for the TG, DG, and SG devices. The volumetric power density of a 20-nm device is on the order of 10 TW/cm^3, and a 6-nm device at the end of the roadmap is expected to increase by one order of magnitude. The calculations assume a uniform current distribution and that heat generation in the S/D fin extension dominates, as will be discussed in Section IV.

Manuscript received June 12, 2007; revised October 9, 2007. This work was supported in part by the Semiconductor Research Corporation (SRC) under Task 1043. The work of J. A. Rowlette is supported by an Intel Graduate Fellowship. The review of this paper was arranged by Editor J. Welser.

J. A. Rowlette is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305-3030 USA (e-mail: rowlette@stanford.edu).

K. E. Goodson is with the Department of Mechanical Engineering, Stanford University, Stanford, CA 94305-3030 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2007.911043

0018-9383/$25.00 © 2008 IEEE

Authorized licensed use limited to: Stanford University. Downloaded on November 18, 2008 at 12:51 from IEEE Xplore. Restrictions apply.

This paper focuses on the fundamental heat generation and transport mechanisms in silicon devices and shows why the nanoscale and nonequilibrium thermal phenomena are being factored into device technology decisions. This paper is organized into four main sections. Sections II and III deal with the detailed electron transport and the generation and transport of heat at length and time scales less than 100 nm and 10 ps, respectively. In Section IV, we close the transport loop by describing efficient simulation techniques for coupling the heat and charge transport, which is an essential requirement to understand the thermal impact on electrical characteristics in future devices. In Section V, we use the results of Sections II–IV to understand the implications of nonequilibrium coupled charge–heat transport at nanometer length scales and their impact on leakage power, electrical drive current, and reliability. Additionally, we address the topics of anomalous temperature rise near nanometer-scale heat sources as well as the issue of hot optical phonons in silicon, both topics being heavily debated over the past two decades.

II. Heat Generation

Within the transistor, thermal energy is predominantly stored and transported by the vibrational modes of the lattice, or phonons, of the semiconducting material. Heat generation is the result of electrons transferring their excess energy gained from the electric field to the phonon population by means of scattering. To model the details of the heat-generation process, we use an electron Monte Carlo simulator (e-MC) developed by Pop *et al.* [6] and later modified by Rowlette *et al.* [7]. While the models are summarized here, the reader is referred to the original works as well as a large body of work describing the general Monte Carlo technique [8], [9]. The e-MC code employs a six-valley, analytic, nonparabolic, single conduction-band model described by [8]

$$E_k(1+\alpha E_k) = \frac{\hbar^2}{2m_e}\sum_i \frac{(k_i - k_{\nu i})^2}{m^*_{\nu i}}, \quad i = x, y, z, \text{ and } \nu = 1, 2, 3, \ldots, 6. \quad (1)$$

The effective mass $m^*_{\nu i}$ is along the direction index i for valley ν, and the nonparabolicity factor α is taken to be 0.5 eV^{-1} for silicon. The valleys are centered at the six equivalent $(2\pi/a)$ [0.85, 0, 0] points within the first Brillouin Zone (BZ). The analytic electron-band structure is shown in Fig. 2 (red) along with (blue) an accurate full-band (FB) model [10]. The density of states (DOS) for the analytic band closely matches that for the FB description below about 1.5 eV [6]. Since gate voltages are not expected to rise above 1 V for future nanotransistor technologies, the analytic nonparabolic band (NPB) provides reasonable accuracy while achieving a significant reduction in computational cost compared with the more accurate FB codes [11]–[13]. By reducing the complexity of the electron-band model, the e-MC program is able to efficiently incorporate a more detailed dispersion-relation model for both acoustic and optical phonons for computing scattering probabilities and

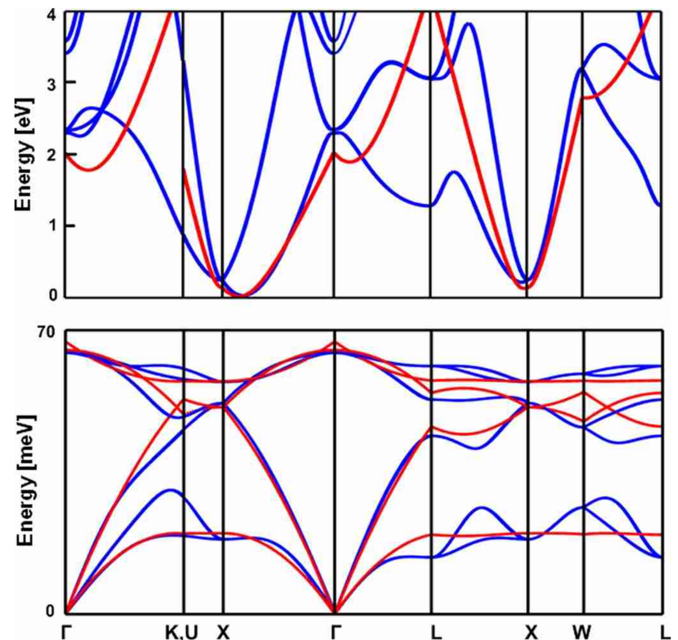

Fig. 2. (Top) Electron-band structure for silicon which is computed using (blue) an empirical tight-binding parameter set [11] and (red) a single six-valley analytic nonparabolic electron band [10] used in this paper. (Bottom) Phonon dispersion for silicon which is computed using (blue) a VFF model [12] compared with (red) an isotropic model using simple quadratics [8], as used in this paper.

energy- and crystal-momentum-conserving final states. The dispersion for each polarization branch s is modeled by a quadratic relation given by $\omega_{q,s} = \omega_0^s + \nu^s q + c^s q^2$, where q is the wave-vector magnitude along an arbitrary direction in the crystal [6]. The bottom panel of Fig. 2 shows (red) the isotropic dispersion relation for silicon for directions of high symmetry along with a FB phonon dispersion calculated using a six-parameter empirical valence-force-field (VFF) model [14]. The quadratic parameters used in the e-MC simulations here were optimized to fit the dispersion relation along $\langle 100\rangle$ directions [6].

In addition to elastic ionized impurity scattering, both intravalley and intervalley phonon scattering are treated inelastically.[1] The intervalley scattering is modeled by three g-type (equivalent valley) and three f-type (nonequivalent valley) transitions, both of which are Umklapp (U) processes [15]. The g-type phonons are directed along $\langle 100\rangle$ and are located at $(2\pi/a)$ [0.3, 0, 0] and at equivalent points within the BZ, whereas the f-type phonon is directed 11° off the $\langle 100\rangle$ equivalent directions at $(2\pi/a)$ [1, 0.15, 0.15] and at equivalent points within the BZ. To enable the isotropic phonon-dispersion assumption, the f-type phonons are considered to be directed along $\langle 100\rangle$ and positioned at the edge of the BZ at the $\mathbf{X}$ point [6].

Electron–phonon scattering is treated in the standard way using Fermi's Golden Rule. The forms of the scattering rates

[1] Unlike in early Monte Carlo codes, all phonon scattering processes, including the intravalley acoustic scattering, are treated inelastically. This is important because an appreciable amount of the heat is generated in small wave-vector acoustic modes, even at large power densities.

Authorized licensed use limited to: Stanford University. Downloaded on November 18, 2008 at 12:51 from IEEE Xplore. Restrictions apply.

TABLE I
SUMMARY OF ELECTRON- AND PHONON-TRANSPORT-MODEL PARAMETERS USED IN SIMULATIONS DISCUSSED IN THIS PAPER

Electron Transport Parameters							
Δ_f^{TA} [eV/cm]	$\Delta_f^{LA/LO}$ [eV/cm]	Δ_f^{TO} [eV/cm]	Δ_g^{TA} [eV/cm]	Δ_g^{LA} [eV/cm]	Δ_g^{LO} [eV/cm]	D_{LA} [eV]	D_{TA} [eV]
0.5x10^8	3.5x10^8	1.5x10^8	0.3x10^8	1.5x10^8	6.0x10^8	6.39	3.01
Phonon Transport Parameters							
$A_{N,LA}$ [s-K^{-3}]	$A_{N,TA}$ [K^{-4}]	$A_{U,TA}$ [s^{-1}]	$A_{\delta M}+A_{\delta R}$ [s^3] n_0=1x10^{20} cm^{-3}		ω_1 [THz]	ω_2 [THz]	τ_{op} [ps]
2.0x10^{-24}	9.3x10^{-13}	5.5x10^{-18}	1.3x10^{-44}(n/n$_0$)		23.56	27.49	0.5-10

for both the intravalley acoustic and the intervalley acoustic and optical scattering are, respectively [6]

$$\Gamma_i = \frac{D_{a,s}^2 m_d}{4\pi\rho\hbar^2 k_s} \int\limits_q \frac{1}{\omega_{q,s}} \left(N_{q,s} + \frac{1}{2} \mp \frac{1}{2} \right) I_q^2 q^3 dq \tag{2a}$$

$$\Gamma_{if} = \frac{\pi \Delta_{if}^2 Z_f}{2\rho\omega_{q,s}} \left(N_{q,s} + \frac{1}{2} \mp \frac{1}{2} \right) g_{df}(E_k \pm \hbar\omega_{q,s}). \tag{2b}$$

The values for the effective intravalley and intervalley deformation potentials D_a and Δ_{if} used in this paper are reported in Table I. The upper and lower signs correspond to absorption and emission processes, respectively, throughout this paper. $N_{q,s}$ is the average phonon occupation number given by the Bose–Einstein (B–E) distribution $N_{q,s} = [\exp(\hbar\omega_{q,s}/k_B T) - 1]^{-1}$ at equilibrium. Under nonequilibrium conditions, $N_{q,s}$ must be determined by solving the Boltzmann transport equation (BTE), which will be discussed in the subsequent sections.

The present calculations employ the multivalley NPB model because of its computational simplicity and do not attempt to resolve the specific limitations of this approximate method. However, we must caution the reader that the set of phonon deformation potentials used in this paper (Table I) is not unique and that a wide range of values have been reported over the past three decades (cf. [6] and [16]). FB models typically yield lower (~2–3x) phonon deformation potentials due to the differences in the electron-phonon joint DOS (JDOS). We will further discuss the potential shortcomings of making the NPB model approximation in Section V.

