Techniques for Modeling Ultrasmall Quantum Devices

D. K. Ferry[1] and H. L. Grubin[2]

[1]Center for Solid State Electronics Research
Arizona State University
Tempe, AZ 85287-6206

[2]Scientific Research Associates, Inc.
P. O. Box 1058
Glastonbury, CT 06033

Abstract
As semiconductor devices continue to shrink in size, quantum effects are beginning to become important in their operation. In this paper, we discuss some approaches for including the quantum effects in device modeling.

I. Introduction

The transport of carriers in semiconductors and in ultra-small semiconductor devices has long been a subject of much interest, not only for material evaluation, but also in the realm of device modeling and, more importantly, as an illuminating tool for delving into the physics governing the interaction of electrons (or holes) with their environment. The scaling of ULSI device dimensions to future chips indicates that we will eventually see devices with gate lengths at the 0.05 μm level. Very few laboratories have produced even working research transistors with gate lengths on the 50 nm scale and little is understood about the limitations (from the physics) that will determine whether or not these devices are practical. On the other hand, when devices of 30 nm (or less) gate lengths are made, it is found that their performance is different from that of current FETs. Research devices with gate lengths of 25-80 nm clearly show that tunneling through the gate depletion region is a major contributor (if not the dominant contributor) to current, and gate control is much reduced due to this effect [1-3]. Even in MOSFETs, quantization is found to occur in the channel, which affects the overall performance of the devices [4]. In consequence, it appears that more detailed modeling of quantum effects needs to be included in device modeling for future ultrasmall devices.

There are several approaches that have been used to model quantum effects in semiconductor devices (of varying levels of device complexity). In this paper, we will try to give a short overview of some of these approaches, and indicate how they are similar and how they differ. In the next section, we discuss how quantum modeling differs from semi-classical modeling. We then turn to a description of the various quantum "distribution" functions, discuss their equations of motion, and the levels of complexity. In each case, simple examples are described where the approach has been used with some effectiveness.

II. How does Quantum Modeling Differ from Classical Modeling?

Generally, modeling of quantum phenomena is more complicated than modeling of classical and/or semi-classical phenomena. For instance, the energy-conserving delta function used in computing scattering rates with the Fermi golden rule is no longer valid, as energy and momentum become separate dynamical variables. Thus, we are forced to add a method of computing the *spectral density*, which relates the energy to the momentum, in addition to having to compute non-equilibrium distribution functions (or various moments of these). To

be sure, in some approaches this is finessed by using single-time functions, such as the density matrix and the Wigner distribution function, which essentially integrate out the spectral function. But, this is accompanied by the full non-local nature of the potential interactions becoming explicit in the dynamical variables; i.e., the potential becomes a two-point function. Let us consider further how this nonlocality arises. Consider a simple potential barrier $V(x) = V_0 u(-x)$, where $u(x)$ is the Heavyside step function. We assume that there is some density existing in the region $x > 0$, and the question is how the density varies near the barrier, a quite typical problem in introductory quantum mechanics, except here we have a statistical mixed state to describe. In Fig. 1, we show the Wigner distribution function for this case (for parameters appropriate to GaAs, with $n = 2 \times 10^{17}$ cm^{-3}) for $V_0 \rightarrow \infty$. We note that far from the barrier, the distribution approaches the classical Maxwellian form, but near the barrier, it differs greatly. The repulsion from the barrier is required by the vanishing of the wave function at the barrier, but the first peak away from the barrier (in the wave function) occurs closer to the barrier for higher momentum states. This leads to much of the complication evident in the figure. The overshoot occurs to accomodate the need for total charge neutrality. Classically, in the absence of self-consistency, the density would be uniform up to the barrier, and the differences are the result of the quantum mechanics. This variation exists over a distance of the order of several thermal de Broglie wavelengths, which provides a spatial scale length. In GaAs, at 300 K, this is about 5 nm for electrons, and of course increases with the inverse square root of the temperature as the thermal de Broglie wavelength is given by $\lambda_D = (\hbar^2/2mk_BT)^{1/2}$. Thus, nonlocal variations can be expected over a range of 10-20 nm even at room temperature!

