Multidimensional Modeling of Nanotransistors

M. P. Anantram and A. Svizhenko

(Invited Paper)

***Abstract*—In this paper, we review our recent work using the nonequilibrium Green's function method to model nanotransistors. After presenting a motivation for the need of quantum mechanical modeling, an account of the equations and implementation is given for both 1-D and 2-D modeling. Examples are given to highlight the use of the developed models. Finally, possible future directions in quantum mechanical modeling of transport in nanotransistors are highlighted along with computational challenges.**

***Index Terms*—Algorithm, computational electronics, contact resistance, gate leakage, Green's function, modeling, MOSFETs, nanoelectronics, nanomaterials, nanotechnology, non equilibrium quantum transport, parallel processing, phonons, quantum effects, quantum theory, review, scattering, semiconductors, silicon, simulator, software.**

I. Introduction

BALLISTIC and quantum–mechanical effects are becoming important in transistors with nanoscale channel lengths. Predictive modeling methods using physically accurate equations to capture the device physics will help both in choosing between emerging technologies and in understanding the limitations of downscaling conventional technology. More specifically, building modeling infrastructure will help throw light on the following issues.

1) Ballistic transport: The channel length of the silicon transistors is projected to approach 10 nm in the year 2015 for the 22-nm technology node [1]. This is comparable to the wavelength of an electron with thermal velocity and mean free path (due to electron–phonon scattering), which are both also around 10 nm. The concept of mobility within the context of the drift-diffusion equation is no longer well defined, and transport in the channel is quasi-ballistic.
2) Scattering from impurities: With decreasing channel lengths, the device to device fluctuations in transistor current–voltage characteristics due to scattering from impurities become important [2]. Scattering from multiple impurities will not be equivalent to the series combination of scattering from independent impurities, as assumed in semiclassical modeling.
3) Tunneling and capacitance corrections: The thickness of ultrathin silicon oxides will approach 1 nm unless new high-κ dielectrics are used. The tunneling current through oxides is expected to increase with a decreasing channel length. In addition, the location of the inversion layer in the silicon channel typically peaks around 1 nm from the oxide–silicon interface. This decreases the capacitance of the MOS structure due to a nonnegligible increase in effective oxide thickness [3].
4) Novel channel materials: In emerging prototype transistors based on nanomaterials such as carbon nanotubes and nanowires, a straightforward application of the traditional effective mass approach is of limited validity. A more accurate method based on the tight binding/*ab initio* Hamiltonian can handle a wide variety of channel materials.
5) Novel device physics: Nanotransistors may exploit band to band tunneling (between conduction and valence bands) in the channel [4], [5]. These tunnel transistors can have an inverse subthreshold slope that is smaller than 60 mV per decade at room temperature, which is the minimum possible value in conventional transistors.
6) Quantum corrections to semiclassical simulations: Semiclassical models have incorporated quantum corrections to account for quantum mechanical features that become important with miniaturization [6]–[8]. Quantum mechanical simulations are playing an important role in benchmarking these quantum corrections [9], [10].

In this review paper, we will discuss our approach to quantum mechanical modeling of transistors using the nonequilibrium Green's function (NEGF) approach. Simply stated, the advantage of using the NEGF approach in transistor modeling stems from its ability to calculate the nonequilibrium distribution function of electrons in the presence of tunneling, scattering mechanisms (that cause energy, momentum, and phase relaxation), and coupling of the device to contacts.

To put NEGF modeling into context, we will first discuss the procedure to calculate the charge and current densities in a device using Schrodinger's equation. The wave function in the MOSFET is obtained by solving Schrodinger's equation

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r})\right]\phi(\vec{r}) = E\phi(\vec{r}) \tag{1}$$

for the structure consisting of the device (D) coupled to the source, drain, and gate (G) contacts, as shown in Fig. 1. The

Manuscript received January 16, 2007; revised May 16, 2007. The work of M. P. Anantram was supported by the NIST. The review of this paper was arranged by Editor A. Asenov.

M. P. Anantram is with the Waterloo Nanotechnology Center, ECE Department, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

A. Svizhenko is with the Silvaco International, Santa Clara, CA 95054 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.2007.902857

0018-9383/$25.00 © 2007 IEEE

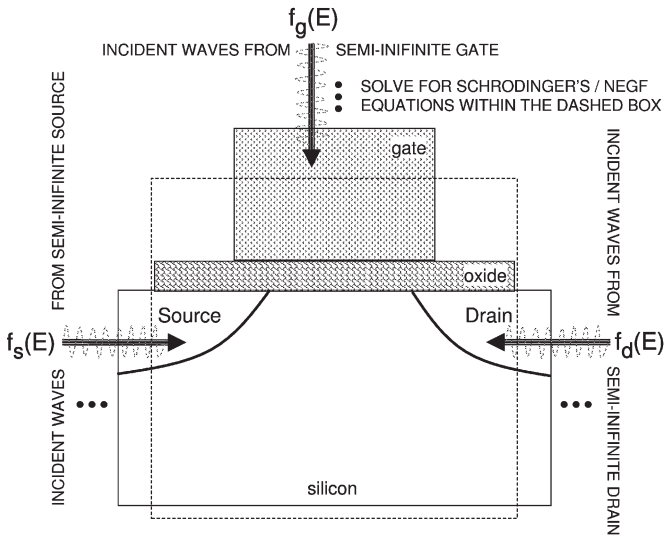


Fig. 1. Wave function in the device (D) region shown by the dashed box consists of waves incident from the source ($\phi_D^{(\mathrm{source})}$), drain ($\phi_D^{(\mathrm{drain})}$), gate ($\phi_D^{(\mathrm{gate})}$), and localized states ($\Psi_D^{(\mathrm{loc})}$). Waves incident from the substrate are neglected. The charge density inside the dashed box is obtained by solving the Schrodinger's/NEGF equations by treating the role of contacts as boundary conditions. Electron waves are incident from the semi-infinite source, drain, and gate, with a probability given by the Fermi factor in the contacts, f_s, f_d, and f_d, respectively.

wave function $\phi(\vec{r})$ at energy E consists of four orthonormal sets of wave functions [11]: 1) traveling waves incident from the source ($\phi_{D,k}^{(\mathrm{Source})}$); 2) traveling waves incident from the drain ($\phi_{D,k}^{(\mathrm{Drain})}$); 3) traveling waves incident from the gate ($\phi_{D,k}^{(\mathrm{Gate})}$); and 4) states localized in the device region ($\phi_D^{(\mathrm{loc})}$). Note that the subscript D represents the device region, and k is the quantum number of a traveling wave incident from contacts to the device. The potential in the source and drain contacts is assumed to be constant along the transport direction (horizontal direction in Fig. 1). Similarly, the potential of the gate is assumed to be constant along the vertical direction in Fig. 1. The resistance of these highly doped contact regions can be included as classical series resistances, although caution must be exercised in choosing the boundary between the device (D) and contact regions, which will be discussed in Section IV.

The contribution to electron (n) and current (J) densities at location $\vec{r}$ are given by

$$n(\vec{r}) = \sum_k \left[\left| \Psi_{D,k}^{(\mathrm{Source})}(\vec{r}) \right|^2 f_s(E) + \left| \Psi_{D,k}^{(\mathrm{Drain})}(\vec{r}) \right|^2 \times f_d(E) + \left| \Psi_{D,k}^{(\mathrm{Gate})}(\vec{r}) \right|^2 f_g(E) \right] \quad (2)$$

and

$$\vec{J}(\vec{r}) = \sum_k \frac{e\hbar}{2\mathrm{mi}} \Big[\Psi_{D,k}^{(\mathrm{Source})}(\vec{r})^\dagger \nabla \Psi_{D,k}^{(\mathrm{Source})}(\vec{r}) f_s(E) + \Psi_{D,k}^{(\mathrm{Drain})}(\vec{r})^\dagger \nabla \Psi_{D,k}^{(\mathrm{Drain})}(\vec{r}) f_d(E) + \Psi_{D,k}^{(\mathrm{Gate})}(\vec{r})^\dagger \nabla \Psi_{D,k}^{(\mathrm{Gate})}(\vec{r}) f_g(E) - c.c \Big] \quad (3)$$

where f_s, f_d, and f_g are the Fermi factors in the source, drain, and gate contacts.[1] A computational approach based on solving (1)–(3) coupled with Poisson's equation can account for all the effects in nanotransistors stemming from phase coherent transport (phase coherent transport here refers to evolution of the electron wave, as given by the Schrodinger's equation without decoherence). The following quantum mechanical features can be modeled using (1)–(3):

1) ballistic transport and tunneling from the source to the drain;
2) tunneling across the gate;
3) discrete energy levels and correct shape of the electron density in the inversion layer;
4) elastic scattering due to dopants, impurities, and surface.

However, in a device operating at finite bias and room temperature, electron–phonon and electron–electron scatterings are inevitable. We have shown that transistors having channel lengths of 10 and 25 nm are far from ballistic when electron–phonon scattering is included [12]. References [13] and [14] has more comprehensively evaluated the scattering rates associated with various mechanisms for scattering and pointed out the importance of scattering in nanotransistors. The major reason for going beyond the Schrodinger's equation in modeling nanotransistors is the importance of including inelastic scattering mechanisms (electron–phonon and electron–electron) which leads to loss of phase coherence.

In Section II, we present the NEGF equations as applied to nanotransistors in the context of a general framework. In Sections III and IV, we present the application of the general equations discussed in Section II to our 2-D work in the ballistic limit and quasi-1-D work with electron–phonon scattering. Section V highlights details of the computational procedure and code parallelization efficiency. Some challenges and outlook are presented in Section VI.

II. NEGF Equations for Layered Structures

In the presence of phase-breaking scattering, (2) and (3) cannot be used to compute the charge and current densities. This is because scattering causes the various eigen states of the Hamiltonian (1) to mix with arbitrary quantum mechanical phase factors that cannot be determined. Furthermore, electron–phonon scattering drives the distribution function in the device away from equilibrium in a manner that the charge and current densities cannot be determined in a simple manner from the Fermi-distribution function of the contacts, as given by (2) and (3). The NEGF approach gives us a clear recipe in calculating the electron and current densities in the presence of an applied bias by accounting for electron–phonon scattering. It has been used to model transistors in the context of device structures that can be broken down into layers, with each layer consisting of grid points [12], [15]. Fig. 2 shows layered structures corresponding to 2-D (conventional MOSFET) and 1-D [linear chain of grid points used to model a dual-gate MOSFET (DGMOSFET) in mode space] devices. The single particle Hamiltonian of the layered structure is a block tridiagonal

[1] $c.c$ stands for complex conjugate and † stands for Hermitian conjugate.

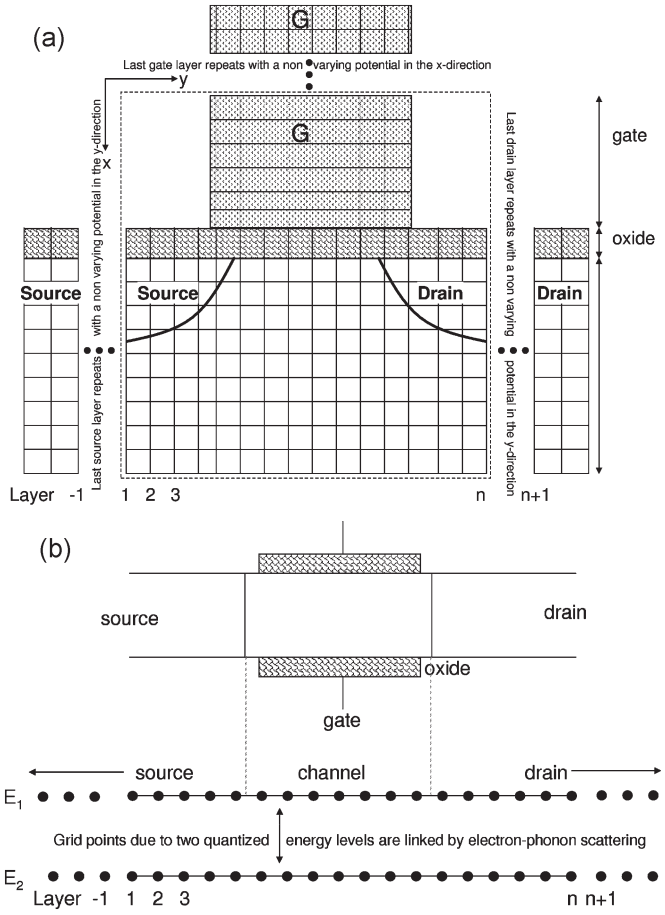


Fig. 2. (a) MOSFET consisting of n layers along the y-direction. Each layer consists of all grid points in the x-direction, which can vary with y. The dashed box corresponds to the dashed box in Fig. 1, where the NEGF equations are solved. The semi-infinite source, drain, and gate contacts are included as self-energies. (b) DGMOSFET in (top) real space is transformed into a 1-D problem in (bottom) mode space with n layers. Each of the n layers consists of M grid points corresponding to quantized energy levels in the channel of the DGMOSFET. These M grid points are linked only by electron–phonon interaction in our treatment. The figure only shows the grid points corresponding to two quantized energy levels ($M = 2$).

matrix, which is shown at the bottom of the page, where diagonal blocks H_q represent the Hamiltonian of layer q and off-diagonal blocks $T_{q,q+1}$ represent interaction between layers q and $q+1$.

