Journal of
Computational and Theoretical Nanoscience
Vol. 5, 1–32, 2008

REVIEW

Semiconductor Device Modeling

D. Vasileska*, D. Mamaluy, H. R. Khan, K. Raleva, and S. M. Goodnick

Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287-5706, USA

In this review paper we describe a hierarchy of simulation models for modeling state of the art devices. Within the semiclassical simulation arena, emphasis is placed on particle-based device simulations that can model devices operating from diffusive down to ballistic regime. In here, we also describe in detail the proper inclusion of the short-range Coulomb interactions using real-space approach that eliminates double-counting of the Coulomb interaction (due to its partial inclusion via the solution of the Poisson equation). Regarding the quantum transport approaches, emphasis is placed on the description of the CBR method that is implemented in ASU's 2D and 3D NEGF device simulator (that is used for modeling 10 nm gate length FinFETs, which are likely to be the next generation of devices that the Industry will be mass-producing in year 2015). Comparison with existing experimental data is presented to verify the accuracy and speed of the quantum transport simulator. We conclude this review paper by emphasizing what kind of semiconductor tools will be needed to model next generation devices.

Keywords: Semiclassical and Quantum Transport, Particle-Based Device Simulations, Boltzmann Transport Equation, Electron–Electron and Electron–Ion Interactions, Quantum Transport, Landauer's Approach, Green's Functions, Contact Block Reduction Method, FinFET Devices.

CONTENTS

1. COMPUTATIONAL ELECTRONICS

As semiconductor feature sizes shrink into the nanometer scale regime, even conventional device behavior becomes increasingly complicated as new physical phenomena at short dimensions occur, and limitations in material properties are reached.[1] In addition to the problems related to the understanding of actual operation of ultra-small devices, the reduced feature sizes require more complicated and time-consuming manufacturing processes. This fact signifies that a pure trial-and-error approach to device optimization will become impossible since it is both too time consuming and too expensive. Since computers are considerably cheaper resources, simulation is becoming an indispensable tool for the device engineer. Besides offering the possibility to test hypothetical devices which have not (or could not) yet been manufactured, simulation offers unique insight into device behavior by allowing the observation of phenomena that can not be measured on real devices. *Computational Electronics*[2–4] in this context refers to the physical simulation of semiconductor devices in terms of charge transport and the corresponding electrical behavior. It is related to, but usually separate from process simulation, which deals with various physical processes such as material growth, oxidation, impurity diffusion, etching, and metal deposition inherent in device fabrication[5] leading to integrated circuits. Device simulation can be thought of as one component of technology for computer-aided design (TCAD), which provides a basis for device modeling, which deals with compact behavioral models for devices and sub-circuits relevant for circuit simulation in commercial packages such as SPICE.[6] The relationship between various simulation design steps that have to be followed to achieve certain customer need is illustrated in Figure 1.

The goal of *Computational Electronics* is to provide simulation tools with the necessary level of sophistication to capture the essential physics while at the same time minimizing the computational burden so that results may

*Author to whom correspondence should be addressed.

Dragica Vasileska received the B.S.E.E. (Diploma) and the M.S.E.E. Degree form the University Sts. Cyril and Methodius (Skopje, Republic of Macedonia) in 1985 and 1992, respectively, and a Ph.D. Degree from Arizona State University in 1995. From 1995 until 1997 she held a Faculty Research Associate position within the Center of Solid State Electronics Research at Arizona State University. In the fall of 1997 she joined the faculty of Electrical Engineering at Arizona State University. In 2002 she was promoted to Associate Professor and in 2007 to Full Professor. Her research interests include semiconductor device physics and semiconductor device modeling, with strong emphasis on quantum transport and Monte Carlo particle-based device simulations. She is a Senior Member of both IEEE and APS. Professor D. Vasileska has published more than 100 publications in prestigious scientific journals, over 80 conference proceedings refereed papers, has given numerous invited talks and is a co-author on a book on Computational Electronics with Professor S. M. Goodnick. She has many awards including the best student award from the School of Electrical Engineering in Skopje since its existence (1985, 1990). She is also a recipient of the 1998 NSF CAREER Award. Her students have won the best paper and the best poster award at the LDSD conference in Cancun, 2004. Dragica Vasileska is a Senior Member of IEEE and is listed in Strathmore's Who's-Who.

REVIEW

Denis Mamaluy (M'05) was born in Kharkov, USSR in 1975. He received the MS in physics (with honors) from Kharkov State University in 1997, the MA in philosophy from the UNESCO department at Kharkov University in 1997, and the Ph.D. in physics and mathematics from the B. Verkin Institute for Low Temperature Physics and Engineering, Kharkov, Ukraine in 2000. From 2000 to 2002 he worked as Post-doctorate Fellow in Walter Schottky Institute, Technische Universität München, Munich, Germany. In 2003 he joined Arizona State University as Faculty Research Associate. In 2006 he was promoted to Research Assistant Professor. He is one of the authors of the contact block reduction (CBR) method for efficient quantum transport modeling in open nano-systems: D. Mamaluy et al., Phys. Rev. B 71, 245321 (2005); D. Mamaluy et al., J. App. Phys. 93, 4628 (2003). He has published over 30 peer-review journal papers, contributed to more than 40 conference papers and presentations, and given a number of invited talks on quantum transport simulation for industries and academia. He also worked on theory and applications of surface electromagnetic waves, induced by magneto-electric interaction. His current research interests include quantum transport in nano-structures and nano-devices, semiconductor device computer modeling. Dr. Mamaluy has won the best poster presentation award on 7th International Conference on Complex Media (Bianisotropics '98), in Braunschweig, Germany, 1998.

Hasanur R. Khan (S'06) was born in Dhaka, Bangladesh. He received B.Sc. in electrical and electronic engineering from the Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh in 1999 and MSE in electrical engineering with major in solid state electronics from Arizona State University (ASU), Tempe, Arizona in 2002. He is currently pursuing the Ph.D. degree in electrical engineering at ASU. He served as a lecturer in the department of Electrical engineering at BUET from 1999 to 2000. His current research interests include simulation, analyses and modeling of quantum effects in ultra-scaled semiconductor devices with an emphasis on double and tri-gate FinFET.

Katerina Raleva (M'99) received the B.S.E.E. (Diploma) and the M.S.E.E. Degree from the University Sts. Cyril and Methodius (Skopje, Republic of Macedonia) in 1991 and 2001, respectively. From 1993 she held a Teaching and Research Assistant position within the Department of Electronics at Faculty of Electrical Engineering and Information Technologies (FEIT), Skopje, Macedonia. She is currently pursuing the Ph.D. degree in electrical engineering at FEIT. Her Ph.D. research work is done in collaboration with the Center of Solid State Electronics Research at ASU. Her research of interests include, semiconductor physics, semiconductor device modeling and modeling on a circuit level.

Steve Goodnick is associate vice president for research at Arizona State University. In this position, he is responsible for matters involving the strategic investment of TRIF/Proposition 301 funds, allocation of space in ASU's new research facilities, and oversight of university research support functions including the Office for Research and Sponsored Projects Administration, Department of Animal Care, and other units. He will also help coordinate a series of other important technical initiatives at ASU, including alternative energy and the MacroTechnology Works. As one of ASU's most successful researchers over the past decade, Goodnick's research specializations lie in solid-state device physics, semi-conductor transport, quantum and nanostructure devices and device technology, and high frequency devices. He will maintain his leadership of ASU's nanoelectronics efforts as director while in this post. Goodnick previously served as the interim deputy dean for the Ira A. Fulton School of Engineering at ASU, and earlier as chair of the Fulton School's Department of Electrical Engineering, one of ASU's most active and successful units, and served as President of the Electrical and Computer Engineering Department Heads Association from 2003–2004. He received his B.S. in engineering science from Trinity University in 1977, and his M.S. and Ph.D. degrees in electrical engineering from Colorado State University in 1979 and 1983, respectively. Germany, Japan and Italy are among the countries he has served as a visiting scientist. Goodnick is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE) and an Alexander von Humboldt Research Fellow. Other honors and awards he has received include the IEEE Phoenix Section Society Award for Outstanding Service (2002), the Colorado State University College of Engineering Achievement in Academia Award (1998), and the College of Engineering Research Award (Oregon State University, 1996). He is a member of IEEE, the American Physical Society, American Association for the Advancement of Science, and the American Society of Engineering Education. His publication record includes more than 165 refereed journal articles, books and book chapters related to transport in semiconductor devices and microstructures.

REVIEW

be obtained within a reasonable time frame. Figure 2 illustrates the main components of semiconductor device simulation at any level. There are two main kernels, which must be solved self-consistently with one another, the transport equations governing charge flow, and the fields driving charge flow. Both are coupled strongly to one another, and hence must be solved simultaneously. The fields arise from external sources, as well as the charge and current densities which act as sources for the time varying electric and magnetic fields obtained from the solution of Maxwell's equations. Under appropriate conditions, only the quasi-static electric fields arising from the solution of Poisson's equation are necessary.

The fields, in turn, are driving forces for charge transport as illustrated in Figure 3 for the various levels of approximation within a hierarchical structure ranging from compact modeling at the top to an exact quantum mechanical description at the bottom. At the very beginnings of semiconductor technology, the electrical device characteristics could be estimated using simple analytical models (gradual channel approximation for MOSFETs) relying on the drift-diffusion (DD) formalism. Various approximations had to be made to obtain closed-form solutions, but the resulting models captured the basic features of the devices.[7] These approximations include simplified doping

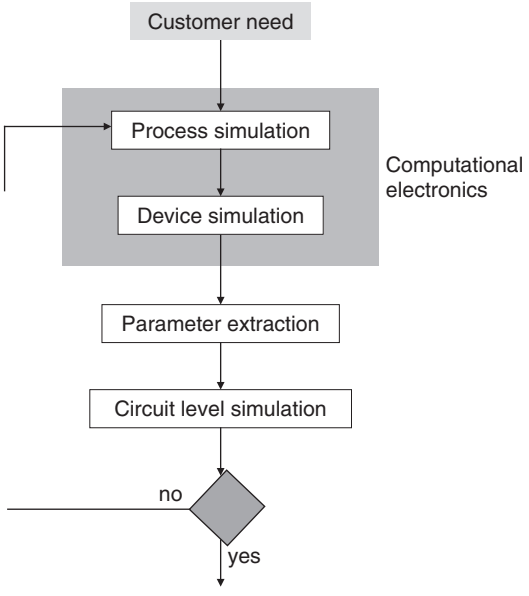

Fig. 1. Design sequence to achieve desired customer need.

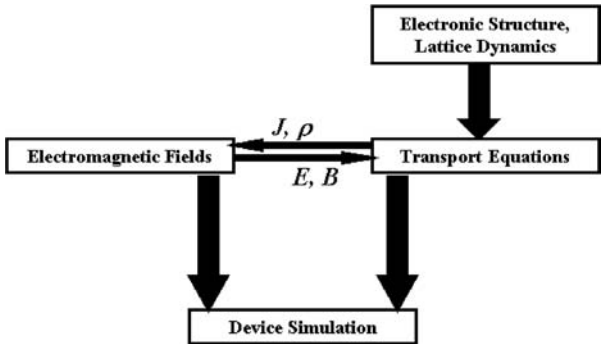

Fig. 2. Schematic description of the device simulation sequence.

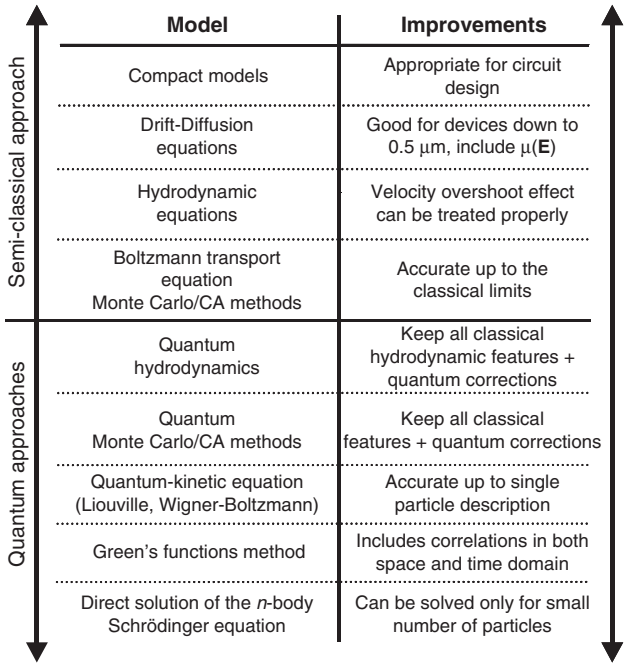

Fig. 3. Illustration of the hierarchy of transport models.

profiles and device geometries. With the ongoing refinements and improvements in technology, these approximations lost their basis and a more accurate description was required. This goal could be achieved by solving the DD equations numerically. Numerical simulation of carrier transport in semiconductor devices dates back to the famous work of Scharfetter and Gummel,[8] who proposed a robust discretization of the DD equations which is still in use today.

However, as semiconductor devices were scaled into the submicrometer regime, the assumptions underlying the DD model lost their validity. Therefore, the transport models have been continuously refined and extended to more accurately capture transport phenomena occurring in these devices. The need for refinement and extension is primarily caused by the ongoing feature size reduction in state-of-the-art technology. As the supply voltages can not be scaled accordingly without jeopardizing the circuit performance, the electric field inside the devices has increased. A large electric field, which rapidly changes over small length scales, gives rise to non-local and hot-carrier effects which begin to dominate device performance. An accurate description of these phenomena is required and is becoming a primary concern for industrial applications.

To overcome some of the limitations of the DD model, extensions have been proposed which basically add an additional balance equation for the average carrier energy.[9] Furthermore, an additional driving term is added to the current expression which is proportional to the gradient of the carrier temperature. However, a vast number of these models exist, and there is a considerable amount of confusion as to their relation to each other. It is now a common practice in industry to use standard hydrodynamic models in trying to understand the operation of as-fabricated devices, by adjusting any number of phenomenological parameters (e.g., mobility, impact ionization coefficient, etc.). However, such tools do not have predictive capability for ultra-small structures, for which it is necessary to relax some of the approximations in the Boltzmann transport equation.[10] Therefore, one needs to move downward to the quantum transport area in the hierarchical map of transport models shown in Figure 3, where, at the very bottom we have the Green's function approach.[11–13] The latter is the most exact, but at the same time the most difficult of all. In contrast to, for example, the Wigner function approach (which is Markovian in time), the Green's functions method allows one to consider simultaneously correlations in space and time, both of which are expected to be important in nano-scale devices. However, the difficulties in understanding the various terms in the resultant equations and the enormous computational burden needed for its actual implementation make the usefulness in understanding quantum effects in actual devices of limited values. For example, the only successful utilization of the Green's function approach commercially is the NEMO (Nano-Electronics Modeling) simulator,[14] which is effectively 1D and is primarily applicable to resonant tunneling diodes.

From the discussion above it follows that, contrary to the recent technological advances, the present state of the art in device simulation is currently lacking in the ability to treat these new challenges in scaling of device dimensions from conventional down to quantum scale devices. For silicon devices with active regions below 0.2 microns in diameter, macroscopic transport descriptions based on drift-diffusion models are clearly inadequate. As already noted, even standard hydrodynamic models do not usually provide a sufficiently accurate description since they neglect significant contributions from the tail of the phase space distribution function in the channel regions.[15, 16] Within the requirement of self-consistently solving the coupled transport-field problem in this emerging domain of device physics, there are several computational challenges, which limit this ability. One is the necessity to solve both the transport and the Poisson's equations over the full 3D domain of the device (and beyond if one includes radiation effects). As a result, highly efficient algorithms targeted to high-end computational platforms (most likely in a multi-processor environment) are required to fully solve even the appropriate field problems. The appropriate level of approximation necessary to capture the proper non-equilibrium transport physics, relevant to a future device model, is an even more challenging problem both computationally and from a fundamental physics framework.

2. SEMICLASSICAL TRANSPORT APPROACHES

2.1. Drift-Diffusion Model

In Section 1, we discussed the various levels of approximations that are employed in the modeling of semiconductor devices. The direct solution of the full BTE is challenging computationally, particularly when combined with field solvers for device simulation. Therefore, for traditional semiconductor device modeling, the predominant model corresponds to solutions of the so-called drift-diffusion equations, which are 'local' in terms of the driving forces (electric fields and spatial gradients in the carrier density), i.e., the current at a particular point in space only depends on the instantaneous electric fields and concentration gradient at that point. The complete drift-diffusion model is based on the following set of equations:

(1) Current equations

$$J_n = qn(x)\mu_n E(x) + qD_n \frac{dn}{dx}$$
$$J_p = qp(x)\mu_p E(x) - qD_p \frac{dn}{dx} \quad (1)$$

(2) Continuity equations

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla \cdot \mathbf{J}_n + U_n$$
$$\frac{\partial p}{\partial t} = -\frac{1}{q}\nabla \cdot \mathbf{J}_p + U_p \quad (2)$$

(3) Poisson's equation

$$\nabla \cdot (\varepsilon \nabla V) = -(p - n + N_D^+ - N_A^-) \quad (3)$$

where U_n and U_p are the net generation-recombination rates. The continuity equations are the conservation laws for the carriers. A numerical scheme which solves the continuity equations should

(1) Conserve the total number of particles inside the device being simulated.

(2) Respect local positive definite nature of carrier density. Negative density is unphysical.

(3) Respect monotonicity of the solution (i.e., it should not introduce spurious space oscillations).

Conservative schemes are usually achieved by subdivision of the computational domain into patches (boxes) surrounding the mesh points. The currents are then defined on the boundaries of these elements, thus enforcing conservation (the current exiting one element side is exactly equal to the current entering the neighboring element through the side in common). In the absence of generation-recombination terms, the only contributions to the overall device current arise from the contacts. Remember that, since electrons have negative charge, the particle flux is opposite to the current flux. When the equations are discretized, using finite differences for instance, there are limitations on the choice of mesh size and time step:[17]

(1) The mesh size Δx is limited by the Debye length.

(2) The time step is limited by the dielectric relaxation time.

A mesh size must be smaller than the Debye length where one has to resolve charge variations in space. A simple example is the carrier redistribution at an interface between two regions with different doping levels. Carriers diffuse into the lower doped region creating excess carrier distribution which at equilibrium decays in space down to the bulk concentration with approximately exponential behavior. The spatial decay constant is the Debye length

$$L_D = \sqrt{\frac{\varepsilon k_B T}{q^2 N}} \quad (4)$$

where N is the doping density. In GaAs and Si, at room temperature the Debye length is approximately 400 Å when $N \approx 10^{16}$ cm^{-3} and decreases to about only 50 Å when $N \approx 10^{18}$ cm.$^{-3}$

The dielectric relaxation time, on the other hand, is the characteristic time for charge fluctuations to decay under the influence of the field that they produce. The dielectric relaxation time may be estimated using

$$t_{dr} = \frac{\varepsilon}{qN\mu} \quad (5)$$

The drift-diffusion semiconductor equations constitute a coupled nonlinear set. It is not possible, in general, to obtain a solution directly in one step, but a nonlinear iteration method is required. The two most popular methods for solving the discretized equations are the Gummel's iteration method[18] and the Newton's method.[19] It is very difficult to determine an optimum strategy for the solution, since this will depend on a number of details related to the particular device under study.

Finally, the discretization of the continuity equations in conservation form requires the determination of the currents on the mid-points of mesh lines connecting neighboring grid nodes. Since the solutions are accessible only on the grid nodes, interpolation schemes are needed to determine the currents. The approach by Scharfetter and Gummel[8] has provided an optimal solution to this problem, although the mathematical properties of the proposed scheme have been fully recognized much later.

2.2. Hydrodynamic Model

The current drive capability of deeply scaled MOSFETs and, in particular, *n*-MOSFETs has been the subject of investigation since the late 1970s. First it was hypothesized that the effective carrier injection velocity from the source into the channel would reach the limit of the saturation velocity and remain there as longitudinal electric fields increased beyond the onset value for velocity saturation. However, theoretical work indicated that velocity overshoot can occur even in silicon,[20] and indeed it is

REVIEW

routinely seen in the high-field region near the drain in simulated devices using energy balance models or Monte Carlo. While it was understood that velocity overshoot near the drain would not help current drive, experimental work[21, 22] claimed to observe velocity overshoot near the source, which of course would be beneficial and would make the drift-diffusion model invalid.

In the computational electronics community, the necessity for the hydrodynamic (HD) transport model is normally checked by comparison of simulation results for HD and DD simulations. Despite the obvious fact that, depending on the equation set, different principal physical effects are taken into account, the influence on the models for the physical parameters is more subtle. The main reason for this is that in the case of the HD model, information about average carrier energy is available in form of carrier temperature. Many parameters depend on this average carrier energy, e.g., the mobilities and the energy relaxation times. In the case of the DD model, the carrier temperatures are assumed to be in equilibrium with the lattice temperature, that is $T_C = T_L$, hence, all energy dependent parameters have to be modeled in a different way.

