CONTRIBUTED
P A P E R

Modeling of Nanoscale Devices

Devices and components subject to quantum and atomistic effects, such as layered semiconductor structures, nanoscale transistors, carbon nanotubes and nanowires may be modeled using quantum analysis and simulation methods.

By M. P. Anantram, Mark S. Lundstrom, *Fellow IEEE*, and Dmitri E. Nikonov, *Senior Member IEEE*

ABSTRACT | We aim to provide engineers with an introduction to the nonequilibrium Green's function (NEGF) approach, which is a powerful conceptual tool and a practical analysis method to treat nanoscale electronic devices with quantum mechanical and atomistic effects. We first review the basis for the traditional, semiclassical description of carriers that has served device engineers for more than 50 years. We then describe why this traditional approach loses validity at the nanoscale. Next, we describe semiclassical ballistic transport and the Landauer–Buttiker approach to phase-coherent quantum transport. Realistic devices include interactions that break quantum mechanical phase and also cause energy relaxation. As a result, transport in nanodevices is between diffusive and phase coherent. We introduce the NEGF approach, which can be used to model devices all the way from ballistic to diffusive limits. This is followed by a summary of equations that are used to model a large class of structures such as nanotransistors, carbon nanotubes, and nanowires. Applications of the NEGF method in the ballistic and scattering limits to silicon nanotransistors are discussed.

KEYWORDS | Electron transport; Green's function; nanoelectronics; nonequilibrium; phonons; quantum mechanics; scattering; semiconductors; simulation; transistor

Manuscript received February 23, 2007; revised April 8, 2008. M. S. Lundstrom was supported by the National Science Foundation through the Network for Computational Nanotechnology and by the semiconductor industry through the Semiconductor Research Consortium, the Focus Center on Materials, Structures, and Devices, and the Nanoelectronics Research Initiative. M. P. Anantram was supported by NASA Ames Research Center and National Institute of Standards and Technology (Gaithersburg).
M. P. Anantram is with the Nanotechnology Program, Electrical and Computer Engineering Department, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: anant@uwaterloo.ca).
M. S. Lundstrom is with the Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 49097 USA (e-mail: lundstro@purdue.edu).
D. E. Nikonov is with the Intel Corporation, Santa Clara, CA 95052 USA (e-mail: dmitri.e.nikonov@intel.com).

Digital Object Identifier: 10.1109/JPROC.2008.927355

I. INTRODUCTION

Semiconductor devices operate by controlling the flow of electrons and holes through a device, and our understanding of charge carrier transport has both benefited from and driven their development. When Shockley wrote, "Electrons and Holes in Semiconductors" [22], semiconductor physics was at the frontier of research in condensed matter physics. Over the years, the essential concepts were clarified and simplified into the working knowledge of device engineers. The treatment of electrons and holes as semiclassical particles with an effective mass was usually adequate. Electronic devices were made of materials (e.g., silicon, gallium arsenide, etc.) with properties (e.g., bandgap, effective mass, etc.) that could be looked up. For most devices, the engineer's drift-diffusion equation provided a simple but adequate description of carrier transport. Today things are changing. Device dimensions have shrunk to the nanoscale. The properties of materials can be engineered by intentional strain and size effects due to quantum confinement. Devices contain a countable number of dopants and are sensitive to structure at the atomistic scale. In addition to familiar devices like the metal–oxide–semiconductor field-effect transistor (MOSFET), which have been scaled to nanometer dimensions, new devices built from carbon nanotubes, semiconductor nanowires, and organic molecules are being explored. Device engineers will need to learn to think about devices differently. To describe carrier transport in nanoscale devices, engineers must learn how to think about charge carriers as quantum mechanical entities rather than as semiclassical particles, and they must learn how to think at the atomistic scale rather than at a continuum one. Our purpose in this paper is to provide engineers with an introduction to the nonequilibrium Green's function (NEGF) approach [5], [8], [9], which provides a powerful conceptual tool and a practical analysis method to treat small electronic devices quantum mechanically and atomistically. We first review the basis for the traditional, semiclassical description of carriers that has served device engineers for more than 50 years in Section II. We then describe why this

0018-9219/$25.00 ©2008 IEEE

traditional approach loses validity at the nanoscale. Next, we describe semiclassical ballistic transport in Section III and the Landauer–Buttiker approach to phase-coherent quantum transport in Section IV. Realistic devices include interactions that break quantum mechanical phase and also cause energy relaxation. As a result, transport in nanodevices is between diffusive and phase-coherent. We introduce the NEGF approach, which can be used to model devices all the way from ballistic to diffusive limits, in Section V. This is followed by a summary of equations that are used to model a large class of structures such as nanotransistors, carbon nanotubes, and nanowires in Section VI. An application of the NEGF method in the ballistic and scattering limits to silicon nanotransistors is discussed in Sections VII and VIII, respectively. We conclude with a summary in Section IX. The Dyson's equations and algorithms to solve for the Green's functions of layered structures developed in [23] are presented in Appendixes II and III. These appendixes can be skipped by the reader whose aim is to gain a basic understanding of the NEGF approach to device modeling.

II. SEMICLASSICAL TRANSPORT: DIFFUSIVE

Electrical engineers have commonly treated electrons as semiclassical particles that move through a device under the influence of an electric field and random scattering potentials. As sketched in Fig. 1, electrons move along a trajectory in phase space (position and momentum space). In momentum space, the equation of motion looks like Newton's law for a classical particle

$$\frac{d\hbar\vec{k}}{dt} = -\nabla_r E_C(\vec{r},t) \tag{1}$$

where $\vec{k}$ is the crystal momentum and E_C is the bottom of the conduction band. In position space, the equation of motion is

$$\frac{d\vec{r}}{dt} = \frac{1}{\hbar}\nabla_k E(\vec{k}) \tag{2}$$

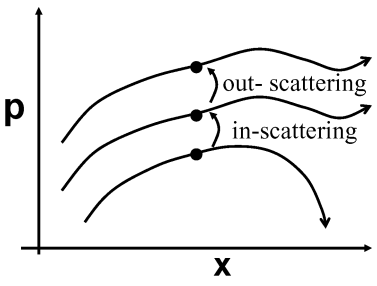


Fig. 1. *Carrier trajectories in phase space showing free flights along a trajectory interrupted by scattering events that begin another free flight. p is momentum and x is position.*

where $E(k)$ describes the band structure of the semiconductor. The right-hand side of (2) is simply the velocity of the semiclassical particle, and in the simplest case it is just $\hbar k/m^*$. By solving (1) and (2), we trace the trajectory of a carrier in phase space as shown in Fig. 1.

Equations (1) and (2) describe the ballistic transport of semiclassical carriers. In practice, carriers frequently scatter from various perturbing potentials (defects, ionized impurities, lattice vibrations, etc.). The result is that carriers hop from one trajectory in phase space to another as shown in Fig. 1. The average distance between scattering events, the mean-free-path l, has (until recently) been much smaller than the critical dimensions of a device. Carriers undergo a random walk through a device with a small bias in one direction imposed by the electric field. To describe this scattering-dominated (so-called diffusive) transport, we should add a random force $(F_S(\vec{r},t))$ to the right-hand side of (1)

$$\frac{d\hbar\vec{k}}{dt} = -\nabla_r E_C(\vec{r},t) + F_S(\vec{r},t). \tag{3}$$

It is relatively easy to solve (1) and (3) numerically. One solves the equations of motion (1) and (2) to move a particle through phase space. Random numbers are chosen to mimic the scattering process and occasionally kick a carrier to another trajectory. By averaging the results for a large number of simulated trajectories, these so-called Monte Carlo techniques provide a rigorous, though computationally demanding, description of carrier transport in devices as described in [7].

Device engineers are primarily interested in average quantities such as the average electron density, current density, etc. (There are some exceptions; noise is important too.) Instead of simulating a large number of particles, we can ask: what is the probability that a state at position $\vec{r}$, with momentum $\hbar\vec{k}$, is occupied at time t? The answer is given by the distribution function, $f(\vec{r},\vec{k},t)$, which can be computed by averaging the results of a large number of simulated trajectories. Alternatively, we can adopt a collective viewpoint instead of the individual particle viewpoint and formulate an equation for $f(\vec{r},\vec{k},t)$. The result is known as the Boltzmann transport equation (BTE)

$$\frac{\partial f}{\partial t} + \vec{v}\cdot\nabla_r f - \frac{q\vec{E}}{\hbar}\nabla_k f = \hat{C}f \tag{4}$$

where $\vec{E}$ is the electric field and $\hat{C}f$ describes the effects of scattering. In equilibrium, $f(\vec{r},\vec{k},t)$ is simply the Fermi function, but in general, we need to solve (4) to find f. Once $f(\vec{r},\vec{k},t)$ is known, quantities of interest to the device engineer are readily found. For example, to find the average electron density in a volume Ω centered at

position $\vec{r}$, we simply add up the probability that all of the states in Ω are occupied and divide by the volume

$$n(\vec{r},t)=\frac{1}{\Omega}\sum_{k} f(\vec{r},\vec{k},t). \tag{5}$$

Similarly, we find

Current Density

$$\vec{J}(\vec{r},t)=\frac{1}{\Omega}\sum_{k}(-q)\vec{v}f(\vec{r},\vec{k},t) \tag{6}$$

Kinetic Energy Density

$$W(\vec{r},t)=\frac{1}{\Omega}\sum_{k} E(\vec{k})f(\vec{r},\vec{k},t) \tag{7}$$

Energy Current Density

$$\vec{J}_E(\vec{r},t)=\frac{1}{\Omega}\sum_{k} E(\vec{k})\vec{v}f(\vec{r},\vec{k},t) \tag{8}$$

where q is the absolute value of the electron charge. This approach provides a clear and fairly rigorous description of semiclassical carrier transport, but solving the six-dimensional BTE is enormously difficult. One might ask if we cannot just find a way to solve directly for the quantities of interest in (5)–(8). The answer is yes, but some simplifying assumptions are necessary.

Device engineers commonly describe carrier transport by a few low-order moments of the Boltzmann transport equation (4). A mathematical prescription for generating moment equations exists, but to formulate them in a tractable manner, numerous simplifying assumptions are required [13]. Moment equations provide a phenomenological description of transport that gives insight and quantitative results when properly calibrated.

The equation for the zeroth moment of $f(\vec{r},\vec{k},t)$ gives the well-known continuity equation for the electron density $n(\vec{r},t)$

$$\frac{\partial n(\vec{r},t)}{\partial t}=-\nabla_r F_n+G_n-R_n \tag{9}$$

where F_n is the electron flux

$$J_n=-qF_n. \tag{10}$$

G_n the electron generation rate and R_n the electron recombination rate. Equation (9) states that the electron density at a location increases with time if there is a net flux of electrons into the region (as described by the first term minus the divergence of the electron flux) or if carriers are being generated there. Recombination causes the electron density to decrease with time. Any physical quantity must obey a conservation law like (9).

The equation for the first moment of $f(\vec{r},\vec{k},t)$ gives the equation for average current density [(6)] projected on the x-axis

$$\frac{\partial J_{nx}}{\partial t}=\frac{2q}{m}\frac{dW_{xx}}{dx}+\frac{nq^2}{m^*}E_x-\frac{J_{nx}}{\tau_m}. \tag{11}$$

Each term on the right-hand side of (11) is analogous to the corresponding terms in (9). The current typically changes slowly on the scale of the momentum relaxation time τ_m (of order of subpicosecond time) so the time derivative can be ignored and (11) solved for

$$J_{nx}=nq\mu_n E_x+\frac{2}{3}\mu_n\frac{dW_{xx}}{dx} \tag{12}$$

where

$$\mu_n=\frac{q\tau_m}{m^*} \tag{13}$$

is the electron mobility and we have assumed equipartition of energy so that $W_{xx}=W/3$, where W is the total kinetic energy density. This assumption can be justified when there is a lot of isotropic scattering, which randomizes the carrier velocity. Equation (12) is a drift-diffusion equation; it says that electrons drift in electric fields and diffuse down kinetic energy gradients. Near equilibrium

$$W=\frac{3}{2}nk_BT \tag{14}$$

so when T is uniform, (12) becomes

$$J_{nx}=nq\mu_n E_x+k_BT\mu_n\frac{dn}{dx}=nq\mu_n E_x+qD_n\frac{dn}{dx} \tag{15}$$

the drift-diffusion equation. By inserting (15) in the electron continuity equation (9), we get an equation for the electron density that can be solved for the electron density within a device. This is the traditional and still most common approach for describing transport in semiconductor devices [18].

Since most devices contain regions with high electric fields, the assumption that $W=3nk_BT/2$ is not usually a good one. The carrier energy enters directly into the

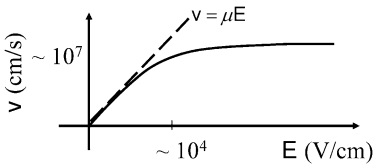


Fig. 2. *The average velocity (v) versus electric field (E) for electrons in bulk silicon at room temperature.*

second term of the transport equation (12) but also enters indirectly because the mobility is energy dependent. To treat transport more rigorously, we need an equation for the electron energy.

The second moment of $f(\vec{r},\vec{k},t)$ gives the carrier energy density $W(\vec{r},t)$ according to (8). The second moment of the BTE gives a continuity equation for the energy density [13]

$$\frac{\partial W}{\partial t} = -\frac{dJ_W}{dx} + J_{nx}E_x - \frac{W - W_0}{\tau_E} \tag{16}$$

where W_0 is the equilibrium energy density and τ_E the energy relaxation time. Note that the energy relaxation time is generally longer than the momentum relaxation time because phonon energies are small, so that it takes several scattering events to thermalize an energetic carrier but only one to randomize its momentum.

To solve (16), we need to specify the energy current. The third moment of $f(\vec{r},\vec{k},t)$ gives the carrier energy flux $J_W(\vec{r},t)$ according to (8). The third moment of the BTE gives a continuity equation for the energy flux

$$J_W = W\mu_E E_x + \frac{d(D_E W)}{dx} \tag{17}$$

where μ_E and D_E are appropriate energy transport mobility and diffusion coefficient [13].

Equations (9) and (15)–(17) can now be solved self-consistently to simulate carrier transport. Fig. 2 sketches the result for uniformly doped, bulk silicon with a constant electric field. At low electric fields, $W \sim 3nk_BT/2$, and we find that $\langle v_x \rangle \sim \mu_n E_x$. For electric fields above $\sim 10^4$ V/cm, the kinetic energy increases, which increases the rate of scattering and lowers the mobility so that at high fields, the velocity saturates at $\sim 10^7$ cm/s. In a bulk semiconductor, there is a one-to-one relation between the magnitude of the electric field and the kinetic energy, so the mobility and diffusion coefficient can be parametrized as known functions of the local electric field. The result is that, for bulk semiconductors or for large devices in which the electric field changes slowly, there is no need to solve all four equations; we need to solve the carrier continuity and drift-diffusion equations with field-dependent parameters.

Electric fields above 10^4 V/cm are common in nanoscale devices. This is certainly high enough to cause velocity saturation in the bulk, but in a short, high field region, transients occur. Fig. 3 illustrates what happens for a hypothetical situation in which the electric field abruptly jumps from a low value to a high value and then back to a low value again. Electrons injected from the low field region are accelerated by the high electric field, but energy relaxation times are longer than momentum relaxation times, so the energy is slow to respond. The result is that the mobility is initially high (even though the electric field is high), so the velocity can be higher than the saturated value shown in Fig. 2. As the kinetic energy increases, however, scattering increases, the mobility drops, and the velocity eventually decreases to $\sim 10^7$ cm/s, the saturated velocity for electrons in bulk silicon. The spatial width of the transient is roughly 100 nm; modern devices frequently have dimensions on this order, and strong velocity overshoot should be expected.

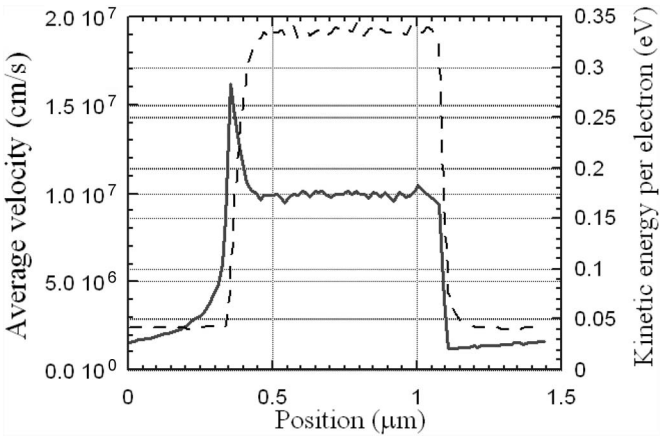


Fig. 3. *The average steady-state velocity (solid line) and kinetic energy (dashed line) versus position for electrons injected into a short slab of silicon with low-high-low electric field profile.*

The example shown in Fig. 3 demonstrates that it is better to think of the mobility and diffusion coefficient as functions of the local kinetic energy rather than the local electric field. What this means is that (9) and (15)–(17) should all be solved self-consistently to simulate carrier transport in small devices. Device simulation programs commonly permit two options: 1) the solution of (9) and (15) self-consistently with the Poisson equation using mobilities and diffusion coefficients that depend on the local electric field (the so-called drift-diffusion approach) or 2) the solution of (9) and (15)–(17) self-consistently with the Poisson equation using mobilities and diffusion coefficients that depend on the local kinetic energy (so-called energy transport or hydrodynamic approaches). Solving the four equations self-consistently is more of a computational burden, but it is necessary for when the electric field changes rapidly on the scale of a mean-free-path for scattering. Actually, numerous simplifying assumptions are also necessary to even write the current and energy flux equations as (15) and (17) [13]. The most rigorous (and computationally demanding) simulations (so-called Monte Carlo simulations) go back to the individual particle picture and track carriers trajectories according to (2) and (3). Drift-diffusion and energy transport approaches for treating carrier transport in semiconductor devices have two things in common: the first is the assumption that carriers can be treated as semiclassical particles and the second is the assumption that there is a lot of scattering. Both of these assumptions are losing validity as devices shrink.

III. SEMICLASSICAL TRANSPORT: BALLISTIC

Consider the "device" sketched in Fig. 4(a), which consists of a ballistic region attached to two leads. The left lead (source) injects a thermal equilibrium flux of carriers into the device; some carriers reflect from the potential barriers within the device, and the rest transmit across and enter the right lead (drain). A similar statement applies to the drain lead. The source and drain leads are assumed to be perfect absorbers, which means that carriers impinging them from the device travel without reflecting back into the device. To compute the electron density, current, average velocity, etc., within the device, we have two choices. The first choice treats the carriers as semiclassical particles, and the Boltzmann equation is solved to obtain the distribution function $f(\vec{r}, \vec{k}, t)$, as discussed in the previous section. The second choice treats the carriers quantum mechanically, as discussed in the next section. In this section, we will use a semiclassical description in which the local density-of-states within the device is just that of a bulk semiconductor but shifted up or down by the local electrostatic potential. This approximation works well when the electrostatic potential does not vary too rapidly, so that quantum effects can be ignored. To

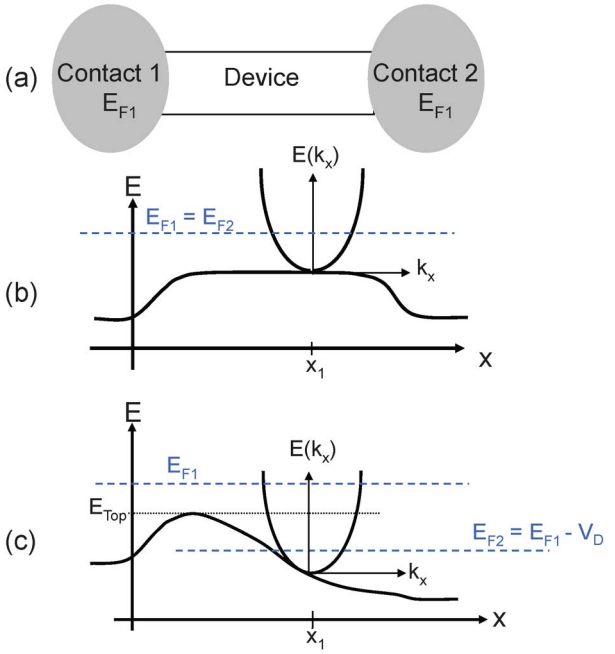


Fig. 4. *Sketch of a ballistic device with two leads that function as reservoirs of thermal equilibrium carriers. (a) Device and the two leads. (b) Energy band diagram under equilibrium conditions ($V_D = 0$). (c) Energy band diagram under bias ($V_D > 0$).*

find how the k-states within the ballistic device are occupied, we solve the Boltzmann transport equation (4). Because the device is ballistic, there is no scattering, and $\hat{C}f = 0$. It can be shown [13] that the solution to the BTE with $\hat{C}f = 0$ is a function of the electron's total energy

$$E = E_C(x) + E(k) \tag{18}$$

where $E_C(x)$ is the conduction band minimum versus position and $E(k)$ is the band structure for the conduction band. We know that under equilibrium conditions sketched in Fig. 4(b), the proper function of total energy is the Fermi function

$$f(E) = \frac{1}{1 + \exp\left(\dfrac{E - E_F}{k_B T}\right)} \tag{19}$$

where the Fermi level E_F and temperature T are constant in equilibrium.

Now consider the situation in Fig. 4(c), where a drain bias has been applied to the ballistic device. Although two thermal equilibrium fluxes are injected into the device, it is now very far from equilibrium. Since scattering is what drives the system to equilibrium, the ballistic device is as far from equilibrium as it can be. Nevertheless, for the

ballistic device, the relevant steady-state Boltzmann equation is the same equation as in equilibrium. The solution is again a function of the carrier's total kinetic energy. At the leads, we know that the solution is a Fermi function, which specifies the functional dependence on energy. For the ballistic device, therefore, the probability that a k-state is occupied is given by an equilibrium Fermi function. The only difficulty is that we have two Fermi levels, so we need to decide which one to use.

Return again to Fig. 4(c) and consider how to fill the states at $x = x_1$. We know that the probability that a k-state is occupied is given by a Fermi function, so we only need to decide which Fermi level to use for each k-state. For the positive k-states with energy above E_{TOP}, the top of the energy barrier, the states can only have been occupied by injection from the source, so the appropriate Fermi level to use is the source Fermi level. Similarly, negative k-states with energy above E_{TOP} can only be occupied by injection from the drain, so the appropriate Fermi level to use is the drain Fermi level. Finally, for k-states below E_{TOP}, both positive and negative velocity states are populated according to the drain Fermi level. The negative velocity k-states are populated directly by injection from the drain, and the positive k-states are populated when negative velocity carriers reflect from the potential barrier.

Ballistic transport can be viewed as a special kind of equilibrium. Each k-state is in equilibrium with the lead from which it was populated. Using this reasoning, one can compute the distribution function and any moment of it (e.g., carrier density, carrier velocity, etc.) at any location within the device. Fig. 5 shows that computed distribution function in a ballistic nanoscale MOSFET under high gate and drain bias [21]. A strong ballistic peak develops as carriers injected from the source are accelerated in the high electric field near the drain. Each k-state is in equilibrium with one of the two leads, but the overall carrier distribution is very different from the equilibrium Fermi–Dirac distribution. When scattering dominates, carriers quickly lose their "memory" of which lead they were injected from, but for ballistic transport, there are two separate streams of carriers: one injected from the source and one from the drain.

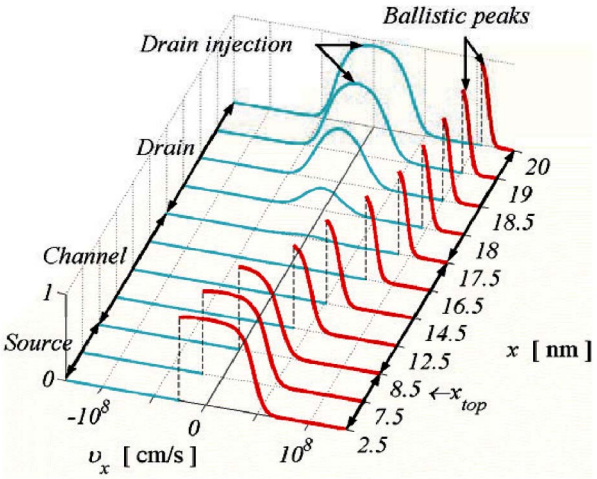


Fig. 5. *The computed carrier distribution function within a nanoscale MOSFET under high gate and drain bias. (From [21].)*

To evaluate the electron density versus position within the ballistic device, we should compute a sum like (5), but we must do two sums: one for the states filled from the left lead and another for the states filled from the right lead

$$n(x) = 2\sum_{k_L} \Theta_L(x) f_L(E) + 2\sum_{k_R} \Theta_R(x) f_R(E) \qquad (20)$$

where f_L and f_R are the equilibrium Fermi functions of leads L and R and $\Theta_{L,R}(x)$ is a function that selects out the k-states at position x that can be filled by lead L or R according to the procedure summarized in Fig. 4. The factors of "2" in (20) correspond to summation over spin states. It is often convenient to do the integrals in energy space rather than in k-space, in which case (20) becomes

$$n(x) = \int dE[\mathrm{LDOS}_L(x,E) f_L(E) + \mathrm{LDOS}_R(x,E) f_R(E)] \qquad (21)$$

where $\mathrm{LDOS}_{L,R}(x,E)$ is the local density of states at energy E, fillable from lead L or R. The density of state contains summation over spin states. For diffusive transport, we deal with a single density-of-states and fill it according to a source quasi-Fermi level, but for ballistic devices, the density of states separates into parts fillable from each lead.

