Monte Carlo Device Simulations

Dragica Vasileska[1], Katerina Raleva[2], Stephen M. Goodnick[1], Christian Ringhofer[1], Shaikh S. Ahmed[3], Nabil Ashraf[1], Arif Hossain[1], Raghuraj Hathwar[1] Ashwin Ashok[4] and Balaji Padmanabhan[1]

[1] *Arizona State University, Tempe AZ*
[2] *University Sts Cyril and Methodius, Skopje, Republic of Macedonia*
[3] *Southern Illinois University, Carbondale, IL*
[4] *Intel Corp., Hillsboro, AZ*

Table of Contents

Abstract

As semiconductor devices are scaled into nanoscale regime, first velocity saturation starts to limit the carrier mobility due to pronounced intervalley scattering, and when the device dimensions are scaled to 100 nm and below, velocity overshoot (which is a positive effect) starts to dominate the device behavior leading to larger ON-state currents. Alongside with the developments in the semiconductor nanotechnology, in recent years there has been significant progress in physical based modeling of semiconductor devices. First, for devices for which gradual channel approximation can not be used due to the two-dimensional nature of the electrostatic potential and the electric fields driving the carriers from source to drain, drift-diffusion models have been exploited. These models are valid, in general, for large devices in which the fields are not that high so that there is no degradation of the mobility due to the electric field. The validity of the drift-diffusion models can be extended to take into account the velocity saturation effect with the introduction of field-dependent mobility and diffusion coefficients. When velocity overshoot becomes important, drift diffusion model is no longer valid and hydrodynamic model must be used. The hydrodynamic model has been the workhorse for technology development and several high-end commercial device simulators have appeared including Silvaco, Synopsys, Crosslight, etc. The advantages of the hydrodynamic model are that it allows quick simulation runs but the problem is that the amount of the velocity overshoot depends upon the choice of the energy relaxation time. The smaller is the device, the larger is the deviation when using the same set of energy relaxation times. A standard way in calculating the energy relaxation times is to use bulk Monte Carlo simulations. However, the energy relaxation times are material, device geometry and doping dependent parameters, so their determination ahead of time is not possible. To avoid the problem of the proper choice of the energy relaxation times, a direct solution of the Boltzmann Transport Equation (BTE) using the Monte Carlo method is the best method of choice. That is why the focus of this review paper is on explaining basic Monte Carlo device simulator and then the focus will be shifted on the inclusion of various higher order effects that explain particular physical phenomena or processes.

The Monte Carlo book chapter is organized as follows. First, the idea behind the Monte Carlo technique is outlined by revoking the path integral method for the solution of the BTE. This approach naturally leads to the free-flight-scatter sequence that is used in solving the BTE using the Monte Carlo method. Various scattering mechanisms relevant for different materials are given to completely specify the collision integral in the BTE. A discussion followed with the presentation of a generic flow-chart for implementing bulk Monte Carlo code is presented. Note that bulk Monte Carlo approach is suitable for the characterization of materials, but in order to study behavior of semiconductor devices coupling of the Monte Carlo transport kernel with a Poisson equation solver which gives the self-consistent field that moves the carriers around is needed. Important ingredients in describing particle-based device simulators are the particle-mesh coupling, treatment of the Ohmic contacts and calculation of the current. Again, a generic flowchart of a particle-based device simulator will be provided.

Having described the basic principles of the particle-based device simulators, we will then describe how one implements higher order effects within a particle-based device simulation scheme to study specific phenomena such as degeneracies, short-range Coulomb interactions, and quantum-mechanical size quantization effects, to mention just a few. We want to point out that the ASU team has been a leader in the world in the inclusion of short-range Coulomb interactions within a particle-based device simulation scheme and has thoroughly investigated the role of random dopants in conventional MOSFETs, the role played by unintentional dopants in alternative device technologies and the influence of the charged random trap on the threshold voltage and on-current fluctuations. We have also been leaders in including quantum corrections within a particle-based device simulation scheme. Numerous simulation results that address all of these advanced issues are provided at the end of the book chapter. Recently, we have emerged as leaders in the incorporation of the self-heating effects within a particle-based device simulation scheme and have applied our approach to investigate the role of self-heating effects in the operation of fully-depleted (FD) silicon on insulator (SOI) devices, nanowire transistors and GaN HEMTs.

1. Importance of MC Particle-Based Device Simulations

1.1 Industry Trends and the Need for Modeling and Simulation

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* [2,3,4] 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 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.

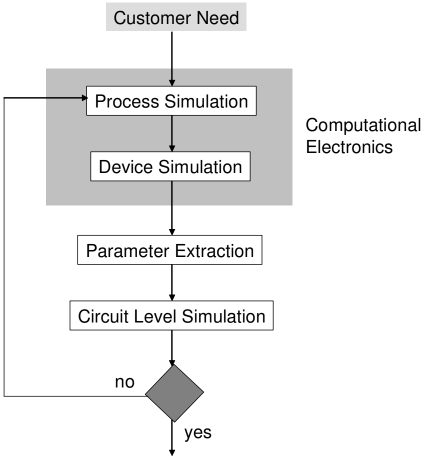


Figure 1. Design sequence to achieve desired customer need.

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 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.

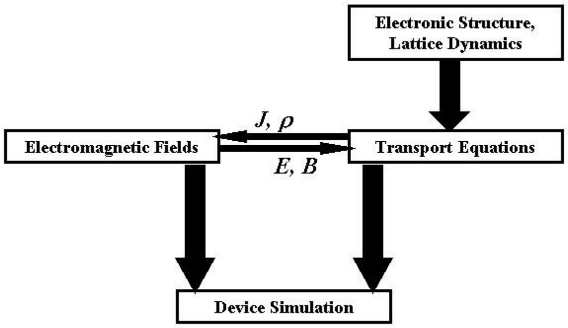
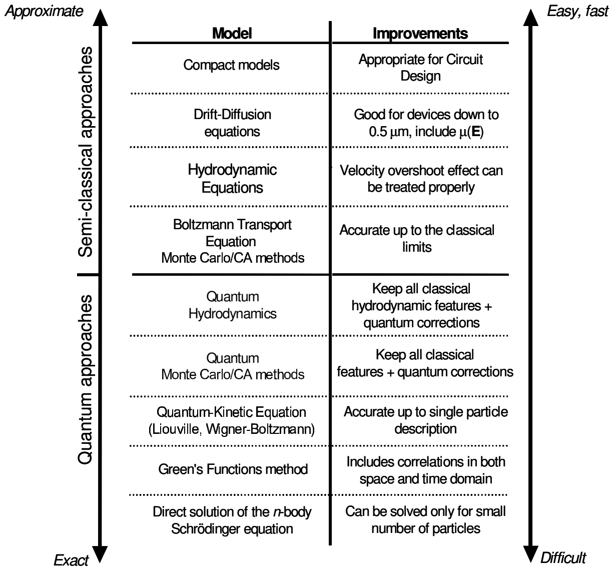


Figure 2. Schematic description of the device simulation sequence.


Approximate
Exact
Semi-classical approaches
Quantum approaches
Easy, fast
Difficult

Model	Improvements
Compact models	Appropriate for Circuit Design
Drift-Diffusion equations	Good for devices down to 0.5 μm, include μ(**E**)
Hydrodynamic Equations	Velocity overshoot effect can be treated properly
Boltzmann Transport Equation Monte Carlo/CA methods	Accurate up to the classical limits
Quantum Hydrodynamics	Keep all classical hydrodynamic features + quantum corrections
Quantum Monte Carlo/CA methods	Keep all classical features + quantum corrections
Quantum-Kinetic Equation (Liouville, Wigner-Boltzmann)	Accurate up to single particle description
Green's Functions method	Includes correlations in both space and time domain
Direct solution of the *n*-body Schrödinger equation	Can be solved only for small number of particles

Figure 3. Illustration of the hierarchy of transport models**.**

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 [11,12,13]. 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.

1.2 Drift-Diffusion and Hydrodynamic Models

In Section 1.1 above, 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:

Current equations:

$$\begin{aligned} J_n &= qn(x)\mu_n E(x) + qD_n \frac{dn}{dx} \\ J_p &= qp(x)\mu_p E(x) - qD_p \frac{dn}{dx} \end{aligned} \tag{1}$$

Continuity equations:

$$\begin{aligned} \frac{\partial n}{\partial t} &= \frac{1}{q}\nabla\cdot\mathbf{J}_n + U_n \\ \frac{\partial p}{\partial t} &= -\frac{1}{q}\nabla\cdot\mathbf{J}_p + U_p \end{aligned} \tag{2}$$

Poisson's equation:

$$\nabla\cdot(\varepsilon\nabla V) = -\left(p - n + N_D^+ - N_A^-\right), \tag{3}$$

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

- Conserve the total number of particles inside the device being simulated.
- Respect local positive definite nature of carrier density. Negative density is unphysical.
- 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]:

- The mesh size Δx is limited by the Debye length.
- 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

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

where N is the doping density, ε is the dielectric constant, k_B is the Boltzmann constant, T is the lattice temperature and q is the elementary charge. 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

$$t_{dr} = \frac{\varepsilon}{qN\mu} \tag{5}$$

where μ is the carrier mobility.

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.

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.

1.2.1 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 the electron current density equation as

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

where D_T is the thermal diffusivity and D_n is the diffusion constant.

1.2.2 Stratton's Approach

One of the first derivations of extended transport equations was performed by Stratton [20]. First the distribution function is split into the even and odd parts

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

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

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

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

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

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

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

Equation for the current density can be rewritten as:

$$J = q\mu\left(nE + \frac{k_B}{q}T_n\nabla n + \frac{k_B}{q}n\left(1+\nu_n\right)\nabla T_n\right) , \tag{12a}$$

with

$$\nu_n = \frac{T_n}{\mu}\frac{\partial\mu}{\partial T_n} = \frac{\partial\ln\mu}{\partial\ln T_n} \tag{12b}$$

which is commonly used as a fit parameter with values in the range [-0.5,-1.0]. For ν_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.

1.2.3 Balance Equations Model

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

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

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

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

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

$$
\mathbf{F}_W = \mathbf{v}W + nk_B \vec{\vec{T}} \cdot \mathbf{v} + \mathbf{Q} . \tag{16}
$$

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

$$
\begin{aligned}
\frac{\partial J_z}{\partial t} &= \frac{2e}{m^*}\sum_i \frac{\partial}{\partial x_i}\left(K_{iz} + \frac{1}{2}nk_B T_{iz} \right) + \frac{ne^2}{m^*}E_z - \left\langle\left\langle \frac{1}{\tau_m} \right\rangle\right\rangle J_z \\
\frac{\partial W}{\partial t} &= -\nabla \cdot \left(\mathbf{v}W + \mathbf{Q} + nk_B \vec{\vec{T}} \cdot \mathbf{v} \right) + \mathbf{E}\cdot\mathbf{J}_n - \left\langle\left\langle \frac{1}{\tau_E} \right\rangle\right\rangle (W - W_0)
\end{aligned}
\tag{17}
$$

For displaced-Maxwellian approximation for the distribution function, the heat flux **Q** = 0. However, Blotekjaer [21] 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 **Q** actually leads to a closed set of equations in which the energy balance equation is of the form

$$
\begin{aligned}
\frac{\partial W}{\partial t} &= -\nabla \cdot \left(\mathbf{v}W - \kappa \nabla T_c + nk_B T_c \mathbf{v} \right) + \mathbf{E}\cdot\mathbf{J}_n \\
&\quad - \left\langle\left\langle \frac{1}{\tau_E} \right\rangle\right\rangle (W - W_0)
\end{aligned}
\tag{19}
$$

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, Alam and Lundstrom [22] have suggested a new form of closure

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

where **J** is the current density and *r* is a tunable parameter less than unity. Now using

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

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

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

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

$$\begin{aligned} \frac{\partial n}{\partial t} &= \frac{1}{e}\nabla\cdot J_n + S_n \\ \frac{\partial J_z}{\partial t} &= \frac{e}{m^*}\frac{\partial}{\partial x_z}\left(nm^* v_{dz}^2 + nk_B T_c\right) \\ &\quad + \frac{ne^2}{m^*}E_z - \left\langle\left\langle\frac{1}{\tau_m}\right\rangle\right\rangle J_z \\ \frac{\partial W}{\partial t} &= -\frac{\partial}{\partial x_z}\left[(W + nk_B T_c)v_{dz} - \kappa\frac{\partial T_c}{\partial x_z}\right] \\ &\quad + J_z E_z - \left\langle\left\langle\frac{1}{\tau_E}\right\rangle\right\rangle (W - W_0) \end{aligned} \tag{23}$$

where

$$\begin{aligned} J_z &= -env_{dz} = -\frac{e}{m^*}P_z \\ W &= \frac{1}{2}nm^* v_{dz}^2 + \frac{3}{2}nk_B T_c \end{aligned} \tag{24}$$

1.3 Failure of the Drift-Diffusion and Hydrodynamic Models

To understand the advantages and the limitations of the drift-diffusion and of the hydrodynamic model, let us consider the following examples: Fully Depleted (FD) SOI devices with channel lengths 25, 45 and 90 nm. The oxide thickness and the doping of the channel of the three devices considered are summarized in Table 1. In Figures 4 and 5 (a-c) we compare the output characteristics of the three devices when using the drift-diffusion and the hydrodynamic model.

Table 1. Geometrical dimensions and applied biases of the fully-depleted SOI nMOSFETs simulated here.

Feature (channel length)	14 nm	25 nm	90 nm
Tox	1 nm	1.2 nm	1.5 nm
V_{DS}	1V	1.2 V	1.4 V
Overshoot EB/HD without series resistance	233% / 224%	139% / 126%	31% /21%
Overshoot EB/DD with series resistance	153%/96%	108%/67%	39%/26%

Source/drain doping = 10^{20} cm^{-3} and 10^{19} cm^{-3} (series resistance (SR) case)

Channel doping = 10^{18} cm^{-3}

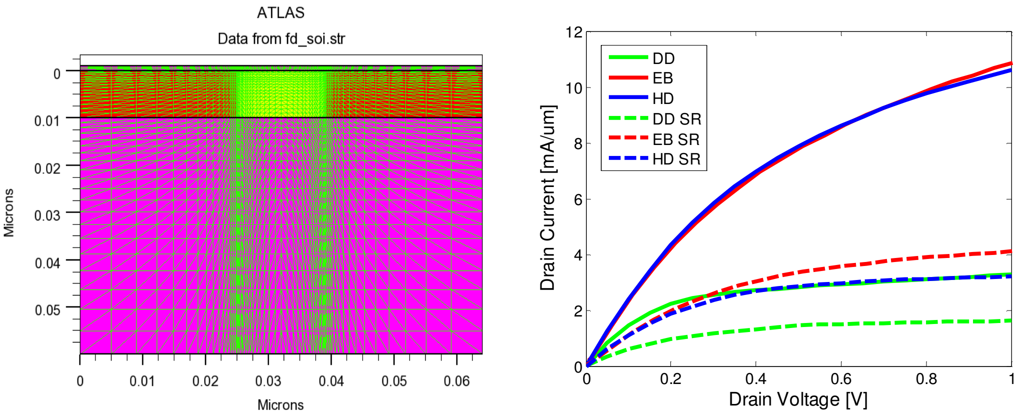
Overshoot= $(ID_{HD}-ID_{DD})/ID_{DD}$ (%) ; ID is the on-state current

Here we use the commercial Silvaco Atlas (PISCEC) simulation package [23] that includes hydrodynamic modeling with momentum and energy relaxation times of 0.2 ps, Auger generation/recombination (important for the proper modeling of the heavily doped source and drain contacts), and the Schockley-Read-Hall (SRH) generation-recombination mechanism are included here for completeness, although the latter is not really important for this device structure. Impact ionization is not included in these simulations. In the hydrodynamic calculation, it is important that one uses the NEWTON method for solving the coupled set of equations, otherwise the simulation will not converge due to the strong coupling of the equations at high drain biases. We consider both the simplified energy balance (EB) model and the complete hydrodynamic model (HD). We present simulation results for the following two cases:

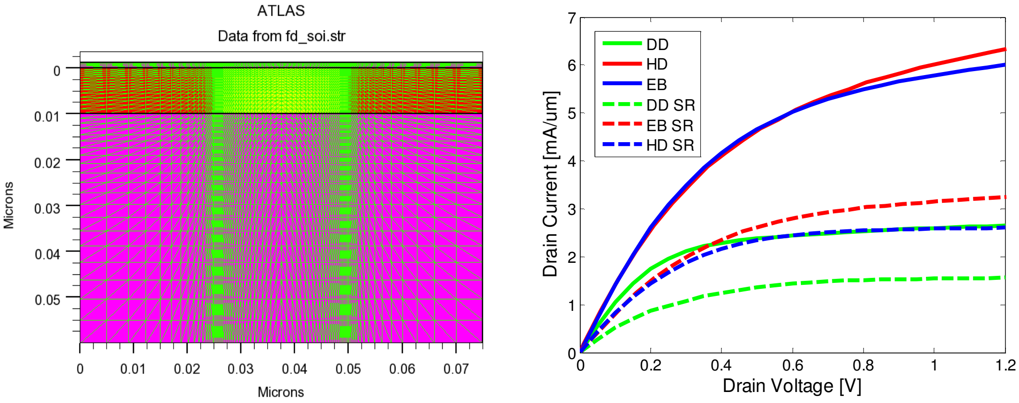
1. Source and drain doping of 10^{20} and 10^{19} cm^{-3} to examine series resistance effects. This is very important to know as in prototypical Monte Carlo device simulations source and drain regions are usually doped up to 10^{19} cm^{-3} to reduce the computational cost (total number of particles simulated). In these simulations we assume that the energy relaxation time is 0.2 ps, which is a typical value used for the silicon material system. The results from these simulations are presented in Figure 4 for the 14 nm, 25 nm and 90 nm channel length devices. On the left panel, we show the meshing used in these simulations and on the right panel we show the output characteristics for the appropriate on-state gate bias and drift-diffusion and hydrodynamic transport models.
2. In this second case we perform only hydrodynamic simulations to investigate the sensitivity of the hydrodynamic model to variations in the energy relaxation time which, in principle, is a material and device geometry dependent parameter which makes it almost impossible to determine analytically. This variation for the three technology nodes of devices is shown in Figure 5.

The results first show that the source/drain doping plays an important role in terms of the drive current, which is primarily and effect of series resistance. From the results presented it is evident that non-stationary transport plays smaller role in 90 nm gate-length FD SOI devices, whereas the importance of non-stationary transport and the velocity overshoot associated with it increases drastically for 14 nm gate length FD SOI device. These results suggest that one must include energy balance equation if proper modeling of nano-scale devices with gate lengths less than 100 nm is to be achieved.

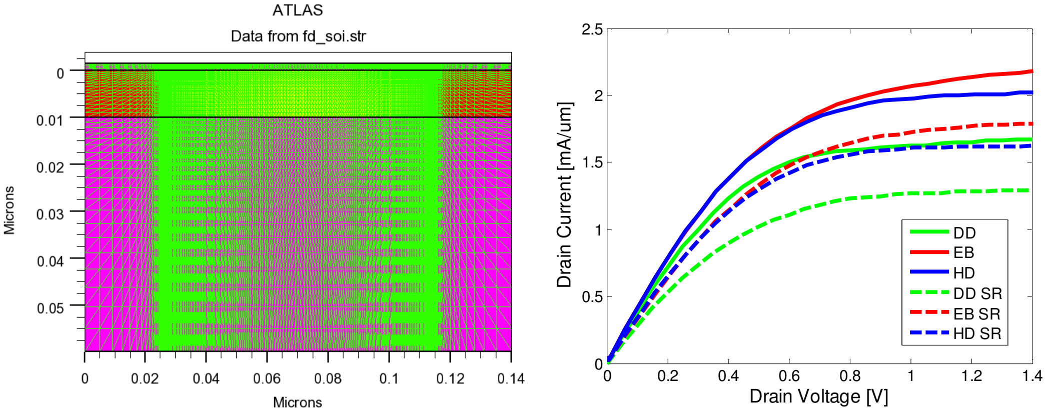
Yet another issue that deserves further attention is the dependence of the simulation results upon the choice of the energy relaxation time. In Figure 5 we plot the output characteristics of 14, 25 and 90 nm gate length FD SOI devices in which parameter is the energy relaxation time. We see strong dependence of the on-current upon the choice of the energy relaxation time for the smallest structure being simulated which suggests that proper determination of the energy relaxation time is needed. The energy relaxation time, in turn, is bias and geometry dependent parameter and its exact determination is impossible. The inability to properly determine the energy relaxation time in hydrodynamic/energy balance models has been the main motivation for the development of particle-based simulators discussed in Chapters 2 and 3.



(a) channel length = 14 nm. V_G=1 V. SR stands for series resistance.

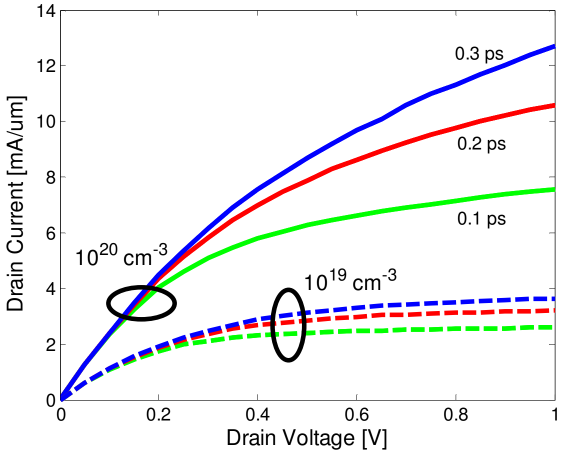


(b) channel length = 25 nm. V_G=1.2 V. SR stands for series resistance.

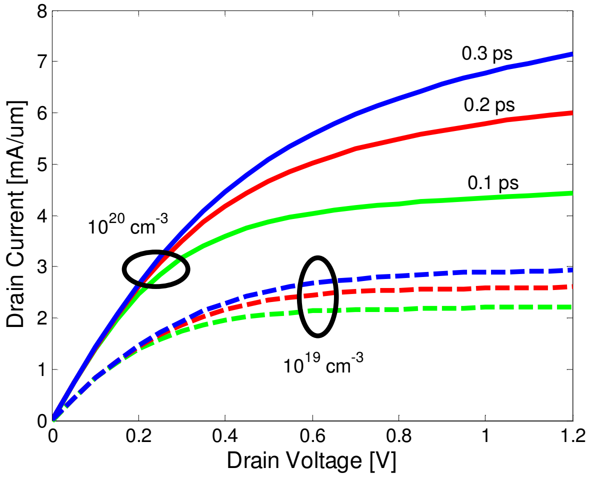


(c) channel length = 90 nm. V_G=1.4 V. SR stands for series resistance.

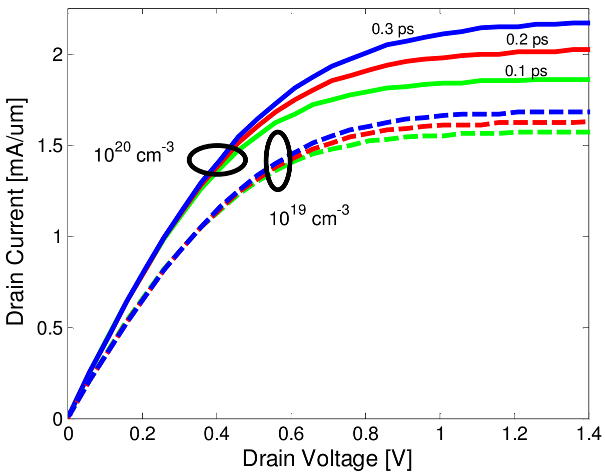
Figure 4. Mesh and output characteristics of 14, 25 and 90 nm channel length FD SOI devices in the on-state when using drift-diffusion, energy balance, and hydrodynamic models.



(a) Simulated characteristics for different energy relaxation times for two different source/drain dopings for a channel length of 14 nm, V_G=1 V.



(b) Simulated characteristics for different energy relaxation times for two different source/drain dopings, for a channel length of 25 nm, V_G=1.2 V.



(c) Simulated characteristics for different energy relaxation times for two different source/drain dopings, for a channel length of 90 nm, V_G=1.4 V.

Figure 5. Dependence of the on-state current upon the choice of the energy relaxation time for three different channel length FD SOI devices.

2. Bulk Monte Carlo Method

In the previous section 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 to the BTE, it was shown that in the case of small devices (see Section 1.3 above), 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 [24,25,26]. 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 [27].

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 semiclassical 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 [28]. 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. In the following, we first describe the derivation of the free-flight scatter sequence using the path-integral method (section 2.1) and then we describe the standard Monte Carlo algorithm used to simulate charge transport in semiconductors (section 2.2). 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 (section 3)

2.1 Monte Carlo and Path-Integral Methods

The path-integral method for solving the BTE is a rather a useful and an intuitive procedure for describing the Monte Carlo method. In its general form the BTE is:

$$\begin{aligned}&\frac{\partial f}{\partial t}+\vec{v}\cdot\nabla_r f+(-e)\vec{\varepsilon}\cdot\nabla_p f=\frac{\partial f}{\partial t}\Big|_{coll}=\\&=\sum_{i=1}^{N}\sum_{\vec{p}'}[S_i(\vec{p}',\vec{p})f(\vec{r},\vec{p}',t)-S_i(\vec{p},\vec{p}')f(\vec{r},\vec{p},t)]\end{aligned} \quad . \qquad (25)$$

The first term on the right-hand side (RHS) of Eq. (25) gives the scattering into state $\vec{p}$, while the second term on the RHS of Eq. (25) is the scattering out of state $\vec{p}$. This form of the BTE is valid for non-degenerate semiconductors. The collision integral on the RHS can also be expressed as:

$$RHS=\sum_{\vec{p}'}\sum_{i=1}^{N}[S_i(\vec{p}',\vec{p})f(\vec{r},\vec{p}',t)]-f(\vec{r},\vec{p},t)\sum_{\vec{p}'}\sum_{i=1}^{N}S_i(\vec{p},\vec{p}')\,, \qquad (26)$$

where, $\frac{1}{\tau(\vec{p})}=\sum S_i(\vec{p},\vec{p}')$ is the total scattering rate out of state $\vec{p}$.

Hence:

$$\begin{aligned}RHS&=\sum_{\vec{p}'}\sum_{i=1}^{N}[S_i(\vec{p}',\vec{p})f(\vec{r},\vec{p}',t)]-f(\vec{r},\vec{p},t)\left[\frac{1}{\tau(\vec{p})}+\Omega(\vec{p})\right]+f(\vec{r},\vec{p},t)\Omega(\vec{p})=\\&=\sum_{\vec{p}'}\sum_{i=1}^{N}[S_i(\vec{p}',\vec{p})f(\vec{r},\vec{p}',t)]-\Gamma(\vec{p})f(\vec{r},\vec{p},t)+\sum_{\vec{p}'}f(\vec{r},\vec{p}',t)\Omega(\vec{p})\delta_{\vec{p},\vec{p}'}\end{aligned} \quad . \qquad (27)$$

In this last expression we have added a term $\Omega(\vec{p})$ such that the total scattering rate out of a state p is constant. To understand the meaning of this term we need to go backwards, i.e. write $\Gamma(\vec{p})$ as:

$$\begin{aligned}\Gamma(\vec{p})&=\frac{1}{\tau(\vec{p})}+\Omega(\vec{p})=\sum_{\vec{p}'}\sum_{i=1}^{N}S_i(\vec{p},\vec{p}')+\sum_{\vec{p}'}\Omega(\vec{p})\delta_{\vec{p},\vec{p}'}=\\&=\sum_{\vec{p}'}\left[\sum_{i=1}^{N}S_i(\vec{p},\vec{p}')+\Omega(\vec{p})\delta_{\vec{p},\vec{p}'}\right]\end{aligned} \quad . \qquad (28)$$

We now define an effective transition rate:

$$S_{eff}(\vec{p},\vec{p}')=\sum_{i=1}^{N}S_i(\vec{p},\vec{p}')+\Omega(\vec{p})\delta_{\vec{p},\vec{p}'}\,, \qquad (29)$$

which consists of the sum of the N physical transition rates plus a term that has a momentum conserving δ-function. This second term has no effect on the carrier momentum and energy and it is a fictitious scattering process which is called self-scattering. The self-scattering can be calculated from:

$$\Omega(\vec{p}) = \Gamma(\vec{p}) - \frac{1}{\tau(\vec{p})} . \tag{30}$$

With the above definition for the self-scattering term, the BTE becomes:

$$\frac{\partial f}{\partial t} + \vec{v}\nabla_r f + (-e)\vec{\varepsilon}\nabla_p f + \Gamma(\vec{p}) f(\vec{r},\vec{p},t) = \sum_{\vec{p}'}\left[\sum_{i=1}^{N} S_i(\vec{p}',\vec{p}) + \left(\Gamma(\vec{p}) - \frac{1}{\tau(\vec{p})}\right)\delta_{\vec{p},\vec{p}'}\right] f(\vec{r},\vec{p}',t) . \tag{31}$$

For homogenous samples, the BTE reduces to:

$$\begin{aligned} &\frac{\partial f}{\partial t} + \vec{v}\nabla_r f + (-e)\vec{\varepsilon}\nabla_p f + \Gamma(\vec{p}) f(\vec{r},\vec{p},t) = \tilde{I}(\vec{p},t) = \\ &= \sum_{\vec{p}'}\left[\sum_{i=1}^{N} S_i(\vec{p}',\vec{p}) + \left(\Gamma(\vec{p}) - \frac{1}{\tau(\vec{p})}\right)\delta_{\vec{p},\vec{p}'}\right] f(\vec{p}',t) \end{aligned} . \tag{32}$$

In the last formulation of the BTE, the coordinate space (phase space) is fixed and the electrons move along given trajectory in response to the applied forces. With the introduction of variables:

$$\begin{cases} \tilde{t} = t \\ \tilde{p} = \vec{p} + e\vec{\varepsilon}t \end{cases} , \tag{33}$$

we go to a description in which electrons are frozen in their positions and the coordinate system is moving. Then,

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial \tilde{t}} \cdot \frac{\partial \tilde{t}}{\partial t} + \frac{\partial f}{\partial \tilde{p}} \cdot \frac{\partial \tilde{p}}{\partial t} = \frac{\partial f}{\partial \tilde{t}} + e\vec{\varepsilon} \cdot \frac{\partial f}{\partial \tilde{p}} , \tag{34a}$$

$$\frac{\partial f}{\partial \vec{p}} = \frac{\partial f}{\partial \tilde{t}} \cdot \frac{\partial \tilde{t}}{\partial \vec{p}} + \frac{\partial f}{\partial \tilde{p}}\frac{\partial \tilde{p}}{\partial \vec{p}} = \frac{\partial f}{\partial \tilde{p}} . \tag{34b}$$

In this notation, the BTE becomes:

$$\frac{\partial f}{\partial \tilde{t}} + \Gamma \cdot f(\tilde{p} - e\vec{\varepsilon}\tilde{t}, \tilde{t}) = \tilde{I}(\tilde{p} - e\vec{\varepsilon}\tilde{t}, \tilde{t}) . \tag{35}$$

The solution of the homogenous equation of the form:

$$\frac{\partial f}{\partial \tilde{t}} + \Gamma \cdot f(\tilde{p} - e\vec{\varepsilon}\tilde{t}, \tilde{t}) = 0 , \tag{36}$$

can be found using a separation of variables method, to be:

$$\ln f(\tilde{p} - e\vec{\varepsilon}\tilde{t}, \tilde{t})\Big|_0^{\tilde{t}} = -\Gamma\int_0^{\tilde{t}} d\tilde{t} , \tag{37a}$$

or:

$$\ln\left[f(\tilde{p} - e\vec{\varepsilon}\tilde{t}, \tilde{t}) / f(\tilde{p},0)\right] = -\Gamma\tilde{t} , \tag{37b}$$

to get:

$$f(\tilde{p} - e\vec{\varepsilon}\tilde{t}, \tilde{t}) = f(\tilde{p},0)e^{-\Gamma\tilde{t}} . \tag{38}$$

Going back to the original coordinate system gives:

$$f(\vec{p},t) = f(\vec{p} + e\vec{\varepsilon}t, 0)e^{-\Gamma t} . \tag{39}$$

This term is the transient term. It states that an electron initially in a state $(\vec{p}+e\vec{\varepsilon}t)$ at time $t=0$, has arrived in a state $\vec{p}$ at time t without scattering. This event occurs with a transition probability:

$$P(\vec{p},t,0)=e^{-\Gamma t}\text{, for }\Gamma(\vec{p})=\Gamma=const. \tag{40}$$

For general case, when $\Gamma(\vec{p})$ is not a constant, one would have had:

$$P(\vec{p},t,0)=\exp\left[-\int_0^t \Gamma(\vec{p})dt'\right]. \tag{41}$$

Also, if the initial momentum of the electron is $(\vec{p}+e\vec{\varepsilon}t)$, because all of the drift motion and the acceleration by the electric field, the final electron momentum at time t equals to: $\vec{p}'=\vec{p}+e\vec{\varepsilon}t-e\vec{\varepsilon}t=\vec{p}$.