The local volumetric heat-generation rate $Q'''(\overleftarrow{r})$ (in watts per cubic centimeter) is equal to the energy-weighted difference of emitted (ems) and absorbed (abs) phonons times the ratio of the electron density $n_e(\vec{r})$ to the number of simulated electrons $N_{\rm sim}$ (typically, 10 000) divided by the simulation time-interval step Δt [17], [18]

$$Q'''(\vec{r}) = \sum_{q,s} Q'''_{q,s}(\vec{r}) = \frac{n_e(\vec{r})}{N_{\rm sim}\Delta t} \sum_{q,s} \left(\hbar\omega_{q,s}^{\rm ems} - \hbar\omega_{q,s}^{\rm abs} \right)\Big|_{\vec{r}}. \tag{3}$$

Here, we define the modal volumetric heat-generation rate as

$$Q'''_{q,s}(\vec{r}) \equiv \frac{n_e(\vec{r})}{N_{\rm sim}\Delta t} \left(\hbar\omega_{q,s}^{\rm ems} - \hbar\omega_{q,s}^{\rm abs} \right)\Big|_{\vec{r}}. \tag{4}$$

The frequency spectrum of net-emitted phonons for both bulk and strained silicon was computed as a function of electric field and for various doping conditions by Pop *et al.* [17]. At low fields, the heat generation is restricted to the intravalley acoustic modes as the electrons do not have sufficient energy to transfer between the valleys. At intermediate fields, sharp peaks centered around the signature g- and f-type phonons are observed. At higher fields, the emission spectrum broadens about each of the intervalley peaks as a result of the finite phonon dispersion which leads to a gradual relaxation of the k-conservation rule associated with the f- and g-type phonons.

III. PHONON TRANSPORT

The electron system cools in the drain end of the transistor through net phonon emission. The phonon population then proceeds to evolve in a manner which tends to bring the phonon population back toward the equilibrium B–E distribution. This evolution of the phonon distribution can be described by a phonon BTE. Scattering mechanisms include phonon–phonon (p–p), phonon–electron (p–e), phonon–impurity/vacancy (p–i), and phonon–boundary (p–b) types. In this paper, we use a split-flux form of the phonon BTE (p-SFBTE) introduced by Sinha *et al.* [19] to describe the phonon transport. The p-SFBTE was derived under the relaxation-time approximation and ensures macroscopic energy conservation. It captures ballistic phonon conduction near the transistor hotspot and also yields a convenient interface to continuum calculations (i.e., diffusive conduction) far from the hotspot. In effect, the phonon distribution is split into two populations. The first is a near-equilibrium component $N_{q,s}(T_F)$ which has the B–E distribution corresponding to a temperature T_F that obeys Fourier's heat conduction law. The second population $n_{q,s}$ is a nonequilibrium departure component, which dominates the transport near the hotspot and is determined by solving the phonon BTE in the relaxation-time approximation given by

$$\nu_{q,s} \cdot \nabla_r n_{q,s} = -\frac{n_{q,s}}{\tau_{q,s}} + \dot{n}_{q,s}. \tag{5}$$

Here, $\nu_{q,s}$, $\tau_{q,s}$, and $\dot{n}_{q,s}(\vec{r})$ are the modal group velocity, lifetime, and source term (in numbers per second), respectively, for a phonon of mode q and branch s. $\dot{n}_{q,s}(\vec{r})$ is determined from the e-MC simulation output and is directly related to the modal volumetric heat-generation rate $Q'''_{q,s}(\vec{r})$ given by (4)

Authorized licensed use limited to: Stanford University. Downloaded on November 18, 2008 at 12:51 from IEEE Xplore. Restrictions apply.

through a division by the number of phonon states per unit volume $g(\omega_{q,s})\Delta\omega$

$$\dot{n}_{q,s}(\vec{r}) = \frac{Q'''_{q,s}(\vec{r})}{\hbar\omega_{q,s} g(\omega_{q,s})\Delta\omega}. \quad (6)$$

Once $n_{q,s}$ is determined, macroscopic energy conservation is used to determine T_F based on

$$\frac{1}{(2\pi)^3}\int \frac{n_{q,s}(\vec{r})}{\tau_{q,s}}\hbar\omega_{q,s} d\,\vec{q} + \overleftrightarrow{k}\cdot\nabla_r^2 T_F(\vec{r}) = 0. \quad (7)$$

Here, $\overleftrightarrow{k}$ is the thermal-conductivity tensor, which for thin films can be appropriately modified to account for increased boundary scattering.[2] The boundary scattering becomes important if the smallest dimension of a material domain is comparable to the mean-free-path (Λ) of a phonon mode, which is given by the product of the modal group velocity and lifetime $(\Lambda_{q,s} = \nu_{q,s} \cdot \tau_{q,s})$ [20], [21]. For LA modes, which conduct most of the heat, Λ is on the order of 100 nm near room temperature. However, for the optical phonons, $\Lambda < 10$ nm, and therefore, only in ultrathin films or inversion layers will the boundary scattering play an important role in the optical phonon transport.

Once $N_{q,s}(T_F(\vec{r}))$ and $n_{q,s}(\vec{r})$ have been determined from (5) and (7), an effective temperature T_{eff} can be defined by equating the total energy density at a particular location to that for an equivalent B–E-distributed population and by integrating over the appropriate polarization branches and wave-vector space according to (8)

$$\frac{1}{(2\pi)^3}\sum_s \int N_{q,s}(T_{\text{eff}})\hbar\omega_{q,s} d\,\vec{q} = \frac{1}{(2\pi)^3}\sum_s \int \left(N_{q,s}(T_F) + n_{q,s}\right)\hbar\omega_{q,s} d\,\vec{q}\,. \quad (8)$$

If we restrict the integration to a particular polarization branch, then we obtain an effective branch temperature. If we further restrict the integration to include only a single wave vector and a branch, then we obtain an effective temperature for a single mode at $\omega_{q,s}$.[3] The use of an effective temperature is merely a convenient means for communicating the degree to which a particular segment of the phonon population deviates from the equilibrium. Ultimately, what matters is the modal occupation number since this determines the strength of the scattering processes through (2).

The transient form of the p-SFBTE can also be found in [19]. In that work, Sinha *et al.* examined the important issue of phonon population buildup between successive clock cycles using a typical phonon-generation spectrum calculated by Monte Carlo simulations. They concluded that the optical phonon lifetimes were sufficiently short to prevent phonon accumulation from cycle to cycle for typical operating frequencies. Although the transient problem is important, we will continue to focus our attention on the steady-state solutions within this paper. We now turn our attention to the determination of the modal lifetimes τ which are essential for capturing the transport physics of the microscopic system.

Empirically determined phenomenological scattering rates $(\tau \sim 1/\Gamma)$ for the acoustic phonons for each type of scattering mechanism were discussed extensively in [22]. The forms of the equations are based on the early works of Klemens [23], Callaway [24], and Holland [25] and arise from limiting forms of the BTE and the assumption of an isotropic dispersion relation. The general forms for normal (N) and Umklapp (U) phonon–phonon (p–p) as well as phonon–impurity/defect (p–i) scattering rates for acoustic phonons are summarized in

$$\Gamma_{p-p,N}^{\text{LA}} = A_{N,\text{LA}}\omega^2 T^3 \quad (9a)$$

$$\Gamma_{p-p,N}^{\text{TA}} = A_{N,\text{TA}}\omega T^4 \quad (9b)$$

$$\Gamma_{p-p,U}^{\text{TA}} = \begin{cases} A_{U,\text{TA}}/\sinh(\hbar\omega/k_B T); & \omega_1 < \omega < \omega_2 \\ 0; & \omega < \omega_1 \end{cases} \quad (9c)$$

$$\Gamma_{p-i} = (A_{\delta M} + A_{\delta R})\omega^4 \quad (9d)$$

where the set of A coefficients is taken to be independent of frequency and temperature. The best known values for these coefficients, along with the frequency parameters ω_1 and ω_2 appearing in (9c) for the TA Umklapp scattering, are listed in Table I. In (9d), the coefficients $A_{\delta M}$ and $A_{\delta R}$ correspond to the impurity scattering caused by mass differences and local lattice distortion, respectively, and are both taken to be proportional to the impurity concentration.

The thermal energy is transported out away from the transistor hotspot primarily through low-energy acoustic modes which have group velocities between 5000 and 9000 m/s in silicon. However, as was shown in [17], a significant amount of the thermal energy (as much as 2/3)[4] is initially stored in the optical phonon modes which have group velocities less than about 1000 m/s. Thermal conduction, therefore, has the potential to be impeded locally as a result of the additional energy decay step required for optical modes to decay into the acoustic modes. Many researchers have cited the potential of an energy bottleneck arising from a relatively long relaxation time for optical phonons compared with the electron–phonon scattering time ($\sim$100 fs) [26]. Such an intermediate decay process is believed to set an upper limit on the frequency performance for some important III–V quantum-well optoelectronic devices [27]. The lifetimes of the optical modes are therefore very important parameters in understanding nonequilibrium heat conduction near the transistor hotspot. In particular, the g-type

[2]It is important to recognize that while the phonon-boundary scattering can reduce the effective thermal conductivity of thin films, this additional scattering mechanism may not necessarily force the phonon system farther from the equilibrium. On the contrary, the additional scattering may serve to allow the local phonon system to more closely approximate the B–E distribution at a given temperature. In contrast, the introduction of specific phonon modes due to the electron scattering is inherently disruptive to the phonon-distribution function and can cause a severe departure from the equilibrium.

[3]In this limit, we are essentially determining the appropriate temperature that, when included in the argument of the B–E distribution function for a given frequency, will produce the appropriate occupation number in the nonequilibrium condition.