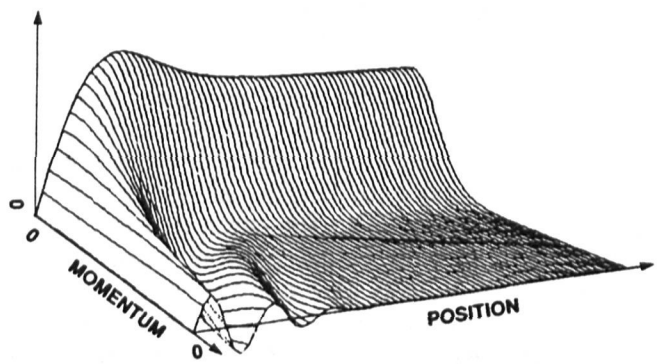


Figure 1. The Wigner distribution function for an infinite barrier, in arbitrary units [5].

It is clear that the density no longer varies simply as $\exp(-\beta V)$, where β is the inverse temperature, and that modifications to the statistical mechanics need to be made. The development of quantum corrections to statistical thermodynamics, especially in equilibrium, has a rich and old history. Unfortunately, there is no consensus as to the form of the quantum potential correction to this simple exponential. If we could find such a correction, it could be utilized in the semi-classical hydrodynamic equations, and most further quantum complications ignored. The various forms of the quantum potential, for use in classical hydrodynamic equations, has recently been reviewed [6]. One form that has been used [7] introduces a *quantum pressure* term as a modification of the electron temperature, through

$$\frac{3}{2}k_B T_{eff} = \frac{3}{2}k_B T_e - \frac{\hbar^2}{8m^*}\nabla^2 \ln(n)\,, \tag{1}$$

although other work has reduced the last term by a factor of 3 [8,9]. The form of the last term in (1) was originally derived by us and is often termed the Wigner potential [10]. Although the results obtained using this model are in agreement with the intuitive expectations, it should be noted that the correction term is an average and does not have the

momentum dependence expected from Fig. 1. A more recent derivation overcomes this limitation [6], but has not been tested in actual simulations as yet.

Another problem with the use of quantum approaches to device modeling is that most quantum discussions, especially those of quantum transport, tend to revolve around *closed* systems, whereas most devices are *open* systems. In treating such open systems quantum mechanically, it is quite difficult to properly define the reservoir (thermal equilibrium) regions, as well as the contact regions between the reservoirs and the active device region. Because of the nonlocal nature of the quantum system, errors in defining the contact region will propagate throughout the device, often leading to spurious results.

III. Approaches to Quantum Distributions

Why don't we just solve the Schrödinger equation for a given potential distribution, and then weight various solutions with a Fermi-Dirac distribution? This approach actually works well in equilibrium situations, or when we know in detail the properties of all contact regions and the exact potential structure within the active area. However, finding the steady-state solutions with the above approach often entails more work than using one of the techniques for directly finding a quantum distribution self-consistently within the entire device domain. A more important reason for not a simple Schrödinger equation solver (plus Fermi-Dirac weighting factors) for a distribution is that the actual distribution in the active region is a very non-equilibrium distribution which we must find as part of the modeling problem. But, the approaches discussed below are developed from the Schrödinger equation; we are solving a more general function, which incorporates the solution of this equation for an entire functional set. In addition, since the Schrödinger equation just defines a wave function, which is one part of the density (or a representation for an electron), it is quite difficult to incorporate dissipation through scattering mechanisms. Nevertheless, the starting point for all of our approaches lies in a mixed state wave function $\Psi(\mathbf{x},t)$, which is taken to be a field operator describing the degree of excitation of the various states of the system (this is one method of conveniently describing the mixed state of the system). Depending upon the Hamiltonian, this wave function can be a one-electron wave function or a many-body wave function.