The next task is to find a solution of the BTE for homogenous systems. The homogenous solution suggests that the general solution will also involve an exponentials. For this purpose we define a function:

$$f_1(\tilde{p}-e\vec{\varepsilon}\tilde{t},\tilde{t})=f(\tilde{p}-e\vec{\varepsilon}\tilde{t},\tilde{t})e^{\Gamma\tilde{t}}, \tag{42a}$$

which leads to:

$$f(\tilde{p}-e\vec{\varepsilon}\tilde{t},\tilde{t})=f_1(\tilde{p}-e\vec{\varepsilon}\tilde{t},\tilde{t})e^{-\Gamma\tilde{t}}. \tag{42b}$$

Then:

$$\frac{\partial f}{\partial\tilde{t}}=\frac{\partial f_1}{\partial\tilde{t}}e^{-\Gamma\tilde{t}}+\Gamma\cdot f_1(\tilde{p}-e\vec{\varepsilon}\tilde{t},\tilde{t})e^{-\Gamma\tilde{t}}. \tag{43}$$

Substituting this result back into the BTE gives:

$$\frac{\partial f_1}{\partial\tilde{t}}=e^{\Gamma\tilde{t}}\tilde{I}(\tilde{p}-e\vec{\varepsilon}\tilde{t},\tilde{t}). \tag{44}$$

Solving the last equation for f_1 finally leads to:

$$f_1(\tilde{p}-e\vec{\varepsilon}\tilde{t},\tilde{t})\Big|_0^{\tilde{t}}=\int_0^{\tilde{t}}d\tilde{t}_1\cdot e^{\Gamma\tilde{t}_1}\tilde{I}(\tilde{p}-e\vec{\varepsilon}\tilde{t}_1,\tilde{t}_1), \tag{45a}$$

or:

$$f_1(\tilde{p}-e\vec{\varepsilon}\tilde{t},\tilde{t})=f_1(\tilde{p},0)+\int_0^{\tilde{t}}d\tilde{t}_1\cdot e^{\Gamma\tilde{t}_1}\tilde{I}(\tilde{p}-e\vec{\varepsilon}\tilde{t}_1,\tilde{t}_1). \tag{45b}$$

Multiplying the last equation by $e^{-\Gamma\cdot\tilde{t}}$ one gets:

$$f(\tilde{p}-e\vec{\varepsilon}\tilde{t},\tilde{t})=f(\tilde{p},0)\cdot e^{-\Gamma\tilde{t}}+\int_0^{\tilde{t}}d\tilde{t}_1\cdot e^{-\Gamma(\tilde{t}-\tilde{t}_1)}\tilde{I}(\tilde{p}-e\vec{\varepsilon}\tilde{t}_1,\tilde{t}_1). \tag{46}$$

Returning to the original coordinate system gives:

$$f(\vec{p},t)=f(\vec{p}+e\vec{\varepsilon}t,0)e^{-\Gamma t}+\int_0^t dt_1\cdot e^{-\Gamma(t-t_1)}\tilde{I}(\vec{p},t_1), \tag{47}$$

where $\tilde{I}(\vec{p},t)=\sum_{\vec{p}'}S_{eff}(\vec{p}',\vec{p})f(\vec{p}',t)$. Substituting this back gives:

$$f(\vec{p},t)=f(\vec{p}+e\vec{\varepsilon}t,0)e^{-\Gamma t}+\int_0^t dt_1\sum_{\vec{p}'} f(\vec{p}',t_1)S_{eff}(\vec{p}',\vec{p}+e\vec{\mathcal{E}}(t-t_1))e^{-\Gamma(t-t_1)}\,, \tag{48}$$

where:

- $f(\vec{p}+e\vec{\varepsilon}t,0)e^{-\Gamma t}$ is the transient term;
- $f(\vec{p}',t_1)$ is the probability that at time t_1 a state $\vec{p}'$ is occupied by an electron;
- $S_{eff}(\vec{p}',\vec{p}+e\vec{\mathcal{E}}(t-t_1))$ is the transition rate (probability) from state $\vec{p}'$ to state $\vec{p}+e\vec{\mathcal{E}}(t-t_1)$.
- $e^{-\Gamma(t-t_1)}$ is the probability that an electron will not undergo collision event in interval $(t-t_1)$.

This last expression is known as Chambers-Rees path integral. Rees [29] innovation is the introduction of the fictitious scattering term. Ignoring the transient term, one can find the solution of the distribution function using the following iterative procedure that is obtained by time discretization, i.e. using $t=N\cdot\Delta t$ and $t_n=n\cdot\Delta t$. Then,

$$f_N(\vec{p})=\Delta t\sum_{m=0}^{N-1}\sum_{\vec{p}'} f_m(\vec{p}')S_{eff}(\vec{p}',\vec{p}+e\vec{\varepsilon}(N-m)\Delta t)e^{-\Gamma(N-m)\Delta t}\,. \tag{49}$$

The two step procedure is then found by using N=1, which means that $t=\Delta t$, i.e.,

$$f_1(\vec{p})=\Delta t\sum_{\vec{p}'} f_0(\vec{p}')S_{eff}(\vec{p}',\vec{p}+e\vec{\mathcal{E}}\Delta t)e^{-\Gamma\Delta t}\,, \tag{50}$$

where,

- $g_0(\vec{p}+e\vec{\mathcal{E}}\Delta t)=f_0(\vec{p}')\cdot S_{eff}(\vec{p}',\vec{p}+e\vec{\mathcal{E}}\Delta t)$ is the intermediate function that describes the occupancy of a state $\vec{p}+e\vec{\mathcal{E}}\Delta t$ at time $t=0$, which can be changed due to in-scattering events;
- $e^{-\Gamma\cdot\Delta t}$ is the probability that no scattering occurred within time integral Δt (free-flight).

Now assume that $t=2\Delta t$. This then gives:

$$\begin{aligned} f_2(\vec{p}) &= \Delta t\sum_{m=0}^{1}\sum_{\vec{p}'} f_m(\vec{p}')S_{eff}(\vec{p}',\vec{p}+e\vec{\varepsilon}(2-m)\Delta t)e^{-\Gamma(2-m)\Delta t} = \\ &= \Delta t\left\{\sum_{\vec{p}'} f_0(\vec{p}')S_{eff}(\vec{p}',\vec{p}+e\vec{\varepsilon}(2\Delta t))e^{-\Gamma(2\Delta t)} + \right. \\ &\qquad \left. + \sum_{\vec{p}'} f_1(\vec{p}')S_{eff}(\vec{p}',\vec{p}+e\vec{\mathcal{E}}\Delta t)e^{-\Gamma\Delta t}\right\} \end{aligned} \tag{51}$$

These examples suggest that the evaluation of $f_{n+1}(\vec{p})$ involves integration over trajectories and the exponential factors just give the probability that no scattering has occurred.

2.2 Bulk Monte Carlo Method

According to the description provided in Section 2.1 above, in the bulk 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 **k** due to acceleration by internal and external fields. For electrons subject to time independent electric and magnetic fields, the time evolution of **k** between collisions is represented as

$$\mathbf{k}(t) = \mathbf{k}(0) - \frac{e(\mathbf{E} + \mathbf{v} \times \mathbf{B})t}{\hbar}, \tag{52}$$

where **E** is the electric field, **v** is the electron velocity and **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 $\exp\left(-\int_0^t \Gamma\left[\mathbf{k}\left(t'\right)\right]dt'\right)$. Thus, the probability of scattering in the time interval dt after a free flight of time t may be written as the joint probability

$$P(t)\,dt = \Gamma[\mathbf{k}(t)] \exp\left[-\int_0^t \Gamma\left[\mathbf{k}\left(t'\right)\right]dt'\right]dt. \tag{53}$$

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 [24]), random flight times sampled from $P(t)$ may be generated according to

$$r = \int_0^{t_r} P(t)dt, \tag{54}$$

where r is a uniformly distributed random number and t_r is the desired free flight time. Integrating Eq. (54) with $P(t)$ given by Eq. (53) above yields

$$r = 1 - \exp\left[-\int_0^{t_r} \Gamma\left[\mathbf{k}\left(t'\right)\right]dt'\right]. \tag{55}$$

Since $1-r$ is statistically the same as r, Eq. (55) may be simplified to

$$-\ln r = \int_0^{t_r} \Gamma\left[\mathbf{k}\left(t'\right)\right]dt'. \tag{56}$$

Eq. (56) 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 **k** between scattering events (for example in ultrafast photo-excitation experiments with no applied bias), the time dependence vanishes, and the integral is trivially evaluated. As noted in the previous section, in the general case where this simplification is not possible, it is expedient to introduce the so called self-scattering method [29], in which one introduces 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

$$\Gamma = \Gamma\left[\mathbf{k}\left(t'\right)\right] + \Gamma_{self}\left[\mathbf{k}\left(t'\right)\right], \tag{57}$$

where $\Gamma_{self}[\mathbf{k}(t')]$ is the self-scattering rate (see the discussion in Section 2.1 above). 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. (56) such that the free flight is given by

$$t_r = -\frac{1}{\Gamma}\ln r. \tag{58}$$

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 [30].

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

$$\Gamma = \Gamma_{self}\left[n,\mathbf{k}\right] + \Gamma_1\left[n,\mathbf{k}\right] + \Gamma_2\left[n,\mathbf{k}\right] + \ldots \Gamma_N\left[n,\mathbf{k}\right], \tag{59}$$

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 6.

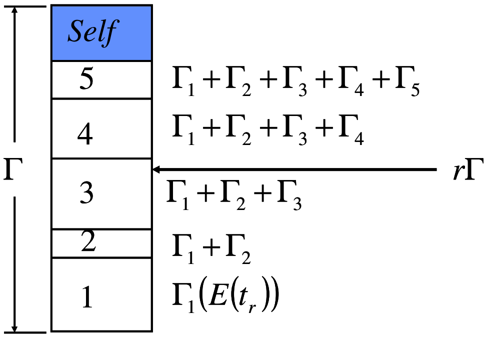

Self
5
4
3
2
1
$\Gamma_1+\Gamma_2+\Gamma_3+\Gamma_4+\Gamma_5$
$\Gamma_1+\Gamma_2+\Gamma_3+\Gamma_4$
$\Gamma_1+\Gamma_2+\Gamma_3$
$\Gamma_1+\Gamma_2$
$\Gamma_1(E(t_r))$
Γ
$r\Gamma$

Figure 6. Selection of the type of scattering terminating a free flight in the Monte Carlo algorithm.

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 **k** and **k**' are the particle wavevectors before and after scattering. Defining a spherical coordinate system around the initial wavevector **k**, the final wavevector **k**' is specified by $|\mathbf{k}'|$ (which depends on conservation of energy) as well as the azimuthal and polar angles, φ and θ around **k**. Typically, the scattering rate, $\Gamma_j[n,\mathbf{k};m,\mathbf{k}']$, only depends on the angle θ between **k** and **k**'. Therefore, φ may be chosen using a uniform random number between 0 and 2π (i.e. $2\pi r$), while θ is chosen according to the angular dependence for scattering arising from $\Gamma_j[n,\mathbf{k};m,\mathbf{k}']$. 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. (54). Otherwise, a rejection technique (see, for example, [24,25]) may be used to select random angles according to $P(\theta)$. Scattering mechanisms that contribute to transport are summarized in Figure 7. The corresponding scattering rates for general non-parabolic bands are summarized in Table 2.

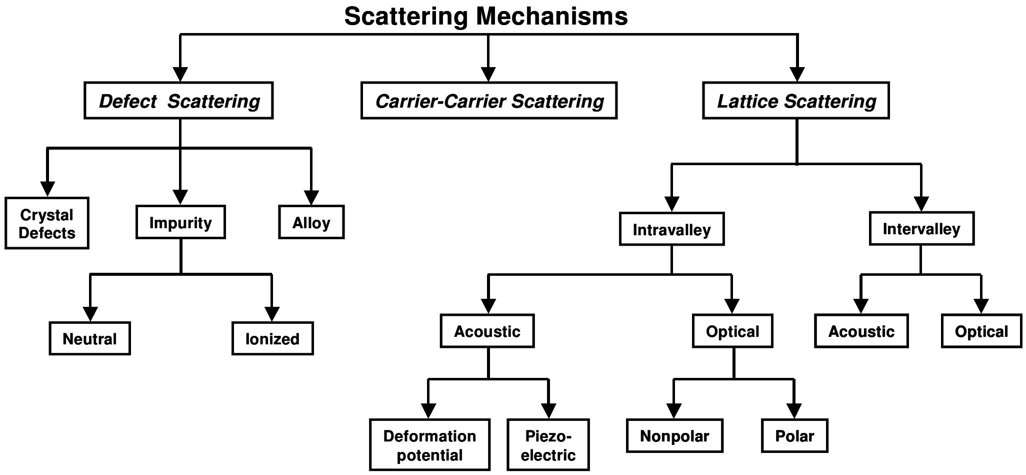


Figure 7. Scattering mechanisms in a typical semiconductor.

A general Monte Carlo code is developed as follows. First a subroutine is typically called that contains all material and scattering rates parameters for the scattering mechanisms included in the theoretical model. After the material and run parameters are read in, in the first step of the Monte Carlo simulation procedure it is necessary to construct scattering tables for the Γ, L and X valleys (for GaAs as a prototypical example) that initializes a series of events that are summarized in Figure 8. At each energy, the cumulative scattering rates for each valley are stored in separate look-up tables, and renormalized according to the maximum scattering rate (including self-scattering) that occurs over the range of energies stored. The structure of these subroutines is such that adding additional scattering event has to be trivial.

Table 2. Scattering rates expressions for non-parabolic bands.

1. Acoustic Phonon Scattering
$W(E)=\left(\frac{2\pi D_{ac}^2 K_B T_L}{\hbar C_l}\right)*\left(\frac{(2m_d)^{\frac{3}{2}}\sqrt{E(1+\alpha E)}}{4\pi^2\hbar^3}\right)*(1+2\alpha E)$
2. Intervalley Phonon Scattering
$W(E)=\left(\frac{\pi D_{ij}^2 Z_j}{\rho w_{ij}}\right)*\left(n(w_{ij})+\frac{1}{2}\mp\frac{1}{2}\right)*\left(\frac{(2m_d)^{\frac{3}{2}}\sqrt{E_f(1+\alpha E_f)}}{4\pi^2\hbar^3}\right)*(1+2\alpha E_f)$ $E_f = E\pm\hbar w_{ij}-\Delta E_{ij}$
3. Ionized Impurity Scattering
$W(E)=\left(\frac{\sqrt{2}e^4 N_I m_d^{\frac{3}{2}}}{\pi\varepsilon_s^2\hbar^4}\right)*\left(\sqrt{E(1+\alpha E)}*(1+2\alpha E)\right)*\left(\frac{1}{q_D^2\left(q_D^2+\frac{8m_d E(1+\alpha E)}{\hbar^2}\right)}\right)$ $q_D=\sqrt{\frac{e^2 N_I}{\varepsilon K_B T_L}}$
4. Polar Optical Phonon Scattering
$W(E)=\left(\frac{\sqrt{m_d}\,e^2 w_{LO}}{4\sqrt{2}\pi\hbar\varepsilon_p}\right)*\left(N_o+\frac{1}{2}\mp\frac{1}{2}\right)*\left(\frac{1+2\alpha E_k'}{\gamma_k}\right)*F(E_k,E_k')$ $N_o=\frac{1}{e^{\frac{\hbar w_{LO}}{K_B T_L}}-1}$ $\quad\varepsilon_p=\frac{1}{\frac{1}{\varepsilon_{high}}-\frac{1}{\varepsilon_{low}}}$ $\quad F(E_k,E_k')=\ln\left(mod\left(\frac{\sqrt{\gamma_k}+\sqrt{\gamma_k'}}{\sqrt{\gamma_k}-\sqrt{\gamma_k'}}\right)\right)$ $\gamma_k=E_k(1+\alpha E_k)$ $E_k'=E_k\pm\hbar w_{LO}$

5. Piezoelectric Scattering

$$W(E)=\left(\frac{m_d^{\frac{1}{2}}K_BT_L}{4\sqrt{2}\pi\rho v_s^2\hbar^2}\right)*\left(\frac{1+2\alpha E}{\sqrt{E(1+\alpha E)}}\right)*\left(\frac{ee_{pz}}{\varepsilon_\infty}\right)^2*\ln\left(1+\frac{8m_dE(1+\alpha E)}{\hbar^2q_D^2}\right)$$

$$q_D=\sqrt{\frac{e^2N_I}{\varepsilon K_BT_L}}$$

6. Dislocation Scattering (e.g. GaN)

$$W(E)=\left(\frac{N_{dis}m_de^4}{4\hbar^3\varepsilon^2c^2}\right)*\left(\frac{\lambda^4}{\left(1+\frac{8\lambda^2m_dE(1+\alpha E)}{\hbar^2}\right)^{\frac{3}{2}}}\right)*\left(1+\frac{4\lambda^2m_dE(1+\alpha E)}{\hbar^2}\right)*(1+2\alpha E)$$

$$\lambda=\sqrt{\frac{\varepsilon K_BT_L}{e^2n'}}$$

where *n'* is the effective screening concentration

N_{dis} is the Line dislocation density

7. Alloy Desorder Scattering ($Al_xGa_{1-x}As$)

$$W(E)=\left(\frac{x(1-x)a^3}{\pi}\right)*\left(\frac{D_{alloy}^2d}{\hbar^4}\right)*m_d\sqrt{2m_dE(1+\alpha E)}*(1+2\alpha E)$$

Where: *d* is the lattice disorder ($0\le d\le 1$)

D_{alloy} is the alloy disorder scattering potential

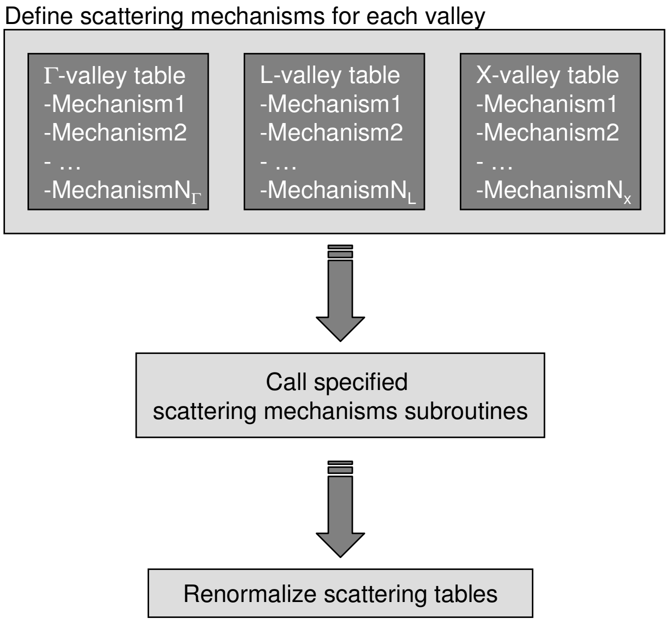


Figure 8. Procedure for the creation of the scattering tables.

Having constructed the scattering table and after renormalizing the table, examples of which are given in Figure 9 and Figure 10 for the Γ, L, and X valley, the next step is to initialize carriers wavevector and energy and the initial free-flight time. This is accomplished by calling the initialization subroutine. Energy and wavevector histograms of the initial carrier energy and the components of the wave-vector along the x-, y-, and z-axes are shown in Figure 11. For good statistics, the number of particles simulated is 10000, and one can see the statistical fluctuation of these average quantities associated with the finite number of particles. Notice that the initial y-component for the wavevector is symmetric around the y-axis which means that the average wavevector along the y-axis is zero, which should be expected since the electric field along the y-component is zero at *t*=0. Identical distributions have been obtained for the x- and for the z-components of the wavevector.

Also note that the energy distribution has the Maxwell-Boltzmann form as it should be expected. One can also estimate from this graph that the average energy of the carriers is on the order of (3/2)$k_B T$.

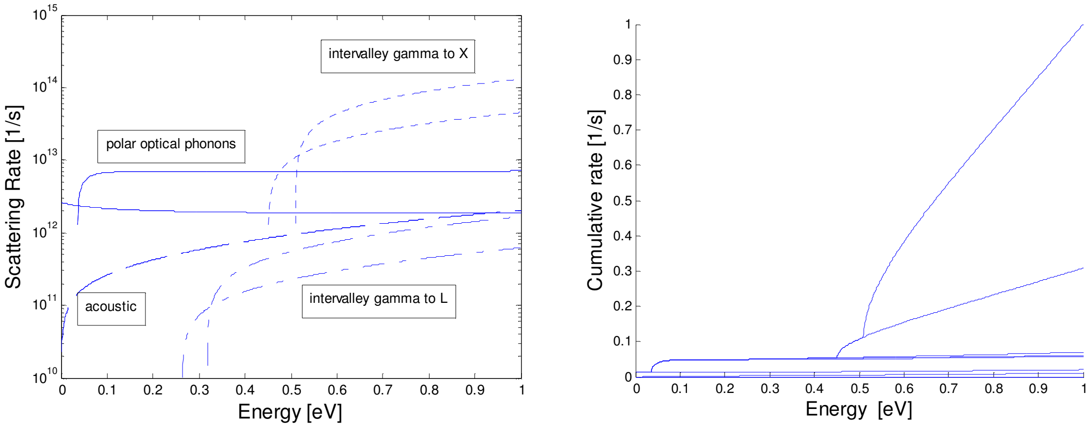


Figure 9. Left panel: scattering rates for the Γ-valley. For simplicity we have omitted Coulomb scattering in these calculations. In the left figure, the dashed line corresponds to the acoustic phonon scattering rate, solid lines correspond to polar optical phonon scattering (absorption and emission), and the dashed-dotted line corresponds to intervalley scattering from Γ-valley to L-valley. Since the L-valley is along the [111] direction, there are 8 equivalent directions and since these valleys are shared there are a total of 4 equivalent L valleys. The dotted line corresponds to scattering from the Γ-valley to X-valleys. The X-valleys are at the [100] direction and since there are 6 equivalent [100] directions and the valleys are shared between Brillouin zones, there are 3 equivalent X valleys. Right panel: normalized cumulative scattering table for the Γ-valley. Everything above the top line up to Γ=1 is self-scattering so it is advisable when checking the scattering mechanisms to first check whether the scattering mechanism chosen is self-scattering or not. This is in particular important for energies below 0.5 eV for this particular scattering table when the Γ to X intervalley scattering (absorption and emission) takes over.

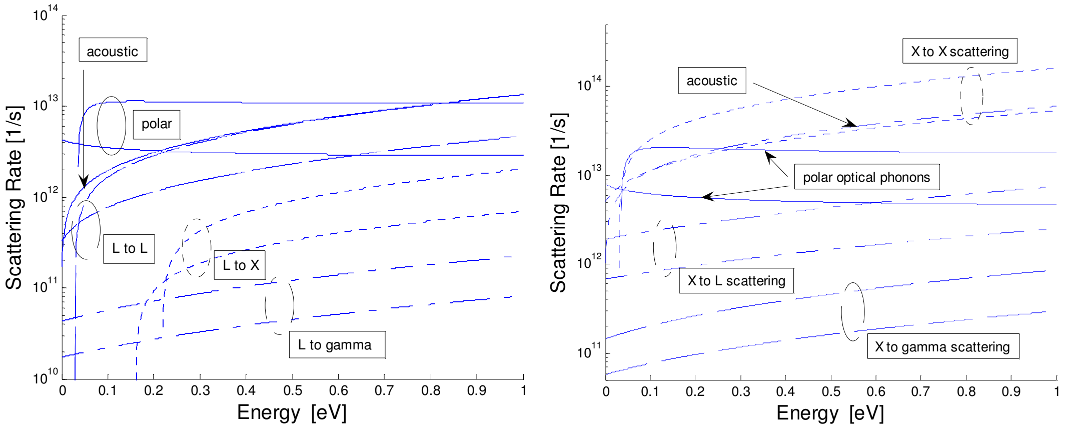


Figure 10. Scattering rates for the L (left panel) and X (right panel) valleys used to create the corresponding normalized scattering tables (not shown here).

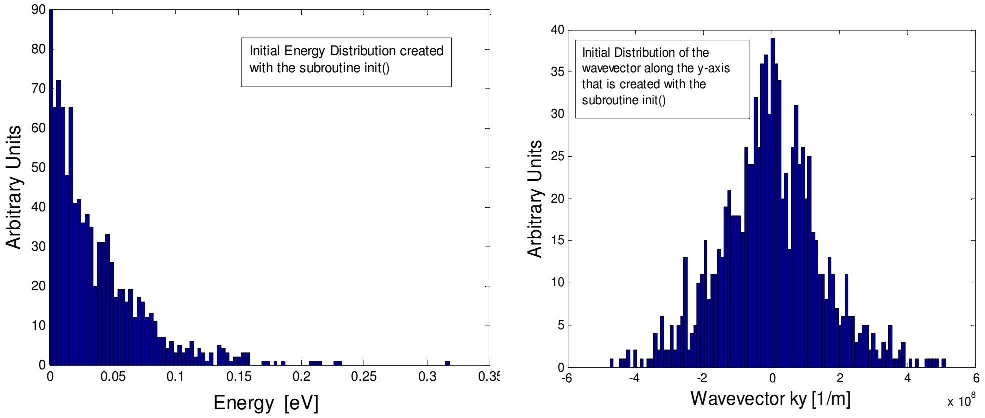


Figure 11. Initial carrier distribution for an ensemble of 10000 Particles. Left panel: distribution of wavevector k_y. Right panel: energy distribution.

When the initialization process is finished, the main free-flight-scatter procedure takes place until the completion of the simulation time. There are two components in this routine; first the carriers accelerate freely due to the electric field, accomplished by calling the **drift()** subroutine, and then their free-flights are interrupted by random scattering events that are managed by the **scatter_carrier()** subroutine. The flow-chart for performing the free-flight-scatter process within one time step Δ*t* is shown diagrammatically in Figure 12.

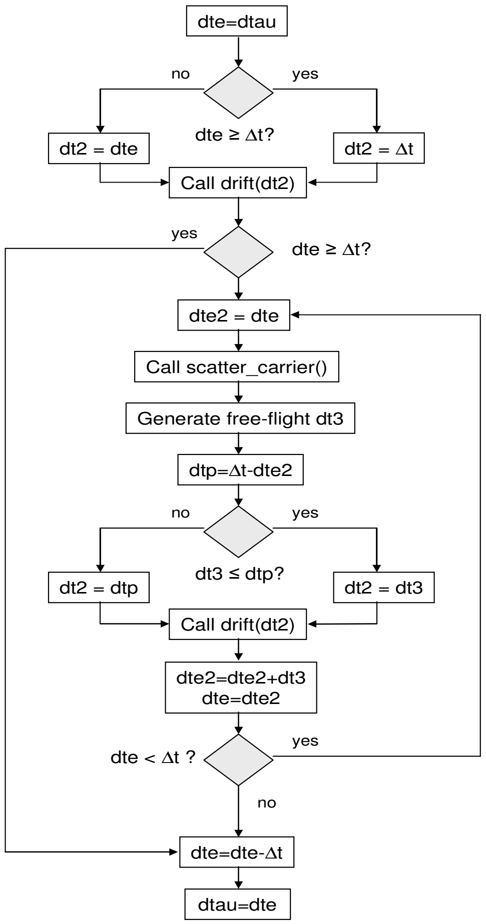


Figure 12. Free-flight-scatter procedure within one time step.

In the **scatter_carrier()** subroutine, first the scattering mechanism terminating the free flight is chosen, to which certain attributes are associated such as the change in energy after scattering. For inelastic scattering processes, we have the change in energy due to emission or absorption of phonons, for example. Also, the nature of the scattering process is identified: isotropic or anisotropic. Note that when performing acoustic phonon and intervalley scattering for GaAs, both of which are isotropic scattering processes, no coordinate system transformation is needed to determine the final wavevector after scattering. Because polar optical phonon and Coulomb scattering mechanisms are anisotropic, it is necessary to do a rotation of the coordinate system, scatter the carrier in the rotated system and then perform inverse coordinate transformation. This procedure is needed because it is much easier to determine final carrier momentum in the rotated coordinate system in which the initial wavevector k is aligned with the z-axis. For this case, one can calculate that the final polar angle for scattering with polar optical phonons for parabolic bands in the rotated coordinate system is

$$\cos\theta = \frac{(1+\xi)-(1+2\xi)^r}{\xi}, \quad \xi = \frac{2\sqrt{E_k\left(E_k \pm \hbar\omega_0\right)}}{\left(\sqrt{E_k}-\sqrt{E_k \pm \hbar\omega_0}\right)^2} \tag{60}$$

where E_k is the carrier energy, $\hbar\omega_0$ is the polar optical phonon energy and r is a random number uniformly distributed between 0 and 1. The final angle for scattering with ionized impurities (Coulomb scattering) and for parabolic bands is

$$\cos\theta = 1 - \frac{2r}{1+4k^2L_D^2(1-r)} \tag{61}$$

where **k** is the carrier wavevector, and L_D is the Debye screening length. The azimuthal angle for both scattering processes is simply calculated using $\varphi = 2\pi r$. The importance of properly calculating the angle θ after scattering to describe small angle deflections in the case of Coulomb or polar optical phonon scattering is illustrated in

Figure 13 (from 0 to π=3.141592654) where we plot the histogram of the polar angle after scattering for electron-polar optical phonon scattering, where we can clearly see the preference for small angle deflections that are characteristic for any Coulomb type interaction (polar optical phonon is in fact electron-dipole interaction). Graphical representation of the determination of the final angle after scattering for both isotropic and anisotropic scattering processes is given in Figure 14.

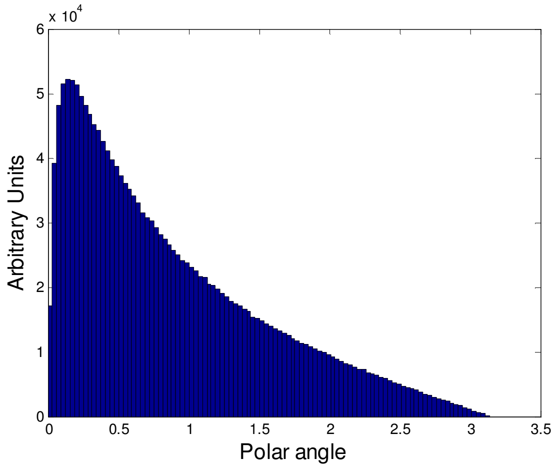
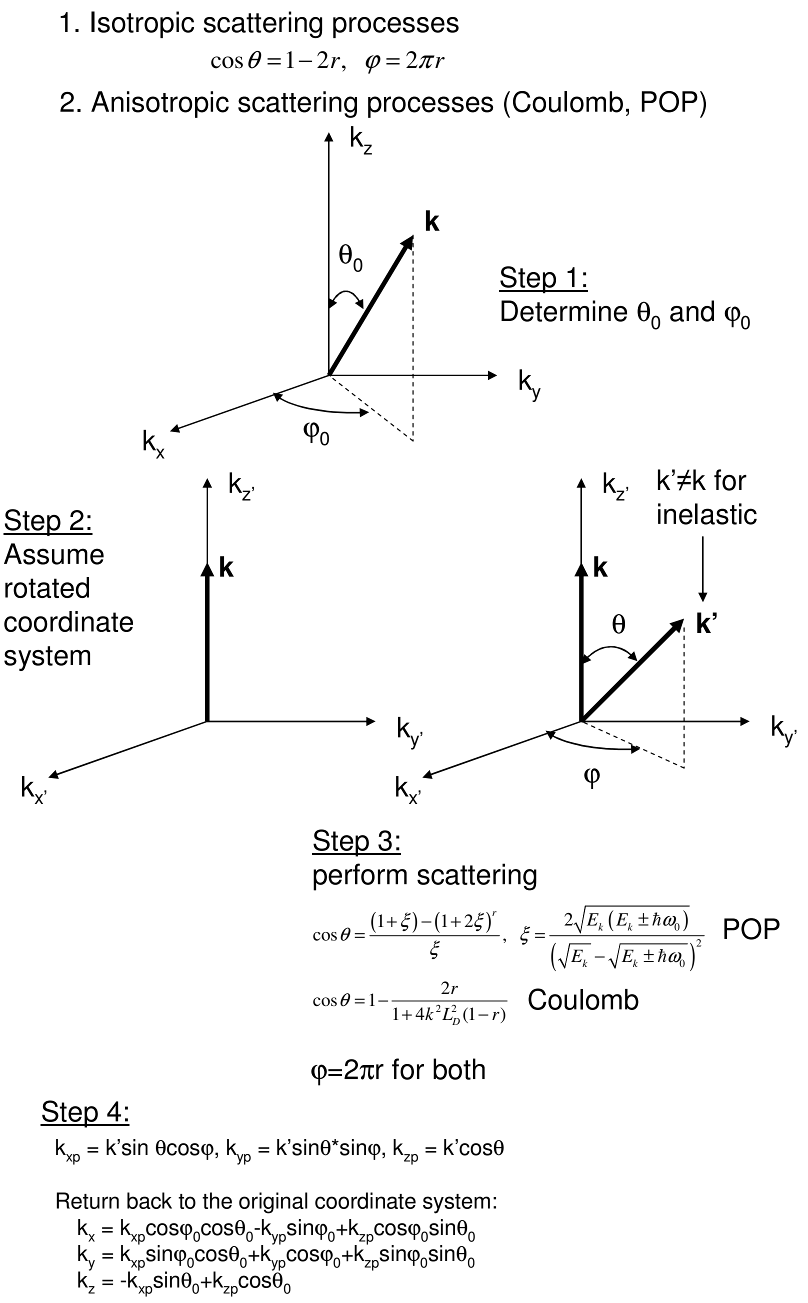


Figure 13. Histogram of the polar angle for electron – polar optical phonon scattering.