The current flowing from source to drain (drain to source) lead is simply the transmission probability $T(E)$ times the Fermi function of the source (drain) lead. The net current flowing in the device is then

$$I = \frac{2e}{h}\int dE T(E)[f_L(E) - f_R(E)]. \qquad (22)$$

For the semiclassical example of Fig. 4, $\mathrm{T(E)} = 0$ for $E < E_{TOP}$ and $\mathrm{T(E)} = 1$ for $E > E_{TOP}$.

IV. PHASE COHERENT QUANTUM TRANSPORT: THE LANDAUER–BUTTIKER FORMALISM

Quantum mechanically, the electron is a wave and the wave function $\Psi(\vec{r})$ is obtained by solving Schrodinger's equation

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r})\right]\Psi(\vec{r}) = E\Psi(\vec{r}) \qquad (23)$$

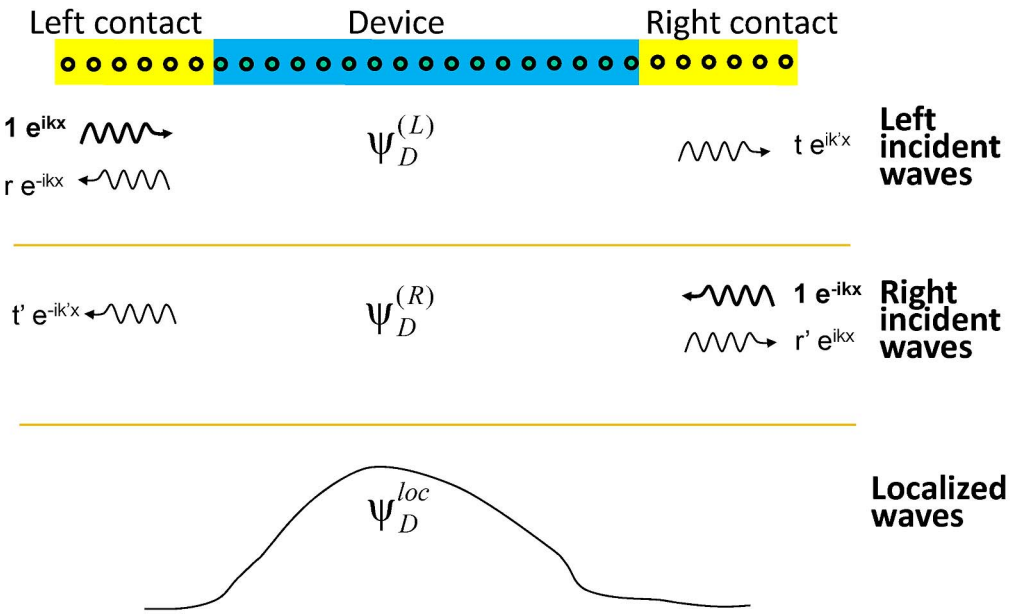


Fig. 6. ***All wave functions in the device can be represented as left incident*** $(\Psi_D^{(L)})$***, right incident*** $(\Psi_D^{(R)})$***, or localized states*** $(\Psi_D^{(\mathrm{loc})})$**.**

where E is the energy. Consider a device connected to two leads as shown in Fig. 6, where the leads are assumed to have a constant electrostatic potential. In a manner identical to the discussion of semiclassical modeling in the previous section, where we considered corpuscular electrons incident from the left and right leads, in quantum mechanical modeling, we need to consider electron waves incident from the left and right leads. The electron wave function in device region D can be thought to arise from the following.

- Waves incident from the left lead (L) of the form e^{ikx}, which have transmitted and reflected components $te^{ik'x}$ and re^{-ikx} in the right and left leads, respectively. The wave function in the device region D due to this wave is represented by $\Psi_D^{(L)}$. More properly, we must attach a subscript k (or E) to denote which state (or energy level) in the left lead induced the state $\Psi_D^{(L)}$ in the device, but we leave it out for compactness.
- Waves incident from the right lead (R) of the form $e^{-ik'x}$, which have transmitted and reflected components $t'e^{-ikx}$ and $r'e^{ik'x}$ in the left and right leads, respectively. The wave function in the device region D due to this wave is represented by $\Psi_D^{(R)}$. Again k (or E) is implicit.
- States localized in device region D represented by $\Psi_D^{(\mathrm{loc})}$. Localized and quasi-localized states are filled up by scattering due to electron–phonon and electron–electron interaction. We will assume here that localized states are absent.

The Landauer–Buttiker approach expresses the expectation value of an operator in terms of the left and right incident electrons from the leads and their distribution functions. The expectation value of operator $\hat{Q}$ is

$$Q = \sum_{k,s} \left[\left\langle \Psi_D^{(L)} | \hat{Q} | \Psi_D^{(L)} \right\rangle f_L(E) + \left\langle \Psi_D^{(R)} | \hat{Q} | \Psi_D^{(R)} \right\rangle f_R(E) \right]. \tag{24}$$

Here the summation is performed over the momentum k and the spin s states.

The operator $\hat{Q}$ can be just the number 1, in which case the summation gives the number of occupied states; or it can be the momentum operator $-i\hbar(d/dx)$. The crucial point here is that we are able to simply add the contributions from the left and right leads in the absence of scattering. One may argue that this cannot be done since electrons are subject to Pauli's principle and two electrons injected into the device cannot occupy the same state at the same time. However, since the Hamiltonian of the whole system (drain, source, and reservoirs) is Hermitian, the states injected from the source and the drain are orthogonal to each other. Since they are now two distinct states with no overlap, two electrons can occupy them without violating Pauli's principle. Another implication of the orthogonality of the source- and drain-injected states is that, instead of taking a linear combination of $\Psi_D^{(L)}$ and $\Psi_D^{(R)}$ to extract the expectation values of $\hat{Q}$, we can consider them separately (this is not true if we have a new

interaction in the original phase-coherent Hamiltonian H, like the electron–phonon interaction, which couples a source-injected state and a drain-injected state).

Since in this paper we do not consider any spin-dependent phenomena, the spin summation translates into a factor of two. Equation (24) has contributions from two physically different sources. The first term corresponds to contribution from waves incident from the left lead ($\Psi_D^{(L)}$) at energy E, weighted by the Fermi factor of the left lead (f_L). The second term corresponds to waves incident from the right lead ($\Psi_D^{(R)}$) weighted by the Fermi factor of the right lead (f_R). More generally, if device region D is connected to a third lead G, then the expectation value of operator $\hat{Q}$ is

$$Q = \sum_{k,s}\left[\left\langle \Psi_D^{(L)}|\hat{Q}|\Psi_D^{(L)}\right\rangle f_L(E) + \left\langle \Psi_D^{(R)}|\hat{Q}|\Psi_D^{(R)}\right\rangle f_R(E) + \left\langle \Psi_D^{(G)}|\hat{Q}|\Psi_D^{(G)}\right\rangle f_G(E)\right] \quad (25)$$

where $\Psi_D^{(G)}$ corresponds to the wave function in the Device due to waves incident from lead G and f_G is the Fermi factor of lead G. Using (24), the contribution to electron (n) and current (J) densities at x in the device region D are given by (neglecting gate lead)

$$n(x) = \sum_{k,s}\left[\left|\Psi_D^{(L)}(x)\right|^2 f_L(E) + \left|\Psi_D^{(R)}(x)\right|^2 f_R(E)\right] \quad (26)$$

and

$$J(x) = \sum_{k,s}\frac{e\hbar}{2mi}\left[\Psi_D^{(L)}(x)^\dagger \frac{d\Psi_D^{(L)}(x)}{dx} f_L(E) + \Psi_D^{(R)}(x)^\dagger \frac{d\Psi_D^{(R)}(x)}{dx} f_R(E) - \text{c.c}\right] \quad (27)$$

where "c.c." represents complex conjugate of the terms to its left. The quantum mechanical density of states at energy E due to waves incident from the left (LDOS_L) and right (LDOS_R) are

$$\text{LDOS}_L(x,E) = 2\sum_{k_l}\left|\Psi_D^{(L)}(x)\right|^2 \quad (28)$$

$$\text{LDOS}_R(x,E) = 2\sum_{k_r}\left|\Psi_D^{(R)}(x)\right|^2 \quad (29)$$

where k_l and k_r are states with energy E incident from the left and right leads, respectively. Then the electron density can be written in the same form as (21)

$$n(x) = \int \text{LDOS}_L(x,E) f_L(E) dE + \int \text{LDOS}_R(x,E) f_R(E) dE \quad (30)$$

except that the expressions for the local density of states are different. Similarly, (27) can be expressed in a form identical to (22). The above formalism can be extended to calculate noise (shot and Johnson-Nyquist) in nanodevices [3]. Device modeling in the phase-coherent limit involves solving Schrodinger's equation to obtain the electron density self-consistently with Poisson's equation.

V. QUANTUM TRANSPORT WITH SCATTERING: THE NEED FOR GREEN'S FUNCTIONS

The description in the previous section is valid only in the phase-coherent limit. The terminology "phase coherent" refers to a deterministic evolution of both the amplitude and phase of $\Psi_n(\vec{r})$ as given by Schrodinger's equation. The quantum mechanical wave function evolves phase coherently only in the presence of rigid scatterers, a common example of which is the electrostatic potential felt by an electron in the device. The wave function of an electron loses phase coherence due to scatterers that have an internal degree of freedom such as phonons. Phase-incoherent scattering involves irreversible loss of phase information to phonon degrees of freedom. Naturally, including loss of phase information is important when device dimensions become comparable to the scattering lengths due to phonons and other phase-breaking mechanisms. Accurate modeling of nanodevices should have the ability to capture:

- interference effects;
- quantum mechanical tunneling;
- discrete energy levels due to confinement in two-dimensional (2-D) and three-dimensional device geometries;
- scattering mechanisms (electron–phonon, electron–electron).

The first three effects can be modeled by solving Schrodinger's equation in a rigid potential as discussed in Section IV. While in the semiclassical device modeling, the Boltzmann equation accounts for the energy and momentum relaxation due to scattering mechanisms, in quantum mechanical device modeling, the NEGF approach is necessary to account for energy, momentum, and quantum mechanical phase relaxation.

The semiclassical approaches to transport can be derived from quantum mechanics [2], [10]. In its current form of implementation, the quantum mechanical approaches,

however, need a considerable amount of work to model a broad class of realistic devices. We will now briefly discuss three topics that require additional work. The first topic is the inclusion of scattering mechanisms between electrons and phonons and other electrons (including plasmons). Here, the current implementations of scattering mechanisms in the Monte Carlo method to device modeling are significantly more comprehensive [6]. While there is no fundamental roadblock to including these scattering mechanisms, efficient methods of determining the self-energies and computing the Green's functions are required to make quantum transport simulation practical. The second area that requires further work is a careful understanding of the coupling between the device and the leads. In most implementations of quantum transport, it is assumed that the role of leads can be added as a simple series resistance. This assumption usually fails unless the leads correspond to wide regions consisting of a large number of modes compared to the number of modes in the device. In a number of practical situations, an abrupt demarcation between the device and leads does not exist. In such cases, considerable care must be exerted to determine the physically optimum demarcation, which can sometimes lead to computationally intractable problems. The third area that requires further work is three-dimensional modeling. Compared to the sophisticated three-dimensional modeling possible in the drift-diffusion framework, the computational modeling of quantum transport is in its infancy. The algorithms required to handle experimentally relevant three-dimensional nanodevices currently do not exist. Most work on quantum transport that has yielded insights into experiments is based on solving the quantum transport equations in reduced dimensions using detailed device physics known to experts. The field is far away from the state where one can define a three-dimensional real space or tight-binding grid and model experiments.

The NEGF approach is based on the rigorous many-body quantum theory [16]. It is designed to describe carrier transport with scattering. In its derivation [4], it makes the assumptions similar to those of the quantum Boltzmann equation [or Keldysh–Kadanoff–Baym approach (KKB)] [8] and [9]; also see [16]. The main assumptions are: i) a single particle approach and ii) a mean-field approximation. Note that NEGF goes beyond KKB in treating interactions with reservoirs, e.g., electric leads, in a manner similar to Landauer approach; see [5]. When applying the NEGF approach to devices, additional assumptions are frequently made, such as the self-consistent Born approximation in the perturbation theory and the neglect of off-diagonal elements in the scattering matrices. In practice, the NEGF approach provides a good description for many devices. When scattering is strong and potential variations slow, it can be shown to reduce to the Boltzmann transport equation [15], which forms the basis for semiclassical modeling of devices. There are, however, important problems that the NEGF approach cannot describe. These have to do with strongly correlated transport in devices that display so-called single electron charging effects [28].

In the remainder of this section, we explain the NEGF approach in the phase-coherent limit by starting from Schrodinger's equation [5]. We will start by an explanation of the tight-binding Hamiltonian and relate this Hamiltonian to a device with open boundary conditions (Section V-A). The open boundary conditions lead to an infinite dimensional matrix. We will describe a procedure to fold the effect of the open boundaries into the finite device region in Section V-B. This will allow us to deal with small matrices where the open boundaries are modeled by *lead self-energies*. The Green's functions, self-energies, and their relationship to current and electron density are derived in Section V-C. Then in Section V-D, we extend the discussion in Section V-C to include electron–phonon interaction, which is where the NEGF approach is really essential.

A. Tight-Binding Hamiltonian for a One-Dimensional Device

Consider a system described by a set of one-dimensional (1-D) grid/lattice points with uniform spacing a. Further assume that only nearest neighbor grid points are coupled. A spatially *uniform system* with a constant potential has the Hamiltonian shown in (31) at the bottom of the page or

$$-t\Psi_{q-1} + (E-\epsilon)\Psi_q - t\Psi_{q+1} = 0 \tag{32}$$

where E is the energy and Ψ_q is the wave function at grid point q. The Hamiltonian matrix is tridiagonal because of nearest neighbor interaction. The diagonal and off-diagonal elements of the Hamiltonian ϵ and t represent the potential and interaction between nearest neighbor grid points q and $q+1$, respectively.

$$(E-H)\Psi = 0 \rightarrow \begin{pmatrix} \bullet & \bullet & \bullet & & & & & \\ & \bullet & \bullet & \bullet & & & & \\ & & -t & E-\epsilon & -t & & & \\ & & & & -t & E-\epsilon & -t & \\ & & & & & -t & E-\epsilon & -t \\ & & & & & & \bullet & \bullet & \bullet \end{pmatrix} \begin{pmatrix} \bullet \\ \bullet \\ \Psi_{q-1} \\ \Psi_q \\ \Psi_{q+1} \\ \bullet \\ \bullet \end{pmatrix} = 0 \tag{31}$$

The solution of (32) can be easily verified using Bloch theorem to be

$$E = \epsilon + 2t\cos(ka) \tag{33}$$

$$\Psi_q = e^{ikqa} \tag{34}$$

and the group velocity is

$$v = \frac{1}{\hbar}\frac{\partial E}{\partial k} = -\frac{2at}{\hbar}\sin(ka). \tag{35}$$

The uniform tight-binding Hamiltonian in (32) can be extended to a *general nearest neighbor tight-binding* Hamiltonian given by

$$-t_{q,q-1}\Psi_{q-1} + (E - \epsilon_q)\Psi_q - t_{q,q+1}\Psi_{q+1} = 0 \tag{36}$$

where ϵ_q is the on-site potential at grid point q and $t_{q,q+1}$ is the Hamiltonian element connecting grid points q and $t_{q+1,q} = t^{\dagger}_{q,q+1}1$. $t_{q+1,q} = t^{\dagger}_{q,q+1}$ because the Hamiltonian is Hermitian.

In the special case of the discretized Schrodinger equation on a uniform grid

$$t = t_{q,q+1} = t_{q+1,q} = -\frac{\hbar^2}{2ma^2} \text{ and } \epsilon_q = V_q + \frac{\hbar^2}{ma^2} \tag{37}$$

where a is the grid spacing and V_q is the electrostatic potential at grid point q.

B. Eliminating the Left and Right Semi-Infinite Leads

A typical nanodevice can be conceptually divided into three regions (Fig. 7):

- left semi-infinite lead (L) with a constant potential ϵ_l;
- device (D) with an arbitrary potential;
- right semi-infinite lead (R) with a constant potential ϵ_r.

The potential of the left (right) lead $\epsilon_l(\epsilon_r)$ and the hopping parameter $t_l(t_r)$ are assumed to be constant, which signifies that the leads are highly conducting and uniform. Then the Hamiltonian of the device and leads (36) is an infinite dimensional matrix that can be expanded as

$$\begin{aligned}
&\bullet\\
&\bullet\\
-t_l\Psi_{l3} + (E-\epsilon_l)\Psi_{l2} - t_l\Psi_{l1} &= 0\\
-t_l\Psi_{l2} + (E-\epsilon_l)\Psi_{l1} - t_{l,d}\Psi_1 &= 0 \qquad (38)\\
-t_{d,l}\Psi_{l1} + (E-\epsilon_1)\Psi_1 - t_{1,2}\Psi_2 &= 0 \qquad (39)\\
-t_{2,1}\Psi_1 + (E-\epsilon_2)\Psi_2 - t_{2,3}\Psi_3 &= 0 \qquad (40)\\
&\bullet\\
&\bullet\\
-t_{n,n-1}\Psi_{n-1} + (E-\epsilon_n)\Psi_n - t_{d,r}\Psi_{r1} &= 0 \qquad (41)\\
-t_{r,d}\Psi_n + (E-\epsilon_r)\Psi_{r1} - t_r\Psi_{r2} &= 0 \qquad (42)\\
-t_r\Psi_{r1} + (E-\epsilon_r)\Psi_{r2} - t_r\Psi_{r3} &= 0\\
&\bullet\\
&\bullet
\end{aligned}$$

where the top and bottom bullets represent the semi-infinite left and right leads. The subscript $lm(rm)$ refers to grid point m in the left (right) lead. However, to find the electron density in (26), the wave function is only required at the device grid points. We will now discuss a procedure to fold the influence of the left and right semi-infinite leads into the device region.

Terminating the Semi-Infinite Left and Right Leads: The wave function in the leads due to waves incident from the

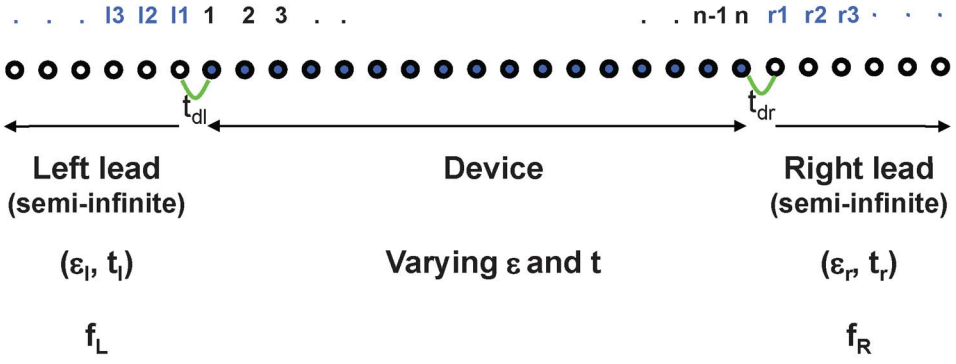


Fig. 7. *A one-dimensional device connected to two semi-infinite leads. While the potential in the leads (L and R) is held fixed, the potential in the device (D) can vary spatially.*

left lead is

$$\Psi_{ln} = (e^{+ik_l x_{ln}} + s_{ll} e^{-ik_l x_{ln}}) \text{ in region L} \quad (43)$$

$$\Psi_{rn} = s_{rl} e^{ik_r x_{rn}} \text{ in region R} \quad (44)$$

where x_{ln} and x_{rn} correspond to integer times grid spacing (a). The normalization constant has been neglected in the above equations. The corresponding eigenvalues are [(33)]

$$E - \epsilon_l = 2t_l \cos(k_l a) = t_l(e^{ik_l a} + e^{-ik_l a}) \quad (45)$$

and, similarly for the right lead, with the indexes r. s_{ll} and s_{rl} are the reflection and transmission amplitudes. Substituting (43) and (45) in (38) yields

$$s_{ll} = t_l^{-1}(-t_l + t_{ld}\Psi_1). \quad (46)$$

Substituting (43) and (46) in (39), we obtain

$$(E - \epsilon_1 - t_{d,l} e^{+ik_l a} t_l^{-1} t_{l,d})\Psi_1 - t_{1,2}\Psi_2 = -2it_{d,l}\sin(k_l a). \quad (47)$$

Equation (47) is a modification of Schrodinger's equation centered at grid point 1 of the device [(39)] to include the influence of the entire semi-infinite left lead. Similarly, substituting (44) and $E - \epsilon_r = 2t_r \cos(k_r a)$ into (42), we get

$$s_{rl} = \frac{e^{-ikna}}{t_r} t_{r,d}\Psi_n. \quad (48)$$

Now, substituting (44) and (48) into (41), we can terminate the right semi-infinite region to yield

$$-t_{n,n-1}\Psi_{n-1} + \left(E - \epsilon_n - t_{d,r}\frac{e^{ik_r a}}{t_r} t_{r,d}\right)\Psi_n = 0. \quad (49)$$

Equation (49) is a modification of Schrodinger's equation centered at grid point n of the device (41) to include the influence of the entire semi-infinite right lead.

The influence of the semi-infinite left and right leads have been folded into grid points 1 and n of the device for waves incident from the left lead [(47) and (49)]. Now the wave function in the device due to waves incident from the left lead can be obtained (to within a phase factor) by solving the following n-dimensional matrix instead of the infinite dimensional matrices in (38)–(42)

$$A\Psi_D^{(L)} = i_L \quad (50)$$

where A is a square matrix of dimension n and $\Psi_D^{(L)}$ and i_L are n by 1 vectors. i_L is the source function at (k, E) due to the left lead. Matrix A is

$$A = EI - H_D - \Sigma_{\text{lead}} \quad (51)$$

and the nonzero elements of Σ_{lead} are

$$\Sigma_{\text{lead}_{1,1}} = t_{d,l} e^{+ik_l a} t_l^{-1} t_{l,d} = \Sigma_L \quad (52)$$

$$\Sigma_{\text{lead}_{n,n}} = t_{d,r} e^{+ik_r a} t_r^{-1} t_{r,d} = \Sigma_R. \quad (53)$$

Σ_L and Σ_R are called the self-energies, and they represent the influence of the semi-infinite left and right leads on the device, respectively. The real part of self-energy shifts the on-site potential at grid point 1 from ϵ_1 to $\epsilon_1 + \text{Re}(\Sigma_L)$. The imaginary part of self-energy multiplied by -2 is the scattering rate of electrons from grid point 1 of the device to the left lead (scattering rate $= -2\text{Im}[\Sigma_L]$) in the weak coupling limit.

