Theory of the Monte Carlo Method for Semiconductor Device Simulation

Hans Kosina, *Member, IEEE*, Michail Nedjalkov, and Siegfried Selberherr, *Fellow, IEEE*

***Abstract*—A brief review of the semiclassical Monte Carlo (MC) method for semiconductor device simulation is given, covering the standard MC algorithms, variance reduction techniques, the self-consistent solution, and the physical semiconductor model. A link between physically based MC methods and the numerical method of MC integration is established. The integral representations the transient and the steady-state Boltzmann equations are presented as well as the corresponding conjugate equations. The structure of the iteration terms of the Neumann series and their evaluation by MC integration is discussed. Using this formal mathematical approach, the standard algorithms and variety of new algorithms are derived. The basic ideas of the weighted ensemble MC and the MC backward algorithms are explained.**

***Index Terms*—Boltzmann transport equation, event biasing, Monte Carlo algorithms, semiconductor device simulation, small signal analysis.**

I. Introduction

OVER the last three decades the Monte Carlo (MC) method has evolved to a reliable and frequently used tool that has been successfully employed to investigate a great variety of transport phenomena in semiconductors and semiconductor devices.

The method consists of a simulation of the motion of charge carriers in the six-dimensional (6-D) phase space formed by the position and wave vectors. Subjected to the action of an external force, the pointlike carriers follow trajectories determined by Newton's law and the carrier's dispersion relation. Due to imperfections of the crystal lattice, the drift process is interrupted by scattering events, which are considered to be local in space and instantaneous in time. The duration of a drift process and the type of scattering mechanism are selected randomly according to given probabilities that describe the microscopic process.

In principle, such a procedure yields the carrier distribution in phase space. Integrating the distribution function over some phase space volume gives a measure for the relative number of carriers in that volume. Macroscopic quantities are obtained as mean values of the corresponding single-particle quantities. Moreover, the distribution function satisfies a Boltzmann equation (BE).

The method of generating sequences of drift and scattering appears so transparent from a physical point of view that it is frequently interpreted as a direct emulation of the physical process rather than as a numerical method. The main MC algorithms used to date were originally devised from merely physical considerations, viewing an MC simulation as a simulated experiment. The proof that the used algorithms implicitly solve the BE was carried out later.

The alternative way to use the BE explicitly and to formulate MC algorithms for its solution was reported only one decade ago in the literature [1], [2]. New MC algorithms are derived which typically differ from the common ones by the fact that additional statistical variables appear, such as weights, that do not have a counterpart in the real statistical process.

Manuscript received March 20, 2000. This work was supported in part by the "Fonds zur Förderung der wissenschaftlichen Forschung," Project P13333-TEC. The review of this paper was arranged by Editor K. Hess.

The authors are with the Institute for Microelectronics, Technical University Vienna, A-1040 Vienna, Austria (e-mail: kosina@iue.tuwien.ac.at).

Publisher Item Identifier S 0018-9383(00)07785-6.

II. Application of the MC Method to Semiconductor Devices

Based upon the physical picture of the MC technique, it has been possible to apply the method to the simulation of a great variety of semiconductor devices. When the need for variance reduction techniques emerged, these have again been devised from physical considerations, such as splitting a particle into light ones by conserving the charge. For the MC simulation of devices, there are two algorithms generally employed, known as the *ensemble MC* (EMC) and the *one-particle MC* (OPMC) algorithms.

A. MC Algorithms

For a systematic description of the MC algorithms, let us begin with the homogeneous case. Transient phenomena occur when the carrier system evolves from an initial to some final distribution. Accordingly, the evolution of an ensemble of test particles is simulated starting from a given initial condition. The physical characteristics are obtained in terms of ensemble averages, giving rise to the name "ensemble" MC [3], [4]. For example, the distribution function in a given phase space point at a given time is estimated by the relative number of particles in a small volume around the point.