The retarded Green's function equation in the presence of electron–phonon scattering is

$$\left[EI - H - \Sigma^r_{\text{Phonon}}\right] G^r = I \tag{4}$$

where Σ^r_{Phonon} is the self-energy due to electron–phonon scattering. We partition the structure into the device, source (s), drain (d), and gate (g) leads, as shown in Fig. 2. The current and nonequilibrium electron density in the device can be calculated without computing any nonequilibrium quantity in the contacts. That is, the role of contacts is accounted for only by including the effect of "isolated contacts" (that are not connected to the device) in thermal equilibrium as boundary conditions.

We first define

$$A' = \left[EI - H - \Sigma^r_{\text{Phonon}}\right].$$

Then, (4) can be written as

$$\begin{aligned} A'G^r = I \rightarrow & \begin{pmatrix} A'_{\text{ss}} & A'_{\text{sD}} & O & O \\ A'_{\text{Ds}} & A'_{\text{DD}} & A'_{\text{Dd}} & A'_{\text{Dg}} \\ O & A'_{\text{dD}} & A'_{\text{dd}} & O \\ O & A'_{\text{gD}} & O & A'_{\text{gg}} \end{pmatrix} \\ & \times \begin{pmatrix} G^r_{\text{ss}} & G^r_{\text{sD}} & G^r_{\text{sd}} & G^r_{\text{sg}} \\ G^r_{\text{Ds}} & G^r_{\text{DD}} & G^r_{\text{Dd}} & G^r_{\text{Dg}} \\ G^r_{\text{Ds}} & G^r_{\text{dD}} & G^r_{\text{dd}} & G^r_{\text{dg}} \\ G^r_{\text{gs}} & G^r_{\text{gD}} & G^r_{\text{gD}} & G^r_{\text{gg}} \end{pmatrix} \\ & = \begin{pmatrix} I & O & O & O \\ O & I & O & O \\ O & O & I & O \\ O & O & O & I \end{pmatrix} \end{aligned} \tag{5}$$

where

$$A'_{\text{ss}} = \begin{pmatrix} \bullet & \bullet & \bullet & & & \\ \bullet & \bullet & \bullet & & & \\ & & -T^\dagger_{s3} & A'_{s3} & -T_{s2} & \\ & & & T^\dagger_{s1} & A'_{s2} & -T_{s1} \\ & & & & -T^\dagger_{s1} & A'_{s1} \end{pmatrix}$$

and

$$A'_{\text{dd}} = \begin{pmatrix} A'_{d1} & -T_{d1} & & & & \\ -T^\dagger_{d1} & A'_{d2} & -T_{d2} & & & \\ & -T^\dagger_{d2} & A'_{d3} & -T_{d3} & & \\ & & \bullet & \bullet & \bullet & \\ & & & \bullet & \bullet & \bullet \end{pmatrix}$$

are semi-infinite dimension matrices corresponding to the semi-infinite source and drain contacts, respectively. The symbol O is used to represent zero matrices. The matrix A'_{DD}, which is shown at the bottom of the next page, corresponds to the device

$$H = \begin{pmatrix} \bullet & \bullet & \bullet & & & & & & \\ \bullet & H_1 & T_{12} & & & & & & \\ & T^\dagger_{12} & H_2 & T_{2,3} & & & & & \\ & & \bullet & \bullet & \bullet & & & & \\ & & & \bullet & \bullet & \bullet & & & \\ & & & & \bullet & \bullet & & & \\ & & & & T^\dagger_{n-2,n-1} & H_n & T_{n-1,n} & & \\ & & & & & T^\dagger_{n-1,n} & H_n & \bullet & \\ & & & & & & \bullet & \bullet & \bullet \end{pmatrix}$$

region. Each of the matrices A'_1 through A'_n corresponds to the n layers of the device; A'_{s1}, A'_{s2}, and so on correspond to the layers of the source contact, and A'_{d1}, A'_{d2}, and so on correspond to the layers of the drain contact. The matrices

$$A'_{\mathrm{Ds}} = \begin{pmatrix} \bullet & \bullet & \bullet & \bullet & O & -T_{\mathrm{Ds}} \\ \bullet & \bullet & \bullet & \bullet & O & O \\ \bullet & \bullet & \bullet & \bullet & O & O \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & O & O \end{pmatrix}$$

and

$$A'_{\mathrm{Dd}} = \begin{pmatrix} O & O & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ O & O & \bullet & \bullet & \bullet & \bullet \\ O & O & \bullet & \bullet & \bullet & \bullet \\ -T_{\mathrm{Dd}} & O & \bullet & \bullet & \bullet & \bullet \end{pmatrix}$$

represent coupling of the source and drain contacts to the device, respectively. The dimension of A'_{Ds} and A'_{Dd} is equal to the number of grid points in the first and last vertical layers in the device by the infinite number of grid points in the source and drain contacts, respectively. The only nonzero elements of $A'_{\mathrm{Ds}}(A'_{\mathrm{Dd}})$ correspond to the connection between vertical layer 1 (n) of the device to the vertical layer of the source (drain) that is closest to the device.

The gate is represented in a similar way as the source and drain, except that each element corresponds to the horizontal layer of the gate contact

$$A'_{\mathrm{gg}} = \begin{pmatrix} A'_{g1} & -T_{g1} & & & & \\ -T^{\dagger}_{g1} & A'_{g2} & -T_{g2} & & & \\ & -T^{\dagger}_{g2} & A'_{g3} & -T_{g3} & & \\ & & \bullet & \bullet & \bullet & \\ & & & \bullet & \bullet & \bullet \end{pmatrix}.$$

The coupling of the gate to the device is represented by

$$A'_{\mathrm{ss}} = \begin{pmatrix} -T_{D_1 g} & O & \bullet & \bullet & \bullet & \bullet \\ -T_{D_2 g} & O & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ -T_{D_{n-1} g} & O & \bullet & \bullet & \bullet & \bullet \\ -T_{D_n g} & O & \bullet & \bullet & \bullet & \bullet \end{pmatrix}.$$

The size of A'_{Dg} is equal to the number of grid points in the horizontal layer of the device times the infinite number of grid points in the gate contact. Note that the gate connects to all layers of the device (Fig. 2).

The Green's function in the device region can be solved using simple algebraic operations on (5) to give

$$\left[A'_{\mathrm{DD}} - A'_{\mathrm{Ds}} A'^{-1}_{\mathrm{ss}} A'_{\mathrm{sD}} - A'_{\mathrm{Dd}} A'^{-1}_{\mathrm{dd}} A'_{\mathrm{dD}} - A'_{\mathrm{Dg}} A'^{-1}_{\mathrm{gg}} A'_{\mathrm{gD}}\right] G^r_{\mathrm{DD}} = I. \quad (6)$$

The second, third, and fourth terms of (6), which represent coupling of the device region to the source, drain, and gate contacts, respectively, are called self-energy functions.

The Green's functions of the isolated semi-infinite contacts by definition are

$$A'_{\mathrm{ss}} g^r_s = I,\ A'_{\mathrm{dd}} g^r_d = I \text{ and } A'_{\mathrm{gg}} g^r_g = I.$$

The surface Green's function of the source (drain) contact is the element of g^r_s corresponding to the edge layer $-1(n+1)$ in the source (drain) contact

$$g^r_{s_{-1,-1}} = \left(A'^{-1}_{\mathrm{ss}}\right)_{1,1} \qquad \text{and} \qquad g^r_{d_{n+1,n+1}} = \left(A'^{-1}_{\mathrm{dd}}\right)_{1,1}.$$

Equation (6) can now be rewritten as

$$\left[EI - H - \Sigma^r_{\mathrm{Phonon}} - \Sigma^r_{\mathrm{source}} - \Sigma^r_{\mathrm{drain}} - \Sigma^r_{\mathrm{gate}}\right] G^r_{\mathrm{DD}} = I. \quad (7)$$

The only nonzero contributions due to the source and drain contacts are in the first and nth block of (7), respectively

$$\begin{aligned} \Sigma^r_{\mathrm{source}_{1,1}} &= T_{\mathrm{Ds}}\, g^r_{s_{-1,-1}} & T_{\mathrm{sD}} &= \Sigma^r_s \\ \Sigma^r_{\mathrm{drain}_{n,n}} &= T_{\mathrm{Dd}}\, g^r_{d_{n+1,n+1}} & T_{\mathrm{dD}} &= \Sigma^r_d \end{aligned}.$$

The gate self-energy $(\Sigma^r_{\mathrm{gate}})$

$$\Sigma^r_{\mathrm{gate}_{i,i'}} = T_{D_i g} g^r_{g_{1,1}} T_{g D_{i'}}$$

is a little more complex as summation over the gate (g) grid points are implied. Note that $T_{\mathrm{Ds}} = T^{\dagger}_{\mathrm{sD}}$ and $T_{\mathrm{Dd}} = T^{\dagger}_{\mathrm{dD}}$. $\Sigma^r_{\mathrm{source}}$, $\Sigma^r_{\mathrm{drain}}$, and Σ^r_{gate} are known as the self-energies due to the source, drain, and gate contacts, respectively; they are all square matrices with a dimension equal to the number of grid points in the device. The only nonzero elements of $\Sigma^r_{\mathrm{source}}(\Sigma^r_{\mathrm{drain}})$ are shown earlier, and they correspond to the connection between the grid points in layer 1 (n) and the source (drain), respectively. The value of $\Sigma^r_{\mathrm{source}}$ and $\Sigma^r_{\mathrm{drain}}$ at all other device grid points is zero. $T_{D_i g}$ is the coupling

$$A'_{\mathrm{DD}} = \begin{pmatrix} A'_1 & -T_{1,2} & & & & & & \\ -T^{\dagger}_{1,2} & A'_2 & -T_{2,3} & & & & & \\ & \bullet & \bullet & \bullet & & & & \\ & & \bullet & \bullet & \bullet & & & \\ & & & \bullet & \bullet & \bullet & & \\ & & & & -T^{\dagger}_{n-2,n-1} & A'_{n-1} & -T_{n-1,n} & \\ & & & & & -T^{\dagger}_{n-1,n} & A'_n & \end{pmatrix}$$

between layer i in the device and gate. If the tunneling across the gate oxide is neglected, then Σ^r_{gate} is zero. By defining

$$A_{\text{DD}} = EI - H - \Sigma^r_{\text{Phonon}} - \Sigma^r_{\text{source}} - \Sigma^r_{\text{drain}} - \Sigma^r_{\text{gate}} \quad (8)$$