In a manner identical to the derivation of (50), for waves incident from the right lead, the wave function in the device $(\Psi_D^{(R)})$ can be obtained by solving

$$A\Psi_D^{(R)} = i_R \quad (54)$$

where i_R is the source function due to the right semi-infinite lead. The only nonzero elements of A, i_L and i_R are

$$A(1,1) = E - \epsilon_1 - \Sigma_L \text{ and } A(n,n) = E - \epsilon_n - \Sigma_R \quad (55)$$

$$A(i,i) = E - \epsilon_i,$$

$$A(i, i+1) = -t_{i,i+1} \text{ and } A(i+1, i) = -t_{i,i+1}^{\dagger} \quad (56)$$

$$i_L(1) = -2it_{d,l}\sin(k_l a) \quad (57)$$

$$i_R(n) = -2it_{d,r}\sin(k_r a). \quad (58)$$

C. Electron and Current Densities Expressed in Terms of Green's Functions

The Green's function corresponding to Schrodinger's equation $([E - H]\Psi = 0)$ for the device and leads is

$$[E - H + i\eta]G = I \quad (59)$$

where η is an infinitesimally small positive number that pushes the poles of G to the lower half-plane in complex energy, and H is the Hamiltonian. This Green's function is defined in a spirit similar to the Green's function for Poisson's equation. In the rest of this paper, η is implicit. The Green's function of device region D with the influence of the leads included is

$$AG = I \tag{60}$$

where

$$A = EI - H_D - \Sigma_{\text{lead}}(E) \tag{61}$$

is an n-dimensional matrix defined in (55) and (56).

Using the definition for G in (50) and (54), the wave function in region D due to waves incident from left and right leads can be written as

$$\Psi_D^{(L)} = G i_L \tag{62}$$

$$\Psi_D^{(R)} = G i_R. \tag{63}$$

As i_L and i_R are nonzero only at grid points 1 and n, the full G matrix is not necessary to find the wave function in the device; only the two columns $G(:,1)$ and $G(:,n)$ are necessary.

The *electron density* at grid point q can now be written using (62) and (63) in (26) as

$$n_q = \sum_{k,s} G_{q,1} i_L i_L^{\dagger} (G^{\dagger})_{1,q} f_L + G_{q,n} i_R i_R^{\dagger} (G^{\dagger})_{n,q} f_R \tag{64}$$

$$= \sum_{k,s} G_{q,1} \left[4 t_{d,l} \sin^2(k_l a) t_{l,d} f_L\right] (G^{\dagger})_{1,q} + G_{q,n} \left[4 t_{d,r} \sin^2(k_r a) t_{r,d} f_R\right] (G^{\dagger})_{n,q} \tag{65}$$

where $G^{\dagger}$ is the Hermitian conjugate of the Green's function. The summation over k can be converted to an integral over E by,

$$\sum_k \rightarrow \int \frac{dE}{2\pi} \left|\frac{dk}{dE}\right|. \tag{66}$$

Using (66) and $|dE/dk| = 2a|t||\sin(ka)|$ (where $k \in k_l, k_r$ and $t \in t_l, t_r$), (65) becomes

$$n_q = 2 \int \frac{dE}{2\pi} \left[G_{q,1}(E) \Sigma_L^{\text{in}}(E) (G^{\dagger})_{1,q}(E) + G_{q,n}(E) \Sigma_R^{\text{in}}(E) (G^{\dagger})_{n,q}(E) \right] \tag{67}$$

where

$$\Sigma_L^{\text{in}}(E) = 2 t_{d,l} \frac{1}{|t_l|} |\sin(k_l a)| t_{l,d} f_L(E) \tag{68}$$

$$\Sigma_R^{\text{in}}(E) = 2 t_{d,r} \frac{1}{|t_r|} |\sin(k_r a)| t_{r,d} f_R(E). \tag{69}$$

k_l and k_r at energy E are determined by $E = \epsilon_l + 2t_l \cos(k_l a)$ and $E = \epsilon_r + 2t_r \cos(k_r a)$ [(33)], respectively. It can be seen from (52), (53), (68), and (69)

$$\Sigma_L^{\text{in}}(E) = -2 Im[\Sigma_L(E)] f_L(E) \tag{70}$$

$$\Sigma_R^{\text{in}}(E) = -2 Im[\Sigma_R(E)] f_R(E). \tag{71}$$

The electron density [(67) or (72)] can then be written as

$$n_q = 2 \int \frac{dE}{2\pi} G(E) \Sigma_{\text{lead}}^{\text{in}}(E) G^{\dagger}(E)\big|_{q,q} \tag{72}$$

where the nonzero elements of $\Sigma_{\text{lead}}^{\text{in}}$ are

$$\Sigma_{\text{lead}1,1}^{\text{in}}(E) = \Sigma_L^{\text{in}}(E) \tag{73}$$

$$\Sigma_{\text{lead}n,n}^{\text{in}}(E) = \Sigma_R^{\text{in}}(E). \tag{74}$$

Σ_L^{in} and Σ_R^{in} defined above in (70) and (71) are called the *in-scattering self-energies* due to leads. These self-energies physically represent in-scattering of electrons from the semi-infinite leads to the device and so play an important role in determining the charge occupancy in the device. They depend on the Fermi–Dirac factor/occupancy in the leads f_L and f_R and the strength of coupling between leads and device $Im[\Sigma_L(E)]$ and $Im[\Sigma_R(E)]$.

It is easy to see that the electron density in (67) and (72) can also be written as

$$n_q = \int dE [\text{LDOS}_L(q,E) f_L(E) + \text{LDOS}_R(q,E) f_R(E)] \tag{75}$$

where $\text{LDOS}_L(q,E)$ ($\text{LDOS}_R(q,E)$) is the density of states due to waves incident from the left (right) lead at grid point q and

$$\text{LDOS}_L(q,E) = \frac{G_{q,1} \Gamma_L (G^{\dagger})_{1,q}}{\pi} \tag{76}$$

$$\text{LDOS}_R(q,E) = \frac{G_{q,n} \Gamma_R (G^{\dagger})_{q,n}}{\pi} \tag{77}$$

where

$$\Gamma_L(E) = -2Im[\Sigma_L(E)] \tag{78}$$
$$\Gamma_R(E) = -2Im[\Sigma_R(E)]. \tag{79}$$

Note that (75) is identical to (21) for semiclassical ballistic transport.

The *current density* between grid points q and $q+1$ per unit energy can be written as [(27)]

$$J_{q\to q+1}(E) = \frac{e\hbar}{2mai} 2\Big[\Big(\Psi_q^{(L)\dagger}\Psi_{q+1}^{(L)} - \Psi_{q+1}^{(L)\ \dagger}\Psi_q^{(L)}\Big)f_L(E) + \Big(\Psi_q^{(R)\dagger}\Psi_{q+1}^{(R)} - \Psi_{q+1}^{(R)\ \dagger}\Psi_q^{(R)}\Big)f_R(E)\Big]. \tag{80}$$

Now, following the derivation for electron density above [(67)], it is straightforward to derive that the current density

$$\begin{aligned} J_{q\to q+1} = \frac{ie\hbar}{2ma} 2\int \frac{dE}{2\pi} \times \Big[& G_{q,1}(E)\Sigma_L^{\text{in}}(E)(G^\dagger)_{1,q+1}(E) \\ & + G_{q,n}(E)\Sigma_R^{\text{in}}(E)(G^\dagger)_{n,q+1}(E) \\ & - G_{q+1,1}(E)\Sigma_L^{\text{in}}(E)(G^\dagger)_{1,q}(E) \\ & - G_{q+1,n}(E)\Sigma_R^{\text{in}}(E)(G^\dagger)_{n,q}(E)\Big]. \end{aligned} \tag{81}$$

The current density is given by

$$J_{q\to q+1} = \frac{ie\hbar}{2ma} 2\int \frac{dE}{2\pi}\Big[G(E)\Sigma_{\text{lead}}^{\text{in}}(E)G^\dagger(E)\big|_{q,q+1} - G(E)\Sigma_{\text{lead}}^{\text{in}}(E)G^\dagger(E)\big|_{q+1,q}\Big]. \tag{82}$$

1) Electron Correlation Function: More generally, we define the electron correlation function G^n, which is the solution to

$$AG^n = \Sigma_{\text{lead}}^{\text{in}}G^\dagger. \tag{83}$$

Noting that $G = A^{-1}$, it is easy to obtain (72) $G^n = G\Sigma_{\text{lead}}^{\text{in}}G^\dagger$.

The expressions for electron [(72)] and current [(82)] densities at energy E in the phase-coherent case at finite applied biases can now be written as

$$n_q(E) = 2\frac{G_{q,q}^n(E)}{2\pi} \tag{84}$$
$$J_{q\to q+1}(E) = \frac{ie\hbar}{2ma} 2\frac{1}{2\pi}\Big[G_{q,q+1}^n(E) - G_{q+1,q}^n(E)\Big]. \tag{85}$$

That is, the diagonal and first off-diagonal elements of G^n are related to the electron and current densities, respectively. Note that these equations are equivalent to (26) and (27) appearing in the Landauer–Buttiker approach. Equations (72) and (84) have to be divided by the grid spacing a for the density to have units of per unit length.

2) Hole Correlation Function: In the absence of phase-breaking scattering, the Green's function (G) and the electron correlation function (G^n) are sufficient for device modeling. Scattering introduces the need for the hole correlation function (G^p), whose role will become clearer in Section V-D. While the G^n Green's function is directly proportional to the density of occupied states, the hole correlation function is proportional to the density of unoccupied states.

The density of unoccupied states at grid point q is also obtained by applying the Landauer–Buttiker formalism. For this, we simply replace the probability of finding an occupied state in the lead $f_{L,R}$ by the probability of finding an unoccupied state in the lead $1-f_{L,R}$ in (26). Then following the derivation leading to (26), we obtain

$$h_q = 2\int \frac{dE}{2\pi}\Big[G_{q,1}(E)\Sigma_L^{\text{out}}(E)G_{1,q}^\dagger(E) + G_{q,n}(E)\Sigma_R^{\text{out}}(E)G_{n,q}^\dagger(E)\Big] \tag{86}$$
$$h_q = 2\int \frac{dE}{2\pi}G(E)\Sigma_{\text{lead}}^{\text{out}}G^\dagger(E)\big|_{q,q} \tag{87}$$

where the only nonzero elements of $\Sigma_{\text{lead}}^{\text{out}}$ are

$$\begin{aligned} \Sigma_{\text{lead}_{1,1}}^{\text{out}}(E) &= \Sigma_L^{\text{out}}(E) = 2t_{d,l}\frac{1}{|t_l|}|\sin(k_l a)|t_{l,d}(1-f_L) \\ &= -2Im[\Sigma_L(E)][1-f_L(E)] \end{aligned} \tag{88}$$
$$\begin{aligned} \Sigma_{\text{lead}_{n,n}}^{\text{out}}(E) &= \Sigma_R^{\text{out}}(E) = 2t_{d,r}\frac{1}{|t_r|}|\sin(k_r a)|t_{r,d}(1-f_R) \\ &= -2Im[\Sigma_R(E)][1-f_R(E)]. \end{aligned} \tag{89}$$

Akin to (83) and (84), the density of unoccupied states at energy E at grid point q can be expressed as the diagonal elements of G^p

$$h_q(E) = 2\frac{G_{q,q}^p(E)}{2\pi} \tag{90}$$

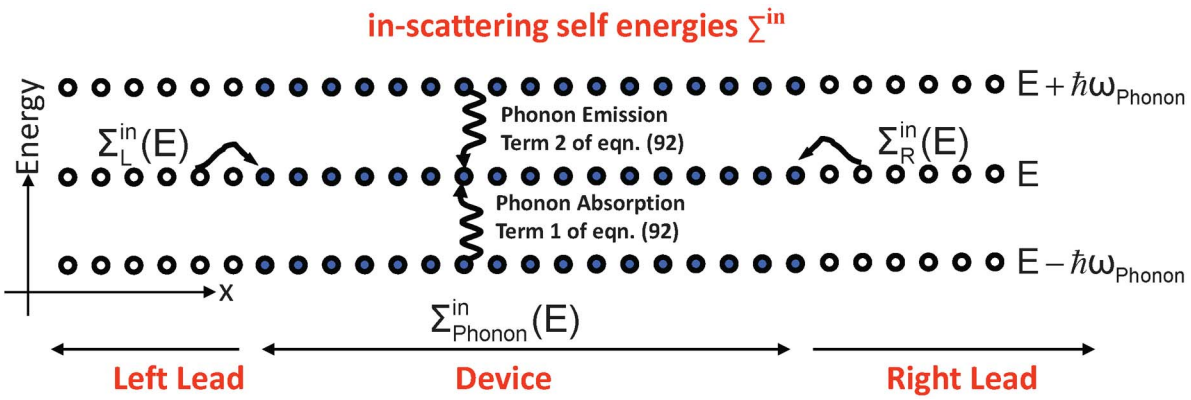


Fig. 8. *Pictorial representation of the two in-scattering self-energies that appear in this paper.* $\Sigma_L^{\text{in}}(E)$ *and* $\Sigma_R^{\text{in}}(E)$ *are the self-energies of the leads and are nonzero only at the first and last device grid points.* $\Sigma_{\text{Phonon}}^{\text{in}}(E)$ *is the self-energy due to electron–phonon interaction and is nonzero at all device grid points.*

where G^p is in general given by

$$AG^p = \Sigma_{\text{lead}}^{\text{out}} G^\dagger. \tag{91}$$

Equations (86) and (90) have to be divided by the grid spacing a for the density to have units of per unit length.

D. Electron-Phonon Scattering

In Section V-C, we defined the self-energies in the device arising from coupling of the device to the external leads. The self-energy Σ_L^{in} represents in-scattering of electrons (in-scattering rate) from the semi-infinite left lead to the device, assuming that grid point 1 of the device is empty. A similar statement applies to Σ_R^{in}. The in-scattering self-energy of the leads depends on their Fermi distribution functions and surface density of states.

A second source for in-scattering to grid point q and energy E is electron–phonon interaction. The self-energy at (q, E) has two terms corresponding to in-scattering from $(q, E + \hbar\omega_{\text{phonon}})$ and $(q, E - \hbar\omega_{\text{phonon}})$, as shown in Fig. 8. Intuitively, the in-scattering self-energy (in-scattering rate) at (x, E) should depend on the Bose factor for phonon occupancy, the deformation potential for electron–phonon scattering and the availability of electrons at energies $E + \hbar\omega_{\text{phonon}}$ and $E - \hbar\omega_{\text{phonon}}$. It follows rigorously, within the Born approximation, that the in-scattering self-energy at energy E and grid point q is [15]

$$\Sigma_{\text{Phonon}_{q,q}}^{\text{in}}(E) = \sum_\eta D_q^\eta \Big[n_B(\hbar\omega_{\text{phonon}}) G_{q,q}^n(E - \hbar\omega_{\text{phonon}}) + \big(n_B(\hbar\omega_{\text{phonon}}) + 1\big) G_{q,q}^n(E + \hbar\omega_{\text{phonon}}) \Big]. \tag{92}$$

D_q^η represents the electron–phonon scattering strength at grid point q due to a phonon mode η. The first term of (92) represents in-scattering to E from $E - \hbar\omega_{\text{phonon}}$ (phonon absorption). n_B is the Bose distribution function for phonons of energy $\hbar\omega_{\text{phonon}}$ and $G_q^n(E - \hbar\omega_{\text{phonon}})$ is the electron density at $E - \hbar\omega_{\text{phonon}}$. The first and second terms of (92) represent in-scattering of electrons from $E - \hbar\omega_{\text{phonon}}$ (phonon absorption) and $E + \hbar\omega_{\text{phonon}}$ (phonon emission) to E, respectively. The in-scattering rate at grid point q is given by

$$\text{In-scattering rate at grid point } q : \frac{\hbar}{\tau_q^{\text{in}}(E)} = \Sigma_{q,q}^{\text{in}}(E) \tag{93}$$

where $\Sigma_{q,q}^{\text{in}}$ is the sum of all in-scattering self energies at grid point q.

The out-scattering self-energy Σ_L^{out} in (88) represents out-scattering of electrons from grid point 1 in the device to the semi-infinite left lead, assuming that grid point 1 of the device was occupied. The out-scattering self-energy due to the left lead Σ_L^{out} depends on the probability of finding an unoccupied state in the left lead $1 - f_L$ and the surface density of states of the left lead. A similar statement applies to Σ_R^{out}. A second source for out-scattering of electrons from an occupied state at (q, E) is electron–phonon interaction, which leads to scattering to $(q, E + \hbar\omega_{\text{phonon}})$ and $(q, E - \hbar\omega_{\text{phonon}})$ as represented in Fig. 9. Intuitively, the out-scattering self-energy (out-scattering rate) at (q, E) should depend on the Bose factor for phonon occupancy, the deformation potential for electron–phonon scattering, and the availability of unoccupied states at energies $E + \hbar\omega_{\text{phonon}}$ and $E - \hbar\omega_{\text{phonon}}$. It follows rigorously, within the Born

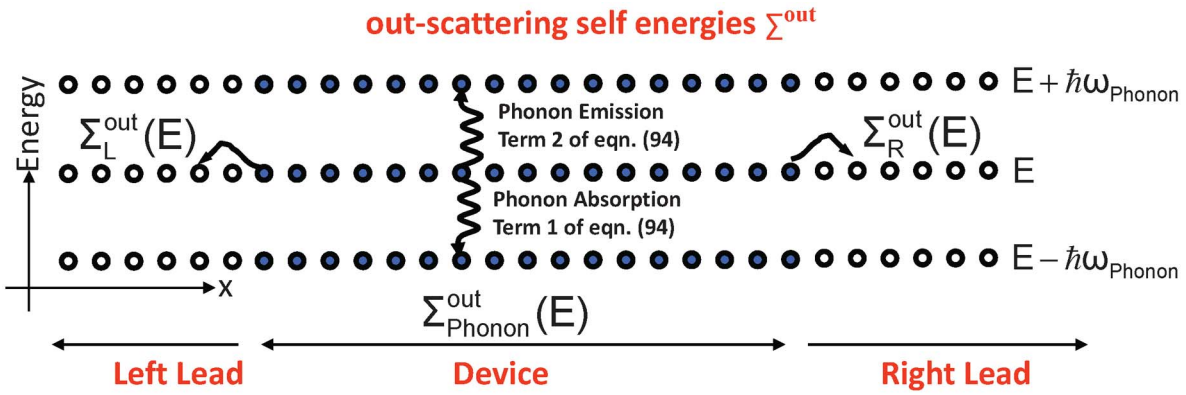


Fig. 9. ***Pictorial representation of the two out-scattering self-energies that appear in this paper.*** $\Sigma_{\text{lead}_q}^{\text{out}}(E)$ ***is self-energy due to leads, which is nonzero only at the first and last device grid points.*** $\Sigma_{\text{Phonon}_q}^{\text{out}}(E)$ ***is self-energy due to electron–phonon interaction, which is nonzero at all device grid points.***

approximation, that the out-scattering self energy at (q, E) is [15]

$$\Sigma^{\text{out}}_{\text{Phonon}q,q}(E) = \sum_{\eta} D_q^{\eta}\Big[\big(n_B(\hbar\omega_{\text{phonon}})+1\big)G^p_{q,q}(E-\hbar\omega_{\text{phonon}}) + n_B(\hbar\omega_{\text{phonon}})G^p_{q,q}(E+\hbar\omega_{\text{phonon}})\Big]. \quad (94)$$

In the above equation, $G_q^p(E-\hbar\omega_{\text{phonon}})$ and $G_q^p(E+\hbar\omega_{\text{phonon}})$ are the densities of unoccupied states at $E-\hbar\omega_{\text{phonon}}$ and $E+\hbar\omega_{\text{phonon}}$. So the first and second terms of (94) represent out-scattering of electrons from E to $E-\hbar\omega_{\text{phonon}}$ (phonon emission) and $E+\hbar\omega_{\text{phonon}}$ (phonon absorption), respectively. The out-scattering rate at grid point q is given by

$$\text{Out-scattering rate at grid point } q : \frac{\hbar}{\tau_q^{\text{out}}(E)} = \Sigma^{\text{out}}_{q,q}(E) \quad (95)$$

where Σ_q^{out} is the sum of all out-scattering self energies at grid point q.

We now discuss how the in-scattering self-energy due to electron–phonon scattering affects the expression for electron density. The electron density at grid point q in the phase-coherent case [(67) or (72)] is the sum of two terms

$$n_q = 2\int \frac{dE}{2\pi}\Big[G_{q,1}(E)\Sigma_L^{\text{in}}(E)(G^{\dagger})_{1,q}(E) + G_{q,n}(E)\Sigma_R^{\text{in}}(E)(G^{\dagger})_{n,q}(E)\Big]. \quad (96)$$

The first term represents in-scattering of electrons from the left lead $\Sigma_L^{\text{in}}(E)$, which is propagated to grid point q via the term $G_{q,1}(E)G_{1,q}^{\dagger}(E)$. The interpretation of the second term is similar except that it involves the right lead. In the presence of electron–phonon interaction, the in-scattering functions $\Sigma^{\text{in}}_{\text{Phonon}}$ is nonzero at all grid points. As a result, an electron can scatter from (q', E') to (q', E) and then propagate to grid point (q, E) via the term $G_{q,q'}(E)G^{\dagger}_{q',q}(E)$. The expression for the electron density can be generalized to include such terms

$$\begin{aligned} n_q &= 2\int \frac{dE}{2\pi}\Big[G_{q,1}(E)\Sigma_L^{\text{in}}(E)G_{1,q}^{\dagger}(E) + G_{q,n}(E)\Sigma_R^{\text{in}}(E)G_{n,q}^{\dagger}(E) + \sum_{q'} G_{q,q'}(E)\Sigma^{\text{in}}_{q',\text{Phonon}}(E)G^{\dagger}_{q',q}(E)\Big] \\ &= 2\int \frac{dE}{2\pi}\big[G(E)\Sigma^{\text{in}}_{\text{lead}}(E)G^{\dagger}(E) + G(E)\Sigma^{\text{in}}_{\text{Phonon}}(E)G^{\dagger}(E)\big]\big|_{q,q} \\ &= 2\int \frac{dE}{2\pi}G^n_{q,q}(E) \end{aligned} \quad (97)$$

where the third term corresponds to propagation of electrons from grid point q' to q after a scattering event at q', as shown in Fig. 10. The in-scattering self-energies due to phonon scattering are given by (92). More generally, G^n is given by

$$G^n(E) = G(E)\Sigma^{\text{in}}(E)G^{\dagger}(E) \quad (98)$$

$$[E-H-\Sigma(E)]G^n(E) = \Sigma^{\text{in}}(E)G^{\dagger}(E) \quad (99)$$

where Σ^{in} is the sum of self-energies due to leads and electron–phonon interaction. The reader can compare the above two equations to (73) and (83), which are valid in the phase-coherent limit.

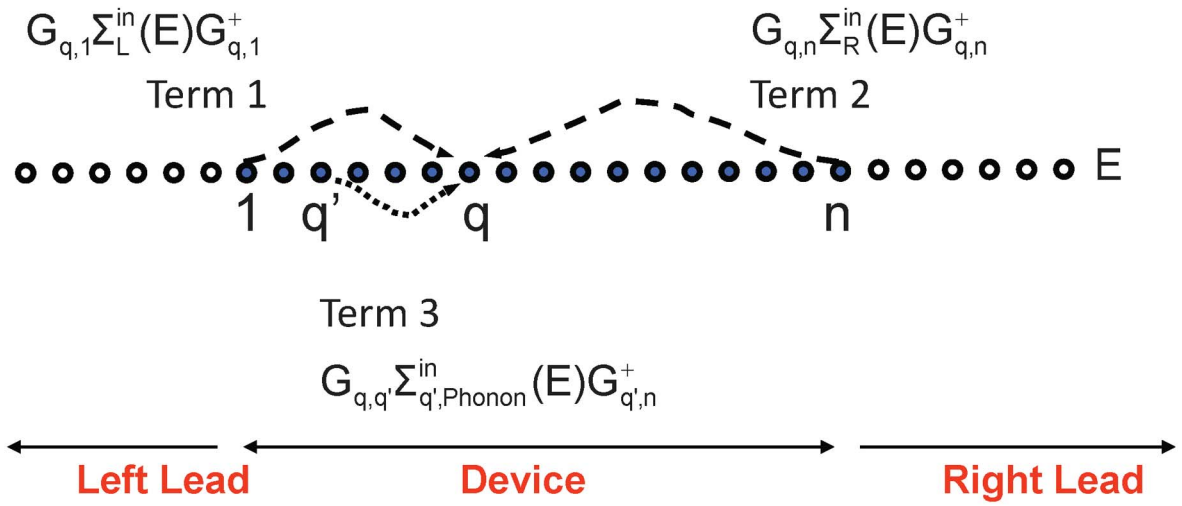


Fig. 10. *Contributions to electron density from leads and electron–phonon scattering. The three terms of (97) are shown.*

The density of unoccupied states can be written in a manner identical to (97) as

$$\begin{aligned} h_q &= 2\int \frac{dE}{2\pi}\Big[G_{q,1}(E)\Sigma_L^{\text{out}}(E)G_{1,q}^{\dagger}(E) \\ &\qquad + G_{q,n}(E)\Sigma_R^{\text{out}}(E)G_{n,q}^{\dagger}(E) \\ &\qquad + \sum_{q'} G_{q,q'}(E)\Sigma_{q',Phonon}^{\text{out}}(E)G_{q',q}^{\dagger}(E)\Big] \\ &= 2\int \frac{dE}{2\pi}\left[G(E)\Sigma_{\text{lead}}^{\text{out}}(E)G^{\dagger}(E)\right. \\ &\qquad \left. + G(E)\Sigma_{\text{Phonon}}^{\text{out}}(E)G^{\dagger}(E)\right]\Big|_{q,q} \\ &= 2\int \frac{dE}{2\pi}G^{p}_{q,q}(E). \end{aligned} \tag{100}$$

More generally, the G^p matrix is given by

$$G^p(E) = G(E)\Sigma^{\text{out}}(E)G^{\dagger}(E) \tag{101}$$

$$[E - H - \Sigma(E)]G^p(E) = \Sigma^{\text{out}}(E)G^{\dagger}(E) \tag{102}$$

where Σ^{out} is the sum of self-energies due to leads, electron–phonon interaction, and all other processes.