The EMC algorithm can also be applied under stationary conditions. In this situation, for large evolution times the final distribution approaches a steady state, and the information introduced by the initial condition is lost entirely. Alternatively, the ergodicity of the process can be exploited to replace the ensemble average by a time average. Since it is sufficient to simulate one test particle for a long period of time, the algorithm is called "one-particle" MC. The effect of the particle's initial state vanishes for long simulation times. The distribution function in a given phase space point is estimated by the time spent by the particle in a fixed, small volume around the point divided by the total time the trajectory was followed. Another method of obtaining steady-state averages has been introduced by Price [5]. With the

synchronous-ensemble or before-scattering method averages are formed by sampling the trajectory at the end of each free flight, which is in many cases easier a task than evaluating a path integral over each free flight. When the EMC algorithm is applied to the simulation of a stationary phenomenon, the steady state is reached after the initial transient has decayed. The ensemble average is taken at the end of the simulation time, while, on the contrary, with the OPMC algorithm averages are recorded during the whole time of simulation.

For the inhomogeneous situation, the mathematical model needs to be augmented by boundary conditions. A Neumann type of boundary is simply realized by reflecting the particles at the boundary. Physical models for ohmic contacts typically enforce local charge neutrality [6]. The EMC and the OPMC algorithms remain basically unchanged in the inhomogeneous case. Whenever in an OPMC simulation a particle leaves the simulation domain through an ohmic contact it is reinjected through one of the contacts, selected according to the probabilities of the underlying model.

It has been proven that the described algorithms yield a distribution function which satisfy the respective BE. Such proofs have been given by Fawcett *et al.*for the homogeneous OPMC algorithm and the steady-state BE [7], by Baccarani *et al.* for the inhomogeneous OPMC algorithm and the steady-state BE [8], and by Reklaitis [9] for the EMC algorithm and the transient BE. The latter proof can be found also in [10].

The established technique to study the small signal ac characteristics of a device consists of an EMC simulation of the transient response. Elements of the impedance matrix are obtained by applying a steplike voltage signal at a given port, and calculating the Fourier transform of the response currents through all ports of interest. Since the EMC simulation captures the general nonlinear behavior of a device, the voltage increment must be sufficiently small in order to stay in the linear response regime. An MC algorithm for the solution of the BE linearized with respect to the field would have significant advantages over the above described method. Yet, to date, such algorithms exist only for the homogeneous case. MC algorithms for small signal analysis are reported in [11]–[15]. A review can be found in [16].

B. Physical Models

The work of Kurosawa in 1966 [17] is considered as the first account of an application of the MC method to high-field transport in semiconductors. The following decade has seen considerable improvement of the method and application to a variety of materials [6]. Early papers deal with gallium arsenide [7] and germanium [18]. In the mid-1970s, a physical model of silicon was developed, capable of explaining major macroscopic transport characteristics [19], [20]. The used band structure models were represented by simple analytic expressions accounting for nonparabolicity and anisotropicity. With the increase of the energy range of interest the need for accurate, numerical band structure models arose [21]–[25]. For electrons in silicon, the most thoroughly investigated case, it is believed that only recently a satisfactory understanding of the basic scattering mechanisms at high energies has been reached, giving rise to a new "standard model" [26]. Various technologically significant semiconductors and alloys are investigated in [27].

Several attempts were made to circumvent the usage of the full, anisotropic band structure by employing simpler, analytic band models whose free parameters are fixed by best fitting a given density of states [28]–[30]. Although having such models of intermediate complexity is very desirable, a comparative study, however, revealed that they fail to reproduce the high energy part of the distribution function [31].

The interaction of the carriers with the crystal imperfections is generally considered as weak enough such as to allow a treatment by first order perturbation theory. Dominant scattering mechanisms are due to various modes of thermal lattice vibrations, ionized impurities [32], [33], and at high energies impact ionization [21], [34]–[36]. Other mechanisms, such as plasmon scattering [37], [38] and carrier–carrier scattering [39], as well as the Pauli exclusion principle [40], [41] make the transport problem strongly nonlinear.

C. Self-Consistent Coupling Schemes

In a semiconductor device, the particles move in an electric field that originates from fixed background charges and from the charge carried by the particles themselves. In a simulation a self-consistent solution of the transport problem and the Poisson equation is achieved by means of iteration. On each step the particle-mesh force calculation is carried out, consisting of the following four principal steps:

1) assign the charge to the mesh;
2) solve the field equation on the mesh;
3) calculate the mesh-defined force field;
4) interpolate to find forces on the particles [3].