(7) can be written as

$$A_{\text{DD}} G^r_{\text{DD}} = I.$$

The Green's function G^r described does not contain information about the nonequilibrium electron and current densities. However, G^r is necessary to calculate these quantities. At equilibrium, the electron density (n) at spatial location $\vec{r}$ and energy E is related to G^r by

$$n(\vec{r}, E) = N(\vec{r}, E) f(E)$$

where $N(\vec{r}, E) = -(1/\pi)\text{Im}[G^r(r, E)]$ is the density of states at location $\vec{r}$, and $f(E)$ is the Fermi factor. Away from equilibrium, the electron density is equal to

$$n(\vec{r}, E) = -iG^<(\vec{r}, \vec{r}, E) \quad (9)$$

where $G^<$ is called the less-than Green's function. The governing equation for $G^<$ is

$$[EI - H - \Sigma^r_{\text{Phonon}}]\, G^< = \Sigma^<_{\text{Phonon}} G^a \quad (10)$$

where G^a is the Hermitian conjugate of the retarded Green's function G^r. $\Sigma^<_{\text{Phonon}}(r, E)$ is the less-than self-energy which represents in-scattering of electrons to $(\vec{r}, E)$ due to electron–phonon scattering. The dimension of (10) is essentially infinite due to the semi-infinite source and drain leads. It can, however, be converted to a matrix with dimensions equal to the number of grid points in the n device layers. In a manner similar to the derivation of (7) for the retarded Green's function, the role of the contacts can be folded into the device to obtain

$$[EI - H - \Sigma^r_{\text{Phonon}} - \Sigma^r_{\text{source}} - \Sigma^r_{\text{drain}} - \Sigma^r_{\text{gate}}]\, G^<_{\text{DD}} = \Sigma^<_{\text{DD}} G^a_{\text{DD}} \quad (11)$$

or

$$A_{\text{DD}} G^<_{\text{DD}} = \Sigma^<_{\text{DD}} G^a_{\text{DD}}.$$

The self-energy $\Sigma^<_{\text{DD}}$ has contributions due to both electron–phonon interactions and leads

$$\begin{aligned}
\Sigma^<_{\text{DD}_{11}} &= \Sigma^<_{\text{Phonon}_{11}} + \Sigma^<_s + \Sigma^<_{\text{gate}_{11}} \\
\Sigma^<_{\text{DD}_{nn}} &= \Sigma^<_{\text{Phonon}_{nn}} + \Sigma^<_d + \Sigma^<_{\text{gate}_{nn}} \\
\Sigma^<_{\text{DD}_{ii}} &= \Sigma^<_{\text{Phonon}_{ii}} + \Sigma^<_{\text{gate}_{ii}}
\end{aligned}$$

where $i = 2, 3, 4, \ldots, n-1$. The in-scattering self-energies due to the contacts, $\Sigma^<_s$, $\Sigma^<_d$, and $\Sigma^<_{\text{gate}}$ are

$$\begin{aligned}
\Sigma^<_s(E) &= -i2\,\text{Im}\left[\Sigma^r_s(E)\right] f_s(E) = i\Gamma_s(E) f_s(E) \\
\Sigma^<_d(E) &= -i2\,\text{Im}\left[\Sigma^r_d(E)\right] f_d(E) = i\Gamma_d(E) f_d(E), \\
\Sigma^<_{\text{gate}}(E) &= -i2\,\text{Im}\left[\Sigma^r_{\text{gate}}(E)\right] f_g(E) = i\Gamma_{\text{gate}}(E) f_g(E),
\end{aligned}$$

where

$$\begin{aligned}
\Gamma_s(E) &= -2\,\text{Im}\left[\Sigma^r_s(E)\right] \\
\Gamma_d(E) &= -2\,\text{Im}\left[\Sigma^r_d(E)\right] \\
\Gamma_{\text{gate}}(E) &= -2\,\text{Im}\left[\Sigma^r_{\text{gate}}(E)\right].
\end{aligned}$$

f_s, f_d, and f_g are the distribution functions (Fermi–Dirac) in the source, drain, and gate contacts, respectively. The dimension of $\Sigma^r_s(E)$, $\Sigma^r_d(E)$, and $\Sigma^r_{\text{gate}}(E)$ are equal to the number of grid points in layer 1, layer n, and the entire device, respectively.

While the calculation of the diagonal elements of $G^<$ gives us the electron density as indicated in (9), the calculation of the self-energy $\Sigma^{r,<}_{\text{phonon}}$ requires another correlation function $G^>$, which is called the greater-than (hole) Green's function. The diagonal elements of $G^>(E)$ represent the density of unoccupied states (in the valence band, this would give the hole density)

$$h(\vec{r}, E) = iG^>(\vec{r}, \vec{r}, E).$$

The governing equation for $G^>(E)$,

$$[EI - H - \Sigma^r_{\text{Phonon}}]\, G^> = \Sigma^>_{\text{Phonon}} G^a \quad (12)$$

can be written as

$$[EI - H - \Sigma^r_{\text{Phonon}} - \Sigma^r_{\text{source}} - \Sigma^r_{\text{drain}} - \Sigma^r_{\text{gate}}]\, G^>_{\text{DD}} = \Sigma^>_{\text{DD}} G^a_{\text{DD}}. \quad (13)$$

Equation (13) can also be written as

$$\begin{aligned}
A_{\text{DD}} G^>_{\text{DD}} &= \Sigma^>_{\text{DD}} G^a_{\text{DD}} \\
\Sigma^>_{\text{DD}_{11}} &= \Sigma^>_{\text{Phonon}_{11}} + \Sigma^>_s + \Sigma^>_{\text{gate}_{11}} \\
\Sigma^>_{\text{DD}_{nn}} &= \Sigma^>_{\text{Phonon}_{nn}} + \Sigma^>_d + \Sigma^>_{\text{gate}_{nn}} \\
\Sigma^>_{\text{DD}_{ii}} &= \Sigma^>_{\text{Phonon}_{ii}} + \Sigma^>_{\text{gate}_{ii}}
\end{aligned}$$

where $i = 2, 3, 4, \ldots, n-1$. $\Sigma^>_{\text{Phonon}}$ is the out-scattering self-energy due to phonon scattering. The out-scattering self-energies due to the leads $\Sigma^>_s$, $\Sigma^>_d$, and $\Sigma^>_{\text{gate}}$ are

$$\begin{aligned}
\Sigma^>_s(E) &= i2\,\text{Im}\left[\Sigma^r_s(E)\right](1 - f_s(E)) \\
&= -i\Gamma_s(E)\,(1 - f_s(E)) \\
\Sigma^>_d(E) &= i2\,\text{Im}\left[\Sigma^r_d(E)\right](1 - f_d(E)) \\
&= -i\Gamma_d(E)\,(1 - f_d(E)) \\
\Sigma^>_{\text{gate}}(E) &= i2\,\text{Im}\left[\Sigma^r_{\text{gate}}(E)\right](1 - f_g(E)) \\
&= -i\Gamma_{\text{gate}}(E)\,(1 - f_g(E))
\end{aligned}$$

where $1 - f_s(E)$, $1 - f_d(E)$, and $1 - f_g(E)$ are the probability of finding an unoccupied state in the source, drain, and gate contacts, respectively, at energy E. We also state that G^r, G^a, $G^<$, and $G^>$ are related by $G^r - G^a = G^> - G^<$ and $\Sigma^r - \Sigma^a = \Sigma^> - \Sigma^<$.

It is important to note that the full G^r_{DD} matrix is not needed to calculate the diagonal elements of $G^<$ (electron density)

and $G^>$, which are repetitively needed in modeling of devices. Computing the full $G^r_{\rm DD}$ matrix approximately requires $N^{2.7}$ operations. $N = M \cdot N_{\rm layer}$ is the total number of grid points, where M is the matrix size of each layer (equal to the number of grid points) and $N_{\rm layer}$ is the number of layers in the structure. In contrast, the use of the recursive algorithm discussed in reference [15], albeit more difficult to implement, significantly cuts down the operation count to $M^{2.7} \cdot N_{\rm layer}$. The main assumption that should be satisfied so as to use the recursive algorithm developed in [15] is that $A_{\rm DD}$ is block tridiagonal.

III. 2-D Modeling of Transistors: Real Space

In this section, we will discuss 2-D modeling of nanotransistors and define the A' matrices of Section II. While [15] considers the more general case, the discussion here is restricted to a single valley of silicon with effective masses m_x, m_y, and m_z. The Hamiltonian is

$$H(\vec{r}) = -\frac{\hbar^2}{2}\left[\frac{d}{dx}\left(\frac{1}{m_x}\frac{d}{dx}\right) + \frac{d}{dy}\left(\frac{1}{m_y}\frac{d}{dy}\right) + \frac{d}{dz}\left(\frac{1}{m_z}\frac{d}{dz}\right)\right] + V(\vec{r}) \quad (14)$$

where m_x, m_y, and m_z are components of the effective mass. The equation of motion for the Green's function (G^r) and electron correlation function ($G^<$) is

$$[E - H(\vec{r}_1)]\, G^r(\vec{r}_1, \vec{r}_2, E) - \int d\vec{r}\, \Sigma^r(\vec{r}_1, \vec{r}, E) G^r(\vec{r}, \vec{r}_2, E) = \delta(\vec{r}_1 - \vec{r}_2) \quad (15)$$

and

$$[E - H(\vec{r}_1)]\, G^{<(>)}(\vec{r}_1, \vec{r}_2, E) - \int d\vec{r}\, \Sigma^r(\vec{r}_1, \vec{r}, E) G^{<(>)}(\vec{r}, \vec{r}_2, E) = \int d\vec{r}\, \Sigma^{<(>)}(\vec{r}_1, \vec{r}, E) G^a(\vec{r}, \vec{r}_2, E). \quad (16)$$

$\vec{r}$, $\vec{r}_1$, and $\vec{r}_2$ span only the device in (15) and (16). The influence of the semi-infinite source, drain, and gate leads, and electron–phonon interaction are included via self-energy terms Σ^r, $\Sigma^<$, and $\Sigma^>$.

The electrostatic potential varies in the (x, y) plane in Fig. 2, and the system is translationally invariant along the z-axis. Therefore, all quantities $A(\vec{r}_1, \vec{r}_2, E)$ only depend on the difference coordinate $z_1 - z_2$. By using the relation

$$A(\vec{r}_1, \vec{r}_2, E) = \int \frac{dk_z}{2\pi} e^{ik_z(z_1 - z_2)} A(x_1, y_1, x_2, y_2, k_z, E)$$

(15) and (16) simplify to

$$\left[E - \frac{\hbar^2 k_z^2}{2m_z} - H(\vec{r}_1)\right] G^r(\vec{r}_1, \vec{r}_2, k_z, E) - \int d\vec{r}\, \Sigma^r(\vec{r}_1, \vec{r}, k_z, E) G^r(\vec{r}, \vec{r}_2, k_z, E) = \delta(\vec{r}_1 - \vec{r}_2) \quad (17)$$

and

$$\left[E - \frac{\hbar^2 k_z^2}{2m_z} - H(\vec{r}_1)\right] G^{<(>)}(\vec{r}_1, \vec{r}_2, k_z, E) - \int d\vec{r}\, \Sigma^r(\vec{r}_1, \vec{r}, k_z, E) G^{<(>)}(\vec{r}, \vec{r}_2, k_z, E) = \int d\vec{r}\, \Sigma^{<(>)}(\vec{r}_1, \vec{r}, k_z, E) G^a(\vec{r}, \vec{r}_2, k_z, E) \quad (18)$$

where $\vec{r} \to (x, y)$ for the remainder of this section.

Equations (17) and (18) can be written in matrix form as

$$A' G^r = \lambda \quad \text{and} \quad A' G^{<(>)} = \Sigma^{<(>)} G^a. \quad (19)$$

The self-energies due to the source and drain are nonzero only at layer 1 and n, respectively [Fig. 2(a)]. The self-energy due to the gate is nonzero at all layers of the device region [dashed box in Fig. 2(a)], which make contact to the semi-infinite gate. The A' matrix is ordered such that all grid points at the y-coordinate (layer) correspond to a diagonal block of dimension N_x, and there are N_y of such blocks. In the notation adopted, $A'_{j_1,j_2}(i, i')$ is the off-diagonal entry corresponding to grid points (x_i, y_{j_1}) and $(x_{i'}, y_{j_2})$. Here, j_1 and j_2 correspond to the y-direction, and i and i' correspond to the x-direction. The nonzero elements of the diagonal blocks of the A' matrix are

$$\begin{aligned} A'_{j,j}(i,i) &= E' - V_{i,j} - T_{j,j}(i+1,i) - T_{j,j}(i-1,i) \\ &\quad - T_{j+1,j}(i,i) - T_{j-1,j}(i,i) - \Sigma^r_s(x_i,x_i)\delta_{j,1} \\ &\quad - \Sigma^r_d(x_i,x_i)\delta_{j,N_y} - \Sigma^r_{\rm gate}(y_j,y_j)\delta_{i,1} \\ A'_{j,j}(i \pm 1, i) &= T_{j,j}(i \pm 1, i) - \Sigma^r_s(x_{i\pm1}, x_i)\delta_{j,1} \\ &\quad - \Sigma^r_d(x_{i\pm1}, x_i)\delta_{j,N_y} \\ A'_{j,j}(i,i') &= -\Sigma^r_s(x_i,x_{i'})\delta_{j,1} - \Sigma^r_d(x_i,x_{i'})\delta_{j,N_y}, \\ &\quad \text{for } i' \neq i,\ i \pm 1 \end{aligned}$$

where $E' = E - \hbar^2 k_z^2/2m_z$, and $V_{i,j} = V(x_i, y_j)$. The off-diagonal blocks are

$$A'_{j,j'}(i,i') = 0, \qquad \text{for } j' \neq j.$$

The nonzero elements of the T matrix are

$$T_{j,j}(i \pm 1, i) = \frac{\hbar^2}{2m^{\pm x}} \frac{2}{x_{i+1} - x_{i-1}} \frac{1}{|x_{i\pm1} - x_i|}$$

$$T_{j\pm1,j}(i,i) = \frac{\hbar^2}{2m^{\pm y}} \frac{2}{y_{j+1} - y_{j-1}} \frac{1}{|y_{j\pm1} - y_j|}$$

