Macroscopic Quantum Carrier Transport Modeling

Zhiping Yu, Robert W. Dutton, and Daniel W. Yergeau
Center for Integrated Systems
Stanford University, Stanford, CA 94305, USA
yu@ee.stanford.edu

Mario G. Ancona
Naval Research Laboratory, Washington, D.C. 20375, USA

Abstract

It has been established [1]-[4] that the density gradient (DG) model is the lowest order, in terms of $\hbar$, approximation of the Wigner function approach to including quantum mechanical (QM) effects in carrier transport. In this paper, we report a five-equation PDE system (reduced to three-equation at thermal equilibrium) which preserves the numerical stability of classical drift-diffusion (DD) model, yet faithfully manifests QM corrections. Tunneling through the gate oxide (or barrier region) is modeled by ballistic transport with each type of carrier (electrons or holes) further split into forward and backward moving species and solved for separately. The entire device, including semiconductor and barrier regions, is solved self-consistently. Terminal characteristics, either *dc* or small signal *ac*, for realistic, multi-dimensional device structures can be simulated using this model. An SOI device example is simulated and the comparison with microscopic (Schrödinger/Poisson) results is excellent. A DG prediction of a dipole in the poly gate near the poly/gate-oxide interface is also confirmed by microscopic simulation. Both *I-V* and *C-V* for MOS devices including SOI are shown.

1 Introduction

Quantum mechanical (QM) effects in semiconductor devices are manifestations of the wave nature of highly confined electrons. For scaled MOSFETs, two QM effects are most prominent:

1. Quantum confinement in the inversion or accumulation layer in the substrate adjacent to the Si-SiO_2 interface.

2. Tunneling through the thin gate oxide (for thicknesses below approximately 2nm).

An "exact" analysis of these phenomena requires quantum mechanics. In its simplest form, a microscopic treatment of the confinement effects involves an equilibrium problem in one-dimension and can be solved using the Hartree approximation. This is the conventional approach; one finds that the energy bands split into discrete sub-bands with the wave functions associated with each energy level spread over the breadth of the well. Because of the large chemical potential barrier at the Si-SiO_2 interface, the wave functions essentially vanish at that interface; the peak carrier concentration is thus located away from the substrate surface. Qualitatively, all of this remains true in a real device but the details change because the device is actually three-dimensional, and the degree of confinement varies with position along the channel.

The microscopic treatment of gate tunneling is even more problematic because it is in essence a non-equilibrium problem. One-dimensional, one-electron treatments have long been used but these lack self-consistency. Much more sophisticated one-dimensional treatments such as those based on non-equilibrium Green's functions [6] capture the physics quite well for one-dimension but are too computationally demanding to be extended to multiple dimensions and to be used for everyday engineering analysis.

For many important problems there is a way out of this dilemma that comes with the recognition that a full-scale treatment of the quantum mechanics is overkill. For MOS scaling problems, quantum interference phenomena are typically not important. Furthermore, in many device situations, a large number of the sub-bands are filled. Therefore, the details of sub-band structure have little consequence. This is the motivation for developing a macroscopic transport model with the lowest-order QM corrections. One such approach is the so-called density gradient (DG) model [1].

2 The evolution of DG model from Wigner function approach

The DG model can be developed using either macroscopic arguments [2] or from quantum mechanics [1]. The macroscopic approach introduces lowest-order effects of non-locality by having the internal energy of the electron gas depend not only on the density of electrons, but also on the gradient of the density. Alternatively, one can derive the DG model from the Wigner function, which is the probability density function of particles in real and momentum space. The Wigner function is constructed from the wave function as [7]

$$f(\boldsymbol{r},\boldsymbol{p}) = \frac{1}{(\hbar\pi)^l}\int_{-\infty}^{\infty}\Psi^*(\boldsymbol{r}+\boldsymbol{r}')\Psi(\boldsymbol{r}-\boldsymbol{r}')e^{2i\boldsymbol{p}\cdot\boldsymbol{r}'/\hbar}d\boldsymbol{r}' \tag{1}$$

where $\boldsymbol{p}$ is the momentum, Ψ is the wave function, and l is the dimensionality, e.g., for 3D, $l = 3$. One can view the above expression as a partial Fourier transform of the density matrix.

