VLSI Design
1995, Vol. 3, No. 2, pp. 159–177
Reprints available directly from the publisher
Photocopying permitted by license only

2-D Simulation of Quantum Effects in Small Semiconductor Devices Using Quantum Hydrodynamic Equations*

JING-RONG ZHOU and DAVID K. FERRY
Center for Solid State Electronics Research, Arizona State University, Tempe, AZ 85287-6206

(Received November 1, 1993)

We discuss the basis of a set of quantum hydrodynamic equations and the use of this set of equations in the two-dimensional simulation of quantum effects in deep submicron semiconductor devices. The equations are obtained from the Wigner function equation-of-motion. Explicit quantum correction is built into these equations by using the quantum mechanical expression of the moments of the Wigner function, and its physical implication is clearly explained. These equations are then applied to numerical simulation of various small semiconductor devices, which demonstrate expected quantum effects, such as barrier penetration and repulsion. These effects modify the electron density distribution and current density distribution, and consequently cause a change of the total current flow by 10–15 per cent for the simulated HEMT devices. Our work suggests that the inclusion of quantum effects into the simulation of deep submicron and ultra-submicron semiconductor devices is necessary.

Key Words: *Device simulation: Numerical simulation; Quantum modeling.*

I. INTRODUCTION

Since the advent of the integrated circuit in the late 1950's, the number of devices contained on a single chip has approximately doubled every three years as a result of the tendency for semiconductor devices to become smaller. As devices become small, some physical effects (such as quantum effects) which are not important for large devices may change the device operation significantly. The physical effects inherent in the operation of ultra-small devices are based on the fact that the critical length (e.g. the gate length or the depletion length) becomes so small that it approaches the coherence length of the electrons that provide the operation, which suggests that such small devices must be treated as quantum mechanical objects [1–4]. The coherence length, or the inelastic mean free path, can be more than 1 μm at low temperatures and as much as 0.1 μm at room temperature in high-quality heterojunction structures. This is much larger than the gate length ($\sim$ 20–25 nm) of the smallest transistors that have been made [5–8]. Due to the quantum interference within the devices, as well as between the devices, these physical effects may greatly modify the operation of a single device as well as an integrated circuit. It is very important to fully understand these quantum effects on the device and circuit operations.

The classical semiconductor transport theory is based on the Boltzmann transport equation (BTE). Numerous analytical and numerical methods have been developed for solving this equation in various semiconductor problems [9]. The Monte Carlo method provides the most accurate and detailed solution but is limited in practical engineering applications by its computational expense [10]. As an alternative, a reduced description of the BTE based upon moment equations has played a significant role in advanced device modeling [11–15]. As device feature sizes are reduced to the ultra-submicron regime, and sometimes with a narrow quantum well structure feature, device simulation faces new chal-

* Work supported by the Army Research Office.

lenges. Even the hydrodynamic model, upon which the moment equations are based and used to investigate non-stationary and hot electron dynamics through the distinction of the momentum and energy relaxation times, must be improved. Some efforts have been made in including quantum effects in the simulation of quantum well devices [16]. These generally are a combination of a classical description (either drift-diffusion equation or Monte Carlo method) with a quantum treatment in one dimension normal to the heterojunction interface. However, this does not appear to affect device performance beyond mobility modifications. As the device structures are made smaller, the 1-D treatment of the 2-D electron gas is no longer accurate, for the quantum well is not uniform along the channel and a single quantum well model is not valid. One improvement uses a set of quantum moment equations developed from a Wigner function prototype [17, 18], which preserves explicit quantum corrections as well as the classical hydrodynamic model features.

This paper is a review of our work on the modeling of quantum hydrodynamic equations and simulation of quantum effects in small semiconductor devices (including our work in [19–21]). In section II, we describe the formulation of the quantum hydrodynamic equations based upon the Wigner function equation-of-motion; In section III, we discuss the physics of quantum effects in small semiconductor devices; In section IV, we present the numerical technique and structure model for the simulation; And in section V and VI, we discuss our simulation results for MESFET and HEMT structures; finally, we will give our summary and conclusion.

II. QUANTUM HYDRODYNAMIC EQUATIONS

In principle, large-scale devices can be modeled classically, with an accurate description given by the Boltzmann transport equation. This equation time-evolves a complete single-particle phase-space distribution. However, the accurate simulation of ultra-small devices requires quantum effects such as tunneling and quantum repulsion (complementary to barrier penetration) to be included. A quantum phase-space distribution function, analogous to the Boltzmann distribution function, is useful for use in the existing mathematical methods for the classical theories [22]. A full quantum description, at the single particle level, can fruitfully be based on the Wigner distribution function (WDF) [23]

$$f_W(\mathbf{x},\mathbf{p},t) = \frac{1}{(2\pi\hbar)^3}\int_{-\infty}^{\infty} d\mathbf{y} \times e^{i\mathbf{p}\cdot\mathbf{y}/\hbar}\rho\left(\mathbf{x}+\frac{\mathbf{y}}{2},\mathbf{x}-\frac{\mathbf{y}}{2}\right), \quad (1)$$

a transformation of the density matrix $\rho(\mathbf{x},\mathbf{x}') \equiv \langle\psi^\dagger(\mathbf{x})\psi(\mathbf{x}')\rangle$ ($\psi^\dagger$ and ψ are field operators in a position representation), which is a natural generalization of the classical phase-space distribution function. Here $\mathbf{x}$ is the space coordinate, $\mathbf{p}$ is the momentum coordinate, and $\hbar$ is Planck's constant (reduced by 2π). The WDF satisfies the collisionless time-evolution equation

$$\frac{\partial}{\partial t}f_W(\mathbf{x},\mathbf{p},t) + \frac{\mathbf{p}}{m}\cdot\frac{\partial}{\partial\mathbf{x}}f_W(\mathbf{x},\mathbf{p},t) + \theta\cdot f_W(\mathbf{x},\mathbf{p},t) = 0, \quad (2)$$

where the operator $\theta\cdot f_W(\mathbf{x},\mathbf{p},t)$ is

$$\theta\cdot f_W(\mathbf{x},\mathbf{p},t) = \frac{1}{(\pi\hbar)^3}\int_{-\infty}^{\infty} d\mathbf{y}\,d\mathbf{P} \times\left[V\left(\mathbf{x}+\frac{\mathbf{y}}{2}\right)-V\left(\mathbf{x}-\frac{\mathbf{y}}{2}\right)\right] \times\sin\left(\frac{y}{\hbar}\cdot\mathbf{P}\right)f_W(\mathbf{x},\mathbf{p}+\mathbf{P},t), \quad (3)$$

which can be derived from Schrödinger's equation or Liouville's equation and has a form similar to that of the BTE, but with quantum corrections built-in by including both static potential $V(\mathbf{x})$ and momentum non-localities into the equation. The Wigner distribution function has been successfully used in simulation of a resonant tunneling diode in one dimension [24, 25], but it is not expected to be directly used for multi-dimensional device simulation because of its expense in memory storage and computation time. For a device simulation with a higher-dimensional description, the practical alternative is the reduced description of the Wigner distribution function, i.e. its moments, which are very useful because the lowest several moments represent the basic physical quantities such as density, momentum, and energy of a physical system. The equation of motion of the distribution function then results in the hydrodynamic equations. Following the same procedure as that for the classical BTE, we

have (with a relaxation time approximation)

$$\frac{\partial n}{\partial t} + \nabla \cdot \left(\frac{\langle \mathbf{p} \rangle}{m^*} \right) = 0, \tag{4}$$

$$\frac{\partial \langle \mathbf{p} \rangle}{\partial t} + \nabla \cdot \left(\frac{\langle \mathbf{pp} \rangle}{m^*} \right) = -nq\mathbf{E} - \frac{\langle \mathbf{p} \rangle}{\tau_m}, \tag{5}$$

$$\frac{\partial \langle \mathbf{p}^2 \rangle}{\partial t} + \nabla \cdot \left[\frac{\langle \mathbf{pp}^2 \rangle}{m^*} \right] = -2q\mathbf{E} \cdot \langle \mathbf{p} \rangle - \frac{\langle \mathbf{p}^2 \rangle - \langle \mathbf{p}^2 \rangle 0}{\tau_W}, \tag{6}$$

with

$$\langle \mathbf{p}^n \rangle \equiv \int_{-\infty}^{\infty} d\mathbf{p}\, f_W(\mathbf{x}, \mathbf{p}, t) \mathbf{p}^n, \tag{7}$$

where n is an integer, $\mathbf{E}$ is the electric field, m^* is electron effective mass, τ_m is the momentum relaxation time, τ_w is the energy relaxation time. For $n = 0$, Eq. (7) gives the density. However, the lowest three moment equations above are formally identical to their classical analog under the relaxation-time approximation and do not contain explicit quantum corrections, which are expected [26]. The key step to preserve quantum corrections in the lowest three moment equations relies on the method of decoupling the energy equation from higher-order moment equations and the treatment of the tensor in the momentum equation. In order to get explicit quantum corrections into the hydrodynamic equations, several different methods have been proposed [17, 27–30]. We adopt the method in [17]. By writing the WDF in the following form

$$f_{\mathrm{W}}(\mathbf{x}, \mathbf{p}, t) = \frac{1}{\hbar^3} \int_{-\infty}^{\infty} d\mathbf{y}\, e^{i\mathbf{p}\cdot\mathbf{y}/\hbar} \psi^{\dagger}\left(\mathbf{x} + \frac{\mathbf{y}}{2}\right) \psi\left(\mathbf{x} - \frac{\mathbf{y}}{2}\right), \tag{8}$$

the various moments can be evaluated. We remark that, in this equation and the following text, $\psi^{\dagger}$ and ψ are treated as wave functions instead of field operators. By taking the zero moment we get the density

$$n = \langle \mathbf{p}^0 \rangle \equiv \int_{-\infty}^{\infty} d\mathbf{p}\, f_{\mathrm{W}}(\mathbf{x}, \mathbf{p}, t) = \psi^{\dagger}(\mathbf{x}) \psi(\mathbf{x}). \tag{9}$$

The first moment carries the information of the current density

$$m^* n \mathbf{v} = \langle \mathbf{p} \rangle \equiv \int_{-\infty}^{\infty} d\mathbf{p}\, \mathbf{p} f_{\mathrm{W}}(\mathbf{x}, \mathbf{p}, t) = \left(\frac{\hbar}{2i} \right) \left[\psi(\mathbf{x}) \nabla \psi^{\dagger}(\mathbf{x}) - \psi^{\dagger}(\mathbf{x}) \nabla \psi(\mathbf{x}) \right]. \tag{10}$$

The energy density can be derived from the second moment

$$\langle \mathbf{p}^2 \rangle \equiv \int_{-\infty}^{\infty} d\mathbf{p}\, \mathbf{p}^2 f_W(\mathbf{x}, \mathbf{p}, t) = \left(\frac{\hbar}{2i} \right)^2 \left[\psi(\mathbf{x}) \nabla^2 \psi^{\dagger}(\mathbf{x}) - 2\nabla \psi^{\dagger}(\mathbf{x}) \cdot \nabla \psi(\mathbf{x}) + \psi^{\dagger}(\mathbf{x}) \nabla \psi(\mathbf{x}) \right]. \tag{11}$$