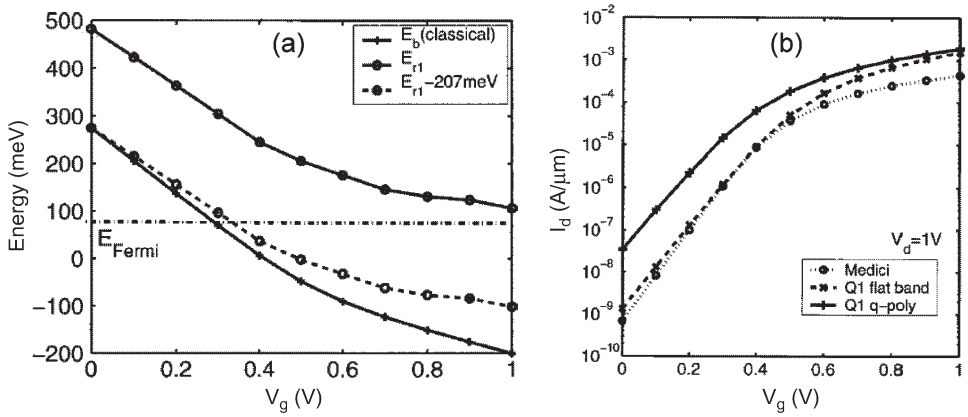


Fig. 3. (a) Resonant energy level versus gate voltage. (b) Current–voltage relationship using a classical and our quantum simulator (with permission from [15]).

where $m^{\pm x} = (m_{i\pm1,j} + m_{i,j})/2$ and $m^{\pm y} = (m_{i,j\pm1} + m_{i,j})/2$. λ is a diagonal matrix whose elements are given by

$$\lambda_{i,j;i,j} = \frac{4}{(x_{i+1} - x_{i-1})(y_{i+1} - y_{i-1})}.$$

The λ appearing in (19) corresponds to the delta function in (17).

Finally, we comment on the procedure that is developed to calculate the gate current. A proper treatment of the gate contact would require the inclusion of gate self-energy into the Hamiltonian matrix $A'_{\rm DD}$. This would, however, break the block-tridiagonal form of $A'_{\rm DD}$ by introducing far off-diagonal matrix elements, which couple remote points of the gate contact. The recursive algorithm, which is used to calculate the charge and current densities [15], requires a block tridiagonal form of $A'_{\rm DD}$. Therefore, we set $\Sigma^r_{{\rm gate}_{j,j'\neq j}} = 0$. That is, the self-energy elements connecting two different layers is zero. We therefore use an approximate treatment of the gate contact by including only the diagonal elements of the gate self-energy. The equations for self-energies of the gate contact are

$$\Sigma^r_{{\rm gate}_{j,j'}}(i,i')(E) = \delta_{{\rm jj}'}\delta_{ii'}\delta_{i1}\delta_{i'1}T_{\rm ss}g^r_g T_{\rm gD}$$
$$\Sigma^<_{{\rm gate}_{j,j'}}(i,i')(E) = i\Gamma_{\rm gate}(E)f_g(E)$$
$$\Sigma^>_{{\rm gate}_{j,j'}}(i,i')(E) = -i\Gamma_{\rm gate}(E)\left[1 - f_g(E)\right]$$

where

$$\Gamma_{\rm gate}(E) = -2\,{\rm Im}\left[\Sigma^r_{\rm gate}(E)\right].$$

The electron density at (x_i, y_j, k_z, E) is given by

$$n(x_i, y_j, k_z, E) = -iG^<_{j,j}(i,i) \quad \text{evaluated at } (k_z, E).$$

The current density flowing in the y-direction from (x_i, y_j, k_z, E) to (x_i, y_{j+1}, k_z, E) is

$$J_y(x_i, y_j, k_z, E) = \frac{2e}{\hbar}\left[T_{j+1,j}(i,i)G^<_{j,j+1}(i,i) - T_{j,j+1}(i,i)G^<_{j+1,j}(i,i)\right] \quad \text{evaluated at } (k_z, E). \tag{20}$$

The current density flowing in the x-direction from (x_i, y_j, k_z, E) to (x_{i+1}, y_j, k_z, E) is

$$J_x(x_i, y_j, k_z, E) = \frac{2e}{\hbar}\left[T_{j,j}(i+1,i)G^<_{j,j}(i,i+1) - T_{j,j}(i,i+1)G^<_{j,j}(i+1,i)\right] \quad \text{evaluated at } (k_z, E).$$

A. Numerical Results

By using the equations presented earlier and by following the computational procedure outlined in Section V, we review our results of the 2-D simulations. Our approach can account for quantization in the inversion/depletion layer, gate tunneling, and ballistic transport. However, electron–phonon scattering is not included due to the large computational cost (see the discussion in Section VI). Consider the MIT-25 MOSFET [16] which has a channel length of 25 nm. The current as a function of the gate voltage calculated using our quantum simulator is compared to results from MEDICI [17]. The subthreshold slope from the quantum simulator is smaller than that predicted by MEDICI [Fig. 3(a)]. To understand this, we note that, while classically, the drain current at small gate biases is determined by the bottom of the conduction band at the source injection barrier, quantum mechanically, it is determined by the energy of the resonant level at the source injection barrier. With an increase in gate bias, the band bottom of the quantum well decreases at the source injection barrier, as shown by E_b (classical) in Fig. 3(a). In comparison, the energy of the lowest resonant level at the source injection barrier E_{r1} decreases more slowly, due to a stronger confinement in the channel, with an increase in gate voltages. As the drain current in the subthreshold region exponentially depends on barrier heights, which are E_b (classical) and E_{r1} in the classical and quantum cases, respectively, it immediately follows that the subthreshold slope is smaller in the quantum case.

The quantum mechanically calculated OFF current is higher than the semiclassical value when the polysilicon gate of the MIT-25 device is quantum mechanically treated just like the silicon in the channel (see "Q1 q-poly" and "Medici" in Fig. 3(a) at $V_G = 0$). The reason for this is that the calculated band bending in the polysilicon gate is opposite to the semiclassical

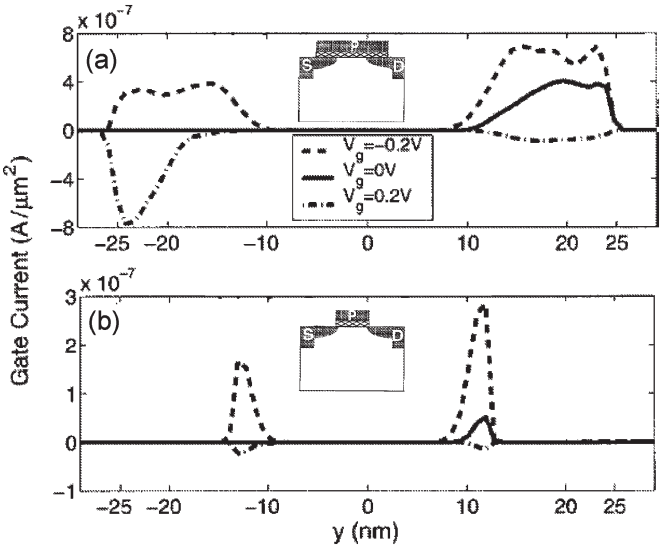


Fig. 4. Gate-tunneling current flowing at different points along the channel for devices with gate lengths of (a) 50 and (b) 25 nm (with permission from [15]).

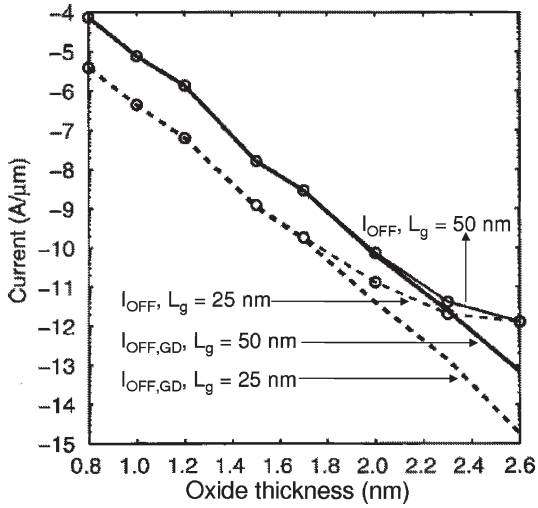


Fig. 5. Gate current versus oxide thickness for two different gate lengths of 25 and 50 nm (with permission from [15]).

prediction. Semiclassically, the conduction band in the polysilicon gate monotonically bends upward as the gate oxide is approached from deep inside the polysilicon gate in Fig. 2(a). In contrast, the quantum mechanically calculated conduction band bends downward as the gate oxide is approached from deep inside the gate. This is because the wave function is very small close to the gate-oxide interface, which makes the background doping density larger than the electron density. The lower conduction band of the polysilicon gate, in turn, causes a lower value of the conduction band bottom in the silicon channel when compared to the classical case, which causes the significantly higher quantum mechanically calculated OFF current. The OFF current that is obtained when all bands bending in the polysilicon gate are neglected ["Q1 flat band" in Fig. 3(a)] is in close agreement with MEDICI. Finally, the quantum mechanically calculated ON current is larger than the MEDICI value because it is calculated in the ballistic limit.

The simulator is also capable of calculating the current flowing through the gate using (20). The polysilicon gate length (L_g) in the MIT-25 device is 50 nm. There are significant gate–source and gate–drain overlaps because the channel length is only 25 nm. The gate current tunneling across the oxide at various locations along the channel is shown in Fig. 4(a) at a drain bias of 0.1 V and for three different gate biases. At negative gate biases, the tunnel current flows into the gate at locations corresponding to both the gate–source and gate–drain overlap regions, as shown by the dashed line. At zero gate voltage, the tunnel current primarily flows into the gate at the gate–drain overlap region, as shown by the solid line. At positive gate biases, the tunnel current flows into the channel at locations corresponding to both the gate–source and gate–drain overlap regions, as shown by the dash-dotted line; this current is larger in the gate–source overlap region because the voltage between the gate and source is larger than that between the gate and drain. Based on the profile of the tunneling current along the channel in Fig. 4(a), the simulator was used to investigate if the gate current can be decreased without changing the drive current by decreasing the gate length. The gate-tunneling current that is obtained by reducing the gate length to 25 nm is shown in Fig. 4(b). It is shown that the gate current in the overlap region decreases by about ten times. At the same time, the electrostatics near the injection barrier does not significantly change, and the corresponding drive current decreases by less than 2%. The simulator can also be used to find how the source–drain $(I_{\mathrm{OFF,SD}})$ and gate–drain $(I_{\mathrm{OFF,GD}})$ components contribute to the total OFF current (I_{OFF}) with decreasing oxide thickness. I_{OFF} can be written as

$$I_{\mathrm{OFF}} = I_{\mathrm{OFF,SD}} + I_{\mathrm{OFF,GD}}.$$

As the oxide thickness decreases, $I_{\mathrm{OFF,GD}}$ rapidly increases because of quantum mechanical tunneling across the oxide and eventually becomes much larger than $I_{\mathrm{OFF,SD}}$. Fig. 5 shows the results obtained from a number of calculations for the MIT-25 device with gate lengths of 25 and 50 nm. For the device with $L_g = 25/50$ nm, the OFF current is essentially determined by $I_{\mathrm{OFF,GD}}$ when the oxide thickness is smaller than 1.7/2.1 nm. It is also shown in Fig. 5 that I_{OFF} is about an order of magnitude smaller for the device with the smaller gate length.

IV. Quasi-1-D Modeling of Transistors With Electron–Phonon Scattering in Mode Space

Realistic nanotransistor modeling will involve phase-breaking scattering such that transport is between the ballistic and diffusive limits. In this regime, carriers are not thermally relaxed in the drain-end of the transistor, in contrast to long channel devices. The reflection of hot carriers from the drain-end toward the source-end, both above and below the source injection barrier, should be explicitly included in computing the drive current. It is in this intermediate regime that the NEGF method has a distinct advantage over self-consistently solving

Schrodinger's and Poisson's equations. To illustrate modeling of electron–phonon scattering in silicon nanotransistors using the NEGF approach, we consider a DGMOSFET with an ultrathin body. As 2-D calculations are very computationally expensive, our work with electron–phonon scattering is performed within the framework of a quasi-1-D mode space approach.