The Wigner function is real, but not necessarily positive, which poses difficulties if it is used directly to find carrier concentration, etc. To derive the DG model from the Wigner function, one notes that if Ψ is the solution to the Schrödinger equation then the Wigner function must satisfy

$$\frac{\partial f}{\partial t} + \boldsymbol{v}\cdot\nabla_{\boldsymbol{r}} f - \frac{1}{\hbar}\nabla_{\boldsymbol{r}} U\cdot\nabla_{\boldsymbol{k}} f + \sum_{\alpha=1}^{\infty}\frac{(-1)^{\alpha+1}}{\hbar 4^{\alpha}(2\alpha+1)!}(\nabla_{\boldsymbol{r}}\cdot\nabla_{\boldsymbol{k}})^{2\alpha+1}Uf = \left.\frac{\partial f}{\partial t}\right|_C \tag{2}$$

where U is the potential, and all other symbols have their conventional meaning. By utilizing only the lowest order term of the series expansion in the above equation, one obtains

$$\frac{\partial f}{\partial t} + \boldsymbol{v}\cdot\nabla_{\boldsymbol{r}} f - \frac{1}{\hbar}\left[\nabla_{\boldsymbol{r}} U\cdot\nabla_{\boldsymbol{k}} f - \frac{1}{24}(\nabla_{\boldsymbol{r}}\cdot\nabla_{\boldsymbol{k}})^3 Uf\right] = \left.\frac{\partial f}{\partial t}\right|_C \tag{3}$$

Or under further assumptions such as isotropic effective mass, this equation can be recast as

$$\frac{\partial f}{\partial t} + \frac{\hbar}{m^*}\boldsymbol{k}\cdot\nabla_{\boldsymbol{r}} f - \frac{1}{\hbar}\nabla_{\boldsymbol{r}}\left(U - \frac{\hbar^2}{8m^*}\nabla_{\boldsymbol{r}}^2 \ln n\right)\cdot\nabla_{\boldsymbol{k}} f = \left.\frac{\partial f}{\partial t}\right|_C \tag{4}$$

This equation is called the quantum corrected Boltzmann transport equation (QBTE) because of its close resemblance to the conventional Boltzmann transport equation (BTE). In fact, by comparing the above equation to the conventional BTE and taking the zeroth moment of the equation, one gets a carrier continuity equation which is identical to the classical drift-diffusion (DD) model, except that the carrier flux now has the form of

$$\boldsymbol{F}_n = D_n\nabla n - \mu_n n\nabla\psi - 2\mu_n n b_n \nabla\left(\frac{\nabla^2\sqrt{n}}{\sqrt{n}}\right) \tag{5}$$

where

$$b_n = \hbar^2/(4lqm_n^*) \tag{6}$$

where l can again be viewed as the space dimensionality (although it needs not be an integer value). The additional term in the drift part of the model is due to the quantum correction and is sometimes lumped into the potential with the term

$$2b_n\frac{\nabla^2\sqrt{n}}{\sqrt{n}} \tag{7}$$

being the so-called quantum potential. Note that the unusual $\sqrt{n}$ enters this expression by virtue of the transformation

$$-\frac{1}{2}\frac{1}{n^2}\nabla n\cdot\nabla n + \frac{1}{n}\nabla^2 n = 2\frac{\nabla^2\sqrt{n}}{\sqrt{n}} \tag{8}$$

3 DG model formulation

3.1 Five-equation set in semiconductor regions

The quantum correction of DG theory obviously raises the order of the differential system, and in order to avoid higher order (no higher than 2) spatial mathematical operators, the carrier continuity equation (for either electrons or holes) is split into two equations with two independent variables, n and ϕ_n. Thus, the conventional DD set of semiconductor equations, which include Poisson's equation and the carrier continuity equations, is expanded from three to five equations, all being second order PDEs. The following five-equation PDE system is solved for in terms of the five independent variables, ψ, $\sqrt{n}$ (or n), $\sqrt{p}$ (or p), ϕ_n, and ϕ_p:

$$\nabla\cdot(\epsilon\nabla\psi) + q(p - n + N_D^+ - N_A^-) = 0 \tag{9}$$

$$\nabla\cdot(b_n\nabla\sqrt{n}) + \frac{\sqrt{n}}{2}\left(\psi - \frac{kT}{q}\ln\frac{n}{n_i} - \phi_n\right) = 0 \tag{10}$$