[4]In this paper, we find that $\sim$2/3 of the energy is dissipated by optical phonon modes with the remaining energy going into the acoustic modes. The FB MC codes, which compute the individual matrix elements using pseudopotential theory, have shown that the situation is reversed with the acoustic modes receiving the $\sim$2/3 majority of the thermal energy.

Authorized licensed use limited to: Stanford University. Downloaded on November 18, 2008 at 12:51 from IEEE Xplore. Restrictions apply.

longitudinal-optical (g-LO) phonon decay rate is thought to play an important role in the nonequilibrium energy relaxation in the drain of silicon-based transistors and in determining the onset of hot phonon effects [19], [26], [28]. This is because of its strong coupling with high-energy electrons[5] and its relatively low-modal heat capacity as determined by the inverse of its group velocity. However, despite the importance of the decay processes of the g-LO phonon, among others, few researchers have attempted to calculate the lifetime or to illuminate the decay channels available to this mode (cf. [29]). One of the main reasons for this is that the g-LO phonon is not optically active like the zone-center Raman active LO–TO mode (R-LTO). Furthermore, because of the reduced symmetry of the g-LO mode, simulations of these phonons require large supercells and, therefore, extensive computational resources. Sinha *et al.* [30] recently performed detailed molecular-dynamic (MD) simulations of the g-LO phonon wavepackets. The key results from that work were that normal three-phonon processes dominated in the relaxation process and that the primary decay channels were of LO $\rightarrow$ LA + TA type. Despite the complexity and rigor of the MD simulations, a subtle limitation in the choice of interatomic potentials was that they do not reproduce the exact harmonic eigen frequencies of the Si crystal. Hence, the decay channels described in that work are necessarily different than what can occur in Si. To provide additional insight into the physics of the decay process for an arbitrary optical phonon mode in silicon, and particularly the g-LO phonon, we calculate the density of final states[6] $g_2(\omega, \omega_0 - \omega)$ for pairs of phonons which conserve both energy $(\omega_{s_0}(\vec{q}_0) = \omega_{s_1}(\vec{q}_1) + \omega_{s_2}(\vec{q}_2))$ and crystal momenta $(\vec{q}_0 = \vec{q}_1 + \vec{q}_2 + \vec{G})$ for an optical phonon with initial wave vector $\vec{q}_0$ and branch index s_0. This "final-state spectrum" for the three-phonon processes arises naturally from third-order anharmonic perturbation theory as applied to the calculation of the energy-relaxation time for a single mode. Such calculations were rigorously performed using density functional theory (DFT) for the R-LTO mode of silicon by Debernardi *et al.* [31]. The expression for the transition rate under the assumption that only three-phonon processes are present is given by[7]

$$\Gamma\left(\omega_{s_0}(\vec{q}_0)\right) \propto \sum_{\vec{q}_1, \vec{q}_2, s_1, s_2} \left| U\begin{pmatrix} \vec{q}_0, & \vec{q}_1, & \vec{q}_2 \\ s_0, & s_1, & s_2 \end{pmatrix} \right|^2 \times \frac{N\left(\omega_{s_1}(\vec{q}_1)\right) + N\left(\omega_{s_2}(\vec{q}_2)\right) + 1}{\omega_{s_0}(\vec{q}_0)\omega_{s_1}(\vec{q}_1)\omega_{s_2}(\vec{q}_2)} \times \delta\left(\omega_{s_0}(\vec{q}_0) - \omega_{s_1}(\vec{q}_1) - \omega_{s_2}(\vec{q}_2)\right) \quad (10)$$

where U is the third-order matrix element. We retain only third-order processes in which the initial phonon decays into tow lower energy modes, a reasonable approximation considering the phonon dispersion relation for silicon [31].

Using a tetrahedral BZ integration method [32]–[35], we calculate $g_2(\omega, \omega_0 - \omega)$ for various initial LO phonon modes along $\Gamma - X$ (including the R-LTO mode, the g-LO mode, and approximately the f-LO mode, as discussed in Section II) by integrating over the entire BZ using the FB phonon dispersion relation in Fig. 2 [14].[8] We restrict our calculations to normal processes, i.e., $\vec{G} = 0$. Fig. 3 shows the results of the calculations for the LO phonons with initial wave vectors of $\vec{q}_0 = (2\pi/a)\ [0, 0, \alpha]$, where $\alpha = 0$, 0.3, 0.5, 0.7, and 1.0, which are labeled as (a)–(e), respectively. The single-phonon DOS function $g(\omega)$ for silicon is superimposed on top of the $g_2(\omega, \omega_0 - \omega)$ plot for the R-LTO mode for easy comparison. We now point out several key features in these plots. First, $g_2(\omega, \omega_0 - \omega)$ is always even with respect to $\omega = \omega_0/2$. The monotonic red shift in $\omega_0/2$ seen in going from (a) to (e) is a consequence of the LO dispersion along $\Gamma - X$. Second, we see that $g_2(\omega, \omega_0 - \omega)$ is strongly peaked at regions where both $g(\omega)$ and $g(\omega_0 - \omega)$ have comparable strength. This is why combinations of low- and high-frequency modes contribute negligibly to $g_2(\omega, \omega_0 - \omega)$ in silicon despite the very large $g(\omega)$ at high frequencies. Finally, we point out the relatively weak central peak around $\omega_0/2$ for the R-LTO mode. This is the so-called "Klemens channel" [36]. As shown in Fig. 3, this channel has a relatively small $g_2(\omega, \omega_0 - \omega)$, and hence, it should not dominate in the decay for the R-LTO mode unlike Klemens' initial postulation. These results are consistent with the work of Debernardi *et al.* [31].[9]

Because of the stated importance of the g-LO phonon, we examine its $g_2(\omega, \omega_0 - \omega)$ spectrum more closely in the bottom panel of Fig. 3. We identify four dominant final-state phonon pairs, which are labeled as 1–4, respectively, and their representative wave vectors and energies are summarized in Table II. For each of these pairs, we observe only combinations of the form LO $\rightarrow$ LA + TA which is consistent with the detailed MD calculations of [30].

Finally, we take our calculations one step further and estimate the intrinsic phonon lifetime for each of the modes shown in Fig. 6 as a function of temperature. We do so by computing (12) and assuming the third-order matrix elements to be a constant equal to U_0 which we fit to the Raman linewidth data of [37]. After obtaining U_0 for the R-LTO mode, we apply this same factor to the remaining modes of interest and compute the lifetime as a function of temperature. These results are shown in Fig. 4. The inset of Fig. 4 shows the R-LTO linewidth as

[5]The exact strength of the coupling, as determined by the deformation potential, is controversial. The FB MC codes have yielded values of about a factor of two lower than the results used in this paper but are still of the same order of magnitude.

[6]The subscript reminds us that the DOS is for phonon pairs with energies E and $E_0 - E$.

[7]Summing over the third-order matrix elements over all normal modes yields an additional delta function of the form $\vec{q}_0 = \vec{q}_1 + \vec{q}_2 + \vec{G}$, where G is a reciprocal lattice vector. The delta functions appearing in (10) explicitly and implicitly constrain the summation over the BZ to include only the pairs of phonons which satisfy the energy- and momentum-conservation relations.

[8]We note that only for the R-LTO mode that we would be able to take advantage of the full symmetry of the crystal to reduce the size of the calculation; for a phonon of arbitrary initial wave vector, integrations must be performed over the entire BZ. However, in our analysis, we restrict our calculations to the LO phonons with an initial wave vector falling along the $\Gamma - X$ direction which allows us to reduce the volume of the integration by a factor of 1/8 of the BZ compared with 1/48 for the irreducible wedge.

[9]We made similar calculations for diamond and found that the Klemens channel is a dominant decay pathway, which is also consistent with the results of [31].

Authorized licensed use limited to: Stanford University. Downloaded on November 18, 2008 at 12:51 from IEEE Xplore. Restrictions apply.

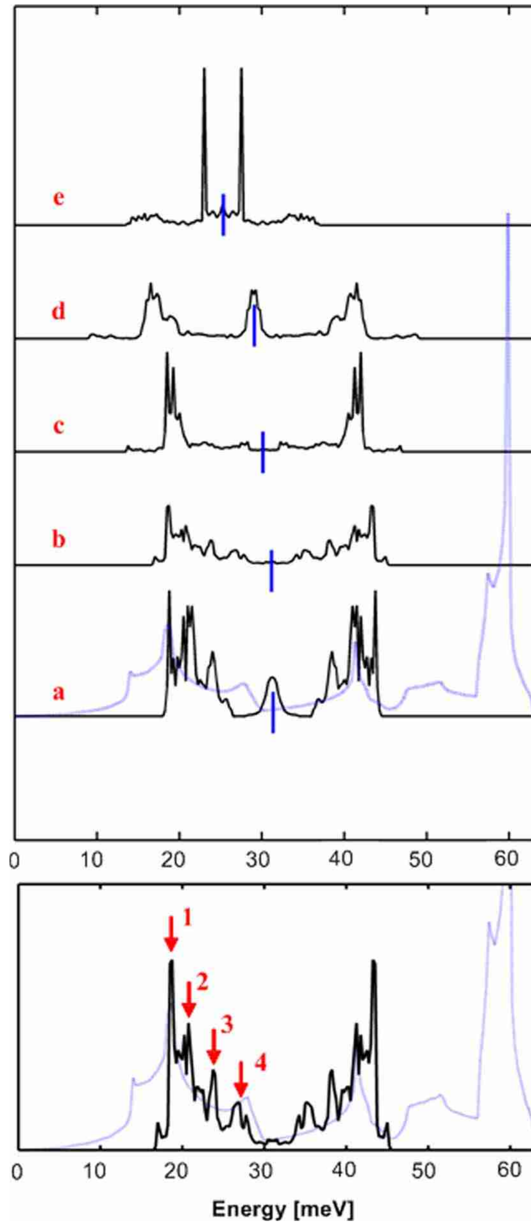

Fig. 3. $g_2(E, E_0 - E)$ for an LO phonon of initial energy E_0 and wave vector directed along $\Gamma - X$ at the points $(2\pi/a)$ $[0, 0, \alpha]$, where $\alpha = 0$, 0.3, 0.5, 0.7, and 1.0 (labeled (a)–(e), respectively), decaying into two lower energy- and crystal-momentum-conserving normal modes of energies E and $E_0 - E$. The single-phonon DOS $g(E)$ is shown in light blue for reference on top of (a). (Bottom) Zoomed-in view of the $g_2(E, E_0 - E)$ spectrum for the g-LO $(\alpha = 0.3)$ phonon. The vertical axes are in arbitrary units.