A. *The Density Matrix*

The density matrix is formed from the composite of two such wave functions described above. It may be written as

$$\rho(\mathbf{x},\mathbf{x}',t) = \Psi(\mathbf{x},t)\Psi^{+}(\mathbf{x}',t)\,, \tag{2}$$

where the "+" symbol on the second wave function indicates the Hermitian adjoint function. This is an *equal time* function and describes the correlation between events at positions **x** and **x'**. Obviously, $\langle\rho(\mathbf{x},\mathbf{x},t)\rangle = \langle\Psi(\mathbf{x},t)\Psi^{+}(\mathbf{x},t)\rangle = n(\mathbf{x},t)$ defines the local density of particles. Here, we have taken an expectation of the density *operator*, since the definition in (2) is obviously that of an operator. The equation of motion arises from the Liouville equation. It may be written as (in the absence of dissipative processes)

$$i\hbar\frac{\partial\rho}{\partial t} = -\frac{\hbar^2}{m}\frac{\partial^2\rho}{\partial\mathbf{R}\partial\mathbf{s}} + 2\left[\sinh\left(\frac{1}{2}\mathbf{s}\cdot\nabla\right)V\right]\rho\,, \tag{3}$$

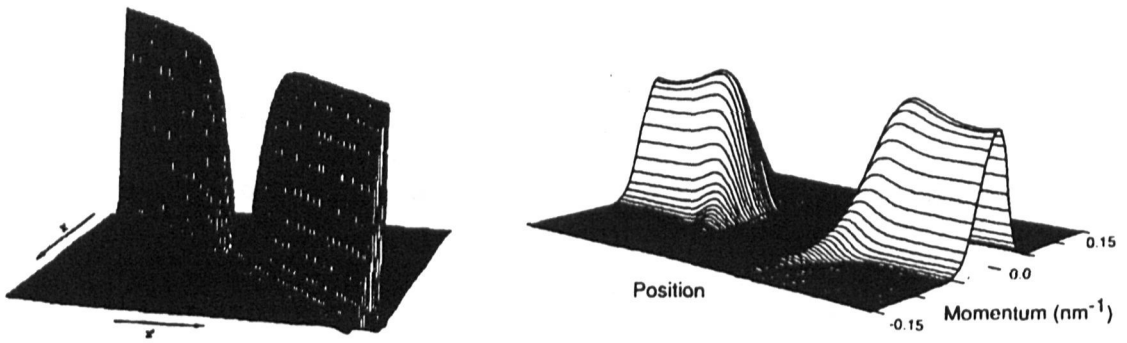
where the last term, in the square brackets, is a short-hand notation for

$$[\sinh(\tfrac{1}{2}\mathbf{s}\cdot\nabla)\,V] = \tfrac{1}{2}[V(\mathbf{R}+\tfrac{\mathbf{s}}{2}) - V(\mathbf{R}-\tfrac{\mathbf{s}}{2})]\ , \tag{4}$$

and we have introduced the coordinate transformations

$$\mathbf{R} = \tfrac{1}{2}(\mathbf{x}+\mathbf{x}') \quad , \quad \mathbf{s} = \mathbf{x} - \mathbf{x}' \quad . \tag{5}$$

The density matrix has been used directly to study a number of devices. In Fig. 2(a), we show the density matrix for a double-barrier resonant tunneling diode (DBRTD) in equilibrium. For comparison, we also show the Wigner distribution function (described below) for this structure in Fig. 2(b). Both calculations are for barriers 0.3 eV high, 5 nm wide, and separated by 5 nm. Both are within lightly doped regions adjacent to the barriers, 5 nm wide for the Wigner function and 7.5 nm wide for the density matrix. In both calculations, the nominal density was 10^{18} cm^{-3}, and Fermi statistics were applied at the boundaries. For the density matrix in Fig. 2(a), this is represented by a dampled oscillation in the nonlocal coordinate, whose period decreases as the density increases. Density is obtained from the diagonal component, which for the DBRTD, shows a small buildup of charge within the quantum well. The peak of this charge is approximately 2×10^{16} cm^{-3}.



(a) (b)

Figure 2. The double-barrier resonant tunneling diode in equilibrium: (a) the density matrix [10], (b) the Wigner distribution function [13].