1. Isotropic scattering processes

$$\cos\theta = 1 - 2r, \quad \varphi = 2\pi r$$

2. Anisotropic scattering processes (Coulomb, POP)

Step 1:
Determine θ_0 and φ_0

Step 2:
Assume rotated coordinate system

k'≠k for inelastic

Step 3:
perform scattering

$$\cos\theta = \frac{(1+\xi)-(1+2\xi)^r}{\xi}, \quad \xi = \frac{2\sqrt{E_k(E_k \pm \hbar\omega_0)}}{\left(\sqrt{E_k} - \sqrt{E_k \pm \hbar\omega_0}\right)^2} \quad \text{POP}$$

$$\cos\theta = 1 - \frac{2r}{1+4k^2 L_D^2 (1-r)} \quad \text{Coulomb}$$

φ=2πr for both

Step 4:

$k_{xp} = k'\sin\theta\cos\varphi$, $k_{yp} = k'\sin\theta{*}\sin\varphi$, $k_{zp} = k'\cos\theta$

Return back to the original coordinate system:

$k_x = k_{xp}\cos\varphi_0\cos\theta_0 - k_{yp}\sin\varphi_0 + k_{zp}\cos\varphi_0\sin\theta_0$

$k_y = k_{xp}\sin\varphi_0\cos\theta_0 + k_{yp}\cos\varphi_0 + k_{zp}\sin\varphi_0\sin\theta_0$

$k_z = -k_{xp}\sin\theta_0 + k_{zp}\cos\theta_0$

Figure 14. Description of final angle selection for isotropic and anisotropic scattering processes using the direct technique.

The direct technique described above can be applied when the integrals describing cosθ can be analytically calculated. For most cases of interest, the integral cannot be easily inverted. In these cases a rejection technique may be employed. The procedure of the rejection technique goes as follows:

- Choose a maximum value C, such that $C > f(x)$ for all x in the interval (a,b).
- Choose pairs of random numbers, one between a and b ($x_1 = a + r_1(b-a)$) and another $f_1 = r_1'C$ between 0 and C, where r_1 and r_1' are random numbers uniformly distributed between zero and 1.
- If $f_1 \le f(x_1)$, then the number x_1 is accepted as a suitable value, otherwise it is rejected.

The three steps described above are schematically shown in the figure below (Figure 15). For $x = x_1$, r_1C is larger than $f(x_1)$ and in this case if this represents the final polar angle for scattering, this angle is rejected and a new sequence of two random numbers is generated to determine x_2 and r_2C. In this second case, $f(x_2) > r_2C$ and the polar angle $\theta = x_2$ is selected (for polar angle selection $a = 0$ and $b = \pi$).

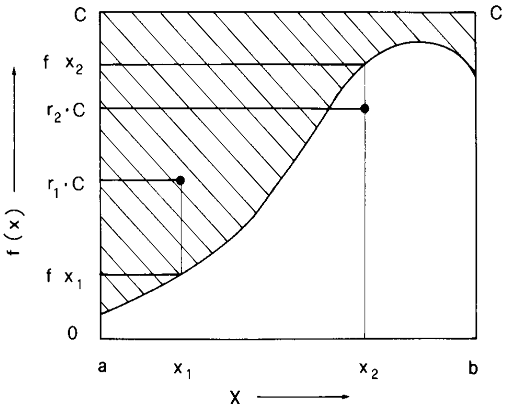


Figure 15. Schematic description of the rejection technique**.**

After the simulation is completed, typical results to check are the velocity-time, the energy-time and the valley occupation versus time characteristics, such as those shown in Figure 16 , where the velocity time characteristics for applied electric fields ranging from 0.5 to 7 kV/cm, with an electric field increment of 0.5 kV/cm, are shown. These clearly demonstrate that after a transient phase, the system reaches a stationary steady state, after which time we can start taking averages for calculating steady-state quantities.

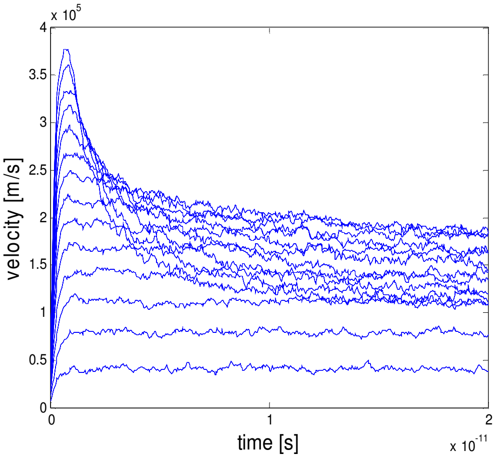


Figure 16. Time evolution of the drift velocity for electric field strengths ranging between 0.5 and 7 kV/cm, in 0.5 kV/cm increments**.**

From the results shown in Figure 16, one can see that steady-state is achieved for larger time intervals when the electric field value is increased and the carriers are still sitting in the Γ-valley. Afterwards the time needed to get to steady-state decreases. This trend is related to the valley repopulation and movement of the carriers from the Γ, into the X and finally into the L valley. The steady-state velocity-field and valley population versus electric field characteristics are shown in Figure 17 and Figure 18, respectively. One can clearly see on the velocity-field characteristics that a low-field mobility of about 8000 cm^2/V-s is correctly reproduced for GaAs without the use of any adjustable parameters.

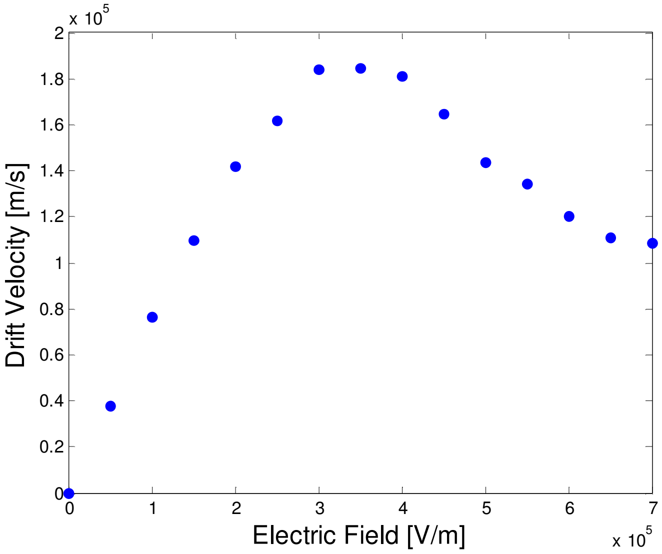


Figure 17. Steady state drift velocity vs. electric field.

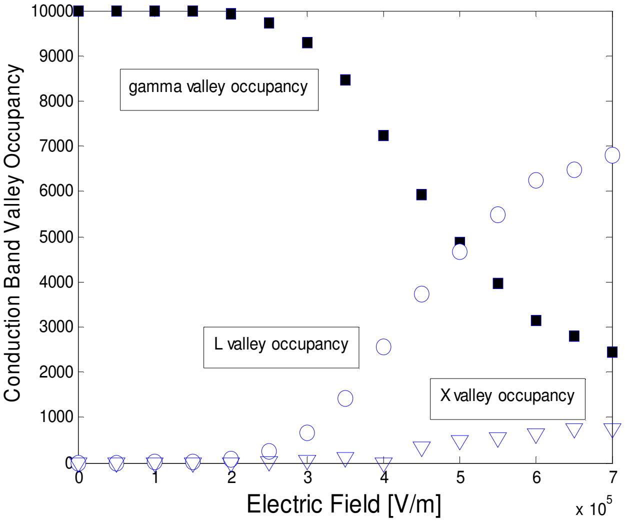


Figure 18. Different valley occupancy vs. electric field**.**

At this point, it is advisable to check the energy and wavevector histograms (Figure 19) to ensure that the energy range chosen in the scattering tables is correct or not for the particular maximum electric field strength being considered, which gives the worst case scenario. Since, as already noted, we apply the electric field in the y-direction, for comparative purposes we plot the histograms of the x-component of the wavevector, y-component of the wavevector, and the histogram of the final carrier energy distribution for which a drifted Maxwellian form is evident. Since there is no field applied in the x-direction, we see that the average wavevector in the x-direction is 0. Due to the application of the field in the y-direction, there is a finite positive shift in the y-component of the velocity, which is yet another signature for the displaced Maxwellian form of the energy distribution in the bottom histogram.

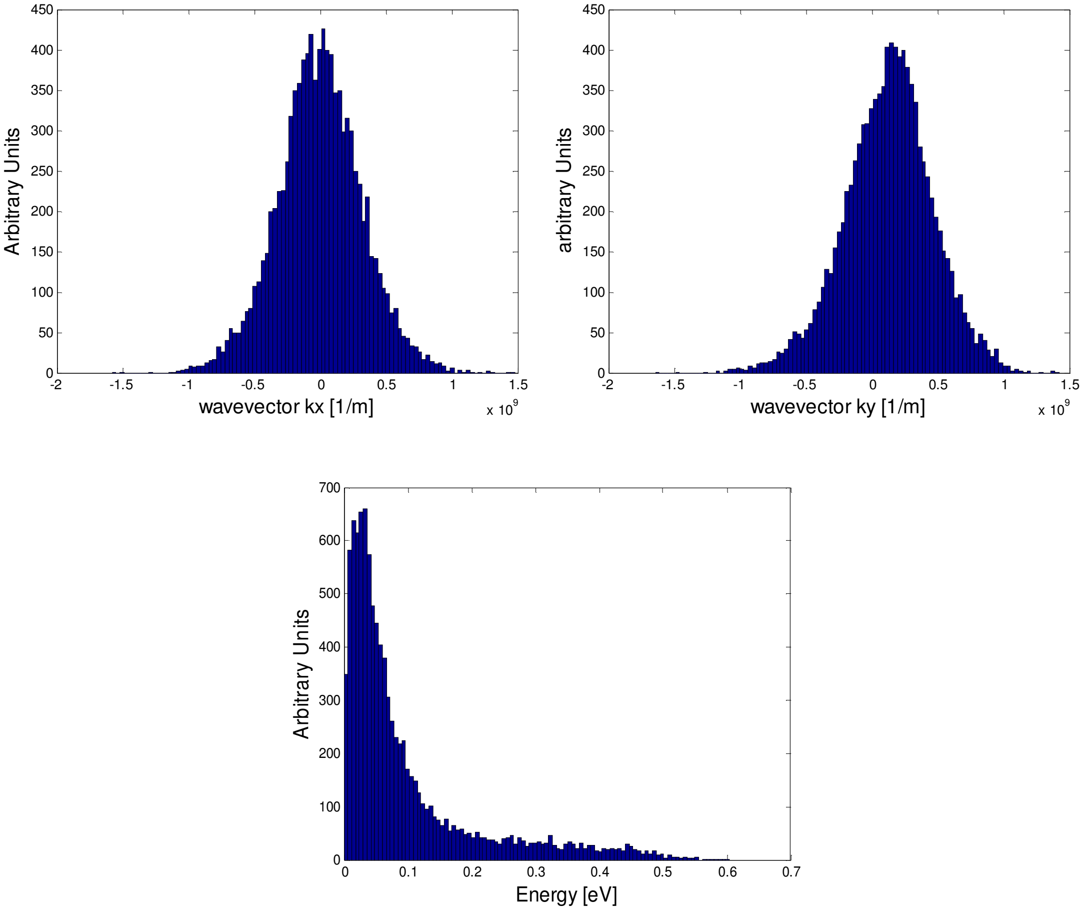


Figure 19. Top left panel: histogram of the x-component of the wavevector. Top right panel: Histogram of the y-component of the wavevector. Bottom panel: histogram of the carrier energy. Applied electric field is 7kV/cm.

3. Particle-Based Device Simulation

In Section 2.2, we introduced the numerical solution of the BTE using Monte Carlo method. Within a device, both the transport kernel and the field solver are coupled to each other (see Figure 2). The field associated with the potential coming from Poisson's equation is the driving force accelerating particles in the Monte Carlo phase, for example, while 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. Below we give an extensive description of the Monte Carlo particle-based device simulators with emphasis on the particle-mesh coupling.

Within the particle-based EMC method with its time-marching algorithm, Poisson's equation may be decoupled from the BTE 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 (see discussion in Ref. [45]). Note that Poisson's equation is solved on a mesh, whereas the solution of charge motion using EMC occurs over a continuous range of coordinate space in terms of the particle position. Therefore, a particle-mesh (PM) coupling is needed for both the charge assignment and the force interpolation. The PM coupling is broken into four steps: (1) assign particle charge to the mesh; (2) solve the Poisson equation on the mesh; (3) calculate the mesh-defined forces; and (4) interpolate to find forces on the particle. There are a variety of schemes that can be used for the PM coupling and these are discussed in Section 3.4.

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

$$\begin{aligned}\mathbf{v}(t) &= \frac{d\mathbf{r}}{dt} = \frac{1}{\hbar}\nabla_{\mathbf{k}} E\left(\mathbf{k}(t)\right) \\ \frac{d\mathbf{k}}{dt} &= \frac{q\mathbf{E}(\mathbf{r})}{\hbar}\end{aligned} \tag{62}$$

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, **k**(*t*), is related to the velocity through the gradient of *E* with respect to **k.** If one has to use the full band-structure of the semiconductor, then integration of these equations to find **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.

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = \frac{\hbar\mathbf{k}}{m^*}; \quad \frac{d\mathbf{k}}{dt} = \frac{q\mathbf{E}(\mathbf{r})}{\hbar} \tag{63}$$

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

$$x(t) = x(0) + v_x(0)\,t + \frac{qE_x^0 t^2}{2m^*} \tag{64}$$

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 20**.**

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 21 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 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.

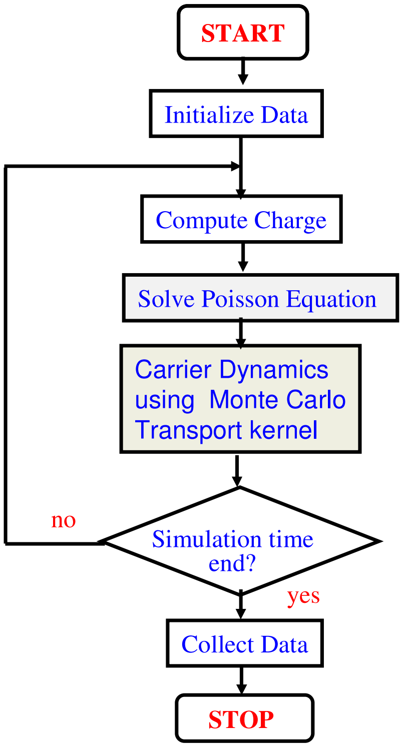


Figure 20. Flow-chart of a typical particle based device simulation.

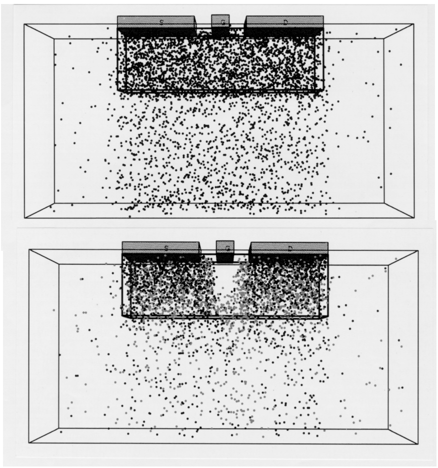


Figure 21. Example of the particle distribution in a MESFET structure simulated in 3D using an EMC approach. The upper plot is the device with zero gate voltage applied, while the lower is with a negative gate voltage applied, close to pinch-off.

3.1 Calculation of the current

The device output current can be determined using two different yet consistent methods. *First*, by keeping track of the charges entering and exiting each terminal/contact, the net number of charges over a period of the simulation can be used to calculate the terminal current. The net charge crossing a terminal boundary is determined by

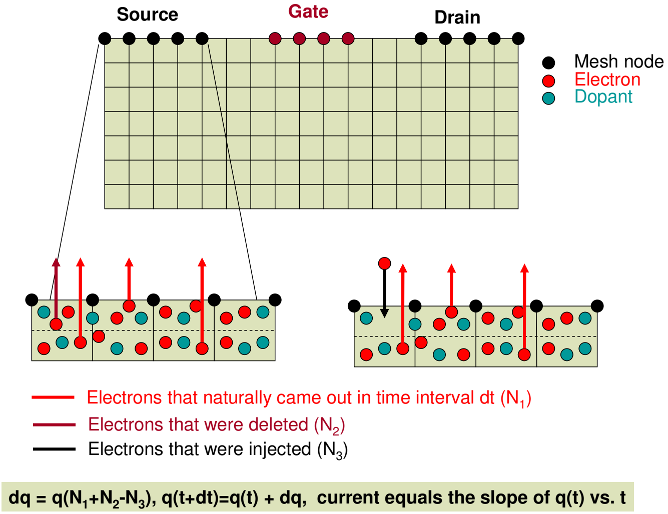
$$Q(t)=e\left(n_{abs}(t)-n_{injec}(t)\right)+\varepsilon\int E_y(x,t)\,dy, \tag{65}$$

where n_{abs} is the number of particles that are absorbed by the contact (exit), n_{injec} is the number of particles that have been injected at the contact, E_y is the vertical field at the contact. The second term in Eq. (65) on the right-hand-side is used to account for the displacement current due to the changing field at the contact. Eq.

(65) assumes the contact is at the top of the device and that the fields in the x and z direction are negligible. The charge e in Eq. (65) should be multiplied by the particle charge if it is not unity. The slope of $Q(t)$ versus time gives a measure of the terminal current. In steady state, the current can be found by

$$I = \frac{dQ(t)}{dt} = \frac{e(n_{net})}{\Delta t}, \tag{66}$$

where n_{net} is the net number of particles exiting the contact over a fixed period of time Δt. The method is quite noisy, due to the discrete nature of the electrons. An example of calculation of the current and keeping the ohmic contacts charge neutral is given Figure 22.



Source
Gate
Drain
● Mesh node
● Electron
● Dopant
Electrons that naturally came out in time interval dt (N_1)
Electrons that were deleted (N_2)
Electrons that were injected (N_3)
dq = q(N_1+N_2-N_3), q(t+dt)=q(t) + dq, current equals the slope of q(t) vs. t

Figure 22. Keeping charge neutrality at the ohmic contacts and contributions of various terms to the current.

In a *second* method, the sum of the electron velocities in a portion of the channel region of the device is used to calculate the current. The electron current density through a cross-section of the device is given by

$$J = env_d, \tag{67}$$

where v_d is the average electron drift velocity and n is the carrier concentration. If there are a total of N particles in a differential volume, $dV = dL \cdot dA$, the current found by integrating Eq. (67) over the cross-sectional area, dA, is

$$I = \frac{eNv_d}{dL}, \quad \text{or} \quad I = \frac{e}{dL}\sum_{i=1}^{N} v_x(i), \tag{68}$$

where $v_x(i)$ is the velocity along the channel of the i^{th} electron. The device is divided into several sections along the x-axis, and the number of electrons and their corresponding velocity is added for each section after each free-flight. The total x-velocity in each section is then averaged over several timesteps to determine the current for that section. Total device current can be determined from the average of several sections, which gives a much smoother result compared to counting terminal charges. By breaking the device into sections, individual currents can be compared to verify that there is conservation of particles (constant current) throughout the device. In addition, sections near the source and drain regions may have a high y-component in their 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 can be observed to ensure the proper physical characteristics. The two methods for the calculation of the current are illustrated in Figure 23 on the example of a 50 nm channel length MOSFET device.

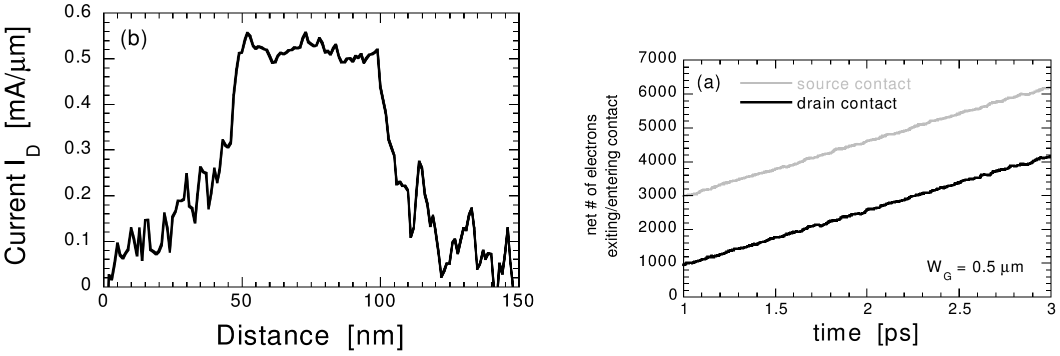


Figure 23. (Left panel) Net charge entering/exiting the source/drain contact. (Right panel) Average current along the channel. The gate-length of the device being modeled equals 50 nm. We use $V_G = 1.4$ V and $V_D = 1$ V in these simulations.

Extrapolating the slope of the curve shown in

Figure 23 (left panel), that represents the cumulative electron charge that enters/exits the source/drain contact, leads to source/drain current of 0.5205/0.5193 mA/μm. When compared with the results shown in

Figure 23 (right panel), it is evident that both the current measurement techniques discussed in this section give current values with relative error less than 2 %.

3.2 Ohmic Contacts

Another issue that has to be addressed in particle-based simulations is the real space boundary conditions for the particle part of the simulation. Reflecting boundary conditions are usually imposed at the artificial boundaries. As far as the ohmic contacts are concerned, they require more careful consideration because electrons crossing the source and drain contact regions contribute to the corresponding terminal current. Commonly employed models for the contacts include [31]:

- 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 [32].
- 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 is a very good model most commonly employed in actual device simulations.
- 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 [31], but at the expense of increased computational time due to the extra electrons simulated. It is an excellent model employed in few 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].

3.3 Time Step

As in the case of solving the Drift-Diffusion, Hydrodynamic or full Maxwell's equations, for a 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 stability. For example, as discussed in the context of solving Drift-Diffusion simulations, the time step Δt must be related to the plasma frequency

$$\omega_p = \sqrt{\frac{e^2 n}{\varepsilon_s m^*}} \text{ ,} \tag{69}$$

where n is the carrier density. From the viewpoint of the stability criterion, Δt must be much smaller than the inverse plasma frequency. The highest carrier density specified in the device model is used to estimate Δt. If the material is a multi-valley semiconductor, the smallest effective mass to be experienced by the carriers must be used in Eq. (69) as well. In the case of GaAs, with the doping of 5×10^{17} cm^{-3}, $\omega_p \cong 5\times10^{13}$; hence, Δt must be smaller than 0.02 ps.

The mesh size for the spatial resolution of the potential is dictated by the 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, given as

$$\lambda_D = \sqrt{\frac{\varepsilon_s k_B T}{e^2 n}} \text{ .} \tag{70}$$

The highest carrier density specified in the model should be used to estimate λ_D from the stability criterion. The mesh size must be chosen to be smaller than the value given by Eq. (70). In the case of GaAs, with the doping density of 5×10^{17} cm^{-3}, $\lambda_D \cong 6$ nm.

Based on the discussion above, the time step (Δt), and the mesh size (Δx, Δy, and/or Δz) can be specified separately. However, the Δt chosen must be checked again by calculating the distance l_{max}, defined as

$$l_{\mathrm{max}} = \mathbf{v}_{\mathrm{max}} \times \Delta t \text{ ,} \tag{71}$$

where $\mathbf{v}_{\mathrm{max}}$ 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). Therefore, the distance l_{max} is regarded as the maximum distance the carriers can propagate during Δt. The time step chosen must be small enough so that l_{max} is smaller than the spatial mesh size chosen using Eq. (71). This is because large Δt chosen may cause substantial change in the charge distribution, while the field distribution in the simulation is only updated every Δt.

3.4 Particle-mesh (PM) coupling

As mentioned earlier, the position of charge as described by the EMC algorithm is continuous, whereas Poisson's equation is solved on a mesh, hence the charge associated with the individual particles must be mapped onto the field mesh in some fashion. 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]. 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 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 particle-mesh coupling sequence is presented in Figure 24.

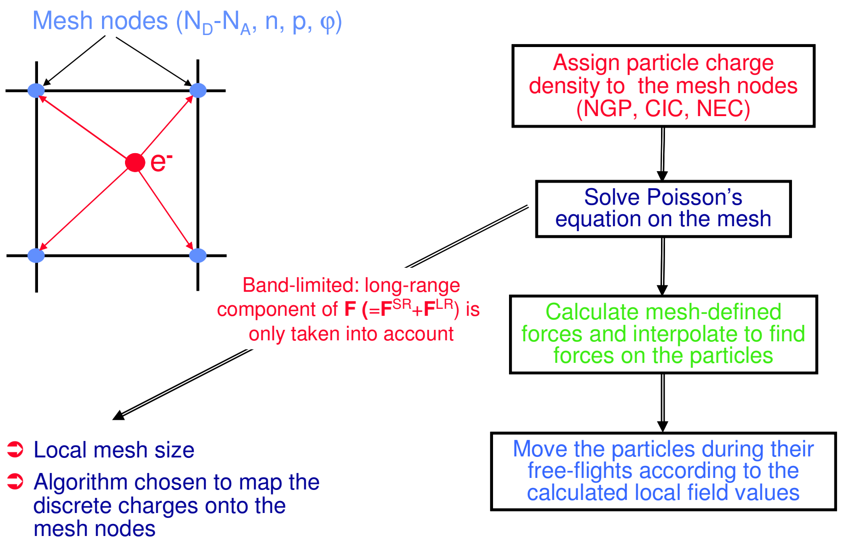

Mesh nodes (N_D-N_A, n, p, φ)
e^-
Assign particle charge density to the mesh nodes (NGP, CIC, NEC)
Solve Poisson's equation on the mesh
Band-limited: long-range component of **F** (=**F**^{SR}+**F**^{LR}) is only taken into account
Calculate mesh-defined forces and interpolate to find forces on the particles
Move the particles during their free-flights according to the calculated local field values
➲ Local mesh size
➲ Algorithm chosen to map the discrete charges onto the mesh nodes

Figure 24. Particle-mesh coupling sequence.

To better understand the NGP and the CIC scheme, consider a tensor product mesh with mesh lines $x_i, i=1,\ldots,N_x$ and $y_j, j=1,\ldots,N_y$. If the mesh is uniformly spaced in each axis direction, then $(x_{l+1}-x_l)=(x_{l+2}-x_{l+1})$. The permittivities are considered constant within each mesh element and are denoted by ε_{kl}, $k=1,\ldots,N_x-1$ and $l=1,\ldots,N_y-1$. Define centered finite-differences of the potential ψ in the *x*- and *y*-axis at the midpoints of element edges as follows:

$$\begin{cases} \Delta^x_{k+\frac{1}{2},l} = -\dfrac{\psi_{k+1,l}-\psi_{k,l}}{x_{k+1}-x_k} \ , \\ \Delta^y_{k,l+\frac{1}{2}} = -\dfrac{\psi_{k,l+1}-\psi_{k,l}}{y_{l+1}-y_l} \ , \end{cases} \tag{72}$$

where the minus sign is included for convenience because the electric field is negative of the gradient of the potential. Consider now a point charge in 2-D located at (x, y) within an element $\langle i, j\rangle$. If the restrictions for the permittivity (P) and the tensor-product meshes with uniform spacing in each direction (M) apply, the standard NGP/CIC schemes in two dimensions can be summarized by the following four steps:

Charge assignment to the mesh: The portion of the charge ρ_L assigned to the element nodes (*k*,*l*) is $w_{kl}\rho_L$, *k*=*i*, *i*+1 and *l*=*j*, *j*+1, where w_{kl} are the four charge weights which sum to unity by charge conservation. For the NGP scheme, the node closest to (*x*,*y*) receives a weight w_{kl} =1, with the remaining three weights set to zero. For the CIC scheme, the weights are $w_{ij}=w_x w_y$, $w_{i+1,j}=(1-w_x)w_y$, $w_{i,j+1}=w_x(1-w_y)$, and $w_{i+1,j+1}=(1-w_x)(1-w_y)$, $w_x=(x_{i+1}-x)/(x_{i+1}-x_i)$ and $w_y=(y_{j+1}-y)/(y_{j+1}-y_j)$.

Solve the Poisson equation: The Poisson equation is solved by some of the numerical techniques discussed in Ref. [34].

Compute forces on the mesh: The electric field at mesh nodes (*k*,*l*) is computed as: $E^x_{kl}=\left(\Delta^x_{k-\frac{1}{2},l}+\Delta^x_{k+\frac{1}{2},l}\right)/2$ and $E^y_{kl}=\left(\Delta^y_{k,l-\frac{1}{2}}+\Delta^y_{k,l+\frac{1}{2}}\right)/2$, for k = *i*, *i*+1 and l=*j*, *j*+1.

Interpolate to find forces on the charge: Interpolate the field to position (*x*,*y*) according to $E^x=\sum_{kl} w_{kl}E^x_{kl}$ and $E^y=\sum_{kl} w_{kl}E^y_{kl}$, where *k* = *i*, *i*+1, *l* =*j*, *j*+1 and the w_{ij} are the NGP or CIC weights from step 1.

The requirements (P) and (M) severely limit the scope of devices that may be considered in device simulations using the NGP and the CIC schemes. Laux [35] 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 new steps of the pure NEC PM scheme are:

(1') *Charge assignment to the mesh:* Divide the line charge ρ_L equally to the four mesh points of the element $\langle i, j \rangle$.

(2') Solve the Poisson equation.

(3') *Compute forces on the mesh:* Calculate the fields $\Delta^x_{i+\frac{1}{2},l}$, l=*j*, *j*+1, and $\Delta^y_{k,j+\frac{1}{2}}$, k=*i*, *i*+1.

(4') *Interpolate to find force on the charge:* Interpolate the field according to the following $E^x = \left(\Delta^x_{i+\frac{1}{2},j} + \Delta^x_{i+\frac{1}{2},j+1}\right)/2$ and $E^y = \left(\Delta^x_{i,j+\frac{1}{2}} + \Delta^x_{i+1,j+\frac{1}{2}}\right)/2$.

The NEC designation derives from the appearance, in step (1') 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.

3.5 Higher order effects

Multi-particle effects relate to the interaction between particles in the system, which is a nonlinear effect when viewed in the context of the BTE, due to the dependence of such effects on the single particle distribution function itself. Most algorithms developed to deal with such effects essentially linearize the BTE by using the previous value of the distribution function to determine the time evolution of a particle over the successive time-step. Multi-carrier effects may range from simple consideration of the Pauli exclusion principle (which depends on the exact occupancy of states in the system), to single particle and collective excitations in the system. Inclusion of carrier-carrier interactions in Monte Carlo simulation has been an active area of research for quite some time and is briefly discussed below. Another carrier-carrier effect, that is of considerable importance when estimating leakage currents in MOSFET devices, is impact ionization, which is a pure generation process involving three particles (two electrons and a hole or two holes and an electron). The latter is also discussed below.

3.5.1 Pauli exclusion principle

The Pauli exclusion principle requires that the bare scattering rate be modified by a factor $1 - f_m(\mathbf{k}')$ in the collision integral of the BTE, where $f_m(\mathbf{k}')$ is the one-particle distribution function for the state $\mathbf{k}'$ in band (subband) *m* after scattering. Since the net scattering rate including the Pauli exclusion principle is always less than the bare scattering rate, a self-scattering rejection technique may be used in the Monte Carlo simulation as proposed by Bosi and Jacoboni [36] for one particle simulation and extended by Lugli and Ferry [37] for EMC. In the self-scattering rejection algorithm, an additional random number *r* is generated (between 0 and 1), and this number is compared to $f_m(\mathbf{k}')$, the occupancy of the final state (which is also between 0 and 1 when properly normalized for the numerical **k**-space discretization). If *r* is greater than $f_m(\mathbf{k}')$, the scattering is accepted and the particle's momentum and energy are changed. If this condition is not satisfied, the scattering is rejected, and the process is treated as a self-scattering event with no change of energy or momentum after scattering. Through this algorithm, it is clear that no scattering occurs if the final state is completely full.