2.2.1. Extensions of the Drift-Diffusion Model

In the DD approach, the electron gas is assumed to be in thermal equilibrium with the lattice temperature. ($T_n = T_L$) However, in the presence of a strong electric field, electrons gain energy from the field and the temperature T_n of the electron gas is elevated. Since the pressure of the electron gas is proportional to nk_BT_n, the driving force now becomes the pressure gradient rather then merely the density gradient. This introduces an additional driving force, namely, the temperature gradient besides the electric field and the density gradient. Phenomenologically, one can write

$$J = q(n\mu_n E + D_n \nabla n + nD_T \nabla T_n) \tag{6}$$

where D_T is the thermal diffusivity.

2.2.2. Stratton's Approach

One of the first derivations of extended transport equations was performed by Stratton.[23] First the distribution function is split into the even and odd parts

$$f(k, r) = f_0(k, r) + f_1(k, r) \tag{7}$$

From $f_1(-k, r) = -f_1(k, r)$, it follows that $\langle f_1 \rangle = 0$. Assuming that the collision operator C is linear and invoking the microscopic relaxation time approximation for the collision operator

$$C[f] = -\frac{f - f_{\text{eq}}}{\tau(\varepsilon, r)} \tag{8}$$

the BTE can be split into two coupled equations. In particular f_1 is related to f_0 via

$$f_1 = -\tau(\varepsilon, r)\left(v \cdot \nabla_r f_0 - \frac{q}{\hbar} E \cdot \nabla_k f_0\right) \tag{9}$$

The microscopic relaxation time is then expressed by a power law

$$\tau(\varepsilon) = \tau_0 \left(\frac{\varepsilon}{k_B T_L}\right)^{-p} \tag{10}$$

When f_0 is assumed to be heated Maxwellian distribution, the following equation system is obtained

$$\begin{aligned} \nabla \cdot J &= q\frac{\partial n}{\partial t} \\ J &= qn\mu E + k_B \nabla(n\mu T_n) \\ \nabla \cdot (nS) &= -\frac{3}{2}k_B \partial(nT_n) + E \cdot J - \frac{3}{2}k_B n \frac{T_n - T_L}{\tau_\varepsilon} \\ nS &= -\left(\frac{5}{2} - p\right)\left(\mu n k_B T_n E + \frac{k_B^2}{q}\nabla(n\mu T_n)\right) \end{aligned} \tag{11}$$

Equation for the current density can be rewritten as:

$$J = q\mu\left(nE + \frac{k_B}{q}T_n \nabla n + \frac{k_B}{q} n(1 + \nu_n)\nabla T_n\right) \tag{1}$$

with

$$\nu_n = \frac{T_n}{\mu}\frac{\partial \mu}{\partial T_n} = \frac{\partial \ln \mu}{\partial \ln T_n} \tag{2}$$

which is commonly used as a fit parameter with values in the range $[-0.5, -1.0]$. For $\nu_n = -1.0$, the thermal distribution term disappears. The problem with Eq. (10) for τ is that p must be approximated by an average value to cover the relevant processes. In the particular case of impurity scattering, p can be in the range $[-1.5, 0.5]$, depending on charge screening. Therefore, this average depends on the doping profile and the applied field; thus, no unique value for p can be given. Note also that the temperature T_n is a parameter of the heated Maxwellian distribution, which has been assumed in the derivation. Only for parabolic bands and a Maxwellian distribution, this parameter is equivalent to the normalized second-order moment.

2.2.3. Balance Equations Model

The first three balance equations, derived by taking moments of Boltzmann Transport Equation (BTE), take the form

$$\begin{aligned} \frac{\partial n}{\partial t} &= \frac{1}{e}\nabla \cdot J_n + S_n \\ \frac{\partial J_z}{\partial t} &= \frac{2e}{m^*}\sum_i \frac{\partial W_{iz}}{\partial x_i} + \frac{ne^2}{m^*}E_z - \left\langle\left\langle \frac{1}{\tau_m} \right\rangle\right\rangle J_z \\ \frac{\partial W}{\partial t} &= -\nabla \cdot F_W + E \cdot J - \left\langle\left\langle \frac{1}{\tau_E} \right\rangle\right\rangle (W - W_0) \end{aligned} \tag{13}$$

The balance equation for the carrier density introduces the carrier current density, which balance equation introduces the kinetic energy density. The balance equation for the kinetic energy density, on the other hand, introduces the energy flux. Therefore, a new variable appears in the hierarchy of balance equations and the set of infinite balance equations is actually the solution of the BTE. The momentum and energy relaxation rates, that appear in Eq. (13) are ensemble averaged quantities. For simple scattering mechanisms one can utilize the drifted-Maxwellian form of the distribution function, but for cases where several scattering mechanisms are important, one must use bulk Monte Carlo simulations to calculate these quantities.

One can express the energy flux that appears in Eq. (13) in terms of the temperature tensor. The energy flux, is calculated using

$$\mathbf{F}_W = \frac{1}{V}\sum_{\mathbf{p}} \mathbf{v}E(\mathbf{p})f(\mathbf{r},\mathbf{p},t) \tag{14}$$

which means that the i-th component of this vector equals to

$$F_{Wi} = v_{di}W + nk_{\mathrm{B}}\sum_j T_{ij}v_{dj} + Q_i \tag{15}$$

where Q_i is the component of the heat flux vector which describes loss of energy due to flow of heat out of the volume. To summarize, the kinetic energy flux equals the sum of the kinetic energy density times velocity plus the velocity times the pressure, which actually represents the work to push the volume plus the loss of energy due to flow of heat out. In mathematical terms this is expressed as

$$\mathbf{F}_W = \mathbf{v}W + nk_{\mathrm{B}}\overleftrightarrow{T}\cdot\mathbf{v} + \mathbf{Q} \tag{16}$$

With the above considerations, the momentum and the energy balance equations reduce to

$$\frac{\partial J_z}{\partial t} = \frac{2e}{m^*}\sum_i \frac{\partial}{\partial x_i}\left(K_{iz} + \frac{1}{2}nk_{\mathrm{B}}T_{iz}\right) + \frac{ne^2}{m^*}E_z - \left\langle\left\langle\frac{1}{\tau_m}\right\rangle\right\rangle J_z \tag{17}$$

$$\frac{\partial W}{\partial t} = -\nabla\cdot(\mathbf{v}W + \mathbf{Q} + nk_{\mathrm{B}}\overleftrightarrow{T}\cdot\mathbf{v}) + \mathbf{E}\cdot\mathbf{J}_n - \left\langle\left\langle\frac{1}{\tau_E}\right\rangle\right\rangle(W - W_0)$$

For displaced-Maxwellian approximation for the distribution function, the heat flux $\mathbf{Q} = 0$. However, Blotekjaer[24] has pointed out that this term must be significant for non-Maxwellian distributions, so that a phenomenological description for the heat flux, of the form described by Franz-Wiedermann law, which states that

$$\mathbf{Q} = -\kappa\nabla T_c \tag{18}$$

is used, where κ is the thermal or heat conductivity. In silicon, the experimental value of κ is 142.3 W/mK. The above description for $\mathbf{Q}$ actually leads to a closed set of equations in which the energy balance equation is of the form

$$\frac{\partial W}{\partial t} = -\nabla\cdot(\mathbf{v}W - \kappa\nabla T_c + nk_{\mathrm{B}}T_c\mathbf{v}) + \mathbf{E}\cdot\mathbf{J}_n - \left\langle\left\langle\frac{1}{\tau_E}\right\rangle\right\rangle(W - W_0) \tag{19}$$

It has been recognized in recent years that this approach is not correct for semiconductors in the junction regions, where high and unphysical velocity peaks are established by the Franz-Wiedemann law. To avoid this problem, Stettler et al.[25] have suggested a new form of closure

$$\mathbf{Q} = -\kappa\nabla T_c + \frac{5}{2}(1-r)\frac{k_{\mathrm{B}}T_L}{e}\mathbf{J} \tag{20}$$

where $\mathbf{J}$ is the current density and r is a tunable parameter less than unity. Now using

$$\frac{\partial}{\partial x}(2K_{iz}) = \frac{\partial}{\partial x_i}(nm^*v_{di}v_{dz}) = nm^*\frac{\partial}{\partial x}(v_{di}v_{dz}) = nm^*\left[\frac{\partial v_{di}}{\partial x_i}v_{dz} + v_{dz}\frac{\partial v_{dz}}{\partial x_z}\right] \tag{21}$$

and assuming that the spatial variations are confined along the z-direction, we have

$$\frac{\partial}{\partial x_z}(2K_{iz}) = \frac{\partial}{\partial x_z}(nm^*v_{dz}^2) \tag{22}$$

To summarize, the balance equations for the drifted-Maxwellian distribution function simplify to

$$\frac{\partial n}{\partial t} = \frac{1}{e}\nabla\cdot J_n + S_n$$

$$\frac{\partial J_z}{\partial t} = \frac{e}{m^*}\frac{\partial}{\partial x_z}(nm^*v_{dz}^2 + nk_{\mathrm{B}}T_c) + \frac{ne^2}{m^*}E_z - \left\langle\left\langle\frac{1}{\tau_m}\right\rangle\right\rangle J_z \tag{23}$$

$$\frac{\partial W}{\partial t} = -\frac{\partial}{\partial x_z}\left[(W + nk_{\mathrm{B}}T_c)v_{dz} - \kappa\frac{\partial T_c}{\partial x_z}\right] + J_zE_z - \left\langle\left\langle\frac{1}{\tau_E}\right\rangle\right\rangle(W - W_0)$$

where

$$J_z = -env_{dz} = -\frac{e}{m^*}P_z$$
$$W = \frac{1}{2}nm^*v_{dz}^2 + \frac{3}{2}nk_{\mathrm{B}}T_c \tag{24}$$

2.3. Particle Based Device Simulation Methods

In the previous sections we have considered continuum methods of describing transport in semiconductors, specifically the drift-diffusion and hydrodynamic models, which are derived From moments of the semi-classical Boltzmann Transport Equation (BTE). As approximations

REVIEW

to the BTE, it is expected that at some limit, such approaches become inaccurate, or fail completely. Indeed, one can envision that, as physical dimensions are reduced, at some level a continuum description of current breaks down, and the granular nature of the individual charge particles constituting the charge density in the active device region becomes important.

The microscopic simulation of the motion of individual particles in the presence of the forces acting on them due to external fields as well as the internal fields of the crystal lattice and other charges in the system has long been popular in the chemistry community, where *molecular dynamics* simulation of atoms and molecules have long been used to investigate the thermodynamic properties of liquids and gases. In solids, such as semiconductors and metals, transport is known to be dominated by random scattering events due to impurities, lattice vibrations, etc., which randomize the momentum and energy of charge particles in time. Hence, stochastic techniques to model these random scattering events are particularly useful in describing transport in semiconductors, in particular the Monte Carlo method.

The Ensemble Monte Carlo techniques have been used for well over 30 years as a numerical method to simulate nonequilibrium transport in semiconductor materials and devices and has been the subject of numerous books and reviews.[26–28] In application to transport problems, a random walk is generated using the random number generating algorithms common to modern computers, to simulate the stochastic motion of particles subject to collision processes. This process of random walk generation is part of a very general technique used to evaluate integral equations and is connected to the general random sampling technique used in the evaluation of multi-dimensional integrals.[29]

The basic technique as applied to transport problems is to simulate the free particle motion (referred to as the free flight) terminated by instantaneous random scattering events. The Monte Carlo algorithm consists of generating random free flight times for each particle, choosing the type of scattering occurring at the end of the free flight, changing the final energy and momentum of the particle after scattering, and then repeating the procedure for the next free flight. Sampling the particle motion at various times throughout the simulation allows for the statistical estimation of physically interesting quantities such as the single particle distribution function, the average drift velocity in the presence of an applied electric field, the average energy of the particles, etc. By simulating an *ensemble* of particles, representative of the physical system of interest, the non-stationary time-dependent evolution of the electron and hole distributions under the influence of a time-dependent driving force may be simulated.

This particle-based picture, in which the particle motion is decomposed into free flights terminated by instantaneous collisions, is basically the same approximate picture underlying the derivation of the semi-classical Boltzmann Transport Equation (BTE). In fact, it may be shown that the one-particle distribution function obtained from the random walk Monte Carlo technique satisfies the BTE for a homogeneous system in the long-time limit.[30] This semi-classical picture breaks down when quantum. mechanical effects become pronounced, and one cannot unambiguously describe the instantaneous position and momentum of a particle, a subject which we will comment on later. In the following, we develop the standard Monte Carlo algorithm used to simulate charge transport in semiconductors. We then discuss how this basic model for charge transport within the BTE is self-consistently solved with the appropriate field equations to perform particle based device simulation.

2.3.1. Free Flight Generation

In the Monte Carlo method, particle motion is assumed to consist of free flights terminated by instantaneous scattering events, which change the momentum and energy of the particle after scattering. So the first task is to generate free flights of random time duration for each particle. To simulate this process, the probability density, $P(t)$, is required, in which $P(t)dt$ is the joint probability that a particle will arrive at time t without scattering after a previous collision occurring at time $t = 0$, and then suffer a collision in a time interval dt around time t. The probability of scattering in the time interval dt around t may be written as $\Gamma[\mathbf{k}(t)]dt$, where $\Gamma[\mathbf{k}(t)]$ is the scattering rate of an electron or hole of wavevector $\mathbf{k}$. The scattering rate, $\Gamma[\mathbf{k}(t)]$, represents the sum of the contributions from each individual scattering mechanism, which are usually calculated quantum mechanically using perturbation theory, as described later. The implicit dependence of $\Gamma[\mathbf{k}(t)]$ on time reflects the change in $\mathbf{k}$ due to acceleration by internal and external fields. For electrons subject to time independent electric and magnetic fields, the time evolution of $\mathbf{k}$ between collisions is represented as

$$\mathbf{k}(t) = \mathbf{k}(0) - \frac{e(\mathbf{E} + \mathbf{v} \times \mathbf{B})t}{\hbar} \tag{25}$$

where $\mathbf{E}$ is the electric field, $\mathbf{v}$ is the electron velocity and $\mathbf{B}$ is the magnetic flux density. In terms of the scattering rate, $\Gamma[\mathbf{k}(t)]$, the probability that a particle has not suffered a collision after a time t is given by $\exp(-\int_0^t \Gamma[\mathbf{k}(t')]\,dt')$. Thus, the probability of scattering in the time interval dt after a free flight of time t may be written as the joint probability

$$P(t)dt = \Gamma[\mathbf{k}(t)] \exp\left[-\int_0^t \Gamma[\mathbf{k}(t')]\,dt'\right] dt \tag{26}$$

Random flight times may be generated according to the probability density $P(t)$ above using, for example, the pseudo-random number generator implicit on most modern computers, which generate uniformly distributed random numbers in the range [0, 1]. Using a direct method (see, for example Ref. [26]), random flight times sampled from $P(t)$ may be generated according to

$$r = \int_0^{t_r} P(t)\,dt \tag{27}$$

where r is a uniformly distributed random number and t_r is the desired free flight time. Integrating Eq. (27) with $P(t)$ given by Eq. (26) above yields

$$r = 1 - \exp\left[-\int_0^{t_r} \Gamma\left[\mathbf{k}(t')\right] dt'\right] \tag{28}$$

Since $1-r$ is statistically the same as r, Eq. (28) may be simplified to

$$-\ln r = \int_0^{t_r} \Gamma\left[\mathbf{k}(t')\right] dt' \tag{29}$$

Equation (29) is the fundamental equation used to generate the random free flight time after each scattering event, resulting in a random walk process related to the underlying particle distribution function. If there is no external driving field leading to a change of $\mathbf{k}$ between scattering events (for example in ultrafast photoexcitation experiments with no applied bias), the time dependence vanishes, and the integral is trivially evaluated. In the general case where this simplification is not possible, it is expedient to introduce the so called self-scattering method,[31] in which we introduce a fictitious scattering mechanism whose scattering rate always adjusts itself in such a way that the total (self-scattering plus real scattering) rate is a constant in time

$$\Gamma = \Gamma\left[\mathbf{k}(t')\right] + \Gamma_{\text{self}}\left[\mathbf{k}(t')\right] \tag{30}$$

where $\Gamma_{\text{self}}[\mathbf{k}(t')]$ is the self-scattering rate. The self-scattering mechanism itself is defined such that the final state before and after scattering is identical. Hence, it has no effect on the free flight trajectory of a particle when selected as the terminating scattering mechanism, yet results in the simplification of Eq. (29) such that the free flight is given by

$$t_r = -\frac{1}{\Gamma}\ln r \tag{31}$$

The constant total rate (including self-scattering) Γ, must be chosen at the start of the simulation interval (there may be multiple such intervals throughout an entire simulation) so that it is larger than the maximum scattering encountered during the same time interval. In the simplest case, a single value is chosen at the beginning of the entire simulation (constant gamma method), checking to ensure that the real rate never exceeds this value during the simulation. Other schemes may be chosen that are more computationally efficient, and which modify the choice of Γ at fixed time increments.[32]

2.3.2. Final State After Scattering

The algorithm described above determines the random free flight times during which the particle dynamics is treated semi-classically. For the scattering process itself, we need the type of scattering (i.e., impurity, acoustic phonon, photon emission, etc.) which terminates the free flight, and the final energy and momentum of the particle(s) after scattering. The type of scattering which terminates the free flight is chosen using a uniform random number between 0 and Γ, and using this pointer to select among the relative total scattering rates of all processes including self-scattering at the final energy and momentum of the particle

$$\Gamma = \Gamma_{\text{self}}[n,\mathbf{k}] + \Gamma_1[n,\mathbf{k}] + \Gamma_2[n,\mathbf{k}] + \cdots \Gamma_N[n,\mathbf{k}] \tag{32}$$

with n the band index of the particle (or subband in the case of reduced-dimensionality systems), and k the wavevector at the end of the free-flight. This process is illustrated schematically in Figure 4.

Once the type of scattering terminating the free flight is selected, the final energy and momentum (as well as band or subband) of the particle due to this type of scattering must be selected. For elastic scattering processes such as ionized impurity scattering, the energy before and after scattering is the same. For the interaction between electrons and the vibrational modes of the lattice described as quasi-particles known as phonons, electrons exchange finite amounts of energy with the lattice in terms of emission and absorption of phonons. For determining the final momentum after scattering, the scattering rate, $\Gamma_j[n, \mathbf{k}; m, \mathbf{k}']$ of the jth scattering mechanism is needed, where n and m are the initial and final band indices, and $\mathbf{k}$ and $\mathbf{k}'$ are the particle wavevectors before and after scattering. Defining a spherical coordinate system as shown in Figure 5 around the initial wavevector $\mathbf{k}$, the final wavevector $\mathbf{k}'$ is specified by $|\mathbf{k}'|$ (which depends on conservation of energy) as well as the azimuthal and polar angles, φ and θ around $\mathbf{k}$. Typically, the scattering rate, $\Gamma_j[n, \mathbf{k}; m, \mathbf{k}']$, only depends on the angle θ between $\mathbf{k}$ and $\mathbf{k}'$. Therefore, φ may be chosen using a uniform random number between 0 and 2π (i.e., $2\pi r$), while θ is

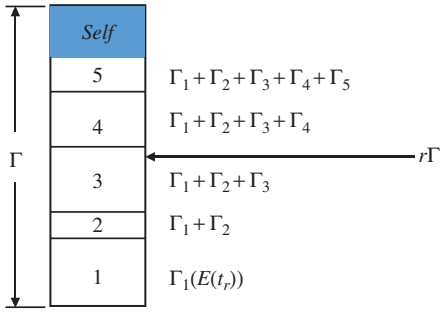

Fig. 4. Selection of the type of scattering terminating a free flight in the Monte Carlo algorithm.

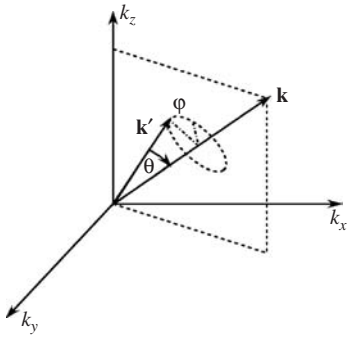

Fig. 5. Coordinate system for determining the final state after scattering.

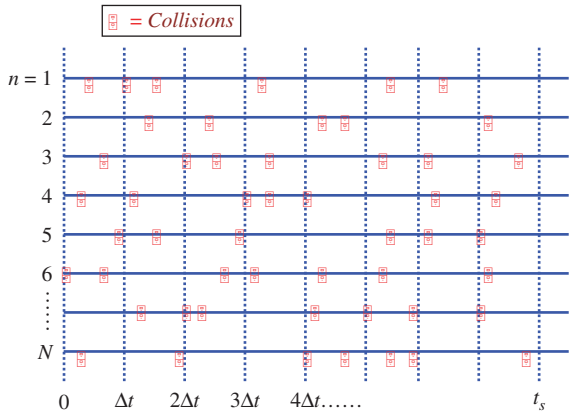

Fig. 6. Ensemble Monte Carlo simulation in which a time step, Δt, is introduced over which the motion of particles is synchronized. The squares represent random scattering events.

chosen according to the angular dependence for scattering arising from Γ_j $[n, \mathbf{k}; m, \mathbf{k}']$. If the probability for scattering into a certain angle $P(\theta)d\theta$ is integrable, then random angles satisfying this probability density may be generated from a uniform distribution between 0 and 1 through inversion of Eq. (27). Otherwise, a rejection technique (see, for example, Refs. [26, 27]) may be used to select random angles according to $P(\theta)$.