The quasi-1-D mode space approach consists of solving a 1-D Green's function equation for each subband that arises from quantization in the ultrathin body [quantization in the x-direction in Fig. 2(b)]. The electrostatics, however, is considered by solving the 2-D Poisson's equation, which will become clear toward the end of this section. As has been established by [17], in the mode space approach intersubband scattering due to the electrostatic potential is neglected in the phase-coherent/ballistic limit. In our treatment, we include intersubband scattering due to electron–phonon scattering within the Born approximation, as will be described below [12].

The 3-D effective mass Hamiltonian is the same as (14). As the z direction is infinite, the wave function can be expanded as $\Phi(x,y,z) = e^{ik_z z}\phi(x,y)$. The Schrodinger's equation is then

$$\left[E - \frac{\hbar^2 k_z^2}{2m_z^b} - \left(-\frac{\hbar^2}{2}\frac{d}{dy}\left(\frac{1}{m_y^b}\frac{d}{dy}\right) - \frac{\hbar^2}{2}\frac{d}{dx}\left(\frac{1}{m_x^b}\frac{d}{dx}\right) + \mathrm{V}(x,y)\right)\right]\phi(x,y) = 0.$$

In the first step of the mode space approach, the 1-D Schrodinger's equation

$$\left[-\frac{\hbar^2}{2m_x^b}\frac{d}{dx}\left(\frac{1}{m_x^b}\frac{d}{dx}\right) + \mathrm{V}(x,y)\right]\Psi_n(x,y) = E_n(y)\Psi_n(x,y) \tag{21}$$

is solved at every y-cross section [along the source–drain direction in Fig. 2(b)]. This gives the eigen values $E_n(y)$ that arise due to quantization along the body thickness (x-direction). $n = \{\nu, b\}$, where ν is the quantum number due to quantization in the x-direction and $b = 1, 2$, and 3 is the band index representing the valley of silicon.

In the second step, the Green's function equations for G^r, $G^<$, and $G^>$ are solved in each subband n using the values of $E_n(y)$ found in the first step. Note that $E_n(y)$ is effectively an electrostatic potential for electrons in subband n. The equations solved are

$$\left[E - \frac{\hbar^2 k_z^2}{2m_z^n} - \left(-\frac{\hbar^2}{2}\frac{d}{d}\left(\frac{1}{m_y^n}\frac{d}{dy}\right) + E_n(y)\right)\right] \times G_n^r(y,y',k_z,E) - \int dy_1 \Sigma_n^r(y,y_1,k_z,E) \times G_n^r(y_1,y',k_z,E) = \delta(y-y') \tag{22}$$

and

$$\left[E - \frac{\hbar^2 k_z^2}{2m_z^n} - \left(-\frac{\hbar^2}{2}\frac{d}{dy}\left(\frac{1}{m_y^n}\frac{d}{dy}\right) + E_n(y)\right)\right] G_n^{<(>)}(y,y',k_z,E) - \int dy_1 \Sigma_n^r(y,y_1,k_z,E) G_n^{<(>)}(y_1,y',k_z,E) = \int dy_1 \Sigma_n^{<(>)}(y,y_1,k_z,E) G_n^a(y_1,y',k_z,E) \tag{23}$$

m_y^n and m_z^n are the effective masses of silicon in the y and z directions.

The self-energies $\Sigma_n^\alpha (\alpha \in r, <, >)$ are given by

$$\Sigma_n^\alpha = \Sigma_{n,C}^\alpha + \Sigma_{n,P}^\alpha \qquad \text{where}$$
$$\Sigma_{n,P}^\alpha = \Sigma_{n,\mathrm{el}}^\alpha + \Sigma_{n,\mathrm{inel}}^\alpha.$$

$\Sigma_{n,C}^\alpha$ is the self-energy due to the contacts. The phonon self-energy $\Sigma_{n,P}^\alpha$ consists of two terms, i.e., $\Sigma_{n,\mathrm{el}}^\alpha$ due to elastic scattering and $\Sigma_{n,\mathrm{inel}}^\alpha$ due to inelastic scattering. Scattering between the lowest three subbands due to electron–phonon interaction is included, and all other intersubband scatterings are neglected. The self-energy due to contacts is nonzero only at the first (source) and nth (drain) grid points [Fig. 2(b)].

Assuming an isotropic scattering, an equilibrium phonon bath, and the self-consistent Born approximation, the self-energies due to electron–phonon scattering at grid point y_i are [19]

$$\Sigma_{\mathrm{el},n}^\alpha(y_i,E) = \sum_{n'} D_{n,n'}^{\mathrm{el}} \frac{\sqrt{m_z^{n'}}}{\pi\hbar\sqrt{2}} \int dE_z \frac{1}{\sqrt{E_z}} G_{n'}^\alpha(y_i,E_z,E)$$

and

$$\Sigma_{\mathrm{inel},n}^<(y_i,E) = \sum_{n',\eta} D_{n,n'}^\eta \frac{\sqrt{m_z^{n'}}}{\pi\hbar\sqrt{2}} \int dE_z \frac{1}{\sqrt{E_z}} \times \left[n_B(\hbar\omega_\eta) G_{n'}^<(y_i,E_z,E-\hbar\omega_\eta) + (n_B(\hbar\omega_\eta)+1) G_{n'}^<(y_i,E_z,E+\hbar\omega_\eta)\right]$$

$$\Sigma_{\mathrm{inel},n}^>(y_i,E) = \sum_{n',\eta} D_{n,n'}^\eta \frac{\sqrt{m_z^{n'}}}{\pi\hbar\sqrt{2}} \int dE_z \frac{1}{\sqrt{E_z}} \times \left[n_B(\hbar\omega_\eta) G_{n'}^>(y_i,E_z,E+\hbar\omega_\eta) + (n_B(\hbar\omega_\eta)+1) G_{n'}^>(y_i,E_z,E-\hbar\omega_\eta)\right].$$

$\Sigma_{\mathrm{inel},n}^r$ can be obtained using the Kramers–Krönig relationship

$$\mathrm{Re}\left[\Sigma_{\mathrm{inel},n}^r(y_i,E)\right] = \frac{1}{\pi}\mathrm{P}\int dE' \frac{\mathrm{Im}\left[\Sigma_{\mathrm{inel},n}^r(y_i,E')\right]}{E'-E}$$

and

$$\mathrm{Im}\left[\Sigma_{\mathrm{inel},n}^r(y_i,E)\right] = \frac{1}{2i}\left[\Sigma_{\mathrm{inel},n}^>(y_i,E) - \Sigma_{\mathrm{inel},n}^<(y_i,E)\right]$$

where P stands for the principal part of the integral. Note that the self-energies due to electron–phonon scattering depend only

on the total energy E (and not on E_z) due to the assumption of isotropic scattering. In silicon, assuming isotropic scattering is a reasonable approximation. Isotropic scattering, in turn, implies that the self-energy matrices are diagonal in the spatial coordinate.

Here, η represents the phonon modes, and the square of the matrix elements for phonon scattering is given by [20]

$$D^{\rm el}_{n,n'} = \left(\delta_{\nu,\nu'} + \frac{1}{2}\right)\delta_{b,b'}\frac{D_A^2 kT}{\rho v^2}$$

$$D^{\eta}_{n,n'} = \left(\delta_{\nu,\nu'} + \frac{1}{2}\right)\left[\delta_{b,b'}\frac{D_{g\eta}^2\hbar}{2\rho\omega_{g\eta}} + (1-\delta_{b,b'})\frac{D_{f\eta}^2\hbar}{\rho\omega_{f\eta}}\right].$$

The first expression represents intravalley scattering of 2-D electrons on bulk acoustic phonons within an elastic approximation. The second expression corresponds to inelastic g- and f-type intervalley scattering processes in silicon [21]. D_A, $D_{g\eta}$, and $D_{f\eta}$ are the deformation potentials, and $\omega_{g\eta}$ and $\omega_{f\eta}$ are the phonon frequencies. ρ is the mass density, k is the Boltzmann constant, T is the temperature, and v is the velocity of sound.

In the numerical solution, n uniformly spaced grid points in the y-direction with the grid spacing equal to Δy are considered. The discretized form of (22) and (23) are straightforwardly found to be

$$A_{i,i}G_n^r(y_i,y_i',k_z,E) + A_{i,i+1}G_n^r(y_{i+1},y_i',k_z,E) + A_{i,i-1}G_n^r(y_{i-1},y_i',k_z,E) = \frac{\delta_{i,i'}}{\Delta y}$$

and

$$A_{i,i}G_n^\alpha(y_i,y_i',k_z,E) + A_{i,i+1}G_n^\alpha(y_{i+1},y_i',k_z,E) + A_{i,i-1}G_n^\alpha(y_{i-1},y_i',k_z,E) = \Sigma_n^\alpha(y_i,E)G_n^a(y_i,y_i',k_z,E)$$

where

$$A_{i,i} = E - \frac{\hbar^2k_z^2}{2m_z^n} - \frac{\hbar^2}{m_y^n\Delta y^2} - E_n(y_i) - \Sigma_n^r(y_i,k_z,E)$$

and

$$A_{i\pm1,i} = +\frac{\hbar^2}{2m_z^n\Delta y^2}.$$

The contact self-energies are

$$\Sigma^r_{n,C}(y_1,k_z,E) = \left(\frac{\hbar^2}{2m_z^n\Delta y^2}\right)^2 g_s^r(k_z,E)$$

$$\Sigma^r_{n,C}(y_N,k_z,E) = \left(\frac{\hbar^2}{2m_z^n\Delta y^2}\right)^2 g_d^r(k_z,E)$$

$$\Sigma^<_{n,C}(y_1,k_z,E) = -i2{\rm Im}\left(\Sigma^r_{n,C}(y_1,k_z,E)\right)f_s(E)$$

$$\Sigma^<_{n,C}(y_N,k_z,E) = -i2{\rm Im}\left(\Sigma_{n,C}(y_N,k_z,E)\right)f_d(E)$$

$$\Sigma^>_{n,C}(y_1,k_z,E) = i2{\rm Im}\left(\Sigma^r_{n,C}(y_1,k_z,E)\right)[1-f_s(E)]$$

$$\Sigma^>_{n,C}(y_N,k_z,E) = i2{\rm Im}\left(\Sigma^r_{n,C}(y_N,k_z,E)\right)[1-f_d(E)]$$

where y_1 an y_N are the leftmost (source-end) and rightmost (drain-end) grid points, respectively, $g_s(k_z,E)$ and $g_d(k_z,E)$ are the surface Green's functions of the source and drain leads, respectively, and f_s and f_d are the Fermi functions in the source and drain contacts, respectively.

The electron and current densities are

$$n_n(y_i,k_z,E) = -iG_n^<(y_i,y_i,k_z,E)$$

$$J_n(y_i,k_z,E) = \frac{e}{\hbar}\frac{\hbar^2}{2m_y^n\Delta y^2}\left[G_n^<(y_i,y_{i+1},k_z,E) - G_n^<(y_{i+1},y_i,k_z,E)\right].$$

The total electron (n) and current densities at grid point y_i are given by

$$\begin{aligned} n(y_i) &= \sum_n n_n(y_i) \\ &= 4\sum_n \frac{\sqrt{m_z^{n'}}}{\pi\hbar\sqrt{2}}\int\frac{dE}{2\pi}\int dE_z\frac{1}{\sqrt{E_z}}n_n(y_i,E_z,E) \\ J(y_i) &= 4\sum_n \frac{\sqrt{m_z^{n'}}}{\pi\hbar\sqrt{2}}\int\frac{dE}{2\pi}\int dE_z\frac{1}{\sqrt{E_z}}J_n(y_i,E_z,E) \end{aligned} \tag{24}$$

where the prefactor of four in the above two equations accounts for two fold spin and valley degeneracies. $n_n(y_i)$ is the electron density (n) in subband n (subscript) at location y_i. The nonequilibrium electron and current densities are calculated in both the channel and extension regions using the algorithm for $G^<$ presented in [15]. Finally, solving the 1-D Green's function equation gives us only an electron density per unit length. This electron density per unit length is transformed to a 2-D electron density, which will form an input to the solved 2-D Poisson's equation, is taken to be [using (21) and (24)]

$$n(x_i,y_i) = \sum_n n_n(y_i)\left|\Psi_n(x_i,y_i)\right|^2.$$

A. Numerical Results

The equations presented above are applied to study the role of electron–phonon scattering in DGMOSFETs. As the channel length scales down and becomes comparable to the electron–phonon mean free path, transport in the channel transitions from diffusive to ballistic. This transition poses a question on the relative importance of scattering at various locations along the channel of nanotransistors: Is scattering near the source injection barrier significantly more important than scattering near the drain-end of the channel? Our simulations [12] established that, when the channel length is comparable to the scattering length, electron–phonon scattering in the drain-end of the channel causes a reduction in drive current that is comparable to the reduction caused by scattering near the source-end of the channel.