Compared with OPMC, stability requirements for EMC are more severe. Problems associated with self-consistent EMC simulation are the avoidance of self-forces [3], [42] and the proper choice of the field-adjusting time step. The latter needs to be in accordance with the Nyquist theorem, where the highest frequency to be sampled is given by the plasmon frequency [22]. The high frequency of required field updates calls for fast Poisson solvers [3]. EMC is coupled with the linear Poisson equation via the particle density.

In case of OPMC, a stable, self-consistent iteration scheme is obtained by using the quasi-Fermi level and the carrier temperature from the MC simulation in the Poisson equation, which is nonlinear in this case due to the chosen variable set [43]. The so-called MC–drift-diffusion coupling technique provides another scheme for self-consistent iteration [44]. From the particle simulation transport coefficients are extracted, which are then used in a drift-diffusionlike current relation. In each iteration step the coupled set of semiconductor equations is solved, consisting of the Poisson and the carrier continuity equations [45].

In surface inversion layers or channels formed by heterojunctions, the motion of the carriers normal to the interface is quantized and thus governed by the Schrödinger equation [46]. The coupled solution of MC transport parallel to the interface, Poisson equation, and Schrödinger equation needs to be sought. Self-consistency has to be achieved for the potential, the eigenenergies, and subband populations. The strengths of the scattering mechanisms are modified through the overlap integrals [47]–[50].

D. Variance Reduction Techniques

The large variations of the carrier density within a realistic device impose severe problems upon the common MC algorithms. Statistical enhancement methods are required to reduce the variance in rarely visited phase space regions of interest. Trajectory multiplication schemes used in various MC device simulators [51], [52], [45] are extensions of the method of Phillips and Price [53]. Several variable-weight or population control techniques have been developed for the EMC method [54]–[57]. A comparison of statistical enhancement methods is given in [58].

III. MC Algorithms for the Solution of the Boltzmann Equation

The alternative way, namely, to state the transport equation first and to formulate then an MC algorithm for its solution, was reported only one decade ago [1], [2]. A link between physically based MC methods and numerical MC methods for solving integrals and series of integrals has been established.

In [1], the BE is transformed into an integral equation, which is then iteratively substituted into itself. The resulting iteration series is evaluated by a new MC technique, called *MC backward* (MCB) since the trajectories are followed back in time. All trajectories start from the chosen phase space point, and their number is freely adjustable and not controlled by the physical process. The MCB method allows the evaluation of the distribution function in a given point with a desired precision. The algorithm offers advantages when rare events have to be simulated or when the distribution function is needed only in a small phase space domain.

The *weighted ensemble MC* (WEMC) method allows the use of arbitrary probabilities for trajectory construction, such that particles can be guided toward a region of interest [59], [60]. The method evaluates the iteration series of an integral form of the BE. The unbiased estimator for the distribution function contains a product of weights which are given by the ratio of the real and the modified probabilities of the selected events.

With the iteration approach introduced in [2], the MCB and the WEMC algorithms are stated in an unified way. The concept of numerical trajectories is introduced. The common EMC algorithm is obtained as a particular case, in which the numerical trajectories coincide with the physical carrier trajectories, for homogeneous [62] and inhomogeneous [63] conditions.

A. Integral Equations

The semiclassical transport problem is described by a Fredholm integral equation of the second kind for the distribution function f.

$$f(x) = \int f(x')K(x', x)\,dx' + f_0(x). \tag{1}$$

The kernel K and the free term f_0 are given functions. The multi-dimensional variable x stands for $(\boldsymbol{k}, \boldsymbol{r}, t)$ in the transient case and for $(\boldsymbol{k}, \boldsymbol{r})$ in the steady state. The iteration

$$f^{(n+1)}(x) = \int f^{(n)}(x')K(x', x)\,dx' \tag{2}$$

with $f^{(0)}(x) = f_0(x)$ gives a formal solution to (1), known as the Neumann series [64].

$$f = f^{(0)} + f^{(1)} + f^{(2)} + \cdots \tag{3}$$

The integral equation conjugate to (1) is defined by

$$g(x') = \int g(x)K(x', x)\,dx + g_0(x'). \tag{4}$$

To evaluate a linear functional of f, one can either use directly the solution of the integral equation or as an alternative the solution of the conjugate equation due to the following equality:

$$(f, g_0) = (g, f_0). \tag{5}$$

Here, a scalar product of the form $(u, v) = \int u(x)v(x)\,dx$ is used.