The rejection technique for sampling a random variable over some interval corresponds to choosing a maximum probability density (referred to here as a maximizing function) that is integrable in terms of Eq. (27) (for example a uniform or constant probability), and is always greater than or equal to the actual probability density of interest. A sample value of the random variable is then selected using a uniform number between 0 and 1, and then applying Eq. (27) to the maximizing function to select a value of the random variable analytically according to the probability density of the maximizing function. To now sample according to the desired probability density, a second random number is picked randomly between 0 and the value of the maximizing function at the value of the random variable chosen. If the value of this random number is less than the true value of the probability density (i.e., lies below it) at that point, the sampled value of the random variable is 'selected.' If it lies above, it is 'rejected,' and the process repeated until one satisfying the condition of selected is generated. In choosing random samples via this technique, one then samples according to the desired probability density.

2.3.3. Ensemble Monte Carlo Simulation

The basic Monte Carlo algorithm described in the previous sections may be used to track a single particle over many scattering events in order to simulate the steady-state behavior of a system. However, for improved statistics over shorter simulation times, and for transient simulation, the preferred technique is the use of a *synchronous ensemble* of particles, in which the basic Monte Carlo algorithm is repeated for each particle in an ensemble representing the (usually larger) system of interest until the simulation is completed. Since there is rarely an identical correspondence between the number of simulated charges, and the number of actual particles in a system, each particle is really a *super-particle*, representing a finite number of real particles. The corresponding charge of the particle is weighted by this super-particle number. Figure 6 illustrates an ensemble Monte Carlo simulation in which a fixed time step, Δt, is introduced over which the motion of all the carriers in the system is synchronized. The squares illustrate random, instantaneous, scattering events, which may or may not occur during a given time-step. Basically, each carrier is simulated only up to the end of the time-step, and then the next particle in the ensemble is treated. Over each time step, the motion of each particle in the ensemble is simulated independent of the other particles. Nonlinear effects such as carrier–carrier interactions or the Pauli exclusion principle are then updated at each times step, as discussed in more detail below.

The non-stationary one-particle distribution function and related quantities such as drift velocity, valley or subband population, etc., are then taken as averages over the ensemble at fixed time steps throughout the simulation. For example, the drift velocity in the presence of the field is given by the ensemble average of the component of the velocity at the nth time step as

$$\bar{v}_z(n\Delta t) \cong \frac{1}{N}\sum_{j=1}^{N} v_z^j(n\Delta t) \tag{33}$$

where N is the number of simulated particles and j labels the particles in the ensemble. This equation represents an estimator of the true velocity, which has a standard error given by

$$s = \frac{\sigma}{\sqrt{N}} \tag{34}$$

where σ^2 is the variance which may be estimated from Ref. [29]

$$\sigma^2 \cong \frac{N}{N-1}\left\{\frac{1}{N}\sum_{j=1}^{N}(v_z^j)^2 - \bar{v}_z^2\right\} \tag{35}$$

REVIEW

Similarly, the distribution functions for electrons and holes may be tabulated by counting the number of electrons in cells of k-space. From Eq. (35), we see that the error in estimated average quantities decreases as the square root of the number of particles in the ensemble, which necessitates the simulation of many particles. Typical ensemble sizes for good statistics are in the range of 10^4–10^5 particles. Variance reduction techniques to decrease the standard error given by Eq. (35) may be applied to enhance statistically rare events such as impact ionization or electron–hole recombination.[27]

An overall flowchart of a typical Ensemble Monte Carlo (EMC) simulation is illustrated in Figure 7. After initialization of run parameters, there are two main loops, and outer one which advances the time step by increments of ΔT until the maximum time of the simulation is reached, and an inner loop over all the particles in the ensemble (N), where the Monte Carlo algorithm is applied to each particle individually over a given time step.

2.3.4. Device Simulation Using Particles

Within an inhomogeneous device structure, both the transport dynamics and an appropriate field solver are coupled to each other. For quasi-static situations, the spatially varying fields associated with the potential arising from the numerical solution of Poisson's equation are the driving force accelerating particles in the Monte Carlo phase.

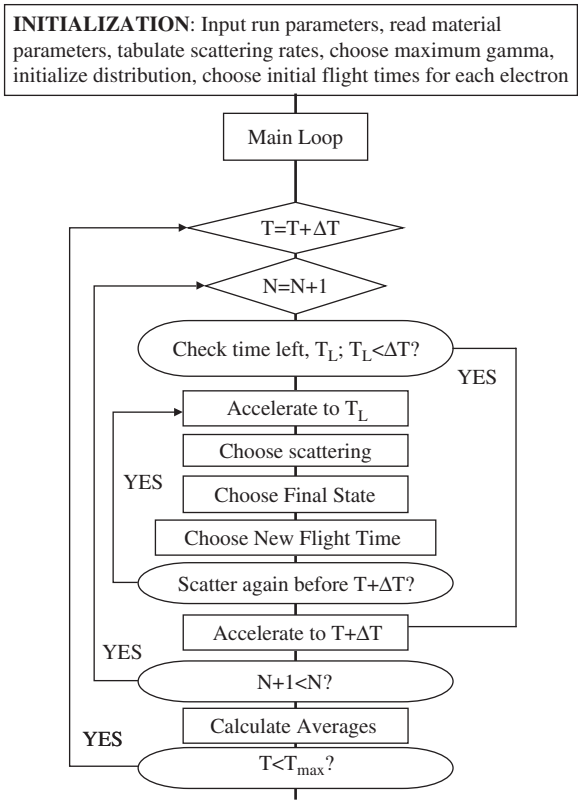

Fig. 7. Flow chart of an ensemble Monte Carlo (EMC) simulation.

Likewise, the distribution of mobile (both electrons and holes) and fixed charges (e.g., donors and acceptors) provides the source of the electric field in Poisson's equation. By decoupling the transport portion from the field portion over a small time interval (discussed in more detail below), a convergent scheme is realized in which the Monte Carlo transport phase is self-consistently coupled to Poisson's equation, similar to Gummel's algorithm. In the following section, a description of Monte Carlo particle-based device simulators is given, with emphasis on the particle-mesh coupling and the inclusion of the short-range Coulomb interactions.

As mentioned above, for device simulation based on particles, Poisson's equation is decoupled from the particle motion (described e.g., by the EMC algorithm) over a suitably small time step, typically less than the inverse plasma frequency corresponding to the highest carrier density in the device. Over this time interval, carriers accelerate according to the frozen field profile from the previous time-step solution of Poisson's equation, and then Poisson's equation is solved at the end of the time interval with the frozen configuration of charges arising from the Monte Carlo Phase. It is important to note that Poisson's equation is solved on a discrete mesh, whereas the solution of charge motion using EMC occurs over a continuous range of coordinate space in terms of the particle position. An illustration of a typical device geometry and the particle mesh scheme is shown in Figure 8. Therefore, a particle-mesh (PM) coupling is needed for both the charge assignment and the force interpolation. The size of the mesh and the characteristic time scales of transport set constraints on both the time-step and the mesh size. We must consider how particles are treated in terms of the boundaries, and how they are injected. Finally, the determination of the charge motion and corresponding terminal currents from averages over the simulation results are necessary in order to calculate the I–V characteristics of a device. These issues are discussed briefly below, along with some typical simulation results.

As in the case of any time domain simulation, for stable Monte Carlo device simulation, one has to choose the appropriate time step, Δt, and the spatial mesh size (Δx, Δy, and/or Δz). The time step and the mesh size may correlate to each other in connection with the numerical

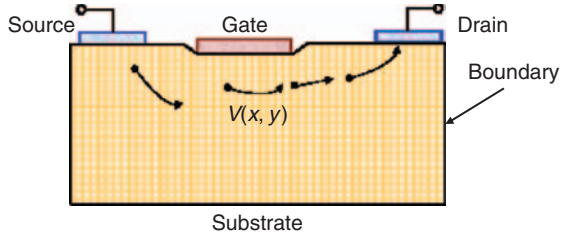

Fig. 8. Schematic diagram of a prototypical three-terminal device where charge flow is described by particles, while the fields are solved on a finite mesh.

REVIEW

stability. For example, the time step Δt must be related to the plasma frequency. From the viewpoint of numerical stability, Δt must be much smaller than the inverse plasma frequency above. Since the inverse plasma frequency goes as $1/\sqrt{n}$, the highest carrier density occurring in the modeled device structure corresponds to the smallest time used to estimate Δt. If the material is a multi-valley semiconductor, the smallest effective mass encountered by the carriers must be used as well.

The mesh size for the spatial resolution of the potential is dictated by the spatial variation of charge variations. Hence, one has to choose the mesh size to be smaller than the smallest wavelength of the charge variations. The smallest wavelength is approximately equal to the Debye length (for degenerate semiconductors the relevant length is the Thomas-Fermi wavelength).

Based on the discussion above, the time step (Δt), and the mesh size (Δx, Δy, an/or Δz) are chosen independently based on the physical arguments given above. However, there are numerical constraints coupling both as well. More specifically, the relation of Δt to the grid size must also be checked by calculating the distance $l_{\max}$, defined as

$$l_{\max} = v_{\max} \times \Delta t \tag{36}$$

where $\mathbf{v}_{\max}$ is the maximum carrier velocity, that can be approximated by the maximum group velocity of the electrons in the semiconductor (on the order of 10^8 cm/s). The distance $l_{\max}$ is the maximum distance the carriers can propagate during Δt. The time step is therefore chosen to be small enough so that $l_{\max}$ is smaller than the spatial mesh size chosen. This constraint arises because for too large of a time step, Δt, there may be substantial change in the charge distribution, while the field distribution in the simulation is only updated every Δt, leading to unacceptable errors in the carrier force.

To illustrate these various constraints, Figure 9 illustrates the range of stability for the time step and minimum grid size. The unshaded region corresponds to stable selections of both quantities. The right region is unstable due to the time step being larger than the inverse plasma frequency, whereas the upper region is unstable due to the grid spacing being larger than the Debye length. The velocity constraint bounds the lower side with its linear dependence on time-step.

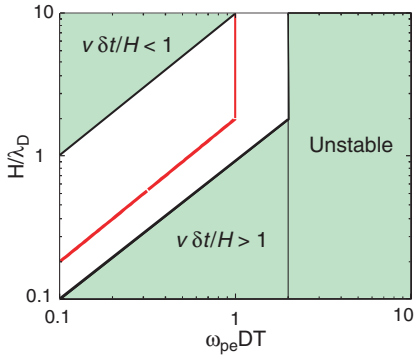

Fig. 9. Illustration of the region of stability (unshaded regions) of the time step, δt, and the minimum grid size, H. ω_{pe} is the plasma frequency corresponding to the maximum carrier density.

An issue of importance in particle-based simulation is the real space boundary conditions for the particle part of the simulation. Reflecting or periodic boundary conditions are usually imposed at the artificial boundaries. For Ohmic contacts, they require more careful consideration because electrons (or holes) crossing the source and drain contact regions contribute to the corresponding terminal currents. In order to conserve charge in the device, the electrons exiting the contact regions must be re-injected. Commonly employed models for the contacts include:[33]

- Electrons are injected at the opposite contact with the same energy and wavevector **k**. If the source and drain contacts are in the same plane, as in the case of MOSFET simulations, the sign of **k**, normal to the contact will change. This is an unphysical model, however.[34]
- Electrons are injected at the opposite contact with a wavevector randomly selected based upon a thermal distribution. This is also an unphysical model.
- Contact regions are considered to be in thermal equilibrium. The total number of electrons in a small region near the contact are kept constant, with the number of electrons equal to the number of dopant ions in the region. This approximation is most commonly employed in actual particle based device simulation.
- Another method uses 'reservoirs' of electrons adjacent to the contacts. Electrons naturally diffuse into the contacts from the reservoirs, which are not treated as part of the device during the solution of Poisson's equation. This approach gives results similar to the velocity-weighted Maxwellian, but at the expense of increased computational time due to the extra electrons simulated. It is an excellent model employed in some of the most sophisticated particle-based simulators. There are also several possibilities for the choice of the distribution function–Maxwellian, displaced Maxwellian, and velocity-weighted Maxwellian.[33]

The particle-mesh (PM) coupling is broken into four steps: (1) assignment of particle charge to the mesh; (2) solution of Poisson's equation on the mesh; (3) calculation of the mesh-defined forces; and (4) interpolation to find the forces acting on the particle. The charge assignment and force interpolation schemes usually employed in self-consistent Monte Carlo device simulations are the nearest-grid-point (NGP) and the cloud-in-cell (CIC) schemes.[35] Figure 10 illustrates both methods. In the NGP scheme, the particle position is mapped into the charge density at the closest grid point to a given particle. This has the advantage of simplicity, but leads to a noisy charge distribution, which may exacerbate numerical instability. Alternately, within the CIC scheme, a finite volume is associated with each particle spanning several cells in the

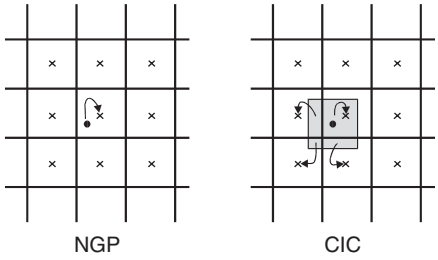

Fig. 10. Illustration of the charge assignment based on the nearest grid point method (NGP) and the cloud in cell method (CIC).

mesh, and a fractional portion of the charge per particle is assigned to grid points according to the relative volume of the 'cloud' occupying the cell corresponding to the grid point. This method has the advantage of smoothing the charge distribution due to the discrete charges of the particle based method, but may result in an artificial 'self-force' acting on the particle, particularly if an inhomogeneous mesh is used.

The requirements for constant permittivity (P) and constant mesh (M) severely limit the scope of devices that may be considered in device simulations using the NGP and the CIC schemes. Laux[36] proposed a new particle-mesh coupling scheme, namely, the nearest-element-center (NEC) scheme, which relaxes the restrictions (P) and (M). The NEC charge assignment/force interpolation scheme attempts to reduce the self-forces and increase the spatial accuracy in the presence of nonuniformly spaced tensor-product meshes and/or spatially-dependent permittivity. In addition, the NEC scheme can be utilized in one axis direction (where local mesh spacing is nonuniform) and the CIC scheme can be utilized in the other (where local mesh spacing is uniform). Such hybrid schemes offer smoother assignment/interpolation on the mesh compared to the pure NEC. The NEC designation derives from the appearance, of moving the charge to the center of its element and applying a CIC-like assignment scheme. The NEC scheme involves only one mesh element and its four nodal values of potential. This locality makes the method well-suited to non-uniform mesh spacing and spatially-varying permittivity. The interpolation and error properties of the NEC scheme are similar to the NGP scheme.

The motion in real space of particles under the influence of electric fields is somewhat more complicated due to the band structure. The velocity of a particle in real space is related to the E-$\mathbf{k}$ dispersion relation defining the bandstructure as

$$\mathbf{v}(t) = \frac{d\mathbf{r}}{dt} = \frac{1}{\hbar}\nabla_{\mathbf{k}} E(\mathbf{k}(t)) \qquad \frac{d\mathbf{k}}{dt} = \frac{q\mathbf{E}(\mathbf{r})}{\hbar} \tag{37}$$

where the rate of change of the crystal momentum is related to the local electric field acting on the particle through the acceleration theorem expressed by the second equation. In turn, the change in crystal momentum, $\mathbf{k}(t)$, is related to the velocity through the gradient of E with respect to $\mathbf{k}$. If one has to use the full bandstructure of the semiconductor, then integration of these equations to find $\mathbf{r}(t)$ is only possible numerically, using for example a Runge-Kutta algorithm. If a three valley model with parabolic bands is used, then the expression is integrable

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = \frac{\hbar\mathbf{k}}{m^*}; \qquad \frac{d\mathbf{k}}{dt} = \frac{q\mathbf{E}(\mathbf{r})}{\hbar} \tag{38}$$

Therefore, for a constant electric field in the x direction, the change in distance along the x direction is found by integrating twice

$$x(t) = x(0) + v_x(0)t + \frac{qE_x^0 t^2}{2m^*} \tag{39}$$

To simulate the steady-state behavior of a device, the system must be initialized in some initial condition, with the desired potentials applied to the contacts, and then the simulation proceeds in a time stepping manner until steady-state is reached. This process may take several picoseconds of simulation time, and consequently several thousand time-steps based on the usual time increments required for stability. Clearly, the closer the initial state of the system is to the steady state solution, the quicker the convergence. If one is, for example, simulating the first bias point for a transistor simulation, and has no a priori knowledge of the solution, a common starting point for the initial guess is to start out with charge neutrality, i.e., to assign particles randomly according to the doping profile in the device and based on the super-particle charge assignment of the particles, so that initially the system is charge neutral on the average. For two-dimensional device simulation, one should keep in mind that each particle actually represents a rod of charge into the third dimension. Subsequent simulations at the same device at different bias conditions can use the steady state solution at the previous bias point as a good initial guess. After assigning charges randomly in the device structure, charge is then assigned to each mesh point using the NGP or CIC or NEC particle-mesh methods, and Poisson's equation solved. The forces are then interpolated on the grid, and particles are accelerated over the next time step. A flow-chart of a typical Monte Carlo device simulation is shown in Figure 11.

As the simulation evolves, charge will flow in and out of the contacts, and depletion regions internal to the device will form until steady state is reached. The charge passing through the contacts at each time step can be tabulated, and a plot of the cumulative charge as a function of time gives the steady-state current. Figure 12 shows the particle distribution in 3D of a MESFET, where the dots indicate the individual simulated particles for two different gate biases. Here, the heavily doped MESFET region (shown by the inner box) is surrounded by semi-insulating GaAs

REVIEW

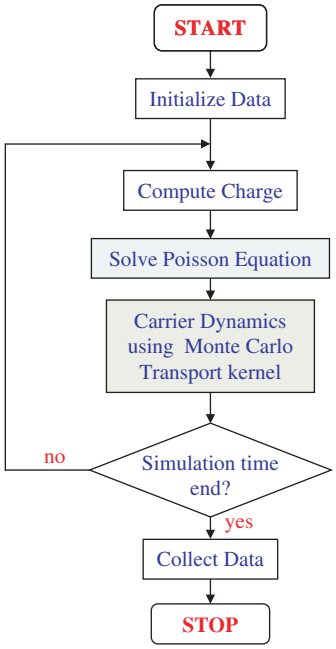

Fig. 11. Flow-chart of a typical particle based device simulation.

forming the rest of the simulation domain. The upper curve corresponds to no net gate bias (i.e., the gate is positively biased to overcome the built-in potential of the Schottky contact), while the lower curve corresponds to a net negative bias applied to the gate, such that the channel is close to pinch-off. One can see the evident depletion of carriers under the gate under the latter conditions.

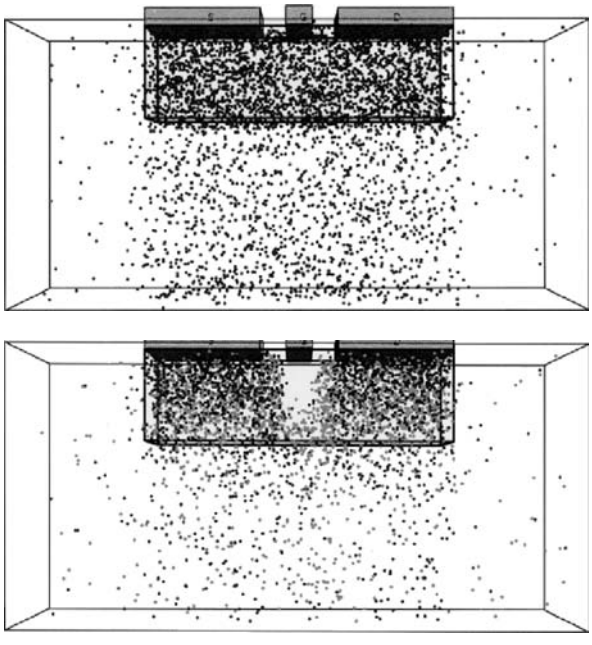

Fig. 12. Example of the particle distribution in a MESFET structure simulated in 3D using an EMC approach. The upper plot is the device with zero gate voltage applied, while the lower is with a negative gate voltage applied, close to pinch-off.

After sufficient time has elapsed, so that the system is driven into a steady-state regime, one can calculate the steady-state current through a specified terminal. The device current can be determined via two different, but consistent methods. First, by keeping track of the charges entering and exiting each terminal, the net number of charges over a period of the simulation can be used to calculate the terminal current. This method, however, is relatively noisy due to the discrete nature of the carriers, and the fact that one is only counting the currents crossing a 2D boundary in the device, which limits the statistics. A second method uses the sum of the carrier velocities in a portion of the device are used to calculate the current. For this purpose, the device is divided into several sections along, for example, the x-axis (from source to drain for the case of a MOSFET or MESFET simulation). The number of carriers and their corresponding velocity is added for each section after each free-flight time step. The total x-velocity in each section is then averaged over several time steps to determine the current for that section. The total device current can be determined from the average of several sections, which gives a much smoother result compared to counting the terminal charges. By breaking the device into sections, individual section currents can be compared to verify that the currents are uniform. In addition, sections near the source and drain regions of a MOSFET or a MESFET may have a high y-component in their velocity and should be excluded from the current calculations.