For the density matrix under dynamic current flow conditions, dissipation is incorporated and serves to couple the real and imaginary parts of the density matrix. Current boundary conditions are represented by a *displaced distribution function*, similar to a displaced Fermi distribution. Dissipation is introduced as a phenomenological scattering potential whose diagonal components have the properties of a dynamic quasi-Fermi level [11]. At low values of bias, the scattering potential has the form $(\mathbf{x}-\mathbf{x}')[J/\tau\rho(\mathbf{x},\mathbf{x})]\rho(\mathbf{x},\mathbf{x}')$, which is similar to that discussed in [12]. Here J and τ represent current and scattering time, respectively. This form of scattering conserves the total number of particles. For simple barriers, the current-voltage characteristics display the expected exponential dependence on potential energy, with accumulation at the emitter side of the barrier and depletion on the collector side.

The computational procedures are described in detail in [12], and briefly may be described as re-expressing (3) as a coupled first-order system of equations, and seeking solutions along characteristic directions for the coupled equations. All of the calculations incorporate equally spaced grids, and a coupled Poisson solver. For simple barriers, the current-voltage characteristics display the expected exponential dependence on potential energy, with accumulation at the emitter side of the barrier and depletion on the collector side. This is shown in Fig. 3 for a 0.3 eVbarrier, 15 nm thick, which is embedded in a 30 nm lightly doped region (GaAs). In Fig. 3(a), the real part of the density matrix, which is symmetric about the diagonal and shows charge accumulation on the emitter side. The imaginary part is shown in

Fig. 3(b), and is anti-symmetric, as its derivative along the nonlocal direction yields the current. We illustrate the computed current-voltage relationship in Fig. 3(c).