calculated using our semiempirical method which is compared with the more rigorous calculations of [31] and the experimental data of [37]. Our results are within 10% of the experimental data over a range of nearly 0 K–700 K using the single fit parameter U_0. The success of this calculation supports the conclusion that the third-order matrix-element magnitudes are only weakly dependent on the wave vector. From these results, we see that the g-LO mode is estimated to be around 8.5 ps near 0 K and reduces to about 5 ps at room temperature. Over any operating temperature typical of integrated circuits, the g-LO lifetime is seen to be about twice that of the R-LTO mode but is still of the same order of magnitude. At high power densities, this lifetime is expected to decrease on account of higher phonon occupation. The temperature dependence of the lifetimes arises from the occupation factors in (10) which are given by the B–E distribution at equilibrium. For the nonequilibrium conditions, it is straightforward to compute these lifetimes provided that we know the appropriate nonequilibrium occupation factors to apply.[10]

IV. Fully Coupled Electron–Phonon Transport

To model the effects of self-heating in a transistor, the electron and phonon systems must be fully coupled together. It is not sufficient for the electron–phonon scattering to be simply included in an electron-transport model since the phonons that are generated during the simulation are not "sensed" by the simulated electrons. There needs to be a way to feed the updated occupation numbers back into the calculation of the e–p scattering rates. Furthermore, the phonons generated during the simulation must be allowed to propagate and decay as they would in a real device, as discussed in Section III. The complexity and magnitude of such a task has prevented truly rigorous solutions of the coupled transport physics at such length and time scales. Various approximations in either the electron or phonon models are typically necessary to make the problem tractable. In [38]–[40], either moments of the phonon BTE or the use of a ballistic-diffusive form of the BTE using major simplifications in the electron and phonon dispersions was performed. Lake and Datta [41] used a nonequilibrium Green's function formalism to study a detailed energy transfer between the electrons and the phonons in a mesoscopic diode. As discussed briefly in [7] and [42], we have chosen to fully couple the electron and phonon populations by combining the e-MC technique described in Section II with the p-SFBTE described in Section III and by solving the two transport problems in sequential iteration. Having been provided with the simplifications made in modeling both the electron and phonon systems as well as in the solution techniques for solving the phonon BTE, we are able to examine the coupled transport physics for realistic devices while retaining valuable physical insights into the spectral decomposition of the heat within the device at all segments of the calculation. For each iteration, two independent simulations are performed: one for the electron system and the other for the phonon system. Outputs from each simulation are fed back into the other, and the simulation proceeds until satisfactory convergence is achieved. We find that this method typically achieves convergence within five iterations. The coupled simulation begins with an isothermal (300 K) e-MC simulation whose initial conditions are given by a drift-diffusion device simulator such as MEDICI. The e-MC computes electron transport self-consistently with the electric field by solving the Poisson equation at all steps across the device grid. Net-phonon-generation rates as a function of position and phonon frequency are gathered from the e-MC

[10]The simplicity of the semiempirical methodology enables an efficient computation of modal lifetimes by integrating the occupation number-weighted DOS over frequency. This would enable one to perform self-consistent phonon-transport studies efficiently.

Authorized licensed use limited to: Stanford University. Downloaded on November 18, 2008 at 12:51 from IEEE Xplore. Restrictions apply.

TABLE II
ENERGY AND REPRESENTATIVE WAVE VECTORS FOR FOUR DOMINANT LA + TA PHONON PAIRS WHICH A g-TYPE LO PHONON MAY CREATE DURING SPONTANEOUS DECAY WHILE CONSERVING ENERGY AND CRYSTAL MOMENTA.

	Initial	Pair I		Pair II		Pair III		Pair IV	
	LO	LA	TA	LA	TA	LA	TA	LA	TA
$q_x/(2\pi/a)$	0.0	0.95	-0.95	0.75	-0.75	0.63	-0.63	0.53	-0.53
$q_y/(2\pi/a)$	0.0	0.23	-0.23	0.48	-0.48	0.18	-0.18	0.05	-0.05
$q_z/(2\pi/a)$	0.3	-0.13	-0.43	0.25	0.05	0.53	-0.23	0.68	-0.38
Energy [meV]	62	44	18	41	21	38	24	35	27

A relatively course tetrahedral q-space grid, with a mesh size of $0.1(2\pi/a)$ filling 1/8 of the BZ volume, was used for these calculations. Energy uncertainty (σ_e) arising from the finite resolution of the BZ grid was determined to be less than 1.5 meV.

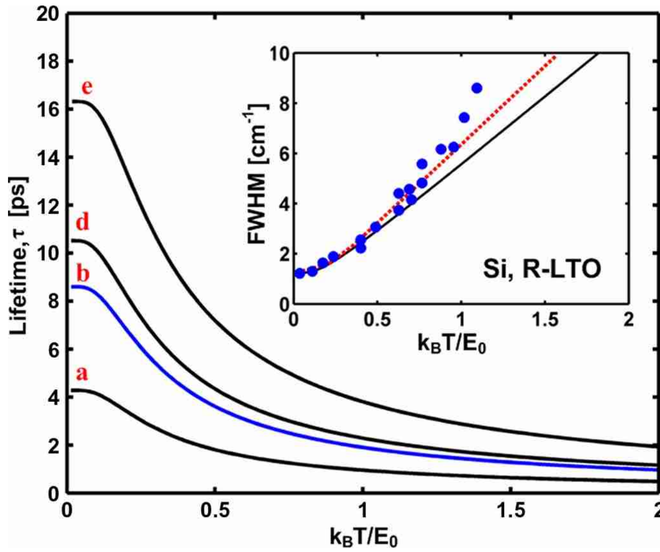

Fig. 4. Calculated lifetime for the LO phonon modes along $\Gamma-X$. The inset shows the calculated spectral full-width at half-maximum for the R-LTO mode calculated using (black solid) semiempirical approach and (red-dotted curve) the DFT calculations of [31] compared with (solid blue dots) the experimental data [37]. Note that the axis ranges for the inset are the same as that shown in [37, Fig. 2].

and fed into the p-SFBTE portion of the code. In the latter, the phonons are allowed to propagate in the absence of the electron system, and only the p–p and p–i scatterings are included. The scattering rates for the acoustic modes are determined at the beginning of the p-SFBTE simulation using (9) of Section III. The temperature dependence of the p–p scattering rates is treated by using the temperature field calculated from the classical heat diffusion equation and by taking the volumetric power-generation rate from the e-MC simulation as the source term. Because of a lack of experimental data, we have typically assigned a single lifetime for all optical modes to be in the range of 0.5–10 ps which is consistent with the Raman data and the theoretical calculations. At the end of the p-SFBTE calculation, an updated distribution of phonons as a function of position is computed. This distribution of phonons is then used to compute the electron–phonon scattering rates for the subsequent e-MC simulation by updating the phonon occupation numbers. It can be impractical to use separate phonon occupation numbers for all modes and branches and for all grid points in computing the scattering rates. Therefore, we typically compute effective temperatures for the dominant optical f- and g-type phonons as well as for the LA and TA branches. In all, we compute and pass six position-dependent effective temperature vectors back to the e-MC. These temperature vectors are then used to adjust the scattering rates in a manner that we will discuss shortly. Before doing so, we note that aside from the added complexity and the reduction in computational speed, there is nothing fundamentally preventing the use of additional temperatures to account for the occupation of individual phonon modes or ranges of phonon modes during the scattering-rate calculations.

The maximum scattering rate for each electron–phonon scattering type is then calculated using the corresponding maximum effective temperature, and the simulation begins. To include the dependence of the local phonon occupation, we then employ a temperature-based (or rather occupation-number-based) rejection algorithm, a technique that is commonly used in the Monte Carlo technique [8]. When an electron at a grid location r_i is chosen to scatter with a particular phonon of type j, the local effective temperature $T_{\mathrm{eff},j}(r_i)$ is compared with the maximum effective temperature $T_{\mathrm{eff},j,\mathrm{MAX}}$ via

$$\eta_j(r_i) = \frac{N_{q,s}\left(T_{\mathrm{eff},j}(r_i)\right) + \frac{1}{2} \mp \frac{1}{2}}{N_{q,s}(T_{\mathrm{eff},j,\mathrm{MAX}}) + \frac{1}{2} \mp \frac{1}{2}} \tag{11}$$

where $0 \leq \eta_j(r_i) \leq 1$. A random number X with a uniform probability density over the unit interval [0, 1] is then generated, and if $\eta_j(r_i) < X$, then the scattering event is allowed to take place. Otherwise, it is rejected, and the electron continues on its initial trajectory unperturbed, i.e., is treated as a self-scattering event. With the phonon-generation source-term output from the e-MC and the electron–phonon scattering rates being adjusted to account for phonon occupation (via the effective temperature and the rejection algorithm), the electrons and the phonons form a closed-loop system. We now discuss the application of this algorithm to the simulation of a simple 1-D silicon device.

Fig. 5 shows a 1-D n+/n/n+ silicon device along with its electrical characteristics which we have used extensively in developing the e-MC, the p-SFBTE, and the coupled-simulation algorithms [7], [28], [42]. Although the device is infinite in extent in the transverse plane and lacks a gate terminal, such a device structure resembles the core of transistor structures such as the FinFET. The band diagram along the channel is similar to that along the channel of typical CMOS devices, particularly the DG devices where vertical (transverse) symmetry

Authorized licensed use limited to: Stanford University. Downloaded on November 18, 2008 at 12:51 from IEEE Xplore. Restrictions apply.