2.3.5. Simulation Example

The fully depleted (FD) Silicon-On-Insulator (SOI) MOSFET (Fig. 13) was of much interest a decade ago, because of its projected superiority over the partially depleted (PD) and the bulk-silicon counterparts.[37] Its advantages, due mainly to the gate-substrate charge coupling enabled by the thin FD Si-film body on a thick buried oxide (BOX)[38, 39] included higher drive current/transconductance, near-ideal subthreshold slope, low-junction capacitance and suppression of the floating-body effects. However, in the deep sub-micrometer regime, because of velocity saturation, two-dimensional effects in the BOX and the technological limits of scaling the SOI-body thickness, these advantages diminished,[40] the interest subsided and classical (i.e., bulk-Si and PD SOI) CMOS prevailed.

Now, as the scaling of classical CMOS approaches its limit, interest in non-classical FD devices-particularly double-gate (DG) and ultra-thin-body (UTB) MOSFETs—is rapidly growing. These devices deliver fundamental improvements over the performance of the bulk Si MOSFET devices.[41] And, while DG FinFETs[42] seem most promising, their complex and immature process technology has led to a renewed interest in single-gate FD SOI UTB MOSFETs.[43]

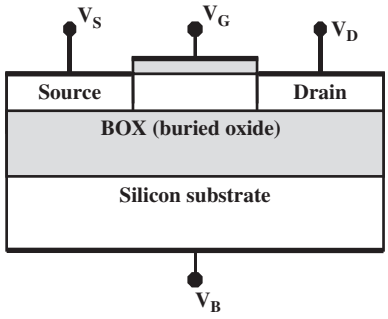

Fig. 13. FD SOI MOSFET device structure.

The dimensions of the *n*-channel FD SOI MOSFET being investigated using the particle-based device simulator discussed in the previous section are: the channel length is 25 nm, the silicon film width, which is equal to the source/drain junction depth is 10 nm, the gate oxide width is 2 nm, the BOX width is 50 nm, the source/drain doping is 1×10^{19} cm^{-3} and the channel doping is 1×10^{18} cm^{-3}.

To simulate the steady-state state behavior of a device, the system is started in some initial condition, with the desired potential applied to the contacts, and then the simulation proceeds in a time stepping manner until steady-state is reached. This takes several picoseconds of simulation time and consequently several thousand time steps based on the usual time increments required for stability. A common starting point for the initial guess is to start out with charge neutrality, i.e., to assign particles randomly according to the doping profile in the device, so that initially the system is charge neutral on the average. After assigning charges randomly in the device structure, charge is then assigned to each mesh point using an adequate PM coupling method, and Poisson's equation is solved. The forces are then interpolated on the grid, and particles are accelerated over the next time step. At this stage, typical simulation result that is shown is the variation of the scattering rates of the various scattering mechanisms included in the model. In our case, we have included acoustic phonon scattering, *g*- and *f*-intervalley phonon scattering (see Fig. 14). Afterwards, bias is applied and the carriers undergo the free-flight scattering sequence until steady-state is achieved. At this point, it is important to present the simulation results for the average drift velocity and the average carrier energy in the channel region of the device. These results are shown in Figures 15(a) and (b), respectively. They demonstrate the need for performing Monte Carlo device simulations that are more time consuming then solving either the drift-diffusion or the hydrodynamic model discussed previously.

The choice of the Monte Carlo device simulation is justified with the fact that in the devices simulated, we observe significant velocity overshoot near the drain end of the channel. Namely, the saturation velocity of the electrons in Si is 1.1×105 m/s and from the results shown in

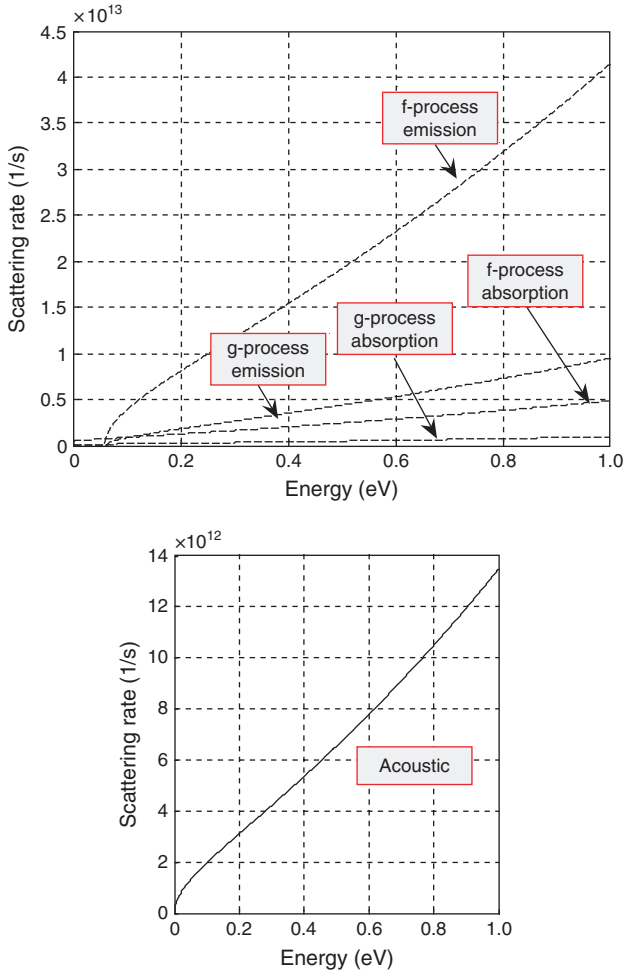

Fig. 14. Scattering rate variation versus energy for the various scattering mechanisms included in the model.

Figure 15(a) it is evident that the electrons are in the overshoot regime near the drain end of the channel and their average drift velocity exceeds 2×105 m/s. Proper modeling of the velocity overshoot effect, which leads to larger current drive, is only possible via a Monte Carlo device simulation scheme. Another issue that is worth mentioning is the fact that the average carrier energy in the channel region of the device is less than 0.5 eV which justifies the use of the non-parabolic model that is adopted in this work.

After sufficient time has elapsed, so that the system is driven into a steady-state regime, one can calculate the steady-state current through a specified terminal. As already discussed, the device current can be determined via two different, but consistent methods. First, by keeping track of the charges entering and exiting each terminal, the net number of charges over a period of the simulation can be used to calculate the current (Fig. 16(a)). The method is quite noisy due to the discrete nature of the carriers. In a second method, the sum of the carrier velocities in a portion of the device are used to calculate the

REVIEW

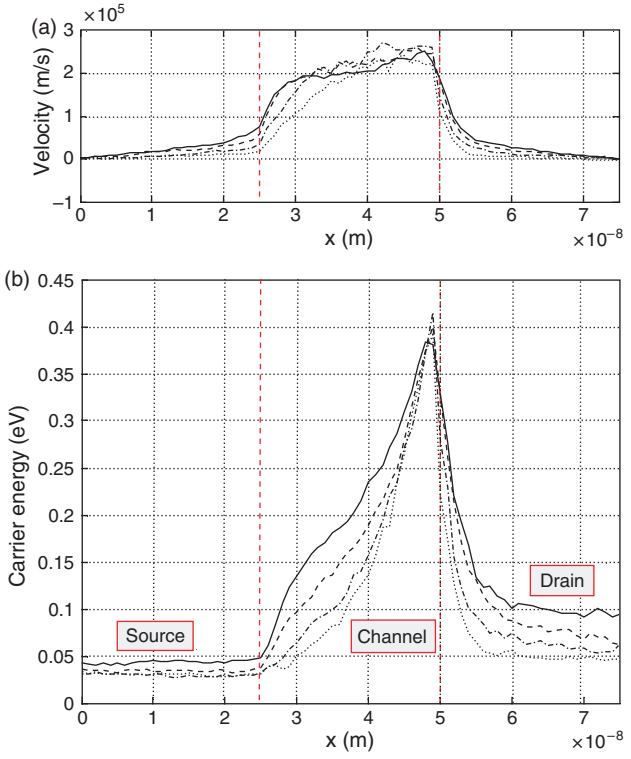

Fig. 15. (a) Average carrier drift velocity along the channel. (b) Average carrier energy. Different currents correspond to the different gate biases denoted on Figure 16.

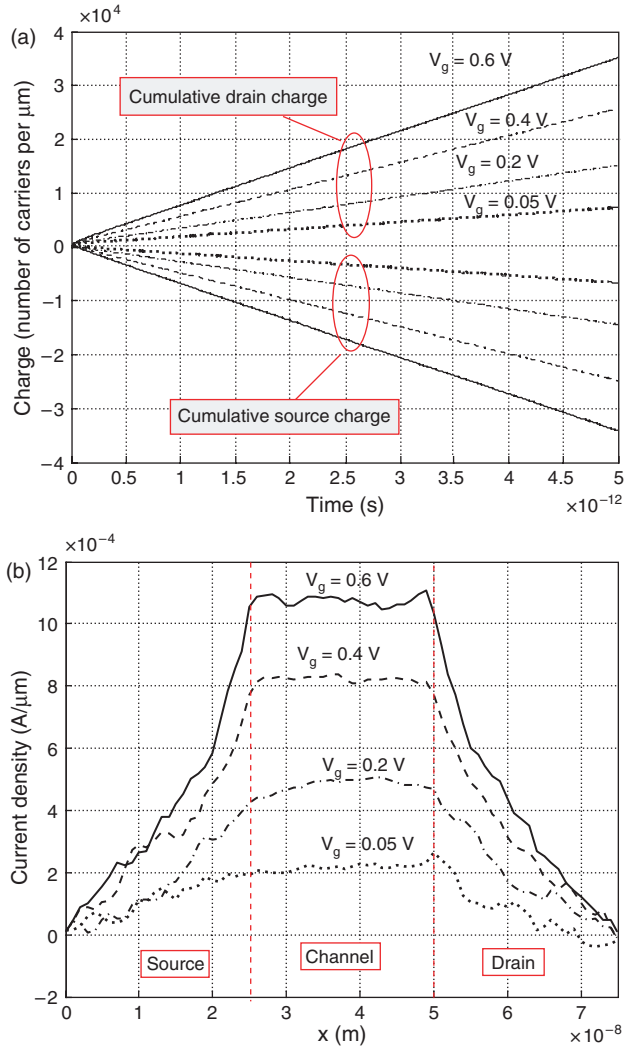

Fig. 16. (a) Cumulative charge versus time for drain bias $V_d = 0.6$ V and different gate biases. The slope of the curve gives the source and drain currents. (b) Current density calculated by using the average drift velocity of the carriers in the x-direction. Both methods give the same value of the current through the device which suggests that conservation of particles in the system is being preserved.

current (Fig. 16(b)). For this purpose, the device is divided into several sections along, for example, the x-axis (from source to drain for the case of a MOSFET simulation).

The number of carriers and their corresponding velocity is added for each section after each free flight time step. The total x-velocity in each section is then averaged over several time steps to determine the current for that section. The total device current can be determined from the average of several sections, which gives a much smoother result compared to counting the terminal charges. By breaking the device into sections, individual section currents can be compared to verify that the currents are uniform. In addition, sections near the source and the drain regions of a MOSFET may have a high y-velocity and should be excluded from the current calculations. Finally, by using several sections in the channel, the average energy and velocity of electrons along the channel is checked to ensure proper physical characteristics. The two ways of determining current through the device are demonstrated in Figures 16(a and b).

In Figure 17 we show the device transfer characteristics for different drain biases. It is obvious from the results presented that the threshold voltage shifts due to the Drain Induced Barrier Lowering (DIBL) of the source barrier (see Fig. 18). This observation also demonstrates the need of using computer simulations for modeling semiconductor devices as fields and potentials are two-dimensional quantities and one-dimensional models can not properly capture effects such as DIBL.

Finally a snapshot of the electron density in the channel, when the transistor is turned on, is shown in Figure 19(a). We see the existence of electrons in the channel region of the device. The corresponding conduction band profile, for the same biasing conditions, that is smoothed over time, is shown in Figure 19(b) and demonstrates the two-dimensional character of the potential and the electric field profiles in the active portion of the device.

2.3.6. Direct Treatment of Inter-Particle Interaction

In modern deep-submicrometer devices, for achieving optimum device performance and eliminating the so-called punch-through effect, the doping densities must be

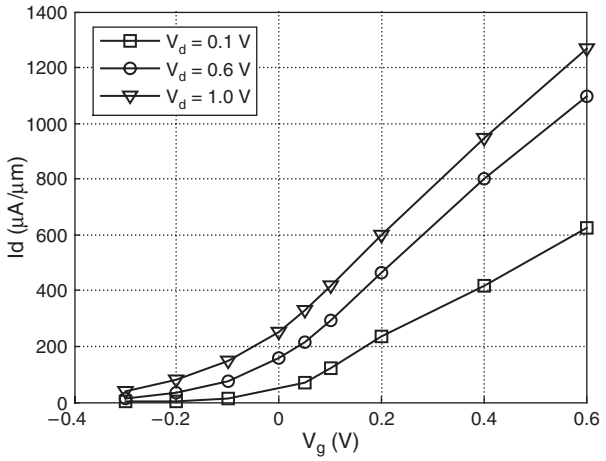

Fig. 17. Transfer characteristics of the device shown in Figure 1. Notice the increase in threshold voltage with increasing drain bias.

quite high. This necessitates a careful treatment of the electron–electron (*e–e*) and electron–impurity (*e–i*) interactions, an issue that has been a major problem for quite some time. Many of the approaches used in the past have included the short-range portions of the *e–e* and *e–i* interactions in the **k**-space portion of the Monte Carlo transport kernel, thus neglecting many of the important inelastic properties of these two interaction terms.[44,45] An additional problem with this screened scattering approach in devices is that, unlike the other scattering processes, *e–e* and *e–i* scattering rates need to be re-evaluated frequently during the simulation process to take into account the changes in the distribution function in time and spatially. The calculation and tabulation of a spatially inhomogeneous distribution function may be highly CPU and memory intensive. Furthermore, ionized impurity scattering is usually treated as a simple two-body event, thus ignoring the multi-ion contributions to the overall scattering

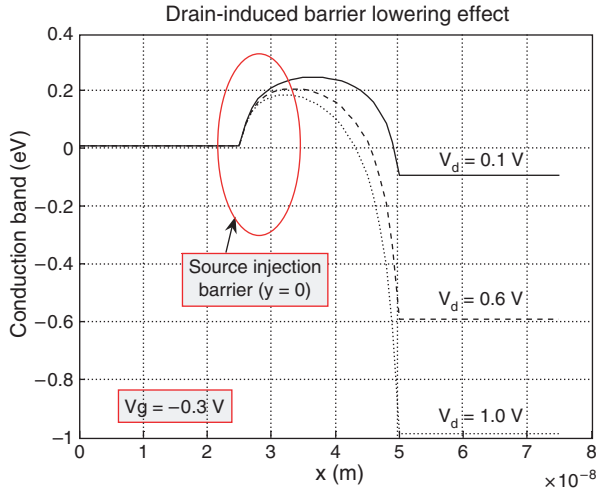

Fig. 18. Demonstration of drain induced barrier lowering for gate bias $V_g = -0.3$ V and different drain bias.

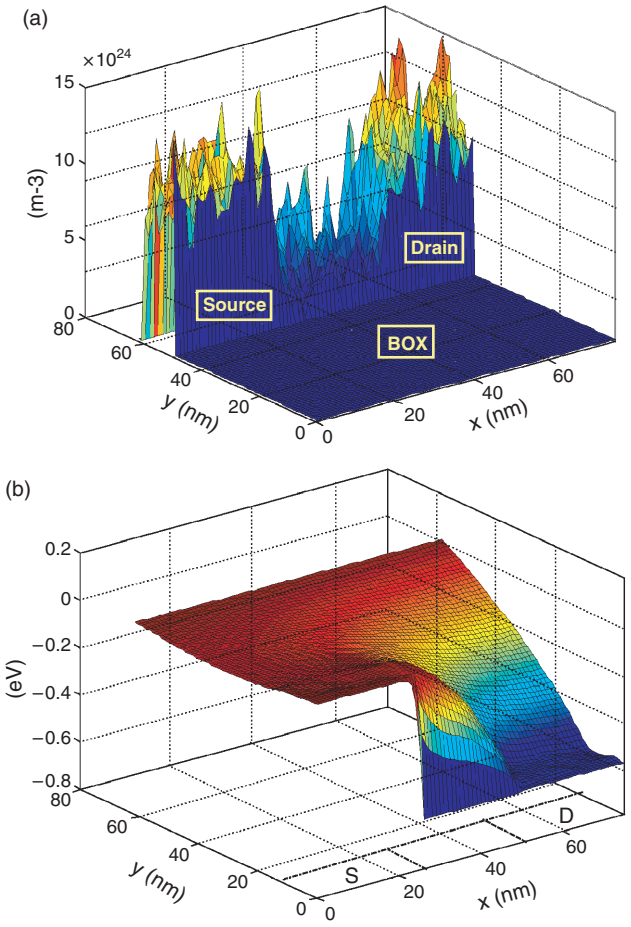

Fig. 19. (a) Snapshot of the Electron density in the device. We use $V_g = 0.6$ V and $V_d = 0.6$ V in these simulations. (b) Variation of the conduction band edge for the same bias conditions.

potential. A simple screening model is usually used that ignores the dynamical perturbations to the Coulomb fields caused by the movement of the free carriers. To overcome the above difficulties, several authors have advocated coupling of the semi-classical molecular dynamics approach to the ensemble Monte-Carlo approach.[46–48] Simulation of the low field mobility using such a coupled approach results in excellent agreement with the experimental data for high substrate doping levels.[48] However, it is proven to be quite difficult to incorporate this coupled ensemble Monte-Carlo-molecular dynamics approach when inhomogeneous charge densities, characteristic of semiconductor devices, are encountered.[45,49] An additional problem with this approach in a typical particle-based device simulation arises from the fact that both the *e–e* and *e–i* interactions are already included, at least within the Hartree approximation (long-range carrier–carrier interaction), through the self-consistent solution of the Poisson equation via the PM coupling discussed in the previous section. The magnitude of the resulting mesh force that arises from the force interpolation scheme, depends upon the volume of the cell, and,

REVIEW

for commonly employed mesh sizes in device simulations, usually leads to double-counting of the force.

To overcome the above-described difficulties of incorporation of the short-range *e–e* and *e–i* force into the problem, one can follow two different paths. One way is to use the P^3M scheme introduced by Hockney and Eastwood.[35] An alternative to this scheme is to use the corrected-Coulomb approach due to Gross et al.[50–53]

2.3.6.1. The P^3M Method. The particle-particle-particle-mesh (P^3M) algorithms are a class of hybrid algorithms developed by Hockney and Eastwood.[35] These algorithms enable correlated systems with long-range forces to be simulated for a large ensemble of particles. The essence of the method is to express the interparticle forces as a sum of two component parts; the short range part $\mathbf{F}_{sr}$, which is nonzero only for particle separations less than some cutoff radius $\mathbf{r}_e$, and the smoothly varying part $\mathbf{F}$, which has a transform that is approximately band-limited. The total short-range force on a particle $\mathbf{F}_{sr}$ is computed by direct particle–particle (PP) pair force summation, and the smoothly varying part is approximated by the particle-mesh (PM) force calculation.

2.3.6.2. The Corrected Coulomb Approach. This second approach is a purely numerical scheme that generates a corrected Coulomb force look-up table for the individual *e–e* and *e–i* interaction terms. To calculate the proper short-range force, one has to define a 3D box with uniform mesh spacing in each direction. A single (fixed) electron is then placed at a known position within a 3D domain, while a second (target) electron is swept along the 'device' in, for example, 0.2 nm increments so that it passes through the fixed electron. The 3D box is usually made sufficiently large so that the boundary conditions do not influence the potential solution. The electron charges are assigned to the nodes using one of the charge-assignment schemes discussed previously.[36] A 3D Poisson equation solver is then used to solve for the node or mesh potentials. At self-consistency, the force on the swept electron $\mathbf{F} = F_{mesh}$ is interpolated from the mesh or node potential. In a separate experiment, the Coulomb force $\mathbf{F}_{tot} = F_{coul}$ is calculated using standard Coulomb law. For each electron separation, one then tabulates F_{mesh}, F_{coul} and the difference between the two $F' = F_{coul} - F_{mesh} = \mathbf{F}_{sr}$, which is called the corrected Coulomb force or a short-range force. The later is stored in a separate look-up table.