Note that, in general, G^n and G^p are full matrices, the diagonal elements of which correspond to density of occupied and unoccupied states, respectively, and the first off-diagonal elements of G^n and G^p correspond to the current density.

The Green's function G in the device region is obtained by solving

$$[E - H - \Sigma_{\text{lead}}(E) - \Sigma_{\text{Phonon}}(E)]G = I \tag{103}$$

which is similar to (60) for the phase-coherent case, except for the additional self-energy due to phonon scattering.

VI. NONEQUILIBRIUM GREEN'S FUNCTION EQUATIONS FOR LAYERED STRUCTURES

The previous section dealt with a simple one-dimensional Hamiltonian. In this section, we will present the NEGF equations for a family of more realistic structures called *layered structures*. A layer can be considered to be a generalization of a single grid point/orbital (Fig. 7) to a set of grid points/orbitals per layer. Note that we will use "grid points" to represent both orbitals and the conventional grid points that follow from discretization of a differential equation in a real space grid. For example, consider the structure that consists of two grid points per layer labeled by a and b, as shown in Fig. 11(a). For this structure, the form of the Hamiltonian remains the same as in (36) except that ϵ_q and $t_{q,q+1}$ become (2×2) matrices $\begin{pmatrix} \epsilon_q^a & \epsilon_q^{ab} \\ \epsilon_q^{ba} & \epsilon_q^b \end{pmatrix}$ and $\begin{pmatrix} t_{q,q+1}^{aa} & t_{q,q+1}^{ab} \\ t_{q,q+1}^{ba} & t_{q,q+1}^{bb} \end{pmatrix}$, which represents the Hamiltonian element of layer q and the coupling between layers q and $q+1$, respectively. ϵ_q^a and ϵ_q^b are the diagonal elements of the Hamiltonian at grid points a and b, respectively, in layer q. $t_{q,p}^{ij}$ is the Hamiltonian element representing interaction between grid point i in layer q and grid point j in layer p, where $i, j \in a, b$. Ψ_q becomes $\begin{pmatrix} \psi_q^a \\ \psi_q^b \end{pmatrix}$, which is a (2×1) vector representing the wavefunction in layer q. Its components ψ_q^a and ψ_q^b are the wave functions at grid points a and b in layer q, respectively. For the case in Fig. 11(a), the equivalent of Schrodinger's

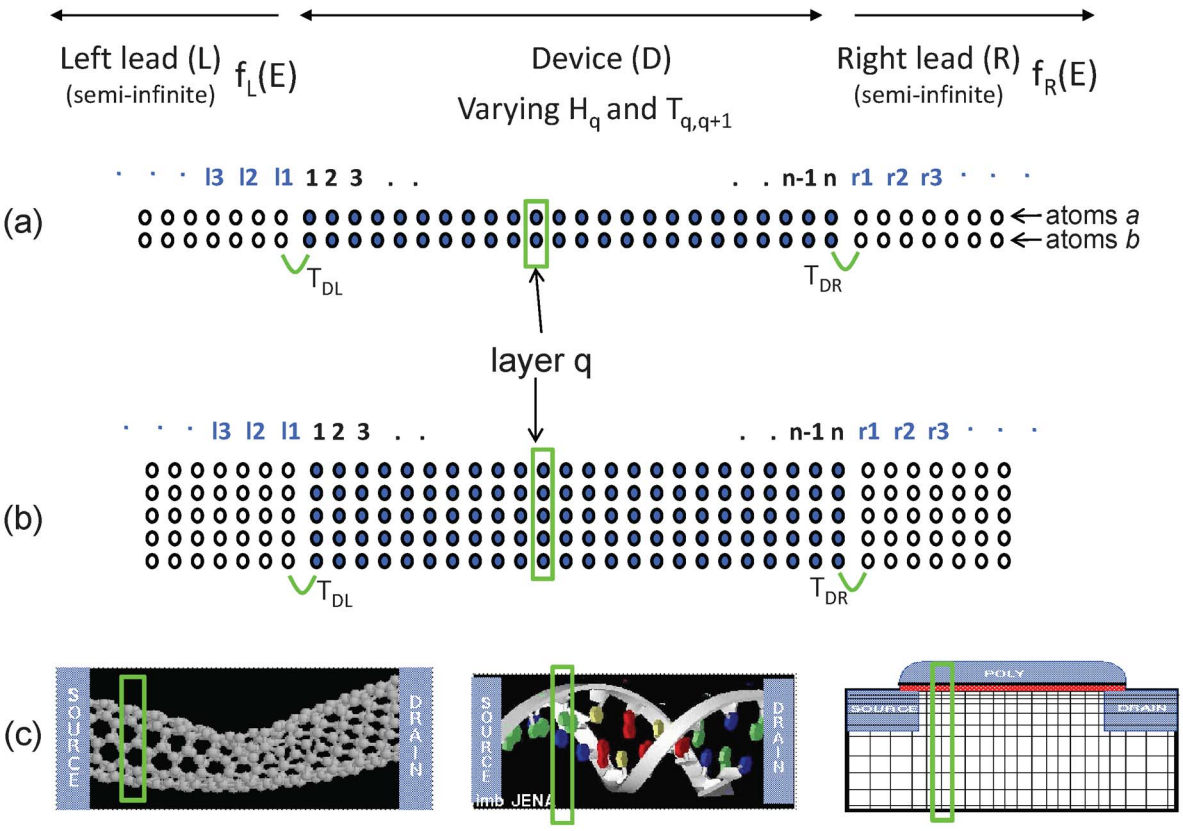


Fig. 11. *(a) and (b) Scheme of simulation domains in the layered structure: device, left and right leads. (c) Examples of the layered structures: carbon nanotube, DNA molecule, MOSFET.*

equation for the single grid points per layer in (36) becomes

$$-\begin{pmatrix} t^{aa}_{q,q-1} & t^{ab}_{q,q-1} \\ t^{ba}_{q,q-1} & t^{bb}_{q,q-1} \end{pmatrix}\begin{pmatrix} \psi^a_{q-1} \\ \psi^b_{q-1} \end{pmatrix} + \left(\begin{pmatrix} E & 0 \\ 0 & E \end{pmatrix} - \begin{pmatrix} \epsilon^a_q & \epsilon^{ab}_q \\ \epsilon^{ba}_q & \epsilon^b_q \end{pmatrix}\right) \times \begin{pmatrix} \psi^a_q \\ \psi^b_q \end{pmatrix} - \begin{pmatrix} t^{aa}_{q,q+1} & t^{ab}_{q,q+1} \\ t^{ba}_{q,q+1} & t^{bb}_{q,q+1} \end{pmatrix}\begin{pmatrix} \psi^a_{q+1} \\ \psi^b_{q+1} \end{pmatrix} = 0.$$

Now, if we consider a structure with five grid points per layer as shown in Fig. 11(b), the (2×2) matrices in the above equation become (5×5) matrices.

Three examples of layered structures are shown in Fig. 11(c). In each of these structures, a *layer* consists of the orbitals/grid points between the dashed lines in Fig. 11. A common approximation used to describe the Hamiltonian of such layered structures consists of including interaction only between nearest neighbor layers. That is, each layer q interacts only with itself and its nearest neighbor layers $q-1$ and $q+1$. Then, the single particle Schrodinger's equation of the layered system becomes

$$-T_{q,q-1}\Psi_{q-1} + (EI - H_q)\Psi_q - T_{q,q+1}\Psi_{q+1} = 0$$

where the size of H_q is equal to N_q and the size of $T_{q,q+1}$ is equal to $(N_q \times N_{q+1})$, where $N_q(N_{q+1})$ is the number of grid points in layer $q(q+1)$. I_q is an identity matrix of dimension N_q. The Hamiltonian of the layered structure is a block tridiagonal matrix, 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 $q+1$, can be written as shown in (104) at the bottom of the next page, where we have made use of the fact that $T_{i+1,i} = T^{\dagger}_{i,i+1}$ because the Hamiltonian is Hermitian. The other elements of the Hamiltonian matrix are zero.

In the previous section, we derived the equations for the Green's function in the device by including the influence of leads as self-energies rather rigorously within the single particle picture. The self-energies due to electron–phonon interaction, however, were introduced as an afterthought, with the lead self-energies uninfluenced by electron–phonon interaction [see (52) and (53)]. This is not correct, however, because the electron–phonon interaction is present in the D, L, and R regions. The Green's function equation for the entire device and leads in the presence of electron–phonon scattering is given by [16]

$$[EI - H - \Sigma_{\text{Phonon}}]G = I \qquad (105)$$

where Σ_{Phonon} is the self-energy due to electron–phonon scattering. Partitioning of the device into the left lead (L), device (D), and right lead (R) is a mathematical construct that is motivated by the physics of the device. We partition the layered structure into L, D, and R regions as shown in Fig. 11. The device corresponds to the region where we solve for the nonequilibrium electron density. The leads are the highly conducting regions connected to the nanodevice. While the device region, where we seek the nonequilibrium density, consists of only n layers, the matrix equation corresponding to (105) is infinite dimensional due to the semi-infinite leads. We will now show how the influence of the semi-infinite leads can be folded into the device region. In a manner akin to the previous section, the influence of the semi-infinite leads is to affect layers 1 and n of the device region. An important difference is that the derivation here includes electron–phonon scattering and does not assume a flat potential in the semi-infinite leads. We first define

$$A' = [EI - H - \Sigma_{\text{Phonon}}]. \tag{106}$$

Noting that the Hamiltonian of the device can be partitioned into the sub-Hamiltonians of the D, L, and R regions and coupling between them, and noting that the Hamiltonian terms coupling L and R are zero, (105) can be written as

$$A'G = I \rightarrow \begin{pmatrix} A'_{LL} & A'_{LD} & O \\ A'_{DL} & A'_{DD} & A'_{DR} \\ O & A'_{RD} & A'_{RR} \end{pmatrix} \begin{pmatrix} G_{LL} & G_{LD} & G_{LR} \\ G_{DL} & G_{DD} & G_{DR} \\ G_{RL} & G_{RD} & G_{RR} \end{pmatrix} = \begin{pmatrix} I & O & O \\ O & I & O \\ O & O & I \end{pmatrix} \tag{107}$$

where we have (108)–(111) as shown at the bottom of the next page, where $T_{LD} = T_{l1,1}$ and $T_{RD} = T_{r1,n}$ are the coupling between the left and right leads and device, respectively. Note that $A'_{DL} = {A'_{LD}}^{\dagger}$, $A'_{DR} = {A'_{RD}}^{\dagger}$, A'_{LD}, and A'_{DL} (A'_{RD} and A'_{DR}) are sparse matrices. Their only nonzero entry represents coupling between the left (right) lead and device. O represents zero matrices. From (107), we have

$$G_{LD} = -A'^{-1}_{LL} A'_{LD} G_{DD} \tag{112}$$
$$G_{RD} = -A'^{-1}_{RR} A'_{RD} G_{DD} \tag{113}$$
$$A'_{DL} G_{LD} + A'_{DD} G_{DD} + A'_{DR} G_{RD} = I. \tag{114}$$

Substituting (112) and (113) in (114), we obtain a matrix equation with dimension corresponding to total number of grid points/orbitals in the n layers of the device

$$\left[A'_{DD} - A'_{DL} A'^{-1}_{LL} A'_{LD} - A'_{DR} A'^{-1}_{RR} A'_{RD}\right] G_{DD} = I. \tag{115}$$

The second and third terms of (115) are self-energies due to coupling of the device region to left and right leads, respectively.

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

$$A'_{LL} g_L = I \text{ and } A'_{RR} g_R = I. \tag{116}$$

The *surface Green's function* of the left and right leads are the Green's function elements corresponding to the edge

$$H = \begin{pmatrix} \bullet & \bullet & \bullet & & & & & & & & & & & & \\ & \bullet & \bullet & \bullet & & & & & & & & & & & \\ & & \bullet & \bullet & \bullet & & & & & & & & & & \\ & & & T^{\dagger}_{l4,l3} & H_{l3} & T_{l3,l2} & & & & & & & & & \\ & & & & T^{\dagger}_{l3,l2} & H_{l2} & T_{l2,l1} & & & & & & & & \\ & & & & & T^{\dagger}_{l2,l1} & H_{l1} & T_{l1,1} & & & & & & & \\ & & & & & & T^{\dagger}_{l1,1} & H_1 & T_{12} & & & & & & \\ & & & & & & & T^{\dagger}_{12} & H_2 & T_{2,3} & & & & & \\ & & & & & & & & \bullet & \bullet & \bullet & & & & \\ & & & & & & & & & \bullet & \bullet & \bullet & & & \\ & & & & & & & & & & \bullet & \bullet & \bullet & & \\ & & & & & & & & & & T^{\dagger}_{n-2,n-1} & H_{n-1} & T_{n-1,n} & & \\ & & & & & & & & & & & T^{\dagger}_{n-1,n} & H_n & T_{n,r1} & \\ & & & & & & & & & & & & T^{\dagger}_{n,r1} & H_{r1} & T_{r1,r2} \\ & & & & & & & & & & & & & T^{\dagger}_{r1,r2} & H_{r2} & T_{r2,r3} \\ & & & & & & & & & & & & & & T^{\dagger}_{r2,r3} & H_{r3} & T_{r3,r4} \\ & & & & & & & & & & & & & & & \bullet & \bullet & \bullet \\ & & & & & & & & & & & & & & & & \bullet & \bullet & \bullet \\ & & & & & & & & & & & & & & & & & \bullet & \bullet & \bullet \end{pmatrix} \tag{104}$$

layers $l1$ and $r1$, respectively

$$g_{L_{l1,l1}} = \left(A'^{-1}_{LL}\right)_{1,1} \text{ and } g_{R_{r1,r1}} = \left(A'^{-1}_{RR}\right)_{1,1}. \tag{117}$$

Equation (115) can now be rewritten in a form very similar to (60)

$$[EI - H - \Sigma_{\text{Phonon}} - \Sigma_{\text{lead}}]G_{DD} = I \tag{118}$$

where

$$\Sigma_{\text{lead}1,1} = T_{DL}g_{L_{l1,l1}}T_{LD} = \Sigma_L \tag{119}$$

$$\Sigma_{\text{lead}n,n} = T_{DR}g_{R_{r1,r1}}T_{RD} = \Sigma_R. \tag{120}$$

All other elements of Σ_{lead} are zero. Σ_L and Σ_R are *self-energies* due to the left and right leads, respectively, and A and $T_{DR} = T^{\dagger}_{RD}$. Finally, defining matrix A to be (over the device layers 1 through n)

$$A = EI - H - \Sigma_{\text{Phonon}} - \Sigma_{\text{lead}} \tag{121}$$

(118) can be written as

$$AG = I. \tag{122}$$

Solving (122) gives us the Green's function G over the device layers 1 through n. The difference between and A and A' matrices above is that the former is defined only over the device layer and has the effect of the semi-infinite left (L) and right (R) leads included in it as self-energies.

The self-energies Σ_{lead} allowed us to derive (122) in the device from (105), which was valid in the device and leads. We stress that the elements of G in (122) are exactly equal to the elements of G in the device region obtained from (105). The main information needed to solve (118) is the surface Green's functions of g_L and g_R. We will discuss two methods to obtain this surface Green's functions for a constant potential in the left and right leads. When the potential does not vary, A'_{LL} and A'_{RR} are semi-infinite

$$A'_{LL} = \begin{pmatrix} \bullet & \bullet & \bullet & & & \\ & \bullet & \bullet & \bullet & & \\ & & -T^{\dagger}_{l4,l3} & A'_{l3} & -T_{l3,l2} & \\ & & & -T^{\dagger}_{l3,l2} & A'_{l2} & -T_{l2,l1} \\ & & & & -T^{\dagger}_{l2,l1} & A'_{l1} \end{pmatrix} \text{ corresponds to the left semi-infinite lead} \tag{108}$$

$$A'_{RR} = \begin{pmatrix} A'_{r1} & -T_{r1,r2} & & & \\ -T^{\dagger}_{r1,r2} & A'_{r2} & -T_{r2,r3} & & \\ & -T^{\dagger}_{r2,r3} & A'_{r3} & -T_{r3,r4} & \\ & & \bullet & \bullet & \bullet \\ & & & \bullet & \bullet \end{pmatrix} \text{ corresponds to the right semi-infinite lead} \tag{109}$$

$$A'_{DD} = \begin{pmatrix} A'_1 & -T_{12} & & & & & \\ -T^{\dagger}_{12} & A'_2 & -T_{2,3} & & & & \\ & \bullet & \bullet & \bullet & & & \\ & & \bullet & \bullet & \bullet & & \\ & & & \bullet & \bullet & \bullet & \\ & & & & -T^{\dagger}_{n-2,n-1} & A'_{n-1} & -T_{n-1,n} \\ & & & & & -T^{\dagger}_{n-1,n} & A'_n \end{pmatrix} \text{ corresponds to the device region} \tag{110}$$

$$A'_{LD} = \begin{pmatrix} O & O & \bullet & \bullet & O & O \\ O & O & \bullet & \bullet & O & O \\ O & O & \bullet & \bullet & O & O \\ O & O & \bullet & \bullet & O & O \\ -T_{LD} & O & O & \bullet & \bullet & O \end{pmatrix} \text{ and } A'_{RD} = \begin{pmatrix} O & O & \bullet & \bullet & O & -T_{RD} \\ O & O & \bullet & \bullet & O & O \\ O & O & \bullet & \bullet & O & O \\ O & O & \bullet & \bullet & O & O \\ O & O & O & \bullet & \bullet & O \end{pmatrix} \tag{111}$$

periodic matrices with all diagonal/off-diagonal blocks being equal

$$A'_{l1} = A'_{l2} = A'_{l3} = \ldots = A'_{l} \tag{123}$$
$$A'_{r1} = A'_{r2} = A'_{r3} = \ldots = A'_{r}$$
$$T_{l1,l2} = T_{l2,l3} = T_{l3,l4} = \ldots = T_{l} \tag{124}$$
$$T_{r1,r2} = T_{r2,r3} = T_{r3,r4} = \ldots = T_{r}.$$

$g_{L_{n,n}}$ is obtained by solving the matrix quadratic equation

$$\left[A'_l - T_l^{\dagger} g_{Ll1,l1} T_l\right] g_{Ll1,l1} = I. \tag{125}$$

This equation can be solved iteratively by

$$\left[A'_l - T_l^{\dagger} g_{Ll1,l1}^{\langle m-1\rangle} T_l\right] g_{Ll1,l1}^{\langle m\rangle} = I \tag{126}$$

where the superscript of g_L represents the iteration number. Note that the solution to (125) is analytic when the dimension of A_l is one. A second simpler solution to obtain $g_{L_{n,n}}$ involves transforming to an eigenmode basis using an unitary transformation (S), such that

$$S^{-1} A'_l S = A_{L-\mathrm{diag}} \quad \text{and} \quad S^{-1} T_l S = T_{L-\mathrm{diag}} \tag{127}$$

where both $A_{L-\mathrm{diag}}$ and $T_{L-\mathrm{diag}}$ are diagonal matrices. The surface Green's function in this new basis is simply a diagonal matrix, whose elements are obtained by solving the scalar quadratic version of (125). The Green's function in the original basis (in which A'_l is not diagonal) can be obtained using the inverse unitary transformation.

1) Electron (G^n) and Hole (G^p) Green's Function: The electron density at any location in D, L, or R is equal to (see the discussion of electron density in Section V-C)

$$n(\vec{r}, E) = 2\frac{G^n(\vec{r}, \vec{r}, E)}{2\pi}. \tag{128}$$

The governing equation for G^n is

$$[EI - H - \Sigma_{\mathrm{Phonon}}] G^n = \Sigma^{\mathrm{in}}_{\mathrm{Phonon}} G^{\dagger} \tag{129}$$

where $G^{\dagger}$ is the Hermitian conjugate Green's function and $\Sigma^{\mathrm{in}}_{\mathrm{Phonon}}$ is the in-scattering self-energy due to phonon scattering. Equation (129) gives the electron density in L, D, and R; its dimension is infinite due to the semi-infinite left and right leads. It can, however, be converted to a finite dimensional matrix with dimension equal to the number of grid points/orbitals corresponding to the n device layers by calculating the self-energy due to the L and R leads. In a manner identical to the derivation of (118) for the Green's function, it can be shown that the leads can be folded into layers 1 and n of the device to yield

$$[EI - H - \Sigma_{\mathrm{Phonon}} - \Sigma_{\mathrm{lead}}] G^n = \Sigma^{\mathrm{in}} G^{\dagger} \tag{130}$$

or

$$A G^n = \Sigma^{\mathrm{in}} G^{\dagger} \tag{131}$$

where A has been defined in (121). Equation (131) gives the electron density only in the device (D) regions. We stress that the elements of G^n in (131) are exactly equal to the elements of G^n in the device region obtained from (129).

The self-energy Σ^{in} defined over device layers 1 through n has contributions due to both electron–phonon interaction and leads

$$\Sigma^{\mathrm{in}}_{1,1} = \Sigma^{\mathrm{in}}_{\mathrm{Phonon}1,1} + \Sigma^{\mathrm{in}}_{L} \tag{132}$$
$$\Sigma^{\mathrm{in}}_{n,n} = \Sigma^{\mathrm{in}}_{\mathrm{Phonon}n,n} + \Sigma^{\mathrm{in}}_{R} \tag{133}$$
$$\Sigma^{\mathrm{in}}_{q,q} = \Sigma^{\mathrm{in}}_{\mathrm{Phonon}q,q}, \text{ where } q = 2, 3, 4, \ldots, n-1. \tag{134}$$

The *in-scattering self-energies* due to the leads Σ^{in}_L and Σ^{in}_R have forms very similar to (70) and (71) of Section V-C

$$\Sigma^{\mathrm{in}}_L(E) = -2\mathrm{Im}[\Sigma_L(E)] f_L(E) = \Gamma_L(E) f_L(E) \tag{135}$$
$$\Sigma^{\mathrm{in}}_R(E) = -2\mathrm{Im}[\Sigma_R(E)] f_R(E) = \Gamma_R(E) f_R(E) \tag{136}$$

where

$$\Gamma_L(E) = -2\mathrm{Im}[\Sigma_L(E)] \tag{137}$$
$$\Gamma_R(E) = -2\mathrm{Im}[\Sigma_R(E)]. \tag{138}$$

f_L and f_R are the distribution functions in the left and right leads, respectively (Fermi factors at equilibrium). The self-energies $\Sigma_L(E)$ and $\Sigma_R(E)$ have been defined in (119) and (120).