The second moment can be expressed in terms of the zeroth and the first moments, by using the following identities [25]

$$\psi(\mathbf{x}) \nabla^2 \psi^{\dagger}(\mathbf{x}) + \psi^{\dagger}(\mathbf{x}) \nabla \psi(\mathbf{x}) = \nabla^2 n - 2\nabla \psi^{\dagger}(\mathbf{s}) \cdot \nabla \psi(\mathbf{x}), \tag{12}$$

and

$$(\nabla n)^2 - \left(\frac{2i}{\hbar} \right)^2 \langle \mathbf{p} \rangle^2 = 4n \nabla \psi^{\dagger}(\mathbf{x}) \cdot \nabla \psi(\mathbf{x}), \tag{13}$$

we have

$$\langle \mathbf{p}^2 \rangle = \left(\frac{\hbar}{2i} \right)^2 \left[\nabla^2 n - \frac{1}{n} \left[(\nabla n)^2 - \left(\frac{2i}{\hbar} \right)^2 \langle \mathbf{p} \rangle^2 \right] \right] = \langle \mathbf{p} \rangle^2 / n - \frac{\hbar^2}{4} \left[\nabla^2 n - \frac{1}{n} (\nabla n)^2 \right] = \langle \mathbf{p} \rangle^2 / n - \frac{\hbar^2}{4} n \nabla^2 \ln n. \tag{14}$$

This leads to the zero temperature energy density

$$\Omega = \frac{1}{2m^*} \langle \mathbf{p}^2 \rangle = \frac{1}{2m^* n} \langle \mathbf{p} \rangle^2 - \frac{\hbar^2}{8m^*} n \nabla^2 \ln n, \tag{15}$$

which consists of drift kinetic energy density and quantum potential energy density. Now we see ex-

plicit quantum correction enters the second moment but not the lower moments. Next, we need to derive the expressions of the tensor term $\langle \mathbf{pp} \rangle$ and the third moment $\langle \mathbf{p}^3 \rangle$. From definition, the tensor term is

$$\langle \mathbf{pp} \rangle \equiv \int_{-\infty}^{\infty} d\mathbf{p}\, \mathbf{pp} f_{\mathrm{W}}(\mathbf{x}, \mathbf{p}, t), \tag{16}$$

where $\mathbf{pp}$ is a tensor. By expanding in detail, we have

$$\begin{aligned} \langle \mathbf{pp} \rangle &\equiv \int_{-\infty}^{\infty} d\mathbf{p} \sum_{ij=1}^{3} p_i p_j f_{\mathrm{W}}(\mathbf{x}, \mathbf{p}, t)\mathbf{ij} \\ &= \sum_{i,j=1}^{3} \int_{-\infty}^{\infty} d\mathbf{p}\, p_i p_j f_{\mathrm{W}}(\mathbf{x}, \mathbf{p}, t)\mathbf{ij}, \end{aligned} \tag{17}$$

where $\mathbf{i}$ and $\mathbf{j}$ are unit vectors. By performing the integration we get the following results

$$\begin{aligned} \langle \mathbf{pp} \rangle &= \sum_{i,j=1}^{3} \mathbf{ij} \left(\frac{\hbar}{2i} \right)^2 \left(\psi \frac{\partial^2 \psi^\dagger}{\partial x_i\, \partial x_j} - \frac{\partial \psi^\dagger}{\partial x_i} \frac{\partial \psi}{\partial x_j} \right. \\ &\qquad \left. - \frac{\partial \psi}{\partial x_i} \frac{\partial \psi^\dagger}{\partial x_j} + \psi^\dagger \frac{\partial^2 \psi}{\partial x_i\, \partial x_j} \right) \\ &= \sum_{i,j=1}^{3} \mathbf{ij} \left(\langle p_i \rangle \langle p_j \rangle / n - \frac{\hbar^2}{4} n \frac{\partial^2 \ln n}{\partial x_i\, \partial x_j} \right) \\ &= \langle \mathbf{p} \rangle \langle \mathbf{p} \rangle / n - \frac{\hbar^2}{4} n \nabla\nabla \ln n, \end{aligned} \tag{18}$$

where $\langle \mathbf{p} \rangle \langle \mathbf{p} \rangle$ and $\nabla\nabla$ are tensors. This is the zero temperature tensor with explicit quantum correction included. Similarly, the third moment is

$$\begin{aligned} \langle \mathbf{p}^3 \rangle &\equiv \int_{-\infty}^{\infty} d\mathbf{p}\, \mathbf{p}^3 f_W(\mathbf{x}, \mathbf{p}, t) \\ &= \sum_{i,j=1}^{3} \mathbf{i} \int_{-\infty}^{\infty} d\mathbf{p}\, p_i p_j^2 f_W(\mathbf{x}, \mathbf{p}, t) \\ &= \sum_{i,j=1}^{3} \mathbf{i} \left(\frac{\hbar}{2i} \right)^3 \cdot \left[\psi \frac{\partial^3 \psi^\dagger}{\partial x_i\, \partial x_j^2} - \frac{\partial \psi}{\partial x_i} \frac{\partial^2 \psi^\dagger}{\partial x_j^2} \right. \\ &\qquad - 2 \frac{\partial^2 \psi^\dagger}{\partial x_i\, \partial x_j} \frac{\partial \psi}{\partial x_j} + 2 \frac{\partial \psi^\dagger}{\partial x_j} \frac{\partial^2 \psi}{\partial x_i\, \partial x_j} \\ &\qquad \left. + \frac{\partial \psi^\dagger}{\partial x_i} \frac{\partial^2 \psi}{\partial x_j^2} - \psi^\dagger \frac{\partial^3 \psi}{\partial x_i\, \partial x_j^2} \right]. \end{aligned} \tag{19}$$

After some tedious work, one find the above equation can be written as

$$\begin{aligned} \langle \mathbf{p}^3 \rangle &= \sum_{i,j=1}^{3} \mathbf{i} \cdot \left(\langle p_i \rangle \langle p_j \rangle^2 / n^2 - \frac{\hbar^2}{4} \left(\langle p_i \rangle \frac{\partial^2 \ln n}{\partial x_j^2} \right. \right. \\ &\qquad \left. \left. + 2 \langle p_j \rangle \frac{\partial^2 \ln n}{\partial x_i\, \partial x_j} + n \frac{\partial^2}{\partial x_i\, \partial x_j} \left(\frac{\langle p_j \rangle}{n} \right) \right) \right) \\ &= \sum_{i,j=1}^{3} \mathbf{i} \cdot \left(\langle p_i p_j \rangle \langle p_j \rangle / n - \frac{\hbar^2}{4} \left(\langle p_i \rangle \frac{\partial^2 \ln n}{\partial x_j^2} \right. \right. \\ &\qquad \left. \left. + \langle p_j \rangle \frac{\partial^2 \ln n}{\partial x_i\, \partial x_j} + n \frac{\partial^2}{\partial x_i\, \partial x_j} \left(\frac{\langle p_j \rangle}{n} \right) \right) \right), \end{aligned} \tag{20}$$

or

$$\begin{aligned} \langle \mathbf{p}^3 \rangle &= \langle \mathbf{p} \rangle \langle \mathbf{p} \rangle^2 / n^2 - \frac{\hbar^2}{4} \left(\langle \mathbf{p} \rangle \nabla^2 \ln n + 2 \langle \mathbf{p} \rangle \right. \\ &\qquad \left. \cdot (\nabla\nabla \ln n) + n (\nabla\nabla) \cdot \left(\frac{\langle \mathbf{p} \rangle}{n} \right) \right). \end{aligned} \tag{21}$$

This is the exact equation for the third moment at zero temperature with quantum correction included.

So far, we have derived the expressions of the first four moments and a tensor moment (Eqs. (9), (10), (14), (18) and (21)) for the WDF at zero temperature. Except for the definition of the WDF Eq. (8), there is no further restriction imposed on the distribution function. In other words, the derivation of the moments above is exact under the definition of Eq. (8). The apparently missing terms in the moments are the temperature terms. We now turn to investigate how the temperature terms should be included in these moments.

At high temperature limit, quantum corrections are negligible, thus a classical distribution function could be used in describing a physical system. Under drifted-symmetric approximation [or $f(\mathbf{x}, \mathbf{p}, t) = f(\mathbf{x}, [\mathbf{p} - \langle \mathbf{p} \rangle / n]^2, t)$] of the distribution function, the classical second, third and tensor moments can be written as follows

$$\langle \mathbf{p}^2 \rangle = 3nm^* k_B T_e + \langle \mathbf{p} \rangle^2 / n, \tag{22}$$

$$\langle \mathbf{pp} \rangle = nm^* k_B T_e (\mathbf{ii} + \mathbf{jj} + \mathbf{kk}) + \langle \mathbf{p} \rangle \langle \mathbf{p} \rangle / n, \tag{23}$$

$$\langle \mathbf{p}^3 \rangle = \langle \mathbf{p}^2 \rangle \langle \mathbf{p} \rangle / n + 2m^* k_B T_e \langle \mathbf{p} \rangle + q, \tag{24}$$

where

$$q = \left\langle (\mathbf{p} - \langle \mathbf{p} \rangle / n)^3 \right\rangle, \tag{25}$$

is zero under the drifted-symmetric approximation of the distribution function, k_B is the Boltzmann constant, T_e is the average electron effective temperature. Extensive discussions on the term q can be found in the literature [14, 15], which regard this term as heat flow. We leave this term in the form of (25), since we believe it may represent more than just the heat flow and could be characterized by other methods [29, 30].

Comparing Eqs. (14), (18) and (21) with these latter three equations, one finds that the first lacks thermal energy and the latter set misses quantum corrections. It is physically acceptable that a combination of the quantum version and classical version of moments should give us a correct set of moments which can be used in describing a quantum mechanical system. As a matter of fact, conceptually, the thermal energy density can be defined as the total kinetic energy density minus the drift kinetic energy density and minus the quantum potential energy density

$$\frac{3}{2} n k_B T_e = \frac{1}{2m^*} \langle \mathbf{p}^2 \rangle - \frac{1}{2m^* n} \langle \mathbf{p} \rangle^2 - \left(-\frac{\hbar^2}{8m^*} n \nabla^2 \ln n \right), \tag{26}$$

where the factor of 3 account for three dimensions. Thus by combining the quantum and classical versions of the moments, we arrive the following moments with explicit quantum corrections included

$$\langle \mathbf{p}^2 \rangle = 3nm^* k_B T_e + \frac{1}{n} \langle \mathbf{p} \rangle^2 - \frac{\hbar^2}{4} n \nabla^2 \ln n, \tag{27}$$

$$\langle \mathbf{pp} \rangle = \langle \mathbf{p} \rangle \langle \mathbf{p} \rangle / n + nm^* k_B T_e (\mathbf{ii} + \mathbf{jj} + \mathbf{kk}) - \frac{\hbar^2}{4} n \nabla \nabla \ln n, \tag{28}$$

$$\langle \mathbf{p}^3 \rangle = \langle \mathbf{p}^2 \rangle \langle \mathbf{p} \rangle / n + 2m^* k_B T_e \langle \mathbf{p} \rangle + q - \frac{\hbar^2}{4} \left(2 \langle \mathbf{p} \rangle \cdot (\nabla \nabla \ln n) + n (\nabla \nabla) \cdot \left(\frac{\langle \mathbf{p} \rangle}{n} \right) \right). \tag{29}$$