As an example, the corresponding fields to these three forces for a simulation experiment with mesh spacing of 10 nm in each direction are shown in Figure 20. It is clear that the mesh force and the Coulomb force are identical when the two electrons are separated several mesh points (30–50 nm apart). Therefore, adding the two forces in this region would result in double-counting of the force. Within 3–5 mesh points, F_{mesh} starts to deviate from F_{coul}. When the electrons are within the same mesh cell, the mesh force approaches zero, due to the smoothing of the

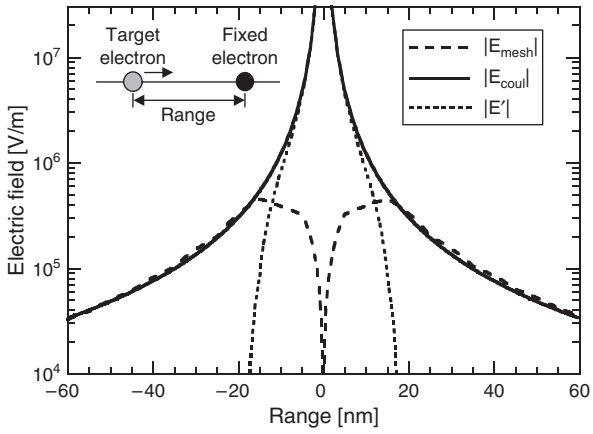

Fig. 20. Mesh, Coulomb and corrected Coulomb field versus the distance between the two electrons. Note: $F = -eE$.

electron charge when divided amongst the nearest node points. The generated look-up table for F' also provides important information concerning the determination of the minimum cutoff range based upon the point where F_{coul} and F_{mesh} begin to intersect, i.e., F' goes to zero.

Figure 21 shows the simulated doping dependence of the low-field mobility, derived from 3D resistor simulations, which is a clear example demonstrating the importance of the proper inclusion of the short-range electron–ion interactions. For comparison, also shown in this figure are the simulated mobility results reported in Ref. [17], calculated with a bulk EMC technique using the Brooks-Herring approach[54] for the *e–i* interaction, and finally the measured data[55] for the case when the applied electric field is parallel to the (100) crystallographic direction. From the results shown, it is obvious that adding the corrected Coulomb force to the mesh force leads to mobility values that are in very good agreement with the experimental data. It is also important to note that, if only the mesh force is used in the free-flight portion of the simulator, the simulation mobility data points are significantly higher than the experimental

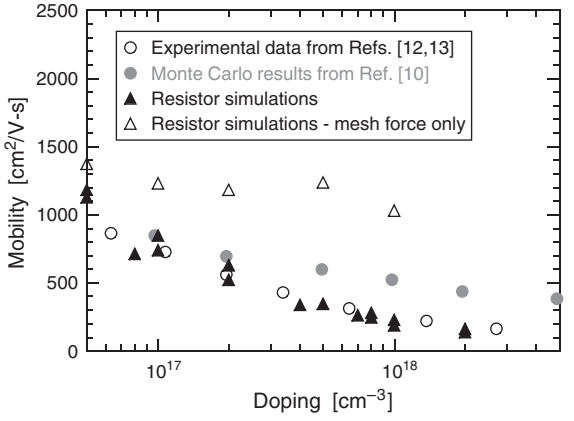

Fig. 21. Low-field electron mobility derived from 3D resistor simulations versus doping. Also shown on this figure are the Ensemble Monte Carlo results and the appropriate experimental data.

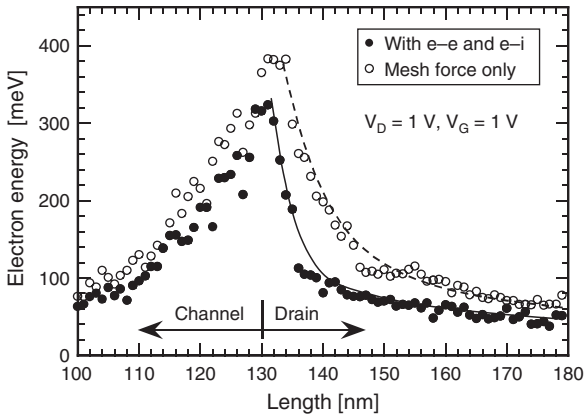

Fig. 22. Average energy of the electrons coming to the drain from the channel. Filled (open) circles correspond to the case when the short-range *e–e* and *e–i* interactions are included (omitted).

ones due to the omission of the short-range portion of the force.

The short-range *e–e* and *e–i* interactions also play significant role in the operation of semiconductor devices. For example, carrier thermalization at the drain end of the MOSFET channel is significantly affected by the short-range *e–e* and *e–i* interactions. This is illustrated in Figure 22 on the example of a 80 nm channel-length *n*-MOSFET. Carrier thermalization occurs over distances that are on the order of few nm when the *e–e* and *e–i* interactions are included in the problem. Using the mesh force alone does not lead to complete thermalization of the carriers along the whole length of the drain extension, and this can lead to inaccuracies when estimating the device on-state current.

3. QUANTUM TRANSPORT

Semiconductor transport in the nanoscale region has approached the regime of quantum transport. This is suggested by two trends: (1) within the effective-mass approximation, the thermal de Broglie wavelength for electrons in semiconductors is on the order of the gate length of nano-scale MOSFETs, thereby encroaching on the *physical optics* limit of wave mechanics; (2) the time of flight for electrons traversing the channel with velocity well in excess of 10^7 cm/sec is in the 10^{-15} to 10^{-12} sec region—a time scale which equals, if not being less than the momentum and energy relaxation times in semiconductors which precludes the validity of the Fermi's golden rule.

The static quantum effects, such as tunneling through the gate oxide and the energy quantization in the inversion layer of a MOSFET are also significant in nanoscale devices. The current generation of MOS devices has oxide thicknesses of roughly 15–20 Å and is expected that, with device scaling deeper into the nanoscale regime, oxides with 8–10 Å thickness will be needed. The most obvious quantum mechanical effect, seen in the very thinnest oxides, is gate leakage via direct tunneling through the oxide. The exponential turn-on of this effect sets the minimum practical oxide thickness (∼10 Å). A second effect due to spatial/size-quantization in the device channel region is also expected to play significant role in the operation of nanoscale devices. To understand this issue, one has to consider the operation of a MOSFET device based on two fundamental aspects: (1) the channel charge induced by the gate at the surface of the substrate, and (2) the carrier transport from source to drain along the channel. Quantum effects in the surface potential will have a profound impact on both, the amount of charge which can be induced by the gate electrode through the gate oxide, and the profile of the channel charge in the direction perpendicular to the surface (the transverse direction). The critical parameter in this direction is the gate-oxide thickness, which for a nanoscale MOSFET device is, as noted earlier, on the order of 1 nm. Another aspect, which determines device characteristics, is the carrier transport along the channel (lateral direction). Because of the two-dimensional (2D), and/or one-dimensional (1D) in the case of narrow-width devices, confinement of carriers in the channel, the mobility (or microscopically speaking, the carrier scattering) will be different from the three-dimensional (3D) case. Theoretically speaking, the 2D/1D mobility should be larger than its 3D counterpart due to reduced density of states function, i.e., reduced number of final states the carriers can scatter into, which can lead to device performance enhancement. A well known approach that takes this effect into consideration is based on the self-consistent solution of the 2D Poisson–1D Schrödinger–2D Monte Carlo, and requires enormous computational resources as it requires storage of position dependent scattering tables that describe carrier transition between various subbands.[56] More importantly, these scattering tables have to be re-evaluated at each iteration step as the Hartree potential (the confinement) is a dynamical function and slowly adjusts to its steady-state value. It is important to note, however, that in the smallest size devices (10 nm feature size), carriers experience very little or no scattering at all (ballistic limit), which makes this second issue less critical when modeling these nanoscale devices (e.g., Refs. [57–59]).

On the other hand, the dynamical quantum effects in nanoscale MOSFETs, associated with energy dissipating scattering in electron transport can be physically much more involved.[60] There are several other fundamental problems one must overcome in this regard. For example, since ultrasmall devices, in which quantum effects are expected to be significant, are inherently three-dimensional (3D), one must solve the 3D open-Schrödinger equation.

Another question that becomes important in nanoscale devices is the treatment of scattering processes. Within the Born approximation, the scattering processes are treated as independent and instantaneous events. It is, however, a

REVIEW

nontrivial question to ask whether such an approximation is actually satisfactory under high temperature, in which the electron strongly couples with the environment (such as phonons and other carriers). In fact, many dynamical quantum effects, such as the collisional broadening of the states or the intra-collisional field effect, are a direct consequence of the approximation employed for the scattering kernel in the quantum kinetic equation. Depending on the orders of the perturbation series in the scattering kernel, the magnitude of the quantum effects could be largely changed. Many of these issues relevant to quantum transport in semiconductors are highlighted in Table I. Note that at present there is no consensus as to what can be "the best" approach to model quantum transport in semiconductors. Density matrices, and the associated Wigner function approach, Green's functions, and Feynman path integrals all have their application strengths and weaknesses.

3.1. Open Systems

A general feature of electron devices is that they are of use only when connected to a circuit, and to be so connected any device must possess at least two terminals, contacts, or leads. As a consequence, every device is an open system with respect to carrier flow.[61] This is the overriding fact that determines which theoretical models and techniques may be appropriately applied to the study of quantum devices. For example, the quantum mechanics of pure, normalizable states, such as those employed in atomic physics, does not contribute significantly to an understanding of devices, because such states describe closed systems.

To understand devices, one must consider the unnormalizable scattering states, and/or describe the state of the device in terms of statistically mixed states, which casts the problem in terms of quantum kinetic theory. As a practical matter of fact, a device is of use only when its state is driven far from thermodynamic equilibrium by the action of the external circuit. The non-equilibrium state is characterized by the conduction of significant current through the device and/or the appearance of a non-negligible voltage drop across the device.

Table I. Quantum effects.

1. Static quantum effects
   - Periodic crystal potential and band structure effects
   - Scattering from defects, phonons
   - Strong electric and magnetic field
   - Inhomogeneous electric field
   - Tunneling-gate oxide tunneling and source-to-drain tunneling
   - Quantum wells and band-engineered barriers
2. Dynamical quantum effects
   - Collisional broadening
   - Intra-collisional field effects
   - Temperature dependence
   - Electron–electron scattering
   - Dynamical screening
   - Many-body effects
   - Pauli exclusion principle

In classical transport theory, the openness of the device is addressed by the definition of appropriate boundary conditions for the differential (or integro-differential) transport equations. Such boundary conditions are formulated so as to approximate the behavior of the physical contacts to the device, typically Ohmic or Schottky contacts.[62] In the traditional treatments of quantum transport theories, the role of boundary conditions is often taken for granted, as the models are constructed upon an unbounded spatial domain. The proper formulation and interpretation of the boundary conditions remains an issue, however. It should be understood that, unless otherwise specified, all models to be considered here are based upon a single-band, effective-mass open-system Schrodinger equation.

3.2. Evaluation of the Current Density

To investigate the transport properties of a quantum system one must generally evaluate the current flow through the system, and this requires that one examine systems that are out of thermal equilibrium. A common situation, in both experimental apparatus and technological systems, is that one has two (or more) physically large regions densely populated with electrons in which the current density is low, coupled by a smaller region through which the current density is much larger. It is convenient to regard the large regions as "electron reservoirs" within which the electrons are all in equilibrium with a constant temperature and Fermi level, and which are so large that the current flow into or out of the smaller "device" represents a negligible perturbation. The reservoirs represent the metallic contacting leads to discrete devices or experimental samples, or the power-supply busses at the system level. Consequently the electrons flowing from a reservoir into the device occupy that equilibrium distribution which characterizes the reservoir. In a simple one-dimensional system with two reservoirs, the electrons flowing in from the left-hand reservoir have $k > 0$ and those flowing from the right-hand reservoir have $k < 0$. Then, the current density is calculated using the Tsu-Esaki formula

$$J = q\int_{V_0}^{\infty} \frac{dE_{\|}}{2\pi\hbar} T(E_{\|}) \ln\left\{\frac{1+\exp\left[-\beta\left(E_{\|}-E_{Fl}\right)\right]}{1+\exp\left[-\beta\left(E_{\|}-E_{Fr}\right)\right]}\right\} \quad (40)$$

where $E_{\|}$ is the in-plane carrier energy, $\beta = 1/k_{\mathrm{B}}T$, $T(E_{\|})$ is the transmission coefficient and E_{Fl} and E_{Fr} are the Fermi levels of the left and the right lead, respectively. Note that this expression is valid in general with respect to the dispersion relation in the x direction, but requires a parabolic dispersion relation in the transverse directions. The separation of variables leading to Eq. (40) is never rigorously valid in a semiconductor heterostructure. The reason for this is that the transverse effective mass $m_{\perp}^{*}$ will vary with semiconductor composition, which varies

in the x direction. In principle, one must do at least a two-dimensional integral (if axial symmetry holds, otherwise a three-dimensional integral). Nevertheless, Eq. (40) is widely used to model the current density in heterostructure devices.

If the transverse dimensions are constrained, but separation of variables is still possible, the transverse motion of the electrons consists of a discrete set of standing waves or normal modes. Such systems are referred to as "one-dimensional" systems, quantum wires, or electron waveguides. The symbol $k_\perp$ is now interpreted as an index for the discrete transverse modes, and the expression for the current density now becomes

$$J_r = 2q \sum_{k_\perp} \int_0^{-\infty} \frac{dE_\parallel}{2\pi} T(k_\parallel, k_\perp) \times \left[f_{FD}(E_\parallel + E_\perp - E_{Fl}) - f_{FD}(E_\parallel + E_\perp - E_{Fr}) \right] \quad (41)$$

3.3. Landauer-Buttiker Formalism and Related Numerical Methods

The Landauer-Buttiker formalism,[63, 64] is the most widely used scheme to calculate the ballistic transport through an open system. Within this approach, current is calculated using the transmission function that is obtained from the solution of the Schrödinger equation with scattering boundary conditions. While this is a significant simplification compared to a rigorous calculation that includes the relaxation of carriers, even this approach becomes computationally very challenging for higher dimensional nanostructures with complex geometry. A number of methods have been developed in the past to calculate the ballistic transport through quantum devices. *Transfer matrix method*,[65, 66] a well-known approach appears to be unstable[67] for larger devices in its original form. However, this drawback was overcome by a series of generalizations developed by Frensley,[68] Lent et al.,[69] and Ting et al.[70] respectively. These approaches take into account the coupling to the leads using the quantum transmitting boundary method (QTBM),[69] and can handle structures of arbitrary geometry. There are QTBM implementations applied to one-dimensional tight-binding Hamiltonian,[70, 71] $k \cdot p$-based multi-band calculations,[72] and two-dimensional single-band calculations.[69, 73] A three-dimensional self-consistent scheme based on QTBM on the assumption of separable device potential has been reported.[74] The boundary element method[75] is computationally efficient, but so far the published applications are limited to wave-guide structures, i.e., structures possessing a flat potential[76] or consisting of piecewise homogeneous materials with constant potentials.[77] The *Recursive Green's function* (RGF) method,[1, 14] an efficient and widely used algorithm, has been successfully implemented for two-dimensional devices,[78, 79] and for small three-dimensional structures such as nano-wires.[80] It is very well suited for 2-terminal devices that can be discretized into cross-sectional slices with nearest neighbor interactions, but has difficulties dealing with additional (i.e., more than two) contacts.[81] A closely related approach, *modular recursive Green's function* method[82] is applicable to devices that can be divided into regions of sufficiently high symmetry, where the Schrödinger equation is separable, and has been recently adopted to include magnetic fields.[83] Another method that is applicable in a case of separable device potentials and when current can flow through two leads (e.g., in the situation of quasi-1D transport) has been termed the *mode-space* approach.[84] This method has been implemented using effective-mass approximation in simulator NanoMOS ($2 \cdot x$ and 3.0) that has been extended recently to include the simulation of devices with channels along arbitrary crystallographic directions.[85, 86] A 3D simulator with effective mass approximation using the mode-space approach for silicon nano-transistor has been reported where scattering is taken into account via Buttiker probes.[87] An application of the mode-space approach to the simulations of ultra small FinFET has been presented[88] in which phonon scattering and surface roughness have been taken into account. Very recently a 3D simulation of silicon nanowires, based on effective-mass approximation and the mode-space approach, has been presented.[89] In that work the simulation has been performed assuming that the device potential is separable in the confinement (mode representation) and the transport direction, along which the potential is assumed to be close to uniform. The resulting quasi-1D transport problem is solved using a simplified NEGF formalism self-consistently. The inter- and intra-valley scattering on phonons has been included in that work[89] via deformation potential theory, which allowed the authors to check the validity of ballistic model on 15 nm gate wire transistor. Finally, a modified version of the QTBM has been developed that expands the scattering solutions in terms of two different closed system wave functions in an efficient way.[90] This 2D scheme is charge-self consistent and has also been implemented for effective-mass Hamiltonians and different crystallographic directions.[91] One of the advantages of this approach is that the transport is considered to be "truly 2D," i.e., the corresponding quantum transport equation is *not* assumed to be separable in the confinement direction. Therefore, using the Laux's method,· gate leakage currents can be calculated self-consistently with the rest of the device.

Thus, despite the significant progress in developing quantum transport simulator, at present there is no simulators that treat the quantum-mechanical transport in three spatial dimensions rigorously, and only few simulators treat transport as truly 2D. In the following, we describe an approach that goes beyond quasi-1D transport modeling and allows us to take into account gate-leakage and other 2D and 3D transport effects.

REVIEW

3.4. Contact Block Reduction Method

An efficient method based on Green's function approach, termed as Contact Block Reduction (CBR) method,[81, 92, 93] that is presented next has been developed at Walter Schottky Institute and ASU and used by the group from ASU to calculate self-consistently transport properties in nanoscale 10 nm gate length FinFET device operating in the ballistic regime. The method rigorously separates the open system problem into the solution of a suitably defined closed system (energy-independent) eigen-problem and the energy-dependent solution of a small linear system of equations of size determined by the contact regions that couple the closed system to the leads. The calculation of the charge density of the open system throughout the device can be performed with an effort comparable to a single calculation of a small percentage of the eigenstates of a closed system.

The CBR method allows one to calculate 2D or 3D ballistic transport properties of a device that may have any shape, potential profile, and most importantly any number of external leads. In this method, quantities like the transmission function and the charge density of an open system can be obtained from the eigenstates of the corresponding closed system defined as $H^0|\alpha\rangle = \varepsilon_\alpha|\alpha\rangle$, and the solution of a very small linear algebraic system for every energy step E. The retarded Green's function $\mathbf{G}^R(E)$ can be calculated via the Dyson equation through a Hermitian Hamiltonian $\mathbf{H^0}$ of a closed system represented by,[94]

$$\mathbf{G}^R(E) = \mathbf{A}^{-1}(E)\mathbf{G^0}(E), \quad \mathbf{A}(E) \equiv \left[\mathbf{I} - \mathbf{G^0}(E)\Sigma(E)\right]$$
$$\mathbf{G^0}(E) \equiv \left[\mathbf{I}E - \mathbf{H}^0\right]^{-1} = \sum_\alpha \frac{|\alpha\rangle\langle\alpha|}{E - \varepsilon_\alpha} \tag{42}$$

The inversion of the matrix $\mathbf{A}$ can be easily performed using the property of the self-energy Σ in real space representation: it is non-zero only at boundary regions of the device, which are in ***contact*** with the external leads. We denote these boundary regions (=*contacts*) with index C, and the rest of the device with index D. As a result, the Green's function matrix of the open system can be written in the following form:

$$\mathbf{G}^R = \begin{bmatrix} \mathbf{G}^R_C & \mathbf{G}^R_{CD} \\ \mathbf{G}^R_{DC} & \mathbf{G}^R_D \end{bmatrix} = \begin{bmatrix} \mathbf{A}_C^{-1}\mathbf{G}^0_C & \mathbf{A}_C^{-1}\mathbf{G}^0_{CD} \\ -\mathbf{A}_{DC}\mathbf{A}_C^{-1}\mathbf{G}^0_C + \mathbf{G}^0_{DC} & -\mathbf{A}_{DC}\mathbf{A}_C^{-1}\mathbf{G}^0_{CD} + \mathbf{G}^0_D \end{bmatrix} \tag{43}$$

The left-upper matrix block $\mathbf{G}^R_C = \mathbf{A}_C^{-1}\mathbf{G}^0_C$ fully determines the transmission function whereas the left-lower block $\mathbf{G}^R_{DC}$ determines the density of states, charge density, etc. The particle density $n(\mathbf{r})$ can be obtained using,

$$n(\mathbf{r}) = \sum_{\alpha,\beta} \langle \mathbf{r} \mid \alpha\rangle\langle\beta \mid \mathbf{r}\rangle \xi_{\alpha\beta} \tag{44}$$

where $\xi_{\alpha\beta}$ is the density matrix and is given by,

$$\xi_{\alpha\beta} = \sum_{\lambda=1}^{L} \int \Xi^{(\lambda)}_{\alpha\beta}(E) f_\lambda(E)\, dE$$
$$\Xi^{(\lambda)}_{\alpha\beta}(E) = \frac{1}{2\pi} \frac{Tr\left(\left[|\beta\rangle\langle\alpha|\right]_C \mathbf{B}_C^{-1}\Gamma_C^{(\lambda)}\mathbf{B}_C^{-1\dagger}\right)}{(E - \varepsilon_\alpha + i\eta)(E - \varepsilon_\beta - i\eta)}\Bigg|_{\eta\to 0+} \tag{45}$$
$$\Gamma_C = i\left[\Sigma_C - \Sigma_C^\dagger\right], \qquad \mathbf{B}_C = \mathbf{1}_C - \Sigma_C \mathbf{G}^0_C$$

In Eq. (45), L denotes the total number of external leads of the device, index λ denotes individual lead number and $f_\lambda(E)$ is the distribution function associated with lead λ. The integration in Eq. (45) is performed over the energy interval, where both the density matrix distribution $\Xi^{(\lambda)}_{\alpha\beta}(E)$ and the distribution function $f_\lambda(E)$ are non-negligible. Consequently, the density matrix distribution defines the lower integration limit, and the distribution function $f_\lambda(E)$ the upper integration limit. The advantage of using Eqs. (43)–(45) for determining electron density is in *splitting numerical costs* between calculation of position-independent density matrix and position-dependent, but energy-independent charge density in Eq. (44). Then the total numerical cost can be estimated as $N_{n(\mathrm{r})} = N^2_{\mathrm{eigen}} N_E + N^2_{\mathrm{eigen}} N_{\mathrm{grids}}$, where N_E is number of energy steps, N_{eigen} is number of eigenstates to be used, and N_{grids} is the number of grid points in real space.[81] Note the absence of a large terms like $N_E \times N_{\mathrm{grids}}$.