The diagonal elements of the hole correlation functions $G^p(E)$ represents the density of unoccupied states

$$h(\vec{r}, E) = 2\frac{G^p(\vec{r}, \vec{r}, E)}{2\pi}. \tag{139}$$

The governing equation for $G^p(E)$ is

$$[EI - H - \Sigma_{\text{Phonon}}]G^p = \Sigma^{\text{out}}_{\text{Phonon}} G^\dagger \tag{140}$$

or

$$A_{DD}G^p = \Sigma^{\text{out}} G^\dagger \tag{141}$$
$$\Sigma^{\text{out}}{}_{1,1} = \Sigma^{\text{out}}_{\text{Phonon}1,1} + \Sigma^{\text{out}}_L \tag{142}$$
$$\Sigma^{\text{out}}{}_{n,n} = \Sigma^{\text{out}}_{\text{Phonon}n,n} + \Sigma^{\text{out}}_R \tag{143}$$
$$\Sigma^{\text{out}}{}_{i,i} = \Sigma^{\text{out}}_{\text{Phonon}i,i}, \text{ where i} = 2, 3, 4, \ldots \text{n} - 1 \tag{144}$$

where $\Sigma^{\text{out}}_{\text{Phonon}}$ is the out-scattering self-energy due to phonon scattering. The *out-scattering self-energies* due to the leads Σ^{out}_L and Σ^{out}_R have forms very similar to (88) and (89)

$$\begin{aligned}\Sigma^{\text{out}}_L(E) &= -2\text{Im}[\Sigma_L(E)](1 - f_L(E)) \\ &= \Gamma_L(1 - f_L(E))\end{aligned} \tag{145}$$
$$\begin{aligned}\Sigma^{\text{out}}_R(E) &= -2\text{Im}[\Sigma_R(E)](1 - f_R(E)) \\ &= \Gamma_R(1 - f_R(E))\end{aligned} \tag{146}$$

where $1-f_L(E)$ and $1-f_R(E)$ are the probabilities of finding an unoccupied state in the left and right lead at energy E.

Equations (131) and (141) for G^n and G^p can be written as

$$G^n = A^{-1}\Sigma^{\text{in}}G^\dagger = G\Sigma^{\text{in}}G^\dagger \tag{147}$$
$$G^p = A^{-1}\Sigma^{\text{out}}G^\dagger = G\Sigma^{\text{out}}G^\dagger. \tag{148}$$

While these equations appear often in literature, we do not suggest using them to compute the diagonal elements of G^n and G^p of layered structures. This is because their use requires knowledge of the entire G matrix when Σ^{in} is nonzero at all grid points. Computation of the entire G amounts to inversion of A. Matrix inversion is computationally expensive. The diagonal elements of G^n and G^p of layered structures can be computed more efficiently [23] without calculating the entire G matrix using the algorithm developed in [23], a simplified version of which is discussed in Appendix II.

2) Current Density: We will now present some expression for current density commonly used in literature. In doing so, for brevity of notation we will drop the subscript DD used in the above sections, and simply remember that the Green's functions represented by G and G^n are over the device layers 1 through n. The current flowing between layers q and $q+1$ is [which has a similar form to (85)]

$$J_{q\to q+1} = \frac{ie}{\hbar}2\int\frac{dE}{2\pi}\text{Tr}\left[T_{q,q+1}G^n{}_{q+1,q}(E) - T_{q+1,q}G^n{}_{q,q+1}(E)\right] \tag{149}$$

where *Tr* stands for trace. Equation (149) frequently appears in the literature in other useful forms that are derived below. Expanding both terms of (149) using (B10) of Appendix I, we get

$$\begin{aligned}J_L = \frac{ie}{\hbar}2\int\frac{dE}{2\pi}\text{Tr}\Big(&\Big[T_{LD}G_{1,1}(E)T_{DL}g^n_{Ll1,l1}(E) \\ &+ T_{LD}G^n{}_{1,1}(E)T_{DL}g^\dagger_{Ll1,l1}(E)\Big] \\ &- \Big[T_{DL}g^n_{Ll1,l1}(E)T_{LD}G^\dagger{}_{1,1}(E) \\ &+ T_{DL}g^\dagger_{Ll1,l1}(E)T_{LD}G^n{}_{1,1}(E)\Big]\Big)\end{aligned} \tag{150}$$

$$\begin{aligned}= \frac{ie}{\hbar}2\int\frac{dE}{2\pi}\text{Tr}\Big(&\left[G_{1,1}(E) - G^\dagger{}_{1,1}(E)\right]T_{DL}g^n_{Ll1,l1}(E)T_{LD} \\ &- G^n{}_{1,1}(E)T_{DL} \\ &\times\left[g_{Ll1,l1}(E) - g^\dagger_{Ll1,l1}(E)\right]T_{LD}\Big).\end{aligned} \tag{151}$$

Using the relationships

$$\Sigma^{\text{in}}_L = T_{DL}g^n_{Ll1,l1}T_{LD} \tag{152}$$
$$-i\Gamma_L = T_{DL}\left[g_{Ll1,l1} - g^\dagger_{Ll1,l1}\right]T_{LD}. \tag{153}$$

Equation (151) can be written as

$$\begin{aligned}J_L = \frac{e}{\hbar}2\int\frac{dE}{2\pi}\text{Tr}\big(&i\left[G_{1,1}(E) - G^\dagger{}_{1,1}(E)\right] \\ &\times\Sigma^{\text{in}}_L(E) - G^n{}_{1,1}(E)\Gamma_L(E)\big).\end{aligned} \tag{154}$$

Equations (149) and (154) are both general expression for current density valid in the presence of electron–phonon scattering in the device [17]. The advantage of using (149) is that the current density can be calculated at every layer of the device. This expression is useful in understanding how the current density is energetically redistributed along the length of the device as a result of scattering.

In the phase-coherent limit, where $\Sigma_{\text{Phonon}} = 0$, we expect to get the Landauer–Buttiker formula, which is a special case of (149) and (154). We define matrices $\tilde{\Gamma}_L$ and $\tilde{\Gamma}_R$, which consist of n diagonal blocks corresponding to the n device layers (dimension of A matrix) and with the following nonzero elements:

$$\tilde{\Gamma}_L|_{1,1} = \Gamma_L, \quad \tilde{\Gamma}_R|_{n,n} = \Gamma_R. \tag{155}$$

Now left-multiplying (122) by $G^\dagger$ and right-multiplying the Hermitian conjugate of (122) by G, and subtracting the

resulting two equations, we have

$$G - G^{\dagger} = G^{\dagger}(\Sigma - \Sigma^{\dagger})G \tag{156}$$

where Σ is the total self-energy due to phonons and the leads. The Hermitian conjugate Green's functions and self-energies are $G^{\dagger}$ and $\Sigma^{\dagger}$. In the absence of phonon scattering, the self-energies only have components due to the leads and so (156) can be written as

$$i[G - G^{\dagger}] = G^{\dagger}(\tilde{\Gamma}_L + \tilde{\Gamma}_R)G \tag{157}$$

where (137) and (138) have been used. It also follows from (135), (136), and (147) that

$$G^n = G(\tilde{\Gamma}_L f_L + \tilde{\Gamma}_R f_R)G^{\dagger}. \tag{158}$$

Now using (157) and (158) in (154), the current in the phase-coherent limit is

$$J_L = \frac{e}{\hbar} 2 \int \frac{dE}{2\pi} T(E)[f_L(E) - f_R(E)]. \tag{159}$$

The total transmission at energy E is identified from (159) to be

$$T(E) = Tr\left[\tilde{\Gamma}_L(E)G(E)\tilde{\Gamma}_R(E)G^{\dagger}(E)\right]. \tag{160}$$

Note that to compute the total transmission using (160), only the elements of G connecting layers 1 and n are required because $\tilde{\Gamma}_L$ and $\tilde{\Gamma}_R$ are nonzero only in layers 1 and n, respectively.

A. Crib Sheet

The algorithmic flow in modeling nanodevices using the nonequilibrium Green's function consists of the following steps (Fig. 12). We first find a guess for the electrostatic potential $V(\vec{r})$ and calculate the self energies due to the leads [(175)–(185)]. The self-energies due to electron–phonon scattering are set to zero. The nonequilibrium Green's function equations for G, G^n, and G^p [(171)–(174)] are then solved. Following this, the self-energies due to electron–phonon scattering and leads [(175)–(185)] are calculated. As the equations governing the Green's functions depend on the self-energies, we iteratively solve for the Green's function and self-energies, as indicated by the inner loop of Fig. 12. Then, the electron density (diagonal elements of G^n) is used in Poisson's equation to obtain a new potential profile. We use this updated electrostatic potential profile as an input to

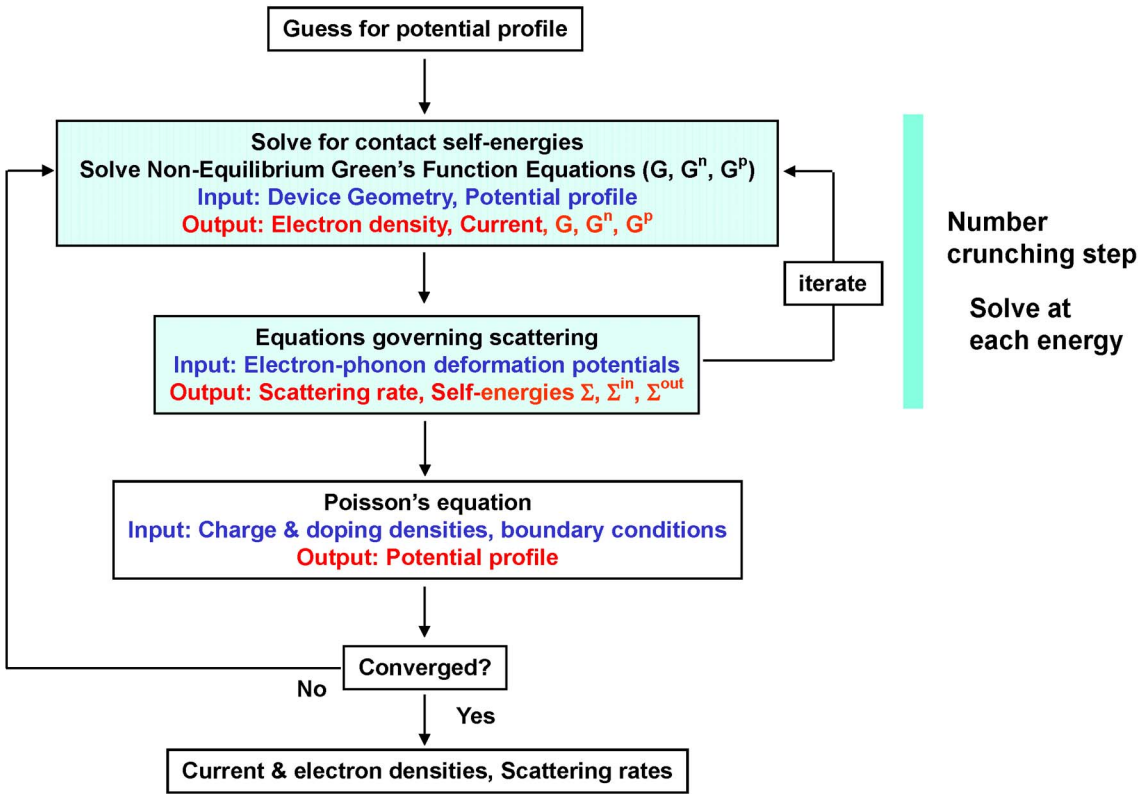


Fig. 12. *Flowchart of a typical simulation involved in modeling of a nanodevice. (Adapted from [1].)*

solve for updated nonequilibrium Green's functions and continue the above process iteratively until convergence is achieved (outer loop of Fig. 12). A number of equations that are repeatedly used in nanodevice modeling are listed below.

1) Physical Quantities:

Scattering Rate

$$\frac{\hbar}{\tau(E)} = \Gamma(E) = -2\,\mathrm{Im}[\Sigma(E)] \tag{161}$$

$$\frac{\hbar}{\tau^{\mathrm{in,out}}(E)} = \Sigma^{\mathrm{in,out}}(E) \tag{162}$$

Density of States (DOS) at $(\vec{r}, E)$. Use recursive algorithm to calculate DOS (Do not invert A).

$$N(\vec{r}, E) = -\frac{1}{\pi}\mathrm{Im}(G(\vec{r}, \vec{r}, E)) \tag{163}$$

Electron/Occupied density at $\vec{r}$. Use recursive algorithm to calculate n (Do not use $G^n = G\Sigma^{\mathrm{in}}G^\dagger$).

$$n(\vec{r}) = 2\int \frac{dE}{2\pi} G^n(\vec{r}, \vec{r}, E) \tag{164}$$

Hole/Unoccupied Density at $\vec{r}$. Use recursive algorithm to calculate h (Do not use $G^p = G\Sigma^{\mathrm{out}}G^\dagger$).

$$h(\vec{r}) = 2\int \frac{dE}{2\pi} G^p(\vec{r}, \vec{r}, E) \tag{165}$$

Current density flowing between layers q and $q+1$ (valid with scattering in device):

$$J_{q\to q+1} = \frac{ie}{\hbar} 2\int \frac{dE}{2\pi}\mathrm{Tr}\left[T_{q,q+1}G^n_{q+1,q}(E) - T_{q+1,q}G^n_{q,q+1}(E)\right] \tag{166}$$

Current density flowing from the left lead into layer 1 of device (valid with scattering in device)

$$J_L = \frac{e}{\hbar} 2\int \frac{dE}{2\pi}\mathrm{Tr}\Big\{i\Big[G_{1,1}(E) - G^\dagger_{1,1}(E)\Big]\Sigma^{\mathrm{in}}_L(E) - G^n_{1,1}(E)\Gamma_L(E)\Big\} \tag{167}$$

$$= \frac{e}{\hbar} 2\int \frac{dE}{2\pi}\mathrm{Tr}\Big\{G^p_{1,1}(E)\Sigma^{\mathrm{in}}_L(E) - G^n_{1,1}(E)\Sigma^{\mathrm{out}}_L(E)\Big\}. \tag{168}$$

Current density flowing from the left lead into layer 1 of device (valid only in the phase coherent limit) is

$$J_L = \frac{2e}{\hbar}\int \frac{dE}{2\pi} T(E)[f_L(E) - f_R(E)] \tag{169}$$

where the total transmission from the left to right lead at energy E is given by

$$T(E) = \mathrm{Tr}\left[\tilde{\Gamma}_L(E)G(E)\tilde{\Gamma}_R(E)G^\dagger(E)\right]. \tag{170}$$

Only elements of G connecting layers 1 and n are necessary.

2) Equations Solved:

Green's Function:

$$[EI - H - \Sigma]G(E) = I \to AG = I \tag{171}$$

Hermitian conjugate Green's Function:

$$G^\dagger(E)[EI - H - \Sigma^\dagger] = I \tag{172}$$

electron correlation function:

$$[EI - H - \Sigma]G^n(E) = \Sigma^{\mathrm{in}}(E)G^\dagger(E) \to AG^n = \Sigma^{\mathrm{in}}G^\dagger \tag{173}$$

hole correlation function:

$$[EI - H - \Sigma]G^p(E) = \Sigma^{\mathrm{out}}(E)G^\dagger(E) \to AG^p = \Sigma^{\mathrm{out}}G^\dagger \tag{174}$$

$$\Sigma^\alpha(E) = \Sigma^\alpha_{\mathrm{lead}}(E) + \Sigma^\alpha_{\mathrm{Phonon}}(E), \quad \text{where } \alpha \in \text{in,out} \tag{175}$$

$$\Sigma^\alpha_{\mathrm{lead}1,1}(E) = \Sigma^\alpha_L(E) \tag{176}$$

$$\Sigma^\alpha_{\mathrm{lead}n,n}(E) = \Sigma^\alpha_R(E) \tag{177}$$

$$\Sigma^\alpha_{\mathrm{lead}i,i} = 0 \quad \forall i \neq 1, n \tag{178}$$

and

$$\Gamma(E) = -2\mathrm{Im}[\Sigma(E)] \tag{179}$$

$$\Gamma_L(E) = -2\mathrm{Im}[\Sigma_L(E)] \tag{180}$$

$$\Gamma_R(E) = -2\mathrm{Im}[\Sigma_R(E)] \tag{181}$$

$$\Sigma^{\mathrm{in}}_L(E) = \Gamma_L(E)f_L(E) \tag{182}$$

$$\Sigma^{\mathrm{in}}_R(E) = \Gamma_R(E)f_R(E) \tag{183}$$

$$\Sigma^{\mathrm{out}}_L(E) = \Gamma_L[1 - f_L(E)] \tag{184}$$

$$\Sigma^{\mathrm{out}}_R(E) = \Gamma_R[1 - f_R(E)]. \tag{185}$$

The diagonal and nearest neighbor off-diagonal elements of G and G^n are computed repeatedly as they correspond to physical quantities such as the density of states, electron density, and current. Nonlocal scattering mechanisms, which require calculation of further off-diagonal elements, are not discussed here.

3) Useful Relationships:

$$i[G-G^\dagger] = G^p + G^n \propto \text{DOS} \quad (186)$$

$$i[\Sigma-\Sigma^\dagger] = \Sigma^{\text{out}} + \Sigma^{\text{in}} = \Gamma \quad (187)$$

$$G-G^\dagger = G[\Sigma^\dagger - \Sigma]G^\dagger \quad (188)$$

$$= G^\dagger[\Sigma^\dagger - \Sigma]G \quad (189)$$

$$= -iG\Gamma G^\dagger = -iG^\dagger\Gamma G \quad (190)$$

$$G^{n\dagger} = G^n \quad (191)$$

$$G^{p\dagger} = G^p. \quad (192)$$

VII. APPLICATION TO A BALLISTIC NANOTRANSISTOR

Quantum mechanics is playing an increasingly important role in modeling transistors with channel lengths in the 10 nm regime for several reasons.

i) *Tunneling* from gate to channel and source to drain determine the off current.

ii) *Ballistic* flow of electrons in the channel is important as the channel length becomes comparable to the electron mean free path.

iii) Classically, the electron distribution in the inversion layer is a sheet charge at the Si-SiO_2 interface. Quantum mechanically, the inversion layer charge is distributed over a few nanometers perpendicular to the Si-SiO_2 interface due to *quantum confinement*. Methods based on the drift-diffusion and Boltzmann equations do not a priori capture the quantum mechanical features mentioned above.

In this section, we will first discuss the NEGF equations involved in the two-dimensional modeling of nanotransistors within the effective mass framework [19], [23] and then compare the quantum mechanical and semiclassical results to point out the importance of the quantum (NEGF) formulation. The related equations and their discretized matrix forms that we solve are discussed in Section VII-A, while the application of the quantum mechanical method to illustrate i) the role of the polysilicon gate depletion, ii) the slopes of the drain current versus gate voltage (I_d–V_g), and iii) the transmission function $T(E)$ are discussed in Section VII-B.

A. Related Equations and Discretization

The schematic of the cross-section of the simulated nanoscale MOSFET is shown in Fig. 13. We consider N_b independent valleys for electrons within the effective mass approximation. The Hamiltonian of valley b is

$$H_b(\vec{r}) = -\frac{\hbar^2}{2}\left[\frac{d}{dx}\left(\frac{1}{m_x^b}\frac{d}{dx}\right) + \frac{d}{dy}\left(\frac{1}{m_y^b}\frac{d}{dy}\right) + \frac{d}{dz}\left(\frac{1}{m_z^b}\frac{d}{dz}\right)\right] + eV(\vec{r}) \quad (193)$$

where (m_x^b, m_y^b, m_z^b) are the (x,y,z) components of the electron effective mass in valley b and $V(r)$ is the potential energy. The equations for the Green's function (G) and electron and hole correlation functions (G^n and G^p) are

$$[E - H_{b_1}(\vec{r}_1)]G_{b_1,b_2}(\vec{r}_1,\vec{r}_2,E) - \int d\vec{r}\,\Sigma_{b_1,b'}(\vec{r}_1,\vec{r},E)G_{b',b_2}(\vec{r},\vec{r}_2,E) = \delta_{b_1,b_2}\delta(\vec{r}_1-\vec{r}_2) \quad (194)$$

and

$$[E - H_{b_1}(\vec{r}_1)]G^{n(p)}_{b_1,b_2}(\vec{r}_1,\vec{r}_2,E) - \int d\vec{r}\,\Sigma_{b_1,b'}(\vec{r}_1,\vec{r},E)G^{n(p)}_{b',b_2}(\vec{r},\vec{r}_2,E) = \int d\vec{r}\,\Sigma^{in(out)}_{b_1,b'}(\vec{r}_1,\vec{r},E)G^\dagger_{b',b_2}(\vec{r},\vec{r}_2,E). \quad (195)$$

The coordinate in (194) and (195) spans only the device. The influence of the semi-infinite source (S), drain (D), and polysilicon gate (G) leads and the electron–phonon interaction are included via self-energy terms $\Sigma_{b_1,b'}$ and $\Sigma^{in(out)}_{b_1,b'}$, as discussed in Section VI. The lead self-energies are diagonal in the band index $\Sigma^\alpha_{b_1,b_2,C} = \Sigma^\alpha_{b_1,C}\delta_{b_1,b_2}$ (C represents contacts = leads).

The electrostatic potential $V(r)$ varies in the (x,y) plane of Fig. 13, and the system is translationally invariant along the z-axis. So, any quantity $Q(\vec{r}_1,\vec{r}_2,E)$ depends only on the difference coordinate $z_1 - z_2$. Using the relation

$$Q(\vec{r}_1,\vec{r}_2,E) = \int \frac{dk_z}{2\pi} e^{ik_z(z_1-z_2)} Q(x_1,y_1,x_2,y_2,k_z,E) \quad (196)$$

the equations for G and $G^{n(p)}$ simplify to

$$\left[E - \frac{\hbar^2 k_z^2}{2m_z} - H_b(\vec{r}_1)\right]G_b(\vec{r}_1,\vec{r}_2,k_z,E) - \int d\vec{r}\,\Sigma_b(\vec{r}_1,\vec{r},k_z,E)G_b(\vec{r},\vec{r}_2,k_z,E) = \delta(\vec{r}_1-\vec{r}_2) \quad (197)$$

and

$$\left[E - \frac{\hbar^2 k_z^2}{2m_z} - H_b(\vec{r}_1)\right]G^{n(p)}_b(\vec{r}_1,\vec{r}_2,k_z,E) - \int d\vec{r}\,\Sigma_b(\vec{r}_1,\vec{r},k_z,E)G^{n(p)}_b(\vec{r},\vec{r}_2,k_z,E) = \int d\vec{r}\,\Sigma^{in(out)}_b(\vec{r}_1,\vec{r},k_z,E)G^\dagger_b(\vec{r},\vec{r}_2,k_z,E) \quad (198)$$

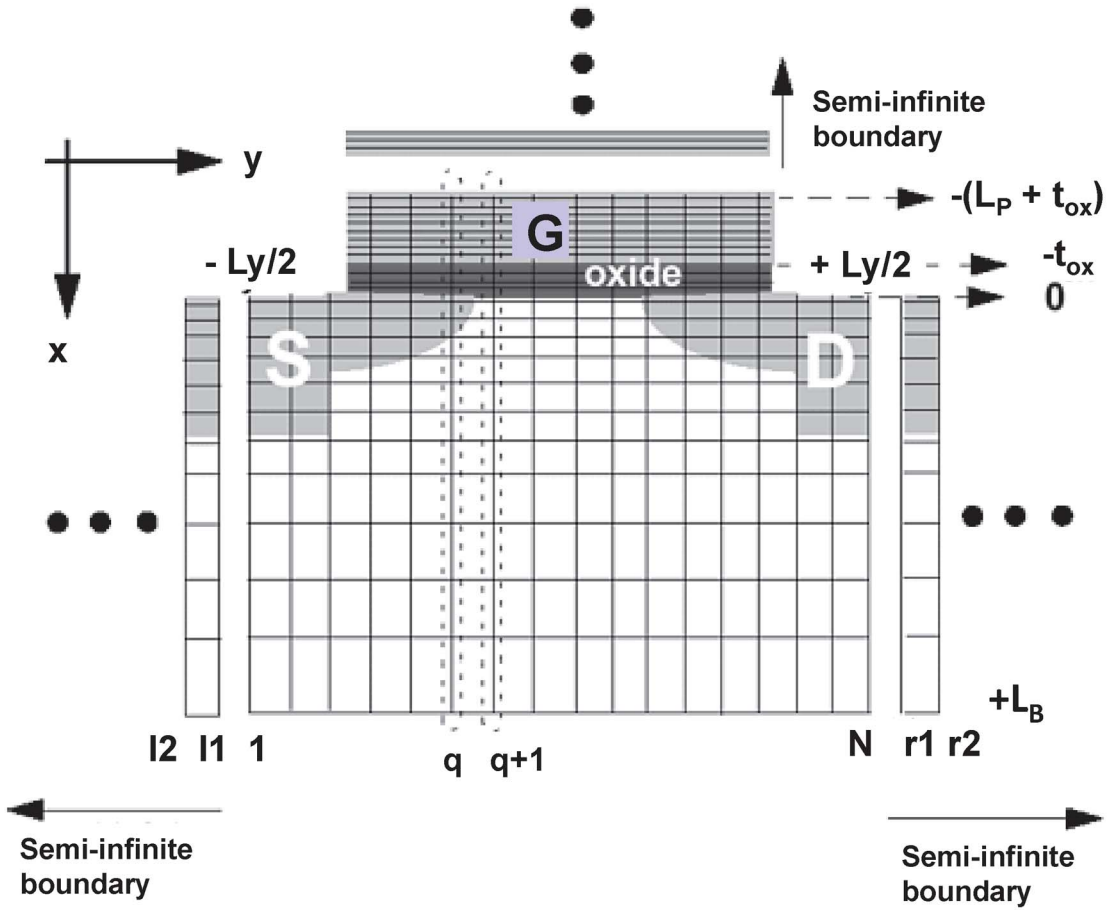


Fig. 13. *Schematic of the cross-section of the nanoscale MOSFET simulated. The equations are solved in a two-dimensional nonuniform spatial grid, with semi-infinite boundaries as shown. Each column q corresponds to the diagonal blocks of the Green's function equations. (Adapted from [23].)*

where $G^{\alpha}_{b} = G^{\alpha}_{b,b'}\delta_{b,b'}$ and $\Sigma^{\alpha}_{b} = \Sigma^{\alpha}_{b,b'}\delta_{b,b'}$, and $\vec{r} \rightarrow (x, y)$ for the remainder of this section.