Substituting these moments into the hydrodynamic equations (4–6), we get the following quantum hydrodynamic equations

$$\frac{\partial n}{\partial t} + \nabla \cdot \left(\frac{\langle \mathbf{p} \rangle}{m^*} \right) = 0, \tag{30}$$

$$\frac{\partial \langle \mathbf{p} \rangle}{\partial t} + \nabla \cdot \left(\frac{\langle \mathbf{p} \rangle \langle \mathbf{p} \rangle}{m^* n} + \hat{\mathbf{U}}_q \right) = -nq\mathbf{E} - \nabla (n k_B T) - \frac{\langle \mathbf{p} \rangle}{\tau_m}, \tag{31}$$

$$\frac{\partial nw}{\partial t} + \nabla \cdot \left(\frac{\langle \mathbf{p} \rangle}{m^*} w + \frac{\langle \mathbf{p} \rangle}{m^*} \cdot \hat{\mathbf{U}}_q + \frac{n}{m^*} U_p \right) = -\frac{q\mathbf{E} \cdot \langle \mathbf{p} \rangle}{m^*} - \nabla \cdot \left(\frac{\langle \mathbf{p} \rangle k_B T}{m^*} \right) - \nabla \cdot \boldsymbol{q} - \frac{nw - nw_0}{\tau_W}, \tag{32}$$

where

$$\hat{\mathbf{U}}_q = -\frac{\hbar^2}{4m^*} \nabla \nabla \ln n, \tag{33}$$

$$U_q = -\frac{\hbar^2}{8m^*} \nabla^2 \ln n, \tag{34}$$

$$U_p = -\frac{\hbar^2}{8m^*} \nabla \left(\nabla \cdot \frac{\langle \mathbf{p} \rangle}{n} \right), \tag{35}$$

$$w = \frac{1}{2m^* n} \langle \mathbf{p} \rangle^2 + \frac{3}{2} k_B T + U_q. \tag{36}$$

Equations (30–36) are our complete quantum hydrodynamic equations. We point out here: 1) The closure for the equations (the decoupling to the higher moments hierarchy) is done as generally as in the classical case. 2) The quantum correction terms are exact under the definition of the WDF in (6), although one may argue that the formalism is for a pure state and lacks the ensemble average, it is still suitable for a general physical system, since we have no restriction on the form of the wave function. Furthermore, Ancona and Iafrate [31], using the expansion of the potential in terms of the ratio of the thermal wavelength to the characteristic length over which the potential varied, obtained the same form of (34) with a reduction factor of 1/3. Grubin *et al.* [32], working with the density matrix, achieved the same reduction factor. Gardner [33], again following the potential expansion, gets exactly the same quantum correction terms as ours except for the factor of 1/3. While the validity of the potential expansion is questionable, we notice that if we have a reduction factor of 1/3, the equations can not

return to zero temperature equations correctly. 3) No clear quantum corrections have been included explicitly into the scattering terms and the term q, and there is no simple method that can introduce a proper quantum correction for the scattering term with even moderate computational effort. Any attempt to evaluate the scattering terms quantum mechanically should return to Levinson's formalism [34], which is the Wigner-Weyl transformation of the interaction terms of the Hamiltonian. A further investigation of the quantum corrections and their origin is discussed in [35].

Comparing to classical equations, the quantum hydrodynamic equations need more computational effort, needless to say much more than the drift-diffusion equations. To investigate the impact of the quantum corrections on the simulation with a clear physical picture and moderate computation, we concentrated on the correction term of the energy, as the modification of the energy directly changes the density distribution which is proportional to the factor of $e^{-(V+W)/k_BT}$, quantum penetration (and repulsion) can be observed. By keeping the major correction for the energy, we approximate the quantum hydrodynamic equations to a simpler set of equations (with temperature representation)

$$\frac{\partial n}{\partial t} + \nabla \cdot (n\mathbf{v}) = 0, \tag{37}$$

$$\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} = -\frac{q\mathbf{E}}{m^*} - \frac{1}{nm^*}\nabla(nk_BT_q) - \frac{\mathbf{v}}{\tau_m}, \tag{38}$$

$$\frac{\partial T}{\partial t} + \frac{1}{3}\mathbf{v} \cdot \nabla(T_q) = -\frac{2}{3}\nabla \cdot (\mathbf{v}T_q) + \frac{m^*\mathbf{v}^2}{3k_B}\left(\frac{2}{\tau_m} - \frac{1}{\tau_w}\right) - \frac{T - T_0}{\tau_w}, \tag{39}$$

where $\mathbf{E}$ is the electric field, T_0 is the lattice temperature and T_q is

$$T_q = T + \frac{2}{3k_B}U_q. \tag{40}$$

This set of equations preserves all classical features except the heat flow property (we leave this for future investigation), and gives explicit quantum corrections. As $\hbar$ goes to zero, the equations return to the full classical hydrodynamic equations. From Eq. (40), one observes that if the thermal energy of the electron is large, the quantum correction has less effect. But as the temperature is lowered, the quantum correction will become dominant. With Poisson's equation

$$\nabla^2 V = \frac{q}{\varepsilon_s}(N_D - n), \tag{41}$$

where V is the electrical potential, q is the absolute electron charge, ε_s is the semiconductor permittivity, and N_D is the doping concentration, Eqs. (37–40) are used in our numerical simulations.

III. ON THE CONCEPT OF QUANTUM CORRECTION IN SMALL DEVICES

In order to demonstrate how the quantum correction (8) works and what this correction means physically, we sketch a single barrier with an approximate electron density distribution in Fig. 1. With the quantum correction included, the total energy can be written as

$$E = E_{\mathrm{cl}} + U_q, \tag{42}$$

where $\mathrm{E_{cl}}$ is the classical total energy and U_q is the quantum correction energy. The density is proportional to the Boltzmann factor

$$\rho \propto \exp\left[-(E_{\mathrm{cl}} + U_q)\right]. \tag{43}$$

For the potential barrier model in Fig. 1(a), without U_q, the classical total energy is constant in and outside the barrier, which results in a constant den-

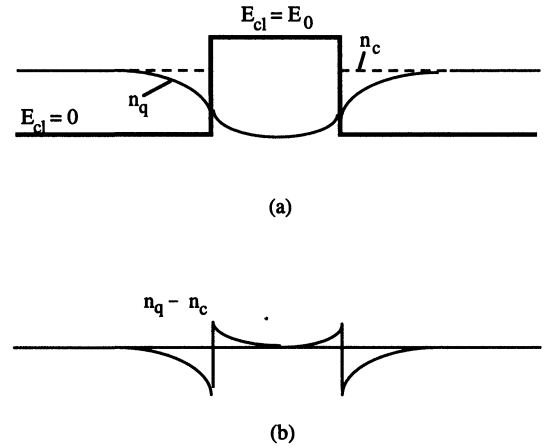

FIGURE 1 Quantum penetration and repulsion by a potential barrier.

sity distribution (n_c) outside the barrier, and a constant zero density inside the barrier. The energy discontinuity creates a density discontinuity at the interface of the barrier, which is quantum mechanically incorrect. By including quantum corrections, the quantum potential energy serves to smooth the actual potential, and modifies the total energy to a smooth transition at the interface, which results in smooth density (n_q) change at the interface transition. In Fig. 1(b), we illustrate the modification of the density distribution by inclusion of the quantum corrections. We distinguish these quantum effects from those of transverse quantization of the electron in a MOSFET or HEMT channel, an effect which does not affect overall device performance [10, 16], and which requires the treatment of each discretized energy level individually. However, since with certain ensemble statistics [28–30], the quantum corrections still take essentially the same form, we expect that certain summation effects of the transverse quantization levels may be included in the formulations.

Besides the abrupt barrier example, the quantum correction arises from the change of the density (the density is reduced when the second derivative of the local log density is negative, and the density is enhanced when the second derivative of the local log density is positive), one can imagine that anywhere a large density change occurs in a short distance, the actual density distribution will be much different from that of a classical picture, even the classical picture gives a continuous density distribution as it occurs in small semiconductor devices.

IV. NUMERICAL IMPLEMENTATION

We use finite difference methods to discretize the quantum hydrodynamic equations and the Poisson's equation in a two-dimension structure (Fig. 2), which is suitable to any planar device structure. The difference schemes used in the simulation are described as follows: a central difference is used for Poisson's equation. In the continuity equation, the gradient term $\nabla \cdot (n\mathbf{v})$ is discretized by using a second-upwind method (or donor cell method) [36], which possesses both conservative and transport properties. In the momentum equation, a central difference is used for the term $\nabla(nk_BT_q)$ at half-grid points, and a first-upwind method is used for the term $(\mathbf{v} \cdot \nabla)\mathbf{v}$. In the energy equation, two terms need to be discretized in space. These are $\mathbf{v} \cdot \nabla T_q$ and $\nabla \cdot (\mathbf{v}T_q)$. The temperature pressure term $\nabla \cdot (\mathbf{v}T_q)$ is in a conservative form, so the donor cell scheme is suitable for this term. The convective term $\mathbf{v} \cdot \nabla T_q$ doesn't have a conservative property, and a first-upwind difference is implemented for this term.

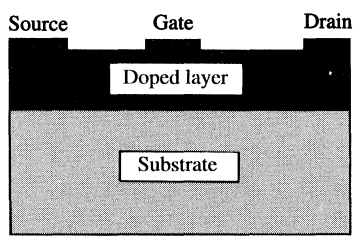

FIGURE 2 Device structure for simulation.

A forward-time difference is adopted for the continuity equation, and an integration and expansion method [15] is used for the momentum and energy equation. Both uniform and nonuniform meshes are used in our simulation. Grid sizes are chosen to satisfy the constraint of the extrinsic Debye length, and a time step is chosen by considering the Courant-Freidrichs-Lewy stability condition. The relaxation times are computed from the velocity-field and energy-field relations [37] which are Monte Carlo simulation results for bulk material. For GaAs material, it is equivalent to convert a multi-valley system to an effective single valley model by this approach. While not strictly accurate, it is suitable for examining the impact of the quantum corrections, which is the aim of the present work. Dirichlet boundary conditions are applied to the contacts and Neumann boundary conditions are used where the perpendicular current flow is zero. An incomplete Cholesky conjugate gradient method [38–40] is used to solve Poisson's equation. Gauss-Seidel iteration is applied to the quantum moment equations. After the initial guess for the potential, density, temperature and velocity, the various equations are solved successively for the corresponding quantities.

All devices simulated here have the structure shown in Fig. 2. However, the interface for heterostructure devices needs to be considered in more detail. For an ideal AlGaAs/GaAs interface, the transition of the conduction band from one material to another is abrupt, and characterized by a distinct band offset. At an equilibrium condition across the interface, due to the diffusion of electron, a field is generated that depletes the donors in the AlGaAs bulk and creates an inversion layer of electrons at the interface on the GaAs side. The potential distribution, coupled with the band offset, forms a barrier on the AlGaAs side and a well on the GaAs side.

Electrons in the quantum well are prevented from drifting into the AlGaAs by the band offset. Electrons can climb over the potential wall only if they have kinetic energy comparable to the conduction band discontinuity. From the interface structure, one may realize an obvious problem in using differential equations in the simulation of this kind of device. The problem is that the partial differential equations can not normally handle the discontinuity, where an infinite field occurs, and this must be carefully handled by the method of discretization. The problem arises not in Poisson's equation, but in the moment equations themselves. The step in the potential can be a source of instability in the computation. While an abrupt change in the transition from one material to another may be ideal, there is a certain transition region to be expected [41]. Thus, an assumption of a narrow transition region can be made, although the estimate of its extent is difficult to determine. The energy transport equation is the main one in the sense that the electron kinetic energy determines the transport of the electrons across the interface. The assumption of a small transition region does not create a significant error. In our simulation, we assume a 0.3 volt potential drop across a 4 nanometer region at the interface for a AlGaAs/GaAs HEMT, and a 0.18 volt potential drop across a 3 nanometer region for a SiGe device (in this case, a double heterojunction is treated). While this seems to be quite wide, it is regarded as a statistical average over the actual transition region and the wavefunction decay at this interface [42], an approach used extensively in statistical physics. However, a further reduce of the grading width at the interface may be considered.