However, a slightly different approach to calculate particle density can be adopted that is also very efficient. This approach appears to be more suitable for self-consistent calculation. For a self-consistent calculation using a predictor-corrector approach described below, it is important to have an expression for the local density of states (LDOS), $\rho(\mathbf{r}, E)$. To obtain the expression for the LDOS using CBR algorithm, we note that the lower-left block $\mathbf{G}^R_{DC}$ of the matrix in Eq. (43) can be also written in the following form,

$$\mathbf{G}^R_{DC} = \mathbf{G}^0_{DC}\mathbf{B}_C^{-1} \tag{46}$$

Next, using the formula $\rho(\mathbf{r}, E) = \langle \mathbf{r}|\mathbf{G}^R\Gamma\mathbf{G}^{R\dagger}|\mathbf{r}\rangle/2\pi$ and performing simple algebraic manipulations one gets

$$\rho(\mathbf{r}, E) = \frac{1}{2\pi}\sum_m |G^R_{\mathbf{r}m}|^2 \Gamma_{mm}$$
$$G^R_{\mathbf{r}m} = \langle r|\mathbf{G}^0_{DC}\mathbf{B}_C^{-1}|m\rangle \tag{47}$$
$$= \sum_{m',\alpha} \frac{\langle \mathbf{r} \mid \alpha\rangle\langle\alpha \mid m'\rangle}{E - \varepsilon_\alpha} \langle m'|\mathbf{B}_C^{-1}|m\rangle$$

The term $G^R_{\mathbf{r}m}$ in Eq. (47) is the retarded Green's function in a mixed space and mode representation,[81] and the second line in this equation is the CBR expression for it. One can check now that the total numerical cost of LDOS using Eq. (47) can be estimated as

$$N_\rho = N_E N_{\mathrm{grids}} N_{\mathrm{eigen}} N_{\mathrm{modes}} \tag{48}$$

where N_{modes} is the number of non-zero elements in Γ_C, which is diagonal in the mode representation (thus N_{modes} is the number of propagating modes[81]). It is usually much more efficient to express the quantities with index C (contacts) in mode representation, due to the possible mode reduction. The advantage of using Eq. (47) is the absence of quadratic and higher order terms with N_{grids} or N_{eigen}.

3.4.1. Bound States Treatment

It is important to note that the density matrix $\xi_{\alpha\beta}$ in Eq. (45) and the derived quantities may also account for bound states, if they are present in the system. Indeed, as it has been shown in Ref. [81], the term $\Xi^{(\lambda)}_{\alpha\beta}(E)$ does not disappear when the coupling to the leads (represented by Σ_C and Γ_C terms) is zero (i.e., when system states are not coupled to the outside world), but instead results in

$$\Xi_{\alpha\beta}(E) \xrightarrow[\Sigma(E)=i\eta\to 0+]{} \delta_{\alpha\beta}\delta(E-E_\alpha) \tag{49}$$

that assures the inclusion of bound states into the total charge density. We point out, however, that in the case of numerical evaluation of Eq. (49), the delta-functions corresponding to the bound states should be integrated analytically, leading to the expression

$$\xi^{\text{TOTAL}}_{\alpha\beta} = \sum_{\gamma\in BS}\delta_{\alpha\beta}\delta_{\alpha\gamma}f(\varepsilon_\gamma) + \sum_{\lambda=1}^{L}\int \Xi^{(\lambda)}_{\alpha\beta}(E)f_\lambda(E)\,dE \tag{50}$$

where the sum with index γ is performed over all bound states (BS) in the system. While in the *idealized* ballistic case, it is generally unclear how these states are occupied if the bias is applied, however, in a presence of small scattering in the system these quasi-bound states (QBS) can be viewed as states that get occupied as a result of scattering of carriers coming from one of the leads $\lambda = 1 \ldots L$. In the later case, if one knew '*from what lead has a carrier come from*,' one could assign to the carrier the corresponding distribution function. Exploring this idea, one can make an assumption that the distribution function $f(\varepsilon_\gamma)$ of the quasi-bound state $|\gamma\rangle$ depends on the "coupling strength" to the outside leads. If a quasi bound state $|\gamma\rangle$ is coupled more strongly to lead λ, then it is reasonable to expect that its distribution function is close to the one of lead λ. Generally, one can speculate that if the scattering is small, then the quasi-bound states can be occupied according to the following approximate formula

$$\xi^{BS}_{\alpha\beta} = \sum_{\gamma\in BS}\delta_{\alpha\beta}\delta_{\alpha\gamma}\sum_{\lambda=1}^{L}F_{\gamma\lambda}f_\lambda(\varepsilon_\gamma)\Big/\sum_{\lambda=1}^{L}F_{\gamma\lambda} \tag{51}$$

where the coupling strength, $F_{\gamma\lambda}$, of state γ to lead λ is given by

$$F_{\gamma\lambda} = \sum_{m=1}^{M_\lambda}|\langle\gamma\,|\,\chi^{(\lambda)}_m\rangle|^2 \tag{52}$$

The summation in Eq. (52) is performed as the squares of the absolute values of projections of states $|\gamma\rangle$ over M_λ transverse modes $\chi^{(\lambda)}_m$ in lead λ. $F_{\gamma\lambda}$ can be used to determine what states $|\gamma\rangle$ should be treated as "quasi-bound" ones. We find this approach to be essential, in particular, for a superior convergence of the self-consistent cycle. An example of using the coupling strength for determining the quasi-bound states is given in Figure 23. The solid circles represent the coupling strength $F_{\gamma\lambda}$ of an eigenstate $|\gamma\rangle$ to the lead λ (for simplicity data for only one (source) lead are shown on Fig. 23). We see that the vast majority of eigenstates are strongly coupled to the lead, except the lower 6 circles, for which $F_{\gamma\lambda} < 0.2$. It is possible, therefore, to introduce a threshold in coupling strength (for example $F_{\text{th}} = 0.19$), so that eigenstates with coupling strength less than the threshold would be identified as QBS. Furthermore, every peak in the DOS corresponds to a certain QBS (there are 6 peaks and 6 QBS shown in Fig. 23). While the former property is not always the case (some QBS do not result in resonant peaks in the DOS), it is generally possible to find a QBS "responsible" for every resonant peak in the DOS. Therefore, most of the hard-to-integrate resonant peaks in the DOS can be eliminated, by excluding the responsible weakly-coupled eigenstates from the eigenstate set $\{|\alpha\rangle\}$, which we use to calculate the retarded Green's function G^0 of a closed device. These excluded states then are taken into account with the following resulting expression for the charge density:

$$n(\mathbf{r}) = \sum_{\alpha\in BS}|\langle\mathbf{r}\,|\,\alpha\rangle|^2\xi^{BS}_{\alpha\alpha} + \sum_{\alpha,\beta\notin BS}\langle\mathbf{r}\,|\,\alpha\rangle\langle\beta\,|\,\mathbf{r}\rangle\xi_{\alpha\beta} \tag{53}$$

If the explicit relation between the charge density and the LDOS is desired, the following formula can be use

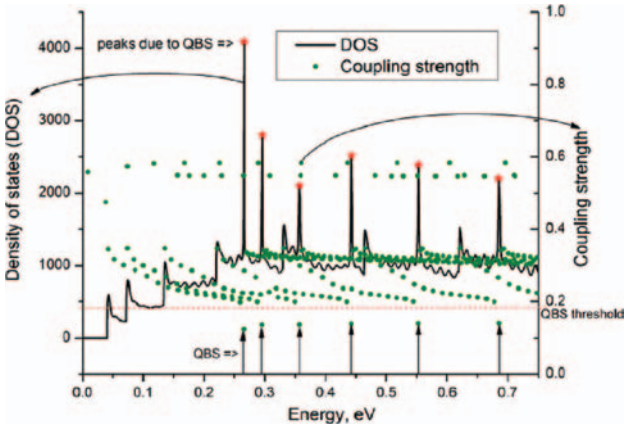

Fig. 23. Quasi-bound state (QBS) detection using the coupling strength in Eq. (52). The graph shows DOS energy dependence (solid curve, left-hand scale) and the coupling strength for the device eigenstates (solid circles, right-hand scale). Note the resonant peak of open-system DOS at each QBS.

instead:

$$n(\mathbf{r}) = \sum_{\alpha \in BS} |\langle \mathbf{r} \mid \alpha \rangle|^2 \xi^{BS}_{\alpha\alpha} + \sum_{\lambda=1}^{L} \int \rho_\lambda(\mathbf{r}, E) f_\lambda(E)\, dE \quad (54)$$

where the LDOS due to lead λ is $\rho_\lambda(\mathbf{r}, E)$.

3.4.2. Energy Discretization

For an efficient numerical implementation of a self-consistent scheme, the choice of the energy grid is of high importance. To integrate the continuous part of the carrier density, the LDOS is discretized in energy space and then a simple numerical integration is done by summing up the values for each energy step weighted by the Fermi distribution and the energy grid spacing ΔE_k with k being the index of the energy grid. Using a regular grid with constant grid spacing, the integral over the peak deriving from the resonant states is very poor since the relative distance between the nearest energy grid point E_k and the resonant energy E_m is, generally, arbitrary. In addition, the resonant energy is slightly shifted with each iteration step, leading to a varying integration error during the self-consistent cycle, which acts as an obstacle against convergence for any self-consistent algorithm. Thus, a solution to this problem is to use the physical information about the system and employ an *adaptive energy grid* that resolves each known peak with a local energy grid of a few tens of grid points that is fixed to the resonant energy E_m. The location of resonant states is easy to find, since the resonant energies are close to (selected) eigenstates of the closed system. Another advantage of the CBR method is that these eigenstates are already known, since in this method the solution for the open system is being expressed in the basis of the closed system. As a result, the integration error is reduced compared to the case of using regular grid and remains constant within the iteration, since the grid is locally fixed to the shifted mode energies.

3.4.3. Self-Consistent Solution

The self-consistent solution of the ballistic or quasi-ballistic transport properties of an open device requires repeated solution of the Schrödinger and Poisson equations. In principle, it is possible to simply iterate the solution of the Schrödinger and Poisson equations and with enough damping this will yield a converged result. However, this approach leads to hundreds of iteration steps for each bias point that do not pose a reasonable scheme. To improve the convergence of a highly non-linear set of coupled equations, such as the Schrödinger-Poisson problem, the Newton algorithm is usually the first choice. However, the exact Jacobian for the Schrödinger-Poisson set cannot be derived analytically, and its numerical evaluation is rather costly (while certainly possible, see e.g., Ref. [90]). In the case of a closed system this problem has been solved using the *predictor-corrector* approach.[95, 96] The aim of this method is to find a good approximation for the quantum density as a function of the electrostatic potential where an expression for the Jacobian is known. In this work we adopted this approach to open systems. At first, the Schrödinger equation is solved for the closed system with the Hartree potential, $\varphi_H(\mathbf{r})$, and the exchange and correlation potential, $\varphi_{XC}(\mathbf{r})$ taken into account. Then the local density of states $\rho(\mathbf{r}, E)$ of the open system is calculated using the CBR method. The Hartree potential φ_H and carrier density n are then used to calculate the residuum, F, of the Poisson equation using,

$$F[\varphi_H] = \mathbf{A}\varphi_H - (n - N_D) \quad (55)$$

where $\mathbf{A}$ is the matrix derived from the discretization of the Poisson equation. If the residuum is smaller than a predetermined threshold the solution is taken to be a converged one. If the residuum is still too large, the correction to the Hartree potential $\Delta\varphi_H(\mathbf{r})$ is calculated in the predictor step, where the *predictor* carrier density $n_{pr}(\mathbf{r})$ is calculated, assuming it to be the functional of the change $\Delta\phi_H(\mathbf{r})$ in the Hartree potential as follows:

$$\begin{cases} n_{pr}(\mathbf{r}) = 2\sum_{\lambda=1}^{L} \int \rho_\lambda(\mathbf{r}, E) f\left(\dfrac{E + \Delta\varphi_H(\mathbf{r}) - E_F^{(\lambda)}}{k_B T}\right) dE \\ \mathbf{A}(\varphi_H(\mathbf{r}) + \Delta\varphi_H(\mathbf{r})) = n_{pr}(\mathbf{r}) - N_D(\mathbf{r}) \end{cases} \quad (56)$$

where $f(x) = [1 + \exp(x)]^{-1}$ for 3D systems or the corresponding Fermi integral for systems with lower dimensions, the energy $E_F^{(\lambda)}$ is the Fermi energy level in lead λ, and a factor of 2 is taken into account for the spin degeneracy of the electrons. Note that the Jacobian for the system Eq. (56) can be easily found analytically:

$$\begin{aligned} J_{\mathbf{r}\mathbf{r}'} &= \frac{\partial F(\mathbf{r})}{\partial \Delta\varphi_H(\mathbf{r}')} = \mathbf{A}_{\mathbf{r}\mathbf{r}'} + \frac{\partial n_{pr}(\mathbf{r})}{\partial \Delta\varphi_H(\mathbf{r}')} \\ &= \mathbf{A}_{\mathbf{r}\mathbf{r}'} + \delta_{\mathbf{r}\mathbf{r}'} \frac{2}{k_B T} \sum_{\lambda=1}^{L} \int \rho_\lambda (f-1) f\, dE \end{aligned} \quad (57)$$

After applying the Newton method, the obtained correction to the Hartree potential $\Delta\varphi_H$ and the corresponding carrier density are used to update the Hartree φ_H, exchange μ_X^{LDA} and correlation μ_C^{LDA} potentials for the next iteration $(i+1)$ as follows:

$$\begin{aligned} \varphi_H^{(i+1)} &= \varphi_H^{(i)} + \Delta\varphi_H^{(i)} \\ \varphi_{XC}^{(i+1)} &= \mu_X^{\mathrm{LDA}}[n_{pr}^{(i)}[\Delta\varphi_H^{(i)}]] + \mu_C^{\mathrm{LDA}}[n_{pr}^{(i)}[\Delta\varphi_H^{(i)}]] \end{aligned} \quad (58)$$

The loop is repeated until convergence is achieved, that is $|\Delta\varphi_H^{(i)}| < \varepsilon$, with ε being the absolute error of the potential. We find that typically only very few (5–7) solutions of the Schrödinger equation are necessary to yield a solution with 3 converged digits in the potential and currents.

3.4.4. Device Hamiltonian, Algorithm and Some Numerical Details

In this work FinFET devices with varying fin width (4 nm∼12 nm) have been simulated. With 12 nm fin width the simulation real space domain is fairly large. While the CBR method for quantum transport simulation can be used with any multi-band Hamiltonians, including the tight-binding and $k \cdot p$,[81] in this work, we choose to adopt the effective mass model and finite difference discretization scheme to be able to simulate relatively 'large' FinFET device within a reasonable time frame. The structure and the size of the corresponding effective mass Hamiltonian are determined by the dimensionality of the transport problem and the number of real space grid points. Due to the presence of non-equivalent valleys in Si, we need to solve the open-system problem for each valley, and then add up the contributions from different valleys (weighting them with the corresponding valley degeneracy).

In ultra-scaled nano-transistors source, drain and gate regions are usually heavily doped, therefore it is important to include quantum-mechanical effects of exchange and correlation. In this work this is done via the local density approximation (LDA). The phenomenological scattering on the phonons using the relaxation time approximation has been taken into account. Since this phonon scattering model relies on phenomenological parameters, in this work we present results that include into account this phenomenological scattering on phonons as well as purely ballistic ones (that do not depend on such parameters).

After the initial guess for the potential and the *initial* number of device eigenstates, the CBR loop is started. For each CBR-Poisson iteration the following tasks are performed: (i) transverse lead modes are calculated; (ii) eigenproblem is solved for closed-system with von Neumann boundary conditions at the contacts; (iii) open-system solution is constructed. The simulator has been modified to incorporate the automatic determination of the required number of device eigenstates and lead modes for each iteration to yield desired accuracy. Due to this dynamic nature of eigenstate and lead modes determination, CPU time can be saved and also memory requirements have been optimized.

The accuracy ε also determines the upper error norm for the functional F; if $\|F\| < \varepsilon$ then the solution is considered to be converged and the next bias point can be processed, otherwise the predictor-corrector approach is invoked to determine correction $\Delta\varphi$ to the potential. With updated potential φ CBR routine is called again and the loop continues until convergence is achieved. Note that the CBR module is called for each non-equivalent Si valley to obtain the LDOS and transmission function for each valley; then the total charge density, currents, etc. are calculated as the corresponding sums. Table II shows the average values of required number of device eigenstates and lead modes in off- and on-state of a FinFET device being simulated. One can see that the CBR method allows us to use a small fraction of device eigenstates and lead modes to get a well-converged solution within 5–6 iterations, on average. It is significant that this excellent rate of convergence has been observed on a wide variety of devices with different doping profiles and geometries. However, in order to achieve this result, a combination of *all* the steps has to be performed. For example, in the absence of QBS detection, the average number of iterations would be about 20–30, and in some cases there could be no converged solution at all (see also Ref. [90]). Similarly, it would be significantly harder to achieve any convergence in the absence of adaptive energy discretization, etc. However, we find that the full scheme presented in this section, resolves convergence problems in most cases. As a real-life example, the convergence of the non-linear Poisson equation for a FinFET with different gate voltages changing from +0.2 to −1.0 V and fixed drain-to-source voltage (0.1 V) is shown in Figure 24. The corresponding

Table II. Convergence data and average number of generalized von Neumann eigenstates used for construction of open system solution.

Parameter\operation regime	Subthreshold	On-state
Number of grid points/mesh size	17169/2.5 Å	17169/2.5 Å
Number of device eigenstates used in calculation (averaged over valleys)	470 (2.7%)	270 (1.5%)
Number of total lead transverse modes used in calculation (averaged over valleys)	39 (20%)	31 (16%)
Average absolute error of potential (eV)/average number of converged digits of the current	10^{-5}/3	10^{-5}/3
Average number of CBR-Poisson iterations	5	6

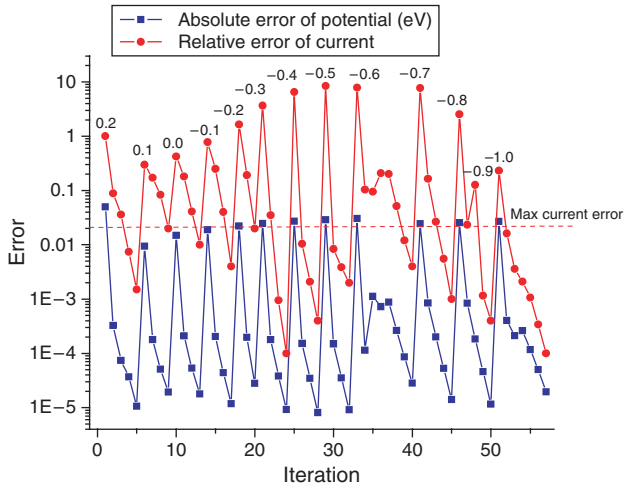

Fig. 24. Residuum of non-linear Poisson equation and corresponding relative error in drain current (w.r.t. converged solution of a higher order) for different gate voltages. The gate voltage values are shown above of each segment of the curve.

REVIEW

error in the source-drain current is also plotted in the same figure. The maximum error in the source-drain current values is 2% for the potential accuracy fixed at 2×10^{-5} eV. No convergence-tuning parameters of any kind have been used in the simulation: the energy grid, energy cut-off, number of eigenstates, lead modes, etc., are automatically determined by the CBR simulator in every iteration and for each bias point.

3.4.5. Simulation Example

Over the decade many novel structures have been proposed for the nanoscale regime of operation, among them fully depleted MOSFETs, in particular Double-gate (DG) MOSFETs emerged as the leading candidate for the ultimate scaling of silicon MOSFETs down to 10 nm. In these devices effective control of the gate over the channel has been enhanced by using multiple gates and thinning of body thickness.[97] For a given insulator thickness theoretical study shows that DG devices can be scaled to the lowest channel length keeping the short channel effects within acceptable limits.[98] Theoretically, cylindrical or surround-gate MOSFET is found to show the best gate control of channel but realization of this structure from fabrication point of view is quite challenging.[99, 100] Different orientation of double-gate MOSFETs have been proposed[101] as shown in Figure 25. In type I device[102] the current direction is in plane but gate-to-gate direction is normal to the wafer plane. The fabrication process with this type of devices is complex and contacting the bottom gate is rather difficult. Type II devices[99] have gate-to-gate direction in plane but the current direction is perpendicular to the plane. This type of devices suffers from inability to easily control the channel and source/drain doping profiles.[103] Type III devices[104] have the advantage of both in-plane gate-to-gate direction and in-plane current direction but the width of the device is normal to the plane.