Equations (197) and (198) can be written in matrix form as

$$AG = \lambda \tag{199}$$

$$AG^{n(p)} = \Sigma^{\text{in(out)}}G^{\dagger}. \tag{200}$$

The self-energies due to S, D, and G are nonzero only along the lines $y = -L_y/2$, $y = +L_y/2$, and $x = -(L_P + t_{\text{ox}})$ of Fig. 13. Definitions of symbols for length variables are given in Table 1. The A matrix is ordered such that all grid points at a y-coordinate (layer) correspond to a diagonal block of dimension N_x and there are N such blocks. In the notation adopted, $A_{q_1,q_2}(i, i')$ is the entry corresponding to grid points (x_i, y_{q_1}) and $(x_{i'}, y_{q_2})$. The index q refers to the layers in Section VI and corresponds to the y-direction in Fig. 13. The nonzero elements of the diagonal blocks of the A matrix are

Table 1 List of Abbreviations: Length Scales

t_{ox}	oxide thickness
L_P	polysilicon gate thickness in x-direction
L_B	boundary of substrate region in x-direction
L_y	Poisson's and NEGF equations are solved from $-L_y/2$ to $+L_y/2$
L_g	length of polysilicon gate region in y-direction

$$A_{q,q}(i,i) = E' - V_{i,q} - T_{q,q}(i+1,i) - T_{q,q}(i-1,i) - T_{q+1,q}(i,i) - T_{q-1,q}(i,i) - \Sigma_S(x_i,x_i)\delta_{q,1} - \Sigma_D(x_i,x_i)\delta_{q,N} - \Sigma_G(y_q,y_q)\delta_{i,1} \tag{201}$$

$$A_{q,q}(i\pm1,i) = T_{q,q}(i\pm1,i) - \Sigma_S(x_{i\pm1},x_i)\delta_{q,1} - \Sigma_D(x_{i\pm1},x_i)\delta_{q,N} \tag{202}$$

$$A_{q,q}(i,i') = -\Sigma_S(x_i,x_{i'})\delta_{q,1} - \Sigma_D(x_i,x_{i'})\delta_{q,N}, \quad \text{for } i' \neq i, i\pm1 \tag{203}$$

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

$$A_{q\pm1,q}(i,i) = T_{q\pm1,q}(i,i) - \Sigma_G(y_q, y_{q\pm1})\delta_{i,1}$$
$$A_{q,q'}(i,i') = 0, \text{ for } q' \neq q, q\pm1 \tag{204}$$

and the nonzero elements of the T matrix are

$$T_{q,q}(i\pm1,i) = \frac{\hbar^2}{2m^{\pm x}} \frac{2}{x_{i+1} - x_{i-1}} \frac{1}{|x_{i\pm1} - x_i|} \tag{205}$$

$$T_{q\pm1,q}(i,i) = \frac{\hbar^2}{2m^{\pm y}} \frac{2}{y_{q+1} - y_{q-1}} \frac{1}{|y_{q\pm1} - y_q|} \tag{206}$$

where $m^{\pm x} = (m_{i\pm1,q} + m_{i,q})/2$ and $m^{\pm y} = (m_{i,q\pm1} + m_{i,q})/2$. Nonzero elements of $\Sigma_G(y_q, y_{q'})$, where $q' \neq q$, $q \pm 1$ are neglected to ensure that A is block tridiagonal. The algorithm to calculate G and G^n (Appendix II) relies on the block tridiagonal form of A. The λ appearing in (199) corresponds to the delta function in (197). λ is a diagonal matrix whose elements are given by

$$\lambda_{q,q}(i,i) = \frac{4}{(x_{i+1} - x_{i-1})(y_{q+1} - y_{q-1})}. \tag{207}$$

B. Results

We now discuss the aspects of the quantum mechanical transport in a two-dimensional ballistic nanotransistor.

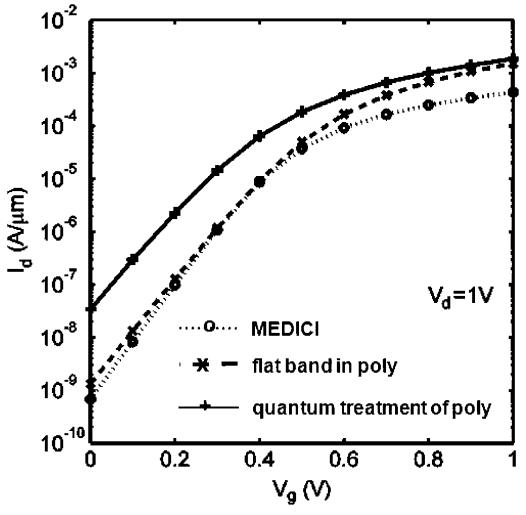


Fig. 14. *Plot of drain current versus gate voltage from the quantum mechanical and drift-diffusion (MEDICI) calculations at V_d = 1 V. At small gate voltages, the drain current from MEDICI is comparable to "flat band in poly." The drain current from "quantum treatment of poly" is, however, significantly different at all gate voltages. (From [23].)*

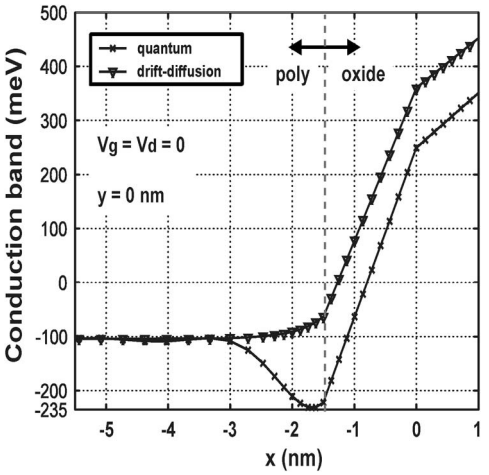


Fig. 15. *Potential profile at the $y = 0$ slice of MIT25, calculated using quantum and drift-diffusion methods by assuming flat band in the polysilicon gate. (From [23].)*

The "MIT well-tempered 25 nm" device,[1] which is referred to here as MIT25, is considered for the purpose of discussions. The $n+$ doping is 2×10^{20} cm^{-3} in both the source and drain regions while the $n+$ doping is 5×10^{20} cm^{-3} in the polysilicon region. MIT25 has a gate width of 50 nm and oxide thickness of 1.5 nm. The effective channel length, defined here as the distance between the source and drain positions where the doping falls down to 2×10^{19} cm^{-3}, is around 25 nm. In all calculations, we assume an isotropic effective mass for electrons. "Quantum treatment of poly" refers to computing the electrostatic potential in the polysilicon region by setting the Poisson boundary condition for the electrostatic potential deep inside the polysilicon region and computing the electron density in the polysilicon region quantum mechanically, while "flat band in poly" refers to neglecting the potential drop in the polysilicon region by setting the boundary condition for the electrostatic profile at the oxide-poly interface (at $y = -tox$).

We first compare the current (I_d) versus voltage (V_g) characteristics from our quantum and drift-diffusion (using MEDICI) simulations, as shown in Fig. 14. In the quantum case, there are higher off-current, higher threshold voltage shift, smaller subthreshold slope, and much higher on-current. The change in the threshold voltage results directly from the very different boundary for potential and the quantum treatment of electrons in the polysilicon gate.

We now compare the conduction band profiles for the two cases in Fig. 15. At polysilicon near the oxide-poly interface, the quantum band bending is opposite to that for the drift-diffusion case. For the quantum case, the conduction

[1]http://www-mtl.mit.edu/researchgroups/Well/.

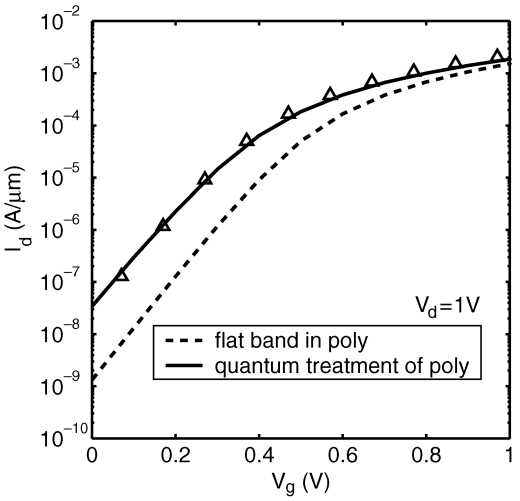


Fig. 16. *Drain current versus gate voltage for $V_d = 1$ V. Quantum mechanical treatment of the polysilicon gate results in much higher current (solid line). The triangles correspond to the $I_d(V_g)$ calculated using a flat band in the polysilicon region shifted by the equilibrium built-in potential of 130 meV in the polysilicon region. (From [23].)*

band is lower by approximately 130 meV. The physical reason for the difference is that, although classically electron density close to the interface is very high, quantum mechanically the electron density there is very tiny. This is because the electron wave function is tiny at the oxide interface because of the large barrier due to the oxide. As the density near the interface is much smaller than the uniform background doping density, the conduction band bends in a direction opposite to that computed classically.

In Fig. 16, the $I_d - V_g$ characteristics is plotted both for "quantum treatment of poly" and "flat band in poly" cases. Also shown is the $I_d - V_g$ characteristic in the "flat band in poly" case shifted by the 1-D equilibrium built-in potential [23]. The gate voltage shift is approximately equal to the band bending in the polysilicon gate. The 1-D built-in potential shifted "flat band in poly" band is close to the "quantum treatment of poly" band, but the difference increases at higher gate voltages.

Fig. 17 plots the heights of the first quantum resonant level (E_{r1}) and the classical source injection barrier (E_b) versus the gate voltage (a). Also plotted in the same figure is the narrowing of the triangular well in the channel with increase in the gate voltage (b). We see that, with increase in V_g, E_{r1} decreases more slowly compared to both E_b. The slower variation of E_{r1} arises due to quantum confinement in the triangular well in the channel that becomes progressively narrower with increase in gate voltage as shown in the right part of Fig. 16. This change in confinement is not an issue in the classical case. Because of the slow variation of E_{r1} with the gate voltage the subthreshold slope $d[log(I_d)]/dV_g$ is smaller in the quantum case compared to the classical case or drift-diffusion case (Fig. 14). We note here that the subthreshold current resulting from the simple intuitive expression

$$I = I_{q0} e^{\frac{-E_{r1}}{kT}} \tag{208}$$

matches the quantum result quite accurately. Here I_{q0} is a prefactor chosen to match the current at $V_g = 0$.

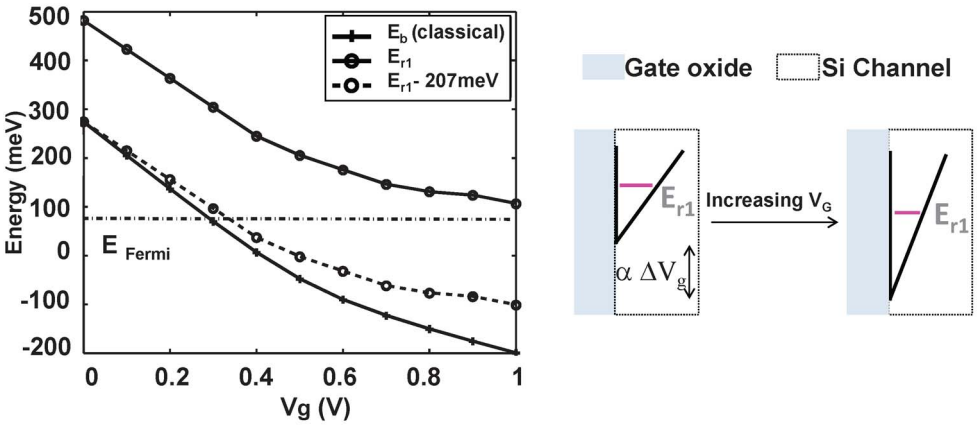


Fig. 17. *(a) Location of the first resonant level E_{r1} and the classical source injection barrier E_b(classical) versus gate voltage. Note that E_{r1} decreases slower than E_b(classical) with gate voltage due to narrowing of channel potential well. (b) Narrowing of the triangular well in the channel with increase in gate bias. E_b(classical) is the bottom of the triangular well, and the resonant level is shown by the horizontal line. (Adapted from [23].)*

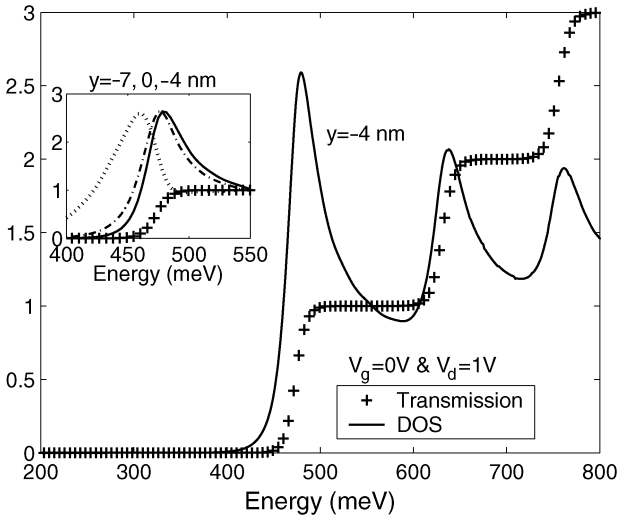


Fig. 18. *Transmission (+) and density of states (solid) versus energy at a spatial location close to the source injection barrier, at $V_g = 0$ V and $V_d = 1$ V. The peaks in the density of states represent the resonant levels in the channel. (Inset) The density of states at three different y-locations and total transmission (+). The points $y = -7$ and 0 nm are to the left and right of the location where the source injection barrier is largest (close to $y = -4$ nm). (From [23].)*

The higher resonant levels do not carry appreciable currents.

We will now address the value of total transmission in a ballistic MOSFET. The transmission is related to drain current (I_d) by (159)

$$I_d = \frac{2e}{h}\int dE\ T(E)[f_S(E) - f_D(E)] \qquad (209)$$

where T is the total transmission from source to drain. f_S and f_D are the Fermi factors in the source and drain, respectively, and the factor of two accounts for spin. As an electron transits from source to drain, the main factors that determine the transmission probability are the tunneling and the scattering in the two-dimensional potential profile. The quantum mechanically computed transmission versus energy is shown in Fig. 18. Four points can be noticed from the transmission curve. First, the transmission increases in a step-wise manner, with the integer values at the plateaus equal to the number of conducting modes in the channel. Secondly, the steps turn on at an energy determined by the effective "subband dependent" source injection barrier (which depends on E_{r1}), that is, the maximum subband energy between the source and drain due to quantization perpendicular to the gate plane (x-direction of Fig. 13). Thirdly, the total transmission assumes integer values at an energy slightly above the maximum in 2-D density of states as shown in the inset of Fig. 18. Fourthly, the transition from one step to the next higher step develops over an energy window of about 50 meV. Interestingly, in the case of MIT25, the current is predominantly carried at energies where the transmission is not an integer.

VIII. APPLICATION TO NANOTRANSISTORS WITH ELECTRON–PHONON SCATTERING

The channel, scattering, and screening lengths become comparable in transistors with diminishing channel lengths and the ballistic transport becomes important (Section VII). However, carrier transport is not fully ballistic. Realistic nanodevice modeling will involve phase-breaking scattering such that transport is between the ballistic and diffusive limits. In this regime, in contrast to long channel devices, carriers are not thermally relaxed in the drain-end of the transistor and are reflected towards the source-end. This reflection of the hot carriers should be explicitly included in the models to compute the drive current. It is in this intermediate regime that the NEGF (with Poisson) method has an advantage over solving the Schrodinger and Poisson equations self-consistently.

In this section, we will first discuss the NEGF equations involved in the two-dimensional modeling of a dual-gate MOSFET (DGMOSFET) with electron–phonon scattering and then illustrate the effect of scattering on the MOSFET characteristics [24]. The related equations, associated approximations and the discretized equations that we solve, are discussed in Section VIII-A, while the application of the NEGF method to demonstrate the role of scattering on transport is discussed in Section VIII-B.

A. Related Equations, Approximations, and Discretization

Fig. 19 shows the schematic of the simulated device. If we consider only the gate-to-gate direction, then we have a 1-D potential well sandwiched between the two oxides so that the quantized energy levels in the channel (well) are approximately given by

$$E_n = \frac{n^2\pi^2\hbar^2}{2m_x T_{ch}^2} \qquad (210)$$

where m_x is the electron effective mass along the x-direction and T_{ch} is the channel thickness, as shown in Fig. 19. When T_{ch} is small, only a few (usually less than four) subbands determine the current–voltage characteristics. In such a case, the mode space approach [27]

gives results comparable to a full 2-D simulation. The mode space approach consists of solving a 1-D Schrodinger equation along the x-direction at each y-position (layer). It is computationally very efficient as it avoids solving a full 2-D Schrodinger equation. The lowest three quantized energy levels for a device with $T_{ch} = 1.5$ nm are 173, 691, and 891 meV above the bulk conduction band. The Fermi energy at the lead doping of 1×10^{20} cm^{-3} is 60 meV above the bulk conduction band. As a result, electrons are injected only into the first subband from the source-end at the operating voltage of $V_d = V_g = 0.6$ V.

The three-dimensional effective mass Hamiltonian is the same as (193). Noting that if the z-direction is infinite, the wave function can be expanded as $\Phi(x,y,z) = e^{ik_z z}\phi(x,y)$. Schrodinger's equation is then

$$\left[E - \frac{\hbar^2 k_z^2}{2m_z^b} - \left(-\frac{\hbar^2}{2}\frac{d}{dy}\left(\frac{1}{m_y^b}\frac{d}{dy}\right) - \frac{\hbar^2}{2}\frac{d}{dx}\left(\frac{1}{m_x^b}\frac{d}{dx}\right) + V(x,y)\right)\right]\phi(x,y) = 0. \quad (211)$$

In the first step of the mode space approach, we solve the following x-directed 1-D Schrodinger equation at each y-position (layer):

$$\left[-\frac{\hbar^2}{2}\frac{d}{dx}\left(\frac{1}{m_x^b}\frac{d}{dx}\right) + V(x,y)\right]\Psi_n(x,y) = E_n(y)\Psi_n(x,y) \quad (212)$$

where $E_n(y)$ is the subband energy and $\Psi_n(x,y)$, the 1-D x-directed wavefunction, both at a certain y. Here $n = \nu, b$, where ν is the quantum number due to quantization in the x-direction and $b = 1, 2$, and 3 are the valley indexes.

Then, the Green's function equations for G, G^n and G^p are solved for each subband n

$$\left[E - \frac{\hbar^2 k_z^2}{2m_z^n} - \left(-\frac{\hbar^2}{2}\frac{d}{dy}\left(\frac{1}{m_y^n}\frac{d}{dy}\right) + E_n(y)\right)\right] \times G_n(y,y',k_z,E) - \int dy_1 \Sigma_n(y,y_1,k_z,E) G_n(y_1,y',k_z,E) = \delta(y-y'), \quad \text{and} \quad (213)$$

$$\left[E - \frac{\hbar^2 k_z^2}{2m_z^n} - \left(-\frac{\hbar^2}{2}\frac{d}{dy}\left(\frac{1}{m_y^n}\frac{d}{dy}\right) + E_n(y)\right)\right] \times G_n^{n(p)}(y,y',k_z,E) - \int dy_1 \Sigma_n(y,y_1,k_z,E) G_n^{n(p)}(y_1,y',k_z,E) = \int dy_1 \Sigma_n^{in(out)}(y,y_1,k_z,E) \times G_n^{\dagger}(y_1,y',k_z,E). \quad (214)$$

m_y^n and m_z^n are the effective masses of silicon in the y- and z-directions that give rise to subband index n. $E_n(y)$ is effectively an electrostatic potential for electrons in subband n. Note that in (213) and (214), the subscript n refers to the subband index while the superscript n refers to the type of Green's function.

The self-energies can be written as

$$\Sigma_n^{\alpha} = \Sigma_{n,C}^{\alpha} + \Sigma_{n,\text{Phonon}}^{\alpha} \quad (215)$$

$$\Sigma_{n,\text{Phonon}}^{\alpha} = \Sigma_{n,el}^{\alpha} + \Sigma_{n,inel}^{\alpha} \quad (216)$$

where $\alpha \in$ (empty), out, in. $\Sigma_{n,C}^{\alpha}$ is the self-energy due to the leads (contacts). The phonon self-energy $\Sigma_{n,\text{Phonon}}^{\alpha}$ consists of two terms: $\Sigma_{n,el}^{\alpha}$ due to elastic and $\Sigma_{n,inel}^{\alpha}$ due to inelastic scattering. Only the lowest three subbands and the electron–phonon scattering among them are considered. All other subbands and the scattering associated with them are neglected because the population of the higher subbands is negligible. The self-energy due to leads is nonzero only at the first (source) and last (drain) grid points.

Assuming isotropic scattering and a phonon reservoir in equilibrium, and using the self-consistent Born

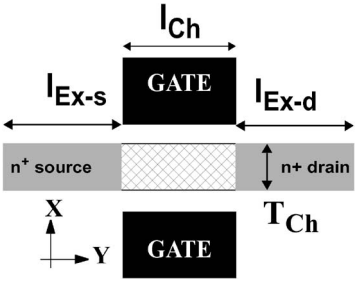


Fig. 19. *Schematic of a DGMOSFET. Ex-s and Ex-d are the extension regions and the hatched region is the channel. The white region between the source/drain/channel and gate is the oxide. The device dimension normal to the page is infinite in extent. (From [24].)*

approximation, the self-energies due to electron–phonon scattering at grid point y_i are [15]

$$\Sigma^{\alpha}_{el,n}(y_i,E) = \sum_{n'} D^{el}_{n,n'} \frac{\sqrt{m_z^{n'}}}{\pi\hbar\sqrt{2}} \int dE_z \frac{1}{\sqrt{E_z}} G^{\alpha}_{n'}(y_i,E_z,E) \tag{217}$$

$$\Sigma^{\text{in}}_{inel,n}(y_i,E) = \sum_{n',\eta} D^{i,\eta}_{n,n'} \frac{\sqrt{m_z^{n'}}}{\pi\hbar\sqrt{2}} \times \int dE_z \frac{1}{\sqrt{E_z}} \left[n_B(\hbar\omega_\eta) G^{n}_{n'}(y_i,E_z,E-\hbar\omega_\eta) + \left(n_B(\hbar\omega_\eta)+1\right) G^{n}_{n'}(y_i,E_z,E+\hbar\omega_\eta) \right] \tag{218}$$

and

$$\Sigma^{\text{out}}_{inel,n}(y_i,E) = \sum_{n',\eta} D^{i,\eta}_{n,n'} \frac{\sqrt{m_z^{n'}}}{\pi\hbar\sqrt{2}} \times \int dE_z \frac{1}{\sqrt{E_z}} \left[n_B(\hbar\omega_\eta) G^{p}_{n'}(y_i,E_z,E+\hbar\omega_\eta) + \left(n_B(\hbar\omega_\eta)+1\right) G^{p}_{n'}(y_i,E_z,E-\hbar\omega_\eta) \right] \tag{219}$$

$$D^{el}_{n,n'} = \left(\delta_{\nu,\nu'} + \frac{1}{2}\right) \delta_{b,b'} \frac{D_A^2 kT}{\rho v^2} \tag{220}$$

$$D^{i,\eta}_{n,n'} = \left(\delta_{\nu,\nu'} + \frac{1}{2}\right) \left[\delta_{b,b'} \frac{D^2_{g\eta}\hbar}{2\rho\omega_{g\eta}} + (1-\delta_{b,b'}) \frac{D^2_{f\eta}\hbar}{\rho\omega_{f\eta}} \right]. \tag{221}$$

Here η represents the phonon modes; n_B represents the Bose factor for phonons in equilibrium; and ω_η, $\omega_{g\eta}$ and $\omega_{f\eta}$ represent phonon frequency due to intraband, g-type, and f-type scattering processes due to mode η, respectively. D_A, $D_{g\eta}$, and $D_{f\eta}$ are the deformation potential for acoustic intraband, g-type, and f-type processes. ρ is the mass density, k is the Boltzmann constant, T is the temperature, and v is the velocity of sound. The values for the above quantities are taken from [13]. b and b' are indexes representing valleys. The following scattering processes are included: acoustic phonon scattering in the elastic approximation and g-type intervalley scattering with phonon energies of 12, 19, and 62 meV. We also remark that the imaginary part of the electron–phonon self-energy, which is responsible for scattering-induced broadening of energy levels, and energetic redistribution of carriers are included but the real part is set to zero. While this approximation has worked well for transistors, it should be used with caution in other situations [24].

