One of the most interesting devices under exploration now is the nanowire transistor. Because of the size of these devices, there have been many approaches to implement full quantum mechanical simulations. To date, most of these approaches have considered only ballistic transport or impurity scattering through the self-consistent potential. However, in order to understand the operation of any type of semiconductor device, one must consider the effects of scattering. For many years, Monte Carlo has been the workhorse that has yielded great insight into the operation of a wide variety of semiconductor devices because of the ease in treating various scattering effects, and these approaches have been modified for simple quantum effects. But in these very small devices, a more exact treatment of the quantum effects is required. To address this problem, we have developed a proper real-space self-energy which can be included within the Hamiltonian formalism for either the recursive scattering matrix approach or the non-equilibrium Green’s function approach. More importantly, this treatment with the proper self-energy is norm-conserving but converts the Hamiltonian to a non-Hermitian form, as expected. When the model is applied to a gated quantum wire transistor structure under bias, we find that the scattering due to phonons causes the electrons to scatter to higher subbands than previously available during simple ballistic transport, with the result that the ballistic-to-diffusive transition in Si devices occurs for gate lengths of the order of 1 nm.

1 Introduction

The Metal-Oxide-Semiconductor Field-Effect Transistor (MOSFET) has been a staple of the semiconductor industry for many years, with progress following a well-known scaling relationship [1]. Nevertheless, it is becoming increasingly clear that the continued scaling of these devices will not meet the expected performance requirements and other solutions are under examination [2]. In the silicon area, many new transistor structures, such as finfets [3] and tri-gate transistors [2] have shown interesting promise. With such new device concepts, more scaling is expected, and 15 nm gate lengths are scheduled for production before the end of this decade. Such short gate length devices have been demonstrated by Intel [2] and AMD [4], and IBM has recently shown a 6 nm gate length p-channel transistor [5]. While the creation of these very small transistors is remarkable enough, the fact that they seem to operate in a quite normal fashion is perhaps even more remarkable. Almost 25 years ago, the prospect of making such small transistors was discussed, and a suggested technique for a 25 nm gate length, Schottky source-drain device, was proposed [6]. At that time, it was suggested that the central feature of transport in such small devices would be that the micro-dynamics could not be treated in isolation from the overall device environment (of a great many similar devices). Rather, it was thought that the transport would by necessity be described by quantum transport and that the array of such small devices on the chip would lead to considerable coherent many-device interactions. While this early suggestion does not seem to have been fulfilled, as witnessed by the quite normal behavior of the above-mentioned devices, there have been many subsequent suggestions for treatment via quantum transport [7,8,9,10,11]. Moreover, there is ample suggestion that the transport will not be normal, but will have significant ballistic transport effects [12] and this, in turn, will lead to quantum transport effects.

Recently, we presented a fully quantum mechanical, and three-dimensional, ballistic simulation of small silicon quantum-wire MOSFETs [13]. In this approach, the full Poisson equation solution is used to determine the local potential, and a recursive scattering matrix approach is used to determine the transport through the device. In this process, for each iteration from one transverse slice of the device to the next, a local Dyson’s equation is solved with the slice Hamiltonian. This means that we can modify this Hamiltonian by the direct inclusion of a slice self-energy as well as a self-energy coupling between the slices where that is appropriate. In this paper, we will review the simulation technique and the inclusion of the self-energy terms necessary to yield the dissipative results.
2 The Recursive Scattering Matrix Approach

There have been many suggestions for different quantum methods to model ultra-small semiconductor devices [14,15,16]. However, in each of these recent approaches, the length and the depth of the device were modeled rigorously, while the third dimension (width) was usually included through the assumption that there is no interesting physics in this dimension (lateral homogeneity). Moreover, it was assumed that the mode does not change shape as it propagates from the source of the device to the drain of the device. Other simulation proposals have simply assumed that only one sub-band in the orthogonal direction is occupied, therefore making higher-dimensional transport considerations unnecessary [10]. These may not be valid assumptions, especially as we approach devices whose width is comparable to the channel length, both of which may be less than 10 nm.

It is important to consider all the modes that may be excited in the source (or drain) region, as this may be responsible for some of the more interesting physics that must be captured. In the source, the modes that are excited are three dimensional (3D) in nature, even in a thin SOI device. These modes are then propagated from the source to the channel, and the coupling among the various modes will be dependent upon the details of the local confining potential at each point along the channel. Moreover, as the doping and the Fermi level in short-channel MOSFETs increases, we can no longer assume that there is only one occupied sub-band. In an effort to provide a more complete simulation method, we have developed a full 3D quantum simulation, based on the use of recursive scattering matrices, and which is being used in our group to simulate short-channel, fully-depleted SOI MOSFET devices [11,13].

The device under consideration is a fully-depleted SOI MOSFET structure, shown schematically in Fig. 1. We orient the $x$ and $z$ directions in order to correspond to the length and the height (thickness of the SOI layer) of the device respectively. In the $x$ direction, the source and the drain contact regions are 10 nm in length and 18 nm in width (lateral direction, the $y$ axis). In an actual device, the length of the source and the drain of a MOSFET would be much longer, but this length captures the important energy relaxation length. We implement open boundary conditions at the ends of the structure and on the sidewalls. The gate length of the device is varied for a number of considerations, but corresponds to a dimension that will allow the gate to fully control that channel of the device. The channel itself is typically 9 nm in width, so that the Si layer is a wide-narrow-wide structure as shown in the figure. The entire structure is on a silicon layer that is taken to be only 6 nm thick, with a 10 nm buried oxide (BOX) layer below this layer. The gate oxide is taken to be 2 nm thick.