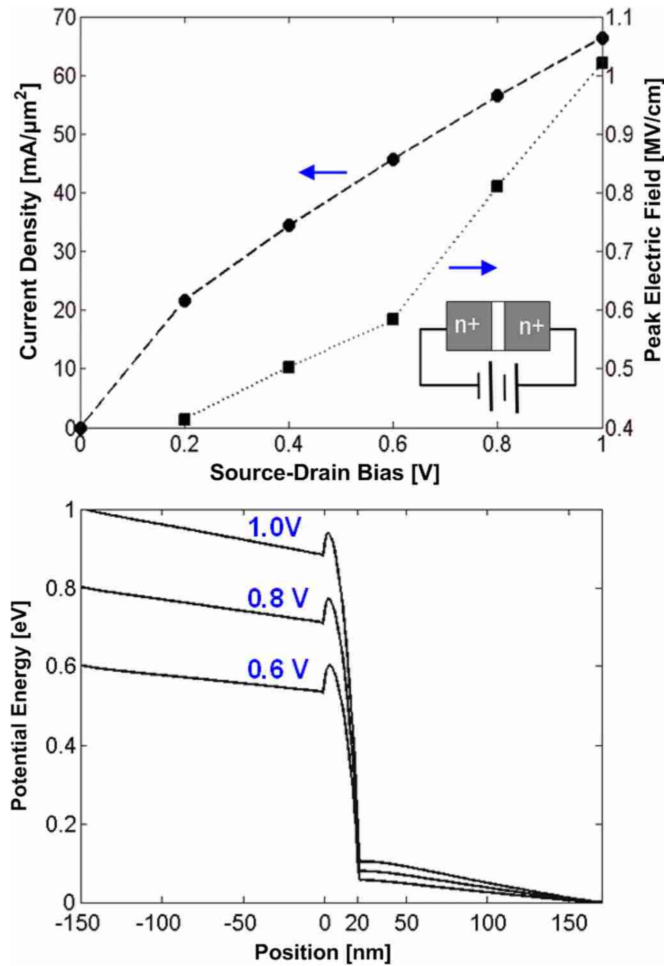

Fig. 5. Electrostatic characteristics of the 1-D n+/n/n+ device simulated in this paper. The device has three regions: two 150-nm $\mathrm{n}+(10^{20}\ \mathrm{cm}^{-3})$ source and drain regions separated by a 20-nm n-type $(10^{16}\ \mathrm{cm}^{-3})$ "channel." The doping is uniform within each region, and 1.25 nm/decade characterizes the doping concentration rolloff between the regions. Top: (Left axis) Current density and (right axis) peak electric field versus source–drain bias voltage. Bottom: Electron potential energy versus position within the device for three bias conditions.

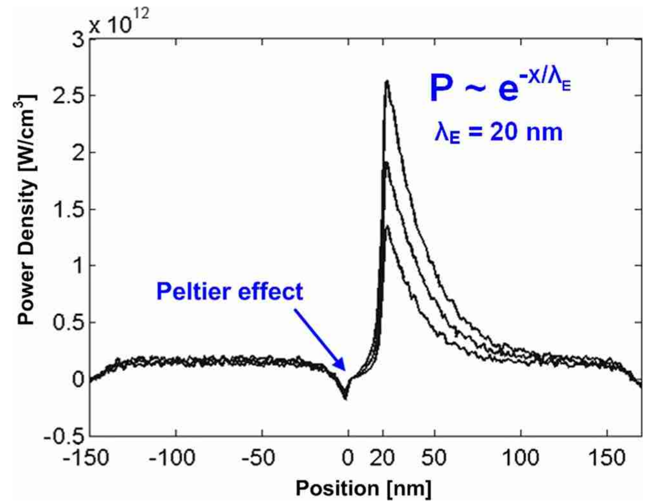

Fig. 6. Volumetric power-generation rate versus position within the 1-D n+/n/n+ device. The power generation at each grid point was computed based on (3).

exists. Additionally, this device structure allows us to extract physical insight into the energy-relaxation processes which can be applied to more sophisticated 3-D device structures.

The device consists of three regions. Two 150-nm "source/drain" regions are doped to $10^{20}\ \mathrm{cm}^{-3}$ and are separated by a 20-nm lightly doped $(10^{16}\ \mathrm{cm}^{-3})$ "channel" region.[11] For thermal boundary conditions, the departure from equilibrium phonon population $n_{q,s}$ (5) is taken to be zero for all modes at the left and right contacts, i.e., the contacts are treated as perfect thermal reservoirs. Furthermore, the temperature T_F, which determines the distribution of the near-equilibrium phonon population, is set to 300 K at both contacts, and its spatial derivative is continuous.

Fig. 6 shows the steady-state power density generated within the device at three different bias conditions (0.6, 0.8, and 1.0 V), which was computed using (3) of Section II. As discussed in [28], nearly all of the power generation within this device structure occurs within the drain, which is a common characteristic for transistors operating in the quasi-ballistic transport regime. Furthermore, the generation profile is exponentially decaying with a characteristic length of about 20 nm which is essentially independent of applied drain voltage. This lack of dependence of drain voltage was attributed to the fact that the average electron velocity, as well as the electron–phonon scattering rates, both scale approximately as $\sim(E-E_g)^{1/2}$, and thus, the energy-relaxation length remains essentially constant. The peak average kinetic energy gained by the electrons across the channel is found to be proportional to the drain-to-source voltage $(E-E_g)\sim 0.4\ (\mathrm{eV_{ds}})$ [28]. The impact on energy-relaxation length by a fully silicided drain located within 20 nm of the channel/drain boundary has not been investigated. However, because of the amorphous nature of the silicide, the scattering length is expected to be on the order of the disorder length, which is $\sim$1 nm. In the case that the fin extension is less than 20 nm, the energy-relaxation length will likely be reduced and will be comparable to the fin-extension length itself. This is why we make the assumption that all of the power is dissipated in the fin extension in arriving at the values shown in Fig. 1.

Fig. 7 shows the phonon-generation spectra computed at four positions within the drain region beginning with the peak power-generation point. As the electron system continues to cool deeper into the drain, the phonon emission spectrum becomes more concentrated about the f- and g-type phonons.

The upper panel of Fig. 8 shows the profiles of the effective temperature for the entire phonon population. Near the hotspot, the temperature is dominated by the effects of ballistic transport, and an anomalous temperature rise is observed. Beyond about 50 nm into the drain, the temperature profiles are predicted well by (dashed lines) the classical heat diffusion equation. In the bottom panel of Fig. 8, we compare the effective temperatures for the LO branch and the isolated g-LO mode obtained by appropriately restricting the integration in (8). From this figure, we can see how poorly an effective

[11] The choice of 150-nm source/drain regions was made in order to ensure that the contacts do not significantly affect the transport near the device hotspot.

Authorized licensed use limited to: Stanford University. Downloaded on November 18, 2008 at 12:51 from IEEE Xplore. Restrictions apply.

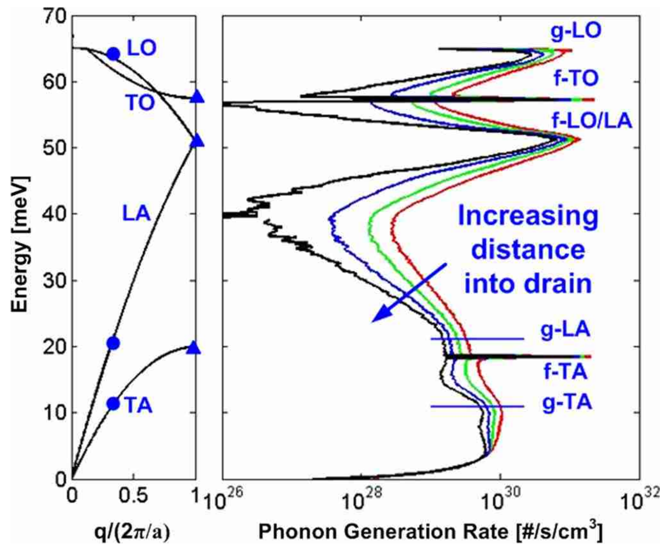

Fig. 7. (Left) Phonon dispersion and (right) phonon-generation spectra computed at four locations within the device for the 1-V case. The red (right-most) curve corresponds to the location of the peak power dissipation. The remaining curves correspond to $r = 10$, 20, and 30 nm displaced from the peak generation point or hotspot within the drain.

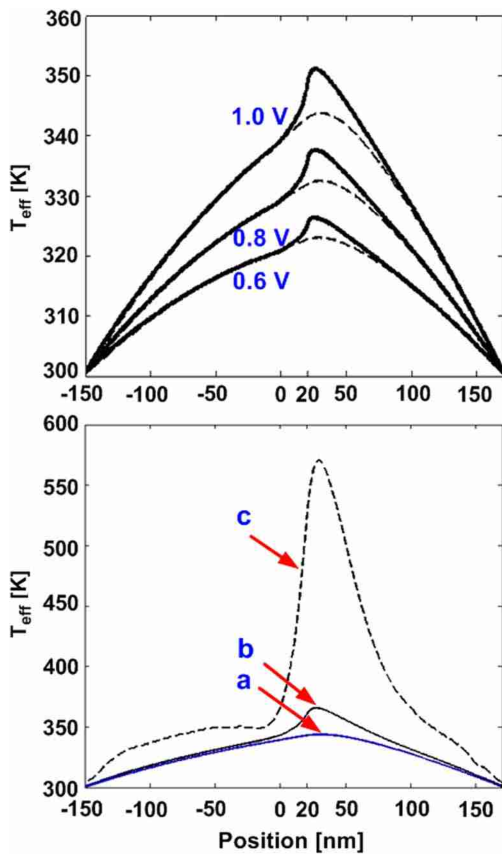

Fig. 8. (Top) Effective lattice temperature ($T_{\rm eff}$) computed using the self-consistent e-MC/p-SFBTE method. (Bottom) $T_{\rm eff}$ for (solid black—b) the LO phonon branch compared with that of (dashed black—c) the g-LO phonon for the 1-V bias condition. The temperature computed from (solid blue—a) the classic heat diffusion equation is also shown for reference.