The conduction path in a quantum well device is from the source contact down to the two-dimensional electron gas, and then through the 2-D conduction channel to the drain region. The source and drain regions are heavily doped, in general. Due to this, transport of electrons across the hetero-interface in the source/drain regions is probably by tunneling through a very thin potential barrier, in which effectively no discontinuity at all is experienced. For consistency with this concept, we assume that the interface discontinuity gradually disappears toward the contact regions. Although this is a conceptual problem, it does not affect the quantum effects that appear in the gate region and is a model of the "ohmic" contact. In order to give a clear picture of the effects of these assumptions, we plot the potential profile of a 24 nm gate length AlGaAs/GaAs HEMT device in Fig. 3 (we plot voltage rather than energy, so that the potential profile under the gate appears inverted), where we can see the transition of the potential at the interface and the reduced potential barrier height towards the source and drain regions.

The results we summarize here are for MESFET devices with gate lengths from 24 nm to 96 nm, AlGaAs/GaAs HEMT devices with gate lengths from 24 nm to 56 nm, and a modulation-doped $Si_{0.7}Ge_{0.3}/Si/Si_{0.7}Ge_{0.3}SiGe$ quantum-well device with a gate length of 0.18 μm. For the MESFET, typical doping in the channel is 1.5×10^{18} cm^{-3}, and a semi-insulating substrate is included. The

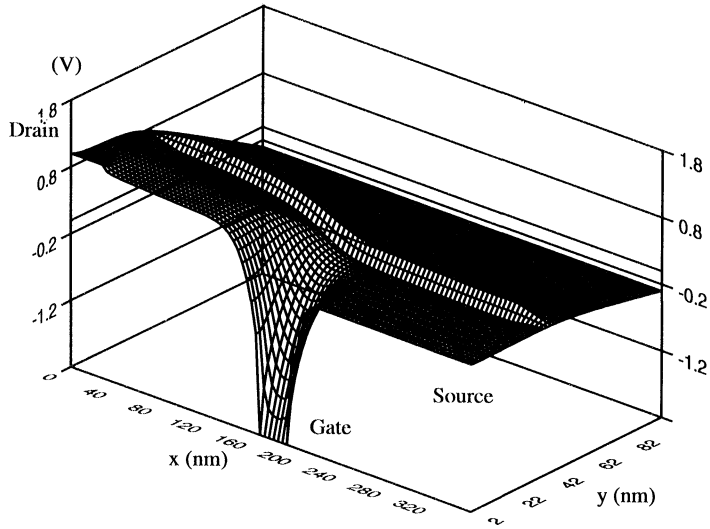

FIGURE 3 The two dimensional potential profile of a HEMT. $L_g = 24$ nm, $V_d = 1$ V, $V_g = -1.5$ V. The interface is at 40 nm.

doping in the HEMT AlGaAs is also 1.5×10^{18} cm^{-3}. Much higher doping (3.5×10^{18} cm^{-3}) is used for the modulation-doped SiGe device in the top $Si_{0.7}Ge_{0.3}$ layer. The lattice temperature is taken to be 300 K in all cases. The total simulation area is 0.36 μm $\times$ 0.1 μm for MESFETs and AlGaAs/GaAs HEMTs, and 1.0 μm $\times$ 0.095 μm for the modulation-doped SiGe device. The thickness of the AlGaAs layer is 39 nm, and the thickness of the top SiGe layer is 19 nm. The strained Si channel is 18 nm.

V. GENERAL DEVICE CHARACTERIZATION

As we mentioned before, the gate lengths of the device we simulate ranges from deep submicron to ultra-submicron, which allows us to understand the small device operation and the effects of the quantum corrections on the device characteristics. We discuss the general device characteristics in this section, and leave the quantum effects to the next section.

Plotted in Fig. 4 and Fig. 5 are the I-V characteristics of 24 nm gate length GaAs MESFET and AlGaAs/GaAs HEMT, respectively. The gate voltage runs from 0 V to -2.5 V, in steps of -0.5 volts. For both devices, the interface (channel to substrate for MESFET, and AlGaAs to GaAs for HEMT) is located 39 nm from the gate. The characteristics of the devices are quit normal, and saturation of the current is obtained. Pinchoff is reached in the MESFET at a gate voltage of -2.5 V, and is determined by checking the density distribution in the channel. The remaining large current at pinchoff is due to substrate current as the electrons are pushed into the substrate. At least another 1.0 V of negative gate bias is required to eliminate this parasitic substrate current. The I-V characteristics for a 0.18 μm gate modulation doped SiGe HEMT device are illustrated in Fig. 6 and the gate biases are 0.7, 0.5, 0.2, and 0 volts, respectively. The small thickness of the top SiGe layer (18 nm) provides a normally-off device, since a Schottky barrier height of 0.9 V (Pt on Si) leads to an estimated depletion width of 18.4 nm. Good saturation with a drain conductance of 4.6 mS/mm at the gate voltage of 0.5 V is obtained. Approximately the same current level was found in a 0.25 μm device. Obviously, for this larger gate length device, the current pinchoff is much easier to achieve.

The transconductance is 450 mS/mm for the MESFET, 480 mS/mm for AlGaAs/GaAs HEMT,

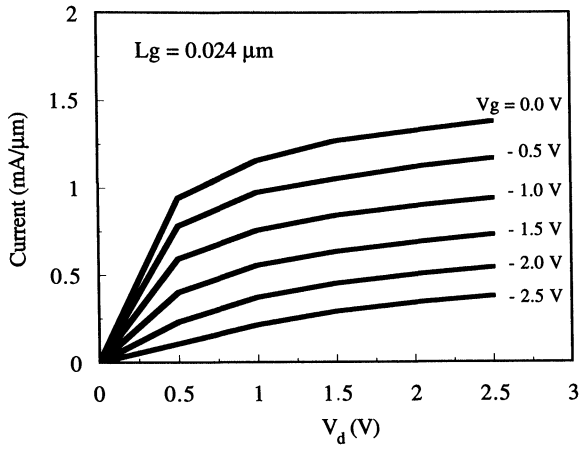

FIGURE 4 I-V characteristics of a 24 nm gate GaAs MESFET device.

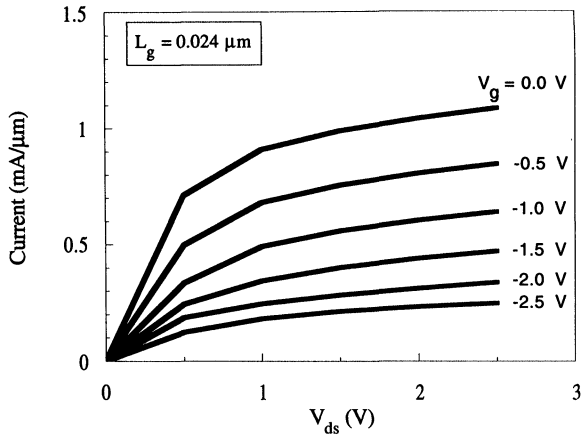

FIGURE 5 I-V characteristics for a 24 nm gate length AlGaAs/GaAs HEMT device.

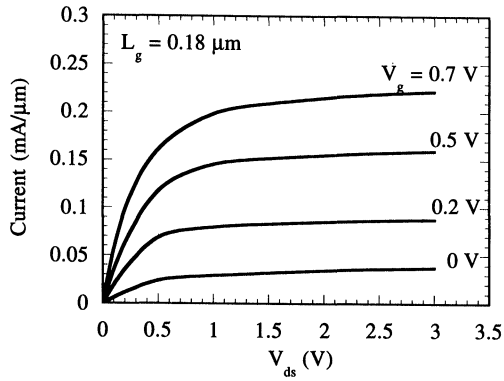

FIGURE 6 I-V characteristics of a 0.18 μm SiGe HEMT device.

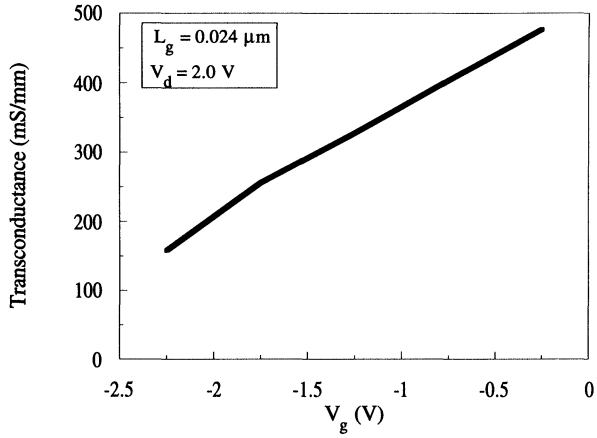

FIGURE 7 Transconductance as a function of gate voltage for a 24 nm gate AlGaAs/GaAs HEMT device at V_d = 2.0 V.

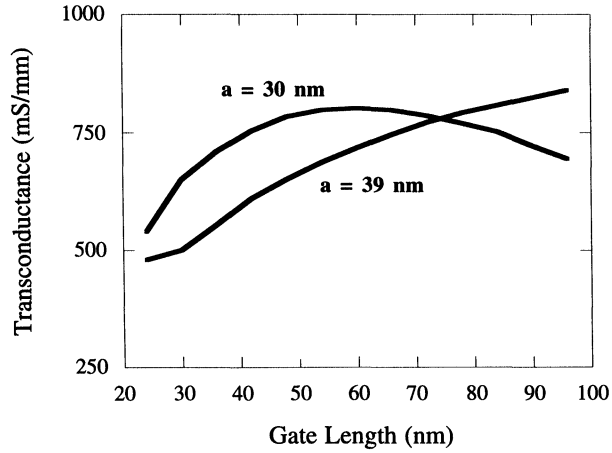

FIGURE 8 Transconductance vs gate length for GaAs MESFET devices (results for two different active layer depths, a = 30 nm and a = 39 nm, are shown).

and 300 mS/mm for the SiGe device discussed above. As for the depletion mode AlGaAs/GaAs HEMT device, the transconductance behavior is slightly different from the general picture of a HEMT. Fig. 7 shows the transconductance versus gate voltage for the 24 nm gate HEMT, at a drain bias of 2.0 V, in which the transconductance decreases linearly from about 480 mS/mm to some 150 mS/mm as gate bias becomes more negative. For most experimental HEMT device operation, one expects that the current flow will be mainly confined to the 2-DF electron gas channel, and the transconductance should have a peak value as the gate bias is varied, with the transconductance also becoming smaller as the gate voltage becomes positive, where the 2-D gas channel is fully open and the gate bias loses control of the current flow in the channel [43]. For the thickness of the AlGaAs, and the doping density, used here, it is certain that there is significant current flow through the AlGaAs for any gate bias larger than -1.0 V for a 24 nm gate length device. Although the mobility in the AlGaAs is low compared to that in the 2-D gas channel, the size structure used here is such that the gate does not really lose control of the channel charge within the range of biases examined. Rather, in this simulation, as the gate voltage begins to lose control of the charge in the 2-D gas, the conduction through the AlGaAs increases sufficiently rapidly that the transconductance continues to increase. We would expect that the transconductance would eventually decrease at positive gate voltages.