An important point relates to the crystal orientation of the device, as indicated in Fig. 1. As is normal, we assume that the device is fabricated on a [100] surface of the Si crystal, and we then orient the channel so that the current will flow along the $<100>$ direction, as shown in the figure. In actual fact, the devices are typically oriented in the $<110>$ direction, which will actually simplify the formulation. Nevertheless, we illustrate the technique with the orientation shown in this figure. Then, all of the principle axes of the conduction band valleys line up with the coordinate axes. This is important so that the resulting quantization will split these ellipsoids into three pairs. Moreover, the choice of axes is most useful as the resulting Hamiltonian matrix will be diagonal. Using our orientation complicates the wave function, as we will see, but allows for simplicity in terms of the amount of memory needed to store the Hamiltonian and to construct the various scattering matrices (as well as the amount of computational time that is required). This is an important point in the computational structure of the simulation.

![Figure 1: Schematic showing the crystal orientation of the SOI MOSFET for the quantum simulation (the directions are not to scale). The overlay shows how the six conduction band valleys of Si line up with the coordinate axes. This is discussed further in the text.](image-url)
We can now write a total wave function which is composed of three major parts, one for each of the three sets of valleys. That is, we can write the wave function as a vector
\[
\Psi_T = \begin{bmatrix}
\Psi^{(x)} \\
\Psi^{(y)} \\
\Psi^{(z)}
\end{bmatrix},
\]
where the superscript refers to the coordinate axis along which the principal axis of the ellipsoid lies (the longitudinal mass direction). Thus, \(\Psi^{(x)}\) refers to the two ellipsoids oriented along the \(x\) axis in Fig. 3 (the \(<100>\) ellipsoids). Each of these three component wave functions is a complicated wave function on its own. Consider the Schrödinger equation for one of these sets of valleys (\(i\) corresponds to \(x\), \(y\), or \(z\) valleys):
\[
-\frac{\hbar^2}{2} \left( \frac{1}{m_x} \frac{d^2}{dx^2} + \frac{1}{m_y} \frac{d^2}{dy^2} + \frac{1}{m_z} \frac{d^2}{dz^2} \right) \Psi^{(i)}(x, y, z) + V(x, y, z) \Psi^{(i)} = E \Psi^{(i)}.
\]
Here, it is assumed that the mass is constant, in order to simplify the equations (for nonparabolic bands, the reciprocal mass enters between the partial derivatives). We have labeled the mass corresponding to the principle coordinate axes, and these take on the values of \(m_L\) and \(m_T\) as appropriate. We then choose to implement this on a finite difference grid with uniform spacing \(a\). Therefore, we replace the derivatives appearing in the discrete Schrödinger equation with finite difference representations of the derivatives. The Schrödinger equation then reads
\[
-t_x \left( \Psi_{i+1, j, k} + \Psi_{i-1, j, k} \right) - t_y \left( \Psi_{i, j+1, k} + \Psi_{i, j-1, k} \right) - t_z \left( \Psi_{i, j, k+1} + \Psi_{i, j, k-1} \right) + (V_{i, j, k} + 2t_x + 2t_y + 2t_z) \Psi_{i, j, k} = E \Psi_{i, j, k},
\]
where \(t_x, t_y\), and \(t_z\) are the hopping energies
\[
t_x = \frac{\hbar^2}{2m_x a^2},
\]
\[
t_y = \frac{\hbar^2}{2m_y a^2},
\]
\[
t_z = \frac{\hbar^2}{2m_z a^2}.
\]
Each hopping energy corresponds with a specific direction in the silicon crystal. The fact that we are now dealing with three sets of hopping energies is quite important.

There are other important points that relate to the hopping energy. The discretization of the Schrödinger equation introduces an artificial band structure, due to the periodicity that this discretization introduces. As a result, the band structure in any one direction has a sinusoidal variation with momentum eigenvalue (or mode index), and the total width of this band is \(4t\). Hence, if we are to properly simulate the real energy band behavior, which is quadratic in momentum, we need to keep the energies of interest below a value where the sinusoidal variation deviates significantly from the parabolic behavior desired. For practical purposes, this means that \(E_{\text{max}} < t\). The smallest value of \(t\) corresponds to the longitudinal mass, and if we desire energies of the order of the source-drain bias \(\sim 1\) V, then we must have \(a < 0.2\) nm. That is, we must take the grid size to be comparable to the Si lattice spacing!

We now seek to obtain the transfer matrices relating adjacent slices in our solution space. For this, we will develop the method in terms of slices, and follow a procedure first put forward by Usuki et al. \[17,18\] and used extensively by our group \[19\]. This is modified here by the two dimensions in the transverse plane. We first note that the transverse plane has \(N_y \times N_z\) grid points. Normally, this would produce a second-rank tensor (matrix) for the wave function, and it would propagate via a fourth-rank tensor. However, we can re-order the coefficients into a \(N_yN_z \times 1\) first-rank tensor (vector), so that the propagation is handled by a simpler matrix multiplication. Since the smaller dimension is the \(z\) direction, we use \(N_z\) for the expansion, and write the vector wave function as
Now, equation (3) can be rewritten as a matrix equation as, with \( s \) an index of the distance along the \( x \) direction, 
\[
H^{(i)} \Psi^{(i)}(s) - T_x^{(i)} \Psi^{(i)}(s-1) - T_x^{(i)} \Psi^{(i)}(s+1) = EP^{(i)}(s) .
\]  
(6)

Here, \( I \) is the unit matrix, \( E \) is the energy to be found from the eigenvalue equation, and
\[
H^{(i)} = \begin{bmatrix}
H^{(i)}_0(r) & \tilde{t}_z^{(i)} & \ldots & 0 \\
\tilde{t}_z^{(i)} & H^{(i)}_0(r) & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots \\
0 & \ldots & \tilde{t}_z^{(i)} & H^{(i)}_0(r)
\end{bmatrix},
\]  
(7a)

\[
T_x^{(i)} = \begin{bmatrix}
\tilde{t}_x^{(i)} & 0 & \ldots & 0 \\
0 & \tilde{t}_x^{(i)} & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots \\
0 & 0 & \ldots & \tilde{t}_x^{(i)}
\end{bmatrix}.
\]  
(7b)