$$\nabla\cdot(b_p\nabla\sqrt{p}) + \frac{\sqrt{p}}{2}\left(\psi + \frac{kT}{q}\ln\frac{p}{n_i} - \phi_p\right) = 0 \tag{11}$$

$$\nabla \cdot (\mu_n n \nabla \phi_n) + \frac{\partial n}{\partial t} + r = 0 \quad (12)$$

$$\nabla \cdot (\mu_p p \nabla \phi_p) - \frac{\partial p}{\partial t} - r = 0 \quad (13)$$

where r is the net electron-hole pair recombination rate. Even though Maxwell-Boltzmann (MB) statistics are assumed here, Fermi-Dirac (FD) can easily be implemented. The boundary conditions for the above system of equations are the same as in DD except that there are explicit boundary conditions that are associated with Eqs. (10-11).

There are two special cases in which the above general equation set can be simplified:

1. Thermal equilibrium. Since the quasi-Fermi levels are known (being zero), Eqs. (12-13) can be dropped and the system reduces to a three-equation set to be solved for the carrier density and potential distributions.

2. Classical case. With $b_n = b_p = 0$ (or $\hbar \to 0$). Eqs. (10-13) can be combined to reduce to the electron/hole continuity equations of DD theory.

The above equation set is readily solved without stabilization by discrete methods, e.g. finite differences, finite volume, or finite element. A noticeable difference from the classical DD model, however, is that the pn product at the thermal equilibrium is no longer equal to n_i^2, i.e., $n_0 p_0 \neq n_i^2$ in the bulk of the semiconductor region. This requires the use of different models for recombination effects than those used with classical DD [8].

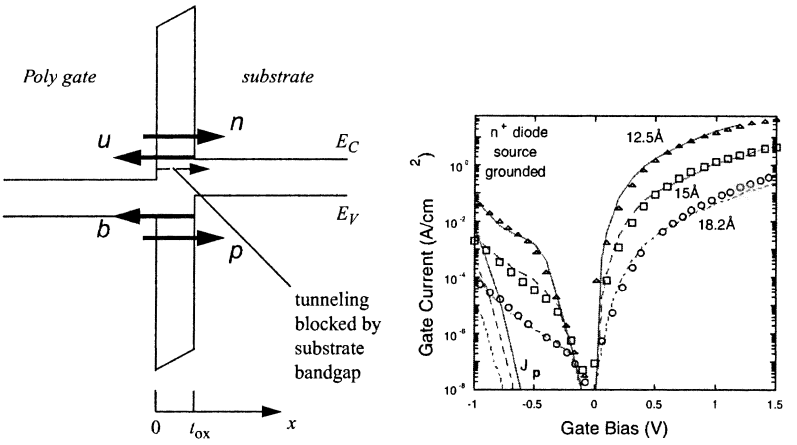
3.2 Equation set in barrier region for tunneling

If tunneling is permitted in barrier regions such as the gate oxide, then one must solve carrier equations in these regions too. We consider only elastic tunneling and therefore the mobility vanishes and the transport is ballistic. Elastic tunneling further implies that carriers (electrons or holes) entering from either end of the barrier do not mix and must therefore be treated separately as shown in Fig. 1 (a). There are thus four independent carrier concentrations to be solved for in the barrier region: two electron populations, n and u, and two hole populations, p and b. Ballistic transport implies that inertia is important. The variables of the problem are the four carrier densities, their respective quasi-Fermi levels and velocities, together with the electrostatic potential. Thus, the number of unknowns in the barrier region is thirteen in principle, for which we must have thirteen equations. But in the one-dimensional case, various approximations can vastly reduce the complexity of the system as well as the number of equations to solve. For details, refer to [8].

The electrostatics is governed by Poisson's equation (assuming no fixed charge in the oxide) which is

$$\nabla \cdot (-\epsilon \nabla \psi) = q(p + b - n - u) \quad (14)$$

where ψ is assumed continuous across the oxide/silicon interface.



(a) Schematic for four tunneling carrier species

(b) Comparison of tunneling data and simulation

Fig. 1: Modeling and simulation of tunneling in DG framework.