The link with the MC method is established by approaching the terms of the Neumann series by MC integration. Consider the task of evaluating an integral

$$I = \int_{-\infty}^{\infty} \phi(x)\,dx = \int_{-\infty}^{\infty} p(x)\psi(x)\,dx \tag{6}$$

where ϕ is absolutely integrable. Suppose that $\phi = p\psi$, where p is nonnegative and $\int_{-\infty}^{\infty} p(x)\,dx = 1$, which means that p is a density function. Then I is the expectation value of a random variable ψ with respect to the density p: $I = E\psi$. An MC estimate for I is obtained by generating a sample $\{x_1, \cdots x_N\}$ from the density p, and taking the sample mean $\overline{\psi} = N^{-1}\sum_{i=1}^{N} \psi(x_i)$. After the central limit theorem, the sample mean $\overline{\psi}$ approaches I for $N \to \infty$.

B. Integral Form of the BE

The BE describes carrier transport by using a semiclassical approach. For device simulation the time- and position-dependent BE needs to be considered.

$$\left(\frac{\partial}{\partial t} + \boldsymbol{v}(\boldsymbol{k}) \cdot \boldsymbol{\nabla}_{\boldsymbol{r}} + \boldsymbol{F}(\boldsymbol{r}, t) \cdot \boldsymbol{\nabla}_{\boldsymbol{k}}\right) f(\boldsymbol{k}, \boldsymbol{r}, t) = Q[f](\boldsymbol{k}, \boldsymbol{r}, t). \tag{7}$$

The force field $\boldsymbol{F}$ takes into account electric and magnetic fields. If only an electric field $\boldsymbol{E}$ is present, $\boldsymbol{F}$ is given by $\boldsymbol{F}(\boldsymbol{r}, t) = q\boldsymbol{E}(\boldsymbol{r}, t)/\hbar$, where q is the charge of the carrier. The carrier's group velocity $\boldsymbol{v}$ is related to the band energy $\epsilon(\boldsymbol{k})$ by $\boldsymbol{v} = \hbar^{-1}\nabla_k\epsilon(\boldsymbol{k})$. The scattering operator $Q = Q_g - Q_l$ consists of a gain term Q_g and a loss term Q_l. In the following the nondegenerate case is considered. If the carrier concentration is small enough, the Pauli exclusion principle has a negligible effect and the scattering operator is linear.

$$Q_g[f](\boldsymbol{k}, \boldsymbol{r}, t) = \int f(\boldsymbol{k}', \boldsymbol{r}, t)S(\boldsymbol{k}', \boldsymbol{k}, \boldsymbol{r}, t)\,d\boldsymbol{k}' \tag{8}$$

$$Q_l[f](\boldsymbol{k}, \boldsymbol{r}, t) = \lambda(\boldsymbol{k}, \boldsymbol{r}, t)f(\boldsymbol{k}, \boldsymbol{r}, t). \tag{9}$$

Here, $S(\boldsymbol{k}', \boldsymbol{k}, \boldsymbol{r}, t)\,d\boldsymbol{k}$ denotes the scattering rate from a state $\boldsymbol{k}'$ to states in $d\boldsymbol{k}$ around $\boldsymbol{k}$, and $\lambda(\boldsymbol{k}, \boldsymbol{r}, t) = \int S(\boldsymbol{k}, \boldsymbol{k}', \boldsymbol{r}, t)\,d\boldsymbol{k}'$ is the total scattering rate.