3.5.2 Carrier-carrier interactions

Carrier-carrier interactions, apart from degeneracy effects, may be treated as a scattering process within the Monte Carlo algorithm on the same footing as other mechanisms. In the simplest case of bulk electrons in a single parabolic conduction band, the process may be treated as a binary collision where the scattering rate for a particle of wavevector $\mathbf{k}_0$ due to all the other particles in the ensemble is given by [38]

$$\Gamma_{ee}(\mathbf{k}_0) = \frac{nm_n e^4}{4\pi\hbar^3 \varepsilon^2 \beta^2} \int d\mathbf{k} f(\mathbf{k}) \frac{|\mathbf{k}-\mathbf{k}_0|}{\left(|\mathbf{k}-\mathbf{k}_0|^2 + \beta^2\right)}, \tag{73}$$

where $f(\mathbf{k})$ is the one-particle distribution function (normalized to unity), ε is the permittivity, n is the electron density, and β is the screening constant. In deriving Eq. (73), one assumes that the two particles interact through a statically screened Coulomb interaction, which ignores the energy exchange between particles in the screening which in itself is a dynamic, frequency-dependent effect. Similar forms have been derived for electrons in 2D [39,40] and 1D [41], where carrier-carrier scattering leads to inter-subband as well as intra-subband transitions. Since the scattering rate in Eq. (73) depends on the distribution function of all the other particles in the system, this process represents a nonlinear term as discussed earlier. One method is to tabulate $f(\mathbf{k})$ on a discrete grid, as is done for the Pauli principle, and then numerically integrate Eq. (73) at each time step. An alternate method is to use a self-scattering rejection technique [42], where the integrand excluding $f(\mathbf{k})$ is replaced by its maximum value and taken outside the integral over $\mathbf{k}$. The integral over $f(\mathbf{k})$ is just unity, giving an analytic form used to generate the free flight. Then, the self-scattering rejection technique is used when the final state is chosen to correct for the exact scattering rate compared to this artificial maximum rate, similar to the algorithm used for the Pauli principle.

The treatment of intercarrier interactions as binary collisions above neglects scattering by collective excitations such as plasmons or coupled plasmon-phonon modes. These effects may have a strong influence on carrier relaxation, particularly at high carrier density. One approach is to make a separation of the collective and single particle spectrum of the interacting many-body Hamiltonian, and treat them separately, i.e. as binary collisions for the single particle excitations, and as electron-plasmon scattering for the collective modes [43]. Another approach is to calculate the dielectric response within the random phase approximation, and associate the damping given by the imaginary part of the inverse dielectric function with the electron lifetime [44].

A semiclassical approach to carrier-carrier interaction, which is fully compatible with the Monte Carlo algorithm, is the use of Molecular Dynamics [45], in which carrier-carrier interaction is treated continuously in real space during the free-flight phase through the Coulomb force of all the particles. A very small time step is required when using Molecular Dynamics to account for the dynamic distribution of the system. A time step on the order of 0.5 *fs* is often sufficiently small for this purpose. The small time step assures that the forces acting on the particles during the time of flight are essentially constant, that is $f(t) \cong f(t+\Delta t)$, where $f(t)$ is the single particle distribution function.

Using Newtonian kinematics, we can write the real space trajectories of each particle as

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + \mathbf{v}\Delta t + \frac{1}{2}\frac{\mathbf{F}(t)}{m}\Delta t^2, \tag{74}$$

and

$$\mathbf{v}(t+\Delta t) = \mathbf{v}(t) + \frac{\mathbf{F}(t)}{m}\Delta t. \tag{75}$$

Here, $\mathbf{F}(t)$ is the force arising from the applied field as well as that of the Coulomb interactions. We can write $\mathbf{F}(t)$ as

$$\mathbf{F}(t) = q\left[\mathbf{E} - \sum_i \nabla\varphi(\mathbf{r}_i(t))\right], \tag{76}$$

where $q\mathbf{E}$ is the force due to the applied field and the summation is the interactive force due to all particles separated by distance $\mathbf{r}_i$, with $\varphi(\mathbf{r}_i)$ the electrostatic potential. As in Monte Carlo simulation, one has to simulate a finite number of particles due to practical computational limitations on execution time. In real space, this finite number of particles corresponds to a particular simulation volume given a certain density of carriers, $V = N/n$, where n is the density. Since the carriers can move in and out of this volume, and since the Coulomb interaction is a long-range force, one must account for the region outside V by periodically replicating the simulated system. The contributions due to the periodic replication of the particles inside V in cells outside has a closed form solution in the form of an Ewald sum [46], which gives a linear as well as $1/r^2$ contribution to the force. The equation for the total force in the Molecular Dynamics technique then becomes

$$\mathbf{F} = \frac{-e^2}{4\pi\varepsilon}\sum_i^N\left(\frac{1}{\mathbf{r}_i^{\,2}}\mathbf{a}_i + \frac{2\pi}{3V}\mathbf{r}_i\right). \tag{77}$$

The above equation is easily incorporated in the standard Monte Carlo simulation discussed up to this point. At every time step the forces on each particle due to all the other particles in the system are calculated from Eq. (77). From the forces, an interactive electric field is obtained which is added to the external electric field of the system to couple the Molecular Dynamics to the Monte Carlo.

The inclusion of the carrier-carrier interactions in the context of particle-based device simulations is discussed in Ref. [47]. The main difficulty in treating this interaction term in device simulations arises from the fact that the long-range portion of the carrier-carrier interaction is included via the numerical solution of the quasi-static Poisson equation. Under these circumstances, special care has to be taken when incorporating the short-range portion of this interaction term to prevent double counting of the force.

3.5.3 Band to Band Impact Ionization

Another carrier-carrier scattering process is that of impact ionization, in which an energetic electron (or hole) has sufficient kinetic energy to create an electron-hole pair. Impact ionization therefore leads to the process of carrier multiplication. This process is critical for example in the avalanche breakdown of semiconductor junctions, and is a detrimental effect in short channel MOS devices in terms of excess substrate current and decreased reliability.

The ionization rate of valence electrons by energetic conduction band electrons is usually described by Fermi's rule in which a screened Coulomb interaction is assumed between the two particles, where screening is described by an appropriate dielectric function such as that proposed by Levine and Louie [48]. In general, the impact ionization rate should be a function of the wavevector of the incident electron, hence of the direction of an electric field in the crystal, although there is still some debate as to the experimental and theoretical evidence. More simply, the energy dependent rate (averaged over all wavevectors on a constant energy shell) may be expressed analytically in the power law form

$$\Gamma_{ii}(E) = P\left[E - E_{th}\right]^a , \tag{78}$$

where E_{th} is the threshold energy for the process to occur, which is determined by momentum and energy conservation considerations, but minimally is the bandgap of the material itself. P and a are parameters which may be fit to more sophisticated models. The Keldysh formula [49] is derived by expanding the matrix element for scattering close to threshold, which gives a=2, and the constant $P=C/E_{th}^{2}$, with C=1.19×10^{14}/s and assuming a parabolic band approximation,

$$E_{th} = \frac{3 - 2m_v/m_c}{1 - m_v/m_c}E_g , \tag{79}$$

where m_v and m_c are the effective masses of the valence and conduction band respectively, and E_g is the bandgap. More complete full-bandstructure calculations of the impact ionization rate have been reported for Si [50,51], GaAs [51,52] and wide bandgap materials [53], which are fairly well fit using using power law model. Within the ensemble Monte Carlo method, the scattering rate given by Eq. (78) is used to generate the free flight time. The state after scattering of the initial electron plus the additional electron and hole must satisfy both energy and momentum conservation within the Fermi rule model, which is somewhat complicated unless simple parabolic band approximations are made.

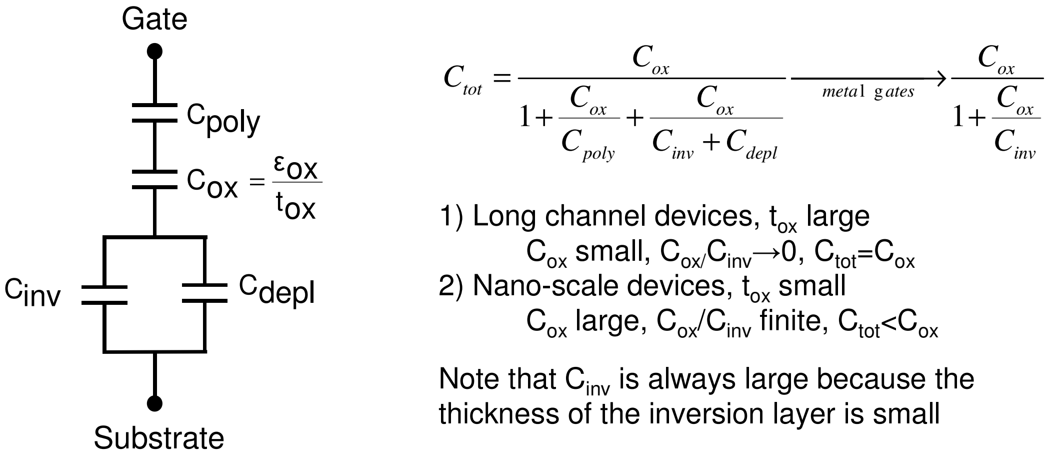
4. Quantum Corrections

Quantum mechanical effects are known to dominate the operation of devices such as resonant tunneling diodes [54], quantum cascade lasers [55], etc. Tunneling through the gate oxide [56], source to drain tunneling and space-quantization effects are expected to be important in nano-scale MOSFETs, and require the solution of the one-dimensional (1D) Schrödinger-Poisson problem. Solutions of the two-dimensional (2D) Schrödinger-

Poisson problem are needed, for example, for describing the channel charge in narrow-width MOSFETs and alternative device technologues such as FinFETs.

From a device modeling point of view, even the solution of the 1D Schrödinger-Poisson problem along slices of the device is difficult in terms of both complexity and computational cost. Because of this, it is common practice in industry to use analytical and macroscopic (in the sense of sticking to the classical transport framework by adding correction terms to account for the quantum-mechanical effects) models that have provided some practical solutions. However, there are a number of problems associated with these approaches and all of them are directly related to the non-stationary nature of carrier transport (velocity overshoot) in deep submicrometer devices. Hence, more sophisticated models are needed that are able to capture the appropriate transport physics of the processes occurring in the smallest device sizes.

Note that successful scaling of MOSFETs towards shorter channel lengths requires thinner gate oxides and higher doping levels to achieve high drive currents and minimized short-channel effects [57,58]. For these nanometer devices it was demonstrated a long time ago that, as the oxide thickness is scaled to 10 nm and below, the total gate capacitance is smaller than the oxide capacitance due to the comparable values of the oxide and the inversion layer capacitances (that arise due to the finite average displacement of the inversion charge from the semiconductor/oxide interface), as illustrated in Figure 25. As a consequence, the device transconductance is degraded relative to the expectations of the scaling theory [59].



$$C_{tot} = \frac{C_{ox}}{1 + \frac{C_{ox}}{C_{poly}} + \frac{C_{ox}}{C_{inv} + C_{depl}}} \xrightarrow{\text{metal gates}} \frac{C_{ox}}{1 + \frac{C_{ox}}{C_{inv}}}$$

1) Long channel devices, t_{ox} large
C_{ox} small, $C_{ox}/C_{inv} \to 0$, $C_{tot} = C_{ox}$
2) Nano-scale devices, t_{ox} small
C_{ox} large, C_{ox}/C_{inv} finite, $C_{tot} < C_{ox}$

Note that C_{inv} is always large because the thickness of the inversion layer is small

Figure 25. Equivalent circuit showing the various contributions to the total gate capacitance in a MOS capacitor. The effect of interface traps has been omitted in the present analysis. If included, it would lead to an additional capacitance component in parallel to the inversion layer and depletion layer capacitances.

The inversion layer capacitance was also identified as being the main cause of the second-order thickness dependence of a MOSFET's *I-V* characteristics [60]. The quantum mechanical inversion layer thickness was estimated experimentally by Hartstein and Albert [61]. The high levels of substrate doping, needed in nano-devices to prevent the punch-through effect, and which enhance the quasi-two-dimensional (Q2D) nature of the carrier transport in the inversion layer, were found responsible for the increased threshold voltage and decreased channel mobility. A simple analytical model that accounts for this effect was proposed by van Dort and co-workers [62,63]. Vasileska and Ferry [64] confirmed these findings by investigating the doping dependence of the threshold voltage in MOS capacitors. The two physical origins of the inversion layer capacitance, due to the finite density of states and due to the quantum mechanical spread of the inversion, were demonstrated experimentally by Takagi and Toriumi [65]. A computationally efficient three-subband model, that predicts both the quantum-mechanical effects in the electron inversion layers and the electron distribution within the inversion layer, was proposed and implemented into the PISCES simulator [66]. The influence of the image and many-body exchange-correlation effects on the inversion layer and the total gate capacitance was studied by Vasileska *et al.* [67]. It was also pointed out that the depletion of the poly-silicon gates considerably affects the magnitude of the total gate capacitance [68].

The above examples outline the advances during the two decades of research on the influence of quantum-effects on the operation on nano-devices. The conclusion is that any state-of-the-art device simulator must take into consideration the quantum-mechanical nature of the carrier transport and the poly-depletion effects to correctly predict the device off- and on-state behavior. As noted by many of these authors, to account for the quantum-mechanical effects, one in principle has to solve the 2D/3D Schrödinger-Poisson problem in conjunction with an appropriate transport kernel. When non-stationary transport and velocity overshoot are pronounced, one has to solve the Schrödinger-Poisson problem with the Boltzmann transport equation e.g., using Ensemble Monte Carlo (EMC) techniques. Because of the importance of the subject matter, in this Section we discuss:

- Effective potential approaches in conjunction with particle-based device simulators.
- Tunneling approaches, oxide charging and inclusion of gate leakage in conjunction with particle-based device simulation scheme.

4.1 Effective Potential Approach

Analogous to the smoothed potential representations discussed for the QHD model in Ref. [69], it is also desirable to define a smooth quantum potential for use in particle based simulation. Ferry [70] suggested an 'effective potential' that is derived from a Gaussian wave packet description of particle motion, where the extent of the wave packet is defined from the range of wavevectors established by the thermalized distribution function (characterized by an electron temperature). The effective potential seen by electrons is given by the convolution of this wave packet with the physical potential

$$V_{eff}(x)=\frac{1}{\sqrt{2\pi}a_0}\int_{-\infty}^{\infty}V(x')\exp\left(-\frac{(x-x')^2}{2a_0^2}\right)dx', \qquad (80)$$

where $V(x')$ is the actual potential, and a_0 is the spatial spread of the wavepacket. The effective potential accounts for the 'size of the electron' and its associated wavepacket, which feels the presence of barriers, etc., at a distance. From this *Ansatz*, the actual particle is treated as point-like in the presence of the effective potential associated with its wave-like nature, leading back to a classical particle simulation scheme. Representative simulation results for assymmetric MOSFET structures (focused ion beam MOSFET (FIBMOS)), which utilize this effective potential approach, are shown in Figure 26 [71,72]. The inclusion of quantum-mechanical space-quantization effects leads to threshold voltage shift of about 150 mV and drain current reduction between 30 and 40 %, depending upon the gate bias. These results are in agreement with experimental findings, thus demonstrating the applicability of such scheme in accurately representing the quantum-mechanical effects in the device channel region.

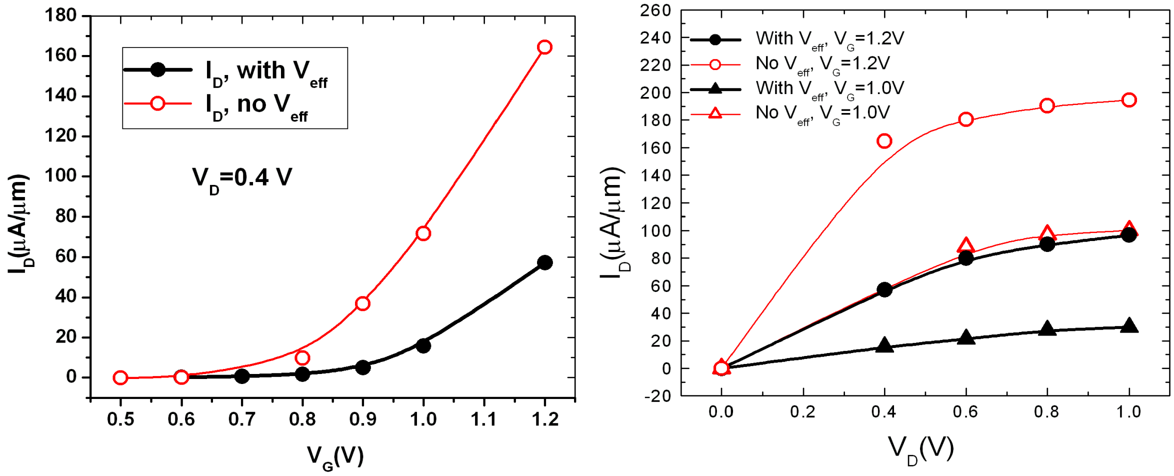


Figure 26. Left panel: transfer characteristics of a FIBMOS device. Right panel: device output characteristics.

4.2 Effective Potential from the Wigner-Boltzmann Equation

The basic idea of the thermodynamic approach to effective quantum potentials is that the resulting semiclassical transport picture should yield the correct thermalized equilibrium quantum state. Using quantum potentials, one generally replaces the quantum Liouville equation

$$[H,\rho]=i\hbar\frac{\partial\rho}{\partial t} \qquad (81)$$

for the density matrix $\rho(x,y)$ by the classical Liouville equation

$$\partial_t f + \tfrac{\hbar}{2m^*} k \cdot \nabla_x f - \tfrac{1}{\hbar} \nabla_x V \cdot \nabla_k f = 0 \,, \tag{82}$$

for the classical density function $f(x,k)$. Here, the relation between the density matrix and the density function (Wigner function) f is given by the Weyl quantization

$$f(x,k) = W[\rho] = \int \rho(x + y/2, x - y/2) \exp(ik \cdot y) dy \,. \tag{83}$$

The thermal equilibrium density matrix in the quantum mechanical setting is given by $\rho^{eq} = e^{-\beta H}$, where $\beta = 1/k_B T$ is the inverse energy, and the exponential is understood as a matrix exponential, i.e. $\rho^{eq}(x,y) = \sum_\lambda \psi_\lambda(x) \exp(-\beta\lambda) \psi_\lambda(y)^*$ holds, with $\{\psi_\lambda\}$ the orthonormal eigensystem of the Hamiltonian H. In the semiclassical transport picture, on the other hand, the thermodynamic equilibrium density function f_{eq} is given by the Maxwellian $f_{eq}(x,k) = \exp\left(-\frac{\beta\hbar^2 |k|^2}{2m^*} - \beta V\right)$. Consequently, to obtain the quantum mechanically correct equilibrium states in the semiclassical Liouville equation with the effective quantum potential V^Q, we set

$$f_{eq}(x,k) = \exp\left(-\tfrac{\beta\hbar^2 |k|^2}{2m^*} - \beta V^Q\right) = W[\rho^{eq}] = \int e^{-\beta H} \rho(x + y/2, x - y/2) \exp(ik \cdot y) dy \,. \tag{84}$$

This basic concept was originally introduced by Feynman and Kleinert [73]. Different forms of the effective quantum potential arise from different approaches to approximate the matrix exponential $e^{-\beta H}$.

In the approach presented below, we represent $e^{-\beta H}$ as the Green's function of the semigroup generated by the exponential. Introducing an artificial dimensionless parameter γ and defining $\rho(x,y,\gamma) = \sum_\lambda \psi_\lambda(x) \exp(-\gamma\beta\lambda) \psi_\lambda(y)^*$, we obtain a heat equation for ρ by differentiating ρ w.r.t. γ and using the eigenfunction property of the wave functions ψ_λ. This heat equation is referred to as the Bloch equation

$$\partial_\gamma \rho = -\frac{\beta}{2}\left(H \cdot \rho + \rho \cdot H\right), \quad \rho(x,y,\gamma = 0) = \delta(x - y) \,, \tag{85}$$

and $\rho^{eq}(x,y)$ is given by $\rho(x,y,\gamma = 1)$. Under the Weyl quantization this becomes, with the usual Hamiltonian $H = -\frac{\hbar^2}{2m^*} \Delta_x + V$ and defining the effective energy E by $f = W[\rho] = e^{-\beta E}$,

$$\begin{aligned}
&\partial_\gamma E = \frac{\beta\hbar^2}{8m^*}\left(\Delta_x E - \beta |\nabla_x E|^2\right) + \frac{\hbar^2 |k|^2}{2m^*} + \\
&\frac{1}{2(2\pi)^3} \sum_{\nu=\pm 1} \int V(x + \nu y/2) \exp\left[\beta E(x,k,\gamma) - \beta E(x,q,\gamma) + iy(k - q)\right] dq dy \,, \\
&E(x,k,\gamma = 0) = 0.
\end{aligned} \tag{86}$$

The effective quantum potential is in this formulation given by $E(x,k,\gamma = 1) = V^Q + \frac{\hbar^2 |k|^2}{2m^*}$. The logarithmic Bloch equation is now solved 'asymptotically', using the *Born approximation*, i.e. by iteratively inverting the highest order differential operator (the Laplacian). This involves successive solution of a heat equation for which the Green's function is well known, giving (see Ref. [74] for the details)

$$V^Q(x,k) = \frac{1}{(2\pi)^3} \int \frac{2m^*}{\beta\hbar^2 k \cdot \xi} \sinh\left(\frac{\beta\hbar^2 k \cdot \xi}{2m^*}\right) \exp\left(-\frac{\beta\hbar^2}{8m^*} |\xi|^2\right) V(y) e^{i\xi \cdot (x - y)} dy d\xi \,. \tag{87}$$

Note that the effective quantum potential V^Q now depends on the wave vector k. For electrons at rest, i.e. for $k = 0$, the effective potential V^Q reduces to the Gaussian smoothing (see Ref. [75]). Also note that there are no fitting parameters in this approach, i.e. the size of the wavepacket is determined by the particle's energy [76].

The potential $V(y)$ that appears in the integral of Eq. (87) can be represented as a sum of two potentials: the barrier potential $V_B(x)$, which takes into account the discontinuity at the Si/SiO$_2$ interface due to the difference in the semiconductor and the oxide affinities, and the Hartree potential $V_H(x)$ that results from the solution of the Poisson equation. Note that the barrier potential is 1D and independent of time and needs to be computed only once in the initialization stage of the code. On the other hand, the Hartree potential is 2D and time-dependent as it describes the evolution of charge from quasi-equilibrium to a non-equilibrium state. Since the evaluation of the effective Hartree potential, as given by Eq. (87), is very time consuming and CPU intensive, approximate solution methods have been pursued to resolve this term within a certain level of error tolerance.

We recall from the above discussion that the barrier potential is just a step-function. Under these circumstances $e\nabla_x V_B(x) = B(1,0,0)^T \delta(x_1)$, where B is the barrier height (on the order of 3.2 eV) and x_1 is a vector perpendicular to the interface. We actually need only the gradient of the potential so that, using the pseudo-differential operators we compute

$$\nabla_x V_B^Q(x,p) = \exp\left[\frac{\beta\hbar^2 |\nabla_x|^2}{8m^*}\right] \frac{2m^* \sin\left(\frac{\beta\hbar p\cdot\nabla_x}{2m^*}\right)}{\beta\hbar p\cdot\nabla_x} \nabla_x V_B(x) \tag{88}$$

This gives

$$e\nabla_x V_B^Q(x,p) = \frac{B}{2\pi}(1,0,0)^T \int \exp\left[-\beta\frac{\hbar^2|\xi_1|^2}{8m^*}\right] \frac{2m^* \sinh\left(\frac{\beta\hbar p_1\cdot\xi_1}{2m^*}\right)}{\beta\hbar p_1\cdot\xi_1} e^{i\xi_1\cdot x_1} d\xi_1 \tag{89}$$

Note that V_B^Q is only a function of (x_1, p_1), i.e. it remains to be strictly one-dimensional, where x_1 and p_1 are the position and the momentum vector perpendicular to the interface. This, when combined with the fact that we have to calculate this integral only once, is a reason why we have decided to tabulate the result given by Eq. (89) on a mesh.

The Hartree potential, as computed by solving the d-dimensional Poisson equation, depends in general upon d particle coordinates. For example, on a rectangular mesh the 2D Hartree potential is given by $V_H(x_1,x_2,t)$, and one has to evaluate $V_H^Q(x_1,x_2,p_1,p_2,t)$ using Eq. (87) N times each time step for all particles position and momenta: x^n, p^n, $n = 1, \ldots, N$ (where N is the number of electrons, which is large). This is, of course, an impossible task to be accomplished in finite time on present state-of-the-art computers. We, therefore, suggest the following scheme. According to (87), we evaluate the quantum potential by multiplying the Hartree potential by a function of $\hbar\nabla_x$, or by multiplying the Fourier transform of the Hartree potential by a function of $\hbar\xi$. We factor the expression in Eq. (87) into

$$V_H^Q(x,k) = \frac{2im^*}{\beta\hbar^2 k\cdot\nabla_x}\sinh\left(\frac{\beta\hbar^2 k\cdot\nabla_x}{2im^*}\right)\exp\left(\frac{\beta\hbar^2}{8m^*}|\nabla_x|^2\right)V_H(x) = \frac{2im^*}{\beta\hbar^2 k\cdot\nabla_x}\sinh\left(\frac{\beta\hbar^2 k\cdot\nabla_x}{2im^*}\right)V_H^0(x) \quad , \tag{90}$$

with

$$V_H^0(x) = \exp\left(\frac{\beta\hbar^2}{8m^*}|\nabla_x|^2\right)V_H(x) \quad . \tag{91}$$

The evaluation of the potential $V_H^0(x)$, which is a version of the Gaussian smoothed potential due to Ferry [75] and is computationally inexpensive since it does not depend on the wavevector k. On the other hand, because of the Gaussian smoothing, $V_H^0(x)$ will be a smooth function of position, even if the Hartree potential $V_H(x)$ is computed via the Poisson equation where the electron density is given by a particle discretization. Therefore, the Fourier transform of the potential $V_H^0(x)$ will decay rapidly as a function of ξ, and it is admissible to use a Taylor expansion for small values of $\hbar\xi$ in the rest of the operator. This gives

$$\frac{2im^*}{\beta\hbar^2 k\cdot\nabla_x}\sinh\left(\frac{\beta\hbar^2 k\cdot\nabla_x}{2im^*}\right)\approx 1-\frac{\beta^2\hbar^4\left(k\cdot\nabla_x\right)^2}{24\left(m^*\right)^2}, \tag{92}$$

or

$$\partial_{x_r} V_H^Q(x^n,p^n)=\partial_{x_r} V_H^0(x^n)-\frac{\beta^2\hbar^2}{24m^{*2}}\sum_{j,k=1}^{2} p_j^n p_k^n \partial x_j \partial x_k \partial x_r V_H^0(x^n), \quad n=1,\ldots,N \tag{93}$$

for all particles. This is done simply by numerical differentiation of the sufficiently smooth grid function V_H^0 and interpolation. The evaluation of Eq. (93) is the price we have to pay when we compare the computational cost of this approach as opposed to the Ferry's approach which uses simple forward, backward or centered difference scheme for the calculation of the electric field. However, with this novel effective potential approach we avoid the use of adjustable parameters.

In the following we illustrate the application of this effective potential approach to modeling of size-effects in SOI MOSFETs in which, because of the nearly undoped channel region, size-quantization effects play a major role in the confined region sandwiched between the two oxide layers. The SOI device modeled here has a gate length of 40 nm, the source/drain length is 50 nm each, the gate oxide thickness is 7 nm with a 2 nm source/drain overlap, the box oxide thickness is 200 nm, the channel doping is uniform at 1×10^{17} cm^{-3}, the doping of the source/drain regions equals 2×10^{19} cm^{-3}, and the gate is assumed to be a metal gate with workfunction equal to the semiconductor affinity. There is a 10 nm spacer region between the gate and the source/drain contacts. The silicon (SOI) film thickness is varied over a range of 1–10 nm for the different simulations that were performed to capture the trend in the variations of the device threshold voltage. Similar experiments were performed in Refs. [77,78] using the Schrödinger-Poisson solver and Ferry's effective potential approaches, respectively. Threshold voltage is extracted from the channel inversion density versus gate bias profile and extrapolating the linear region of the characteristics to a zero value. This method also corresponds well to the linear extrapolation technique using the drain current-gate voltage characteristics.

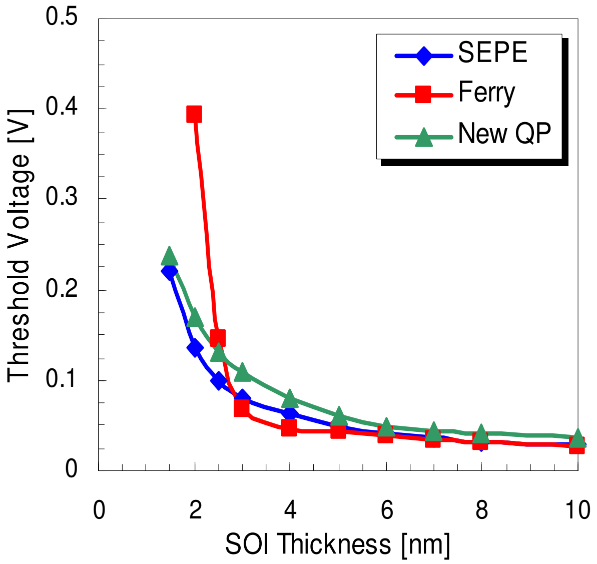


Figure 27. Threshold voltage variation with SOI film thickness.

The results showing the trend in the threshold voltage variation with respect to the SOI film thickness are depicted in Figure 27. One can see that the simple effective potential approach overestimates the threshold voltage for a SOI thickness of 3 nm due to the use of rather approximate value for the standard deviation of the Gaussian wave packet which results in a reduced sheet electron density. As the silicon film thickness decreases, the resulting confining potential becomes more like rectangular from combined effects of both the inversion layer quantization and the SOI film (physical) quantization, which also emphasizes the need for using a more realistic quantum-mechanical wavepacket description for the confined electrons. Of most importance in this figure is the very fact that the new quantum potential approach is free from this large discrepancy and can capture the trend in the threshold voltage as obtained from the more accurate Schrödinger-Poisson solver.

Figure 28 shows a double gate (DG) SOI device structure simulated in this section, which is similar to the devices reported in Ref. [79]. For quantum simulation purposes, only the dotted portion of the device which has been termed as the *intrinsic* device is taken into consideration. DG devices provide better control of the channel charge, hence improved performance and less sensitivity to short channel effects.

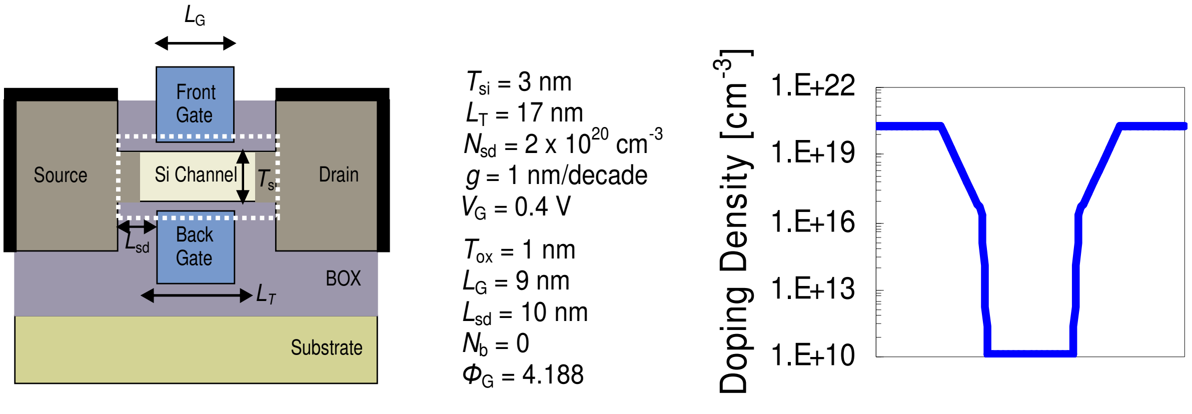


Figure 28. Double gate (DG) device structure.

The intrinsic device consists of two gate stacks (the gate contact and SiO_2 gate dielectric) above and below a thin silicon film. For the intrinsic device, the thickness of the silicon film is 3 nm. Use of a thicker body reduces the series resistance and the effect of process variation, but it also degrades the short channel effects (SCE). From SCE point of view, a thinner body is preferable but it is harder to fabricate and maintain uniform thickness; the same amount of process variation (±10%) may give intolerable fluctuations in the device characteristics. The top and bottom gate insulator thickness is 1 nm, which is expected to be near the scaling limit for SiO_2. As for the gate contact, a metal gate with tunable workfunction, Φ_G, is assumed, where Φ_G is adjusted to provide a specified off-current value of 4 μA/μm. The background doping of the silicon film is taken to be intrinsic, however, due to diffusion of the dopant ions, the doping profile from the heavily doped S/D extensions to the intrinsic channel is graded with a coefficient of g which equals to 1 nm/dec. For convenience, the doping scheme is also shown in

Figure 28. According to the roadmap, the high performance (HP) device should have a gate length of L_G = 9 nm at the year 2016. At this scale, two-dimensional (2D) electrostatics and quantum mechanical effects both play an important role and traditional device simulators may not provide reliable projections. The length, L_T, is an important design parameter in determining the on-current, while gate metal workfunction, Φ_G, directly controls the off-current. The doping gradient, g, affects both on-current and off-current. Values of all the structural parameters of the device are shown in

Figure 28.