LO branch temperature would describe the large departure from equilibrium exhibited by certain individual modes within the LO branch.

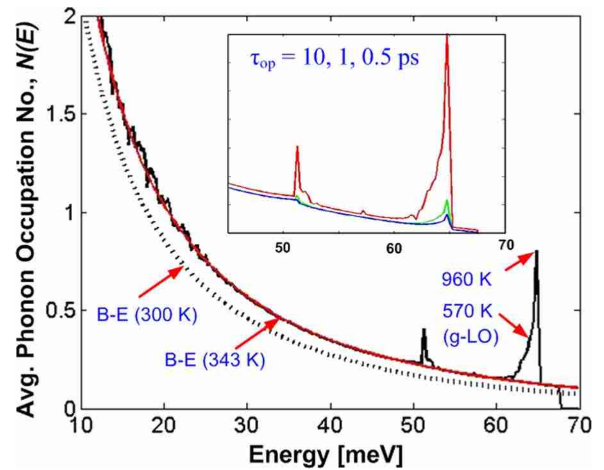

Fig. 9. Phonon distribution at peak power-generation point ($r = 25$ nm). The inset shows a zoomed-in view of the optical-mode occupation as a function of the optical phonon lifetime ($\tau_{\rm op} = 0.5$, 1, and 10 ps).

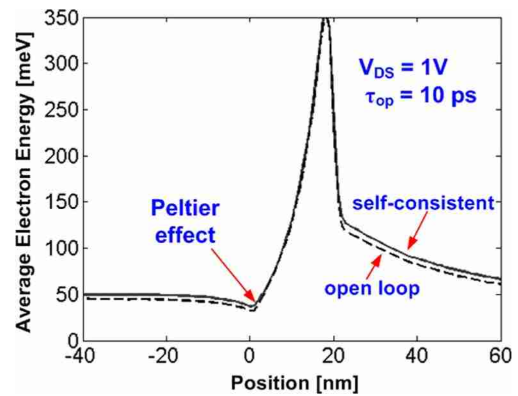

Fig. 10. Average electron energy as a function of position for an optical phonon lifetime of 10 ps for (dashed line) uncoupled and (solid line) fully coupled/closed-loop simulations.

The phonon energy distribution at the peak power-generation point in the device is shown in Fig. 9. The acoustic phonons (< 51 meV) are well behaved in that they closely adhere to the B–E distribution of elevated temperature $T_{\rm eff,(LA,TA)} = 343$ K. However, the optical phonons (> 51 meV) exhibit significant departures from the B–E equilibrium distribution. Temperatures near the zone center approach 1000 K, and the g-LO phonon peaks at about 570 K for an optical phonon lifetime ($\tau_{\rm op}$) of 10 ps. The inset of Fig. 9 shows the dependence of the phonon distribution on the optical phonon-lifetime parameter as it is varied from 0.5 to 10 ps. The deviation of the optical phonon modes from the equilibrium distribution diminishes in direct proportion to the phonon lifetime, and the B–E distribution describes the entire phonon population well for $\tau_{\rm op} < 1$ ps for the power densities simulated.

Finally, the impact on average electron energy is shown in Fig. 10. The average energy increases in both the source and drain regions, but there is little effect in the channel region, which is a result of the near-ballistic transport across the channel. Furthermore, despite an appreciable temperature rise within the device, it is found that the device current is reduced by only about 1% at 1 V. The reduction in drive current is attributed to an increase in the drain scattering, which leads to

Authorized licensed use limited to: Stanford University. Downloaded on November 18, 2008 at 12:51 from IEEE Xplore. Restrictions apply.

a slight increase in the channel barrier height and, ultimately, a reduction in the electron source injection rate. Although we did not include it here, boundary scattering would play an important role in determining the near-equilibrium phonon temperature T_F for a thin-body device such as the FinFET. The effective thermal conductivity for a 22-nm thin film of undoped single-crystalline silicon was measured to be only 20 $\mathrm{Wm^{-1}K^{-1}}$ near room temperature, about an order of magnitude lower than the bulk-undoped value of 148 $\mathrm{Wm^{-1}K^{-1}}$ which was used in this paper. A 10-nm film is expected to be reduced to just 13 $\mathrm{Wm^{-1}K^{-1}}$ [43] as the conductivity scales approximately as $\delta = d_s/\Lambda_b$ [20], where d_s is the film thickness, and Λ_b is the bulk phonon mean-free-path ($\sim$100 nm). The consequences of the thermal boundary scattering for ultrathin-body silicon and germanium devices were recently discussed by Pop *et al.* [43]. Reductions in the effective thermal conductivity for the silicon layer in quasi-1-D devices, such as FinFETs, are expected to be even more severe than the 2-D films.

V. Discussion

In this section, we discuss some of the consequences of nanoscale and nonequilibrium transport which were discussed in detail within the previous three sections on leading edge and future thin-body silicon FET devices. First, we address the origins of what various researchers have referred to as the "size effect" of the heat-generation source in bulk materials.[12] For realistic materials, there may be an increase in a local temperature, as we have defined in (8), relative to the temperature field which would be calculated using diffusive-conduction equations, i.e., Fourier's law. This was quite evident in Fig. 8 for the 1-D device. This rise in effective temperature is the consequence of a skewed phonon distribution near the heat-generation source (Fig. 10) which tends to favor higher energy modes (with lower group velocities) at high power densities. The additional thermal resistance, which leads to this increased temperature rise near the source, is directly related to the energy-relaxation processes between the optical and acoustic phonons, as discussed in Section III. This additional thermal resistance disappears with a vanishing optical phonon lifetime, and the size of the source does not dictate the magnitude of the anomalous temperature rise, unlike what has been hypothesized [19], [44], [45].

The anomalous increase in the effective temperature within the first 20–50 nm of the channel/drain boundary acts to reduce the device current and may have an appreciable impact on the reliability of the transistor. Despite the ballistic nature of nanometer-scale devices, the drain and the source are still electrically coupled. To maintain current continuity, the potential barrier seen by the source electrons will increase in the presence of increased scattering in the drain which acts to reduce the source injection rate. In contrast, an anomalous temperature rise at the source-channel boundary acts to increase the electron injection rate. For the device studied in Section IV, we observed a reduction in the drive current by about 1%, which is a rather weak effect. Although direct evidence of an anomalous increase in the device temperature is lacking, there are some preliminary indications that transistor reliability may be impacted by the effects of the nonequilibrium phonon populations [46].

Steady-state leakage currents are not expected to be influenced by the nonequilibrium effects. This is because the applied voltage across the channel, which is essential for initiating the nonequilibrium phonon population in the first place, vanishes once a switching event has taken place over the course of a few picoseconds. However, the short-circuit leakage current, which arises during the brief fraction of a logic gate switching event in which both the pull-up and the pull-down devices are simultaneously conducting, could conceivably lead to a situation where the leakage currents are temporarily enhanced by the nonequilibrium phonons in the source(s) of the NMOS device(s). Take for example a standard CMOS inverter. A nonequilibrium population of phonons could be generated within the NMOS device as the PMOS device is being turned on, and the short-circuit current dissipates energy in both devices. Now, with the drain voltage at the positive supply rail after the switching event is complete, there remains a driving force for the drain–source leakage current of the NMOS transistor. The time for these nonequilibrium populations to decay would be comparable with the optical phonon decay time < 5 ps. In theory, an elevated phonon population in the drain could be observed through a time-resolved near-field Raman spectroscopy measurement with the local substrate removed. Additionally, if the timing jitter of quantum-limited near-IR single photon detectors could ever be reduced to subpicosecond levels, a measurement of the near-IR light emission strength of an NMOS transistor in a CMOS pair, which is directly proportional to the leakage current, could be used to measure the increase of a nonequilibrium phonon population near the source injection point [47].

Highly nonequilibrium optical phonon populations have the potential to significantly impact the electrical transport and impede the thermal conduction near the hotspot. The effects of hot optical phonons have been widely observed in III–V semiconductor devices through the observation of extended energy-relaxation times in ultrafast optical measurements [48] as well as through the excess noise at microwave frequencies [49]. Recently, negative differential conductance in suspended two-terminal, metallic, and single-wall carbon-nanotube devices was attributed to the hot optical phonons [50]. There, it was postulated that the optical phonon lifetime was significantly enhanced by lack of a substrate. A large population of the nonequilibrium optical phonons in the drain of a transistor may cause the energy-relaxation length to extend beyond the 20-nm characteristic length calculated in this paper. Furthermore, the increased scattering rate would lead to a marked reduction in mobility in the drain and perhaps even lead to negative conductance effects. In order for these types of effects to take place, the optical phonon occupation numbers need to be comparable to or larger than unity such that the absorption processes are comparable to the emission processes. As discussed in Section III, the g-LO phonon has been flagged as a good

[12] We wish not to confuse this topic with the "size effect" associated with the boundary scattering in low-dimensional semiconductors. To be clear, we are talking about an infinite medium in which the heat source is much smaller than the mean-free-path.

Authorized licensed use limited to: Stanford University. Downloaded on November 18, 2008 at 12:51 from IEEE Xplore. Restrictions apply.