We consider the Purdue-10 device, which has a channel length of 10 nm [18]. The body thickness $(t_{\rm ch})$ is 1.5 nm.

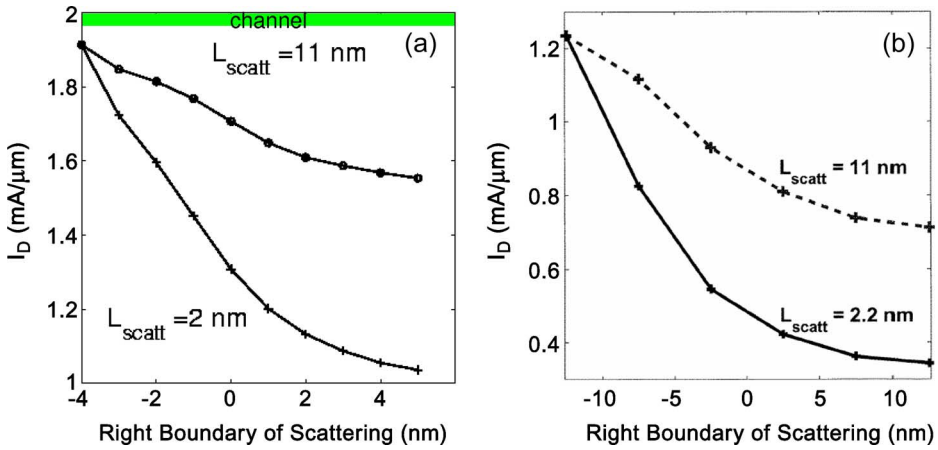


Fig. 6. Drive current versus the right boundary of scattering for devices with channel lengths of (a) 10 and (b) 25 nm. The channel is from -5 to $+5$ nm in (a) and from -12.5 to $+12.5$ nm in (b) (with permission from [12]).

The quantized energy levels in the channel due to the quantization perpendicular to the gate [x-direction in Fig. 2(b)] are

$$E_n = \frac{n^2\pi^2\hbar^2}{2mt_{\rm ch}^2}. \tag{25}$$

When $t_{\rm ch}$ is small, the energy level separation is large, and a very few subbands are occupied in the highly doped extension regions. The lowest three quantized energy levels (25) are at 173, 691, and 891 meV above the bulk conduction band. The Fermi energy at the contact doping of 10^{20} cm^{-3} for the Purdue-10 device is 60 meV above the bulk conduction band. As a result, electrons are injected only into one subband from the source-end at the operating voltage of $V_d = V_g = 0.6$ V. This justifies the use of the mode space approach.

While tunneling from the source to the drain and electron–phonon scattering are included, gate tunneling is neglected. The mean free path for electron–phonon scattering is approximately 11 nm. The channel is from -5 to $+5$ nm, and the peak of the source injection barrier for this device is at -4 nm.

We start with the ballistic (no electron–phonon scattering) limit and perform simulations by moving the "right boundary for scattering" from the source-end to the drain-end of the device. [12] The deformation potential is set to zero in the region to the right of the "right boundary for scattering." The ballistic current is 1.9 mA/μm [Fig. 6(a)]. The current in the presence of scattering from the source injection barrier to the right end of the channel ($+5$ nm) is 1.55 mA/μm. That is, including only electron–phonon scattering makes this device with almost equal channel and scattering lengths have a ballisticity of nearly 81%.[2] As the right boundary of scattering moves toward the drain-end, scattering in the right half of the channel causes about 50% of the decrease in current. That is, scattering in the right half of the channel is as important as scattering in the left half of the channel. We also show the change in drive current as a function of the right boundary of scattering, when the mean free path is reduced to 2 nm. It is interesting to note that, even for the shorter mean free path, scattering in the right half of the channel causes a large reduction in drive current of about 38%.

The reduction in the drive current as a function of the right boundary of scattering is also shown for a device with a 25-nm channel length in Fig. 6(b). The ballistic drive current for this device is approximately 1.23 mA/μm. When electron–phonon scattering is included everywhere in the channel (scattering length is 11 nm), the current decreases to approximately 0.71 mA/μm. This corresponds to a ballisticity of 68%. We again see that, when the scattering length is 11 nm, the reduction in the drain current due to scattering in the right half of the channel is appreciable. However, when the scattering length is decreased to 2 nm (by artificially increasing the deformation potentials), scattering in the right half of the channel causes only a 14% of reduction in ballistic current. Our simulation result in Fig. 6 clearly shows that the relative importance of scattering in the drain-end, in causing a reduction in the drive current from the ballistic limit, decreases as the channel length becomes significantly longer than the mean free path. To understand the reason for this, we plot the current density flowing per unit energy at various cross sections y from source to drain (solid line) for the device with a 25-nm channel length in Fig. 7. The current density is peaked above the source injection barrier for both scattering lengths considered. However, for a scattering length of 2 nm, the carrier density has relaxed to lower energies near the drain-end of the channel [Fig. 7(b)]. The dotted lines correspond to the mean energy for current flow as a function of y. Clearly, the mean energy for current flow follows the potential profile in the bulk of the channel when the scattering length is 2 nm. This indicates that the fraction of carriers reflected in the drain-end which travel back to the source injection barrier is smaller for a scattering length of 2 nm. As a result, scattering in the drain-end is relatively less effective in decreasing the drive current when compared to scattering in the source-end, as the ratio of channel to scattering length becomes larger. The relative importance of drain-end scattering in nanotransistors has also been verified using Monte Carlo simulations [22], [23].

[2]Ballisticity is defined to be the ratio of current with scattering to ballistic current, where both currents have been calculated using the NEGF method.

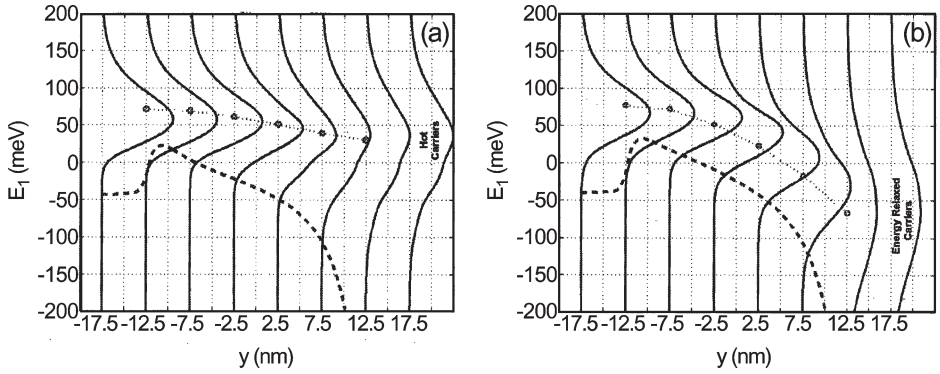


Fig. 7. Current per unit energy at different cross sections for scattering lengths of (a) 11 and (b) 2 nm is shown using solid lines (with permission from [12]).

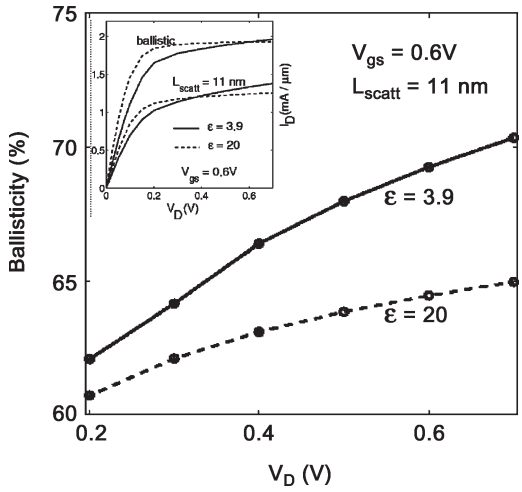


Fig. 8. Ballisticity of two DGMOSFETs with different gate dielectrics, as a function of drain bias. Note that the ballisticity increases with an increase in drain bias.

Another issue we considered is the interplay of channel electrostatics and electron–phonon scattering on the drive current. We consider two devices identical to the Purdue-10 device, except that now the two devices have oxide dielectric constants of 3.9 and 20 (high-κ). The high-κ device has a very small DIBL, as shown in the inset in Fig. 8. The work function of the high-κ device is chosen to make the ballistic currents the same for the two devices at $V_D = V_G = 0.6$ V. The ballisticity increases with an increase in drain bias in Fig. 8 because there is a greater probability that the velocity of carriers (in the drain-end) along the transport direction will decrease after a scattering event so as to force the electron into the drain. The reason for this is the increasing density of states in the drain-end at energies below the source injection barrier, with an increase in drain bias. An interesting finding is that the better electrostatics of the high-κ device prevents an efficient energetic relaxation of carriers to the right of the source injection barrier. As a result, the ballisticity is smaller for the high-κ device Fig. 8.

V. Computational Flowchart, Parallelization, and Efficiency

The flowchart of the computation is shown in Fig. 9 The calculation starts with the initial guess for the electrostatic potential profile $\phi^{(0)}(x_i, y_i)$, which is constructed by analytical functions or from the potential of a nearby bias point. In general, the calculation for one bias point involves two simultaneous iterative processes: 1) Born iterations to find self-consistent electron Green's functions (7), (11), and (13) and electron–phonon self-energies; and 2) Poisson's iterations to find the self-consistent electrostatic potential. The criterion for convergence of the Born iterations is the conservation of current at each cross section of the device. It is usually difficult and not necessary to achieve convergence of Born iterations within one Poisson's iteration. The number of Born iterations N_b is chosen so that the Born and Poisson's iterations converge simultaneously, while the electron–phonon self-energies are stored and updated throughout the calculation.

After N_b Born iterations, the quantum electron density obtained is plugged into Poisson's equation in order to solve for electrostatic potential ϕ. At the nth Poisson's iteration, we need to find the new value of the electrostatic potential $\phi_i^{(n+1)} = \phi_i^{(n)} + \delta\phi_i^{(n)}$. The update for electrostatic potential $\delta\phi_i^{(n)}$ is found by solving

$$-\nabla\left(\epsilon\nabla\left(\delta\phi_i^{(n)}\right)\right) + \frac{\partial n_i}{\partial\phi_i}\delta\phi^{(n)} = \xi,$$
$$\xi = \nabla\left(\epsilon\nabla\left(\phi_i^{(n)}\right)\right) + q(N_{\rm di} - n_i - N_{\rm ai}). \quad (26)$$

The floating boundary conditions $d\phi/dy = 0$ or fixed boundary conditions are used in the contacts. The subscript i means that a quantity is taken at a grid point (x_i, y_i). In the equation mentioned earlier, we neglected the hole density, which is usually much smaller than the electron density in monopolar devices.