The dimension of these two super-matrices is \( N_z \times N_z \), while the basic Hamiltonian terms of (7a) have dimension \( N_y \times N_y \), so that the total dimension of the above two matrices is \( N_y N_z \times N_y N_z \). In general, if we take \( k \) and \( j \) as indices along \( y \), and \( \eta \) and \( \nu \) as indices along \( z \), then
\[
\tilde{t}_z^{(i)}_{\eta \nu} = \delta_{\eta \nu} \tilde{t}_z^{(i)} ,
\]  
(8)

and
\[
H^{(i)}_0(r) = \begin{bmatrix}
V(s, 1, \eta) + W & \tilde{t}_y^{(i)} & \ldots & 0 \\
\tilde{t}_y^{(i)} & V(s, 2, \eta) + W & \ldots & 0 \\
\ldots & \ldots & \ldots & \ldots \\
0 & 0 & \tilde{t}_y^{(i)} & V(s, N_z, \eta) + W
\end{bmatrix}.
\]  
(9)

With this description of the matrices, the general procedure follows that laid out in previous descriptions \[13,20\]. One first solves the eigenvalue problem on slice 0 at the end of the source (away from the channel), which determines the propagating and evanescent modes for a given Fermi energy in this region. The wave function is thus written in a mode basis, but this is immediately transformed to the site basis, and one propagates from the drain end, using the general scattering matrix iteration
\[
C^{(i)}(s+1) = \begin{bmatrix}
C_1^{(i)}(s+1) \\
0 \\
C_2^{(i)}(s+1) \\
1
\end{bmatrix} = \begin{bmatrix}
0 & 1 \\
-1 & (T_x^{(i)})^{-1} \left[ EI - H^{(i)} \right]^{-1} \\
0 & 1
\end{bmatrix} \cdot \begin{bmatrix}
C_1^{(i)}(s) \\
0 \\
C_2^{(i)}(s) \\
1
\end{bmatrix}.
\]  
(10)

The dimension of these matrices is \( 2N_y N_z \times 2N_y N_z \), but the effective propagation is handled by submatrix computations, through the fact that the second row of this equation sets the iteration conditions
\[
C_2^{(i)}(s+1) = P_2^{(i)}(s) = \left[ -C_2^{(i)}(s) + (T_x^{(i)})^{-1} \left[ EI - H^{(i)} \right]^{-1} \right],
\]  
(11)

\[
C_1^{(i)}(s+1) = P_1^{(i)}(s) = P_2^{(i)}(s) C_1^{(i)}(s).
\]

At the source end, \( C_1(0) = 1 \), and \( C_2(0) = 0 \) are used as the initial conditions. These are now propagated to the \( N_x \) slice, which is the end of the active region, and then onto the \( N_x + 1 \) slice. At this point, the inverse of the mode-to-site transformation matrix is applied to bring the solution back to the mode representation, so that the transmission coefficients of each mode can be computed. These are then summed to give the total transmission.
and this is used in a version of equation (2.17) to compute the current through the device (there is no integration over the transverse modes, only over the longitudinal density of states and energy).

The heart of the recursion arises from eqns. (11). The wave function is represented by the excitation matrix $C^{(i)}_j$, where $j = 1, 2$ for propagation directions, and $i$ is the slice index. The recursion is obtained from one slice to the next may be written, in a more direct manner, as

$$C^{(i+1)}_2 = P^{(i)}_2 = \left(H_i - E - H^{(i+1)}_1 C^{(i)}_2 \right)^{-1} H^{(i+1)}_1 C^{(i)}_2,$$

and

$$C^{(i+1)}_1 = P^{(i)}_1 = C^{(i+1)}_2 C^{(i)}_1.$$

Here, $H^{(i+1)}_1$ is the hopping Hamiltonian that describes the standard finite-difference coupling from each element of one slice to the corresponding element of the next slice. The quantity $H_i$ is the bare Hamiltonian for the slice itself and contains “diagonal” terms for the on site terms as well as hopping terms to adjacent sites within the slice. In essence, (12) is a form of the Dyson’s equation for iterative solutions of Green’s function type behavior [21]. In fact, the two $C$ parameters are proper Green’s functions, but are normalized by the hopping terms to be dimensionless. In our simulation, the wave functions themselves are derived from these parameters. Here, $C^{(i)}_2$ plays the role of the on-site Green’s function $G(i,i)$, and $C^{(i)}_1$ plays the role of the on-site Green’s function $G(0,i)$ [22]. Hence, the hopping Hamiltonian already plays the role of the hopping self-energy term that couples one slice to the next. As a result, it will be relatively simple to introduce dissipation through a scattering-derived self-energy addition to the local slice Hamiltonian at each site.

### 3 Ballistic Transport

The application of this recursive scattering matrix, coupled to the self-consistent fields arising from the Poisson equation, has been used to study a number of applications, including open quantum dots [23,24,25] and ballistic transport in the Si nanowire transistor extensively [11,13,20,26,27]. Rather than repeat these results, we will instead look at ballistic transport in a high mobility InAs channel device. Here, the above equations are modified by the presence of only a single valley in the conduction band. The structure of the device with a 30 nm channel length is shown in Fig. 2. The exact device dimensions (multiples of the lattice constant) have been included in this simulation to aid in the inclusion of the discrete dopants. The thickness of the InAs layer is 9.09 nm. The source and drain of the device are $n$-type with a doping density of $6 \times 10^{18}$ cm$^{-3}$, while the channel of the device is considered to be $p$-type, but undoped. The gate material is assumed to be platinum and the gate oxide on each side is composed of 1 nm of hafnium oxide (HfO$_2$). Underneath the device, we have assumed a generic insulating substrate. Once the device geometry is defined, the InAs lattice is scanned, and the dopants are distributed according to the method presented in [28]. Following the distribution of the dopants, they are then mapped back onto the grid of the simulation mesh and the initial self-consistent Poisson solution is obtained. In this case, the full Coulomb potential of the dopants is incorporated. Then, the solution of Poisson’s equation for the local potential is no longer smoothly varying in the source and the drain of the device. The inclusion of discrete dopants causes the formation of potential variation in the source and drain.