We investigated the effect of gate length on the transconductance for the MESFET and AlGaAs/GaAs HEMT by changing the gate length (with the doped layer depth and the doping concentration fixed). For the MESFET, the transconductance is evaluated at a drain voltage of 2.0 V. Fig. 8 illustrates the transconductance characteristics in the range of gate length from 24 nm to 96 nm, for two active layer depths (a = 30 nm and 39 nm). The transconductance for a = 30 nm has a maximum value of about 800 mS/mm at a gate length of 60 nm. As gate length decreases from 96 nm, the transconductance increases until the peak is reached and then decreases with further decrease of the gate length. This transconductance behavior can be explained by velocity overshoot effects for the increasing (longer gate length) transconductance region and small aspect ratio effect for the decreasing (shorter gate length) transconductance region. The transconductance for a = 39 nm decreases monotonically for the entire range as gate length decreases. By comparing these two curves, we see that the small aspect ratio effect begins at larger gate length for thicker active depth devices, an expected physical result. For the AlGaAs/GaAs HEMT, the peak transconductance observed for gate lengths from 24 nm to 56 nm (the thickness of AlGaAs layer is 39 nm) is shown in Fig. 9. These are all evaluated at $V_g = -0.5$ V and $V_d = 2.0$ V. The monotonic decrease in transconductance with decrease of the gate length is directly related to the small aspect ratio (L_g/a) [5, 7, 44]. We note that the devices are not scaled; only L_g is varied in the simulation. In all cases, $L_g/a < 1$, rather than 3–4 that is normally used in longer-channel HEMTs, and this fact alone is felt to lead to the decreasing transconductance.

The transport property of electrons in the conduction channel is critical for the device performance. To see how the velocity changes under the influence

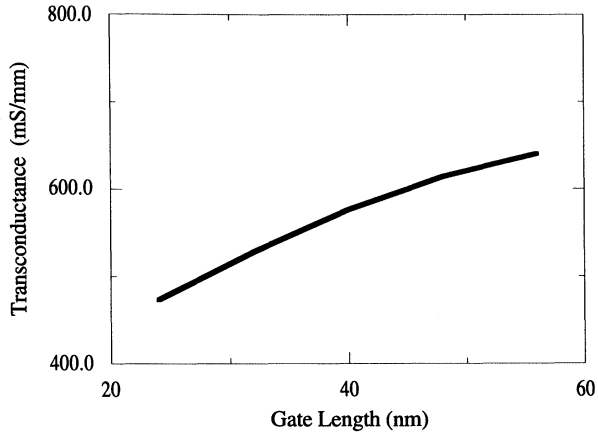

FIGURE 9 Transconductance as a function of gate length AlGaAs/GaAs HEMT device at V_d = 2.0 V.

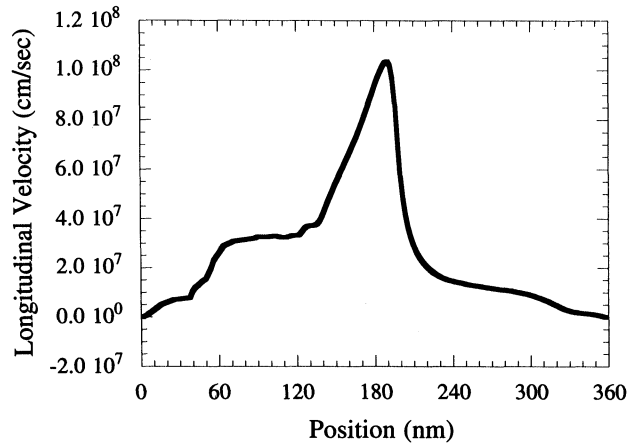

FIGURE 11 The longitudinal velocity in the channel of a AlGaAs/GaAs HEMT. L_g = 24 nm, a = 39 nm, V_d = 2 V, V_g = −1.5 V.

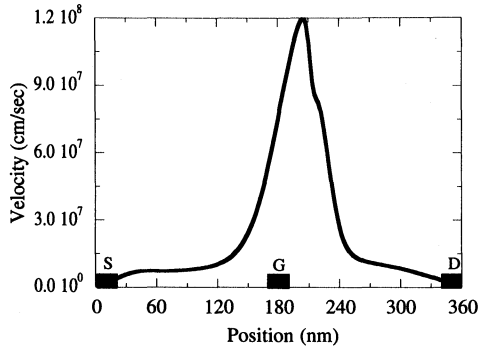

FIGURE 10 The longitudinal velocity in the channel of a GaAs MESFET. L_g = 24 nm, a = 39 nm, V_d = 2 V, V_g = −1.5 V.

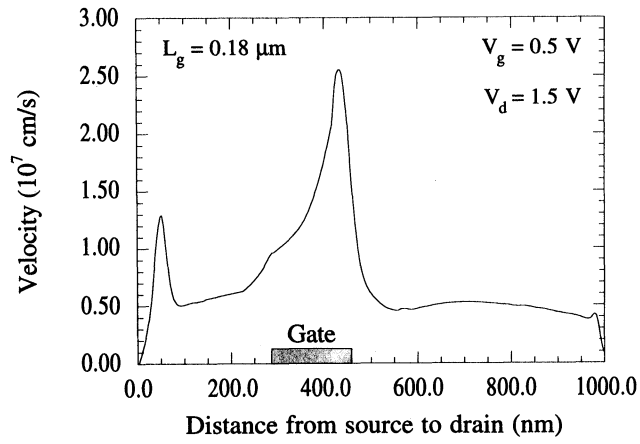

FIGURE 12 The longitudinal velocity in the channel of a SiGe HEMT. L_g = 0.18 μm, V_d = 1.5 V, V_g = 0.5 V.

of the potential barrier induced by the gate, we plot the longitudinal velocity as a function of position along the conduction channels in Fig. 10, 11 and 12, for a 24 nm gate GaAs MESFET, a 24 nm gate AlGaAs/GaAs HEMT and a SiGe HEMT, respectively. Velocity overshoot is obvious, as the peak velocities are much larger than either the saturation velocity or the peak velocity in the velocity-field relations for all cases. The reason for velocity overshoot in SiGe device is due to the high mobility of the electron in the strained Si layer with carrier transfer out of the lower set of valleys and into the upper set of strain split valleys [45]. The velocity overshoot is thought to be important to achieve the high transconductance for these devices. The first velocity peak in the plot of SiGe device is due to the model we used for the change of interface discontinuity (as describe in section III), although it is not practical, it does suggest that the structure can increase the electron velocity between source and gate, which in turn will raise the average velocity through the device and enhance the device performance. It is interesting that there is no similar velocity peak for the AlGaAs/GaAs HEMT close to the source, although the velocity in the later device does increase faster than that in the GaAs MESFET, which implies that with the structure model the double heterojunction creates larger longitudinal electrical field in SiGe device than the single heterojunction does in the AlGaAs/GaAs HEMT. For the GaAs MESFET and AlGaAs/GaAs HEMT devices, the peak velocity exceeds 10^8 cm/sec, which is surprisingly high (near the band structure limit). However, if one notices that the velocity is actually built up within a 60 nm distance, one may realize that quasi-ballistic transport (the

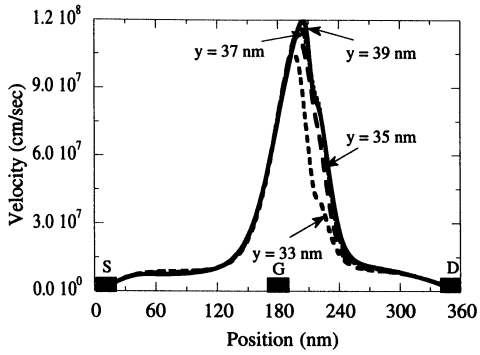

FIGURE 13 The longitudinal velocity in the channel of a GaAs MESFET for different y positions (y is the direction into the device from the surface). $L_g = 24$ nm, $a = 39$ nm, $V_d = 2$ V, $V_g = -1.5$ V.

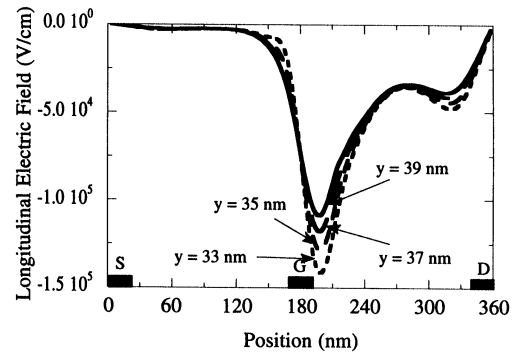

FIGURE 14 The longitudinal electric field in the channel of a GaAs MESFET for different y positions. $L_g = 24$ nm, $a = 39$ nm, $V_d = 2$ V, $V_g = -1.5$ V.

electron experiences few scattering events during the fly) of the electron is quite possible as the length scale is comparable to the electron mean free path. We can estimate the velocity from Newton's law. For a relaxation time around 10^{-13} s, and with an electric field of 10^5 V/cm, an electron with the effective mass used here is able to reach a velocity of 10^8 cm/sec on this time and length scales. Thus, within the accuracy of the relaxation-time approximation, large velocity overshoot is possible. We remark, however, that the actual simulations use energy-dependent relaxation times that fully match detailed ensemble Monte Carlo results; the previous argument is one of justification only. We also remark that these results have good validity, as it has been shown that hydrodynamic equations agree well with Monte Carlo results for transient response when the models are the same [46].

To see the relations among velocity, field and temperature, we take the MESFET as an example. Since the MESFET has a wider conduction channel, it would be informative if we knew how the velocity changes for different transverse positions. To see this, we plot the longitudinal velocity as a function of position along the channel for different transverse positions in Fig. 13, for a 24 nm gate GaAs MESFET (although simulations have been done for many gate lengths and different devices, the details presented here are limited to this later value), for the bias condition of $V_d = 2.0$ V and $V_g = -1.5$ V. The active-layer-to-substrate interface is at 39 nm from the gate. Four velocity curves are plotted at here for different positions which have distances of 8, 6, 4, 2 nm from the interface, respectively. By inspecting the density distribution across the channel, these positions cover the entire channel width on the active layer side. Corresponding longitudinal electric fields are shown in Fig. 14. For the axis direction we choose, the sign of all the field curves are negative. The peak field for each curve reaches 10^5 V/cm just a few nanometers beyond the gate metal edge on the drain side. The field is very low over a long path from the source to the gate and has a very sharp increase in the gate region. This is the key to the very high velocity overshoot, for the driving force increases quickly in a short distance. The decrease of the electric field near the drain is fast but the field remains relatively strong throughout the region from the gate to the drain. In Fig. 15, four curves of effective electron temperature are depicted which take corresponding space positions to the velocity curves under discussion. The temperature is close to the lattice temperature from the source to the center of the gate, then it rises quickly to above 3000 K over the next 60 nm. The low temperature in the source-to-gate region provides the key condition to the high velocity overshoot, since it accounts for less momentum scattering. By

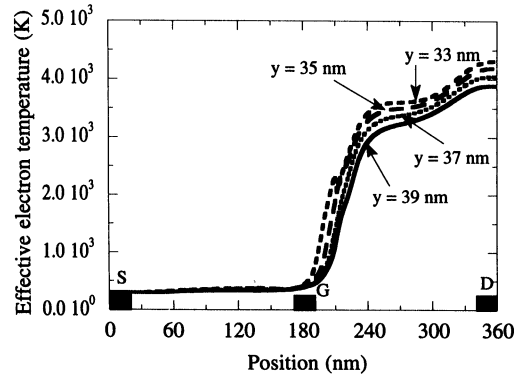

FIGURE 15 Effective electron temperature in the channel of a GaAs MESFET for different y positions. $L_g = 24$ nm, $a = 39$ nm, $V_d = 2$ V, $V_g = -1.5$ V.