The intrinsic device is simulated using the quantum effective potential approach discussed above in order to gauge the impact of size-quantization effects on the DG SOI performance. The results are then compared to that from a full quantum approach based on the non-equilibrium Green's function (NEGF) formalism (NanoMOS–2.5) developed at Purdue University [80]. In this method, scattering inside the intrinsic device is treated by a simple Buttiker probe model, which gives a phenomenological description of scattering and is easy to implement under the Greens' function formalism. The simulated output characteristics are shown in Figure 29. Devices with both 3 nm and 1 nm channel thickness are used with applied gate bias of 0.4 V. The salient features of this figure are: (1) Even with an undoped channel region, the devices achieve a significant improvement with respect to the short channel effects (SCEs) as depicted in flatness of the saturation region. This is due to the use of the two gate electrodes and an ultrathin SOI film which makes the gates gain more control on the channel charge. (2) Reducing the channel SOI film thickness to 1 nm further reduces the SCEs and improves the device performance. However, the reduction in the drive current at higher drain biases is due to series resistance effect pronounced naturally when the drain current increases. (3) Regarding the quantum effects, one can see that quantum-mechanical size quantization does not play a dominant role in degrading the device drive current mainly because of the use of an undoped channel region. Also, looking at the 3 nm (or 1 nm) case alone, one can see that the impact of quantization effects reduces as the drain voltage increases because of the growing bulk nature of the channel electrons. (4) The percentage reduction in the drain current is more pronounced in 1nm case throughout the range of applied drain bias because of the stronger physical confinement arising from the two SiO_2 layers sandwiching the silicon film. (5) Finally, the comparison between the quantum potential formalism and the NEGF approach for the device with 3 nm SOI film thickness shows reasonable agreement which further establishes the applicability of this method in the simulations of different technologically viable nanoscale classical and nonclassical MOSFET device structures.

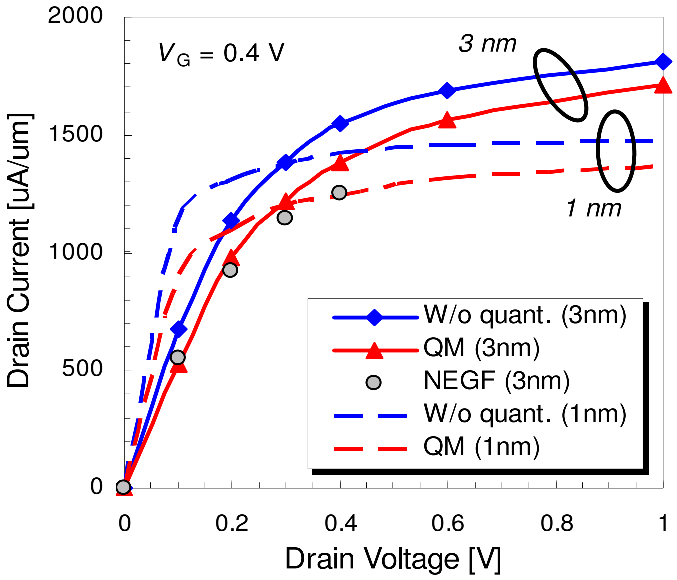


Figure 29. Generic DG SOI device output characteristics.

4.3 Description of Gate Current Models Used in Device Simulations

As already discussed several times, tunneling is one of the most important quantum mechanical phenomena that occurs in mesoscopic devices. In a metal semiconductor contact, it has great impact on the device because electrons are injected through the barrier via the tunneling process. For MESFET devices, in order to calculate gate current by using Monte Carlo simulation accurately, one needs to incorporate this tunneling phenomena. The aim of this section is to offer a treatment of tunneling current through the metal–semiconductor junction within the frameword of ensemble Monte Carlo device simulations.

The band diagram of a Schottky barrier obtained from Monte Carlo simulation is shown in

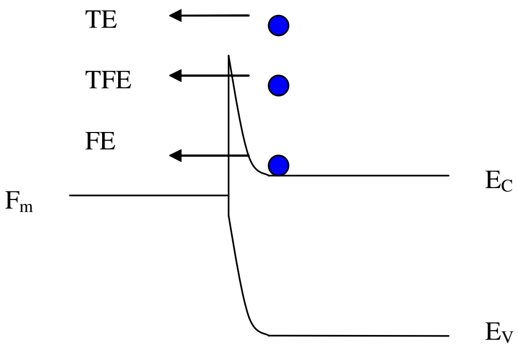


Figure 30. The injection of electron between the Schottky gate and the device channel is handled by using transmission probabilities that are obtained from solving the Schrödinger equation along paths perpendicular to the metal-semiconductor interface. Here, the Schrödinger equation is solved analytically for the approximately triangular barrier using the Airy function approach. The potential profiles along these paths are taken from the solution of the Poisson equation at each self consistent step of the Monte Carlo simulation procedure. Each of these profiles is then considered as an arbitrary one-dimensional piecewise linear potential barrier. When an electron is incident on a given section of the Schottky gate, a random number is generated and if it is smaller than the tunneling probability for the incident electron energy, the electron is allowed to tunnel through the barrier appearing instantaneously on the other side, giving rise to a gate leakage current.

If the exact solution of the Schrödinger equation across a particular potential barrier is not available, several approaches may be considered, depending on the shape of the potential barrier. One commonly used method is the WKB approximation. However, this method is not robust, since it fails to explain certain resonance phenomena observed in a number of tunneling experiments.

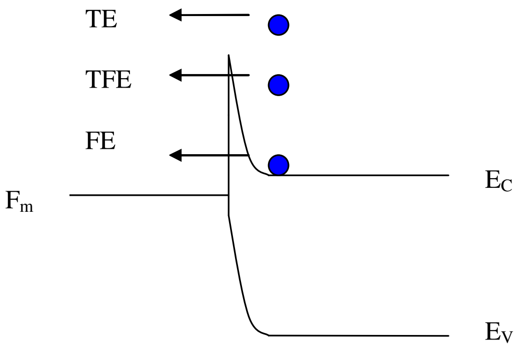


Figure 30. Current transport mechanism in a thin Schottky barrier diode. The various mechanisms are: TE – Thermionic emission over the barrier; TFE – Thermionic field emission; FE – Field emission tunneling.

The methodology we have adopted for the tunneling coefficient calculation is based on the analytical solution of the Schrödinger equation across a linearly varying potential. In this case, the solution can be expressed as linear combination of Airy functions. Proper boundary conditions are imposed at the interface between adjacent linear intervals of the potential using a transfer matrix procedure.

Consider a piecewise linear potential function such that the potential energy profile varies linearly in the region (a_{i-1}, a_i), as shown in Figure 31. The piece-wise linear approximation of the actual 1D potential is given by the expression:

$$\begin{aligned} V(x) &= V(a_{i-1}) + \frac{x-a_{i-1}}{a_i - a_{i-1}}[V(a_i) - V(a_{i-1})] \\ &= V_{i-1} + \frac{V_i - V_{i-1}}{a_i - a_{i-1}}(x - a_{i-1}) \end{aligned} \qquad (94)$$

The electric field profile in the piece-wise linear approximation then equals to:

$$F_i = -\left.\frac{d\varphi}{dx}\right|_i = \frac{1}{q}\left.\frac{dV}{dx}\right|_i = -\frac{V_i - V_{i-1}}{a_i - a_{i-1}}$$, where V_i is in eV. **(95)**

Therefore,

$$V(x) = V_{i-1} + F_i(x - a_{i-1})$$ **(96)**

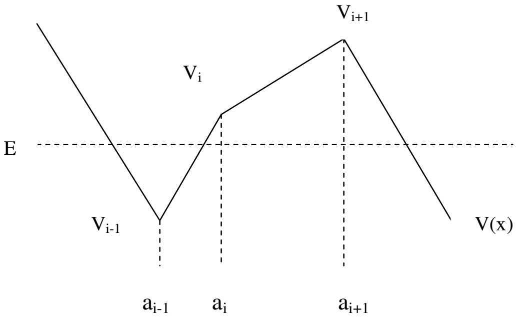


Figure 31. Piecewise linear potential barrier.

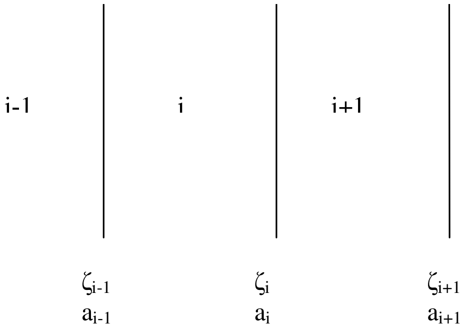


Figure 32. Slicing of the region and corresponding variables in the slices.

Substituting Eq. (94) back into the time independent Schrödinger Wave Equation gives

$$-\frac{\hbar^2}{2m}\frac{d^2\Psi}{dx^2} + V(x)\psi = E\psi ,$$

$$\Rightarrow -\frac{\hbar^2}{2m}\frac{d^2\Psi}{dx^2} + [V_{i-1} + F_i(x - a_i)]\psi = E\psi ,$$

$$\Rightarrow -\frac{\hbar^2}{2m}\frac{d^2\Psi}{dx^2} + F_i x\Psi = (E + F_i a_i - V_{i-1})\Psi ,$$

$$\Rightarrow -\frac{\hbar^2}{2m}\frac{d^2\Psi}{dx^2} + F_i x\Psi = \varepsilon'\Psi.$$ **(97)**

We now define a dimensionless variable ξ such that $\xi = (\frac{2mF_i}{\hbar^2})^{1/3} x - \frac{2m\varepsilon'}{\hbar^2}(\frac{\hbar^2}{2mqF_i})^{2/3}$. Then, the second derivative in Eq. (97) becomes,

$$\frac{d\psi}{dx}=\frac{d\psi}{d\xi}\frac{d\xi}{dx}=(\frac{2mF_i}{\hbar^2})^{1/3}\frac{d\psi}{d\xi}\ ,\quad \frac{d^2\psi}{dx^2}=(\frac{2mF_i}{\hbar^2})^{2/3}\frac{d^2\psi}{d\xi^2}\ . \tag{98}$$

Also, from the variable transformation we have $(\frac{2mF_i}{\hbar^2})^{1/3}x=\xi+\frac{2mE'}{\hbar^2}(\frac{\hbar^2}{2mF_i})^{2/3}$. Substituting back in to the Schrödinger equation leads to $(\frac{2mF_i}{\hbar^2})^{1/3}x=\xi+\frac{2mE'}{\hbar^2}(\frac{\hbar^2}{2mF_i})^{2/3}$. Multiplying both sides by $(\frac{\hbar^2}{2mF_i})^{2/3}$ gives

$$\frac{d^2\psi}{d\xi^2}-\xi\psi(\xi)=0 \ . \tag{99}$$

Now if V is in Joules then F_i is replaced by qF_i, and $\xi=r_i x-\frac{2m\varepsilon'}{\hbar^2}(\frac{\hbar^2}{2mqF_i})^{2/3}$, $\varepsilon'=E+qF_ia_i-V_{i-1}$. The solutions of the reduced equation are the Airy functions and the modified Airy functions. Thus, $\psi_i=C_i^{(1)}A_i(\xi)+C_i^{(2)}B_i(\xi)$, and $\psi_{i+1}(\xi)=C_{i+1}^{(1)}A_i(\xi)+C_{i+1}^{(2)}B_i(\xi)$.

From the continuity and the smoothness conditions for the wave function at $x=a_i$ we get

$$\begin{aligned}&\psi_i(\xi_i)=\psi_{i+1}(\xi_i)\,,\\ &\left.\frac{d\psi_i}{dx}\right|_{a_i}=\left.\frac{d\psi_{i+1}}{dx}\right|_{a_i}\Rightarrow\frac{d\psi_i}{dx}=\left.\frac{d\psi_i}{d\xi}\right|_{\xi_i}\frac{d\xi}{dx}=r_i\frac{d\psi_i}{d\xi}\ ,\\ &\left.\frac{d\psi_{i+1}}{dx}\right|_{a_i}=r_{i+1}\left.\frac{d\psi_{i+1}}{dx}\right|_{\xi_i}.\end{aligned} \tag{100}$$

Therefore,

$$C_i^{(1)}A_i(\xi_i)+C_i^{(2)}B_i(\xi_i)=C_{i+1}^{(1)}A_i(\xi_i)+C_{i+1}^{(2)}B_i(\xi_i)\,, \tag{101a}$$

$$r_iC_i^{(1)}A_i'(\xi_i)+r_iC_i^{(2)}B_i'(\xi_i)=r_{i+1}C_{i+1}^{(1)}A_i'(\xi_i)+r_{i+1}C_{i+1}^{(2)}B_i'(\xi_i). \tag{101b}$$

Rearranging Eqs. (101) and writing them in a matrix form gives,

$$\begin{bmatrix}A_i(\xi_i) & B_i(\xi_i)\\ r_iA_i'(\xi_i) & r_iB_i'(\xi_i)\end{bmatrix}\begin{bmatrix}C_i^{(1)}\\ C_i^{(2)}\end{bmatrix}=\begin{bmatrix}A_i(\xi_i) & B_i(\xi_i)\\ r_{i+1}A_i'(\xi_i) & r_{i+1}B_i'(\xi_i)\end{bmatrix}\begin{bmatrix}C_{i+1}^{(1)}\\ C_{i+1}^{(2)}\end{bmatrix}$$

$$\begin{bmatrix}C_i^{(1)}\\ C_i^{(2)}\end{bmatrix}=M^{-1}\begin{bmatrix}A_i(\xi_i) & B_i(\xi_i)\\ r_{i+1}A_i'(\xi_i) & r_{i+1}B_i'(\xi_i)\end{bmatrix}\begin{bmatrix}C_{i+1}^{(1)}\\ C_{i+1}^{(2)}\end{bmatrix} \tag{102}$$

$$where\ \ M^{-1}=\frac{1}{\det M}\begin{bmatrix}r_iB_i'(\xi_i) & -r_iA_i'(\xi_i)\\ -B_i(\xi_i) & A_i(\xi_i)\end{bmatrix}^T,$$

and $\det M=r_i[A_i(\xi_i)B_i'(\xi_i)-A_i'(\xi_i)B_i(\xi_i)]=\frac{r_i}{\pi}$. As a result $M^{-1}=\frac{\pi}{r_i}\begin{bmatrix}r_iB_i'(\xi_i) & -B_i(\xi_i)\\ -r_iA_i'(\xi_i) & A_i(\xi_i)\end{bmatrix}$

and Eq. (102) becomes

$$\begin{bmatrix} C_i^{(1)} \\ C_i^{(2)} \end{bmatrix} = \frac{\pi}{r_i}\begin{bmatrix} r_i B_i'(\xi_i) & -B_i(\xi_i) \\ -r_i A_i'(\xi_i) & A_i(\xi_i) \end{bmatrix}\begin{bmatrix} A_i(\xi_i) & B_i(\xi_i) \\ r_{i+1}A_i'(\xi_i) & r_{i+1}B_i'(\xi_i) \end{bmatrix}\begin{bmatrix} C_{i+1}^{(1)} \\ C_{i+1}^{(2)} \end{bmatrix} = M_i \begin{bmatrix} C_{i+1}^{(1)} \\ C_{i+1}^{(2)} \end{bmatrix}. \qquad \textbf{(103)}$$

Now consider the case for the initial boundary between region 0 and region 1. In region 0 the wave function is described as plane wave and in region 1 it is a combination of Airy functions. Then

$$\begin{aligned} \psi_0 &= C_0^{(1)} e^{ik_o x} + C_0^{(2)} e^{-ik_o x}, \\ \psi_1(\xi) &= C_1^{(1)} A_i(\xi) + C_1^{(2)} B_i(\xi). \end{aligned} \qquad \textbf{(104)}$$

The continuity of the wave function and of the derivative of the wave function leads to

$$\begin{aligned} & C_0^{(1)} + C_0^{(2)} = C_1^{(1)} A_i(\xi_0) + C_1^{(1)} A_i(\xi_0), \\ & ik_0[C_0^{(1)} - C_0^{(2)}] = r_1 C_1^{(1)} A_i'(0) + r_1 C_1^{(2)} B_i'(0). \end{aligned} \qquad \textbf{(105)}$$

In summary,

$$\begin{bmatrix} C_0^{(1)} \\ C_0^{(2)} \end{bmatrix} = \begin{bmatrix} \frac{1}{2}[A_i(0) + \frac{r_1}{ik_0} A_i'(0)] & \frac{1}{2}[B_i(0) + \frac{r_1}{ik_0} B_i'(0)] \\ \frac{1}{2}[A_i(0) - \frac{r_1}{ik_0} A_i'(0)] & \frac{1}{2}[B_i(0) + \frac{r_1}{ik_0} B_i'(0)] \end{bmatrix}\begin{bmatrix} C_1^{(1)} \\ C_1^{(2)} \end{bmatrix}. \qquad \textbf{(106)}$$

We now consider the other boundary [*N*,*N*+1]. In region N we have a combination of Airy functions and in region *N*+1 we have plane waves. Hence, we have

$$\begin{aligned} \psi_N(\xi) &= C_N^{(1)} A_i(\xi) + C_N^{(2)} B_i(\xi), \\ \psi_{N+1}(\xi) &= C_{N+1}^{(1)} e^{ik_{N+1}x} + C_{N+1}^{(2)} e^{-ik_{N+1}x}. \end{aligned} \qquad \textbf{(107)}$$

The continuity of the wave function and of the derivative of the wave function then implies

$$\begin{aligned} & C_N^{(1)} A_i(\xi_N) + C_N^{(2)} B_i(\xi_N) = C_{N+1}^{(1)} e^{ik_{N+1}a_{N+1}} + C_{N+1}^{(2)} e^{-ik_{N+1}a_{N+1}}, \\ & r_N C_N^{(1)} A_i'(\xi_N) + r_N C_N^{(2)} B_i'(\xi_N) = ik_{N+1}[C_{N+1}^{(1)} e^{ik_{N+1}a_N} - C_{N+1}^{(1)} e^{-ik_{N+1}a_N}. \end{aligned} \qquad \textbf{(108)}$$

In matrix form finally we get,

$$\begin{bmatrix} C_N^{(1)} \\ C_N^{(2)} \end{bmatrix} = \frac{\pi}{r_n}\begin{bmatrix} r_N B_i'(\xi_N) + ik_{N+1}B_i(\xi_N) & r_N B_i'(\xi_N) - ik_{N+1}B_i(\xi_N) \\ -r_N A_i'(\xi_N) + ik_{N+1}A_i(\xi_N) & -r_N A_i'(\xi_N) - ik_{N+1}A_i(\xi_N) \end{bmatrix} M_1 \begin{bmatrix} C_{N+1}^{(1)} \\ C_{N+1}^{(2)} \end{bmatrix}. \qquad \textbf{(109)}$$

Combining Eqs. (103), (106) and (109), we find that the total matrix is given by

$$M_T = M_{FI} M_1 M_2 .... M_{N-1} M_{BI} \begin{bmatrix} e^{ik_{N+1}a_N} & 0 \\ 0 & e^{-ik_{N+1}a_N} \end{bmatrix}$$

$$= \begin{bmatrix} m_{11}^T & m_{12}^T \\ m_{21}^T & m_{22}^T \end{bmatrix} \begin{bmatrix} e^{ik_{N+1}a_N} & 0 \\ 0 & e^{-ik_{N+1}a_N} \end{bmatrix}. \qquad (110)$$

The transmission coefficient is then calculates using

$$T = \frac{k_{N+1}}{k_0} \frac{1}{\left| m_{11}^T \right|^2} \qquad (111)$$

where m_{11}^T is the element of the matrix $M_T = M_{FI} M_1 M_2 .... M_{N-1} M_{BI}$ and

$$M_{FI} = \begin{bmatrix} \frac{1}{2}[A_i(0) + \frac{r_1}{ik_0} A_i^{'}(0)] & \frac{1}{2}[B_i(0) + \frac{r_1}{ik_0} B_i^{'}(0)] \\ \frac{1}{2}[A_i(0) - \frac{r_1}{ik_0} A_i^{'}(0)] & \frac{1}{2}[B_i(0) + \frac{r_1}{ik_0} B_i^{'}(0)] \end{bmatrix},$$

$$M_{BI} = \frac{\pi}{r_n} \begin{bmatrix} r_N B_i^{'}(\xi_N) + ik_{N+1} B_i(\xi_N) & r_N B_i^{'}(\xi_N) - ik_{N+1} B_i(\xi_N) \\ -r_N A_i^{'}(\xi_N) + ik_{N+1} A_i(\xi_N) & -r_N A_i^{'}(\xi_N) - ik_{N+1} A_i(\xi_N) \end{bmatrix},$$

$$M_i = \frac{\pi}{r_i} \begin{bmatrix} r_i B_i^{'}(\xi_i) & -B_i(\xi_i) \\ -r_i A_i^{'}(\xi_i) & A_i(\xi_i) \end{bmatrix} \begin{bmatrix} A_i(\xi_i) & B_i(\xi_i) \\ r_{i+1} A_i^{'}(\xi_i) & r_{i+1} B_i^{'}(\xi_i) \end{bmatrix}. \qquad (112)$$

In the actual implementation of the method outlined above, one considers the electrons in the active region and calculates the potential profile along the thickness of the device by solving Poisson's equation. Then, applying the Airy function transfer matrix method, one calculates the transmission probability in slices. On the basis of particle's position one calculates its potential energy. Then, one compares each particle's energy with the corresponding grid point potential energy. Finally, using a random number, one evaluates whether each particle is going to tunnel through the Schottky barrier or not. If the transmission probability is greater than the random number then tunneling occurs. Once the particle tunnels, the particle is made inactive for the next iterative steps. For each time increment, one counts the number of particles that tunnel through the barrier. After reaching a steady state condition the tunneling current is calculated from the number of tunneled particles. For verification purposes, the above described technique is applied in a nonlinear potential barrier as shown in Figure 33. The calculated transmission probability is shown in

Figure 34. It is observed that our result is properly matched with the calculation previously performed by Lui *et al*. [81].

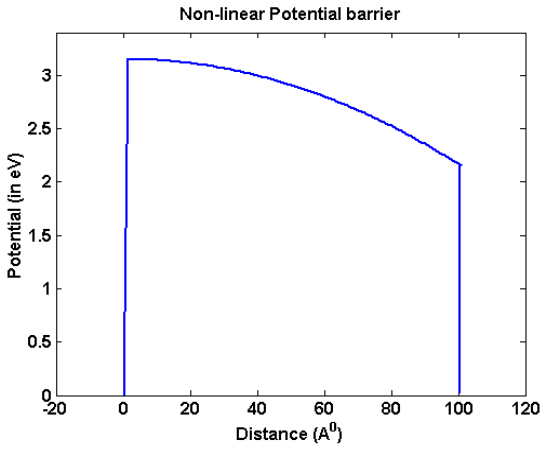


Figure 33. Nonlinear potential barrier is used to calculate quantum mechanical transmission probability.

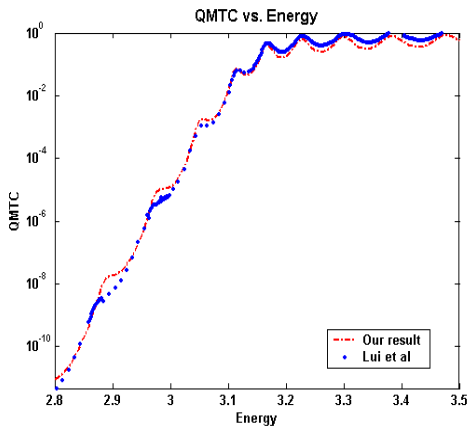


Figure 34. Quantum mechanical transmission probability variation with respect to particle energy compared with analytical results from Lui *et al.* [81].

The methodology described in this section has been applied in the investigation of gate leakage in SOI MESFET device structure shown in the left panel of Figure 35. On the right panel of Figure 35 we show the total gate current and the tunneling component. For small gate voltages, tunneling component of the current dominates the total gate current in this structure.

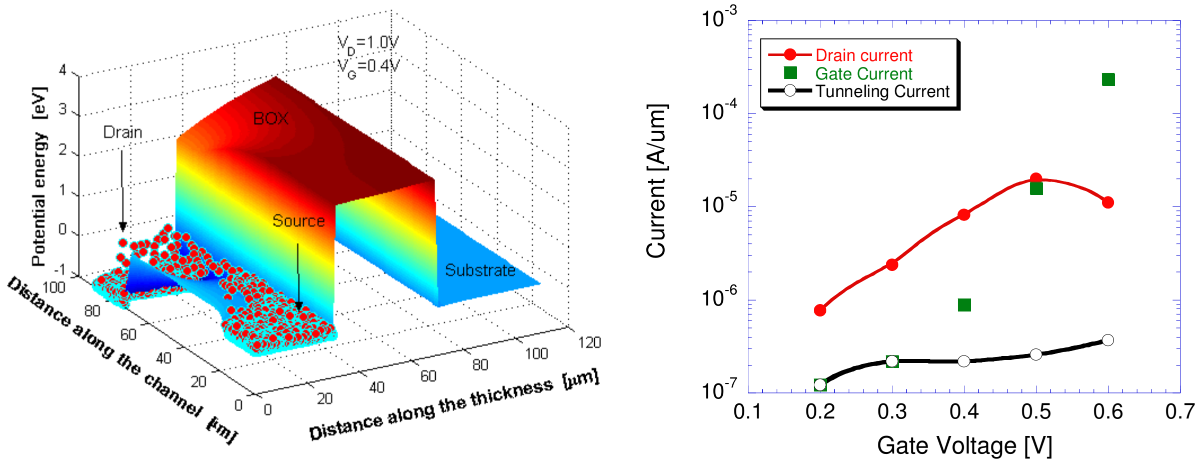


Figure 35. Left panel: Schottky Junction Transistor and electron density. Right panel:Current components.

5. Representative simulation results

5.1 Bulk Monte Carlo Simulations of Different Materials

In most semiconductors in order to properly simulate high field transport it is necessary to consider more than 1 conduction band valley. To calculate the drift velocity along any direction the effective mass along that particular direction is required. In GaAs for example a sub-valley of the L valley lies along the [111] direction. We know the effective masses along the transverse and longitudinal directions of the sub-valley but we do not know the effective mass along the [100] direction. This makes it hard to calculate the drift velocity along the [100] direction.

Assume the Monte-Carlo simulation is run on the x,y,z coordinate system where the x-direction is [100] , y-direction is [010] and z-direction is [001]. Let the 3 perpendicular directions that describe the sub-valley be [a_1,b_1,c_1] , [a_2,b_2,c_2] and [a_3,b_3,c_3]. The electrons in the Monte-Carlo simulation will be drifted according to the x,y,z coordinate system so there will be k_x the wave vector along [100], k_y the wave vector along [010] and k_z the wave vector along [001]. The drift velocity of the electron is first calculated along the 3 directions that describe the sub-valley and along which we know the effective masses.

$$v_{d,[a_1 b_1 c_1]} = \frac{\hbar k_{[a_1 b_1 c_1]}}{m_1(1+2\alpha E)} \tag{113}$$

Here m_1 is the effective mass of the electron along $[a_1,b_1,c_1]$. The final expression of drift velocity along that direction is then calculated by calculating $k_{[a_1 b_1 c_1]}$ using a simple transformation of coordinates to give

$$k_{[a_1 b_1 c_1]} = \frac{k_x a_1}{\sqrt{(a_1^2+b_1^2+c_1^2)}} + \frac{k_y b_1}{\sqrt{(a_1^2+b_1^2+c_1^2)}} + \frac{k_z a_1}{\sqrt{(a_1^2+b_1^2+c_1^2)}} \tag{114}$$

Using similar methods we get $v_{d,[a_2 b_2 c_2]}$ and $v_{d,[a_3 b_3 c_3]}$. The coordinates system is then transformed once again back to the x,y,z coordinate system to get the drift velocities along x,y and z.

$$v_x = \frac{a_1 v_{d,[a_1 b_1 c_1]}}{\sqrt{(a_1^2+b_1^2+c_1^2)}} + \frac{a_2 v_{d,[a_2 b_2 c_2]}}{\sqrt{(a_2^2+b_2^2+c_2^2)}} + \frac{a_3 v_{d,[a_3 b_3 c_3]}}{\sqrt{(a_3^2+b_3^2+c_3^2)}} \tag{115}$$

$$v_y = \frac{b_1 v_{d,[a_1 b_1 c_1]}}{\sqrt{(a_1^2+b_1^2+c_1^2)}} + \frac{b_2 v_{d,[a_2 b_2 c_2]}}{\sqrt{(a_2^2+b_2^2+c_2^2)}} + \frac{b_3 v_{d,[a_3 b_3 c_3]}}{\sqrt{(a_3^2+b_3^2+c_3^2)}}$$

$$v_z = \frac{c_1 v_{d,[a_1 b_1 c_1]}}{\sqrt{(a_1^2+b_1^2+c_1^2)}} + \frac{c_2 v_{d,[a_2 b_2 c_2]}}{\sqrt{(a_2^2+b_2^2+c_2^2)}} + \frac{c_3 v_{d,[a_3 b_3 c_3]}}{\sqrt{(a_3^2+b_3^2+c_3^2)}}$$

Also as the 3 directions are mutually perpendicular we have,

$$a_1 a_2 + b_1 b_2 + c_1 c_2 = 0 \tag{116}$$

$$a_1 a_3 + b_1 b_3 + c_1 c_3 = 0$$

$$a_3 a_2 + b_3 b_2 + c_3 c_2 = 0$$

For N electrons in the simulation an average drift velocity is then calculated as,

$$\langle v_x \rangle = \frac{1}{N}\sum_{i=1}^{N} v_{x,i}\ ,\quad \langle v_y \rangle = \frac{1}{N}\sum_{i=1}^{N} v_{y,i}\ ,\quad \langle v_z \rangle = \frac{1}{N}\sum_{i=1}^{N} v_{z,i} \tag{117}$$

where each drift velocity depends on the sub-valley the electron is currently in. The mobility is calculated as the slope of the velocity versus field curve for low fields. The range of ‘low fields’ varies from material to material, but in all cases it is the range of fields for which the velocity linearly varies with the field.

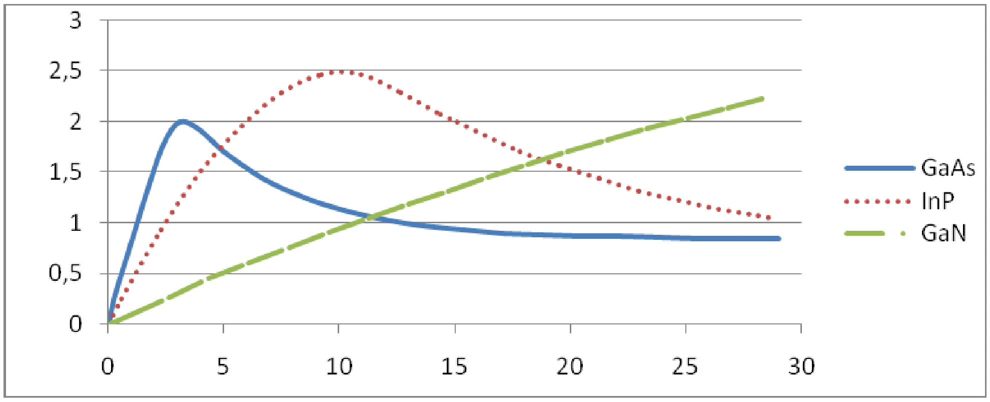


Figure 36. Drift Velocity (10^7cm/s) vs. Electric Field (kV/cm).

The results presented in Figure 36 shows the difference in velocity of electrons for different materials. For GaN the peak velocity of 3x10^7cm/s occurs at around 150kV/cm so that in the electric field range shown in Figure 36 it is clear that for electric fields below 30 kV/cm GaN carriers in the bulk GaN materials are still in the low-field regime.

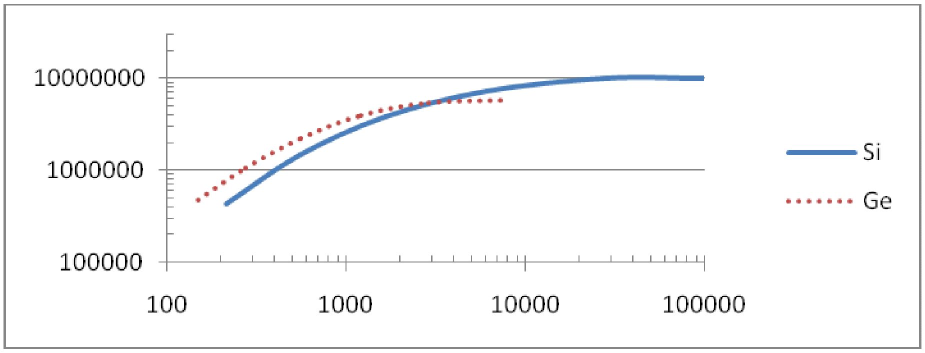
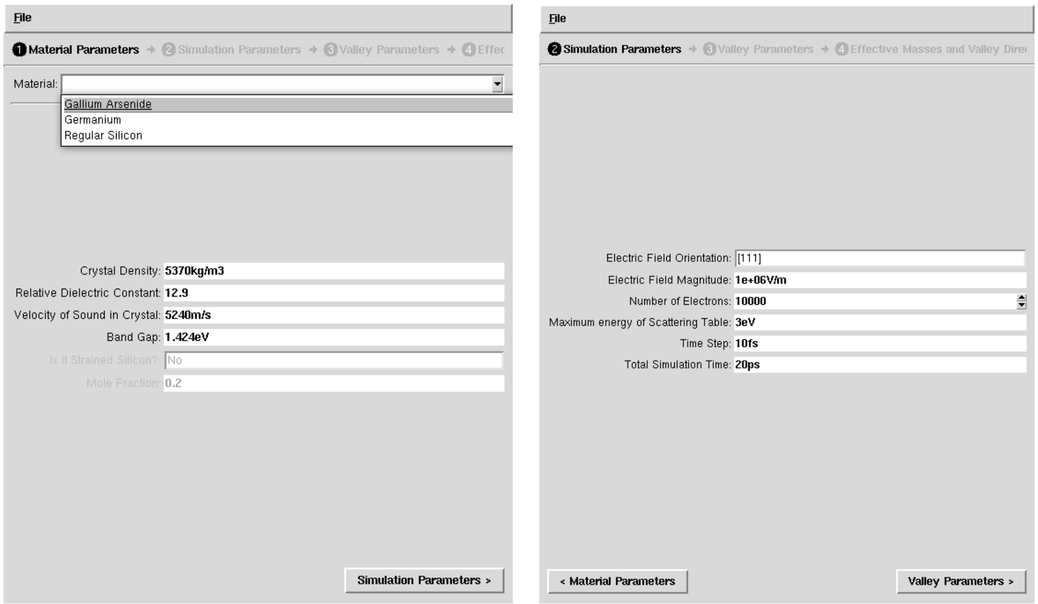


Figure 37. Electron Drift Velocity (cm/s) vs Electric Field (V/cm) in Si and Ge along the principal crystallographic directions.