In Fig. 3, we plot the $I_d-V_g$ and $I_d-V_d$ curves for the 30 nm InAs quantum wire MOSFET. We find that the sub-threshold slope averages about 69 mV/dec and the threshold voltage is 0.46±0.02 V. The $I_{on}/I_{off}$ ratio is $1.9 \times 10^{10}$. These numbers predict a transistor whose currents and charge control should far exceed those of its silicon counterpart. However, the threshold voltage for the devices is too large to be used in the present or the next generation technologies. This is due to the fact that in an InAs system, we have large quantization effects and gate work function dependence. This has both good and bad aspects. First, one normally expects InAs to have a Fermi level pinned roughly 0.4 eV into the conduction band. Thus, proper choice of the metal gate is essential to selecting the proper threshold voltage.
Figure 2: Device structure in the x-y plane for a 30 nm InAs channel MOS device.

Figure 3: (a) \(I_d-V_g\) curves for 2 different 30 nm devices (the dopant distribution is different for the two devices). (b) \(I_d-V_d\) curve for a 30 nm device with \(V_g = 0.4\) V.

We have not optimized the device structure and gate material in these results, but quantization in the channel helps to move the Fermi level out of the conduction band. However, in the present simulation, larger voltages must be applied to the system for conduction to occur. While the threshold voltage is too high in this system, the amount that the threshold voltage varies from device to device is minimal. From Fig. 3(a), two \(I_d-V_g\) curves are compared and these lie almost on top of one another. This invariance to threshold voltage shifts lies in the fact that we do not need to dope these devices to as high a concentration as is required in silicon. With fewer dopants in the source and drain of the device, there is a greatly reduced electron interaction among these dopants, and this results in smaller threshold voltage shifts. Another interesting point found in Fig. 3(a) is the sharp decrease in the output currents found at \(V_g = 0.6\) V. This is due to competing effects between the multiple scattering events undergone by the incident electrons [29] and interference from longitudinal states in the channel of the device [30]. Nevertheless, we can see that it would be advantageous to operate this device at voltages smaller than that for which this interference becomes significant. In Fig. 3(b), we show the \(I_d-V_d\) curve for the 30 nm device with a gate voltage of 0.4 V applied. Here we see a clear downside to the InAs transistor, in that the drain voltage never reaches saturation. This is an effect that can be attributed to the relatively small band gap of InAs. Initially, we see at low drain voltages (<0.15 V) the curve behaves rather normally, but as the drain voltage is increased we find that the current does not saturate.
In Fig. 4, we plot the potential at a depth of 7 nm into the device and with a gate voltage of 0.43 V and drain voltage of 0.6 V. Here, we observe that near the drain end of the device there is no appreciable drop in the potential that would normally be associated with the pinch-off condition observed in many MOSFET devices. Rather, nearly all of the drain potential is dropped in a high field region adjacent to the source. This probably arises as the channel is nominally undoped, and there is a barrier between the n-type source and the channel, which is normal. However, here the gate potential has not dramatically reduced this potential, and nearly all of the drain potential is required to move electrons from the source into the channel by accelerating them over the barrier. Indeed, the actual peak in the source-channel barrier lies slightly within the source region. Once injected into the channel, the carriers move almost ballistically with no further acceleration. The interference, mentioned above, between the source and drain boundaries results in a slight drop in density near the drain. This is accompanied by a slight peak in the local potential, which is of the order of $k_BT$ (the simulation is at room temperature).

4 The Scattering Self-Energy Terms

There have been many different approaches to including dissipation in quantum simulations. Statistical approaches introduce random phase fluctuations into the simulations [31,32]. The major drawback for such an approach is that a large sample space is required over which to average, and this entails a great many runs to have any valuable results. Another method is to add an imaginary term to the Hamiltonian which represents the phase breaking time of the electron in the system under consideration [33,34]. While this approach seems reasonable for a system that is in equilibrium, it is not clear that it is applicable when the system is driven out of equilibrium. The imaginary term is constant throughout the device, and therefore fails to consider the inhomogeneous density in the out of equilibrium system. This approach has also been questioned as not conserving current [35], but this fails to properly consider the entire dissipative current [36]. Dissipation may also be included through the use of Büttiker probes [37]. While this latter approach is an improvement over the use of a phase-breaking related term, in that it is current conserving, it suffers from the fact that an additional loop must be included to insure that the probes do not change the number of electrons in the system. In addition, it does it account for the spatial inhomogeneity of the density and the scattering and a fitting parameter must be used to calibrate the probes to the proper low field mobility. Thus, it is not very useful in real device simulations. Finally, a relaxation time approximation also has been used in approaches utilizing either the density matrix [38,39] or the Wigner function [40].