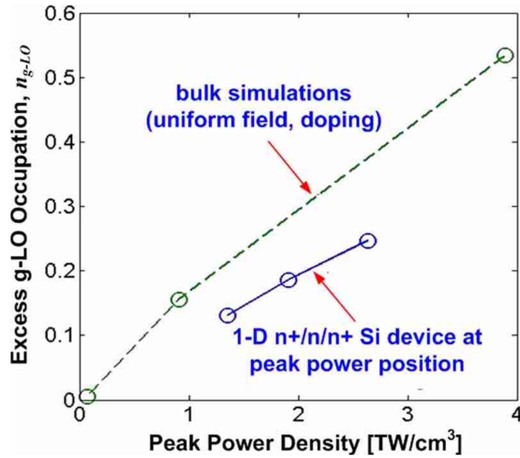

Fig. 11. Excess occupation number for the g-type longitudinal-optical (g-LO) phonon as a function of volumetric power density as calculated for (upper dashed-line data set) a uniform power generation and as calculated at (lower solid-line data set) the peak power-generation point in the 1-D device.

indicator for hot phonons in silicon. A simple expression for the excess occupation number, which is derived from a single mode rate equation, is given by

$$n_{g-\mathrm{LO}} \approx \frac{Q'''_{g-\mathrm{LO}}\tau_{g-\mathrm{LO}}}{\hbar\omega_{g-\mathrm{LO}}g(\omega_{g-\mathrm{LO}})\Delta\omega} \quad (12)$$

where $Q'''_{g\text{-LO}}$ is the volumetric power being transferred to the g-LO phonon by the electrons in the spectral width $\Delta\omega$, and τ_{g-LO} is the modal relaxation time, which is estimated to be about 5 ps at room temperature. This expression is valid only under homogeneous power-generation conditions and tends to be an overestimate of the population by a factor of about two. Fig. 11 shows the excess occupation number for the g-LO mode as a function of the volumetric power density for bulk silicon as well as at the hotspot of the 1-D device of Section IV. We find that the excess occupation number increases approximately as $n_{g-\mathrm{LO}} \sim 1 \times 10^{-10} Q'''^{3/4}_{\mathrm{tot}}$, where Q'''_{tot} is the total volumetric power density measured in watts per cubic centimeter. By reducing this number by a factor of 1/2 to account for nonuniform heat generation, we find that the nonequilibrium occupation number will reach unity for a power density of about 20 TW/cm^3. This assumes that the phonon lifetime does not decrease with an increasing power density. If we assume that the phonon lifetime is reduced to subpicosecond levels at high power densities, as some recent experimental work has indicated it as may be likely [51], then the critical power density would exceed 200 TW/cm^3. According to our estimates in Fig. 1, this level is just beyond the projected power density for well-behaved end-of-the-roadmap FinFET devices. Since the transistor volumetric power density is sensitive to the fin (body) thickness, statistical process variations may cause a significant fraction of the transistors across a chip to cross into the regime where nonequilibrium heating plays a significant role in the electrical transport. This may lead to an enhancement in the variation of performance across die and wafer. In addition to a critical volumetric power density, we can estimate a critical electric field by relating the peak power density for ballistic transport to the peak power density given by the classical Joule heating term as $\max\{Q'''_{\mathrm{ballistic}}\} \sim \max\{\vec{J} \cdot \vec{E}\}/\gamma$, where γ can be taken to be ∼2–3 [28]. By assuming that the current density is given by $|\vec{J}| \sim \beta/L_g$ μA/nm^2, where the magnitude of β is equivalent to the number of gate electrodes (also, ∼2–3) times 1 μA/nm, we arrive at an expression for the critical electric field $|\vec{E}|_{\mathrm{crit}} \sim (\gamma/\beta)\max\{Q'''_{\mathrm{ballistic}}\} \cdot L_g$. For a 10-nm device, a critical power density of 200 TW/cm^3 would correspond to a peak electric field of about 20 MV/cm, which is about 40× the field strength associated with dielectric breakdown in silicon.

Finally, we address the consequences of our choice in using a simplified analytic NPB model. As we mentioned briefly in Section II, the intervalley deformation potentials used in this paper are about twice the values computed using the FB models [12], [16]. This is despite the fact that the NPB model reproduces the electron DOS reasonably well up to about 1.5 eV. Fischetti and Laux [12] attributed the smaller deformation potentials of the FB model to the ability of the electron population to reach a lower energy configuration through equivalent valley repopulation by means of an enhanced contribution from Bloch oscillation, which the NPB model fails to describe adequately. By using the NPB model, it was shown that the ratio of energy being dissipated by the optical and acoustic phonon modes is ∼2 : 1 [17]. However, the FB MC codes have shown the reverse ratio of ∼1:2. To account for the deficiencies of the NPB model, we can somewhat crudely add an additional factor of 1/2 to (12) and increase the estimates for the onset of the hot optical phonons to the power densities approaching 500 TW/cm^3. The FB calculations would undoubtedly improve the accuracy of these estimates, but they should not change one of the more important conclusions of this paper, that hot optical phonons are unlikely to play a significant role in the electrical behavior of well-behaved silicon devices described by the existing technology roadmap.

VI. Conclusion

Transistor designs over the next decade will feature confined geometries with increasing surface-to-volume ratios and rising volumetric power densities. Thermal conductance within the transistor will be dramatically reduced due to increased surface scattering and by the confinement of intrinsically low-conductivity materials. Furthermore, thermal boundary resistance at interfaces between dissimilar materials, including Si and SiGe alloys, will lead to higher junction temperatures and, therefore, higher leakage currents than what may be expected. Reliability will be impacted by the increase in junction temperatures caused by lower thermal conductance and may also be influenced by the nonequilibrium optical phonons in the drain. Mobility reduction due to hot optical phonons does not seem to be a major threat to the near-term evolution of CMOS technology. However, this assessment only stands for well-behaved devices operating under ideal conditions through the end of the current roadmap. In some analog or high-power applications, the hot phonons could play a major role. Based on this work, we estimate that the hot phonon effects will not play a significant role for power densities below about

Authorized licensed use limited to: Stanford University. Downloaded on November 18, 2008 at 12:51 from IEEE Xplore. Restrictions apply.

100 TW/cm^3 for silicon-based devices if we consider optical phonon lifetimes to reduce below 1 ps under typical operating conditions. For silicon-based quantum-well devices, the models used in this paper will need to be modified appropriately, and the issue of hot phonons will need to be reevaluated. Our conclusions are largely dependent on the optical phonon-lifetime parameter. Therefore, experimental validation of the optical phonon lifetimes for realistic device conditions is essential.

ACKNOWLEDGMENT

The authors would like to thank E. Pop, S. Sinha, and M. Panzer for their many contributions to this paper and for the countless discussions on hot-carrier transport in semiconductors.