The major disadvantages of these double gate MOSFETs are (i) non-planer structure as opposed to the planar structure of conventional bulk MOSFETs (ii) self-alignment of the gates with each other and with source/drain and, (iii) formation of ultra thin silicon film. FinFET[105–107] is a special category of type III devices in which the height is reduced to maintain quasi-planar topography for the ease of fabrication.[108] In FinFETs gates are automatically self-aligned with each other[105] and also the packing density is large compared to other double-gate structures.[101]

The geometry of a typical FinFET device is shown in Figure 26. The fin thickness, t_{Si}, is considered to be the most important process parameter as it controls the carrier mobility as well as threshold voltage. The fin is made thin enough when viewed from above, as shown in Figure 26(b), so that both gates simultaneously control the entire fully depleted channel film. Usually the top surface of the fin is covered by a thicker oxide compared to the thickness of the side gates (front and back), t_{ox}; therefore channels form only along the vertical surfaces of the fin. The fin height, h here is equivalent to the "gate width" of the conventional bulk MOSFET. Therefore, the effective channel width in FinFET devices is equal to $2h$ when only side gates are considered. For higher drive current different channel width is achieved by introducing multiple fins in parallel. In that case, the resultant width of the channel can be represented as $2 \times h \times N_{fins}$ with N_{fins} being the number of fins.

In this work we have modified our 2D CBR simulator in such a way that semiconductor devices on wafers

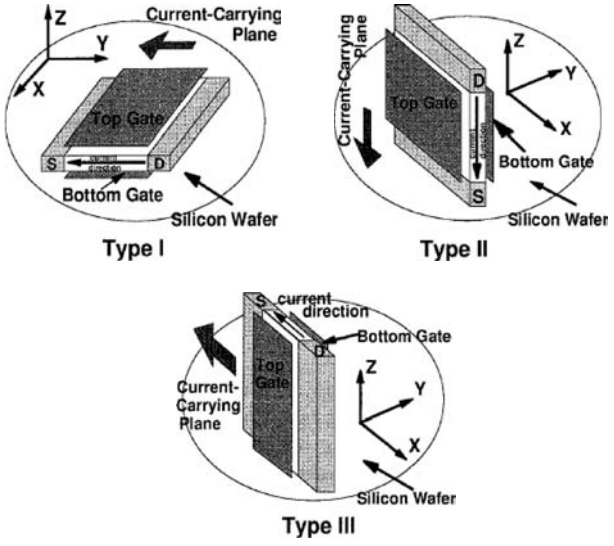

Fig. 25. Three possible orientations of DG MOSFETs in silicon wafer. Adapted with permission from [101], H.-S. P. Wong et al., *IEDM Tech. Dig.* 427 (**1997**). © 1997.

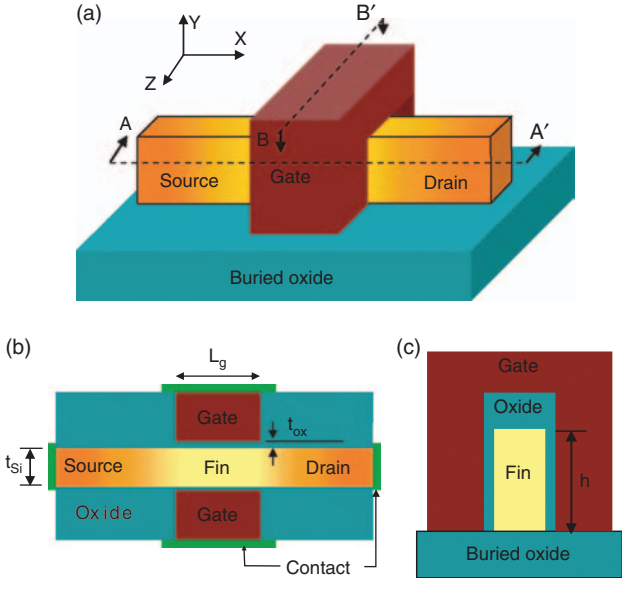

Fig. 26. (a)—3D schematic view of a prototype FinFET, (b)—top view along A-A′ cross section and, (c)—side view along B-B′ cross-section.

of *arbitrary crystallographic orientation* can be simulated. This was necessary to match the experimental data[109] for a FinFET device of which the channel is on (110) wafer plane. The conventional approach assumes wafer in (001) plane, and with the real space axes X, Y, Z being along crystallographic directions [100] and [010] and [001] respectively, the effective mass tensor is diagonal and the Schrödinger equation can be discretized and solved accordingly. However, for FinFET devices with channel oriented in (110) wafer plane, the effective mass tensor is non-diagonal (e.g., Ref. [86]). The resulting Schrödinger equation has mixed second derivative and first derivatives terms of which the coefficients are the non-diagonal element of effective mass tensor. Considering 2D simulation, it is possible to eliminate the mixed second derivative term by rotating the device in real space by a suitable angle.[91] The first derivative term can be eliminated with the wave-function change of variable after the elimination of second derivative terms.[91] As a result, to simulate FinFET devices with channel orientation in (110) wafer plane it is sufficient to use modified effective masses along device coordinates.[86] Note that the above procedures are rigorously valid for 2D (and 1D) transport simulations; a full 3D simulation with the wave-function depending on the device depth (e.g., fin height) would require a somewhat different treatment of the coefficients in the discretized Schrödinger equation containing effective masses. Regarding 2D simulation, however, we assume that the wave-functions depend on device length and width directions, but neglect the explicit height dependence, thus assuming that the *transport* in this 3D FinFET device is two-dimensional (2D).

With the inclusion of the modifications specified above, sets of 2D simulation have been performed in order to match experimental data with fin thickness of 12 nm and physical gate oxide thickness of 1.7 nm. In the experiment the gate electrode consisted of dual doped n^+/p^+ polysilicon. Also the gate insulator is nitrided oxide for which the dielectric constant might not be exactly the same as that of SiO_2.[110] However, in our simulations we use the same device geometry (fin width, gate length and gate oxide thickness) but assume n^+ polysilicon gate and SiO_2 as the gate insulator. The effects of top gate on transport are assumed to be negligible considering much thicker gate oxide compared to side gate oxide. As mentioned earlier, the experimental FinFET device has been fabricated with the channel oriented in (110) wafer plane. In our simulations we also adopt the same wafer plane and assume that carrier propagation is along $[1\bar{1}0]$ crystallographic direction.

In order to obtain the closest match to the experimental results, a series of simulations with different combinations of doping profiles (source/drain doping concentration) and gate-source/drain underlap regions (which defines the doping gradient) have been performed. The doping profile which gives the closest fit of simulation results to the transfer characteristics of experimental FinFET at low drain bias can be described as-source/drain doping of 7×10^{18} cm^{-3} which follow a Gaussian envelope over a gate-source/drain underlap length of 12 nm to reach the body doping of 10^{15} cm^{-3}. The resulting doping gradient is around 3 nm/dec. We use uniform doping of 7×10^{18} cm^{-3} in the gate electrodes. Since the exact doping profile in the gate electrode is not specified, the simulated transfer characteristics can be shifted in voltage-scale (gate voltage) to match experimental data. However, it is important to mention that in selecting the above mentioned doping profiles as the appropriate one to match experimental data we consider simultaneously that (i) the value of subthreshold slope being in good correspondence to that obtained in experiment, (ii) over the gate voltage range of interest (−0.8 V to 0 V) the current values are reasonably close to the experimental data and, (iii) at very low gate voltage transfer characteristics do not show any bending which we do not observe in experimental data.

As one can see from Figure 27, the transfer characteristics obtained using the above mentioned doping profile gives current values close to the experimental ones in the subthreshold regime at a drain voltage of 0.1 V. In order to check that this result is not a coincidence, the transfer characteristics with the same geometry but with a high drain voltage of 1.2 V have been calculated, and found to be in good correspondence with the experimental data (see Fig. 31).

One can see from Figure 27, above threshold, with the increase in gate voltage, the deviation between simulated and experimental data increases rapidly. We predict that presence of very high parasitic series source/drain resistance, a critical issue in FinFET device, might be a reason for the smaller value of the drain current at high gate bias in the experiment. In order to examine the influence of the

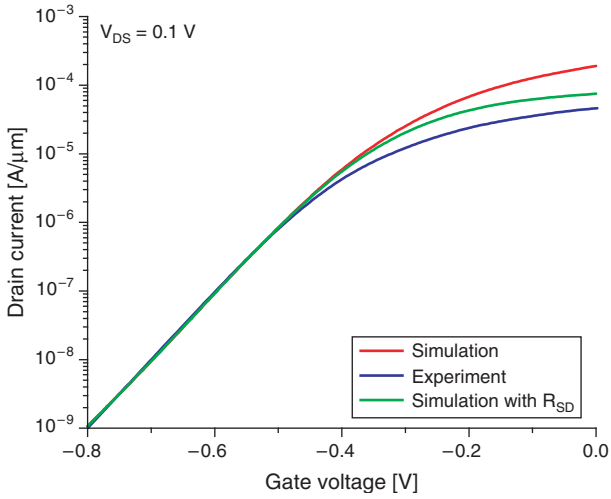

Fig. 27. Comparison of simulated transfer characteristics to the experimental data at low drain bias of 0.1 V.

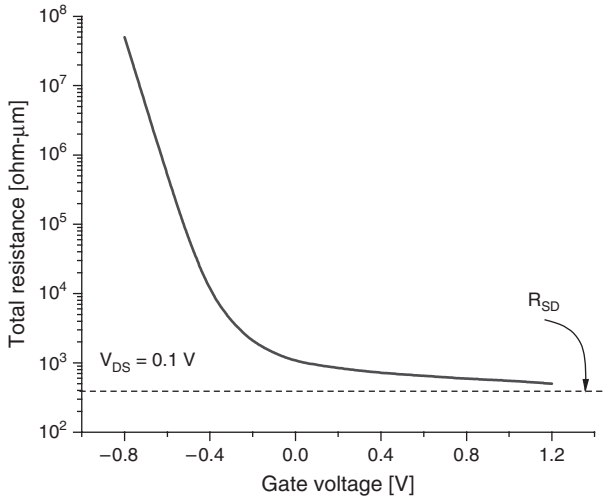

Fig. 28. Total resistance, R_{tot} as a function of gate voltage at low drain bias.

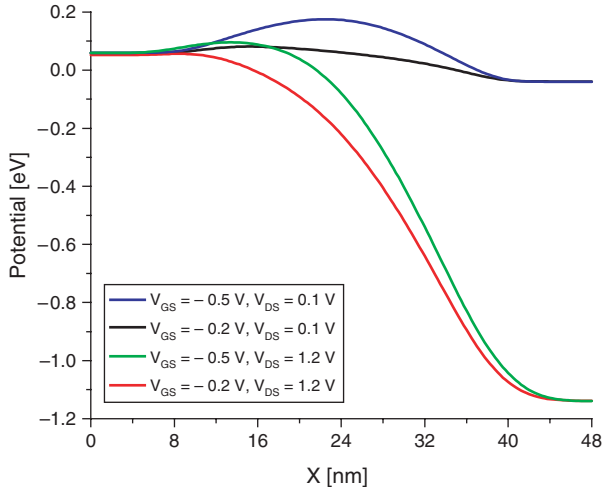

Fig. 29. 1D potential profiles along the length of the device in subthreshold and on-state with low and high drain biases.

series parasitic source/drain resistance, R_{SD} we extract the value of R_{SD} of experimental device from a plot of total resistance, R_{tot} (sum of device resistance, R_{int} and series parasitic source/drain resistance,R_{SD}) versus gate voltage as shown in Figure 28.

For sufficiently large value of gate voltage, R_{int} becomes very small and one can reasonably assume that $R_{tot} \approx R_{SD}$. The value of total parasitic series source/drain resistance extracted for the experimental device is found to be around 400 Ω-μm. Including the effects of R_{SD} the modified transfer characteristics is also shown in Figure 27 and one can see that the simulation result is very close to the experimental findings even at high gate bias. After including the effects of series resistance still we see some deviation of simulation results from the experiment in on-state. It is well known that in nanoscale devices, the presence of an unintentional dopant in the channel is highly probable.[111] Even if the fin is lightly doped, the unavoidable background doping might give rise to a one ionized dopant being present at a random location within the channel. Also, if an electron becomes trapped in a defect state at the interface or in the silicon body, it will introduce a fixed charge in the channel region. Depending on its position and applied bias, this unintentional dopant can significantly alter the device behavior, particularly when the channel is very lightly doped. An unintentional dopant sitting at a random location within the channel introduces a localized barrier which impedes the carrier propagation. The impact is significantly larger for an unintentional dopant sitting at the beginning of the fin near the source end compared to other probable positions.[112]

Figure 29 depicts the effective 1D potential profiles along X direction at the center of the fin in subthreshold regime and on-state at low and high drain bias. Also shown in Figure 30 is the corresponding 1D lateral electric field profiles along X direction. At low drain bias ($V_{DS} = 0.1$ V) and low gate voltage ($V_{GS} = -0.5$ V) the intrinsic barrier is already high enough (as shown in Fig. 29) so that the effects of the localized barrier introduced by the unintentional dopant can be assumed negligible. Therefore, over the subthreshold regime, we see a very good correspondence between simulation and experiment. For higher gate voltages ($V_{GS} = -0.2$ V), the intrinsic barrier is reduced significantly (Fig. 29).

Also the lateral electric field is reduced due to the increased effects of transverse electric field (Fig. 30). Thus, the localized barrier due to unintentional dopant is expected to influence the value of the drain current around device turn-on point. Therefore, at low drain bias ($V_{DS} =$ 0.1 V), the deviation between simulated drain current and experimental value increases with increasing gate voltage (above threshold) up to some cut-off beyond which the inversion electrons start to screen the potential of a single dopant ion. Consequently the influence of unintentional

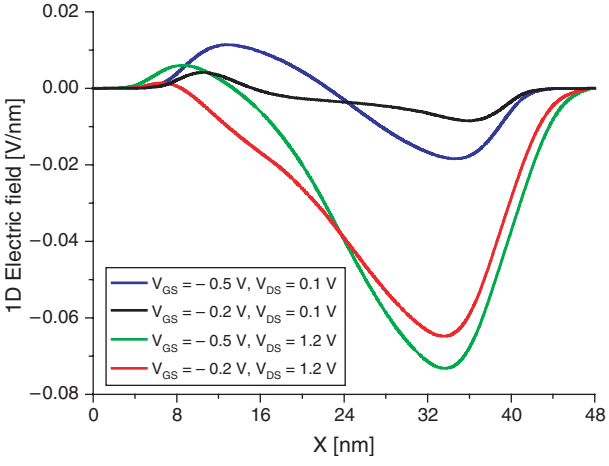

Fig. 30. 1D electric field along the length of the device for the potential profiles in Figure 3.5.

dopant on drain current gradually diminishes at much higher value of the gate voltage beyond threshold voltage which is also evident from Figure 27.

For higher value of drain voltage ($V_{DS} = 1.2$ V), we see some discrepancies between the simulation results and experimental data in *both* subthreshold and at high gate voltages as shown in Figure 31. Inclusion of series parasitic source/drain resistance reduces the drain current, but still the experimental values of drain current remains much smaller than the simulation results. In this case, due to significant DIBL effects, intrinsic barrier reduces, compared to the case with low drain bias, for both low and high gate voltages as shown in Figure 29. In subthreshold regime, the intrinsic barrier is much lower for $V_{DS} = 1.2$ V than for $V_{DS} = 0.1$ V. Consequently, the discrepancy between the experiment and simulation can be explained by more 'noticeable' (with respect to intrinsic barrier) effect of localized barrier due to *unintentional dopant*, which was much less significant for low drain bias. As the gate voltage increases, the effects of unintentional dopant become even more pronounced, which may explain the high voltage trend in Figure 31. We note that the position of the unintentional dopant is crucial in determining its effects on drain current. At high drain bias, unintentional dopant at the source side, will affect the drain current stronger than impurities at other locations.

Finally, we note that the subthreshold slope of 125 mV/dec have been reported for the *n*-FinFET in the experiment.[109] The corresponding value as obtained from our simulation is 120 mV/dec. The value of DIBL(at $I_D = 3 \times 10^{-6}$ A/μm) as extracted from the transfer characteristics of the experimental device is 145 mV/V and the corresponding value calculated from our simulation considering the effects of series parasitic source/drain resistance is 160 mV/V. These numbers clearly show that the experimentally fabricated[109] 10 nm FinFET device was very far from optimal. Consequently, the 10 nm device characteristics could be significantly improved by a proper tuning of device geometry. In the work[113] we have used our CBR simulator to optimize the device geometry and doping profile of a 10 nm FinFET device to meet most of the performance matrices defined by ITRS[114] for high performance 10 nm double-gate devices, which are expected to be commercially available around 2015.

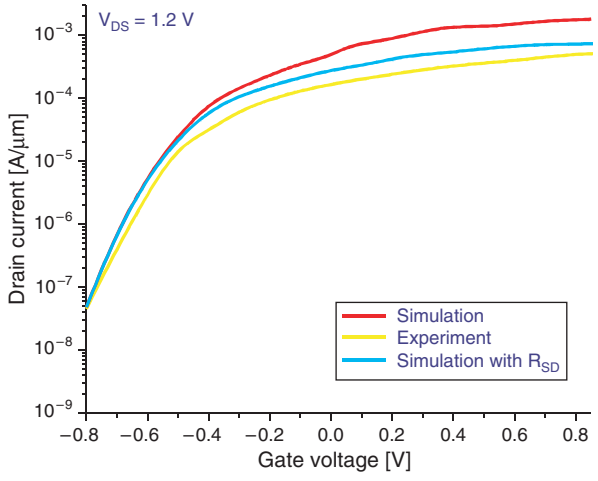

Fig. 31. Comparison of simulated transfer characteristics to the experimental data at $V_{DS} = 1.2$ V.

4. CONCLUSIONS

In this review article we have given a brief description of currently most important and most physically based semiclassical and quantum transport approaches. It is important to note that because of these developments, device simulation has achieved significantly higher maturity level than process simulation. In fact, particle-based device simulators can capture the essential physics up to ballistic transport regime and, when quantum interference effects start to dominate device behavior, quantum transport simulators based on either direct solution of the Schrödinger equation or its counterpart, the Green's functions, have been developed which, with the recent progress of state of the art computers, can simulate 3D nanoscale devices within a reasonable time-frame.

However, nanoelectronic device simulation of the future *must* ultimately include both, the sophisticated physics oriented electronic structure calculations and the engineering oriented transport simulations. Extensive scientific arguments have recently ensued regarding transport theory, basis representation, and practical implementation of a simulator capable of describing a realistic device. Starting from the field of molecular chemistry, Mujica, Kemp, Roitberg, Ratner[115] applied tight-binding based approaches to the modeling of transport in molecular wires. Later, Derosa and Seminario[116] modeled molecular charge transport using density functional theory and Green's functions. Further significant advances in the understanding of the electronic structure in technologically relevant devices were recently achieved through *ab initio* simulation of MOS devices by Demkov and Sankey.[117] Ballistic transport through a thin dielectric barrier was evaluated using standard Green function techniques[118, 119] without scattering mechanisms. However, quantum mechanical simulations of electron transport through 3D confined structures, such as quantum dots, have not yet reached the maturity (it is important, for example, for simulating operation of the next generation quantum dot photodetectors). Early efforts of understanding the operation of coupled quantum dot structures were rate equation based[120–122] where a simplified electronic structure was assumed.

Whereas traditional semiconductor device simulators are insufficiently equipped to describe quantum effects at atomic dimensions, most *ab-initio* methods from condensed matter physics are still computationally too demanding for application to practical devices, even as

REVIEW

small as quantum dots. A number of intermediary methods have therefore been developed in recent years. The methods can be divided into two major theory categories: atomistic and non-atomistic. Atomistic approaches attempt to work directly with the electronic wave function of each individual atom. *Ab-initio* methods overcome the shortcomings of the effective mass approximation; however, additional approximations must be introduced to reduce computational costs. One of the critical questions is the choice of a basis set for the representation of the electronic wave function. Many approaches have been considered, ranging from traditional numerical methods, such as finite difference and finite elements, as well as plane wave expansions,[123–125] to methods that exploit the natural properties of chemical bonding in condensed matter. Among these latter approaches, local orbital methods are particularly attractive. While the method of using atomic orbitals as a basis set has a long history in solid state physics, new basis sets with compact support have recently been developed,[126, 127] and, together with specific energy minimization schemes, these new basis sets result in computational costs which increase linearly with the number of atoms in the system without much accuracy degradation.[128, 129] However, even with such methods, only a few thousand atoms can be described with present day computational resources. NEMO3D uses an empirical tight-binding method[130, 131] that is conceptually related to the local orbital method and combines the advantages of an atomic level description with the intrinsic accuracy of empirical methods. It has already demonstrated considerable success[132, 133] in quantum mechanical modeling of electron transport as well as the electronic structure modeling of small quantum dots.[134] NEMO3D typically uses $sp^3s\Uparrow\leftarrow$ or $sp^3d^5s\Uparrow\leftarrow$ model that consists of five or ten spin degenerate basis states, respectively. Note that for the modeling of quantum dots, three main methods have been used in recent years: $k \leq p$,[135, 136] pseudopotentials,[123] and empirical tight-binding.[134]

As already discussed in Section 3, there are a number of methods developed by solid state theorists over the last several decades to address the issue of quantum transport in nano-devices. Among the most commonly used in nanostructure calculations schemes are the Wigner-function approach,[137] the Pauli master equation,[138] and the non-equilibrium Green's functions (NEGF).[81, 139, 140] The growing popularity of the latest (sometimes referred to as the Keldysh or the Kadanoff–Baym) formalism is conditioned by its sound conceptual basis for the development of the new class of quantum transport simulators.[141] Among its doubtless advantages are the clear physical conceptions, rigorous definitions, well-developed mathematical apparatus and flexibility of the algorithmization.