The details of the computational scheme which we employ were described above. The heart of the recursion arises from eqns. (12) and (13). The wave function is represented by the excitation matrix $C^{ij}$, where $j = 1,2$ for propagation directions, and $i$ is the slice index. Because of the form of these two equations, it becomes quite simple to modify the recursive formulation by the addition of an on-site self-energy in which
The self-energy $\Sigma$ has both real and imaginary parts, with the latter representing the dissipative interactions. In semiconductors, the scattering is weak, and is traditionally treated by first-order time-dependent perturbation theory, which yields the common Fermi golden rule for scattering rates. With such weak scattering, the real part of the self-energy can generally be ignored for the phonon interactions, and that part that arises from the carrier-carrier interactions is incorporated into the solutions of Poisson’s equation by a local-density approximation, which approximately accounts for the Hartree-Fock corrections [41]. In the many-body formulations of the self-energy, the latter is a two-site function in that it is written as [42]

$$\Sigma(r_1, r_2).$$  

(15)

In our case, where we are using transverse modes in the quantum wire, this may be rewritten as

$$\Sigma(i, j; i', j', x_1, x_2).$$  

(16)

Here, the scattering accounts for transitions from transverse mode $i,j$ at position $x_1$ to $i',j'$ at position $x_2$. Generally, one then makes a center-of-mass transformation [42,43]

$$X = \frac{x_1 + x_2}{2}, \quad \xi = x_1 - x_2,$$  

(17)

and then Fourier transforms on the difference variable to give

$$\int \Sigma(i, j; i', j', X, k_x) = \frac{1}{2\pi} \int d\xi \sum_{i,j} \Sigma(i, j; i', j', X, \xi).$$  

(18)

The center-of-mass position $X$ remains in the problem as the mode structure may change as one moves along the channel. At this point, the left-hand side of (18) is the self-energy computed by the normal scattering rates, such as is done in quantum wells and quantum wires previously [44,45]. However, these previous calculation usually used the Fermi golden rule, which is an evaluation of the bare self-energy in (18). In many-body approaches, one normally does not use the energy-conserving delta function that is the central part of the Fermi golden rule. Rather, this function is broadened into the spectral density, through the use of the self-consistent Born approximation [46]. In this way, off shell effects are taken into account through the broadening of the relationship between momentum and energy. In semiconductors, however, we have already noted that the scattering is weak. It has been pointed out that these off-shell corrections are only important in fast processes where we are interested in femtosecond response and their neglect introduces only slight errors for times large compared to the collision duration [47]. Moreover, the broadening of the delta function will not be apparent when we reverse the Fourier transform in (18) to get the $x$-axis variation, and do a mode-to-site unitary transformation to get the self-energy in the form necessary for the recursion. This is the subject of the rest of this section, where we discuss the different phonon processes. Hence, we begin by seeking the imaginary part of the self-energy, which is related to the scattering rate via

$$\text{Im} \left\{ \Sigma(i, j; i', j', X, k_x) \right\} = \frac{\hbar}{\tau} \frac{(\frac{1}{\tau})_{i,j}}{i', j'},$$  

(19)

and it is the latter scattering rate which we calculate. This result will be a function of the $x$-directed momentum (which is related, in turn, to the energy of the carrier) in the quantum wire. Finally, this scattering rate must be converted to the site representation with a unitary transformation

$$\Gamma_{ac} = \text{Im} \{\Sigma\} = U^\dagger \left( \frac{\hbar}{\tau} \right)_{i', j'} U,$$  

(20)

where $U$ is a unitary mode-to-site transformation matrix. The unitary matrix $U^\dagger$ results from the eigenvalue solutions in the transverse slice and are composed of the various eigenfunctions in the site basis. Hence, it represents a mode-to-slice transformation.

The Fermi golden rule scattering rate for acoustic phonons is treated in nearly all textbooks, and the only modification is to account for the transverse modes of the quantum wire. We do not need to redo the general details, as they are given elsewhere [48]. The scattering rate is given by

$$H_0 = H_i \rightarrow H = H_i + \Sigma_i,$$  

(14)
\[
\frac{1}{\tau_{ac}} \left( \frac{1}{\tau_{ac}} \right)_{i,j}^{f',f} = \frac{m_i^* D_{ac}^2 k_B T}{2\hbar^3 \rho_n^2} \left[ L_{i,j}^{f',f} \right] \frac{1}{\sqrt{k^2 + 2m_i^* \Delta_{ij}^{f',f}}} \cdot \theta \left( k^2 + \frac{2m_i^* \Delta_{ij}^{f',f}}{\hbar^2} \right).
\]  

Here, \( \theta \) is the Heavyside step function, and

\[
\Delta_{ij}^{f'} = E_{0,ij} - E_{0,i''j'},
\]

represents the energy difference between the two mode eigen-energies. Now, for our real space quantum transport approach, we need to reverse the Fourier transform in (18). That is, we use the inverse transform to real space from momentum space and obtain the final form for the acoustic deformation potential scattering rate. The Fourier integral is not straightforward, but easily doable. The result is

\[
\left( \frac{1}{\tau_{ac}} \right)_{i,j}^{f',f} (x-x') = \frac{m_i^* D_{ac}^2 k_B T}{2\hbar^3 \rho_n^2} \left[ L_{i,j}^{f',f} \right] \frac{1}{\sqrt{2\pi}} \int_\mathbb{R} \frac{1}{\sinh[\beta(x-x')] \sinh[\beta(x-x')] \sinh[\beta(x-x')]}.\]

The term in curly brackets is sharply peaked around \( x = x' \), which implies the scattering is local with regard to the individual slices in the recursion. There is coupling between the modes within a slice, but this local (to the slice) behavior is just the normal assumption in quasi-classical cases, where the scattering is assumed to be local in space [49]. Yet we need to know the total scattering rate within the slice, so this is achieved by integrating over \( x' \) in order to find the resultant scattering weight

\[
\left( \frac{1}{\tau_{ac}} \right)_{i,j}^{f',f} = \frac{m_i^* D_{ac}^2 k_B T}{4\hbar^3 \rho_n^2} \left[ L_{i,j}^{f',f} \right] \frac{1}{\pi}.\]