As the drift velocities saturate at much smaller electric fields in Silicon and Germanium they are plotted on a separate graph (see Figure 37).

The generalized bulk Monte Carlo tool is a learning tool designed to give the user as much freedom as possible in deciding the nature of the material and the scattering parameters he/she wishes to include. The tool is designed in such a way that a material can have up to 4 different valleys with each having up to 12 different sub-valleys. The tool is in the process of being deployed on the www.nanohub.org. The Monte Carlo code was interfaced with Rappture to make it portable on the nanohub. As can be seen on Figure 38 below (left panel), as of now there are 3 materials loaded into the tool. When it is completed there will be around 10 different materials with pre-loaded values. In addition to this the user can define their own material by choosing the valleys, directions, effective masses, etc.

From the drop down menu the user can choose the material and the values will automatically load themselves. The tool guides you across the various parameters required before simulation. Any electric field direction can be chosen as well as the other parameters listed. One can also speed up the simulation by reducing the Maximum energy of the scattering table if one is running a low field simulation, or by reducing the number of electrons simulated (Figure 38 – right panel).


File
1 Material Parameters → 2 Simulation Parameters → 3 Valley Parameters → 4 Effec
Gallium Arsenide
Germanium
Regular Silicon
Simulation Parameters >
File
2 Simulation Parameters → 3 Valley Parameters → 4 Effective Masses and Valley Dire
< Material Parameters
Valley Parameters >

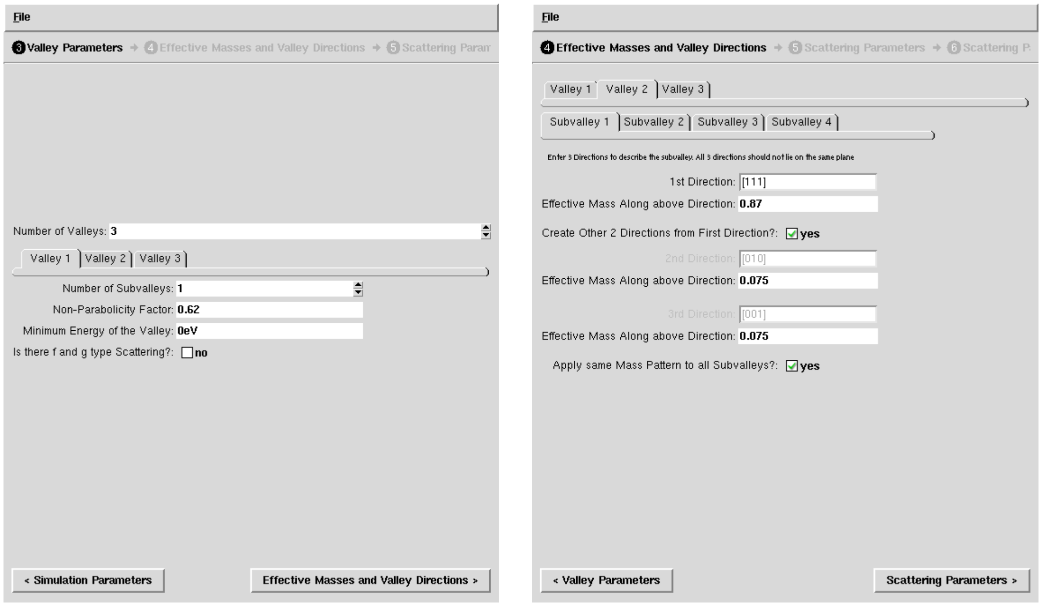
Material:

Crystal Density: 5370kg/m3
Relative Dielectric Constant: 12.9
Velocity of Sound in Crystal: 5240m/s
Band Gap: 1.424eV
Is it Strained Silicon?: No
Mole Fraction: 0.2

Electric Field Orientation: [111]
Electric Field Magnitude: 1e+06V/m
Number of Electrons: 10000
Maximum energy of Scattering Table: 3eV
Time Step: 10fs
Total Simulation Time: 20ps

Figure 38. Choice of material (left panel) and way of imputing simulation parameters (right panel).

The Valley Parameters tab allows you to choose the number of valleys and the properties of those valleys (Figure 39 – left panel). The number of valley tabs shown depends on the number of valleys chosen so as to not clutter the page with too much unnecessary information. The same is done with the number of sub-valley tabs below (Figure 39 – right panel.


File
3 Valley Parameters → 4 Effective Masses and Valley Directions → 5 Scattering Paran
Valley 1
Valley 2
Valley 3
< Simulation Parameters
Effective Masses and Valley Directions >
File
4 Effective Masses and Valley Directions → 5 Scattering Parameters → 6 Scattering P
Valley 1
Valley 2
Valley 3
Subvalley 1
Subvalley 2
Subvalley 3
Subvalley 4
< Valley Parameters
Scattering Parameters >

Number of Valleys: 3
Number of Subvalleys: 1
Non-Parabolicity Factor: 0.62
Minimum Energy of the Valley: 0eV
Is there f and g type Scattering?: ☐ no

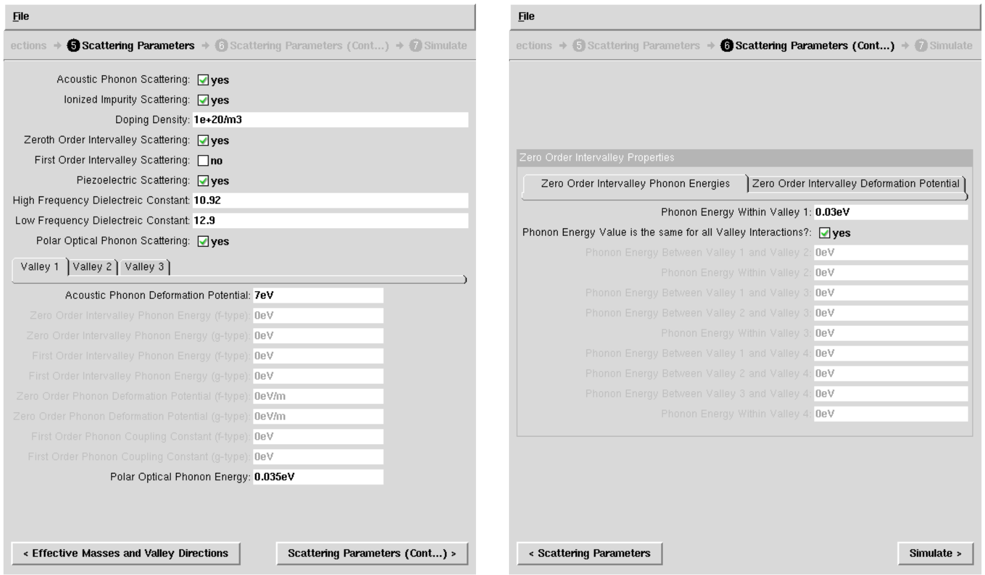
Enter 3 Directions to describe the subvalley. All 3 directions should not lie on the same plane

1st Direction: [111]
Effective Mass Along above Direction: 0.87
Create Other 2 Directions from First Direction?: ☑ yes
2nd Direction: [010]
Effective Mass Along above Direction: 0.075
3rd Direction: [001]
Effective Mass Along above Direction: 0.075
Apply same Mass Pattern to all Subvalleys?: ☑ yes

Figure 39. Graphical user interface for defining valley parameters – left panel. Graphical user interface for defining effective masses and valley orientation – right panel.

To further simplify the input process there are options to let the simulator calculate the transverse directions of the sub-valley so that one is not confined trying to calculate 3 mutually perpendicular directions. There is also an option to apply the same mass pattern of the 1st sub-valley to the other sub-valleys if they are all equivalent. Once these options are ticked the information boxes not required are automatically shaded out so as to not confuse the user. So only the parameters required are left open. The above snapshot shows the L valley description of GaAs.

In this tab (Figure 40) the user can choose the scattering types, and the relevant parameters required are automatically shown. As in the previous tab, the f and g type scattering was not included, thus all the boxes requiring that information in this valley is shaded out. This way the user does not get overwhelmed by unnecessary information.

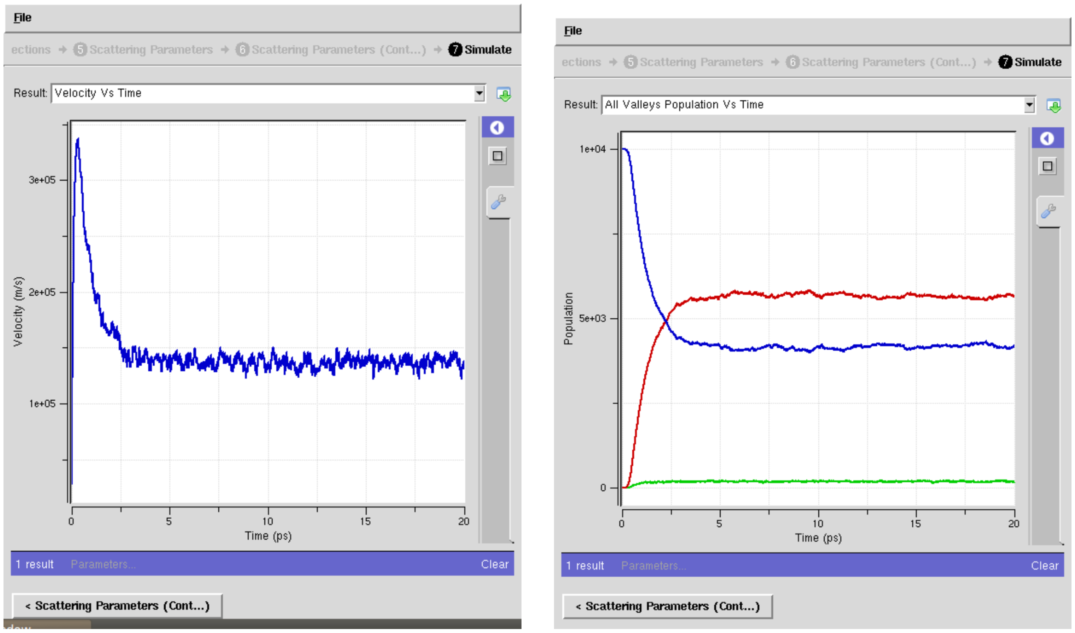

File
ections → 5 Scattering Parameters → 6 Scattering Parameters (Cont...) → 7 Simulate
Valley 1
Valley 2
Valley 3
< Effective Masses and Valley Directions
Scattering Parameters (Cont...) >
File
ections → 5 Scattering Parameters → 6 Scattering Parameters (Cont...) → 7 Simulate
Zero Order Intervalley Properties
Zero Order Intervalley Phonon Energies
Zero Order Intervalley Deformation Potential
< Scattering Parameters
Simulate >

Acoustic Phonon Scattering: ☑ yes
Ionized Impurity Scattering: ☑ yes
Doping Density: 1e+20/m3
Zeroth Order Intervalley Scattering: ☑ yes
First Order Intervalley Scattering: ☐ no
Piezoelectric Scattering: ☑ yes
High Frequency Dielectreic Constant: 10.92
Low Frequency Dielectreic Constant: 12.9
Polar Optical Phonon Scattering: ☑ yes
Acoustic Phonon Deformation Potential: 7eV
Zero Order Intervalley Phonon Energy (f-type): 0eV
Zero Order Intervalley Phonon Energy (g-type): 0eV
First Order Intervalley Phonon Energy (f-type): 0eV
First Order Intervalley Phonon Energy (g-type): 0eV
Zero Order Phonon Deformation Potential (f-type): 0eV/m
Zero Order Phonon Deformation Potential (g-type): 0eV/m
First Order Phonon Coupling Constant (f-type): 0eV
First Order Phonon Coupling Constant (g-type): 0eV
Polar Optical Phonon Energy: 0.035eV

Phonon Energy Within Valley 1: 0.03eV
Phonon Energy Value is the same for all Valley Interactions?: ☑ yes
Phonon Energy Between Valley 1 and Valley 2: 0eV
Phonon Energy Within Valley 2: 0eV
Phonon Energy Between Valley 1 and Valley 3: 0eV
Phonon Energy Between Valley 2 and Valley 3: 0eV
Phonon Energy Within Valley 3: 0eV
Phonon Energy Between Valley 1 and Valley 4: 0eV
Phonon Energy Between Valley 2 and Valley 4: 0eV
Phonon Energy Between Valley 3 and Valley 4: 0eV
Phonon Energy Within Valley 4: 0eV

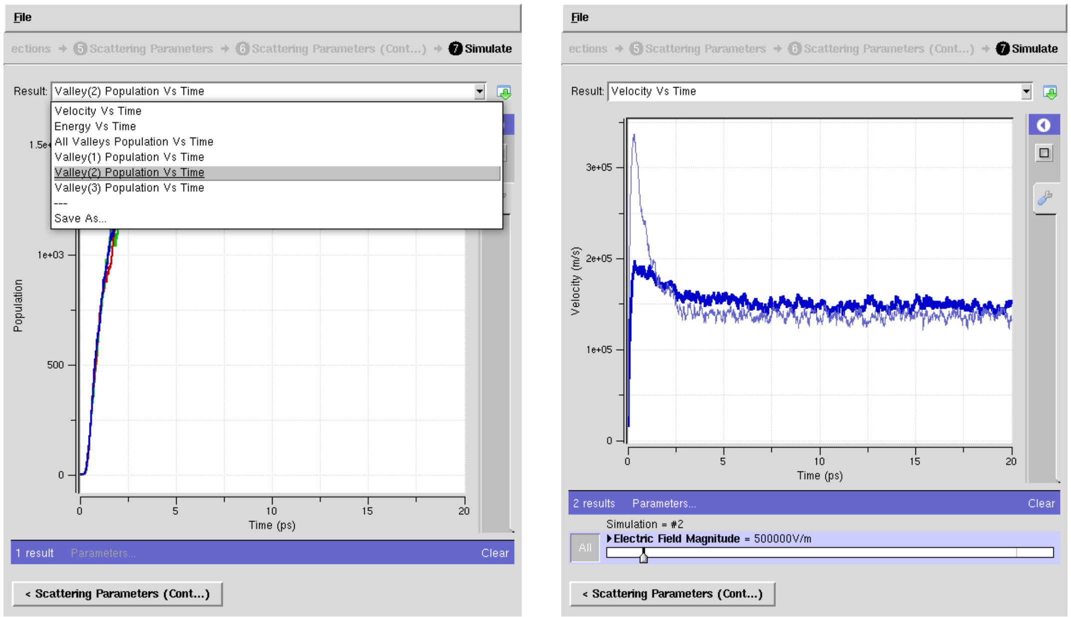
Figure 40. Choice of scattering parameters – acoustic phonon and polar optical phonon scattering – left panel. Choice of scattering – zeroth order intervalley phonon scattering – right panel.

Here as only Zero-Order Intervalley Scattering was included only that information required is shown. There is also an option to allow all valley transitions to have the same phonon energy/deformation potential if needed. There are further adjustments made to the input deck e.g. if f and g type scattering were included in valley 1 there is no need to specify a phonon energy (the independent f and g type phonon energies would have been asked in the previous tab) within the valley , so that option would have been shaded out.


File
ections → 5 Scattering Parameters → 6 Scattering Parameters (Cont...) → 7 Simulate
Velocity (m/s)
3e+05
2e+05
1e+05
0
5
10
15
20
Time (ps)
1 result
Parameters...
Clear
< Scattering Parameters (Cont...)
File
ections → 5 Scattering Parameters → 6 Scattering Parameters (Cont...) → 7 Simulate
Population
1e+04
5e+03
0
0
5
10
15
20
Time (ps)
1 result
Parameters...
Clear
< Scattering Parameters (Cont...)

Result: Velocity Vs Time

Result: All Valleys Population Vs Time


Result: Valley(2) Population Vs Time

Result: Velocity Vs Time

Simulation = #2

Electric Field Magnitude = 500000V/m

Figure 41. Various output that are generated with the bulk Monte Carlo tool.

The above snapshots show the various output curves that are generated. There is an option to observe all valleys' population together or to look within one valley and its sub-valleys' population. This option also changes based on the number of valleys chosen.

5.2 Particle-Based Device Simulations of Random Telegraph Noise Characterization in 45 nm Technology Node Conventional MOSFET Device

Statistical fluctuations of the channel dopant number were predicted by Keyes [82] as a fundamental physical limitation of MOSFETs down-scaling. Entering into the nanometer regime results in a decreasing number of channel impurities whose random distribution leads to significant fluctuations of the threshold voltage and off-state leakage current. These effects are likely to induce serious problems on the operation and performances of logical and analog circuits. Quite recently, the researchers have demonstrated that another fluctuation, random telegraph noise/signal (RTN/RTS), seriously affects operation of digital devices in terms of uncontrollable threshold voltage shifts and transconductance degradation associated with them.

The study of RTS has provided a powerful means of investigating the capture and emission kinetics of single defect in addition to demonstrating the possible microscopic origins of low frequency *1/f* noise in these devices and also providing new insight into the nature of defects at the interface. With each generation of device scaling, the total number of active dopants in the channel region decreases to the extent that, when the device gate length is scaled below sub-100 nm, the dopant distribution can be considered random where the channel is formed. Consequently, a few defects at the Si/SiO$_2$ interface or inside the SiO$_2$ dielectric are sufficient to cause severe RTS related device failure when the dopant distribution becomes fully random across the channel region. Since ΔV_{th} is inversely proportional to the device area, highly integrated digital devices are assumed to be seriously affected by RTN. For instance, Figure 42 illustrates the measurement data from a 90nm SRAM design [83]. The minimum supply voltage (V_{ccmin}), which is highly sensitive to device threshold voltage, exhibits a similar pattern in the time domain as that of RTS. The impact of RTS in this case is more than 200 mV, which is catastrophic to the yield and low-power design of SRAM. Therefore, accurate and physical models of RTS are essential to predict and optimize circuit performance during the design stage. Currently, such models are not available for circuit simulation. The compound between RTS and other sources of variation, such as RDF, further complicates the situation especially in extremely scaled CMOS design.

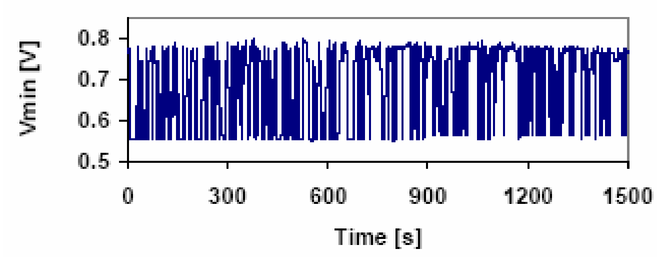


Figure 42. The fluctuation of SRAM V_{cc}(min) due to RTS [83].

5.2.1 Theoretical Model

Three-dimensional Monte Carlo (MC) simulations should provide a more realistic transport description in ultra-short MOSFETs, in particular in the on-state. The MC procedure gives an exact solution of the Boltzmann transport equation. It thus correctly describes the non-stationary transport conditions. Even if microscopic simulations such as the MC method are concerned, the treatment of the electrons and impurities is not straightforward due to, again, the long-range nature of the Coulomb potential. The incorporation of the short range-range Coulomb potential in the MC method has been a long-standing issue [84]. This problem is, in general, avoided by assuming that the electrons and the impurities are always screened by the other carriers so that the short-range part of the Coulomb interaction is effectively suppressed. The complexity of the MC simulation increases as one takes into account more complicated screening processes by using the dynamical and wave-vector dependent dielectric function obtained from, for example, the random phase approximation. Indeed, screening is a very complicated many-body matter [85].

A novel approach has been introduced by the ASU group, in which the MC method is supplemented by a *molecular dynamics* (MD) routine [86]. In this approach, the mutual Coulomb interaction among electrons and impurities is treated in the drift part of the MC transport kernel. Indeed, the various aspects associated with the Coulomb interaction, such as dynamical screening and multiple scatterings, are automatically taken into account. Since a part of the Coulomb interaction is already taken into account by the solution of the Poisson equation, the MD treatment of the Coulomb interaction is restricted only to the limited area near the charged particles. *It is claimed that the full incorporation of the Coulomb interaction is indispensable to reproduce the correct electron mobility in highly doped silicon samples.*

5.2.2 Simulation results for the case of a single charged trap

The simulator described in the previous section is presently being used in the investigation of the random trap fluctuations in 45 nm technology node MOSFET device where, in addition to the randomness of the position and the actual number of the impurity atoms in the whole simulated domain of the device, a random trap or two traps in close proximity [87] are introduced in the middle section of the channel and moved from the source-end to the drain end of the channel. An example of a discrete impurity pattern and a double trap located at the source-end of the channel is shown in **Figure 43**. The effective channel length of 45 nm technology node is taken to be 35 nm.

We consider ensemble of 20 devices with different random dopant distribution. The threshold voltage of each of these devices without the presence of the trap is shown in Figure 44 (left panel). The total variation of the threshold voltage as a function of the trap position in the middle portion of the channel, when the single trap is moved from the source end to the drain end of the channel, is shown in Figure 44 (right panel). We see that the threshold voltage increases from its average value when this trap is located at the source end of the channel. This is due to the fact that carriers see additional large potential barrier due to the presence of the charged trap and are reflected back in the source contact. The threshold voltage rapidly reduces when the trap is moved away from the source injection barrier because when the electrons are injected in the channel, even though the electric field is small (due to small drain bias applied when measuring threshold voltage) they slowly drift towards the drain contact. In Figure 45 we depict the threshold voltage fluctuation taken as a percentage relative to the values in Figure 44 as a function of the trap position when trap is being moved from the middle of the source end of the channel to the middle of the drain end of the channel.

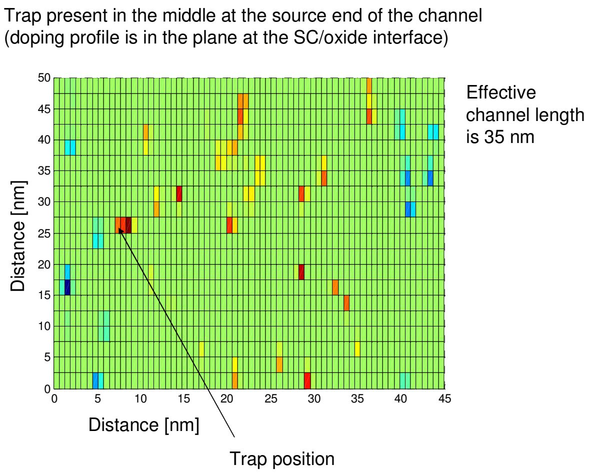


Figure 43. Random dopant distribution and a double trap located the middle of the source end of the channel.

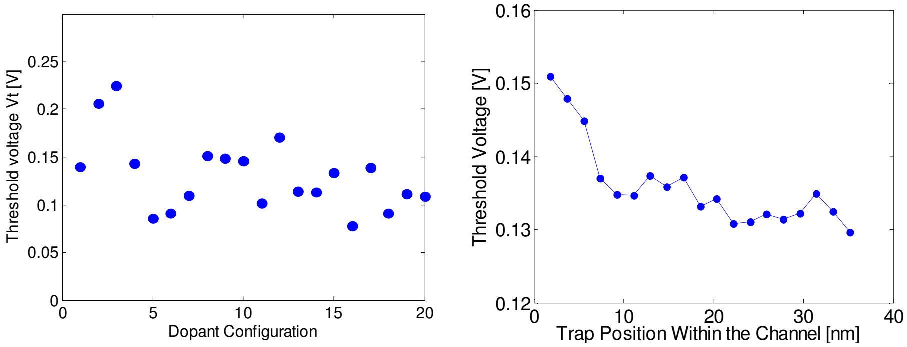


Figure 44. Left Panel: Threshold voltage fluctuations due to random dopant fluctuations (without traps) for a statistical ensemble of 20 devices with different number and different distribution of the impurity atoms. Right Panel: Threshold voltage variation for the case of single trap's position variation averaged over 20 dopant configurations.

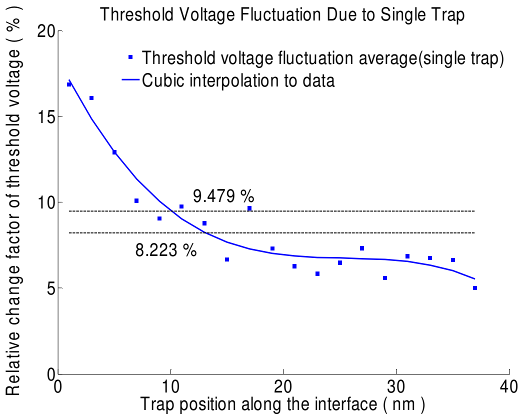


Figure 45. Threshold voltage fluctuation due to single trap's position along the channel (averaged over twenty random dopant configurations).

Figure 46 shows the on-current degradation as a function of the trap position. As depicted on the figure, near the source end of the channel the current degradation due to the presence of trap is smaller when compared to the threshold voltage degradation. Traps near the drain contact, where the electron density is pinched off for the bias conditions used, are not effectively screened and a notable increase of the current degradation of a negatively charged trap is observed. At the drain contact, the degradation drops practically to zero because in this section of the channel traps are effectively screened by the electrons in the contacts. The simulation results presented in this section also confirm that when a significantly large number of dopant configurations are used more reliable predictions for the threshold voltage fluctuations due to charging and discharging of a single trap are obtained. Smoother results will definitely be obtained if a larger statistical ensemble is used. However, the use of a statistical sample of twenty dopant configurations proves to be satisfactory at providing the correct trends and in the physically-based explanation of the results obtained.

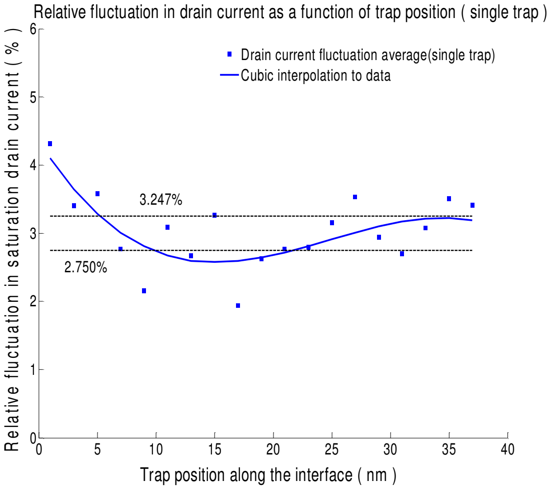


Figure 46. On-current fluctuation as effected by variation of single trap's position (20 random dopant case trap position averaged over 20 random channel dopants.

5.2.3 Simulation results for the case of two charged trap

As the number of traps is increased at the Si/SiO_2 interface, one parameter that aggravates the fluctuation values for the on-current and threshold voltage with its standard deviation is the spacing between the traps. For this purpose, EMC device simulation is performed for enhancing the trap number from single to double but keeping the traps within 1 nm separation from one another. The plots for the threshold voltage fluctuation, the fluctuation in the standard deviation of the threshold voltage and the on-current fluctuations – all underscore a very important feature of closely lying traps. Namely, adjacent traps alter the short range and long range Coulomb potential to the extent that larger number of carriers at the source contact cannot surmount the steep potential barrier due to the presence of closely spaced traps. This results in larger degradation of device parameters such as threshold voltage and drain On-current when compared to the single trap case. The threshold voltage percentage variation for the double-trap case is shown in Figure 47, the on-current degradation is shown in Figure 48 and the threshold voltage standard deviation as a function of two traps position is shown in Figure 49. Comparing the results given in Figure 45(right panel) and Figure 47 for single-trap and double trap, respectively, we see threefold increase in the threshold voltage degradation at the source end of the channel for the case of a double trap. Similarly, comparing the results in Figure 46 and Figure 48 for the case of a single trap and double-trap respectively, we see about a factor of 1.5 increase in the current degradation when the double-trap is placed at the source end of the channel. Therefore, the barrier effect is more pronounced in the off-state but it is still there in the on-state as well.

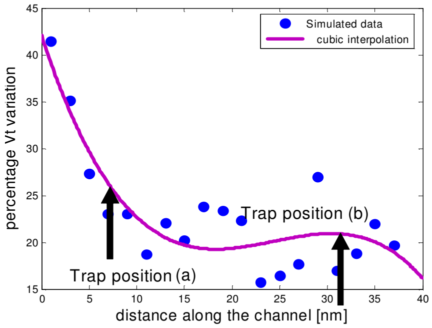


Figure 47. Percentage threshold voltage due to two traps located at the semiconductor/oxide interface and different positions along the middle section of the channel. 20 devices with different random dopant distributions have been averaged out.

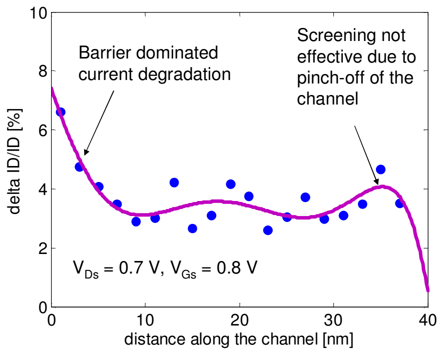


Figure 48. On-current degradation as a function of two traps' positions. The statistical ensemble used here consists of first seven random dopant distributions in both number and positions within the active region of the channel.

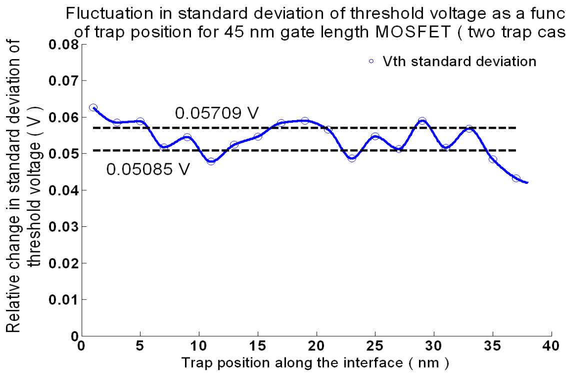


Figure 49. Threshold voltage standard deviation as a function of two traps' positions showing well behaved spatial correlation when sufficient number of random dopants are considered.

5.2.4 Conclusions

Unlike most of the analytical revelation of published articles where it has been reported that fluctuation in drain current amplitude variation and threshold voltage variation tend to diminish at strong inversion and saturation bias conditions owing to the screened out potential due to high inversion charge density at the surface with associated improvement of Coulombic-scattering related mobility, our simulations conducted at saturation bias conditions on a 45 nm channel length MOSFET device reveal that for different random dopant configurations, the fluctuation pattern exhibited by the drain current amplitude variation and the threshold voltage variation (turn-on gate voltage with low drain bias) are truly statistical and random in nature implying that for some specific random dopant distributions, the fluctuation nature is well controlled whereas for some other random dopant distributions, the fluctuation pattern shows significant transitions between local peaks and valleys.

5.3 Modeling Thermal Effects in SOI Devices Using Particle-Based Device Simulator

Although supply voltages are reduced when reducing the size of the channel of nanoscale transistors, there are still sufficiently large electric fields that can accelerate the carriers to average energies of 0.5-0.8 eV. These hot carriers interact with the lattice and transfer their energy to the phonon bath (both acoustic and optical), thus heating the lattice and increasing the lattice vibrations (scattering). This leads to a negative feedback on the carrier drift velocity or mobility (if transport is scattering dominated). Thus, we may conclude that when transistors are in the ON-state, there is a strong interaction of the electrons with the lattice on a very short time scale (in particular with the optical phonons). The optical phonons in a longer time scale decay into acoustic phonons through anharmonic multiple phonon processes, thus removing the heat from the hot spot. Therefore, to properly treat the operation of nanoscale devices it is necessary to, at least, consider the electrons within the Boltzmann Transport Equation (BTE) picture and, in the lowest approximation, via energy balance equations include the acoustic and the optical phonon baths separately. This is what the ASU group has implemented couple of years ago. Our ultimate goal is to have phonon BTE solver coupled with electron BTE solver (project that is currently underway at ASU).