The carrier flux in the barrier region is expressed as

$$\boldsymbol{F}_n = n\boldsymbol{v}_n \tag{15}$$

where $\boldsymbol{v}_n$ is the carrier velocity for forward electrons. The equations solved for n and $\boldsymbol{v}_n$ are: Eq. (10), with the density dependent term, $kT/q \ln(n/n_i)$, replaced by a general dependence of n, for n and the carrier continuity equation for $\boldsymbol{v}_n$,

$$\frac{\partial n}{\partial t} = -\nabla \cdot (n\boldsymbol{v}_n) \tag{16}$$

The equation for ϕ_n is the following equation of motion,

$$m_n^* \frac{d\boldsymbol{v}_n}{dt} \equiv m_n^* \left[\frac{\partial \boldsymbol{v}_n}{\partial t} + (\boldsymbol{v}_n \cdot \nabla)\boldsymbol{v}_n \right] = q\nabla\phi_n \tag{17}$$

Under steady state conditions the equations for the forward-going electrons are the following three equations to be solved for the three independent variables: n, ϕ_n, and $\boldsymbol{v}_n$,

$$\nabla \cdot (b_n \nabla\sqrt{n}) + \frac{\sqrt{n}}{2}\left[\psi + g(n) - \phi_n\right] = 0 \tag{18}$$

$$m_n^*(\boldsymbol{v}_n \cdot \nabla)\boldsymbol{v}_n = q\nabla\phi_n \tag{19}$$

$$\nabla \cdot (n\boldsymbol{v}_n) = 0 \tag{20}$$

where in Eq. (18) we have left the density-dependent part of the equation of state unspecified, using a generic function form $g(n)$ instead. It is probably the best choice to use an adiabatic expression (rather than the MB or FD forms).

To derive the boundary conditions for n and ϕ_n, we illustrate using the one-dimensional case with a barrier extending from 0 to t_{ox} and carriers launched into the barrier the directions shown in Fig. 1. The boundary conditions on these forward-electrons are

$$n(0^+) = n(0^-) \tag{21}$$

$$\left.\frac{dn}{dx}\right|_{x=t_{ox}^-} = 0 \tag{22}$$

$$\phi_n(0^+) = \phi_n(0^-) \tag{23}$$

$$\left.\frac{d\phi_n}{dx}\right|_{x=t_{ox}^-} = 0 \tag{24}$$

$$n(0^-)\mu_n(0^-)\left.\frac{d\phi_n}{dx}\right|_{x=0^-} = n(0^+)v_n(0^+) + u(0^+)v_u(0^+) \tag{25}$$

$$n(t_{ox}^-)v_n(t_{ox}^-) = n(t_{ox}^-)v_{n,TRV}(t_{ox}) \tag{26}$$

where the subscript, *TRV*, indicates the tunneling recombination velocity. At the downstream end ($x = t_{ox}^-$), the normal component of the gradient of both carrier concentration and quasi-Fermi level is assumed to be zero [8].

Equation (26) involves the currents in the barrier region and as discussed in [5] these are modeled by tunneling recombination velocity conditions that describe how the carriers that are accelerated across the barrier by the electric field recombine with the equilibrium carrier population at the downstream end. In the above equations, Eq. (21) is simply a statement that carrier concentration is continuous across the material interface for carriers at the upstream end (i.e., $n(0^+)$). Eq. (23) states that the quasi-Fermi level is continuous across the upstream interface. Eq. (25) refers to the continuity condition of the carrier flux across the interface.

3.3 Interfacial boundary conditions in multi-layer structures

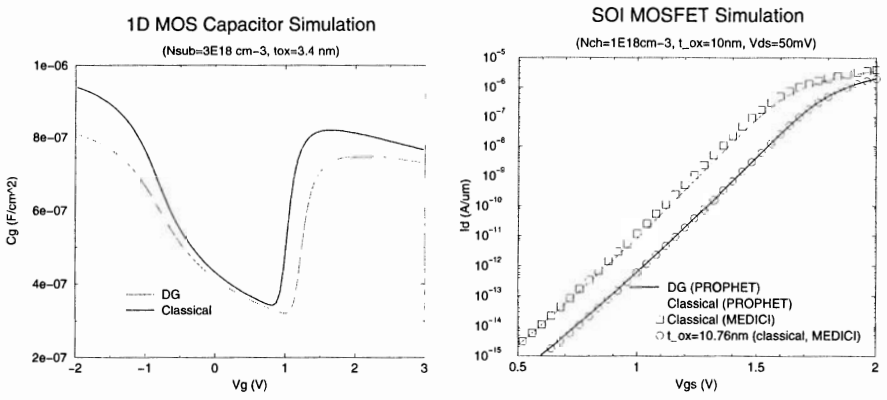
The proper specification of boundary conditions (BC) at material interfaces such as the Si/SiO_2 interface is critical to a robust numerical implementation of the DG model. For problems without tunneling, these interfacial BCs are