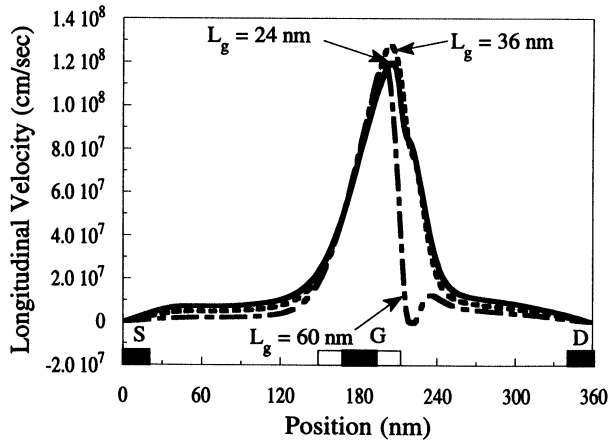

FIGURE 16 The velocity along the channel of a GaAs MESFET for different gate length devices. Here, the velocity is plotted for $y = 39$ nm.

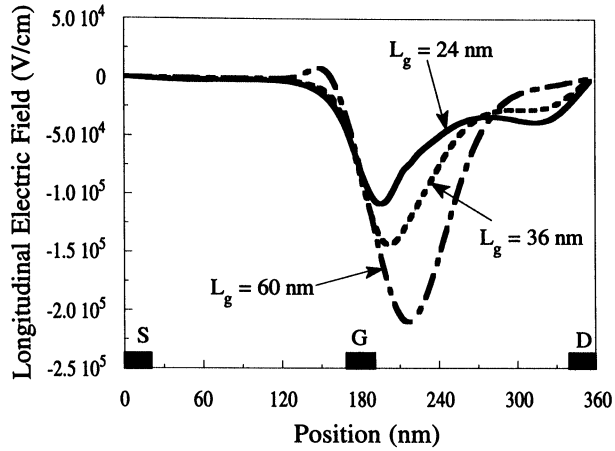

FIGURE 17 The electric field along the channel of a GaAs MESFET for different gate length devices. Here, the field is plotted for $y = 39$.

checking the total energy at the velocity peak positions, one finds that the values are about 0.38 eV, which roughly corresponds to the energy separation from L valley to Γ valley. This means that after the total energy exceeds the energy separation of the two valleys, more intervalley scattering occurs and the velocity is then reduced. The fact that the result is a somewhat higher energy than the actual $\Delta_{\Gamma L}$ is though to be due to the use of a constant effective mass, which leads to an underestimation of the actual relaxation rates.

The longitudinal velocity profiles for MESFETs with several different gate lengths are shown in Fig. 16 (with 2 nm distance from the interface), with the same bias condition above. It can be seen that the peak velocity reached by the carriers is slightly higher in the 36 nm gate length device than in the 24 nm gate length device. This also corresponds to a higher peak electric field (Fig. 17). In essence, the peak velocities reached are limited by the short acceleration lengths in the short gate devices, a result predicted earlier [7, 47]. In nearly all the cases, however, it is clear that the velocity rises almost linearly throughout the region under the gate, so that the overshoot region is essentially defined by the gate length of the device.

V. THE QUANTUM EFFECTS IN SMALL DEVICES

The effect of the quantum correction on the device characteristics can be found by comparing the computed results in the full model with those obtained when $\hbar = 0$, i.e., in the semiclassical hydrodynamic model. Here, we compare these differences in any quantity Q as

$$dQ = Q|_{\text{full model}} - Q|_{\text{no quantum terms}}. \quad (44)$$

As the gate length of the device is reduced to the length scale simulated here, we expect that quantum effects such as barrier repulsion and penetration can be observed. We first examine quantum effects on the density distribution. Fig. 18 shows a two-dimensional plot of the density difference between results obtained with and without the quantum potential correction for a 24 nm gate MESFET. The bias condition is 2.0 V on the drain and −2.5 V on the gate. The increase of the density on the inside of the gate depletion region edge and decrease of the density on the outside of the gate depletion region edge are evident. The same behavior appears at the interface of the active layer and substrate on the source end. This shows that barrier repulsion and penetration do occur. The modifications of the density distribution due to the quantum effects are as large as 4 per cent in the channel and about 8 per cent at the interface of the active layer and substrate in the source side. The equivalent interpretation to the modification of the density distribution is that the quantum corrections serve as a quantum potential which acts to smooth the actual potential, especially at potential barrier edges, which in turn smoothes the electron density distribution. From Fig. 18, one could observe that two factors make the quantum effect important: one is a sharp potential change which also results in a sharp density change, another is low electron thermal energy. We can see that the quantum effect at the interface of the active

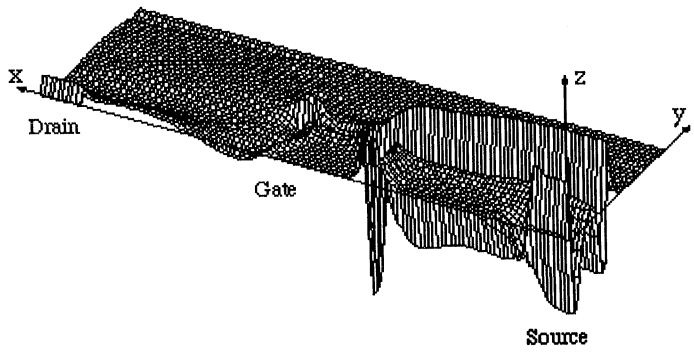

FIGURE 18 A 2-D plot of density difference between the results of with and without quantum correction for a 24 nm gate GaAs MESFET device.

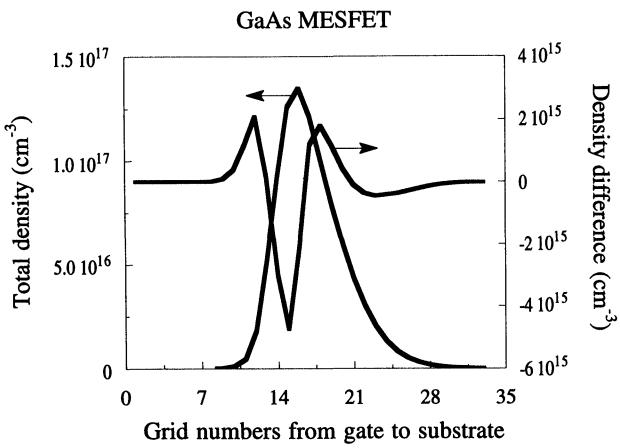

FIGURE 19 Density difference between the results of with and without quantum corrections across the channel for a 24 nm gate GaAs MESFET device.

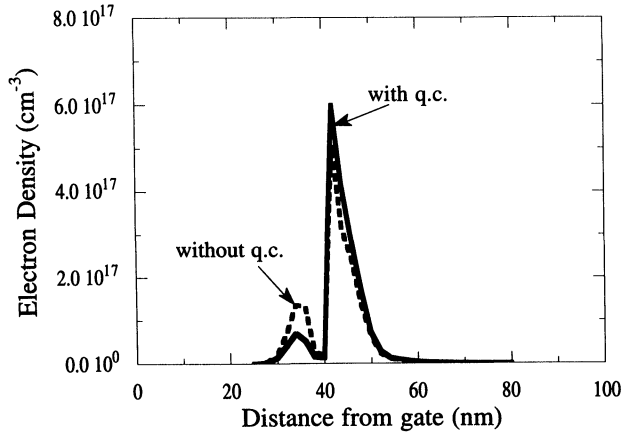

FIGURE 20 Electron density distribution across the conduction channel as a function of distance from the gate contact for a AlGaAs/GaAs HEMT.

layer and substrate decreases from source to drain and disappear at the drain side where the electron effective temperature is high, which reflects the correct physical phenomena.

In order to have more clear observation, we check the density distribution across the conduction channel under the gate for the three kind of devices. The electron density across the conduction channel for the 24 nm MESFET is illustrated in Fig. 19. Here, the total electron density and the density difference are plotted, under the same bias conditions as in Fig. 18. The interface of the active layer and substrate occurs at 39 nm from the gate. From the density difference curve, the density penetration toward the gate and substrate create two peaks, and repulsion from the gate and interface barriers results in the valley. The net effect is that the electron density distribution across the channel is broadened, which corresponds to a smoothing of the actual potential by the effective quantum potential. Following the same concept, one may expect that the same distribution behavior should be found in AlGaAs/GaAs HEMT and SiGe HEMT. However, this is not quite true. In Fig. 20 and 21, we plot the density distribution across the channel under the gate for a AlGaAs/GaAs HEMT and a SiGe HEMT, respectively. Because there is a larger density change along the conduction channel (from source to drain) of the 2-D electron gas, the quantum effect in the channel behaves differently from that for the MESFET case. As can be seen in Fig. 20, there are two peaks across the channel, one in the AlGaAs side that reflects a parasitic MESFET effect, and the expected one in the 2-D gas. For the SiGe HEMT, the density in the top SiGe layer is essentially depleted. The density peak is reduced in the AlGaAs (also in the SiGe top layer) and increased in the quantum wells when the quantum corrections are included. Because of the double heterojunction, the SiGe device has a fatter density

distribution. However, the major portion of the density is still within about 10 nm for both cases, and the density peak is always closer to the gate. The effect of the quantum corrections on the two peaks can be explained as follows: (1) the quantum corrections normal to the interface would soften the interface potential, consequently raising the potential minima in the AlGaAs or the SiGe (a natural effect due effectively to quantization in this potential minimum), and lowering the density, and, (2) the quantum corrections along the channel increase the peak electron density in the quantum well (a reduction in the depleting effect of the gate potential). The total effect of the summation of these two corrections shows that the second dominates in the quantum well. In the AlGaAs or the SiGe, the electron density distribution under the gate (along the channel) is relatively flat, when compared to the direction normal to the interface. The interface potential is broadened (towards the gate) and quantum corrections along the channel direction are small, and these effects result in a lower density peak. Due to the high electron density in the 2-D gas, the quantum correction along the channel direction is dominant, so the net effect is an increase of electron density, resulting in a fatter and higher peak. The relatively large change of the peak density in the AlGaAs is composed of both the local quantum correction and potential modification induced by the change of the electron density in the 2-D gas, which implies that real space transfer occurs. A detailed component analysis may be needed to further prove this behavior. Another obvious effect is that due to the higher density (from the higher doping in the SiGe) in the quantum well for the SiGe device, quantum effects introduce a larger density increase in the SiGe device than that in the AlGaAs/GaAs HEMT device. This large modification of the electron density in the SiGe device with relatively large gate length simulated here was not expected. However, by inspecting the density distribution along the channel, one can find that, although the gate is relatively long, a rapid density change occurs at the gate end close to the drain contact within a region much shorter than the gate length. In light of the quantum correction depending on the density change, the modification of the density by the quantum effects is understandable, since the electron density is high and the density change occurs in a short distance.