5.3.1 Theoretical Model and Computation Details

A schematic description of our 2D/3D electro-thermal simulator is shown in Figure 50. It is important to note that the EMC transport kernel for the electrons is in the loop for the solution of the energy balance equations for the acoustic and optical phonon temperatures. The criterion for convergence of the whole self-consistent loop (measured in terms of Gummel cycles) is convergence in the current to the third significant digit.

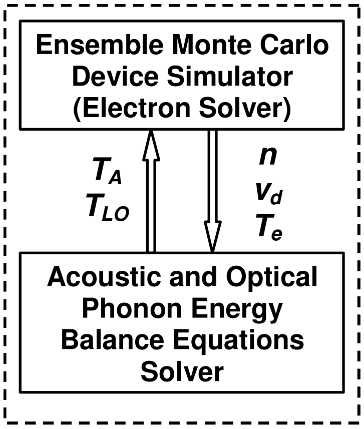

Ensemble Monte Carlo Device Simulator (Electron Solver)
T_A
T_{LO}
n
v_d
T_e
Acoustic and Optical Phonon Energy Balance Equations Solver

Figure 50. Exchange of variables between the two kernels. The electron density (n), electron drift velocity (v_d) and electron temperature (T_e), obtained from the electron solver, are input variables for the phonon kernel. The output variables of the phonon solver are acoustic and optical phonon temperature profiles (T_A and T_{LO}, respectively). They enter in the beginning of the free-flight scattering phase through phonon temperature dependent scattering tables.

Between iterations, there are variables that are being exchanged between the electron and phonon solvers. Namely, after running the EMC for the electrons in the device for about 5 ps for 2D device analysis and 10 ps for the 3D nanowire simulations, the electron transport solver passes to the phonon energy balance solver the information about the spatial variation of the electron density, spatial variation of the average drift velocity and spatial variation of the average electron temperature, where an assumption is made that the drift energy is much smaller than the thermal energy. These variables are then used in the solution of the energy balance equations for the acoustic and optical phonons to get updated values for the spatial variation of the optical phonon and acoustic phonon temperatures which are then used as inputs, in the choice of the scattering table for the electrons (proper scattering table is chosen with the acoustic and optical phonon temperatures at the nearest grid point). Then, when the proper table is selected, based on the energy of the electron and a selection of a random number, scattering mechanism is selected.

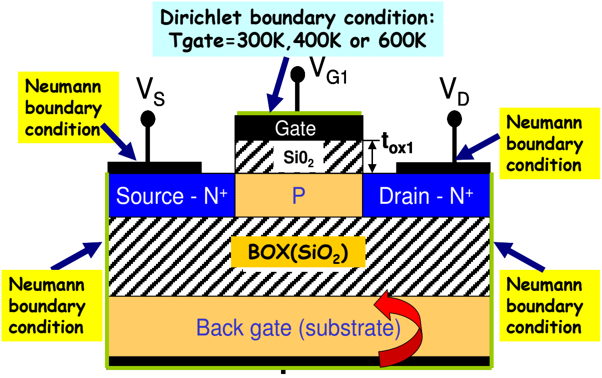

Dirichlet boundary condition:
Tgate=300K,400K or 600K
V_{G1}
V_S
V_D
Neumann boundary condition
Gate
SiO_2
t_{ox1}
Neumann boundary condition
Source - N^+
P
Drain - N^+
BOX(SiO_2)
Neumann boundary condition
Back gate (substrate)
Neumann boundary condition

Figure 51. Schematic of the simulated device structure which illustrates (a) the boundary conditions used in some of the simulations presented, and (b) that the boundary conditions at the bottom of the substrate can be mapped as a boundary condition at the bottom of the BOX.

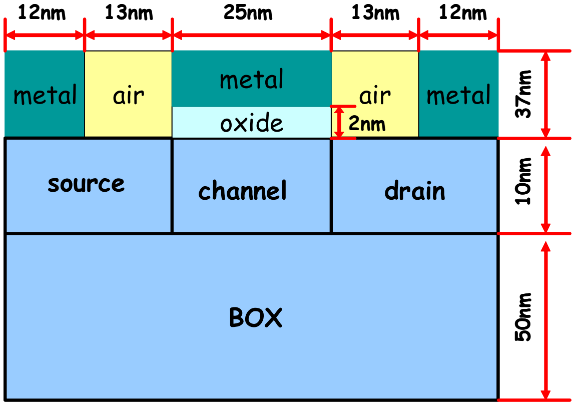


Figure 52. Device structure with extended domain. The geometrical dimensions are for the simulated 25 nm channel length FD-SOI nMOSFET.

The question of the proper boundary conditions for the electronic part of the problem is rather clear and has been discussed in many papers and texts in the literature. The problem in properly specifying the phonon boundary conditions is the selection of acoustic and optical phonon temperatures at either the artificial boundaries or at the contacts. Typically acoustic phonon temperature is equated with the lattice temperature. To better understand the choice of boundary conditions we have taken in various works [88], we want to point out that lattice temperature is analogous to electrostatic potential and heat flux is analogous to current. As it is well known, when considering electrical behavior of the device, at least one node within the structure has to have Dirichlet boundary conditions specified. Analogously, for the lattice temperature, we need at least one node that is a thermal contact and whose temperature is set to 300 K. As illustrated in Figure 51 for the case of fully-depleted (FD) silicon on insulator (SOI) devices, the bottom of the box is assumed to be at thermal equilibrium. Also, the gate contact is assumed to be at thermal equilibrium, but not necessarily at 300 K. For example, in one set of simulations we vary the temperature on the gate to be 300K, 400K, and 600K. Now comes the question: What happens to the source and drain contacts and the side artificial boundaries? Specifying Dirichlet boundary conditions at the ohmic contacts is not accurate from a standpoint that there is current flowing through the contacts, and since the contacts have finite resistance, there will be Joule heating (so the problem becomes unconstrained). The best solution to this, as we have done in several studies, is to extend the metal contact to become part of the simulation domain and to apply at the very ends of the contact isothermal boundary conditions (see Figure 52). With the top and bottom specified, we have vertical transport of heat through the structure. The next question is, is there lateral transfer of heat across the artificial boundaries? Here we can consider two cases: case when the neighboring device is ON, in which case Neumann boundary conditions are appropriate, and case when the neighboring device is OFF, in which case Dirichlet boundary conditions are appropriate. In the case when Neumann boundary conditions are applied across the artificial boundaries, the heat transport remains vertical, but for the case when Dirichlet boundary conditions are applied across the boundary, the heat transport has both vertical and horizontal component. Case of Neumann boundary conditions can happen in analog circuits such as current mirror in which both transistors are simultaneously ON, whereas the case of Dirichlet boundary conditions occurs in digital circuits in which most of the time the transistors are OFF. Yet another issue that deserves attention in getting physically correct results is the proper choice of the thermal conductivity for thin silicon slabs and for nanowires. Asheghi and co-workers [89] and Li Shi and co-workers [90] have demonstrated via experimental measurements that thermal conductivities of thin silicon films and silicon nanowires, respectively, strongly depend of the geometry as for smaller geometries phonon boundary scattering can reduce the thermal conductivity of the silicon film or nanowire by a factor of 10 or more from its bulk value. Moreover, thermal conductivity is temperature dependent quantity. We have done extensive effort, using the theory of Sondhaimer for conductivity of metals [91], to arrive to an empirical formula that simultaneously describes the thickness and temperature dependence of the thermal conductivity. Our empirical expression perfectly matches the experimental data of Asheghi and co-workers [89] In the model, it is assumed that phonon boundary scattering is perfectly diffusive.

It has been shown by the Wisconsin group [92] that the thermal conductivity, in general is a tensor quantity. The temperature dependent thermal conductivity tensors were calculated by Aksamija and co-workers and were given to us. The energy balance equation for acoustic phonons was modified accordingly to take into consideration the temperature dependence of the thermal conductivity tensor. Simulation results were performed using our empirical model and treating the conductivity as a tensor quantity. We did not observe significant differences in the results for (100) crystallographic orientations. To further investigate the problem, we modified the EMC device simulator so that “arbitrary” crystallographic orientations can be considered and the proper temperature dependent thermal conductivity tensors were used. The results we get seem a little bit controversial because for both [100], [110] and [111] transport directions we assume identical values for the specular vs. diffusive scattering ratio at the Si/SiO$_2$ interface that accounts for the interface roughness scattering. More investigation of this issue is needed to give more conclusive comments.

Finally, we come to the last point we want to make and that is: What is the best choice of buried oxide layer? In our analyses we have considered as buried oxide layers SiO$_2$, diamond or AlN.

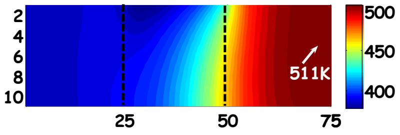
5.3.2 Simulation Results

To validate our premise that self-heating effects become smaller as we scale devices into the nanometer regime (due to more pronounced velocity overshoot effects and thinner BOX layer which leads to smaller thermal resistance), in Table 3 we present the parameters of the devices being simulated. Simulation results for the lattice temperature profile in the silicon film for devices with channel lengths from 25nm to 180 nm are shown in Figure 53. We clearly see that the hot spot is in the channel for 180 nm channel length device and moves into the drain contact for 25 nm channel length device. The current degradation for isothermal temperatures on the gate of 300K, 400K and 600K are shown in Figure 54. These are the worst case scenario results as we use Neumann boundary conditions at the side boundaries.

Table 3. Parameters for various simulated device technology nodes (constant field scaling). Parameters of the simulated structure given in **Figure 51** are: L - Gate Length; tox - Gate Oxide Thickness; t_{Si} - Active Si-Layer Thickness; t_{box} - BOX Thickness; N_{ch} - Channel Doping Concentration; I_D - Isothermal Current Value (300K).

L (nm)	tox (nm)	t_{Si} (nm)	t_{box} (nm)	N_{ch} (cm^{-3})	V_{GS}=V_{DS} (V)	I_D (mA/um)
25	2	10	50	1×10^{18}	1.2	1.82
45	2	18	60	1×10^{18}	1.2	1.41
60	2	24	80	1×10^{18}	1.2	1.14
80	2	32	100	1×10^{17}	1.5	1.78
90	2	36	120	1×10^{17}	1.5	1.67
100	2	40	140	1×10^{17}	1.5	1.57
120	3	48	160	1×10^{17}	1.8	1.37
140	3	56	180	1×10^{17}	1.8	1.23
180	3	72	200	1×10^{17}	1.8	1.03

In obtaining the results presented in Figure 53 and Figure 54, constant thermal conductivity value of 13W/m-K was used in the simulations. In Figure 55 we show the difference in the results obtained when using temperature and thickness dependent thermal conductivity value as calculated with our analytical expression for 25nm and 180 nm channel length FD-SOI devices. The use of constant thermal conductivity model with value of 13W/m-K leads to underestimation (25nm channel length) or overestimation (180nm channel length) of the average maximum temperature of the lattice in the channel region of the device. Also, as can be seen from Figure 55, the hot spot temperature is highly correlated with the model used for the thermal conductivity. To further investigate the role of the thermal conductivity model on the proper estimation of the average maximum temperature and the maximum temperature of the hotspot, we included the thermal conductivity tensor in our code. This model does not have thickness but does have temperature dependence as provided to us by Aksamija and co-workers [92] Simulation results for the lattice temperature profile in 25 nm channel length device obtained with thermal conductivity tensor and with our temperature and thickness dependent thermal conductivity model are presented in Figure 56.



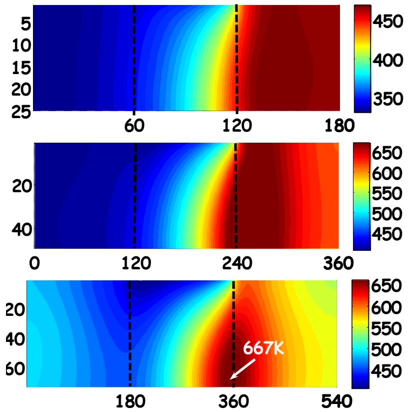


Figure 53. Lattice temperature profile in the active Si-layer for the FD SOI devices from Table 3 ranging from 25 nm (top) to180 nm (bottom) channel length.

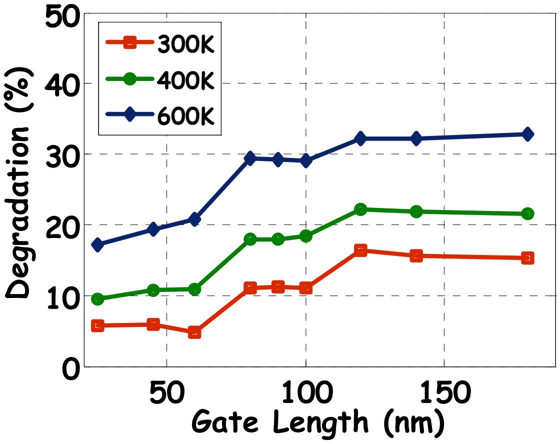


Figure 54. Current degradation versus technology generation for the FD SOI devices from Table 3 ranging from 25 nm to 180 nm channel length. Different curves correspond to different temperatures on the gate electrode (bottom set of results: T_{gate}=300 K, middle set of results: T_{gate} =400 K and top set of results: T_{gate} =600 K).

The position of the hot spot region is the same for both models but the hotspot temperature is slightly lower in the case of thermal conductivity tensor (bottom panel). For the simulation results in Figure 56, the device structure with extended domain as it shown in Figure 52 was used. Dirichlet boundary conditions (300K) at the end of the BOX and at the end of the three electrodes were assumed.

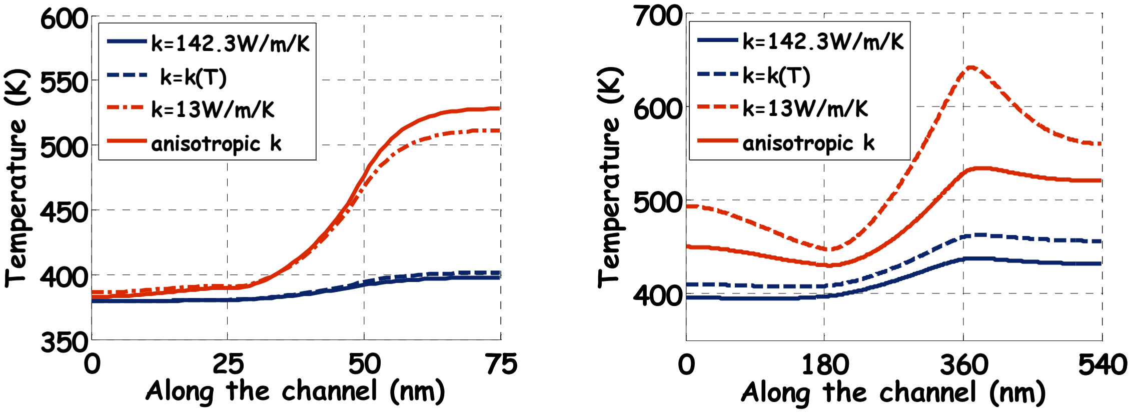


Figure 55. Average lattice temperature profile in the active Si-layer for 25 nm (left panel) and 180 nm (right panel) channel length FD-SOI device for different thermal conductivity models: κ=142.3 W/m/K is the thermal conductivity value of a bulk Si at 300K; κ=κ(T) is the temperature dependence of the thermal conductivity of bulk Si; anistropic κ is our temperature and thickness dependent thermal conductivity model.

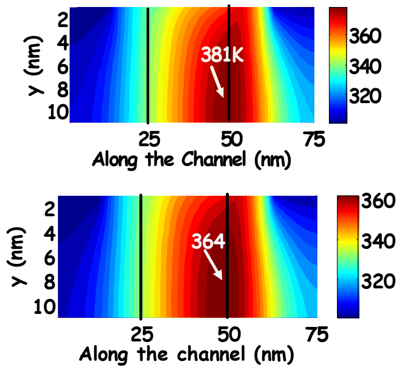


Figure 56. Lattice temperature profile in the active Si-layer for 25nm channel length FD-SOI device using our temperature and thickness dependent thermal conductivity model (top) and thermal conductivity tensor model (bottom).

Finally in Figure 57 we show the role of the choice of the BOX material on the lattice temperature in the channel for the case when the BOX is SiO_2, diamond and AlN. The corresponding current degradations are given in Ref. [93]. We do not show on the figure the BOX due to lack of space, but as can be seen in Ref. [93], even though AlN and diamond give almost the same current degradation, diamond is much better heat spreader so that from a heat removal point of view, when diamond is used as a BOX material then removal of the heat from the active region is more efficient compared to the case when one has AlN as a BOX material.

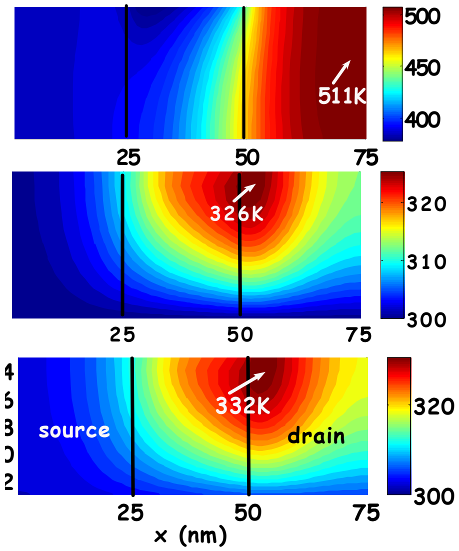


Figure 57. Lattice temperature profile (in Kelvins) in the active Si-layer for 25 nm channel length FD-SOI (top), FD-SOD (middle) and SOAlN (bottom) device from Table 3. Dirichlet boundary conditions (300K) at the end of the BOX and the gate electrode.

5.3.3. Summary of Results

In summary, in this section we have presented modeling of self-heating effects in FD SOI devices. Larger velocity overshoot and smaller BOX thickness lead to smaller degradation due to self-heating effects in 25 nm channel length FD-SOI devices. The amount of self-heating significantly depends upon the magnitude of the thermal conductivity used. Bulk values are inadequate and proper thickness and temperature dependence of the thermal conductivity must be taken into consideration. The choice of the BOX material makes significant impact on the amount of observed self-heating. Both diamond and silicon are good materials to be used as BOX. Diamond, when compared to AlN is a better heat spreader.

5.4 Modeling GaN HEMTs With Particle-Based Device Simulators

5.4.1 History of Gallium Nitride devices and their simulations

Johnson *et al.* [94] first synthesized GaN in 1928, and following this a great deal of work has been carried out in establishing that GaN is an exceedingly stable compound and exhibits significant hardness. This property along with its wider band gap made GaN an attractive candidate for device operation in higher temperature and caustic environments. In 1969, Maruska and Tietjen reported the first single crystal film of GaN on the sapphire substrate [95]. Sapphire substrates were chosen because they have a wurtzite crystal structure and are robust materials which do not react with ammonia. Maruska and Tietjen found out that the undoped GaN crystals have a very high inherent doping, typically up to 10^{19} cm^{-3} and assumed that it was due to the high density of nitrogen vacancies. However, this suggestion has created much controversy over the years. The lack of a suitable substrate material together with difficulties in obtaining p-type doping and in fabrication processing formed early bottlenecks stymieing progress. By the early 1980s, interest in GaN declined due to the above mentioned problems with crystallization and p-type doping.

In the late 1980s, Amano *et al*. reported that high quality GaN films could be obtained by a two-step process, which used an AlN buffer layer before GaN deposition [96]. This paved the way for significant improvement of both the crystal structure and electrical properties of GaN over the next few years. In 1989, the p-type doping problem was solved by post-growth low-energy electron beam irradiation treatment of Mg-doped GaN. In 1992, Nakamura *et al*. replaced this process by a post growth thermal treatment. Following this, the first current injection GaN based laser diodes using a separate confinement heterostructures were reported in 1996 by Nakamura *et al* [97].

The rapid progress of the GaN crystal growth technology has enabled its use for electronic devices especially MESFETs and HEMTs. Khan *et al.* reported the first GaN based MESFET in 1993 [98]. Subsequently, they developed an AlGaN/GaN HEMT with a transconductance G_m of 23 mS/mm and a mobility of 563 cm^2/Vs at 300 K [99]. They also reported microwave measurement results showing a cutoff frequency of 11 GHz and a maximum frequency of operation of 14 GHz. These early HEMTs exhibited poor performance in terms of transconductance and frequency response. As the crystal quality improved, the transconductance, current capacity, and frequency response increased, and presently GaN HEMTS are one of the leading candidates for high power and high frequency device applications.

There are several investigations performed on the theoretical properties of wurtzite phase GaN over the past two decades. However, most of these calculations typically used a modified mobility drift-diffusion model or the Ensemble Monte Carlo (EMC) method using several valleys with analytical band structures. The first transport simulation using Monte Carlo (MC) methods was reported by Littlejohn, Hauser, and Glisson in 1975 [100]. This simulation included a single valley (Gamma Valley) with both parabolic and non-parabolic bands. Acoustic scattering, polar optical phonon scattering, piezoelectric scattering and ionized impurity scattering were taken into account in these calculations. Velocity saturation and negative differential transconductance in GaN were predicted. In 1993, Gelmont, Kim and Shur pointed out that intervalley electron transfer played a dominant role in GaN in high electric field leading to a strongly inverted electron distribution and to a large negative differential conductance [101]. They used a non-parabolic, two valley model including Γ and U valleys. Polar optical phonon, piezoelectric, deformation potential and ionized impurity scattering mechanisms were taken into account. The intervalley coupling coefficient of GaAs was utilized in these calculations. Mansour, Kim and Littlejohn also used a two-valley model to simulate the high-temperature dependence of the electron velocity. They included acoustic phonon, polar optical phonon, intervalley phonon and ionized impurity scatterings. Bhapkar and Shur in 1997 came up with an improved multi-valley model that included a second Γ valley in addition to the Γ and U valleys [102]. The energy gap between the two valleys was modified to 2 eV from the earlier 1.5eV used in all the previous simulations. Scattering mechanism taken into account were acoustic phonon, polar optical phonon, ionized impurity, piezoelectric and inter valley scattering. Kolnik *et al.* reported the first full band MC simulation for both wurtzite and zinc-blende GaN [103]. They considered acoustic, polar optical and intervalley scattering in their calculations. Brennan *et al.* performed full band MC simulations and compared the results for different III-V materials [104]. He reported a higher electron velocity for wurtzite GaN than the previous simulation data. Both these simulations could not verify their results as no experimental velocity data was reported until then.

Barker *et al.* reported recently the measurements of the velocity-field characteristics in bulk GaN and AlGaN/GaN test structures using a pulsed I-V measurement technique [105]. These experimental results are comparable to the theoretical models of Kolnik and Brennan and Yu and Brennan. Most other simulations along these lines have reported lower velocity characteristics than that of Kolnik and Brennan. Some groups like Matulionis *et al.* have suggested that lattice heating could play a very big role in lowering the peak velocity at high electric fields [106]. Some other groups feel that this may be attributed due to the hot phonon effect. Though there seems to be no consensus about this, more work needs to be done in order to better understand the underlying physics.

AlGaN/GaN HEMTs have gained wide recognition as potential devices of choice for ultra-high-power microwave systems and power electronics. However, there are a number of issues, such as current collapse, trap memory effects, piezoelectric effects, and self-heating, where quantitative understanding is not yet achieved [107,108,109]. There have been a lot of publications dedicated to GaN HFET simulations in the recent years [110,111]. However, in almost all simulations reported, two important factors on the device characteristics have been neglected. The first such factor is the role of quantum effects occurring from the formation of a two-dimensional electron gas (2DEG) in the channel of the HFET and electron tunneling through hetero-interfaces. Second, lattice heating has been ignored when employing either the drift-diffusion, hydrodynamic or a particle based transport model in HEMT simulations. Both effects may lead to electron spreading into the bulk GaN and into the barrier which has a considerable impact on the device performance.

Yamakawa *et al.*, at Arizona State University, recently developed an accurate electron transport simulator for III-V nitride materials and devices based on the full electronic band structure and full phonon dispersion [112]. The rigid pseudo-ion model was used to obtain the full electron-phonon interaction. In addition to the scattering mechanisms used in the previous simulations, dislocation scattering was also included which plays an important role at low electric fields. In this work, Yamakawa also incorporated quantum corrections through Ferry's effective potential approach to study their effects on the device performance.

There have been quite a few reports on the self-heating simulation of the GaN-based HEMTs. Wu *et al.* reported a thermal simulation using a 2D steady-state heat conduction model [113,114]. Eastman *et al.* carried out nonlinear three-dimensional heat spreading simulations as functions of the dissipated power and the device geometry [115]. Braga *et al.* have reported 2D steady-state simulations of the GaN-based HEMTs by considering hot electron and quantum effects . Hu *et al.* performed 2D simulations of a GaN-based MOS-HEMT taking into account the piezoelectric and self-heating effects [116]. However, most of these reported thermal simulations include analytical corrections to their transport models to incorporate the self-heating effects. But in reality, one needs to solve a coupled electro-thermal system to achieve self consistent results while including the self-heating and hot phonon effects.

Lai and Majumdar developed a simple coupled thermal and electrical model for sub-micron silicon semiconductor devices comprising of the hydrodynamic equations for electron transport and energy conservation equations for different phonon modes. Their work concurrently studied both the thermal and electrical characteristics of sub-micrometer silicon semiconductor devices by considering the non-equilibrium nature of hot electrons, optical phonons, and acoustic phonons [117]. Yoder and Fichtner were the first to include thermal effects self consistently in a Monte Carlo based device simulator for Silicon MOSFETs [118]. Building upon the approach developed by Lai and Majumdar, Raleva *et al.*, at ASU, investigated the role of self-heating effects on the electrical characteristics of fully depleted Silicon on insulator (FD SOI) devices using a 2-D Monte Carlo device simulator. This electro-thermal simulator included the self-consistent solution of the energy balance equations for both acoustic and optical phonons [119]. Recently, Sridharan and Yoder studied coupled electrical and thermal transport in AlGaN/GaN HEMTs using an ensemble Monte Carlo model [120]. They observed a hot spot in the channel that is localized at low drain-source bias, but expands towards the drain at higher bias, significantly degrading channel mobility.

5.4.2 Polarization Charge

Noncentrosymmetric compound crystals exhibit two different sequences of atomic layering in the two opposing directions parallel to certain crystallographic axes, and consequently crystallographic polarity along these axes can be observed. For binary A B compounds with the wurtzite structure, the atomic layer sequence of A and B is reversed along the [0001] and the $[000\bar{1}]$ directions. The corresponding (0001) and $(000\bar{1})$ faces are referred to as the A-face and B-face, respectively. This is the case for GaN epitaxial layers and GaN-based heterostructures with the most common growth direction normal to the {0001} basal plane, where the atoms are arranged in bi-layers which consist of two closely spaced hexagonal layers, one formed by cations and the other formed by anions, leading to polar faces. Thus, in the case of GaN, a basal surface should be either Ga- or N-faced. It is, however, important to note that the (0001) and $(000\bar{1})$ surfaces of GaN are nonequivalent and differ in their chemical and physical properties [121].

The total macroscopic polarization $\mathbf{P}$ of a GaN or AlGaN layer, in the absence of any external electric field, is the sum of the spontaneous polarization $\mathbf{P}_{\mathbf{SP}}$ in the equilibrium lattice, and the strain-induced or piezoelectric polarization $\mathbf{P}_{\mathbf{PE}}$. There are some quantitative differences in the polarization for GaN and AlN owing to the sensitive dependence of the spontaneous polarization on the structural parameters. The increasing non-ideality of the crystal structure when going from GaN to AlN corresponds to an increase in spontaneous polarization. This non-ideality comes from u_0 (the anion-cation bond length along the (0001) axis in units of c) increases, c/a decreases. Since all epitaxial films and AlGaN/GaN heterostructures are grown along the [0001] axis, let us consider the two polarization components – spontaneous and piezoelectric in the Ga-faced and N-faced AlGaN/GaN heterostructures.

The spontaneous polarization along the c-axis of the wurtzite crystal is given as $P_{SP} = P_{SP}z$. The piezoelectric polarization can be calculated using the piezoelectric coefficients e_{33} and e_{31} as $P_{PE} = e_{33}\epsilon_z + e_{31}(\epsilon_x + \epsilon_y)$ where $\epsilon_z = \frac{c - c_0}{c_0}$ is the strain along the c-axis and $\epsilon_x = \epsilon_y = \frac{a - a_0}{a_0}$ are the isotropic in plane strain along the other two directions. The amount of piezoelectric polarization in the direction of the c axis can be determined by,

$$P_{PE} = 2\frac{(a - a_0)}{a_0}\left(e_{31} - \frac{c_{13}}{c_{33}}e_{33}\right) \tag{118}$$

where, C_{13} and C_{33} are elastic constants. The piezoelectric polarization of AlGaN comes out to be negative for tensile and positive for compressive strained barriers respectively. The spontaneous polarization for both GaN as well as AlN are found to be negative [122] and hence, for Ga(Al)-face heterostructures the spontaneous polarization will point towards the substrate. The alignment of the piezoelectric and spontaneous polarization is parallel in the case of tensile strain, and anti-parallel in the case of compressively strained top layers. This polarization gradient in space will result in a polarization induced charge density given by $\rho_P = \nabla P$. At an abrupt interface of an AlGaN(top)/GaN(bottom) heterostructure the polarization can decrease or increase within a bilayer, causing a net polarization sheet charge density defined by,

$$\sigma = P(Top) - P(Bottom) = [P_{SP}(Top) + P_{PE}(Top)] - [P_{SP}(Bottom) + P_{PE}(Bottom)] \tag{119}$$

The polarization induced sheet charge density is positive in pseudomorphically grown AlGaN/GaN heterostructures and free electrons will tend to compensate the polarization induced charge, thereby forming a two dimensional electron gas (2DEG) at the AlGaN/GaN interface. The following set of linear interpolations between the physical properties of GaN and AlN are utilized to calculate the net polarization induced sheet charge density σ at the AlGaN/GaN in dependence of the Aluminum mole fraction x of the $Al_xGa_{1-x}N$ barrier.

Lattice constant: $a(x) = (-0.077x + 3.189)10^{-10}m$

Elastic constants: $c_{13}(x) = (5x + 103)\,GPa$ $c_{33}(x) = (-32x + 405)GPa$

Piezoelectric constants: $e_{31}(x) = \frac{(-0.11x - 0.49)C}{m^2}$ $e_{33}(x) = \frac{(0.73x + 0.73)C}{m^2}$

Spontaneous polarization: $P_{SP}(x) = \frac{(-0.052x - 0.029)C}{m^2}$

5.4.2.1 Bias dependent Polarization Charge

The coupled formulation is based on the linear piezoelectric constitutive equations for stress and electric displacement [123],

$$\sigma_{ij} = C_{ijkl}\varepsilon_{kl} - e_{kij}E_k \tag{120}$$

$$D_i = e_{ijk}\varepsilon_{jk} - k_{ij}E_j + P_i^S \tag{121}$$

where, σ_{ij} is the stress tensor, C_{ijkl} is the fourth rank elastic stiffness tensor, ε_{kl} is the strain tensor, e_{kij} is the third ranked piezoelectric coefficient tensor, k_{ij} is the second rank permittivity tensor, D_i is the electric displacement, E_k is the electric field and P_i^S is the spontaneous polarization.

The symmetry of the wurtzite crystal structure of GaN and AlGaN reduces the number of independent elastic and piezoelectric moduli. In the pseudomorphically grown AlGaN/GaN heterostructures since the crystals are grown with the c axis normal to the surface in the z direction, one can make the common assumption that the thick GaN layer is unstrained and the biaxial strain of the thin AlGaN layer is $\epsilon_x = \epsilon_y = \frac{a - a_0}{a_0}$, isotropic in the other two dimensions to simplify the equations. The absence of stress along the growth direction (z direction), in the barrier AlGaN layer, allows us to express the areal charge concentration due to piezoelectric polarization in AlGaN as follows:

$$P_{PE}^{AlGaN} = 2\varepsilon_x\left(e_{31} - \frac{c_{13}}{c_{33}}e_{33}\right) + E_z^{AlGaN}\frac{e_{33}^2}{c_{33}} \tag{122}$$

Applying the continuity of perpendicular components of the electric displacement vector at the barrier AlGaN/GaN heterointerface, we can express the electric field in AlGaN as follows:

$$E_z^{AlGaN} = \frac{1}{k^{AlGaN} + \frac{e_{33}^2}{c_{33}}}\left[\sigma_{2D} + \Delta P^S - 2\varepsilon_x\left(e_{31} - \frac{c_{13}}{c_{33}}e_{33}\right)\right] \tag{123}$$

where, k^{AlGaN} is the permittivity of AlGaN layer, $\Delta P^S = P^S_{GaN} - P^S_{AlGaN}$ is the difference in spontaneous polarization charge of GaN and AlGaN layer. The piezoelectric polarization of AlGaN in the coupled formulation as,

$$P_{PE}^{AlGaN} = 2\varepsilon_x\left(e_{31} - \frac{c_{13}}{c_{33}}e_{33}\right)(1-\alpha) + \alpha\left(\sigma_{2D} + \Delta P^S\right), \quad \alpha = \frac{\left(\frac{e_{33}^2}{c_{33}}\right)}{\left(k^{AlGaN} + \frac{e_{33}^2}{c_{33}}\right)} \tag{124}$$

where, α is the measure of electromechanical coupling. $\alpha = 0$ corresponds for the uncoupled formulation.