1. Continuity of ψ across the interface.
2. Vanishingly small carrier densities on the semiconductor side of the semiconductor/insulator interface (in practice a small but non-zero value is specified.)
3. Zero normal component of $\nabla\phi_n$ and $\nabla\phi_p$ across the semiconductor/insulator interface.

4 Implementation and examples

4.1 Model implementation

The above five-equation system has been implemented in PROPHET, a general PDE solver that provided user-specification of PDEs and boundary conditions via a high-



(a) Simulated *C-V* plot for a MOS capacitor using both DD and DG models.

(b) Simulated *I-V* curve for an SOI MOSFET using both classical and DG models. They achieve the same *I-V* plot, the classical model has to use the "electrical" gate oxide which is 0.76 nm thicker than the physical thickness.

Fig. 2: Simulated *C-V* for a MOS capacitor and *I-V* for an SOI MOSFET.

level scripts language, and simulations have been performed for various 1/2/3D MOS structures. Full Newton-Raphson iterations are used to obtain the solution, providing the Jacobian necessary for frequency-dependent small signal *ac* analysis. The convergence rate is acceptable.

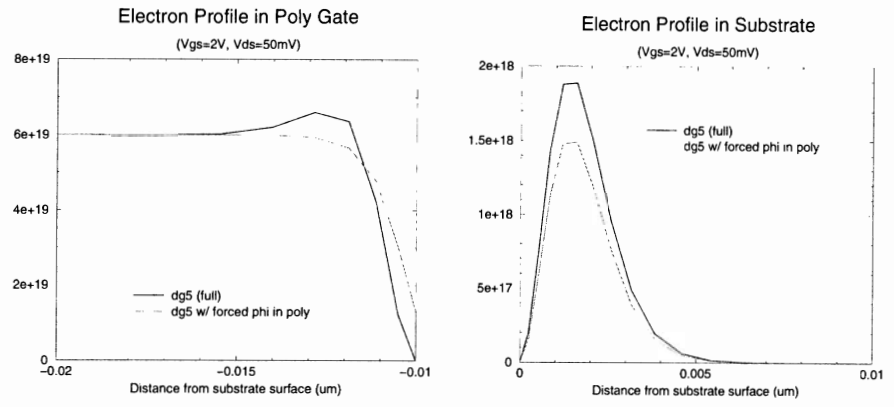
At the present time, tunneling current through a thin oxide has been demonstrated in a separate code [8] and the accuracy of the simulation is demonstrated in Fig. 1 (b).

4.2 Simulation examples

Two simulation examples are provided: quasi-static *ac* analysis of a 1D MOS capacitor and *dc* analysis of a 2D SOI MOSFET with the thickness of the silicon thin film varying from 3 to 20 nm and above. The latter simulation results are also compared to Poisson/Schrödinger equation solutions to establish their accuracy.

Fig. 2 (a) shows a simulated *C-V* curve for a MOS capacitor, comparing the classical DD model with the quantum corrected DG model. The DG model predicts a higher threshold voltage and a lower capacitance in both the accumulation and inversion regions. The *C-V* curve is smooth including in the flat band region where other simulations techniques often have difficulty in achieving convergence due to model singularities.

The second example analyzes an SOI MOSFET. Fig. 2 (b), shows the comparison of *I-V* characteristics. To achieve a similar drain current using conventional DD solutions, the "effective" gate oxide thickness would be 7.6 Å thicker than that of the physical one. This order of t_{ox} "correction" is in agreement with other observations, e.g. [9].