The phase space trajectory $\boldsymbol{K}(\tau)$, $\boldsymbol{R}(\tau)$ is a solution of the equations of motion. Assuming the initial condition $\boldsymbol{K}(t) = \boldsymbol{k}$, $\boldsymbol{R}(t) = \boldsymbol{r}$ the formal solution can be written as

$$\begin{aligned}\boldsymbol{K}(\tau) &= \boldsymbol{k} + \int_t^\tau \boldsymbol{F}(\boldsymbol{R}(y), y)\, dy \\ \boldsymbol{R}(\tau) &= \boldsymbol{r} + \int_t^\tau \boldsymbol{v}(\boldsymbol{K}(y))\, dy. \end{aligned} \tag{10}$$

The left-hand side of (7) represents the total time derivative of $f(\boldsymbol{K}(t), \boldsymbol{R}(t), t)$. In this way, (7) can be rewritten as an ordinary first-order differential equation (11), that has the solution (12)

$$\frac{df(t)}{dt} + \lambda(t) f(t) = Q_g[f](t) \tag{11}$$

$$\begin{aligned} f(t) = &\int_{t_0}^t Q_g[f](t') \exp\left(-\int_{t'}^t \lambda(y)\, dy\right) dt' \\ &+ f(t_0) \exp\left(-\int_{t_0}^t \lambda(y)\, dy\right). \end{aligned} \tag{12}$$

$f(t_0)$ denotes the initial condition at time t_0. Remembering that in (12) every time argument, say τ, stand for $(\boldsymbol{K}(\tau), \boldsymbol{R}(\tau), \tau)$, and by setting $t_0 = 0$, one obtains the integral form of the BE.

$$\begin{aligned} f(\boldsymbol{k}, \boldsymbol{r}, t) = &\int_0^t dt' \int d\boldsymbol{k}' f(\boldsymbol{k}', \boldsymbol{R}(t'), t') \\ &\cdot S(\boldsymbol{k}', \boldsymbol{K}(t'), \boldsymbol{R}(t'), t') \\ &\cdot \exp\left(-\int_{t'}^t \lambda(\boldsymbol{K}(y), \boldsymbol{R}(y), y)\, dy\right) \\ &+ f(\boldsymbol{K}(0), \boldsymbol{R}(0), 0) \\ &\cdot \exp\left(-\int_0^t \lambda(\boldsymbol{K}(y), \boldsymbol{R}(y), y)\, dy\right). \end{aligned} \tag{13}$$

This equation represents the generalized form of Chamber's path integral [61].

C. MC Algorithms for Transient Carrier Transport

To complete the set of basic equations for the transient problem the conjugate equation has to be found. The derivation begins with a transformation of (13) into the standard form (1) by augmenting the kernel.

$$\begin{aligned} f(\boldsymbol{k}, \boldsymbol{r}, t) = &\int_0^\infty dt' \int d\boldsymbol{k}' \int d\boldsymbol{r}' f(\boldsymbol{k}', \boldsymbol{r}', t') \\ &\cdot K(\boldsymbol{k}', \boldsymbol{r}', t', \boldsymbol{k}, \boldsymbol{r}, t) + f_0(\boldsymbol{k}, \boldsymbol{r}, t) \end{aligned} \tag{14}$$

$$\begin{aligned} &K(\boldsymbol{k}', \boldsymbol{r}', t', \boldsymbol{k}, \boldsymbol{r}, t) \\ &= S(\boldsymbol{k}', \boldsymbol{K}(t'), \boldsymbol{r}', t') \exp\left(-\int_{t'}^t \lambda(\boldsymbol{K}(y), \boldsymbol{R}(y), y)\, dy\right) \\ &\cdot \delta(\boldsymbol{r}' - \boldsymbol{R}(t')) H(t - t') \end{aligned} \tag{15}$$

$$\begin{aligned} f_0(\boldsymbol{k}, \boldsymbol{r}, t) = &f_i(\boldsymbol{K}(0), \boldsymbol{R}(0)) \\ &\cdot \exp\left(-\int_0^t \lambda(\boldsymbol{K}(y), \boldsymbol{R}(y), y)\, dy\right). \end{aligned} \tag{16}$$

H denotes the unit step function and f_i the initial distribution. The integral form (13) is immediately recovered from (14) by performing the $\boldsymbol{r}'$ integration and replacing the upper bound of the time integral by t, thus eliminating the unit step function.