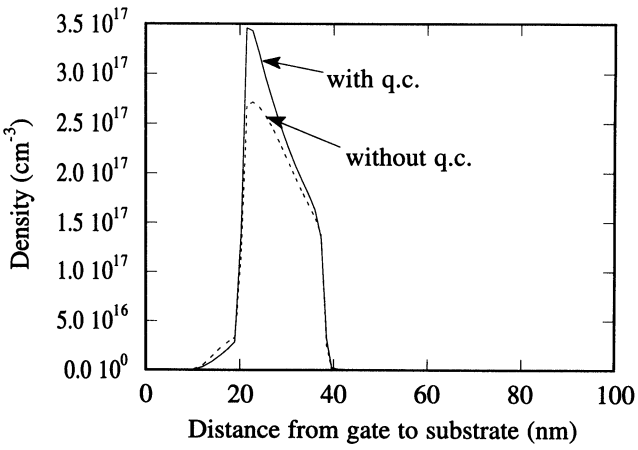

FIGURE 21 Electron density distribution across the conduction channel as a function of distance from the gate contact for a SiGe HEMT.

The modification of the density distribution by the quantum effects in turn changes the current density distribution, and consequently the total current flow through the conduction channel. Although the total current change for the MESFET is quite small (the device is not a quantum device), the change for AlGaAs/GaAs HEMT and SiGe HEMT is appreciably large, which results in a 10 per cent and
15 per cent total current increase, as we plotted in Fig. 22 and Fig. 23, respectively, with the quantum correction included in the simulation. This certainly suggests the importance in including quantum corrections in small device simulation.

Although the transport of electrons is still dominated by classical transport, i.e., most of the electrons pass the potential barrier under the gate by gaining higher energy, the quantum effects do contribute significant changes. The increase of the total current, especially the increase of the peak electron density in the channel for the HEMTs and the current density increase toward the gate due to the density penetration into the gate barrier for MESFETs, suggests tunneling processes (through the depletion region induced by the gate) may occur in the device operation. However, there are two facts suggesting that the current and electron density increase should not be interpreted as tunneling. It is well known that the tunneling current should exponentially decrease with an increase of the potential barrier. Thus, one expects that tunneling will become smaller as we increase the drain voltage or decrease (toward more negative values) the gate voltage. The depletion barrier will be widened in both cases, and will increase in amplitude, both of which should lead to a significant reduction in the tunneling current. This reduction is not observed in our present simulation. As an example, we plot the drain current against gate voltage in Fig. 24 for a

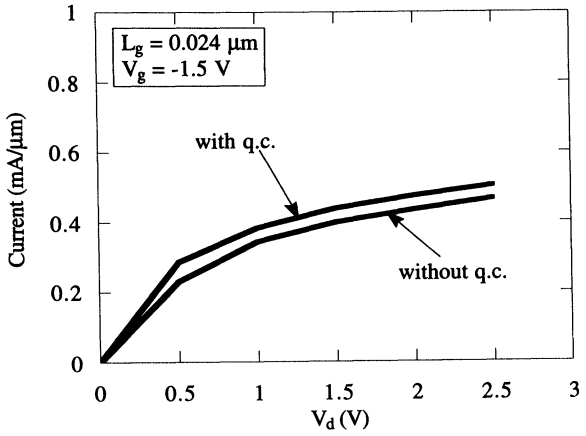

FIGURE 22 Drain output characteristics for a 24 nm gate AlGaAs/GaAs HEMT at $V_g = -1.5$ V. These curves illustrate the effects studied.

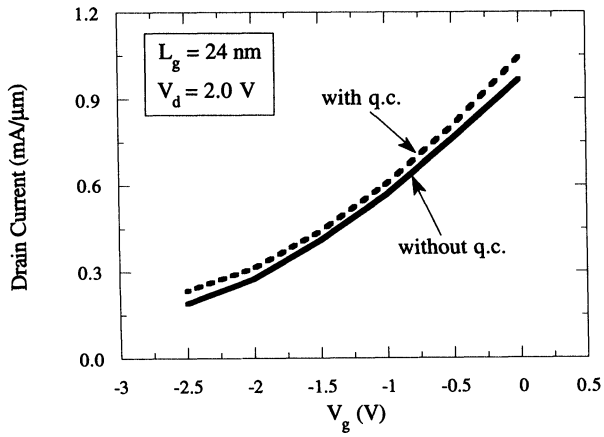

FIGURE 24 Drain current variation with gate voltage for a 24 nm gate AlGaAs/GaAs HEMT at $V_d = 2.0$ V. Two curves show the results with (dashed) and without (solid) quantum corrections included. Quantum effects are relatively insensitive to the gate bias, suggesting that the increase is not due to tunneling.

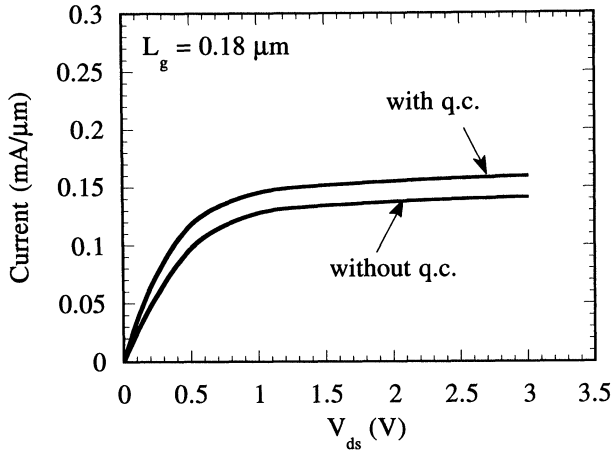

FIGURE 23 Drain output characteristics for a 0.18 μm gate SiGe HEMT at $V_g = 0.5$ V. These curves illustrate the effects studied.

these effects is obviously insensitive to both gate and drain bias. This conclusion differs from early suggestions based upon the experiments [5].

As the physical quantities (density, velocity, temperature ...) are solved self consistently, any change of one quantity is related to other quantities. In Fig. 25, we plot the velocity difference between the velocities with and without the quantum corrections for various space positions for a 24 nm GaAs MESFET. It may be seen that the peak velocity achieved (the value near the drain edge of the gate) is lowered by the quantum potential. The physics in the velocity change is that the quantum potential modifies the electric field distribution, and thus results in the change of the electron temperature

24 nm gate AlGaAs/GaAs HEMT, for simulations both with and without the quantum corrections. As can be seen, the current increase due to quantum effects is relatively insensitive to the gate voltage. A similar property can be found in Fig. 22, in which the current increase due to quantum effects is relatively insensitive to the drain voltage. These results lead us to the conclusion that, if there is any tunneling, its effect must be small. The reason the quantum effects are relatively insensitive to both the gate voltage and drain voltage is that the quantum effects that make major contribution to the increase of the total current primarily soften the gate depletion potential and give a higher electron density distribution along the channel. The current increase due to

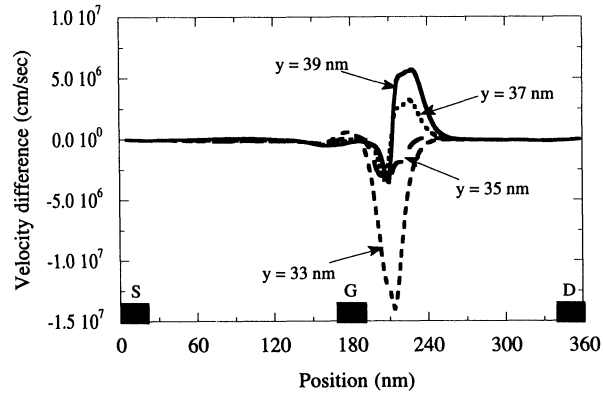

FIGURE 25 Longitudinal velocity difference in the channel of a 24 nm gate GaAs MESFET for different y positions. $a = 39$ nm, $V_d = 2$ V, $V_g = -1.5$ V.

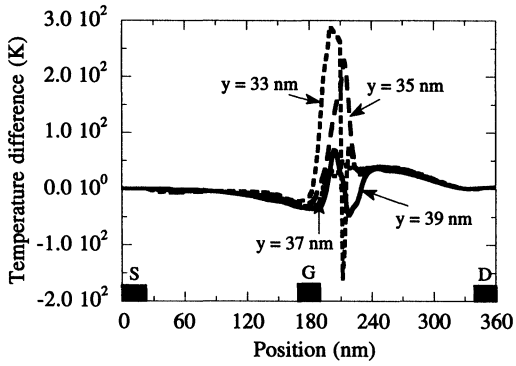

FIGURE 26 Temperature difference in the channel of a 24 nm gate GaAs MESFET for different y positions. a = 39 nm, V_d = 2 V, V_g = −1.5 V.

which subsequently causes the change of the relaxation times and the velocity. The quantum effect on the effective electron temperature has a kind of anti-symmetry with respect to the gate center. This is shown in Fig. 26. As a result of the smoothing effect of the quantum potential on the actual potential, the temperature is reduced around the source side of the gate. At the other side of the gate, due to the broadening of the electron accumulation hill, as for the curve at y = 39 nm, the temperature distribution is broadened to some extent. The major effect on this side is that the higher temperature is pushed towards the source direction. This leads to a reduction in the peak carrier velocity. On the other hand, there is a small increase of the velocity near the drain. This latter is a consequence of the fact that the electric field maintains a high value further from the gate.

VII. SUMMARY AND CONCLUSION

We present a set of 3-D quantum hydrodynamic equations developed from the Wigner function equation-of-motion. Explicit quantum corrections are build into these equations by using quantum mechanical expressions of the moments of the Wigner function. This set of equations returns to full classical hydrodynamic equations as $\hbar$ goes to zero, and retain exact form of the zero temperature equations correctly as temperature goes to zero. Although caution needs to be taken for the lack of ensemble statistics in the derivations, the quantum correction forms are essentially the same as that obtained from other methods, and the use of the equation does give a correct physical effect.

The relation of the mathematical form of the quantum corrections to its physical implication is properly interpreted. The quantum potential energy is introduced from the change of the density distribution, which is also implicitly related to the actual potential profile. Quantum effects tend to smooth the actual potential and density changes. A clear picture is that it tends to reduce the peak and raise the valley of a density distribution. Due to these changes, device operation certainly will be affected.

We have performed a simulation of various small semiconductor devices by using this set of equations. The simulation reveals good device characteristics such as transconductance for all devices simulated here. Velocity overshoot observed in the simulation is important in achieving large transconductance for these devices. Our simulation predicts that very large velocity overshoot can be obtained for ultra-short channel devices, if a low field can be maintained from the source until very close to the gate, since under this condition, the electron could retain a low temperature and thus a higher mobility up to the gate center, where a very high electric field accelerates the electron to very high velocity.

As expected, quantum penetration and repulsion occurs in all simulated devices. The effect causes changes in both density distribution and total current flow. With the inclusion of quantum corrections, the density change in the 24 nm MESFET ranges from 4 per cent to 8 per cent, but the total current is essentially not changed, for the change of the density across the channel is averaged to small. The quantum effects become strong along the 2-D electron channel under the gate for HEMT devices. As the total quantum corrections consist of the effects along the channel direction and perpendicular to the channel direction, the effect along the channel direction dominates under the gate, which is the result of higher density distribution away from the minimum under the gate. Due to the strong effect, the total current flow increases by 10 per cent for a 24 nm gate AlGaAs/GaAs HEMT and 15 per cent for a 0.18 μm gate SiGe HEMT. These results suggest that the inclusion of quantum corrections in deep submicron and ultra-submicron devices is necessary.

Acknowledgments

The authors give special thanks to Dr. A.M. Kriman, Dr. C. Ringhofer, Dr. H. Grubin, and Dr. C. Gardner for useful discussions.

References

[1] D.K. Ferry, "Lateral surface superlattice and the future of ULSI microelectronics," in *Granular Nanoelectronics*, Ed. by D.K. Ferry, J.R. Barker, and C. Jacoboni. Plenum, New York, 1991.