(a) Electron distribution in poly gate

(b) Electron distribution in substrate

Fig. 3: Simulated electron distribution in the poly gate and channel in the linear bias region ($V_{ds} = 50\,\text{mV}$ and $V_{gs} = 2\,\text{V}$) for an SOI MOSFET. Comparison is made between applying DG in all device regions and applying DG only to the silicon film/substrate.

Fig. 3 shows the electron distribution in the poly gate region and in the channel for the DG solutions. One distinctive feature of the simulation results is that there is a dipole in the poly gate region near the poly/oxide interface. This dipole is a result of the physical requirement of quasi-charge neutrality under flat band condition and the repulsion effect of the large SiO_2 barrier. The negative charge of the dipole is due to the pile up of electrons over the background doping. The existence of this dipole has been independently verified by NEMO [6], a program solving Schrödinger equation consistently. Finally, in Fig. 4 the simulated electron profile in the active region for silicon film thickness from 5 to 20 nm is compared with those from SCHRED, a Poisson and Schrödinger equation solver written by Professor D. Vasileska of Arizona State University. The agreement is excellent. It is noteworthy that in order to get a smooth distribution of carriers, SCHRED has to include more than thirty subbands.

Acknowledgment: Authors are grateful to contribution from Dan Connelly of Acorn Technologies. Software prototyping was provided for this research through consortium efforts supported by Texas Instruments and Agere.

References

[1] M.G. Ancona and G.J. Iafrate, "Quantum correction to the equation of state of an electron gas in a semiconductor," p. 9536, *Physical Review B*, vol. 39, no. 13, 1 May 1989.

[2] M.G. Ancona and H.F. Tiersten, "Macroscopic physics of the silicon inversion layer," p. 7959, *Physical Review B*, vol. 35, 1987.

[3] H. Tsuchiya and T. Miyoshi, "Quantum transport modeling of ultrasmall semiconductor devices," p. 880, *IEICE Tr. Electronics*, vol. E82-C, no. 6, June 1999.

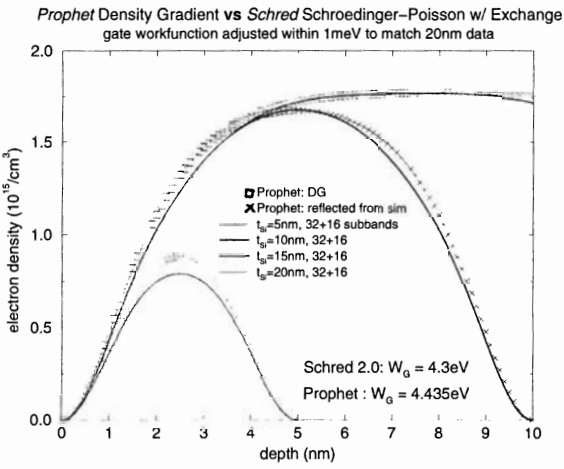


Fig. 4: Simulated electron profile in the active region of double-gate SOI MOSFETs with varying thin film thickness (5, 10, 20 nm). The comparison is made between DG model and the Schrödinger/Poisson's solver, SCHRED.

[4] D. Ferry, R. Akis, and D. Vasileska, "Quantum effects in MOSFETs: Use of an effective potential in 3D Monte Carlo simulation of ultra-short channel devices," p. 287, *IEDM*, Dec. 2000, San Francisco, CA.

[5] M.G. Ancona "Macroscopic description of quantum-mechanical tunneling," p. 1222, *Phys. Rev.*, B42, 1990.

[6] C. Bowen, *et al.*, "Physical oxide thickness extraction and verification using quantum mechanical simulation," p. 859, *IEDM Digest*, Dec. 1997, Washington, DC. Also see R. Lake, *et al.*, p. 7845, *J. Appl. Phys.* **81**, 1997.

[7] E.P. Wigner, "On the quantum correction for thermodynamic equilibrium," p. 749, *Physics Review*, vol. 40, June 1, 1932.

[8] M.G. Ancona, Z. Yu, R.W. Dutton, P.J. Vande Voorde, M. Cao, and D. Vook, "Density-gradient analysis of MOS tunneling," p. 2310, *IEEE T-ED* Dec. 2000.

[9] R. Chau, *et al.*, "30nm physical gate length CMOS transistors with 1.0 ps n-MOS and 1.7 ps p-MOS gate delays," p. 45, *IEDM*, Dec. 2000, San Francisco.