In order to expedite Poisson's convergence, we use a predictor–corrector scheme, which is briefly described here. First, quantum electron density and electrostatic potential are used in a classical expression for electron density to find a quasi-Fermi level everywhere in the device

$$n_i = N_c\mathcal{F}_{1/2}\left(-\frac{\Delta E_{\rm ci} + q\phi_i - E_{\rm fi}}{kT}\right) \quad (27)$$

$$E_{\rm fi} = \Delta E_{\rm ci} + q\phi_i + kT\mathcal{F}_{1/2}^{-1}\left(\frac{n_i}{N_c}\right). \quad (28)$$

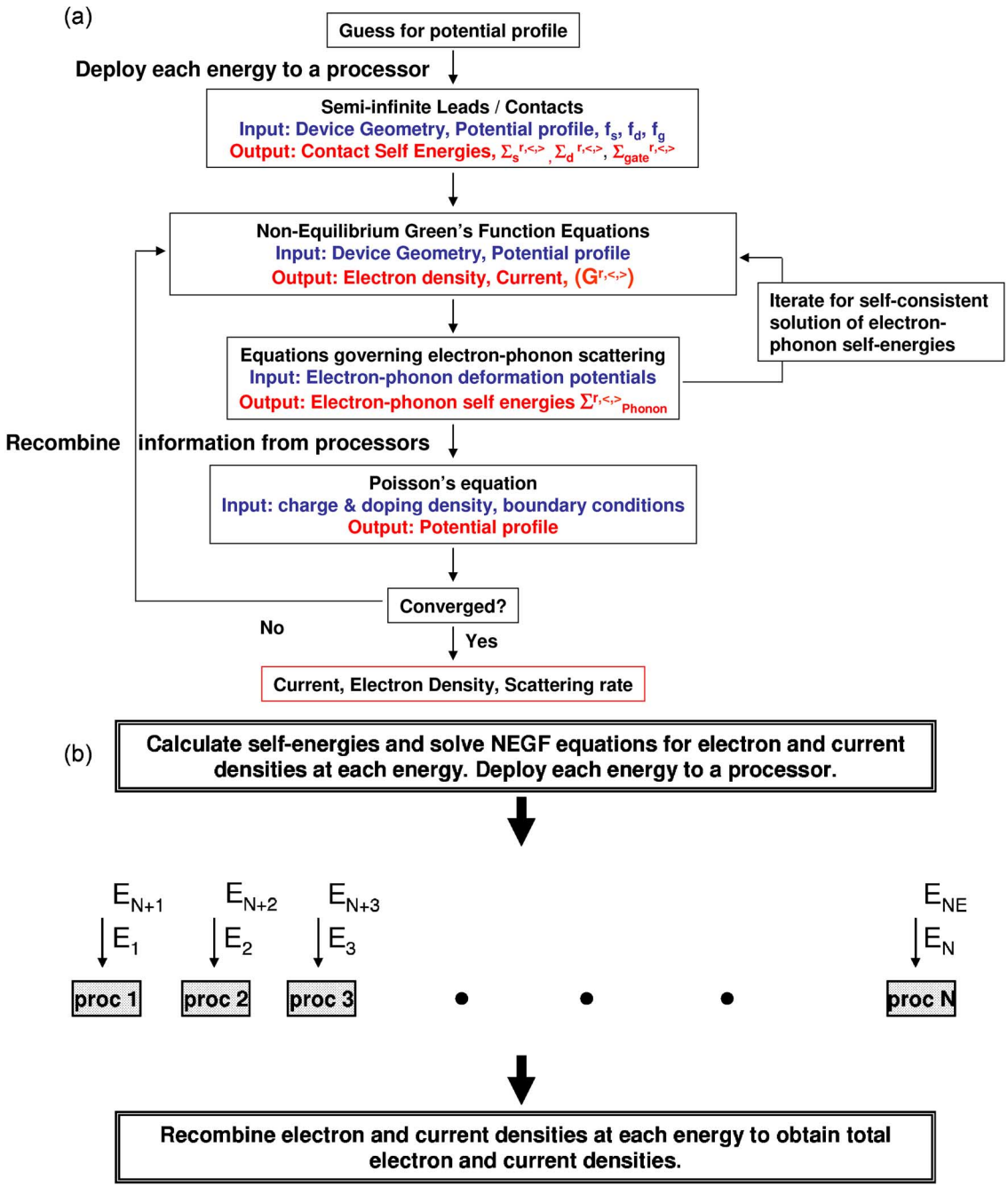


Fig. 9. (a) Flow chart of the computational procedure. (b) Parallel deployment of calculations at different energies to various processors.

We then perform $N_{\rm pp}$ inner Poisson's iterations with classical predictor for electron density (27) and a Jacobian

$$\frac{\partial n_i}{\partial \phi_i} = \frac{q}{kT} N_c \mathcal{F}_{-1/2}\left(-\frac{\Delta E_{\rm ci} + q\phi_i - E_{fi}}{kT}\right). \tag{29}$$

At the end, a new value of the conduction band edge $\Delta E_{\rm ci} + q\phi_i^{(n+1)}$ is plugged back into the transport (4) and (10) in order to find the quantum corrector for the electron density. This procedure is repeated $N_{\rm pc}$ number of times until convergence of the electrostatic potential and electron density is achieved. Such a scheme makes Poisson's iterations more stable and decreases the number of times the quantum transport equations is solved.

The operation count for one bias point is $N_{\rm pc} \cdot (N_E \cdot (M^3 N_{\rm layer}) \cdot N_b + N_{\rm pp} \cdot (N^2))$, where the first term accounts for the outer Poisson's iterations, which require a solution of quantum transport equations, and the second for the inner Poisson's iterations of the predictor–corrector scheme. Here, the number of energy grid points $N_E \simeq 200 - 500$, number of spatial grid points $N = M \cdot N_{\rm layer} \simeq 1000 - 2000$, the number of outer Poisson's iterations $N_{\rm pc} \simeq 7 - 10$, the number of inner Poisson's iterations $N_{\rm pp} \simeq 3 - 5$, and the number of Born iterations $Nb \simeq 3 - 5$.

In the core of the computational flowchart lies the solution of the transport (4), (10), and (12) at N_E energy grid points in order to obtain carrier density and current. This is the

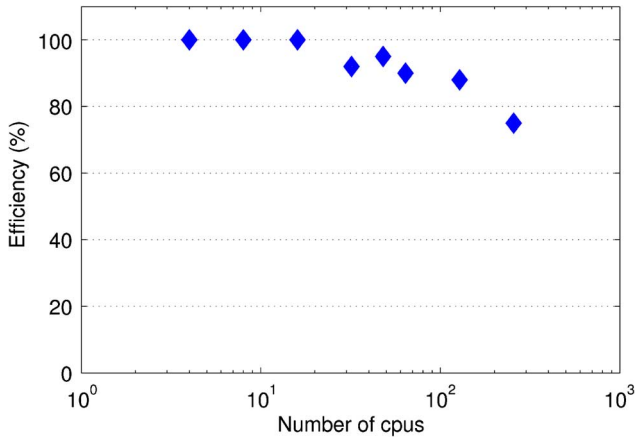


Fig. 10. Scaling efficiency of a single bias point calculation when the NEGF–Poisson's iteration is performed 16 times. The number of processors used varies from 1 to 256.

most critical part because the errors in carrier density may cause an oscillating or even divergent behavior of Poisson's iterations. In order to improve the integration of carrier density and current, a nonlinear energy grid is usually constructed by adding additional grid points near contact Fermi energies and resonant levels in the device. Ultimately, it is the increase in number of energy grid points N_E that makes the Poisson's iteration scheme converge. Since solving the equations at each energy requires the same computational time, a parallelization over N_E energy grid points is highly desirable. The results of OpenMP parallelization on a Columbia supercomputer [24] are shown in Fig. 10, where the scaling efficiency is defined as

$$\text{Efficiency}(N \text{ cpus}) = \frac{\text{Time}(N \text{ cpus})}{N \cdot \text{Time}(1 \text{ cpu})} \cdot 100\% \qquad (30)$$

is plotted versus the number of cpus. The scaling efficiency stays above 80% on up to 128 cpus. The slow decrease is partially due to the decrease of shared time taken by Poisson's equation. Additional parallelization of Poisson's solver may further improve the scaling efficiency. We note that MPI parallelization is also quite applicable.

VI. Challenges and Outlook

We discussed the NEGF approach within the effective mass Hamiltonian for silicon nanotransistors, as applied to 2-D and 1-D mode space simulators. The NEGF approach is, however, powerful enough to handle much more sophisticated Hamiltonians based on tight binding and density function theory approaches. Within the context of transistors, a 1-D simulator based on a tight binding Hamiltonian for silicon [25], carbon nanotubes [26], and molecules [27] has been demonstrated with varying degrees of success. Furthermore, the NEGF method can straightforwardly model Schottky barrier tunneling (from metal to nanotube) [28]–[30] and band-to-band tunneling in nanotube transistors, even in the presence of electron–phonon interaction [5], [30]. These intrinsically quantum mechanical features are accounted for without any extra effort. Furthermore, for someemerging nanomaterials, the effective mass Hamiltonian has a limited validity, except perhaps over a very small energy window close to the band bottom. The NEGF approach has proved useful in rapidly evaluating the current–voltage characteristics of experiments involving new channel materials. It must be admitted that most of these studies are based on performing ballistic simulations, which neglect the effect of electron–phonon and electron–electron scatterings.

The biggest theoretical advantage of the NEGF approach in modeling nanotransistors is in its ability to include phase-breaking scattering processes/electron–phonon and electron–electron interactions, in the presence of applied biases. That is, it gives us a clear quantum mechanical method to calculate the nonequilibrium electron density and is a unified approach in handling ballistic transport, tunneling, and scattering (electron–phonon) by solving the equations for G^r, $G^<$, and $G^>$. However, here, the NEGF approach is prohibitively expensive (operation count wise) in simulations involving more than 1-D models of transistors [12].

Semiclassical approaches to modeling nanotransistors also include quantum effects as a correction. These approaches are the density gradient (DG) [6], [7], quantum potential [8], and the quantum corrected-Monte Carlo (qc-MC) [31] methods. There have been benchmarking comparisons between our NEGF nanotransistor simulations and the DG method. The parameters used in the DG method were obtained by fitting to our quantum mechanically calculated electron density versus position. The results for the obtained current–voltage characteristics qualitatively agreed with the NEGF results, with a much better match in the subthreshold regime. The main difference between the NEGF and DG approaches is in prediction of ON currents, where the NEGF method is better in estimating the quasi-ballistic current. The agreement in the subthreshold regime, even for devices with channel lengths as small as 10 nm, is quite remarkable [9], especially given that the DG approach is orders of magnitude faster than the NEGF approach. More recently, there has been a comparison of qc-MC and the NEGF method in determining the role of scattering in nanoscale DGMOSFETs. The qc-MC approach accounts for quantization of energy levels in the body of the DGMOSFET [22]. These comparisons show that the qc-MC method agrees well with the NEGF approach, especially in determining the ON current. The MC approach, like the NEGF approach, can fully handle ballistic effects. The importance of drain-end electron–phonon scattering in quasi-ballistic nanotransistors is also captured by the qc-MC approach [22]. Currently, the MC approaches are superior in their ability to include anisotropic scattering due to charged impurities and electron–electron scattering. This is a subject that requires further attention by NEGF modelers. The agreement between the NEGF and qc-MC in the subthreshold region is less spectacular, probably because the MC approach is not particularly well suited in determining the OFF current. In comparison to the qc-MC approach, the benefit of the NEGF approach, so far, has been the ease with which tunneling is included, as demonstrated in Section III-A, and in transistors involving band-to-band and Schottky barrier tunneling.

Clearly, a single computational approach in accurately modeling the device physics of all varieties of nanotransistors has

not yet emerged. The drawbacks of the DG, quantum potential, and qc-MC methods are that they can all accurately capture only certain elements in the device physics of nanotransistors. The NEGF method, while not having any fundamental drawbacks, is in need of a major improvement from a computational viewpoint. We discuss some of the important needs here.

1) The first need is about the operation count required for the computation of electron and current densities for the block tridiagonal forms of the $A_{\rm DD}$ matrix that is defined in Section II. Calculation of the electron density only requires the diagonal elements of $G^<$ (9). However, all 1-, 2-, and 3-D ballistic simulations that we are aware of use the algorithm in [15], which is slow because it computes the diagonal blocks of $G^<$. Similarly, the calculation of the current density flowing between layers j and $j+1$ (see Fig. 2) requires only those elements of $G_{\rm jj+1}$ connecting grid points in layers j and $j+1$. The current flowing between layers j and $j+1$ (Fig. 2) can be obtained only from the N_x elements $G_{\rm jj+1}(i,i)$. However, our current algorithm [15] calculates all N_x^2 elements of $G_{\rm jj+1}$, which has an operation count that scales as $N_x^{2.7}N_yN_E$. The memory requirement of this algorithm is 5–7 $N_x^2N_y$. New methods that decrease the operation count to calculate the electron and current density by anywhere from N_x to N_x^2 will have a huge impact on modeling nanotransistors. We emphasize that a mere mathematical scaling with N_x that is better than $N_x^{2.7}N_yN_E$ with large multiplicative prefactors or requiring large memory (larger than the RAM) are not useful.
2) The second drawback is the need to include a more complete model for gate tunneling and scattering processes involving electron-impurity and electron–phonon and electron–electron interactions. All nanotransistor simulations, so far, only assume scattering processes in the isotropic limit. This assumption simplifies the form of the $A_{\rm DD}$ matrix because only the diagonal elements of the self-energies entering the $A_{\rm DD}$ matrix are nonzero. Similarly, off-diagonal elements of self-energy due to the gate that link layers j and j' far away from each other are nonzero. However, we set all self-energy elements, other than the first off diagonal elements, to zero so as to keep the $A_{\rm DD}$ matrix block tridiagonal.
3) The third area that needs further development is the need to model the effect of dopant atom fluctuations in nanotransistors. This is an important area of research, where modeling can help quantify threshold voltage fluctuations as channel lengths and gate/drain voltages become smaller [2]. There has been some limited progress here using extremely narrow cross section nanotransistors based on carbon nanotubes [32] or silicon nanowires [33]. However, progress in the case of large cross-sectional area devices is difficult. This is because a 2-D model will not capture the effect of discrete dopant atoms, while a 3-D model is difficult for computational reasons.
4) The fourth area where the NEGF method needs further development is transient modeling of nanotransistor switching. Solving the time-independent NEGF equations for steady state current–voltage characteristics will suffice for quasi-static analysis. Exploring the high-frequency limits will require solving the time-dependent NEGF equations, which is an unexplored area.
5) Finally, in modeling electron transport using the NEGF approach, it is usually assumed that the phonons are in thermal equilibrium. In the presence of electron transport, the phonon distribution will deviate from the Bose–Einstein factor if phonons emitted (by electron–phonon interaction) do not efficiently dissipate to the bulk contacts or substrate. Recent work has addressed this issue by self-consistently solving for both the nonequilibrium electron and phonon distributions using coupled Boltzmann transport equations for both electrons and phonons [34]–[37]. In a similar way, a potential direction for future work is the self-consistent calculation of both the electron and phonon correlation/distribution functions that are calculated using the NEGF method.