[2] G. Baccarani, M.R. Wordeman, and R.H. Dennard, "Generalized scaling theory and its application to a 1/4 micrometer MOSFET design," *IEEE Trans. Electron Devices*, vol. ED-31, no. 4, pp. 452–462, 1984.

[3] J.R. Barker and D.K. Ferry, "On the physics and modeling of small semiconductor devices-I," *Solid-State Electron.*, vol. 23, no. 6, pp. 519–530, 1980.

[4] J.R. Barker and D.K. Ferry, "On the physics and modeling of small semiconductor devices-II," *Solid-State Electron.*, vol. 23, no. 6, pp. 531–544, 1980.

[5] J. Han, D.K. Ferry, and P. Newman, "Ultra-submicron gate AlGaAs/GaAs HEMTs," *IEEE Electron Device Lett.*, vol. EDL-11, no. 5, pp. 209–211, 1990.

[6] A. Ishibashi, K. Funato, and Y. Mori, "Heterointerface field effect transistor with 200 Å-long gate," *Jpn. J. Appl. Physics*, vol. 27, no. 12, pp. 2382–2384, 1988.

[7] J.M. Ryan, J. Han, A.M. Kriman, D.K. Ferry, and P. Newman, "Overshoot Saturation in Ultra-short Channel FETs Due to Minimum Acceleration Length," in *Nanostructure Physics and Fabrication*, Ed. by M.A. Reed and W.P. Kirk (Academic Press, New York, 1989) pp. 195–200.

[8] D.R. Allee, P.R. de la Houssaye, D.G. Schlom, J.S. Harris, and R.F.W. Pease, "Sub-100-nm gate length GaAs metal-semiconductor field-effect transistors and modulation-doped field-effect transistors fabricated by a combination of molecular-beam epitaxy and electron-beam lithography," *J. Vac. Sci. Technol.* B, vol. 6, no. 1, pp. 328–332, 1988.

[9] D.K. Ferry, *Semiconductors*. Macmillan Publishing Company, New York, 1991.

[10] However, see e.g., M.V. Fischetti, "Monte Carlo simulation of transport in technologically significant semiconductors of the diamond and zinc-blende structures—part I: Homogeneous transport," *IEEE Trans. Electron Devices*, vol. ED-38, no. 3, pp. 634–649, 1991. M.V. Fischetti and S.E. Laux, "Monte Carlo simulation of transport in technologically significant semiconductors of the diamond and zinc-blende structures—part II: Submicrometer MOSFET's," *IEEE Trans. Electron Devices*, vol. ED-38, no. 3, pp. 650–660, 1991.

[11] H. Fröhlich and B.V. Paranjape, "Dielectric breakdown in solids," *Proc. Phys. Soc. B*, vol. 69, no. 433B, pp. 21–32, 1956.

[12] R. Stratton, "The influence of interelectronic collisions on conduction and breakdown in covalent semiconductors," *Proc. Roy. Soc. A*, vol. 242, no. 1230, pp. 355–373, 1957.

[13] R. Stratton, "The influence of interelectronic collisions on conduction and breakdown in polar crystals," *Proc. Roy. Soc. A*, vol. 246, no. 1246, pp. 406–423, 1958.

[14] K. Bløtekjær, "High-frequency conductivity, carrier waves, and acoustic amplification in drifted semiconductor plasmas," *Ericsson Technics*, no. 2, pp. 125–183, 1966.

[15] R. Bosch and H.W. Thim, "Computer simulation of transferred electron devices using the displaced Maxwellian approach," *IEEE Trans. Electron Devices*, vol. ED-21, no. 1, pp. 16–35, 1974.

[16] U. Ravaioli and D.K. Ferry, "MODFET Ensemble Monte-Carlo Model Including the Quasi-Two-Dimensional Electron Gas," *IEEE Trans. Electron Devices*, vol. 33, pp. 677–681, 1986.

[17] G.J. Iafrate, H.L. Grubin, and D.K. Ferry, "Utilization of quantum distribution functions for ultra-submicron device transport," *J. Physique*, Colloq. C-10, vol. 42, no. 10, pp. 307–312, 1981.

[18] H.L. Grubin, D.K. Ferry, G.J. Iafrate, and J.R. Barker, "The numerical physics of micron-length and submicron-length semiconductor devices," in *VLSI Electronics*, Ed. by N.G. Einspruch. Academic Press, New York, 1983.

[19] J-R. Zhou, and D.K. Ferry, "Simulation of ultra-small GaAs MESFET using quantum moment equations," *IEEE Trans. Electron Devices*, vol. 39, No. 3, pp. 473–478, 1992.

[20] J-R. Zhou, and D.K. Ferry, "Simulation of ultra-small GaAs MESFET using quantum moment equations II: Velocity overshoot," in *IEEE Trans. Electron Devices*, vol. 39, No. 8, pp. 1793–1796, 1992.

[21] J.-R. Zhou and D.K. Ferry, "Simulation of quantum effects in ultrasmall HEMT devices," IEEE Trans. Electron Devices, vol. 40, pp. 421–427, 1993.

[22] J.E. Moyal, "Quantum mechanics as a statistical theory," *Proc. Cambridge Phil. Soc.*, vol. 45, pp. 99–124, 1949.

[23] E. Wigner, "On the quantum correction for thermodynamic equilibrium," *Phys. Rev.*, vol. 40, no. 5, pp. 749–754, 1932.

[24] N.C. Kluksdahl, A.M. Kriman, D.K. Ferry, and C. Ringhofer, "Self-consistent study of the resonant-tunneling diode," *Phys. Rev. B*, vol. 39, no. 11, pp. 7720–7735, 1989.

[25] W.R. Frensley, "Wigner-function model of a resonant-tunneling semiconductor device," *Phys. Rev. B*, vol. 36, no. 3, pp. 1570–1580, 1987.

[26] J-R. Zhou, A.M. Kriman and D.K. Ferry, "The conditions of device simulation using full hydrodynamic equations," in *Computational Electronics*, Ed. by K. Hess, J.P. Leburton, and U. Ravaioli. Kluwer Academic Publishers, Boston, pp. 63–66, 1991.

[27] H.L. Grubin and J.P. Kreskovsky, "Quantum moment balance equations and resonant tunneling structures," *Sol.-State Electron.*, vol. 32, no. 12, pp. 1071–1075, 1989.

[28] J-R. Zhou, *Ph.D. Dissertation, Arizona State University*, 1991.

[29] T.-W. Tang, S. Ramaswamy, and J. Nam, "An improved hydrodynamic transport model for silicon," IEEE Trans. Electron Devices, vol. 40, No. 8, pp. 1469–1477, 1993.

[30] L. Reggiani, private communication.

[31] M.G. Ancona, and G.J. Iafrate, "Quantum corrections to the equation of state of an electron gas in a semiconductor," *Phys. Rev. B*, vol. 39, pp. 9536–9540, 1989.

[32] H.L. Grubin, T.R. Govindan, J.P. Kreskovsky, and M.A. Stroscio, *Sol.-State Electron.*, to be published.

[33] C.L. Gardner, "The quantum hydrodynamic model for semiconductor devices," to be published.

[34] I.B. Levinson, "Translational invariance in uniform fields and the equation for the density matrix in the Wigner representation," *Sov. Phys.-JETP*, vol. 30, pp. 362–367, 1970.

[35] D.K. Ferry, and J-R. Zhou, *Phys. Rev. B*, in press.

[36] P.J. Roache, *Computational Fluid Dynamics*. Hermosa, Albuquerque, NM, 1982.

[37] M.S. Shur, "Influence of nonuniform field distribution on frequency limits of field-effect transistors," *Electron. Lett.*, vol. 12, no. 23, p. 615–616, 1976.

[38] S.C. Eisenstat, "Efficient implementation of a class of preconditioned conjugate gradient methods," *SIAM J. Sci. Stat. Comput.*, vol. 2, no. 1, pp. 1–4, 1981.

[39] H.A. van der Vorst, "Iterative solution methods for certain sparse linear systems with a non-symmetric matrix arising from PDE-problems," *J. Comput. Phys.*, vol. 44, no. 1, p. 1–19, 1981.

[40] T. Wada and R.L.M. Dang, "Modification of ICCG method for application to semiconductor device simulator," *Electron. Lett.*, vol. 18, no. 6, p. 265–266, 1982.

[41] T. Shawki, G. Salmer, and O. El-Sayed, "2-D Simulation of Degenerate Hot Electron Transport in MODFET's Including DX Center Trapping," *IEEE Trans. Computer-Added Design.*, vol. 9, no. 11, pp. 1150–1163, 1990.

[42] A.M. Kriman, J-R Zhou, and D.K. Ferry, "Statistical properties of a hard-wall potentials," *Physics Lett. A*, vol. 138, no. 1, pp. 8–12, 1989.

[43] D.K. Ferry and R.O. Grondin, Physics of Sub-Micron Devices, (Plenum Press, New York, 1992) pp. 126–129.

[44] J.R. Hauser, "Characteristics of junction field effect devices with channel length-to-width ratios," *Sol.-State Electron.*, vol. 10, pp. 577–587, 1967.

[45] H. Miyata, T. Yamada, and D.K. Ferry, "Electron transport property of a strained Si layer on a relaxed $Si_{1-x}Ge_x$ substrate by Monte Carlo simulation," *Appl. Phys. Lett.*, Vol. 62, No. 21, pp. 2661–2663, 1993.

[46] D.K. Ferry and J.R. Barker, "On the use of Monte Carlo techniques for the calculation of transient dynamic response in semiconductors," in Phys. Stat. Sol. (b) 100, pp. 683–689, 1980.

[47] See, e.g., E. Constant, "Modeling of Sub-Micron Devices," in *The Physics of Submicron Semiconductor Devices*, Ed. by H.L. Grubin, D.K. Ferry, and C. Jacoboni (Plenum Press, New York, 1988) pp. 1–36.

Biographies

JINGRONG ZHOU (S'87–M'92) received the B.S. degree in applied physics and M.S. degrees in electrical engineering from China, and Ph.D. degree in electrical engineering from Arizona State University in 1982, 1985 and 1991 respectively.

He worked on optical fiber sensor research and optical fiber theory from 1985 to 1986. He is currently a research analyst in the Center for Solid State Electronics Research at Arizona State University, working on the Modeling and simulation of semiconductor devices. His research interests include semiconductor device and process simulation, device characterization and device physics, semiconductor transport theory and quantum transport, scientific visualization.

Dr. Zhou is a member of Sigma Xi, the Scientific Research Society.

DAVID K. FERRY is Regents' Professor of Engineering at Arizona State University. Prior to joining ASU in 1983, he was a faculty member at Colorado State University (1977–83), Texas Tech University (1967–73), and worked at the Office of Naval Research (1973–77). He received his bachelors and masters degrees in electrical engineering from Texas Technological College in 1962 and 1963, respectively, and the Ph.D. degree in electrical engineering from the University of Texas in 1966. He was a National Science Foundation postdoctoral fellow at the University of Vienna during 1966–67. His research encompasses transport physics and modeling of submicron semiconductor devices, particularly the inclusion of quantum effects, and electron-beam lithography for ultra-submicron devices, where his group has fabricated HEMTs with gate lengths as short as 25 nm. He is a fellow of both the American Physical Society and the Institute of Electrical and Electronics Engineers. He is a member of Sigma Xi, Phi Kappa Phi, Golden Key, and Eta Kappa Nu, and is listed in Who's Who in America. He is the author, or co-author of more than 350 refereed publications including three textbooks (two others are in preparation).