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 (213) and (214) is then

$$A_{i,i}G_n(y_i,y_{i'},k_z,E) + A_{i,i+1}G_n\left(y_{i+1},y_i',k_z,E\right) + A_{i,i-1}G_n\left(y_{i-1},y_i',k_z,E\right) = \frac{\delta_{i,i'}}{\Delta y}, \text{ and} \tag{222}$$

$$A_{i,i}G^{\alpha}_n\left(y_i,y_i',k_z,E\right) + A_{i,i+1}G^{\alpha}_n\left(y_{i+1},y_i',k_z,E\right) + A_{i,i-1}G^{\alpha}_n\left(y_{i-1},y_i',k_z,E\right) = \Sigma^{\alpha}_n(y_i,E)G^{\dagger}_n\left(y_i,y_i',k_z,E\right) \tag{223}$$

where

$$A_{i,i} = E - \frac{\hbar^2k_z^2}{2m_z^n} - \frac{\hbar^2}{m_y^n\Delta y^2} - E_n(y_i) - \Sigma_n(y_i,k_z,E) \tag{224}$$

$$A_{i\pm1,i} = \frac{\hbar^2}{2m_z^n\Delta y^2}. \tag{225}$$

The lead self-energies are

$$\Sigma_{n,C}(y_1,k_z,E) = \left(\frac{\hbar^2}{2m_z^n\Delta y^2}\right)^2 g_s(k_z,E) \tag{226}$$

$$\Sigma_{n,C}(y_N,k_z,E) = \left(\frac{\hbar^2}{2m_z^n\Delta y^2}\right)^2 g_d(k_z,E) \tag{227}$$

$$\Sigma^{\text{in}}_{n,C}(y_1,k_z,E) = -2\text{Im}\left(\Sigma_{n,C}(y_1,k_z,E)\right)f_s(E) = \Gamma_s f_s(E) \tag{228}$$

$$\Sigma^{\text{in}}_{n,C}(y_N,k_z,E) = -2\text{Im}\left(\Sigma_{n,C}(y_N,k_z,E)\right)f_d(E) = \Gamma_d f_d(E) \tag{229}$$

$$\Sigma^{\text{out}}_{n,C}(y_1,k_z,E) = -2\text{Im}\left(\Sigma_{n,C}(y_1,k_z,E)\right)[1-f_s(E)] = \Gamma_s[1-f_s(E)] \tag{230}$$

$$\Sigma^{\text{out}}_{n,C}(y_N,k_z,E) = -2\text{Im}\left(\Sigma_{n,C}(y_N,k_z,E)\right)[1-f_d(E)] = \Gamma_d[1-f_d(E)] \tag{231}$$

where y_1 and 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 leads, respectively.

The electron and current densities per energy given by (128) and (149) can be simplified to

$$n_n(y_i, k_z, E) = G^n_n(y_i, y_i, k_z, E) \tag{232}$$

$$J_n(y_i, k_z, E) = \frac{ie}{\hbar}\sum_n \frac{\hbar^2}{2m^n_y \Delta y^2}\left[G^n_n(y_i, y_{i+1}, k_z, E) - G^n_n(y_{i+1}, y_i, k_z, E)\right]. \tag{233}$$

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

$$n(y_i) = 4\sum_n \frac{\sqrt{m^{n'}_z}}{\pi\hbar\sqrt{2}} \times \int \frac{dE}{2\pi}\int dE_z \frac{1}{\sqrt{E_z}} n_n(y_i, E_z, E) \tag{234}$$

$$J(y_i) = 4\sum_n \frac{\sqrt{m^{n'}_z}}{\pi\hbar\sqrt{2}} \times \int \frac{dE}{2\pi}\int dE_z \frac{1}{\sqrt{E_z}} J_n(y_i, E_z, E) \tag{235}$$

where the prefactor of four in the above two equations accounts for twofold spin and valley degeneracies. The nonequilibrium electron and current densities are calculated in both the channel and extension regions using the algorithm for G^n presented in Appendix II.

The Green's function and Poisson's equations are solved self-consistently in the following way. The electrostatic potential is calculated by solving a 2-D Poisson equation using fixed boundary at the gate leads and floating boundary $(dV(y)/dy) = 0$ at the source and drain leads [19]. The applied drain bias corresponds to a difference in the Fermi levels used in the source and drain regions. This potential is used in the mode space calculation to determine the 1-D charge density $n(y)$. Finally, the three-dimensional charge density is determined from

$$n(x_i, y_i, k_z, E) = n_n(y_i, k_z, E)|\Psi_n(x_i, y_i)|^2. \tag{236}$$

The validity of the mode space approach has been tested by [27]. It can be shown that the mode space approach is valid when the wave function $\Psi_n(x, y)$ at various y cross-sections in (212) satisfies $(d\Psi_n(x, y)/dy) \sim 0$. That is, the shape of the wave function at each cross-section should not change significantly along the transport direction. This implies that intersubband scattering due to changes in potential profile are absent. This approximation seems to be valid for channel thickness of less than 5 nm in silicon [27].

B. Results

Using the equations presented in the previous section, we show results illustrating the role of scattering along the channel length of a nanotransistor. First, we show that in devices where the scattering length is comparable to the channel length, the nanotransistor drive current is affected by scattering at all points in the channel. Secondly, we show that when hot electrons enter the drain extension region of a nanotransistor, the drain extension region cannot be modeled as a series resistance. Instead, the drain extension should be included as part of the nonequilibrium simulation region.

The device considered has a channel length of 25 nm, body thickness of 1.5 nm, oxide thickness of 1.5 nm, doping of 1×10^{20} cm^{-3} in the source and drain extension regions, and an intrinsic channel. Scattering is included only from the source end of the channel (-5 nm) to the right boundary of scattering $Y_{R-\mathrm{Scatt}}$ by setting the deformation potential in (220) and (221) to zero to the right of $Y_{R-\mathrm{Scatt}}$. The scattering lengths are decreased by a factor of α by modifying the deformation potential in (220) and (221) by an overall multiplicative factor of $\sqrt{\alpha}$.

Fig. 20 plots the drain current as a function of $(Y_{R-\mathrm{Scatt}})$. When the scattering length due to electron–phonon interaction is 11 nm, the drive current degradation is 30% due to scattering in the right half of the channel. But when the scattering length is smaller (2.2 nm), the drive current degradation is 15%. Thus the scattering in the right half of the channel is important.

To illustrate the drain current degradation due to scattering, we plot the current density $J(y, E)$ as a

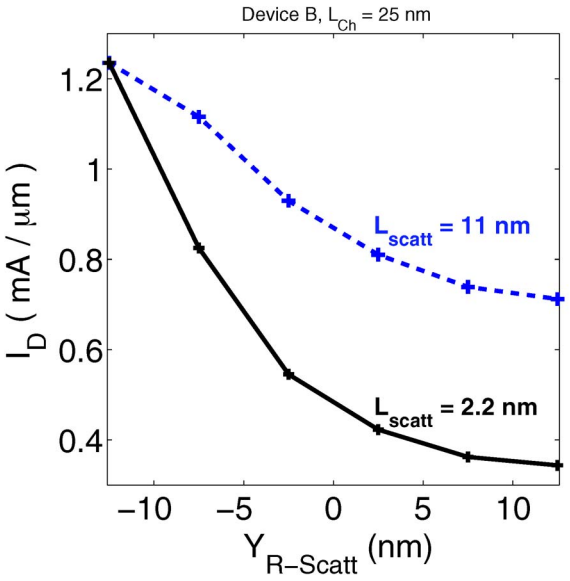


Fig. 20. *Drain current versus $Y_{R-\mathrm{Scatt}}$ for two different scattering lengths. The channel length is 25 nm. (From [24].)*

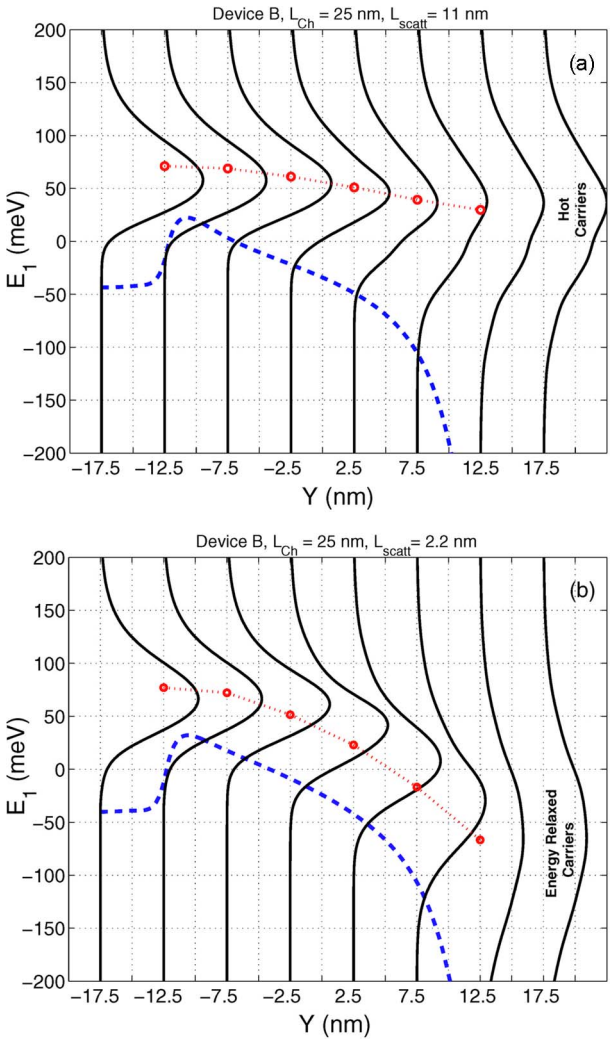


Fig. 21. *Solid lines represent $J(y,E)$ for y equal to -17.5, -12.5, -7.5, -2.5, 2.5, 7.5, 12.5, and 17.5 nm, respectively, when scattering is included every where in the channel. The dashed lines are the first resonant level (E_1) along the channel. The dotted lines represent the first moment of energy with respect to the current distribution function $\int EJ(y,E)dE$. (a) and (b) correspond to $L_{\text{scatt}} = 11$ and 2.2 nm, respectively. (From [24].)*

function of energy at different positions y along the channel (Fig. 21). $J(y,E)$ shows the energetic redistribution of carriers along the channel. When the scattering length is 11 nm, which is comparable to the channel length, $J(y,E)$ in the right-half of the channel is peaked in energy above the source injection barrier, as shown in Fig. 21(a). Scattering causes reflection of these energetic electrons toward the source. These reflected electrons lead to an increase in the channel electron density (classical MOSFET electrostatics). As the charge in the channel should be approximately $C_{\text{ox}}(V_G - V_S)$, the source injection barrier floats to a higher potential energy to compensate for the reflected electrons. The increase in the reflection, and thereby the source injection barrier, leads to a decrease in drain current [14].

To gain further insight into the role of scattering, we now plot $J(y,E)$ for a scattering length of 2.2 nm [Fig. 21(b)], which is five times smaller than the case of 11 nm [Fig. 21(a)]. As the channel length (25 nm) is much larger than 2.2 nm, multiple scattering events now lead to an energy distribution of current (in the right half of the channel) that is peaked well below the source injection barrier.

The first moment of energy with respect to the current distribution function, which is defined as the ratio of $\int EJ(y,E)dE$ to total current, is also shown in Fig. 21 by the dotted line with circles. This quantity is the mean energy at which current flows. When the scattering length is much smaller than the channel length, the carriers relax classically such that the first moment $\int EJ(y,E)dE$ closely tracks the potential profile, as seen in Fig. 21(b).

To further demonstrate the use of NEGF simulations, we study the role of scattering by assuming that the extension regions can be modeled as a classical series resistance. Within the classical series resistance picture, the current with scattering (I_D^{scatt}) can be related to the current without scattering (I_D^{noscatt}) by [26]

$$I_D^{\text{scatt}}(V_D) \sim I_D^{noscatt}(V_D - \delta V_D) \tag{237}$$

where we have assumed that the source extension region and device do not experience scattering. The potential drop in the drain region within the classical series resistance picture is $\delta V_D = I_D^{\text{scatt}}(V_D)R_D$. In Fig. 22, the values of the

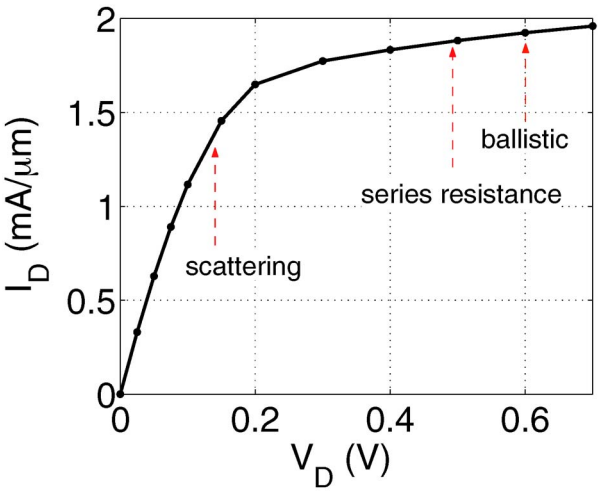


Fig. 22. *The $I_d(V_d)$ plot in the ballistic limit is shown. Marked on this plot are the drive current at $V_d = 0.6$ in the ballistic limit, using the series resistance picture and with scattering included using the NEGF calculations. It should be noted that all three currents marked by the arrows are calculated at $V_d = 0.6$ V. (From [24].)*

drain current versus drain voltage in the ballistic limit are plotted. The currents in the ballistic limit, using the classical series resistance picture and with the NEGF method including scattering, are all marked by arrows. Note that these three values of currents are all calculated at $V_d = 0.6$ V. It is seen from this plot that the current with the NEGF method with scattering is significantly lower than that obtained via the series resistance picture. Clearly, the series resistance picture underestimates the detrimental nature of scattering in the drain end. The physics of the large reduction in drain current was discussed in the context of Fig. 21: When scattering in the channel does not effectively thermalize carriers, the current distribution is peaked at energies above the source injection barrier, upon carriers exiting the channel. Scattering in the drain extension region then causes reflection of electrons toward the source-end. As a result, the source injection increases so as to keep the electron density in the channel approximately at $C_{ox}(V_G - V_S)$. The drain current decreases dramatically as a result of the increase in source injection barrier height.

IX. DISCUSSION AND SUMMARY

Our objectives in this paper have been to:

i) review the underlying assumptions of the traditional, semiclassical treatment of carrier transport in semiconductor devices;
ii) describe how the semiclassical approach can be applied to ballistic transport;
iii) discuss the Landauer–Buttiker approach to quantum transport in the phase coherent limit;
iv) introduce important elements of the non equilibrium Green's function approach using Schrodinger's equation as a starting point;
v) demonstrate the application of the NEGF method to the MOSFET in the ballistic limit and with electron–phonon scattering.

It is appropriate to make a few comments about the computational burden of the various transport models. One reason that device engineers continue to use drift-diffusion simulations rather than the more rigorous Monte Carlo simulations is the enormous difference in computational burden. For semiclassical transport, the fundamental quantity is the carrier distribution function $f(\vec{r}, \vec{k}, t)$. To find $f(\vec{r}, \vec{k})$, we must solve the BTE, which is a six-dimensional equation. The difficulty of solving this six-dimensional equation is one reason that engineers continue to rely on simplified models. For quantum transport, we can take the Green's function $G^n(\vec{r}, \vec{r}', E)$ as the fundamental quantity. The Green's function is a correlation function that describes the phase relationship between the wavefunction at $\vec{r}$ and $\vec{r}'$ for an electron injected at energy E. The quantum transport problem is seven-dimensional, which makes it much harder than the semiclassical problem. We can think of $\vec{r}$ and $\vec{r}'$ as analogous to $\vec{r}$ and $\vec{k}$ in the semiclassical approach, but there is no E in the semiclassical approach. The reason is that, for a bulk semiconductor or in a device in which the potential changes slowly, there is a relation between E and $\vec{k}$, as determined by the semiconductor bandstructure $E(\vec{k})$. When the potential varies rapidly, however, there is no $E(\vec{k})$, and energy becomes a separate dimension. Analysis of electronic devices by quantum simulation is, however, becoming practical because device dimensions are shrinking, which reduces the size of the problem. Quantum simulations are also essential to accurately model devices whose dimensions are comparable to the phase-breaking length, and rely on tunneling and wave interference for operation. The resonant tunneling diode is the most successful example in this category.

We finally remark that the quantum mechanical modeling outlined in this paper reduces to semiclassical modeling when the device dimensions are much longer than the phase-breaking scattering length. The transition from the quantum to semiclassical regime was theoretically addressed by [10] and [15], which derived the Boltzmann transport equation starting from the nonequilibrium Green's function method. Similarly, the transition from the Wigner function to drift-diffusion equations has been established in [2]. The transition from quantum to classical transport in the context of our discussion occurs when the self-energy Σ_{Phonon} in (105), which represents electron–phonon interaction, is nonzero. Electron–phonon interaction causes reflection and breaks the phase-coherent evolution of electron waves incident from the contacts. Apart from this, electron–phonon interaction causes energy dissipation in the device. These features result in a transition from quantum mechanical to classical behavior. ■

APPENDIX I
DYSON'S EQUATION FOR LAYERED STRUCTURES

Partition the device layers into two regions Z and Z' as shown in Fig. 23. Dyson's equation is a very useful method that relates the Green's function of the full system $Z + Z'$ in terms of the subsystems Z, Z' and the coupling between Z and Z'. We will see below that, from a computational point of view, Dyson's equation provides us with a systematic framework to calculate the diagonal blocks of G and G^n without full inversion of the A matrix. The reader should note that Dyson's equation has a significantly more general validity than implied in our application here [16].

A. Dyson's Equation for G

The Green's function equation over the device layers [(122)]

$$AG = I \tag{A1}$$

$$GA = I \tag{A2}$$

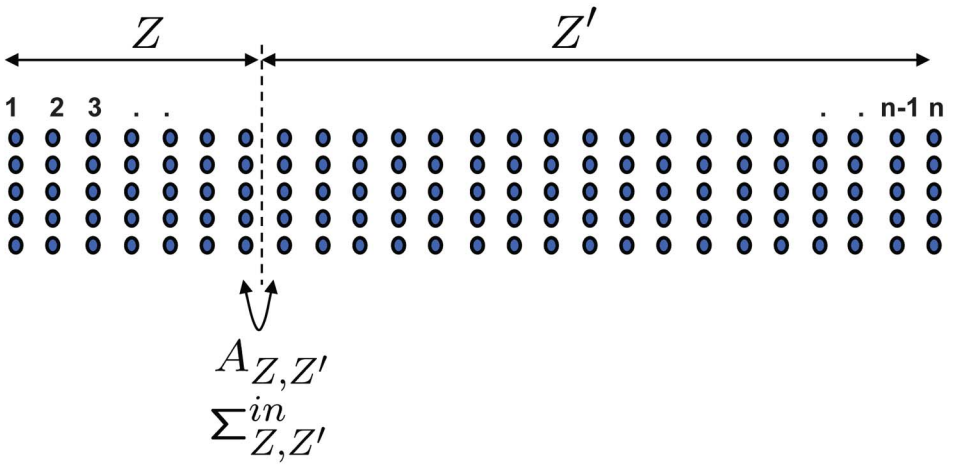


Fig. 23. *Scheme of device for application of Dyson's equation by splitting the device in two parts. (Adapted from [23].)*

can be written as

$$\begin{pmatrix} A_{Z,Z} & A_{Z,Z'} \\ A_{Z',Z} & A_{Z',Z'} \end{pmatrix} \begin{pmatrix} G_{Z,Z} & G_{Z,Z'} \\ G_{Z',Z} & G_{Z',Z'} \end{pmatrix} = \begin{pmatrix} I & O \\ O & I \end{pmatrix}. \quad \text{(A3)}$$

The solution of (A1) is

$$G = G^0 + G^0 UG \quad \text{(A4)}$$
$$= G^0 + GUG^0 \quad \text{(A5)}$$

where

$$G = \begin{pmatrix} G_{Z,Z} & G_{Z,Z'} \\ G_{Z',Z} & G_{Z',Z'} \end{pmatrix}$$
$$G^0 = \begin{pmatrix} G^0_{Z,Z} & O \\ O & G^0_{Z',Z'} \end{pmatrix} = \begin{pmatrix} A^{-1}_{Z,Z} & O \\ O & A^{-1}_{Z',Z'} \end{pmatrix}$$
$$U = \begin{pmatrix} O & -A_{Z,Z'} \\ -A_{Z',Z} & O \end{pmatrix}. \quad \text{(A6)}$$

It is verified by direct substitution of solutions (A5) and (A4) to (A1) and (A2), respectively, and then, using the above defintions

$$(A + U)G^0 = I \quad \text{(A7)}$$

or

$$G^0(A + U) = I. \quad \text{(A8)}$$

The Hermitian conjugate Green's function $(G^\dagger)$ is by definition related to G by

$$G^\dagger = G^{\dagger 0} + G^{\dagger 0} U^\dagger G^\dagger \quad \text{(A9)}$$
$$= G^{\dagger 0} + G^\dagger U^\dagger G^{\dagger 0}. \quad \text{(A10)}$$

Equation (A4) is Dyson's equation for the Green's function.

B. Dyson's Equation for G^n

The electron correlation function equation over the device layers [(131)]

$$AG^n = \Sigma^{\text{in}} G^\dagger \quad \text{(A11)}$$

can be written as

$$\begin{pmatrix} A_{Z,Z} & A_{Z,Z'} \\ A_{Z',Z} & A_{Z',Z'} \end{pmatrix} \begin{pmatrix} G^n_{Z,Z} & G^n_{Z,Z'} \\ G^n_{Z',Z} & G^n_{Z',Z'} \end{pmatrix} = \begin{pmatrix} \Sigma^{\text{in}}_{Z,Z} & \Sigma^{\text{in}}_{Z,Z'} \\ \Sigma^{\text{in}}_{Z',Z} & \Sigma^{\text{in}}_{Z',Z'} \end{pmatrix} \begin{pmatrix} G^\dagger_{Z,Z} & G^\dagger_{Z,Z'} \\ G^\dagger_{Z',Z} & G^\dagger_{Z',Z'} \end{pmatrix}. \quad \text{(A12)}$$

The solution of (A11) is

$$G^n = G^0 U G^n + G^0 \Sigma^{\text{in}} G^\dagger \quad \text{(A13)}$$

where G^0 and U have been defined in (A6). Functions G^n and $G^\dagger$ are readily defined by (A11) and (A10), respectively.

Using $G^{\dagger} = G^{\dagger 0} + G^{\dagger 0}U^{\dagger}G^{\dagger}$, (A13) can be written as

$$G^n = G^{n0} + G^{n0}U^{\dagger}G^{\dagger} + G^0UG^n \quad \text{(A14)}$$

$$= G^{n0} + GUG^{n0} + G^nU^{\dagger}G^{\dagger 0} \quad \text{(A15)}$$

where

$$G^{n0} = G^0\Sigma^{\text{in}}G^{\dagger 0}. \quad \text{(A16)}$$

APPENDIX II
ALGORITHM TO CALCULATE G AND G^n

Why algorithm: A typical simulation of a nanoelectronic device consists of solving Poisson's equation self-consistently with the Green's function equations. The input to Poisson's equation is the charge density, which is obtained by integrating $G^n_{q,q}(E)$ over energy. The index q here runs over the layers of the device. In order to calculate the current density, one requires the elements $G^n_{q,q+1}(E)$ from the diagonals adjacent to the main diagonal. That is, we *do not* require the entire G^n matrix in most situations. Instead, just the three diagonals of elements or of blocks of elements need to be calculated. This also applies for the hole correlation function G^p.