ACKNOWLEDGMENT

The authors would like to thank M. Meyyappan of the NASA Ames Research Center for the support, T. R. Govindan of the Army Research Office for the many useful discussions and collaboration, and M. Lundstrom of Purdue University for his interest in our MOSFET work and for the useful discussions. Supercomputer support from NASA Advanced Supercomputing (NAS) Division puts the authors in a unique position to carry out some of the work.

REFERENCES

[1] Semiconductor Industry Association, *The International Technology Roadmap for Semiconductors*, 2005, Austin, TX: Int. SEMATECH.
[2] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 μm MOSFETs: A 3D 'atomistic' simulation study," *IEEE Trans. Electron Devices*, vol. 45, no. 12, pp. 2505–2513, Dec. 1998.
[3] Y. Taur and T. H. Ning, *Fundamentals of VLSI Design*. Cambridge, U.K.: Cambridge Univ. Press, 1998.
[4] J. Appenzeller, Y.-M. Lin, J. Knoch, Z. Chen, and P. Avouris, "Comparing carbon nanotube transistors—The ideal choice: A novel tunneling device design," *IEEE Trans. Electron Devices*, vol. 52, no. 12, pp. 2568–2576, Dec. 2005.
[5] S. O. Koswatta, M. S. Lundstrom, M. P. Anantram, and D. E. Nikonov, "Simulation of phonon-assisted band-to-band tunneling in carbon nanotube field-effect transistors," *Appl. Phys. Lett.*, vol. 87, no. 25, pp. 253 107-1–253 107-3, Dec. 2005.
[6] M. G. Ancona and H. Iafrate, "Quantum corrections to the equation of state of an electron gas in a semiconductor," *Phys. Rev. B, Condens. Matter*, vol. 39, no. 13, pp. 9536–9539, May 1989.
[7] M. G. Ancona and H. F. Tiersten, "Macroscopic physics of the silicon inversion layer," *Phys. Rev. B, Condens. Matter*, vol. 35, no. 15, pp. 7959–7965, May 1987.
[8] D. K. Ferry, "Effective potentials and the onset of quantization in ultrasmall MOSFETs," *Superlattices Microstruct.*, vol. 28, no. 5, pp. 419–423, Nov. 2000.
[9] J. R. Watling, A. Brown, A. Asenov, A. Svizhenko, and M. P. Anantram, "Simulation of direct source-to-drain tunneling using the density gradient formalism: Non-equilibrium Greens function calibration," in *Proc. SISPAD*, 2002, pp. 267–270. IEEE Cat. 02TH8621.
[10] Z. Yu, D. Yergeau, R. Dutton, A. Svizhenko, and M. P. Anantram, "Quantum transport model for sub-100 nm CMOS devices," *Proc. SPIE*, vol. 4600, pp. 117–122, 2001.
[11] N. C. Kluksdahl, A. M. Kriman, D. K. Ferry, and C. Ringhofer, "Self-consistent study of the resonant-tunneling diode," *Phys. Rev. B, Condens. Matter*, vol. 39, no. 11, pp. 7720–7735, Apr. 1989.
[12] A. Svizhenko and M. P. Anantram, "Role of scattering in nanotransistors," *IEEE Trans. Electron Devices*, vol. 50, no. 6, pp. 1459–1466, Jun. 2003.

[13] M. Fischetti and S. Laux, "Long-range coulomb interactions in small Si devices. Part 1: Performance and reliability," *J. Appl. Phys.*, vol. 89, no. 2, pp. 1205–1231, Jan. 2001.
[14] M. Fischetti, "Long-range coulomb interactions in small Si devices. Part 2: Effective electron mobility in thin-oxide structures," *J. Appl. Phys.*, vol. 89, no. 2, pp. 1232–1250, Jan. 2001.
[15] A. Svizhenko, M. P. Anantram, T. R. Govindan, B. Biegel, and R. Venugopal, "Two-dimensional quantum mechanical modeling of nanotransistors," *J. Appl. Phys.*, vol. 91, no. 4, pp. 2343–2354, Feb. 2002.
[16] Massachusetts Institute of Technology, *Well-Tempered MOSFET*. [Online]. Available: http://www-mtl.mit.edu/researchgroups/Well
[17] *MEDICI Device Simulator*. [Online]. Available: http://www.synopsys.com/products/mixedsignal/taurus/device_sim_ds.html
[18] R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom, and D. Jovanovic, "Simulating quantum transport in nanoscale transistors: Real versus mode-space approaches," *J. Appl. Phys.*, vol. 92, no. 7, pp. 3730–3739, Oct. 2002.
[19] G. D. Mahan, "Quantum transport equation for electric and magnetic fields," *Phys. Rep.*, vol. 145, no. 5, pp. 251–318, 1987.
[20] B. K. Ridley, *Quantum Processes in Semiconductors*, 4th ed. New York: Oxford Univ. Press, 2000.
[21] M. S. Lundstrom, *Fundamentals of Carrier Transport*. Reading, MA: Addison-Wesley, 1990.
[22] H. Tsuchiya, A. Svizhenko, M. P. Anantram, M. Ogawa, and T. Miyoshi, "Comparison of non-equilibrium Green's function and quantum-corrected Monte Carlo approaches in nano MOS simulation," *J. Comput. Electron.*, vol. 4, no. 1/2, pp. 35–38, 2005.
[23] J. S. Martin, A. Bournel, and P. Dollfus, "On the ballistic transport in nanometer-scaled DG MOSFETs," *IEEE Trans. Electron Devices*, vol. 51, no. 7, pp. 1148–1155, Jul. 2004.
[24] *SGI NUMAflex Architecture, 20 SGI Altix 3700 Superclusters, Based on Intel Itanium Processors*. [Online]. Available: http://www.nas.nasa.gov/Resources/Systems/columbia.html
[25] T. Xia, L. F. Register, and S. K. Banerjee, "Quantum transport in double-gate MOSFETs with complex band structure," *IEEE Trans. Electron Devices*, vol. 50, no. 6, pp. 1511–1516, Jun. 2003.
[26] J. Guo, S. Datta, M. P. Anantram, and M. Lundstrom, "Atomistic simulation of carbon nanotube field-effect transistors using non-equilibrium Green's function formalism," *J. Comput. Electron.*, vol. 3, no. 3/4, pp. 373–377, 2004.
[27] Y. Xue, S. Datta, and M. Ratner, "First-principles based matrix Green's function approach to molecular electronic devices: General formalism," *Chem. Phys.*, vol. 281, no. 2/3, pp. 151–170, Aug. 2002.
[28] S. Heinze, J. Tersoff, R. Martel, V. Derycke, J. Appenzeller, and P. Avouris, "Carbon nanotubes as Schottky barrier transistors," *Phys. Rev. Lett.*, vol. 89, no. 10, pp. 106 801-1–106 801-4, Aug. 2002.
[29] J. Appenzeller, J. Knoch, M. Radosavljevic, and P. Avouris, "Multimode transport in Schottky-barrier carbon-nanotube field-effect transistors," *Phys. Rev. Lett.*, vol. 92, no. 22, pp. 226 802-1–226 802-4, Jun. 2004.
[30] A. Svizhenko, "Effect of scattering on transport and electrostatics in metallic CNT," *Phys. Rev. B, Condens. Matter*, vol. 72, pp. 085 430-1–085 430-10, 2005.
[31] B. Winstead and U. Ravaioli, "A quantum correction based on Schrodinger equation applied to Monte Carlo device simulation," *IEEE Trans. Electron Devices*, vol. 50, no. 2, pp. 440–446, Feb. 2003.
[32] N. Neophytou, D. Kienle, E. Polizzi, and M. P. Anantram, "Influence of defects on nanotube transistor performance," *Appl. Phys. Lett.*, vol. 88, no. 24, pp. 242 106-1–242 106-3, Jun. 2006.
[33] A. Martinez, M. Bescond, J. R. Barker, A. Svizhenko, M. P. Anantram, C. Millar, and A. Asenov, "A self-consistent full 3D real-space NEGF simulator for studying non-perturbative effects in nano-MOSFETs," *IEEE Trans. Electron Devices*, vol. 54, no. 9, pp. xxxx–xxxx, Sep. 2007.
[34] R. P. Joshi and D. K. Ferry, "Hot-phonon effects and interband relaxation processes in photoexcited GaAs quantum wells," *Phys. Rev. B, Condens. Matter*, vol. 39, no. 2, pp. 1180–1187, Jan. 1989.
[35] J. M. Barker, D. K. Ferry, S. M. Goodnick, D. D. Koleske, A. Allerman, and R. J. Shul, "High field transport in GaN/AlGaN heterostructures," *J. Vac. Sci. Technol. B, Microelectron. Process. Phenom.*, vol. 22, no. 4, pp. 2045–2050, Jul. 2004.
[36] M. Lazzeri and F. Mauri, "Coupled dynamics of electrons and phonons in metallic nanotubes: Current saturation from hot-phonon generation," *Phys. Rev. B, Condens. Matter*, vol. 73, no. 16, pp. 165419.1–165419.6, Apr. 2006.
[37] S. Hasan, M. A. Alam, and M. Lundstrom, "Simulation of carbon nanotube fets including hot-phonon and self-heating effects," in *IEDM Tech. Dig.*, 2006, pp. 31–34.



M. P. Anantram received the B.Sc. degree in applied sciences from the PSG College of Technology, Coimbatore, India, in 1986, the M.Sc. degree in physics from the University of Poona, Pune, India, in 1989, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, Indiana, in 1995.

He was the Group Lead for Computational Nanoelectronics and, since 2004, a Fellow with the University Affiliated Research Center (UARC), National Aeronautics and Space Administration Ames Research Center, where his group developed the first 2-D quantum simulator for nanotransistors, and he codeveloped an algorithm to compute the charge and current density in nanodevices. He has also worked extensively on the modeling and prediction of electrical and electromechanical properties of nanostructures. Since May 2006, he has been a Professor with the Nanotechnology Engineering Group, Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. His research encompasses theory and computational modeling of semiconductor and molecular nanodevices.

Dr. Anantram is an Associate Editor of the IEEE TRANSACTIONS ON NANOTECHNOLOGY and the Education Chair of the IEEE Nanotechnology Council. He serves on the program and organizing committees of nanotechnology conferences. He has received best paper awards for his work on nanotransistors and electromechanical properties of nanotubes. He coreceived the Computer Sciences Corporation's Highest Technical Achievement in Applied Sciences in 2003.



A. Svizhenko received the B.S. and M.S. degrees from the Moscow Institute of Physics and Technology, Moscow, Russia, in 1995 and 1993, respectively, and the Ph.D. degree in electrical engineering from the University of Nebraska, Lincoln, in 1999.

From 1999 to 2005, he was a Staff Scientist with the National Aeronautics and Space Administration Ames Center for Nanotechnology, Moffett Field, CA. From 2005 to 2006, he was a Research Scientist with Stanford University. Since 2006, he has been with Silvaco International, Santa Clara, CA, as a Senior TCAD Engineer. His research interests include modeling and computation of quantum transport in nanotransistors, silicon nanowires, and carbon nanotubes.

Dr. Svizhenko was the recipient of the 2000 NASA Group Achievement Award and the 2003 and 2004 Computer Sciences Corporation Best Paper Award.