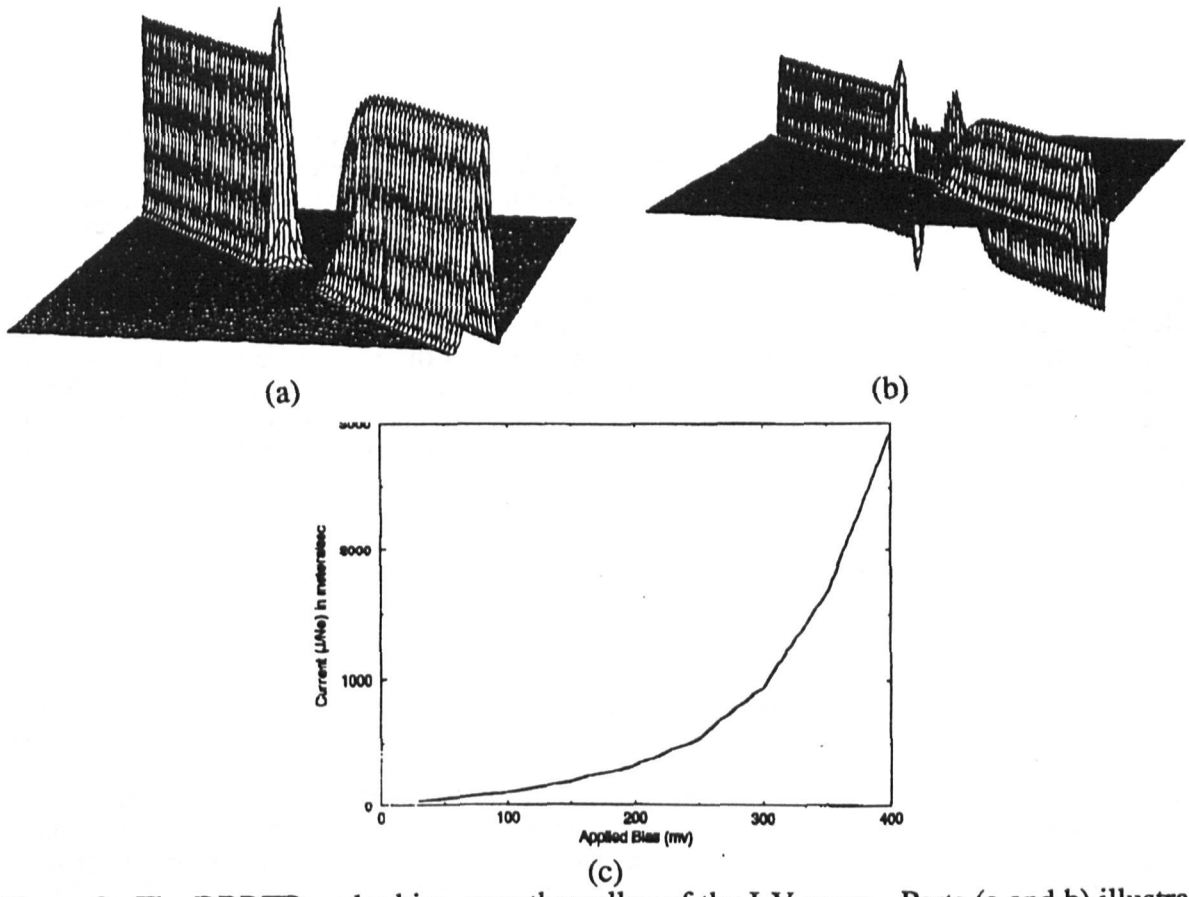


Figure 3. The DBRTD under bias, near the valley of the I-V curve. Parts (a and b) illustrate the real and imaginary parts of the density matrix, respectively, while the I-V curve itself is shown in (c).

B. The Wigner Distribution Function

The Wigner distribution becomes important when the physical problem is one that is better understood in terms of a phase-space distribution, and the carrier distribution function in an inhomogeneous device is one such problem. This phase-space distribution is not easily represented by the density matrix itself, but the Wigner distribution attempts to present an analogy between quantum and classical phase space for statistical mechanics. Since the statistical picture in phase space is well understood, indeed uses the Boltzmann equation for classical mechanics, transforming to a similar picture in quantum statistical mechanics allows the physical picture of a problem to be better understood. Unfortunately, position and momentum do not commute in quantum mechanics, and the two cannot be measured simultaneously to any great accuracy in phase space. This appears in the Wigner picture by regions of the phase space in which the distribution is negative in value. When the distribution is integrated over all space, the probability density in momentum space is recovered, and this quantity is positive definite. When the distribution is integrated over all momentum, the probability density in real space is recovered, and this quantity is also positive definite. In fact, if the Wigner distribution function is coarse-grain averaged over a region of phase space corresponding to a six-dimensional volume element whose size is set

by the uncertainty principle, the result is again a positive-definite "averaged" function. The Wigner distribution function is defined from the density matrix through the Fourier transform, often called a Weyl transform,

$$F_W(\mathbf{R},\mathbf{p},t) = \frac{1}{(2\pi\hbar)^3}\int d^3s\, e^{i\mathbf{p}\cdot\mathbf{s}/\hbar} \langle \rho(\mathbf{x}+\frac{\mathbf{s}}{2},\mathbf{x}-\frac{\mathbf{s}}{2})\rangle \ , \tag{6}$$

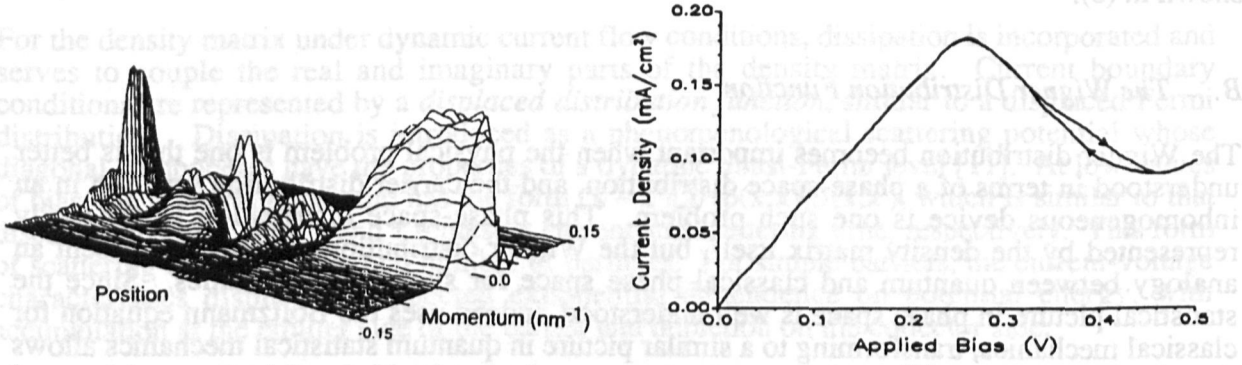
which in a sense is a Fourier transform on the variable that measures the distance from the diagonal in the density matrix $\rho(\mathbf{x}, \mathbf{x}')$. Since the Wigner distribution is a c-number, an expectation has been indicated in (6). This transformation accentuates the correlation that exists in the wave functions separated in position (if the correlation exists). The Wigner function builds in the correlations between different positions that are inherent in the off-diagonal elements of the density matrix. The Wigner function is evaluated at position **R**, but the density matrix terms that are used in the Fourier transform are those at the two positions $\mathbf{R} \pm (\mathbf{s}/2)$. The wave function may actually vanish at **R**, but the Wigner function will have a nonzero value in these areas in which the wave function vanishes, and the values in these such regions are measures of the correlation between the two endpoints on the vector **s**. The equation of motion for the Wigner distribution function is given by (again, in the absence of dissipation)