Scattering between the equivalent valleys of silicon is carried out by high energy optical modes. In general, there are two types of phonons, the \( f \) phonons between valleys on different axes in momentum space and \( g \) phonons between the two valleys along the same coordinate axis [50, 51]. There are several phonons which can contribute to each of these processes [52], but the treatment for each is the same. The differences from the acoustic mode treatment lie in the fact that the energy of the phonon can no longer be ignored in (22), and the phonon distribution function cannot be approximated by a equipartition approximation. Hence, we need to treat the emission and absorption of phonons differently, but the result is very similar. In general, the case for phonon absorption yields the result [analogous to (21)]

\[
\frac{1}{\tau_{ac}} \left( \frac{1}{\tau_{ac}} \right)_{i,j}^{f',f} = \frac{m_i^* D_{ac}^2 N_q}{8\pi^2 \hbar^2 \omega_k L} \left[ L_{i,j}^{f',f} \right] \frac{1}{\sqrt{k^2 + 2m_i^* \Delta_{ij}^{f',f}}} \cdot \theta \left( k^2 + \frac{2m_i^* \Delta_{ij}^{f',f}}{\hbar^2} \right).
\]

where now

\[
\Delta_{ij}^{f'} = E_{0,ij} - E_{0,i''j'} + \hbar \omega_q.
\]

In the case of the emission of a phonon, \( N_q \) in (25) is replaced by \( N_q + 1 \), and the positive sign in front of the phonon energy in (26) is replaced by a negative sign. However, there are further complications in the case of phonon emission which we will address below. For the absorption case, it is apparent that the form of the resulting equation is the same as that of the acoustic phonons, and we can use the same integration results for the inverse Fourier transform and for the site summation to yield the scattering strength. Basically, only the parameters in the leading coefficient are changed. Hence, we can immediately adapt (24) to give the result

\[
\frac{1}{\tau_{ac}} \left( \frac{1}{\tau_{ac}} \right)_{i,j}^{f',f} = \frac{m_i^* D_{ac}^2 N_q}{8\pi^2 \hbar^2 \omega_k L} \left[ L_{i,j}^{f',f} \right] \frac{\pi}{2}.\]

As mentioned above, the emission case has some complications. We must consider three different cases corresponding to different values for the energy difference from the initial subband to the final subband and the emitted phonon energy: \( \Delta_{ij}^{f'} < 0 \), \( \Delta_{ij}^{f'} > 0 \), and \( \Delta_{ij}^{f'} = 0 \). In each case, the Fourier transform is solved through the use of contour integrations. The resultant Fourier integral is
\[ \left( \frac{1}{\tau} \right)_{i,j}^{\prime, \prime} (x-x') = \frac{m^* D^2_{wq} (N_q + 1)}{8\hbar^2 \rho \omega_q L} |t_{i,j}^{\prime, \prime}| F_{\Delta} , \]  

(28)

where

\[
F_{\Delta} = \begin{cases} 
\frac{-\pi}{2} N_0 (x-x') , & \Delta < 0, \\
\frac{1}{\sqrt{2\pi}} K_0 (x-x') , & \Delta > 0, \\
\frac{1}{\sqrt{2\pi}} Ci (x-x') , & \Delta = 0. 
\end{cases}
\]  

(29)

In each case, the function is sharply peaked around \( x = x' \), which is to be expected as there is almost no momentum variation of the scattering rate. Then, we can get the final scattering strength by integrating over \( x' \), which gives the final result

\[
\left( \frac{1}{\tau} \right)_{i,j}^{\prime, \prime} \propto \begin{cases} 
1 , & \Delta < 0, \\
\frac{3}{2} , & \Delta > 0, \\
\frac{1}{\pi^{5/2} \sqrt{2}} , & \Delta = 0. 
\end{cases}
\]  

(30)

Elsewhere, we have used these formulae to compute the mobility of a large cross-section quantum wire, in which the size is essentially bulk. There, we have shown that the results give the normal impurity-dependent bulk mobility found experimentally in silicon. Hence, it is found that we can achieve some confidence with the present approach to treat scattering in quantum structured silicon via our approach. Thus, we have a simple physical approach to including separable scattering mechanisms as a self-energy in a real-space quantum simulation for a nanowire transistor. The inclusion of separable phonon scattering mechanisms is crucial in understanding the role that phonon processes play in electrical transport in quantum confined systems.

5 The Ballistic to Diffusive Crossover

We now turn our attention to the quantum wire transistors of interest. We examine a system where the quantum mechanical nature of the scattering matrices is needed to correctly characterize the system—the silicon quantum wire transistor. In a system such as this, the phonon scattering should have an effect on the transport, as the reduced area of the wire should give rise to an enhanced overlap integral in the scattering calculations. We consider the system shown in Fig. 5. The thickness of the silicon layer is 6.51 nm and the width of the channel is 8.15 nm. The length of the quantum wire that comprises the channel will be varied to study the resistance of the channel. Oxide barriers are placed on either side of the channel to simulate the appearance of a hard wall boundary that would be present in an actual experimental system. The source and drain of the device are 18.47 nm wide and 10.32 nm in length. The latter are discretely doped \( n \)-type with a doping concentration of \( 1 \times 10^{20} \text{ cm}^{-3} \), while the channel is undoped. The quantum wire that forms the channel of the device has metal gates on three sides to form a tri-gate type transistor [53]. The gate oxide thickness (SiO\(_2\)) on this device is 1 nm. All the simulations are carried out at 300 K.