When electromechanical coupling is included, the amount of stress in the c plane of the AlGaN layer is lower, resulting in lower piezoelectric polarization. Secondly, the coupling of the electric field due to spontaneous polarization has the effect of further relaxing the tensile stress of the AlGaN layer, again lowering the piezoelectric polarization.

The application of a gate bias greater than threshold voltage induces a 2DEG concentration, n_{2D}, in the channel of the HEMT. Since the piezoelectric polarization charge in the coupled formulation depends directly on the 2DEG density, any change in change in n_{2D} is accompanied by a proportional change in the piezoelectric charge in the AlGaN layer. When electromechanical coupling is neglected (in the uncoupled formulation), the polarization charge is independent of n_{2D} .When coupling is included, the piezoelectric polarization charge increases with increasing 2DEG concentration for a given Al mole fraction, moving toward the uncoupled value. An increasing 2DEG concentration corresponds to the applied gate bias becoming less negative, the compressive strain along the c axis in the barrier AlGaN layer increases [58]. As we increase the gate bias negatively and move towards the threshold voltage, the difference in net polarization charge between the two formulations increases. This reduced polarization charge in the coupled formulation will manifest as a lower drain current when compared to the drain current in the uncoupled formulation.

5.4.3 Simulation Results

5.4.3.1 Inclusion of Bias Dependent Polarization Charge in the Model

Figure 58 shows the flowchart of our in-house 2D particle-based device simulator for modeling GaN HEMTs. First, the device structure is defined and the simulation variables are initialized. The lookup scattering tables are then generated for the different regions and stored as their normalized values.

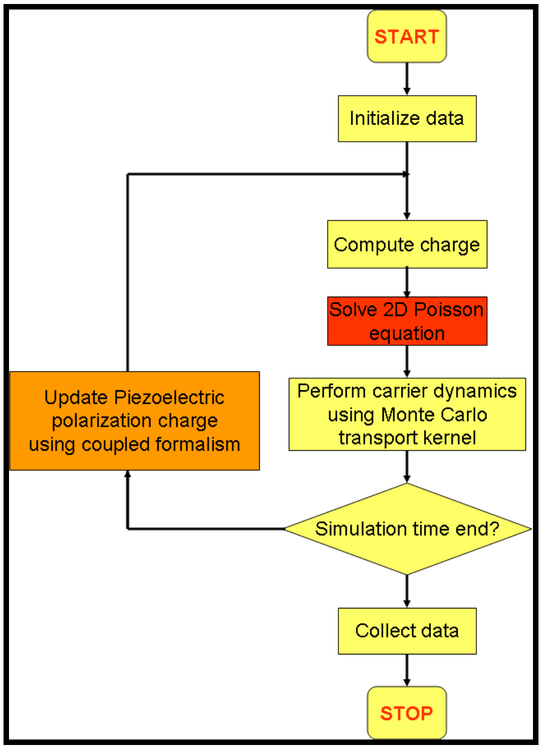

START
Initialize data
Compute charge
Solve 2D Poisson equation
Perform carrier dynamics using Monte Carlo transport kernel
Simulation time end?
Update Piezoelectric polarization charge using coupled formalism
Collect data
STOP

Figure 1:

Figure 58. Flowchart of the 2D Device Simulator for GaN HEMTs [124].

Poisson's equation is solved for the applied gate bias to obtain the equilibrium potential and electron densities. After the application of the drain bias, the EMC routine is used to simulate the non-equilibrium transport of carriers. The updated electric field from the Poisson solver redistributes carriers in the device during each time step. After all the carriers undergo a free flight – scatter procedure, the electron density is updated by counting the number of particles at each mesh, using the Nearest Element Center (NEC) charge assignment scheme. This new electron density is utilized in the Poisson solver to generate the next update in electric field. This process is done for each time step (*0.5fs*) until steady state conditions are attained. A final simulation time of *10ps* is used in these simulations after which ensemble averages like average drift velocity; average energy and current are calculated.

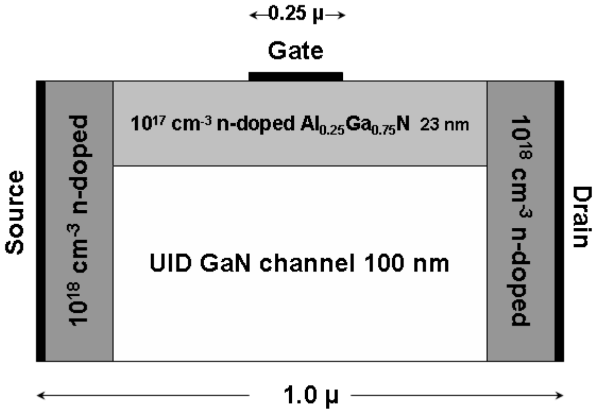

0.25 µ
Gate
10^{17} cm^{-3} n-doped $Al_{0.25}Ga_{0.75}N$ 23 nm
UID GaN channel 100 nm
Source
10^{18} cm^{-3} n-doped
10^{18} cm^{-3} n-doped
Drain
1.0 µ

Figure 2:

Figure 59. Simulated 2D AlGaN/GaN HEMT.

Simulations were performed to generate the $I_D - V_D$ and $I_D - V_G$ characteristics using the two polarization models – "uncoupled" and "coupled" formulations to study the importance of the gate bias induced strain in the heterostructure from Figure 59. A polarization induced sheet charge, $\sigma_i = 0.0165$ C/m^2 is used to fit the experimental on current [125].

Figure 60 shows the output characteristics and

Figure 61 shows the transfer characteristics comparing the two polarization models. The simulated I_D – V_D and I_D – V_G for the uncoupled formulation show good agreement with the experimental data. The simulated drain current in the coupled formulation for a certain gate and drain bias is lower than the uncoupled formulation by around 5 – 10%.

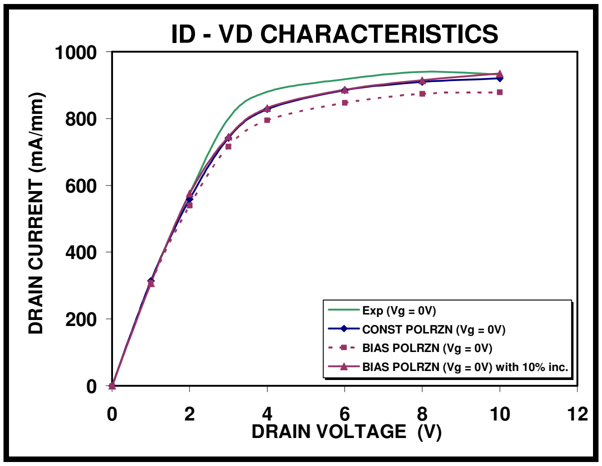


Figure 60. I_D – V_D characteristics of the two polarization models under different gate bias.

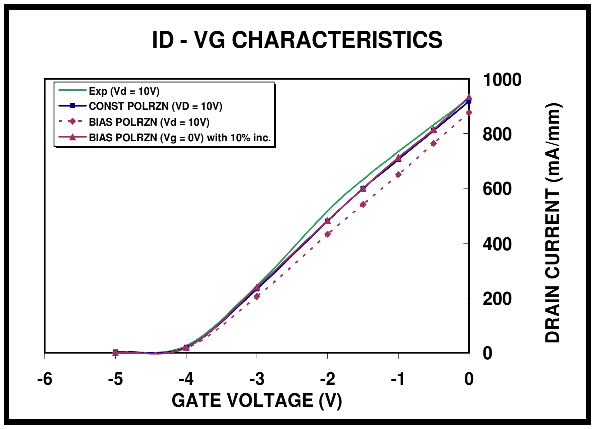


Figure 61. I_D – V_G characteristics of the two polarization models for $V_D = 10V$.

There is a trend that can be clearly seen from the output characteristics – as the gate bias is increased negatively, the change in drain current between the two formulations increases. This decrease in I_d in the coupled formulation can be directly related to a lower sheet density of the 2DEG in the coupled formulation, which is expected while using a fully electromechanically coupled solution for describing the polarization charge. Hence, the coupled formulation degrades the output characteristics as a result of the gate bias induced strain on the polarization sheet charge and this effect becomes more important as we move to larger negative biases, i.e. when we move towards the threshold voltage.

5.4.3.2 Quantum Mechanical Size Quantization Effects

Simulations were run to investigate the effect of including quantum corrections in the device characteristics. For these simulations, the coupled formulation polarization model was utilized and short range interactions were also included. The net polarization charge at the AlGaN/GaN interface was used to fit the on current. The simulations were performed with and without the effective potential correction for different gate and drain voltages. The conduction band profile along the depth is shown in Figure 62. It clearly shows that the effective potential approach tends to lower the electron density as well as introduce a charge setback effect.

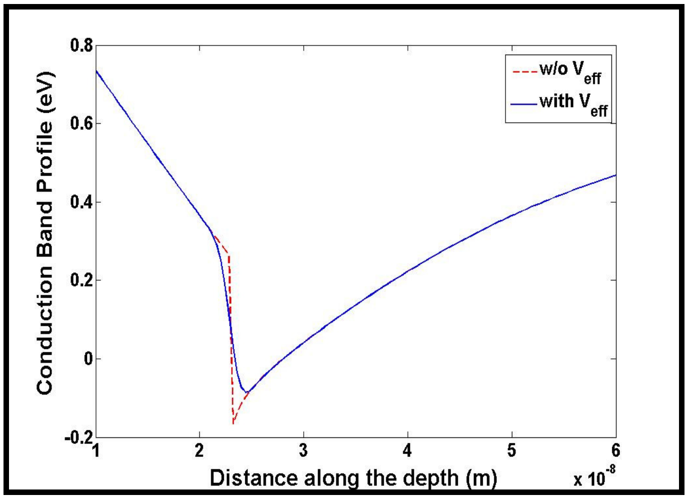


Figure 62. Conduction band profile along the depth with and without effective potential [126].

Figure 63 and Figure 64 show the I_D – V_D and I_D – V_G characteristics respectively. The I_D – V_D characteristics show that the effect of including quantum corrections via the effective potential approach is to lower the drain current. In our simulations, we have used the model with the effective potential approach as our reference in matching the drain current to experiment. Hence, the drain currents are higher for the case when the model does not include quantum corrections.

The effective potential approach lowers the electron density in addition to the charge setback effect. There is about 5 – 10% difference in the drain currents with and without the inclusion of quantum corrections for the various gate voltages. Figure 64 shows the I_D – V_G characteristics with and without the inclusion of quantum corrections. Both models tend to have a good agreement to the experimental threshold voltage although there is a small shift when the quantum corrections are included. The inset in Figure 64 shows the variation of drain current on a log scale which shows the small shift in the threshold voltage.

The average electron velocity and average electron energy along the channel for a gate bias (V_{GS}) of 0V and drain bias (V_{DS}) of 10V are shown in Figure 65 and Figure 66 respectively. The average electron drift velocity with the effective potential correction is slightly lower than the average velocity without the effective potential correction. On the other hand, when one compares the average electron energy along the channel, the model with effective potential correction has a slightly larger energy than the one without the effective potential correction.

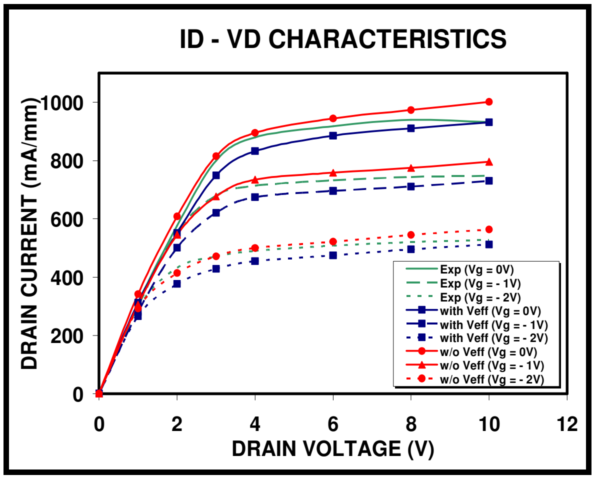


Figure 63. I_D – V_D characteristics with and without the inclusion of quantum corrections.

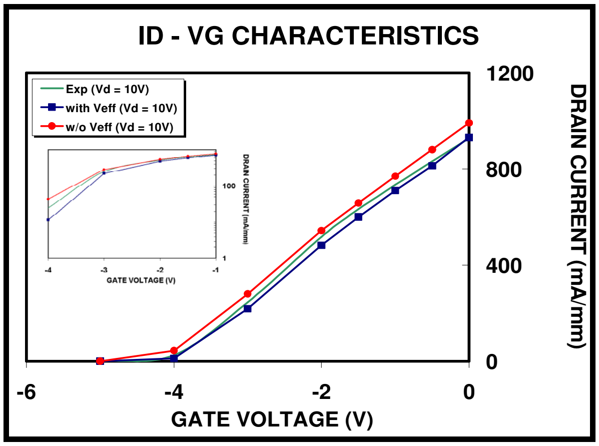


Figure 64. $I_D - V_G$ characteristics with and without the inclusion of quantum corrections.

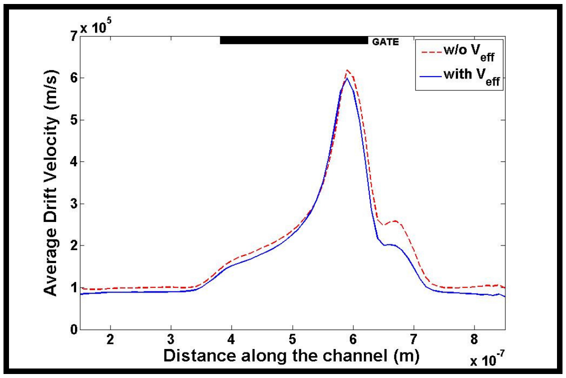


Figure 65. Average electron drift velocity along the channel with and without the effective potential at $V_{GS} = 0$ V and $V_{DS} = 10$ V.

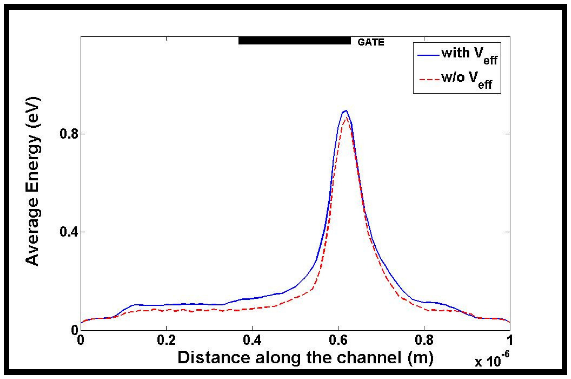


Figure 66. Average electron energy along the channel with and without the effective potential at $V_{GS} = 0$ V and $V_{DS} = 10$ V.

6. What is the future for Monte Carlo device simulations?

References

---

[1] David K. Ferry and Stephen M. Goodnick, *Transport in Nanostructures* (Cambridge Studies in Semiconductor Physics and Microelectronic Engineering, 1997).

[2] Vasileska and S. M. Goodnick, *Materials Science and Engineering, Reports: A Review: Journal*, R38, no. 5, 181 (2002).

[3] S. M. Goodnick and D. Vasileska, *Encyclopedia of Materials: Science and Technology*, Vol. 2, Ed. By K. H. J. Buschow, R. W. Cahn, M. C. Flemings, E. J. Kramer and S. Mahajan, Elsevier, New York, 1456, (2001).

[4] D. Vasileska and S. M. Goodnick, *Computational Electronics* (Morgan and Claypool, 2006).

[5] A. Schütz, S. Selberherr, H. Pötzl, *Solid-State Electronics*, Vol. 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*, Vol. ED-16, 64 (1969).

[9] K. Bløtekjær, *IEEE Trans. Electron Dev.*, Vol. 17, 38 (1970).

[10] M. V. Fischetti and S. E. Laux, “Monte Carlo Simulation of Submicron Si MOSFETs”, *Simulation of Semiconductor Devices and Processes*, vol. 3, G. Baccarani and M. Rudan Eds. (Technoprint, Bologna, 1988), 349.

[11] L. V. Keldysh, *Sov. Phys.—JETP*, Vol. 20, 1018 (1965).

[12] A. L. Fetter, 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*., Vol. 81, 7845 (1997)

[15] G. Baccarani, M. Wordeman, *Solid State Electron*., Vol. 28 , 407 (1985).

[16] S. Cordier, *Math. Mod. Meth. Appl. Sci*., Vol. 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*, Vol. 11, 455 (1964).

[19] T. M. Apostol, *Calculus, Vol. II, Multi-Variable Calculus and Linear Algebra* (Blaisdell, Waltham, MA, 1969) ch. 1.

[20] R. Straton, *Phys. Rev*., Vol. 126, 2002 (1962).

[21] T. Grasser, T.-W. Tang, H. Kosina, and S. Selberherr, *Proceedings of the IEEE*, Vol. 91, 251 (2003).

[22] M.A. Stettler, M.A. Alam, and M.S. Lundstrom, *Proceedings of the NUPAD Conference*, 97 (1992).

[23] www.silvaco.com

[24] C. Jacoboni and L. Reggiani, *Rev. Mod. Phys*., Vol. 55, 645 (1983).

[25] C. Jacoboni and P. Lugli, *The Monte Carlo Method for Semiconductor Device Simulation*, Springer-Verlag, Vienna (1989).

[26] K. Hess, *Monte Carlo Device Simulation: Full Band and Beyond*, (Kluwer Academic Publishing, Boston , 1991).

[27] M. H. Kalos and P. A. Whitlock, *Monte Carlo Methods*, (Wiley, New York, 1986).

[28] D. K. Ferry, *Semiconductors*, (Macmillan, New York, 1991).

[29] H. D. Rees, *J. Phys. Chem. Solids*, Vol. 30, 643 (1969).

[30] R. M. Yorston, *J. Comp. Phys*., Vol. 64, 177 (1986).

[31] T. Gonzalez and D. Pardo, *Solid State Electron*., 39 (1996) 555.

[32] P. A. Blakey, S. S. Cherensky and P. Sumer, *Physics of Submicron Structures*, Plenum Press, New York, (1984).

[33] T. Gonzalez and D. Pardo, *Solid-State Electron*., 39, 555 (1996).

[34] D. Vasileska and S.M. Goodnick, “Computational Electronics”, *Morgan & Claypool*, 2006.

[35] S. E. Laux, *IEEE Trans. Comp.-Aided Des. Int. Circ. Sys*., 15, 1266 (1996).

[36] S. Bosi S and C. Jacoboni, *J. Phys. C*, 9, 315 (1976).

[37] P. Lugli and D. K. Ferry, *IEEE Trans. Elec. Dev*., 32, 2431 (1985).

[38] N. Takenaka, M. Inoue and Y. Inuishi, *J. Phys. Soc. Jap*., 47, 861 (1979).

[39] S. M. Goodnick and P. Lugli, *Phys. Rev. B*, 37 (1988) 2578.

[40] M. Moško, A. Mošková and V. Cambel, *Phys. Rev. B*, 51, 16860 (1995).

[41] L. Rota, F. Rossi, S. M. Goodnick, P. Lugli, E. Molinari and W. Porod, *Phys. Rev. B*, 47,1632 (1993).

[42] R. Brunetti, C. Jacoboni, A. Matulionis and V. Dienys, *Physica B&C*, 134, 369 (1985).

[43] P. Lugli and D. K. Ferry, *Phys. Rev. Lett*., 56, 1295 (1986).

[44] J. F. Young and P. J. Kelly, *Phys. Rev. B*, 47, 6316 (1993).

[45] R. W. Hockney and J. W. Eastwood, *Computer Simulation Using Particles*, Institute of Physics Publishing, Bristol, (1988).

[46] D. J. Adams and G. S. Dubey, *J. Comp. Phys*., 72, 156 (1987).

[47] D. Vasileska, H.R. Khan, S.S. Ahmed, "Modeling Coulomb effects in nanoscale devices", Journal of Computational and Theoretical Nanoscience, Volume 5, Number 9, September 2008, pp. 1793-1827(35).

[48] Z. H. Levine, and S. G. Louie, *Phys. Rev. B*, 25, 6310 (1982).

[49] L. V. Keldysh, Zh. Eksp. *Teor. Fiz*., 37, 713 (1959).

[50] N. Sano and A. Yoshii, *Phys. Rev. B*, 45, 4171 (1992).

[51] M. Stobbe, R. Redmer and W. Schattke, *Phys. Rev. B*, 47, 4494 (1994).

[52] Y. Wang and K. Brennan, *J. Appl. Phys*., 71, 2736 (1992).

[53] M. Reigrotzki, R. Redmer, N. Fitzer, S. M. Goodnick, M. Dür, and W. Schattke, *J. Appl. Phys*., 86, 4458, (1999).

[54] R. Tsu and L. Esaki, Appl. Phys. Lett., 22, 562 (1973).

[55] J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson and A. Y. Cho, Science, 264, 553 (1994).

[56] C.-J. Sheu and S.-L. Jang, Solid-State Electron., 44, 1819 (2000).

[57] R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous and A. R. leBlanc, *IEEE J. Solid-State Circuits* 9, 256 (1974).

[58] J. R. Brews, W. Fichtner, E. H. Nicollian and S. M. Sze, *IEEE Electron Dev. Lett.* 1**,** 2 (1980).

[59] G. Bacarani and M. R. Worderman, in *Proceedings of the IEDM*, 278 (1982).

[60] M.-S. Liang, J. Y. Choi, P.-K. Ko and C. Hu, *IEEE Trans. Electron Devices* 33, 409 (1986).

[61] A. Hartstein and N. F. Albert, *Phys. Rev. B* 38, 1235 (1988).

[62] M. J. van Dort, P. H. Woerlee, A. J. Walker, C. A. H. Juffermans and H. Lifka, *IEEE Trans. Electron Dev*. 39, 932 (1992).

[63] M. J. van Dort, P. H. Woerlee and A. J. Walker, *Solid-State Electronics* 37, 411 (1994).

[64] D. Vasileska and D. K. Ferry, in the *Technical Proceedings of the First International Conference on Modeling and Simulation of Microsystems, Semiconductors, Sensors and Actuators*, 408 (1998).

[65] S. Takagi and A. Toriumi, *IEEE Trans. Electron Devices* 42, 2125 (1995).

[66] S. A. Hareland, S. Krishnamurthy, S. Jallepali, C.-F. Yeap, K. Hasnat, A. F. Tasch, Jr. and C. M. Maziar, *IEEE Trans. Electron Devices* 43, 90 (1996).

[67] D. Vasileska, D. K. Schroder and D. K. Ferry, *IEEE Trans. Electron Devices* 44, 584 (1997).

[68] K. S. Krisch, J. D. Bude and L. Manchanda, *IEEE Electron Dev. Lett*. 17, 521 (1996).

[69] Quantum hydrodynamics

[70] D. K. Ferry, Superlattices and Microstructures, 27 (2000) 61.

[71] I. Knezevic, D. Vasileska, R. Akis, J. Kang, X. He and D. K. Schroder, "Monte Carlo particle-based simulation of FIBMOS: impact of strong quantum confinement on device performance", *Physica B* Vol. 314, pp. 386-390 (2002).

[72] I. Knezevic, D. Vasileska and D. K. Ferry, "Impact of strong quantum confinement on the performance of a highly asymmetric device structure: Monte Carlo particle-based simulation of a focused-ion-beam MOSFET", *IEEE Trans. Electron Devices*, Vol. 49, pp.1019-1026, 2002.

[73] P. Feynman and H. Kleinert, Phys. Rev. A, Vol. 34, pp. 5080-5084, (1986).

[74] C. Ringhofer, C. Gardner and D. Vasileska, *Inter. J. on High Speed Electronics and Systems* **13**, 771 (2003).

[75] D. K. Ferry, *Superlattices and Microstructures*, Vol. 27, pp. 61-66, 2000.

[76] C. Ringhofer, S. Ahmed and D. Vasileska, "Effective potential approach to modeling of 25 nm MOSFET devices", *Journal of Computational Electronics*, Vol. 2, pp. 113-117 (2003).

[77] Y. Omura, S. Horiguchi, M. Tabe, and K. Kishi, *IEEE Elec. Device Lett*. 14, 569 (1993).

[78] S. M. Ramey and D. K. Ferry, *IEEE Transactions on Nanotechnology* 2, (2003).

[79] S. Hasan, J. Wang, and M. Lundstrom, *Solid–State Elect*. 48, 867 (2004).

[80] S. Datta, *Electronic Transport in Mesoscopic Systems* (Cambridge Studies in Semiconductor Physics Series, ISBN 0-521-59943-1 (1998).

[81] Lui, Wayne W. and Fukuma, Masao, J. Appl. Phys. 60, 1555 (1986).

[82] R. W. Keyes, Journal of Applied Physics, Vol. *8*, pp. *251*-259 , (1975).

[83] M. Agnostinelli, J. Hicks, J. Xu, B. Woolery, K. Mistry, K. Zhang, S. Jacobs, J. Jopling, W. Yang, B. Lee, T. Raz, M. Mehalel, P. Kolar, Y. Wang, J. Sandford, D. Pivin, C. Peterson, M. Di Battista, S. Pae, M. Jones, S. Johnson and G. Subramanian, *IEDM Technical Digest,* pp. 655-658, (2005).

[84] N. Sano, K. Matsuzawa, M. Mukai, and N. Nakayama, Microelectronics Reliability, Volume 42, Issue 2, pp. 189-199, (2002).

[85] D. K. Ferry, A. M. Kriman, M. J. Kann, and R. P. Joshi, , Computer Physics Communications, Vol. 67, pp.119-134, (1991).

[86] W. J. Gross, D. Vasileska and D. K Ferry, IEEE Electron Device Letters, Vol. 20, No. 9, pp. 463-465, (1999).

[87] N. Ashraf, D. Vasileska and Z. Ma, , Nanotech Conference and Expo, (2010).

[88] D.Vasileska, K. Raleva and S. M. Goodnick, *IEEE Transactions on Electron Devices*, Vol. 56, No. 12, pp. 3064 – 3071, (2009).

[89] W. Liu and M. Asheghi, *Appl. Phys. Lett.*, vol. 84, no. 19, pp. 3819–3821,(2004).

[90] D. Li, Y. Wu, P. Kim, L. Shi, P. Yang and A. Majumdar, *Appl. Phys. Lett*., vol. 83, pp. 2934-2936, (2003).

[91] E. H. Sondheimer, *Adv. Phys.*, vol. 1, no. 1, pp. 1–42, Jan. 1952, reprinted in Advances in Pysics, 50, pp. 499-537, (2001).

[92] Martin P, Aksamija Z, Pop E, Ravaioli U., Phys. Rev. Lett., vol. 102(12), pp. 125503, (2009).

[93] K. Raleva, D. Vasileska, S. M. Goodnick, Is SOD Technology the Solution to Heating Problems in SOI Devices?, IEEE Electron Device Letters, Volume 29, Issue 6, pp. 621 – 624, (2008).

[94] W. C. Johnson, J. B. Parsons, and M. C. Crew, J. Phys. Chem. 36, 2561 (1932).

[95] H. P. Maruska, W. C. Rhines, D. A. Stevenson, *Mat. Res. Bull.,* vol. 7, 777, (1972).

[96] H. Amano, M. Kito, K. Hiramatsu, and I. Akasaki, , *Jpn. J. Appl. Phys. Lett.,* vol. 28, L2112-2114, (1989).

[97] S. Nakamura, M. Senoh, S. I. Nagahama, N. Iwasa, T. Yamada, T. Matsushita, H. Kiyoku, Y.S. Ugimoto, *App. Phys. Lett.,* vol. 68, 2105-2107, (1996).

[98] M. A. Khan, T. N. Kuznia, A. R. Bhattaraja, D. T. Olson, *App. Phys. Lett.,* vol. 62, 1786-1787, (1993).

[99] M. A. Khan, A. R. Bhattaraja, J. N. Kuznia, D. T. Olson, *App. Phys. Lett.,* vol. 63, 1214-1215, (1993).

[100] M. A. Littlejohn, J. R. Hauser, and T. H. Glisson, *App. Phys. Lett.,* vol. 26, 625-627, (1975).

[101] B. Gelmont, K. Kim, and M. S. Shur, *Journ. Of Appl. Phys.,* vol. 74, 1818-1821, (1993).

[102] U. V. Bhapkar and M. S. Shur, *Journ. Of Appl. Phys.,* vol. 82, 1649-1655, (1997).

[103] J. Kolnik, I. H. Oguzman, K. Brennan, R. Wang, T. Fang, and P. P. Ruden, *Journ. Of Appl. Phys.,* vol. 78, 1033-1038, (1995).

[104] K. F. Brennan, E. Bellotti, M. Farahmand, J. Haralson II, P. P. Ruden, J. D. Albrecht, and A. Sutandi, *Solid-State Electronics,* vol. 44, 195-204, (2000).

[105] J. M. Barker, R. Akis, D. K. Ferry, S. M. Goodnick, T. J. Thornton, D. D. Koleske, A. E. Wickenden, and R. L. Henry,, *Physica B,* vol. 314, 39-41, (2002).

[106] A. Matulionis, R. Katilius, J. Liberis, L. Ardaravičius, L.F. Eastman, J.R. Shealy, and J. Smart, *Journ.of Appl. Phys*., vol. 92, 4490 (2002).

[107] J. A. Mittereder, S. C. Binari, P. B. Klein, J. A. Roussos, D. S. Katzer, D. F. Storm, D. D. Koleske, A. E. Wickenden, and R. L. Henry, "Current collapse induced in AlGaN/GaN high-electron-mobility transistors by bias stress", *Appl. Physc. Lett.,* vol. 83, 1650-1652, (2003).

[108] N. Braga, R. Mickevicius, R. Gaska, X. Hu, M. S. Shur, M. Asif Khan, G. Simin, and J. Yang, *Journ. Of Appl. Physc.,* vol. 95, 6409-6413, (2004).

[109] M. Dyakonov and M. S. Shur, *Journ. of Appl. Phys.* 84, 3726, (1998).

[110] Y. Kawakami, N. Kuze, J.P. Ao, and Y. Ohno, *IEICE Trans. Electron* ,E86-C, 2039,(2003).

[111] G. Verzellesi, R. Pierobon, F. Rampazzo, G. Meneghesso, A. Chini, U. K. Mishra, C. Canali, and E. Zanoni, in Proceedings of International Electron Devices Meeting, (IEEE, Piscataway, NJ), (2002).

[112] S. Yamakawa, S. M. Goodnick, S. Aboud, and M. Saraniti, *Journ. of Comp. Electronics,* vol. 3, 299-3-3, (2004).

[113] Y. F. Wu, B. P. Keller, S. Keller, D. Kapolnek, S. P. Denbaars, and U. K. Mishra, *IEEE Electron Device Lett.,* vol. 17, 455, (1996).

[114] Y. F. Wu, B. P. Keller, S. Keller, D. Kapolnek, P. Kozodoy, S. P. Denbaars, and U. K. Mishra, *Solid-State Electron.*, vol. 41, 1569, (1997).

[115] L. F. Eastman et al., *IEEE Trans. Electron Devices*, vol. 48, 479, (2001).

[116] W. D. Hu, X. S. Chen,a! Z. J. Quan, C. S. Xia, W. Lub and P. D. Ye,, *Journ. of Appl. Physc.,* vol. 100, 074501, (2006).

[117] J. Lai and A. Majumder, *J. Appl. Phys.*, vol. 79, no. 9, 7353–7361, (1996).

[118] Yoder P.D. and Fichtner W., "Simulation of Semiconductor Devices and Processes", Springer, pp. 165–168, (1998) .

[119] K. Raleva, D. Vasileska, S. M. Goodnick and M. Nedjalkov, *IEEE Trans. on Elec. Devices*, vol. 55, no. 6, 1306-1316, (2008).

[120] S. Sridharan, A. Venkatachalam, P.D. Yoder, "Electrothermal analysis of AlGaN/GaN high electron mobility transistors", *Journ. of Compt. Electronics* vol. 7, 236–239, (2008).

---

[121] O. Ambacher, J. Smart, J. R. Shealy, N. G. Weimann, K. Chu, M. Murphy, W. J. Schaff, L. F. Eastman, R. Dimitrov, L. Wittmer, M. Stutzmann, W. Reiger, and J. Hilsenbeck, *Journ. of Appl. Phys.*, vol. 87, 334-344, (2000).

[122] F. Bernardini, V. Fiorentini, and D. Vanderbilt, *Phys. Rev. B*, vol. 56, 10024, (1997).

[123] A. F. M. Anwar, R. T. Webster, and K. V. Smith, *Appl. Phys. Lett.,* vol. 88, 203510, (2006).

[124] A. Ashok, D. Vasileska, S. M. Goodnick and Olin L. Hartin, *IEEE Trans. on Elec. Devices*, vol. 56, no. 5, 998-1006, (2009).

[125] C. Lee et al., *IEEE Elect. Dev. Letters*, vol. 24, 616, (2003).

[126] A. Ashok, D. Vasileska, S. M. Goodnick and Olin L. Hartin, , *IEEE Trans. on Elec. Devices*, vol. 57, no. 3, 562-570, (2010).