$$\frac{\partial F_W}{\partial t} - \frac{1}{m}\mathbf{p}\cdot\nabla F_W = \frac{1}{\pi\hbar}\int d^3p' M(\mathbf{R},\mathbf{p}') F_W(\mathbf{R},\mathbf{p}+\mathbf{p}',t) \ , \tag{7}$$

where

$$M(\mathbf{R},\mathbf{p}') = \int d^3s\, e^{i\mathbf{p}\cdot\mathbf{s}/\hbar}\,[\sinh(\tfrac{1}{2}\mathbf{s}\cdot\nabla)\, V] \ . \tag{8}$$

Equation (7) is quite similar to the streaming terms of the Boltzmann equation, especially if the lowest order term in the expansion of the potential is used. The Wigner distribution has also been used to model the DBRTD [13], and the results are shown in Fig. 4, again for the use of a relaxation time approximation for the dissipation, and for a bias near the valley of the I-V relation.



(a) (b)

Figure 4. (a) Steady-state Wigner distribution at the valley of the I-V curve. (b) The I-V curve for a DBRTD.

The Wigner function in this simulation shows a depletion region in the cathode area, which arises from a contact potential drop and the tendency to form a bound state in this area. It is largely eliminated if a lightly-doped region is introduced adjacent to the barrier layers [13]. Such contact potential drops are typical of most open systems, whether classical or quantum, and are well-known in the Gunn effect device literature [14]. The depletion was also found

to go away with greater amounts of scattering in detailed studies of the role of scattering on the DBRTD by Frensley [15]. Generally, the cathode "barriers" will develop when there is a mismatch between the injection characteristics of the cathode reservoir and the dissipative nature of the active device region. Added dissipation or additional resistance in the active layer (through the lightly-doped regions) reduces the mismatch, and thereby reduces the depletion in the cathode-drop region.

C. Real-Time Green's Functions

Both the density matrix and the Wigner distribution function are *equal-time* functions, and are functions of only seven variables—either two vector postions and the time or the vector position, vector momentum, and the time, respectively. The energy does not enter into either description. This resulted from the definition that was used in (2), but there was no real requirement to have defined it in this manner. We could as easily have written the two wave functions at different times, and it is possible to define another function, which is a function of two times, through

$$G^{<}(\mathbf{x},t,\mathbf{x}',t') = -i\langle\Psi^{+}(\mathbf{x}',t')\Psi(\mathbf{x},t)\rangle \ . \tag{9}$$

The equal time version of (9) is obviously related to the density matrix itself. There are in fact a group of real-time Green's functions, arising from the different ways in which the wave functions can be combined and temporally ordered. Any simulation problem must solve for the independent members of this group (four in number) [16]. The particular Green's function in (9) is the "less than" function, and is closely related to the density in the other distributions above. Introducing the change of variables (5), and the equivalent for the average time T, and the difference time τ, we can introduce the energy ε ($=\hbar\omega$) through an additional Fourier transform, as

$$G^{<}(\mathbf{R},\mathbf{p},\omega,T) = i\int e^{i\mathbf{p}\cdot\mathbf{s}/\hbar - i\omega\tau}\langle\Psi(\mathbf{R}+\frac{\mathbf{s}}{2},T+\frac{\tau}{2})\Psi^{+}(\mathbf{R}-\frac{\mathbf{s}}{2},T-\frac{\tau}{2})\rangle\, d^3s d\tau \ . \tag{10}$$