In general, it has been suggested that the transport in short channel transistors is ballistic, and that once a carrier enters the channel it will continue to the drain, with no chance to scatter back to the source [10,12]. Already, however, Anantram has shown that scattering within the channel will cause second-order effects which do affect the terminal characteristics of the transistor [54]. The search for ballistic behavior has not been so successful, and one has had to estimate the degree to which ballistic transport appears in the device. Here, we want to vary the length of the channel in our nano-wire transistor to determine when the ballistic transport converts to diffusive transport (mobility-dominated transport). When the transport is ballistic the resistance of the channel will be determined by the inverse of the Landauer conductance, as
\[ R_{\text{ballistic}} = \left[ \frac{2e^2 N}{h} \right]^{-1}, \quad (31) \]

where \( N \) is the number of transverse modes propagating through the wire. Importantly, there is no dependence upon the length of the wire. On the other hand, when the resistance is determined by the mobility and carrier density, then the resistance is given by

\[ R_{\text{diffusive}} = \frac{1}{\sigma} \frac{L}{A} = \frac{1}{\text{ne} \mu} \frac{L}{A}, \quad (32) \]

where \( L \) is the channel length and \( A \) is the “cross-sectional area” of the inversion layer. We use the area here, rather than just the width of the two-dimensional layer, as we are dealing with a three-dimensional wire with full quantization in the transverse direction. Clearly, in this latter case, the resistance increases linearly with the channel length. It is the change from (31) to (32) that we seek in the present study.

Figure 5: Schematic of the quantum wire device design in the \( xy \) plane. The device thickness in the \( z \)-direction is 6.51 nm.

Figure 6: Resistance of a silicon quantum wire nanotransistor as the length of the quantum wire in the channel is varied with \( V_g = 0.5 \) V and \( V_d = 10 \) mV. After the initial ballistic region, we see the expected linear increase in the resistance as the channel length is increased. The dashed lines show the maximum and minimum values of the fit to eqn. (32) and the solid line shows the median value taken when the length is calculated through the center of the data.

In Fig. 6, we plot the resistance of our silicon nanowire transistor with the gate and drain voltages of the device held constant at \( V_g = 0.5 \) V and \( V_d = 10 \) mV, respectively. The gate voltage is sufficiently high that we are guaranteed to have an inversion layer formed in the channel of the device. One may wonder why we don’t use a higher drain voltage, but we must operate in the linear region of the device. At higher drain voltages, one begins to get depletion at the drain end, and (32) is no longer valid. In the figure, the resistance of the quantum wire is initially fairly constant when the length of the wire is held to less that 2 nm. When the length of the wire then exceeds roughly 2 nm, we find that the resistance of the device begins to increase linearly as is characteristic for devices in the diffusive transport regime. However, there are oscillations which appear. These can be found not only in the conductance and output current of the quantum wire, but in the density as well. These oscillations arise from the strong quantum interference between the source and drain ends of the device. This oscillatory result of the quantum interference is generally seen in the conductance both experimentally [55] and theoretically [56,57]. Nevertheless, we can estimate the crossover by seeking the crossing point of a line down the center of the linearly increasing part and a line horizontal in the center of the ballistic curve, as shown in the figure (solid curve). The range of possible values is set by the dashed limiting curves. Projection of these
crossings onto the horizontal axis gives us the appropriate dimensions. From this plot, we estimate the ballistic to diffusive crossover occurs at 1.42 ± 0.03 nm.

The reduction of the ballistic to diffusive crossover is summarized in Fig. 7, where we plot the crossover length for three different drain voltages. There is a clear linear trend, within the error associated with the measurements, in the reduction of the ballistic to diffusive crossover by the enhanced phonon scattering in the drain of the silicon nanowire transistor. The more energy the carriers gain as they traverse the length of the nanowire transistor, the more scattering they see. Thus, there is a linear drop in the ballistic to diffusive crossover length with increases in drain voltage, as discussed above.

It is clear that the transition length, where the ballistic transport goes into diffusive transport, is considerably shorter than any previous estimate of this length. However, it is important to note that the channel of the transistor does not act in isolation from other parts of the transistor. Indeed, the key part of the transistor is the barrier between the source and the channel, and this lies partially in the source, as we have seen above (Fig. 4). Thus, we cannot talk about transport in the channel as if the carriers just begin to move at the source end of the channel. Rather, they already have been undergoing some diffusive transport in the source region, and this affects the apparent mean free path in the channel. Moreover, the carriers enter the channel over the source-channel barrier, and this barrier is more than \(k_BT\) high. This means that the carriers entering the channel come from the more energetic part of the distribution in the source. That is, they are relatively high energy carriers, and these will exhibit more scattering than if they were low energy. While they may have a small velocity just as they clear the barrier, they will be accelerated by that part of the potential barrier lying in the channel region back to a higher velocity. Thus, we cannot estimate the transition region length, shown in Fig. 7, from arguments based upon the mean free path in bulk material, where the entire carrier distribution function participates in the transport.

### 6 Summary

In this paper, we have reviewed a recursive scattering matrix approach to simulation of semiconductor devices. This approach fully incorporates Dyson’s equation and is similar to recursive Green’s function approaches, but is easier to utilize for determining carrier densities needed to close the self-consistent loop via Poisson’s equation. Then, we have presented a simple physical approach to including separable scattering mechanisms as a self-energy in this real-space quantum simulation, and applied this to a silicon nanowire transistor. The inclusion of separable phonon scattering mechanisms is crucial in understanding the role that phonon processes play in electrical transport in quantum confined systems. We discussed the scattering contributions for acoustic deformation potential and intervalley phonon scattering with both \(f\) and \(g\) type processes included, and verified that our model captures the experimentally-verified mobility for silicon. When the model is applied to a gated quantum wire structure, we find the largest ballistic to diffusive crossover length in the device occurs with the smallest bias. We find this length to be about 1.4 nm. Increasing the drain bias decreases this ballistic to diffusive crossover length, presumably due to considerably more carrier heating in the channel. These results demonstrate that the limit for ballistic device performance is much less than previously thought.
Parts of this work were supported by the Office of Naval Research and by Intel Corporation.