Thus, in our opinion, *the goal of any future simulation effort is to merge the electronic structure calculations with the quantum transport calculations and develop such a NEGF technique that is numerically efficient and ready for engineering applications in 3D objects on the one hand (such as QDIP), and rigorously quantum-mechanical on the other hand so that it properly incorporates the electronic structure of, for example, regular or disordered quantum dots used in QDIPs.*

The groups from ASU and Purdue are currently working on the development of such simulator in order to be able to calculate all the properties of 3D open quantum systems, particularly QDIPs. The transport kernel of the simulator is based on the Contact Block Reduction (CBR) method[81, 92, 93] and is discussed in more details in Section 3.4 of this review article. As already noted, the CBR method is applicable to fully self-consistent quantum transport calculations in arbitrarily shaped 3D structures using either the effective mass approximation or the mutli-band Hamiltonian description.[81] The band-structure of the QDIP's will be calculated using NEMO3D simulation software.

In summary, from the discussion above it follows that *the ultimate goal of semiconductor transport calculation of future nanoscale devices will be to **merge** the 3D quantum transport approaches with ab-initio band structure calculations.* This will ensure the most accurate simulation and better understanding carrier transport and operation of novel nano-device structures.

References

1. D. K. Ferry and S. M. Goodnick, Transport in Nanostructures, Cambridge Studies in Semiconductor Physics and Microelectronic Engineering (**1997**).
2. D. Vasileska and S. M. Goodnick, *Materials Science and Engineering, Reports: A Review: Journal* R38, 181 (**2002**).
3. S. M. Goodnick and D. Vasileska, Encyclopedia of Materials: Science and Technology, edited by K. H. J. Buschow, R. W. Cahn, M. C. Flemings, E. J. Kramer, and S. Mahajan, Elsevier, New York (**2001**), Vol. 2, p. 1456.
4. D. Vasileska and S. M. Goodnick, Computational Electronics, Morganand Claypool (**2006**).
5. A. Schütz, S. Selberherr, and H. Pötzl, *Solid-State Electron.* 25, 177 (**1982**).
6. P. Antognetti and G. Massobrio, Semiconductor Device Modeling with SPICE, McGraw-Hill, New York (**1988**).
7. M. Shur, Physics of Semiconductor Devices, Prentice Hall Series in Solid State Physical Electronics.
8. D. L. Scharfetter and D. L. Gummel, *IEEE Transaction on Electron Devices* ED-16, 64 (**1969**).
9. K. Bløtekjær, *IEEE Trans. Electron Dev.* 17, 38 (**1970**).
10. M. V. Fischetti and S. E. Laux, Monte Carlo simulation of submicron Si MOSFETs, Simulation of Semiconductor Devices and Processes, edited by G. Baccarani and M. Rudan, Technoprint, Bologna (**1988**), Vol. 3, p. 349.
11. L. V. Keldysh, *Sov. Phys.—JETP* 20, 1018 (**1965**).
12. A. L. Fetter and J. D. Walecka, Quantum Theory of Many-Particle Systems, McGraw-Hill (**1971**).
13. G. D. Mahan, Many-Particle Physics, Kluwer Academic/Plenum Publishers, New York (**2000**).
14. R. Lake, G. Klimeck, R. C. Bowen, and D. Jovanovic, *J. Appl. Phys.* 81, 7845 (**1997**).

15. G. Baccarani and M. Wordeman, *Solid State Electron.* 28, 407 **(1985)**.
16. S. Cordier, *Math. Mod. Meth. Appl. Sci.* 4, 625 **(1994)**.
17. K. Tomizawa, Numerical Simulation of Submicron Semiconductor Devices, The Artech House Materials Science Library.
18. H. K. Gummel, *IEEE Transactions on Electron Devices* 11, 455 **(1964)**.
19. T. M. Apostol, Calculus, Multi-Variable Calculus and Linear Algebra, Blaisdell, Waltham, MA **(1969)**, Vol. II, Chap. 1.
20. J. G. Ruch, *IEEE Trans. Electron Devices* 19, 652 **(1972)**.
21. S. Chou, D. Antoniadis, and H. Smith, *IEEE Electron Device Lett.* 6, 665 **(1985)**.
22. G. Shahidi, D. Antoniadis, and H. Smith, *IEEE Electron Device Lett.* 8, 94 **(1988)**.
23. R. Straton, *Phys. Rev* 126, 2002 **(1962)**.
24. T. Grasser, T.-W. Tang, H. Kosina, and S. Selberherr, *Proceedings of the IEEE* 91, 251 **(2003)**.
25. M. A. Stettler, M. A. Alam, and M. S. Lundstrom, *Proceedings of the NUPAD Conference* **(1992)**, p. 97.
26. C. Jacoboni and L. Reggiani, *Rev. Mod. Phys.* 55, 645 **(1983)**.
27. C. Jacoboni and P. Lugli, The Monte Carlo Method for Semiconductor Device Simulation, Springer-Verlag, Vienna **(1989)**.
28. K. Hess, Monte Carlo Device Simulation: Full Band and Beyond, Kluwer Academic Publishing, Boston **(1991)**.
29. M. H. Kalos and P. A. Whitlock, Monte Carlo Methods, Wiley, New York **(1986)**.
30. D. K. Ferry, Semiconductors, Macmillan, New York **(1991)**.
31. H. D. Rees, *J. Phys. Chem. Solids* 30, 643 **(1969)**.
32. R. M. Yorston, *J. Comp. Phys.* 64, 177 **(1986)**.
33. T. Gonzalez and D. Pardo, *Solid State Electron.* 39, 555 **(1996)**.
34. P. A. Blakey, S. S. Cherensky, and P. Sumer, Physics of Submicron Structures, Plenum Press, New York **(1984)**.
35. R. W. Hockney and J. W. Eastwood, Computer Simulation Using Particles, Institute of Physics Publishing, Bristol **(1988)**.
36. S. E. Laux, *IEEE Trans. Comp.-Aided Des. Int. Circ. Sys.* 15, 1266 **(1996)**.
37. J.-P. Colinge, SOI Technology: Materials to VLSI, Kluwer, Boston **(1997)**.
38. H.-K. Lim and J. Fossum, *IEEE Trans. Electron Devices* 30, 1251 **(1983)**.
39. J. G. Fossum, S. Krishnan, and P.-C. Yeh, *Proc. IEEE Int. SOI Conf.* **(1992)**, p. 132.
40. T. Numata and S. Takagi, *IEEE Trans. Electron Devices* 51, 2161 **(2004)**.
41. R. Chau, J. Kavalieros et al., *IEDM Techn. Dig.* 29, 1 **(2001)**.
42. B. Yu, L. Chang et al., *IEDM Tech. Dig.* 251 **(2002)**.
43. B. Doris, M. Ieong et al., *IEDM Tech. Dig.* 267 **(2002)**.
44. M. E. Kim, A. Das, and S. D. Senturia, *Phys. Rev. B* 18, 6890 **(1978)**.
45. M. V. Fischetti and S. E. Laux, *Phys. Rev. B* 38, 9721 **(1988)**.
46. P. Lugli and D. K. Ferry, *Phys. Rev. Lett.* 56, 1295 **(1986)**.
47. A. M. Kriman, M. J. Kann, D. K. Ferry, and R. Joshi, *Phys. Rev. Lett.* 65, 1619 **(1990)**.
48. R. P. Joshi and D. K. Ferry, *Phys. Rev. B* 43, 9734 **(1991)**.
49. M. V. Fischetti and S. E. Laux, *J. Appl. Phys.* 78, 1058 **(1995)**.
50. W. J. Gross, D. Vasileska, and D. K. Ferry, *IEEE Electron Device Lett.* 20, 463 **(1999)**.
51. W. J. Gross, D. Vasileska, and D. K. Ferry, *VLSI Design* 10, 437 **(2000)**.
52. D. Vasileska, W. J. Gross, and D. K. Ferry, *Superlattices Microstruct.* 27, 147 **(2000)**.
53. W. J. Gross, D. Vasileska, and D. K. Ferry, *IEEE Trans. Electron Dev.* 47, 1831 **(2000)**.
54. H. Brooks, *Phys. Rev.* 83, 879 **(1951)**.
55. C. Canali, G. Ottaviani, and A. Alberigi-Quaranta, *J. Phys. Chem. Solids* 32, 1707 **(1971)**.
56. M. V. Fischetti, Z. Ren, P. M. Solomon, M. Yang, and K. Rim, *J. Appl. Phys.* 94, 1079 **(2003)**.
57. G. Timp, J. Bude, K. K. Bourdelle, J. Garno, A. Ghetti, H. Grossman, M. Green, G. Forsyth, Y. Kim, R. Kleiman, F. Klemens, A. Kornblit, C. Lochstampfor, W. Mansfield, S. Moccio, T. Sorsch, D. M. Tennant, W. Timp, and R. Tung, *IEDM Tech. Dig.* 1999, 55 **(1999)**.
58. P. Palestri, D. Esseni, S. Eminente, C. Fiegna, E. Sangiorgi, and L. Selmi, *IEEE Trans El. Dev.* 52, 2727 **(2005)**.
59. M. M. Chowdhury, V. P. Trivedi, J. G. Fossum, and L. Mathew, *IEEE Trans. El. Dev.* 54, 1125 **(2007)**.
60. N. Sano, A. Hiroki, and K. Matsuzawa, *IEEE Trans. Nanotechnology* 1, 63 **(2002)**.
61. W. R. Frensley, *Rev. Mod. Phys.* 62, 745 **(1990)**.
62. Selberherr, Vasileska, and Goodnick, Private Communication.
63. R. Landauer, *Z. Physik B* 68, 217 **(1987)**.
64. M. Büttiker, *IBM J. Res. Dev.* 32, 63 **(1988)**.
65. E. O. Kane, Tunneling Phenomena in Solids, edited by E. Burstein and S. Lundqvist, Plenum, New York **(1969)**, p. 1.
66. J. N. Schulman and Y. C. Chang, *Phys. Rev. B* 27, 2346 **(1983)**.
67. C. Mailhiot and D. L. Smith, *Phys. Rev. B* 33, 8360 **(1986)**.
68. W. Frensley, *Rev. Mod. Phys.* 62, 745 **(1990)**.
69. C. Lent and D. Kirkner, *J. Appl. Phys.* 67, 6353 **(1990)**.
70. Z.-Y. Ting, E. T. Yu, and T. C. McGill, *Phys. Rev. B* 45, 3583 **(1992)**.
71. C. Strahberger and P. Vogl, *Phys. Rev. B* 62, 7289 **(2000)**.
72. Y. X. Liu, D. Z.-Y. Ting, and T. C. Mcgill, *Phys. Rev. B* 54, 5675 **(1996)**.
73. E. Polizzi, N. Ben Abdallah, O. Vanbésien, and D. Lippens, *J. Appl. Phys.* 87, 8700 **(2000)**.
74. E. Polizzi and N. Ben Abdallah, *Phys. Rev. B* 66, 245301-1 **(2002)**.
75. P. A. Ramachandran, Boundary Element Methods in Transport Phenomena, WIT Press **(1993)**.
76. H. R. Frohne, M. J. McLennan, and S. Datta, *J. Appl. Phys.* 66, 2699 **(1989)**.
77. P. A. Knipp and T. L. Reinecke, *Phys. Rev. B* 54, 1880 **(1996)**.
78. A. Svizhenko, M. P. Anantram, T. R. Govindan, B. Biegel, and R. Venugopal, *J. Appl. Phys.* 91, 2343 **(2002)**.
79. R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom, and D. Jovanovic, *J. Appl. Phys.* 92, 3730 **(2002)**.
80. C. Rivas and R. Lake, *Phys. Stat. Sol. (b)* 239, 94 **(2003)**.
81. D. Mamaluy, D. Vasileska, M. Sabathil, T. Zibold, and P. Vogl, *Phys. Rev. B* 71, 245321-1 **(2005)**.
82. S. Rotter, J. Z. Tang, L. Wirtz, J. Trost, and J. Burgdörfer, *Phys. Rev. B* 62, 1950 **(2000)**.
83. S. Rotter, B. Weingartner, N. Rohringer, and J. Burgdörfer, *Phys. Rev. B* 68, 165302-1 **(2003)**.
84. R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom, and D. Jovanovic, *J. Appl. Phys.* 92, 3730 **(2002)**.
85. A. Rahman, A. Ghosh, and M. Lundstrom, *Tech. Dig.-Int. Electron Devices Meet* 471 **(2003)**.
86. A. Rahman, M. S. Lundstrom, and A. W. Ghosh, *J. Appl. Phys.* 97, 053702 **(2005)**.
87. J. Wang, E. Polizzi, and M. Lundstrom, *J. Appl. Phys.* 96, 2192 **(2004)**.
88. H. Takeda and N. Mori, *J. Comp. El.* 4, 31 **(2005)**.
89. S. Jin, Y. J. Park, and H. S. Min, *J. Appl. Phys.* 99, 123719-1 **(2006)**.
90. S. E. Laux, A. Kumar, and M. V. Fischetti, *J. Appl. Phys.* 95, 5545 **(2004)**.
91. S. E. Laux, *J. Comp. El.* 3, 379 **(2004)**.
92. D. Mamaluy, M. Sabathil, and P. Vogl, *J. Appl. Phys.* 93, 4628 **(2003)**.
93. H. R. Khan, D. Mamaluy, and D. Vasileska, *IEEE Trans El. Dev.* 54, 784 **(2007)**.
94. S. Datta, Ftom Atom to Transistor, Cambridge University Press, Cambridge **(1995)**.
95. A. Trellakis, A. T. Galick, A. Pacelli, and U. Ravaioli, *J. Appl. Phys.* 81, 7880 **(1997)**.

REVIEW

96. R. Lake, G. Klimeck, R. C. Bowen, D. Jovanovic, D. Blanks, and M. Swaminathan, *Phys. Stat. Sol. (b)* 204, 354 **(1997)**.
97. L. Chand and C. Hu, *Superlattices Microstruct.* 28, 351 **(2000)**.
98. D. J. Frank, S. E. Laux, and M. V. Fischetti, *IEDM Tech. Dig.* 553 **(1992)**.
99. H. Takato, K. Sunouchi, N. Okabe, A. Nitayama, K. Hieda, F. Horiguchi, and F. Masuoka, *IEEE Trans El. Dev.* 38, 573 **(1991)**.
100. C. Auth and J. Plummer, *IEEE Elec. Dev. Lett.* 18 **(1997)**.
101. H.-S. P. Wong, K. K. Chan, and Y. Taur, *IEDM Tech. Dig.* 427 **(1997)**.
102. J. Colinge, M. Gao, A. Romano-Rodriguez, H. Maes, and C. Claeys, *IEDM Tech. Dig.* 595 **(1990)**.
103. H.-S. Philip Wong, *Solid-State Electron.* 49, 755 **(2005)**.
104. D. Hisamoto, T. Kaga, Y. Kawamoto, and E. Takeda, *IEDM Tech. Dig.* 833 **(1989)**.
105. D. Hisamoto, W.-C. Lee, J. Kadzierski, H. Takeuchi, K. Asano, C. Kuo, T.-J. King, J. Bokor, and C. Hu, *IEEE Trans El. Dev.* 47, 2320 **(2000)**.
106. Y. Choi, T. King, and C. Hu, *IEEE Trans El. Dev.* 23, 25 **(2002)**.
107. Y.-K. Choi, N. Lindert, P. Xuan, S. Tang, D. Ha, E. Anderson, T.-J. King, J. Bokor, and C. Hu, *IEEE International Electron Device Meeting Technical Digest* 421 **(2001)**.
108. X. Huang, W.-C. Lee, C. Kuo, D. Hisamoto, L. Chang, J. Kadzierski, E. Anderson, H. Takeuchi, Y.-K. Choi, K. Asano, V. Subramanian, T.-J. King, J. Bokor, and C. Hu, *IEEE Trans El. Dev.* 48, 880 **(2001)**.
109. B. Yu, L. Chang, S. Ahmed, H. Wang, S. Bell, C.-Y. Yang, C. Tabery, C. Ho, Q. Xiang, T.-J. King, J. Bokor, C. Hu, M.-R. Lin, and D. Kyser, FinFET Scaling to 10 nm Gate Length, IEDM Tech. Digest IEEE, Piscataway, NJ **(2002)**, pp. 251–254.
110. Y. Y. Chen, C. H. Chien, and J. C. Lou, *Thin Solid Films* 513, 264 **(2006)**.
111. T. Mizuno, J. Okamura, and A. Toriumi, *IEEE Trans El. Dev.* 41, 2216 **(1994)**.
112. H. R. Khan, D. Vasileska, and S. S. Ahmed, *J. Comp. El.* 3, 337 **(2004)**.
113. H. Khan, D. Mamaluy, and D. Vasileska, *IEEE Trans El. Dev.* **(2007)**.
114. International Technology Roadmap for Semiconductors (ITRS), 2006 updated edition, http://public.itrs.net.
115. Mujica, Kemp, Roitberg, and Ratner, *J. Chem. Physics* 104, 7296 **(1996)**.
116. D. Seminario, *J. Phys. Chemistry B* 105, 471 **(2001)**.
117. A. Demkov and O. Sankey, *Phys. Rev. Lett.* 83, 2038 **(1999)**.
118. A. Demkov, L. Zhang, and G. Loechelt, *J. Vac. Sci. Techn. B* 18, 2388 **(2000)**.
119. A. Demkov, Zhang, and Drabold, *Phys. Rev. B* 6412, 5306 **(2001)**.
120. G. Klimeck, R. Lake, S. Datta, and Bryant, *Phys. Rev. B* 50, 5484 **(1994)**.
121. G. Klimeck, Chen, and S. Datta, *Phys. Rev. B* 50, 2316 **(1994)**.
122. Chen et al., *Phys. Rev. B* 50, 8035 **(1994)**.
123. A. Canning, L. W. Wang, A. Williamson, and A. Zunger, *J. Comp. Physics* 160, 29 **(2000)**.
124. L. W. Wang, J. N. Kim, and A. Zunger, *Phys. Rev. B* 59, 5678 **(1999)**.
125. A. J. Williamson, L. W. Wang, and A. Zunger, *Phys. Rev. B* 62, 12963 **(2000)**.
126. R. Martin, *Phys. Rev. B* 1, 4005 **(1970)**.
127. O. Sankey and D. J. Niklewski, *Phys. Rev. B* 40, 3979 **(1989)**.
128. P. Ordej'on, D. A. Drabold, M. P. Grumbach, and R. M. Martin, *Phys. Rev. B* 48, 14646 **(1993)**.
129. F. Ordej'on, G. Galli, and R. Car, *Phys. Rev. B* 47, 9973 **(1993)**.
130. P. Vogl, H. P. Hjalmarson, and J. D. Dow, *J. Phys. Chem. Solids* 44, 365 **(1983)**.
131. J. M. Jancu, R. Scholz, F. Beltram, and F. Bassani, *Phys. Rev. B* 57, 6493 **(1998)**.
132. R. C. Bowen, IEDM 1997, IEEE, New York **(1997)**, p. 869.
133. G. Klimeck et al., *VLSI Design* 8, 79 **(1997)**.
134. Lee, Joensson, and G. Klimeck, *Phys. Rev. B* 63, 195318 **(2001)**.
135. Pryor, *Phys. Rev. B* 57, 7190 **(1998)**.
136. Stier, Grundmann, and Bimberg, *Phys. Rev. B* 59, 5688 **(1999)**.
137. P. Brodone, M. Pascoli, R. Brunetti, A. Bertoni, and C. Jacoboni, *Phys. Rev. B* 59, 3060 **(1998)**.
138. M. V. Fischetti, *Phys. Rev. B* 59, 4901 **(1998)**.
139. A. Haque and A. N. Khondker, *J. Appl. Phys.* 87, 2553 **(2000)**.
140. D. Guan, U. Ravaioli, R. W. Giannetta, M. Hannan, and I. Adesida, M. R. Melloch, *Phys. Rev. B* 67, 205328 **(2003)**.
141. S. Datta, *Superlattices Microstruct.* 28, 253 **(2000)**.

Received: 30 July 2007. Accepted: 29 August 2007.

REVIEW