Provided that N_x is the dimension of the Hamiltonian of each layer and N is the total number of layers, the size of the matrix A equals N_xN. The operation count for the full matrix inversion $G = A^{-1}$ is proportional to $(N_xN)^3$. The computational cost of obtaining the diagonal elements of the G^n matrix at each energy is approximately $N_x^3N^3$ operations if $G^n = G\Sigma^{\text{in}}G^{\dagger}$ is used. Therefore, it is highly desirable to find less expensive algorithms that avoid full inversion of matrix A and specifically take advantage of the fact that only the diagonal elements of Green's functions are required. Another reason to prefer such algorithms is the memory storage. If one had had to retain the whole matrix G in the memory, it might had not fit into the on-chip cache and had required using slower access memory (RAM or a swap file). That would have significantly slowed down the calculations.

One such algorithm to calculate G^n and G^p that is valid for the block tridiagonal form of matrix A developed in [23] is presented in this section. The algorithm to calculate the diagonal blocks of G was developed in [12]. The operation count of this algorithm scales approximately as N_x^3N. The dependence on N_x^3 arises because matrices of dimension of the sub-Hamiltonian of layers should be inverted, and the dependence on N corresponds to one such inversion for each of the N layers.

The algorithm consists of two steps. In the first step, the diagonal blocks of the left connected and full Green's function are evaluated. In the second step, these results are used to evaluate the diagonal blocks of the G^n Green's function.

A. Recursive Algorithm for G

i) Left-connected Green's function (Fig. 24): The left-connected (superscript L) Green's function g^{Lq} is defined by the first q blocks of (122) by [12]

$$A_{1:q,1:q}g^{Lq} = I_{q,q} \quad \text{(B1)}$$

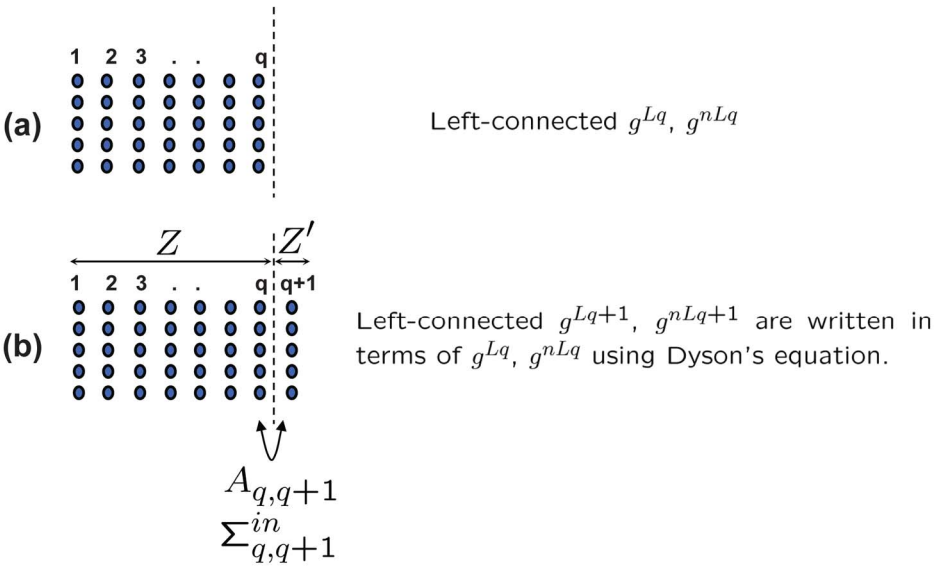


Fig. 24. *Illustration for the relation between the left-connected Green's functions for adjacent layers. (Adapted from [23].)*

where we introduce a shorthand $I_{q,q} = I_{1:q,1:q}$. The matrix g^{Lq+1} is defined in a manner identical to g^{Lq} except that the left-connected system is composed of the first $q+1$ blocks of (122). In terms of (A3), the equation governing g^{Lq+1} can be expressed via the solution g^{Lq} by setting $Z = 1:q$ and $Z' = q+1$. Using Dyson's equation [(A4)], we obtain

$$g^{Lq+1}_{q+1,q+1} = \left(A_{q+1,q+1} - A_{q+1,q} g^{Lq}_{q,q} A_{q,q+1}\right)^{-1}. \quad \text{(B2)}$$

Note that the last element of this progression g^{LN} is equal to the fully connected Green's function G, which is the solution to (122).

ii) Full Green's function in terms of the left-connected Green's function: Consider the special case of (A3) in which $A_{Z,Z} = A_{1:q,1:q}$, $A_{Z',Z'} = A_{q+1:N,q+1:N}$, and $A_{Z,Z'} = A_{1:q,q+1:N}$. Noting that the only nonzero element of $A_{1:q,q+1:N}$ is $A_{q,q+1}$ and using (A4), we obtain

$$G_{q,q} = g^{Lq}_{q,q} + g^{Lq}_{q,q}\left(A_{q,q+1} G_{q+1,q+1} A_{q+1,q}\right) g^{Lq}_{q,q} \quad \text{(B3)}$$

$$= g^{Lq}_{q,q} - g^{Lq}_{q,q} A_{q,q+1} G_{q+1,q}. \quad \text{(B4)}$$

The equations for the adjacent diagonals are obtained similarly as

$$G_{q+1,q} = -G_{q+1,q+1} A_{q+1,q} g^{Lq}_{q,q} \quad \text{(B5)}$$

$$G_{q,q+1} = -g^{Lq}_{q,q} A_{q,q+1} G_{q+1,q+1}. \quad \text{(B6)}$$

Both $G_{q,q}$ and $G_{q+1,q}$ are used in the algorithm for electron density, and so storing both sets of matrices is necessary.

Making use of the above equations, the algorithm to obtain the three diagonals of G is as follows.

1) $g^{L1}_{11} = A^{-1}_{11}$.
2) For $q = 1, 2, \ldots, N-1$, compute (B2).
3) For $q = 1, 2, \ldots, N$, compute $\left(g^{Lq}_{qq}\right)^{\dagger}$.
4) $G_{N,N} = g^{Lq}_{q,q}$.
5) For $q = N-1, N-2, \ldots, 1$, compute (B5), (B6), and (B4) (in this order).
6) For $q = 1, 2, \ldots, N$, compute $(G_{q,q+1})^{\dagger}$ and $(G_{q+1,q})^{\dagger}$.

B. Recursive Algorithm for G^n

i) Left-connected G^n (Fig. 24): The function g^{nLq} is the counterpart of g^{Lq} and is defined by the first q blocks of (131) [23]

$$A_{1:q,1:q}\, g^{nLq} = \Sigma^{\text{in}}_{1:q,1:q}\, g^{\dagger Lq}_{1:q,1:q}. \quad \text{(B7)}$$

g^{nLq+1} is defined in a manner identical to g^{nLq} except that the left-connected system is composed of the first $q+1$ blocks of (131).

The equation governing g^{nLq+1} follows from (A12) by setting $Z = 1:q$ and $Z' = q+1$. Using the Dyson's equations for G [(A4)] and G^n [(A14)], $g^{nLq+1}_{q+1,q+1}$ is recursively obtained as

$$g^{nLq+1}_{q+1,q+1} = g^{Lq+1}_{q+1,q+1}\left[\Sigma^{\text{in}}_{q+1,q+1} + \sigma^{\text{in}}_{q+1}\right] g^{Lq+1\dagger}_{q+1,q+1} \quad \text{(B8)}$$

where $\sigma^{\text{in}}_{q+1} = A_{q+1,q} g^{nLq}_{q,q} A^{\dagger}_{q,q+1}$. Equation (B8) has the physical meaning that $g^{nLq+1}_{q+1,q+1}$ has contributions due to an effective self-energy due to the left-connected structure that ends at q, which is represented by σ^{in}_{q+1} and the diagonal self-energy component at grid point $q+1$ (Σ^{in}_{DD} of (131)).

ii) Full electron correlation function in terms of left-connected Green's function: Consider (A12) such that $A_Z = A_{1:q,1:q}$, $A'_Z = A_{q+1:N,q+1:N}$ and $A_{Z,Z'} = A_{1:q,q+1:N}$. Noting that the only non zero element of $A_{1:q,q+1:N}$ is $A_{q,q+1}$ and using (A14), we obtain

$$G^n_{q,q} = g^{nLq}_{q,q} - g^{nLq}_{q,q} A^{\dagger}_{q,q+1} G^{\dagger}_{q+1,q} - g^{Lq}_{q,q} A_{q,q+1} G^n_{q+1,q}. \quad \text{(B9)}$$

Using (A15), $G^n_{q+1,q}$ can be written in terms of $G^n_{q+1,q+1}$ and other known Green's functions as

$$G^n_{q+1,q} = -G_{q+1,q+1} A_{q+1,q} g^{nLq}_{q,q} - G^n_{q+1,q+1} A^{\dagger}_{q+1,q} g^{\dagger Lq}_{q,q}. \quad \text{(B10)}$$

Substituting (B10) in (B9) and using (A4) and (A5), we obtain

$$G^n_{q,q} = g^{nLq}_{q,q} + g^{Lq}_{q,q}\left(A_{q,q+1} G^n_{q+1,q+1} A^{\dagger}_{q+1,q}\right) g^{\dagger Lq}_{q,q} - \left[g^{nLq}_{q,q} A^{\dagger}_{q,q+1} G^{\dagger}_{q+1,q} + G_{q,q+1} A_{q+1,q} g^{nLq}_{q,q}\right]. \quad \text{(B11)}$$

The terms inside the square brackets of (B11) are Hermitian conjugates of each other. In view of the above equations, the algorithm to compute the diagonal blocks of G^n and G^p is given by the following steps.

1) $g^{nL1}_{11} = g^{L1}_{11} \Sigma^{\text{in}}_{11} g^{L1\dagger}_{11}$.
2) For $q = 1, 2, \ldots, N-1$, compute (B8).
3) $G^n_{N,N} = g^{nLN}_{NN}$.

4) For $q = N-1, N-2, \ldots, 1$, compute (B11) and (B10).
5) Use $G^n_{q,q+1} = (G^n_{q+1,q})^{\dagger}$.
6) Use $G^p = i(G - G^{\dagger}) - G^n$.

The above algorithm is illustrated by a Matlab code in Appendix III.

Challenging problem: The algorithm presented here solves for the three block diagonals of G, G^n, and G^p. Each of n blocks on the main diagonal corresponds to a layer of the device. All blocks in the three diagonals are treated as a full matrix. It is highly desirable to find a more efficient algorithm that finds only the diagonal elements within each block rather than complete blocks.

APPENDIX III
CODE OF THE RECURSIVE ALGORITHM

The listing of Matlab code recursealg3d.m is provided here for illustration of the algorithm described in Appendix II. Open-source simulation tools written in Matlab and using this algorithm are nanoMOS [19], [20], for a 2-D semiconductor transistor, and MOSCNT [11], for a carbon nanotube transistor.

```
function [Grl,Grd,Gru,Gnl,Gnd,Gnu,Gpl,Gpd,Gpu,grL,ginL] ...
= recursealg3d(Np,Al,Ad,Au,Sigin,Sigout)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% function [Grl,Grd,Gru,Gnl,Gnd,Gnu,Gpl,Gpd,Gpu,grL,ginL] ...
% = recursealg3d(Np,Al,Ad,Au,Sigin,Sigout)
% recursive algorithm to solve for the diagonal elements of
% the nonequilibrium Green's function
% "l" = lower diagonal, size (1, Np-1)=[G(2,1) ... G(end,end-1)]
% "d" = main diagonal, size (1,Np) = [G(1,1) ... G(end,end)]
% "u" = upper diagonal, size (1,Np - 1) = [G(1,2) .... G(end - 1,end)]
% Grl, Grd, Gru=retarded Green's function
% Gnl, Gnd, Gnu=electron Green's function
% Gpl, Gpd, Gpu=hole Green's function
% grL=left-connected Green's function
% ginL=left-connected in-scattering function
% Np= size of the matrices
% Al, Ad, Au= matrix of coefficients
% Sigin= matrix of in-scattering self-energies (diagonal)
% Sigout= matrix of out-scattering self-energies (diagonal)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Dmitri Nikonov, Intel Corp. and Siyu Koswatta, Purdue University, 2004
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
flag_Gp = 'no';
Al_cr = conj(Au);
Ad_cr = conj(Ad);                  % Hermitian conjugate of the coefficient matrix
Au_cr = conj(Al);
grL = zeros(1,Np);                 % initialize left-connected function
ginL = zeros(1,Np);                % initialize left-connected in-scattering function
gipL = zeros(1,Np);                % initialize left-connected out-scattering function
Grl = zeros(1,Np - 1);
Grd = zeros(1,Np);                 % initialize the Green's function
Gru = zeros(1,Np - 1);
Gnl = zeros(1,Np - 1);
Gnd = zeros(1,Np);                 % initialize the electron coherence function
Gnu = zeros(1,Np - 1);
Gpl = zeros(1,Np - 1);
Gpd = zeros(1,Np);                 % initialize the hole coherence function
Gpu = zeros(1,Np - 1);
grL(1) = 1/Ad(1);                  % step 1
```

```
for q = 2 : Np                                           % obtain the left-connected function
    grL(q) = 1/(Ad (q) - Al(q - 1) * grL(q - 1) * Au(q - 1));
end
gaL = conj(grL);                                         % advanced left-connected function
Grd(Np) = grL(Np);                                       % step 2
for q = (Np - 1) : -1 : 1
    Grl(q) = -Grd(q + 1) * Al(q) * grL(q);               % obtain the subdiagonal of the Green's function
    Gru(q) = -grL(q) * Au(q) * Grd(q + 1);               % obtain the super-diagonal of the Green's function
    Grd(q) = grL(q) - grL(q) * Au(q) * Grl(q);           % obtain the diagonal of the Green's function
end
Gal = conj(Gru);
Gad = conj(Grd);                                         % advanced Green's function
Gau = conj(Gal);
ginL(1) = grL(1) * Sigin(1) * gaL(1);                    % step 3
for q = 2 : Np
    sla2 = Al(q - 1) * ginL(q - 1) * Au_cr(q - 1);
    prom = Sigin(q) + sla2;
    ginL(q) = real(grL(q) * prom * gaL(q));              % left-connected in-scattering function
end
Gnd(Np) = ginL(Np);                                      % step 4
for q = (Np - 1) : -1 : 1
    Gnl(q) = -Grd(q + 1) * Al(q) * ginL(q) - Gnd(q + 1) *Al_cr(q) * gaL(q);
    % obtain the lower diagonal of the electron Green's function
    Gnd(q) = real(ginL(q) + grL(q) * Au(q) * Gnd(q + 1) * Al_cr(q) * gaL(q) ...
      -(ginL(q) * Au_cr(q) * Gal(q)+ Gru(q) * Al(q) * ginL(q) ));
end
Gnu = conj(Gnl);                                         % upper diagonal of the electron function
switch flag_Gp
    case 'yes'
gipL(1) = grL(1) * Sigout(1) * gaL(1);                   % step 5
for q = 2 : Np
    sla2 = Al(q - 1) * gipL(q - 1) * Au_cr(q - 1);
    prom = Sigout(q) + sla2;
    gipL(q) = real(grL(q) * prom * gaL(q));              % left-connected out-scattering function
end
Gpd(Np) = gipL(Np);                                      % step 6
for q = (Np - 1) : -1 : 1
    Gpl(q) = -Grd(q + 1) * Al(q) * gipL(q) - Gnd(q + 1) *Al_cr(q) * gaL(q);
    % obtain the lower diagonal of the hole Green's function
    Gpd(q) = real(gipL(q) + grL(q) * Au(q) * Gpd(q + 1) * Al_cr(q) * gaL(q) ...
      -(gipL(q) * Au_cr(q) *  Gal(q) + Gru(q) * Al(q) * gipL(q) ));
end
Gpu = conj(Gpl);                                         % upper diagonal of the hole function
    case 'no'
Gpl = i * (Grl - Gal) - Gnl;
Gpd = i * (Grd - Gad) - Gnd;                             % hole Green's function
Gpu = i * (Gru - Gau) - Gnu;
end
Gnd = real(Gnd);
Gpd = real(Gpd);
jnzer = find(Gnd<0);
Gnd(jnzer) = 0;
jpzer = find(Gpd<0);
Gpd(jpzer) = 0;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Acknowledgment

The authors would like to thank S. Datta for many useful and inspirational discussions on the NEGF method. M. S. Lundstrom is indebted to his students who have, over the past several years, applied the NEGF methods to numerous problems in nanotransistors. Anantram is grateful to T. R. Govindan and A. Svizhenko for useful discussions. Anantram would also like to thank A. Fadavi, A. Ramu, D. Shiri, and G. Rabbani for their comments. The authors also acknowledge the IEEE and the *Journal of Applied Physics* for permission to use material from research papers originally published in their journals.

REFERENCES

[1] M. P. Anantram and A. Svizhenko, *IEEE Trans. Electron Devices*, vol. 54, p. 2100, 2007.

[2] M. Ancona and G. J. Iafrate, *Phys. Rev. B*, vol. 39, p. 9536, 1989.

[3] M. Buttiker, "Scattering theory of current and intensity noise correlations in conductors and wave guides," *Phys. Rev. B*, vol. 46, p. 12 485, 1992.

[4] S. Datta, "A simple kinetic equation for steady-state quantum transport," *J. Phys. Condens. Matter*, vol. 2, pp. 8023–8052, 1990.

[5] S. Datta, *Electronic Conduction in Mesoscopic Systems*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[6] M. V. Fischetti and S. E. Laux, *Phys. Rev. B*, vol. 48, p. 2244, 1993.

[7] C. Jacoboni and P. Lugli, *The Monte Carlo Method for Semiconductor Device Simulation*. New York: Springer-Verlag, 1989.

[8] L. Kadanoff and G. Baym, *Quantum Statistical Mechanics: Green's Function Methods in Equilibrium and Non Equilibrium Problems*. New York: W. A. Benjamin, 1962.

[9] L. V. Keldysh, "Diagram technique for nonequil. processes," *Zh. Eksp. Teor. Fiz.*, vol. 47, p. 1515, 1964.

[10] F. S. Khan, J. H. Davies, and J. W. Wilkins, *Phys. Rev. B*, vol. 36, p. 2578, 1987.

[11] S. Koswatta, J. Guo, and D. E. Nikonov. (2006). MOSCNT: Code for carbon nanotube transistor simulation. [Online]. Available: https://www.nanohub.org/resources/1989/

[12] R. Lake, G. Klimeck, R. C. Bowen, and D. Jovanovic, *J. Appl. Phys.*, vol. 81, p. 7845, 1997.

[13] M. S. Lundstrom, *Fundamentals of Carrier Transport*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[14] M. S. Lundstrom, "Elementary scattering theory of the MOSFET," *IEEE Electron Device Lett.*, vol. 18, p. 361, 1997.

[15] G. D. Mahan, *Phys. Rep.*, vol. 145, p. 251, 1987.

[16] G. D. Mahan, *Many Particle Physics*. New York: Plenum, 1990.

[17] Y. Meir and N. S. Wingreen, "Landauer formula for the current through an interacting electron region," *Phys. Rev. Lett.*, vol. 68, p. 2512, 1992.

[18] R. F. Pierret, *Advanced Semiconductor Fundamentals*. Reading, MA: Addison-Wesley, 1987.

[19] Z. Ren, R. Venugopal, S. Goasguen, S. Datta, and M. S. Lundstrom, "nanoMOS 2.5: A two-dimensional simulator for quantum transport in double-gate MOSFETs," *IEEE Trans. Electron. Devices and IEEE Trans. Nanotechnol. (Joint Special Issue on Nanoelectronics*, vol. 50, p. 1914, 2003.

[20] Z. Ren, S. Goasguen, A. Matsudaira, S. S. Ahmed, K. Cantley, and M. Lundstrom. (2006). nanoMOS. [Online]. Available: https://www.nanohub.org/tools/nanomos/

[21] J.-H. Rhew, Z. Ren, and M. Lundstrom, "A numerical study of ballistic transport in a nanoscale MOSFET," *Solid-State Electron.*, vol. 46, p. 1899, 2002.

[22] W. Shockley, *Electrons and Holes in Semiconductors*. New York: Van Nostrand, 1950.

[23] 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, p. 2343, 2002.

[24] A. Svizhenko and M. P. Anantram, "Role of scattering in nanotransistors," *IEEE Trans. Electron Devices*, vol. 50, p. 1459, 2003.

[25] A. Svizhenko and M. P. Anantram, "Effect of scattering and contacts on current and electrostatics in carbon nanotubes," *Phys. Rev. B*, vol. 72, p. 085430, 2005.

[26] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[27] 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, p. 3730, 2002.

[28] N. S. Wingreen and Y. Meir, "Anderson model out of equilibrium: Noncrossing-approximation approach to transport through a quantum dot," *Phys. Rev. B*, vol. 49, p. 11 040, 1994.

ABOUT THE AUTHORS



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, IN, in 1995.

He was the Group Lead for Computational Nanoelectronics and, since 2004, has been a Fellow with the University Affiliated Research Center, National Aeronautics and Space Administration, Ames Research Center, where his group developed the first 2-D quantum simulator for nanotransistors and 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 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. He serves on the program and organizing committees of nanotechnology conferences.

Dr. Anantram is an Associate Editor of the IEEE TRANSACTIONS ON NANOTECHNOLOGY and the Education Chair of the IEEE Nanotechnology Council. He has received best paper awards for his work on nanotransistors and electromechanical properties of nanotubes. He was a corecipient of the Highest Technical Achievement in Applied Sciences from the Computer Sciences Corporation in 2003 for his work on theory and modeling of nanodevices.



Mark S. Lundstrom (Fellow, IEEE) received the B.E.E. and M.S.E.E. degrees from the University of Minnesota, Minneapolis, in 1973 and 1974, respectively, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, in 1980.

From 1974 to 1977, he was with the Hewlett-Packard Corporation, Loveland, CO, working on integrated circuit process development and manufacturing support. In 1980, he joined the School of Electrical and Computer Engineering, Purdue University, where he is currently the Don and Carol Scifres Distinguished Professor of Electrical and Computer Engineering and the founding Director of the Network for Computational Nanotechnology. From 1989 to 1993, he was a Director of the Optoelectronics Research Center, Purdue University, and, from 1991 to 1994, the Assistant Dean of Engineering. His current research interests include the physics of small electronic devices, especially nanoscale transistors, and carrier transport in semiconductor devices.

Prof. Lundstrom is a Fellow of the American Physical Society and the American Association for the Advancement of Science. He is a Distinguished Lecturer of the IEEE Electron Device Society. He received the Frederick Emmons Terman Award from the American Society for Engineering Education in 1992. He (with S. Datta) received the IEEE Cledo Brunetti Award for their work on nanoscale electronic devices and the Semiconductor Research Corporation's Technical Excellence Award in 2002. He received the Semiconductor Industry Association's University Researcher Award for his career contributions to the physics and simulation of semiconductor devices in 2005. Most recently, in 2006, he was the inaugural recipient of the IEEE Electron Device Society's Education Award.



Dmitri E. Nikonov (Senior Member, IEEE) received the M.S. degree in aeromechanical engineering from the Moscow Institute of Physics and Technology, Moscow, Russia, in 1992 and the Ph.D. degree in physics from Texas A&M University, College Station, in 1996.

While at Texas A&M he participated in the demonstration of the world's first laser without population inversion. He joined Intel Corporation in 1998, where he is presently a Project Manager with the Technology Strategy Group, Santa Clara, CA. He is responsible for managing joint research programs with universities on nanotechnology, optoelectronics, and advanced devices. From 1997 to 1998, he was a Research Engineer and Lecturer with the Department of Electrical and Computer Engineering, University of California, Santa Barbara. In 2006, he became an Adjunct Associate Professor of electrical and computer engineering at Purdue University, West Lafayette, IN. He has 38 publications in refereed journals in quantum optics, free-electron, gas and semiconductor lasers, nanoelectronics, spintronics, and quantum devices simulation. He has received 29 patents in optoelectronics, integrated optics, and spintronic devices.

Dr. Nikonov was a Finalist for the Best Doctoral Thesis award from the American Physical Society in 1997.