References

1. G. E. Moore, *Electronics Mag.* 38(8), 114 (April 1965).
2. See, e.g., R. Chau et al., *IEEE Trans. Nanotechnol.* 4, 153 (2005).
3. J. Kedzierski, J. Bokor, and E. Anderson, *J. Vac. Sci. Tech. B* 17, 3244 (1999).
4. B. Doris et al., *Intern. Electron Dev. Mtg. Tech. Dig.* (IEEE Press, New York, 2001) 937.
5. B. Yu et al., *Intern. Electron Dev. Mtg. Tech. Dig.* (IEEE Press, New York, 2002) 267.
6. J. R. Barker and D. K. Ferry, *Sol.-State Electron.* 23, 531 (1980).
7. M. Fischetti, *J. Appl. Phys.* 83, 270 (1998).
8. K. Likharev, in *Advanced Semiconductor and Organic Nano-Techniques*, Ed. by H. Morkoc (Academic Press, New York, 2002).
9. R. Venugopal et al., *J. Appl. Phys.* 92, 3730 (2002).
10. K. Natori, *J. Appl. Phys.* 76, 4879 (1994).
11. M. J. Gilbert et al., *J. Comp. Electron.* 2, 329 (2003).
12. M. Lundstrom, *IEEE Electron Dev. Lett.* 18, 361 (1997).
13. M. J. Gilbert and D. K. Ferry, *J. Appl. Phys.* 95, 7954 (2004).
14. F. G. Pikus and K. K. Likharev, *Appl Phys Lett.* 71, 3661 (1997).
15. S. Datta, *Superlatt Microstruct.* 28, 253 (2000).
16. J. Knoch, B. Lengeler and J. Appenzeller, *IEEE Trans Elec Dev* 49, 1212 (2002).
17. T. Usuki et al., *Phys Rev B* 50, 7615 (1994).
18. T. Usuki et al., *Phys Rev B* 52, 8244 (1995).
19. R. Akis, D. K. Ferry, and J. P. Bird, *Phys Rev B* 54, 17705 (1996).
20. D. K. Ferry et al., in *Silicon Nanoelectronics*, Ed. by S. Oda and D. K. Ferry (Dekker, New York, in press).
21. A. L. Fetter and J. D. Walecka, *Quantum Theory of Many-Particle Systems* (McGraw-Hill, New York, 1971).
22. D. K. Ferry and S. M. Goodnick, *Transport in Nanostructures* (Cambridge University Press, UK, 1997).
23. R. Akis et al., in *Electron Transport in Quantum Dots*, Ed. by J. P. Bird (Kluwer Academic Publishers, Boston, 2003) pp. 209-276.
24. D. K. Ferry, R. Akis, and J. P. Bird, *Phys Rev. Lett.* 93, 026803 (2004).
25. D. K. Ferry, R. Akis, and J. P. Bird, *J. Phys.: Cond. Matt.* 17, S1017 (2005).
26. M. J. Gilbert and D. K. Ferry, *J. Vac. Sci. Technol. B* 22, 2039 (2004).
27. M. J. Gilbert and D. K. Ferry, *IEEE Trans. Nanotechnol.* 4, 355 (2005).
28. M. J. Gilbert and D. K. Ferry, *Superlatt. Microstruct.* 34, 277 (2003).
29. V.A. Sablikov, S.V. Polyakov, and M. Büttiker, *Phys. Rev. B* 61, 13763 (2000).
30. G. Kirczenow, *Phys. Rev. B* 39, 10452 (1989).
31. H. Benisty, *Phys. Rev. B* 51, 13281 (1995).
32. M.G. Pala and G. Iannaccone, *Phys. Rev. Lett.* 93, 256803 (2004).
33. G. Neofotistos, R. Lake, and S. Datta, *Phys. Rev. B* 43, 2442 (1991).
34. R. Akis, J.P. Bird, and D.K. Ferry, *J. Cond. Matt.* 8, L667 (1996).
35. W. R. Frensley, *Rev. Mod. Phys.* 62, 745 (1990).
36. R. Gebauer and R. Car, *Phys. Rev. B* 70, 125324 (2004).
37. M. Büttiker, *Phys. Rev. Lett.* 57, 1761 (1986).
38. H. Dekker, *Phys. Rev. A* 16, 2126 (1977).
39. H. L. Grubin et al., *Semicond. Sci. Technol.* 9, 855 (1994).
40. N. C. Kluksdahl et al., *Phys. Rev. B* 39, 7720 (1989).
41. Y. Wang et al., *Phys. Rev. B* 52, 2738 (1995).
42. L. P. Kadanoff and G. Baym, *Quantum Statistical Mechanics* (Benjamin/Cummings, Reading, MA, 1962).
43. J. Rammer, *Quantum Transport Theory* (Perseus, Reading, MA, 1998).
44. R. Kotlyar et al., *Appl. Phys. Lett.* 84, 5270 (2004).
45. E.G. Emberly and G. Kirczenow, *Phys. Rev. B* 61, 5740 (2000).
See, e.g., R. Lake et al., J. Appl. Phys. 81, 7845 (1997).
47. V. Špička, B. Velický, and A. Kalvová, Physica E, in press.
48. M. J. Gilbert, R. Akis, and D. K. Ferry, submitted for publication.
49. J. R. Barker, in Physics of Nonlinear Transport in Semiconductors, Ed. by D. K. Ferry, J. R. Barker, and C. Jacoboni (Plenum Press, New York, 1980) 126.
50. D. Long, Phys. Rev. 120, 2024 (1960).
51. D. K. Ferry, Phys. Rev. B 12, 2361 (1975).
52. See, e.g., D. K. Ferry, Semiconductors (Macmillan, New York, 1991).
53. H. S. Doyle et al., IEEE Electron Dev. Lett. 24, 263 (2003).
54. A. Svizhenko and M. P. Anantram, IEEE Trans. Electron Dev. 50, 1459 (2003).
55. A.T. Tilke et al., Phys. Rev. B 68, 075311 (2003).
56. V.A. Sablikov, S.V. Polyakov, and M. Büttiker, Phys. Rev. B 61, 13763 (2000).
57. G. Kirczenow, Phys. Rev. B 39, 10452 (1989).