REFERENCES

[1] R. Chau *et al.*, "Silicon nano-transistors and breaking the 10 nm physical gate length barrier," in *Proc. IEEE Device Res. Conf.*, 2003, pp. 123–126.
[2] H. Xuejue *et al.*, "Sub 50-nm FinFET: PMOS," in *IEDM Tech. Dig.*, 1999, pp. 67–70.
[3] R. Chau *et al.*, "A 50 nm depleted-substrate CMOS transistor (DST)," in *IEDM Tech. Dig.*, 2001, pp. 29.1.1–29.1.4.
[4] B. Doyle *et al.*, "Tri-gate fully-depleted CMOS transistors: Fabrication, design and layout," in *VLSI Symp. Tech. Dig.*, 2003, pp. 133–134.
[5] *International Technology Roadmap for Semiconductors (ITRS) 2005 Update*, 2005. [Online]. Available: http://www.itrs.net/Links/2005ITRS/Home2005.htm
[6] E. Pop, R. W. Dutton, and K. E. Goodson, "Analytic band Monte Carlo model for electron transport in Si including acoustic and optical phonon dispersion," *J. Appl. Phys.*, vol. 96, no. 9, pp. 4998–5005, Nov. 2004.
[7] J. Rowlette *et al.*, "Thermal simulation techniques for nanoscale transistors," in *Proc. IEEE/ACM ICCAD*, 2005, pp. 225–228.
[8] C. Jacoboni and L. Reggiani, "The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials," *Rev. Mod. Phys.*, vol. 55, no. 3, pp. 645–705, Jul. 1983.
[9] M. V. Fischetti, S. E. Laux, P. M. Solomon, and A. Kumar, "Thirty years of Monte Carlo simulations of electronic transport in semiconductors: Their relevance to science and to mainstream VLSI technology," in *Proc. 10th Int. Workshop Comput. Electron.*, West Lafayette, IN, Oct. 24–27, 2004, pp. 47–48.
[10] T. B. Boykin, G. Klimeck, and F. Oyafuso, "Valence band effective-mass expressions in the $Sp^3d^5s^*$ empirical tight-binding model applied to a Si and Ge parametrization," *Phys. Rev. B, Condens. Matter*, vol. 69, no. 11, p. 115 201, Mar. 2004.
[11] P. D. Yoder and K. Hess, "First-principles Monte Carlo simulation of transport in Si," *Semicond. Sci. Technol.*, vol. 9, no. 5S, pp. 852–854, May 1994.
[12] M. V. Fischetti and S. E. Laux, "Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects," *Phys. Rev. B, Condens. Matter*, vol. 38, no. 14, pp. 9721–9745, Nov. 1988.
[13] T. Kunikiyo *et al.*, "A Monte Carlo simulation of anisotropic electron transport in silicon including full band structure and anisotropic impact-ionization model," *J. Appl. Phys.*, vol. 75, no. 1, pp. 297–312, Jan. 1994.
[14] R. Tubino, L. Piseri, and G. Zerbi, "Lattice dynamics and spectroscopic properties by a valence force potential of diamondlike crystals: C, Si, Ge, and Sn," *J. Chem. Phys.*, vol. 56, no. 3, pp. 1022–1039, Feb. 1972.
[15] D. Long, "Scattering of conduction electrons by lattice vibrations in silicon," *Phys. Rev.*, vol. 120, no. 6, pp. 2024–2032, Dec. 1960.
[16] A. Abramo *et al.*, "A comparison of numerical solutions of the Boltzmann transport equation for high-energy electron transport silicon," *IEEE Trans. Electron Devices*, vol. 41, no. 9, pp. 1646–1654, Sep. 1994.
[17] E. Pop, R. W. Dutton, and K. E. Goodson, "Monte Carlo simulation of Joule heating in bulk and strained silicon," *Appl. Phys. Lett.*, vol. 86, no. 8, pp. 1–3, Feb. 2005.
[18] E. Pop, S. Sinha, and K. E. Goodson, "Heat generation and transport in nanometer-scale transistors," in *Proc. IEEE*, Aug. 2006, vol. 94, pp. 1587–1601.
[19] S. Sinha, E. Pop, R. W. Dutton, and K. E. Goodson, "Non-equilibrium phonon distributions in sub-100 nm silicon transistors," *Trans. ASME, J. Heat Transf.*, vol. 128, no. 7, pp. 638–647, Jul. 2006.
[20] C. T. M. Flik, "Size effect on the thermal conductivity of high-Tc thin-film superconductors," *Trans. ASME, J. Heat Transf.*, vol. 112, no. 4, pp. 872–881, 1990.
[21] M. Asheghi, Y. K. Leung, S. S. Wong, and K. E. Goodson, "Phonon-boundary scattering in thin silicon layers," *Appl. Phys. Lett.*, vol. 71, no. 13, pp. 1798–1800, Sep. 1997.
[22] M. Asheghi, K. Kurabayashi, R. Kasnavi, and K. E. Goodson, "Thermal conduction in doped single-crystal silicon films," *J. Appl. Phys.*, vol. 91, no. 8, pp. 5079–5088, Apr. 2002.
[23] P. G. Klemens, "The thermal conductivity of dielectric solids at low temperatures (theoretical)," in *Proc. R. Soc. Lond. A, Math. Phys. Sci.*, Aug. 1951, vol. 208, pp. 108–133.
[24] J. Callaway, "Model for lattice thermal conductivity at low temperatures," *Phys. Rev.*, vol. 113, no. 4, pp. 1046–1051, Feb. 1959.
[25] M. G. Holland, "Analysis of lattice thermal conductivity," *Phys. Rev.*, vol. 132, no. 6, pp. 2461–2471, Dec. 1963.
[26] M. Artaki and P. J. Price, "Hot phonon effects in silicon field-effect transistors," *J. Appl. Phys.*, vol. 65, no. 3, pp. 1317–1320, Feb. 1989.
[27] J. Shah, A. Pinczuk, A. C. Gossard, and W. Wiegmann, "Hot carrier energy loss rates in GaAs quantum wells: Large differences between electrons and holes," *Physica B+C*, vol. 134, no. 1–3, pp. 174–178, Nov. 1985.
[28] E. Pop, J. A. Rowlette, R. W. Dutton, and K. E. Goodson, "Joule heating under quasi-ballistic transport conditions in bulk and strained silicon devices," in *Proc. Int. Conf. Simulation Semicond. Processes Devices*, Tokyo, Japan, Sep. 1–3, 2005, pp. 307–310.
[29] S. Narasimhan and D. Vanderbilt, "Anharmonic self-energies of phonons in silicon," *Phys. Rev. B, Condens. Matter*, vol. 43, no. 5, pp. 4541–4544, Feb. 1991.
[30] S. Sinha, P. K. Schelling, S. R. Phillpot, and K. E. Goodson, "Scattering of g-process longitudinal optical phonons at hotspots in silicon," *J. Appl. Phys.*, vol. 97, no. 2, pp. 023 702–023 709, Jan. 2005.
[31] A. Debernardi, S. Baroni, and E. Molinari, "Anharmonic phonon lifetimes in semiconductors from density-functional perturbation theory," *Phys. Rev. Lett.*, vol. 75, no. 9, pp. 1819–1822, Aug. 1995.
[32] G. Gilat and L. J. Raubenheimer, "Accurate numerical method for calculating frequency-distribution functions in solids," *Phys. Rev.*, vol. 144, no. 2, p. 390, Apr. 1966.
[33] G. Lehmann and M. Taut, "On the numerical calculation of the density of states and related properties," *Phys. Stat. Sol. B*, vol. 54, no. 2, pp. 469–477, Dec. 1972.
[34] J. Rath and A. J. Freeman, "Generalized magnetic susceptibilities in metals: Application of the analytic tetrahedron linear energy method to Sc," *Phys. Rev. B, Condens. Matter*, vol. 11, no. 6, pp. 2109–2117, Mar. 1975.
[35] P. B. Allen, "A tetrahedron method for doubly constrained Brillouin zone integrals. Application to silicon optic phonon decay," *Phys. Stat. Sol. B*, vol. 120, no. 2, pp. 529–538, Dec. 1983.
[36] P. G. Klemens, "Anharmonic decay of optical phonons," *Phys. Rev.*, vol. 148, no. 2, p. 845, Aug. 1966.
[37] J. Menendez and M. Cardona, "Temperature dependence of the first-order Raman scattering by phonons in Si, Ge, and alpha-Sn: Anharmonic effects," *Phys. Rev. B, Condens. Matter*, vol. 29, no. 7, pp. 2051–2059, Feb. 1984.
[38] J. Lai and A. Majumdar, "Concurrent thermal and electrical modeling of sub-micrometer silicon devices," *J. Appl. Phys.*, vol. 79, no. 9, pp. 7353–7361, May 1996.
[39] A. Majumdar, K. Fushinobu, and K. Hijikata, "Effect of gate voltage on hot-electron and hot-phonon interaction and transport in a submicrometer transistor," *J. Appl. Phys.*, vol. 77, no. 12, pp. 6686–6694, Jun. 1995.
[40] P. G. Sverdrup, Y. S. Ju, and K. E. Goodson, "Sub-continuum simulations of heat conduction in silicon-on-insulator transistors," *Trans. ASME, J. Heat Transf.*, vol. 123, no. 1, pp. 130–137, Feb. 2001.
[41] R. Lake and S. Datta, "Energy balance and heat exchange in mesoscopic systems," *Phys. Rev. B, Condens. Matter*, vol. 46, no. 8, pp. 4757–4763, Aug. 1992.
[42] J. Rowlette *et al.*, "Thermal phenomena in deeply scaled MOSFETs," in *IEDM Tech. Dig.*, 2005, pp. 984–987.
[43] E. Pop, R. Dutton, and K. Goodson, "Thermal analysis of ultra-thin body device scaling [SOI and FinFet devices]," in *IEDM Tech. Dig.*, Washington, DC, Dec. 8–10, 2003, pp. 36.6.1–36.6.4.
[44] G. Chen, "Nonlocal and nonequilibrium heat conduction in the vicinity of nanoparticles," *Trans. ASME, J. Heat Transf.*, vol. 118, no. 3, pp. 539–545, 1996.

Authorized licensed use limited to: Stanford University. Downloaded on November 18, 2008 at 12:51 from IEEE Xplore. Restrictions apply.

[45] G. Chen, "Ballistic-diffusive heat-conduction equations," *Phys. Rev. Lett.*, vol. 86, no. 11, pp. 2297–2300, Mar. 2001.
[46] Y. Wang, K. P. Cheung, A. S. Oates, and P. Mason, "Ballistic phonon enhanced NBTI," in *Proc. IEEE Int. Rel. Phys. Symp.*, Phoenix, AZ, 2007, pp. 258–263.
[47] S. Polonsky and K. A. Jenkins, "Time-resolved measurements of self-heating in SOI and strained-Si MOSFETs using off-state leakage current luminescence," in *Proc. Int. Semicond. Device Res. Symp.*, Washington, DC, Dec. 10–12, 2003, pp. 359–360.
[48] A. R. Vasconcellos, R. Luzzi, C. G. Rodrigues, and V. N. Freire, "Hot-phonon bottleneck in the photoinjected plasma in GaN," *Appl. Phys. Lett.*, vol. 82, no. 15, pp. 2455–2457, Apr. 2003.
[49] J. Liberis *et al.*, "Hot phonons in Si-doped GaN," *Appl. Phys. Lett.*, vol. 89, no. 20, p. 202 117, Nov. 2006.
[50] E. Pop *et al.*, "Negative differential conductance and hot phonons in suspended nanotube molecular wires," *Phys. Rev. Lett.*, vol. 95, no. 15, pp. 155 504–155 505, Oct. 2005.
[51] J. J. Letcher, K. Kang, D. G. Cahill, and D. D. Dlott, "Effects of high carrier densities on phonon and carrier lifetimes in Si by time-resolved anti-Stokes Raman scattering," *Appl. Phys. Lett.*, vol. 90, no. 25, pp. 252 103–252 104, Jun. 2007.

Jeremy A. Rowlette received the B.S.E.E. degree from the Pennsylvania State University, University Park, in 2000 and the M.S.E.E. degree from Stanford University, Stanford, CA, in 2005. He is currently working toward the Ph.D. degree in electrical engineering at Stanford University. His Ph.D. research is the study of electrothermal and optoelectronic phenomena in low-dimensional semiconductors, which is supported by an Intel Graduate Fellowship. From 2000 to 2004, his graduate studies were supported by the Intel Honors Cooperative Program.

Prior to his doctoral work, Jeremy spent five years, from 1999 to 2004, with the Intel Corporation, developing optical and ion-beam-based circuit diagnostic equipment for microprocessors and flash memory. His interests include solid-state physics and photonics, optical-electronic design, and experimental optics.

Kenneth E. Goodson received the Ph.D. degree in mechanical engineering from the Massachusetts Institute of Technology, Cambridge, in 1993.

He spent two years with the Materials Group, Daimler-Benz AG. He is a Professor of mechanical engineering with the Department of Mechanical Engineering, Stanford University, Stanford, CA, where his group studies nanoscale-transport phenomena relevant for electronic systems. His Stanford research group includes 15 students and research associates studying thermal-transport phenomena relevant for electronic systems, with a focus on those occurring with very small length and time scales. He is a Cofounder and Former Chief Technology Officer of Cooligy, which develops microfluidic cooling technology for computers. He has authored or coauthored more than 120 journal and conference papers and five book chapters.

Dr. Goodson's group has been recognized through the Office of Naval Research Young Investigator Award, the NSF CAREER Award, the Journal of Heat Transfer Outstanding Reviewer Award in 1999, a Japan Society for the Promotion of Science Visiting Professorship at the Tokyo Institute of Technology, Tokyo, Japan, in 1996, as well as the Best Paper Awards at the SEMI-THERM, the Multilevel Interconnect Symposium, and the SRC TECHCON.

Authorized licensed use limited to: Stanford University. Downloaded on November 18, 2008 at 12:51 from IEEE Xplore. Restrictions apply.