It is clear that the $\tau = 0$ limit of (10) will lead to our Wigner distribution function, and

$$F_W(\mathbf{R},\mathbf{p},T) = \int d\omega\, G^{<}(\mathbf{R},\mathbf{p},\omega,T) \ . \tag{11}$$

Much more information is contained within the Green's function formalism, since we can now investigate in detail the spectral density itself, which relates the energy to the momentum. However, very little has done with these Green's functions in actual device modeling. However, they have been used to study high-field transport in homogeneous systems [16], and simplified versions have been used to study the DBRTD [17] for a non-self-consistent potential and weak scattering from phonons (but introduced without resorting to a relaxation time approximation). Nevertheless, the results are suggestive and indicate that quite detailed quantum device modeling can be carried out with the real-time Green's functions.

IV. Conclusions

Over the past few years, many groups have begun to explore quantum methodologies for modeling real semiconductor devices (at real temperatures). Most of the various approaches are closely related to each other, and offer different ways of approaching any given problem. While none of the techniques has become well developed, all of the ones discussed here have

been shown to lead to useful results and have given added insight into the problem. We are now passing the point at which we are trying to understand the methodology, and are in a position where we can now confidently use the methods to study device physics. Even so, many problems of understanding, particularly in the quantum statistical mechanics interpretations still remain, and will lead to many interesting lines of inquiry in the near future.

Acknowledgements

DKF would like to acknowledge support from the Office of Naval Research and the Army Research Office, while HLG would like to acknowledge support from the Office of Naval Research and the Air Force Office of Scientific Research.

References

[1] J. Han, D. K. Ferry, and P. Newman, IEEE Electron Dev. Lett. **11**, 209 (1990).
[2] A. Ishibashi, K. Funato, and Y. Mori, Jpn. J. Appl. Phys. **27**, L2382 (1988).
[3] S. Y. Chou, D. R. Allee, R. F. Pease, and J. S. Harris, Jr., Proc. IEEE **79**, 1131 (1991).
[4] T. Ando, F. Stern, and A. B. Fowler, Rev. Mod. Phys. **54**, 437 (1982).
[5] A. M. Kriman, N. C. Kluksdahl, and D. K. Ferry, Phys. Rev. B **36**, 5953 (1987).
[6] D. K. Ferry and J. R. Zhou, Phys. Rev. B, in press.
[7] J.-R. Zhou and D. K. Ferry, IEEE Trans. Electron Dev. **39**, 473 (1992); **39**, 1793 (1992).
[8] H. L. Grubin and J. P. Kreskovsky, Sol.-State Electron. **32**, 1071 (1989).
[9] M. A. Ancona and G. J. Iafrate, Phys. Rev. B **39**, 9536 (1989).
[10] G. J. Iafrate, H. L. Grubin, and D. K. Ferry, J. Physique (Colloq. C-10) **42**, 307 (1981).
[11] T. R. Govindan and H. L. Grubin, to be published.
[12] H. L. Grubin, T. R. Govindan, J. P. Kreskovsky, and M. A. Stroscio, Sol.-State Electron., in press.
[13] N. C. Kluksdahl, A. M. Kriman, D. K. Ferry, and C. Ringhofer, Phys. Rev. B **39**, 7720 (1989).
[14] M. P. Shaw, H. L. Grubin, and P. R. Solomon, *The Gunn-Hilsum Effect* (Academic Press, New York, 1979).
[15] W. R. Frensley, in *Computational Electronics*, Ed. by K. Hess, J. P. Leburton, and U. Ravaioli (Kluwer, Norwell, MA, 1991) 195.
[16] For a review, see the chapters by G. Mahan and A.-P. Jauho in *Quantum Transport in Semiconductors*, Ed. by D. K. Ferry and C. Jacoboni (Plenum Press, New York, 1992).
[17] R. Lake, G. Klimeck, M. McLennan, and S. Datta, in *Proc. Intern. Workshop on Computational Electronics* (Univ. Ill. Press, Urbana, 1992) 265; R. Lake and S. Datta, Phys. Rev. B **45**, 6670 (1992).