Using the kernel (15) in the defining equation (4) the conjugate equation evaluates to

$$\begin{aligned} g(\boldsymbol{k}', \boldsymbol{r}', t') = &\int_{t'}^\infty d\tau \int d\boldsymbol{k}_a S(\boldsymbol{k}', \boldsymbol{k}_a, \boldsymbol{r}', t') \\ &\cdot \exp\left(-\int_{t'}^\tau \lambda(\boldsymbol{K}(y), \boldsymbol{R}(y), y)\, dy\right) \\ &\cdot g(\boldsymbol{K}(\tau), \boldsymbol{R}(\tau), \tau) + g_0(\boldsymbol{k}', \boldsymbol{r}', t'). \end{aligned} \tag{17}$$

To obtain this equation, in (15) variables are changed, $\boldsymbol{k}_a = \boldsymbol{K}(t')$, $\boldsymbol{r}'' = \boldsymbol{R}(t')$. According to the Liouville theorem the volume element does not change over a trajectory such that $d\boldsymbol{k}\, d\boldsymbol{r} = d\boldsymbol{k}_a\, d\boldsymbol{r}''$. The $\boldsymbol{r}''$ integration is then carried out with the help of the δ-function.

The solution g for free term $g_0 = \delta(\boldsymbol{k} - \boldsymbol{k}')\delta(\boldsymbol{r} - \boldsymbol{r}')\delta(t - t')$ represents the Green's function of the BE. From (5), it follows that the solution f of (14) is given by the scalar product

$$\begin{aligned} f(\boldsymbol{k}, \boldsymbol{r}, t) = &\int_0^\infty dt' \int d\boldsymbol{k}' \int d\boldsymbol{r}' \\ &\cdot g(\boldsymbol{k}', \boldsymbol{r}', t'; \boldsymbol{k}, \boldsymbol{r}, t) f_0(\boldsymbol{k}', \boldsymbol{r}', t'). \end{aligned} \tag{18}$$

1) EMC Method: Assume we are interested in the integral of f over some phase space sub-domain Ω at time t:

$$\begin{aligned} f_\Omega(t) = &\int_0^\infty dt' \int d\boldsymbol{k}' \int d\boldsymbol{r}' \\ &\cdot f(\boldsymbol{k}', \boldsymbol{r}', t') \delta(t - t') \theta_\Omega(\boldsymbol{k}', \boldsymbol{r}') \end{aligned} \tag{19}$$

where θ_Ω denotes the indicator function of the subdomain. Considering this integral as scalar product $f_\Omega(t) = (f, g_0)$, it follows that $g_0 = \delta(t - t')\theta_\Omega(\boldsymbol{k}', \boldsymbol{r}')$. Using the Neumann series of (17), $g = \sum_0^\infty g^{(i)}$, we obtain

$$f_\Omega(t) = (f_0, g) = f_\Omega^{(0)}(t) + f_\Omega^{(1)}(t) + f_\Omega^{(2)}(t) + \cdots. \tag{20}$$

The meaning of the iteration terms can be understood from their structure. The second iteration term, for example, can be expressed as

$$\begin{aligned} f_\Omega^{(2)}(t) = &\int_0^t dt_2 \int_{t_2}^t dt_1 \int d\boldsymbol{k}_2^a \int d\boldsymbol{k}_1^a \int d\boldsymbol{k}_i \int d\boldsymbol{r}_i \\ &\cdot f_i(\boldsymbol{k}_i, \boldsymbol{r}_i) \exp\left(-\int_0^{t_2} \lambda(\boldsymbol{K}_2(y), \boldsymbol{R}_2(y), y)\, dy\right) \\ &\cdot S(\boldsymbol{K}_2(t_2), \boldsymbol{k}_2^a, \boldsymbol{R}_2(t_2), t_2) \\ &\cdot \exp\left(-\int_{t_2}^{t_1} \lambda(\boldsymbol{K}_1(y), \boldsymbol{R}_1(y), y)\, dy\right) \\ &\cdot S(\boldsymbol{K}_1(t_1), \boldsymbol{k}_1^a, \boldsymbol{R}_1(t_1), t_1) \\ &\cdot \exp\left(-\int_{t_1}^t \lambda(\boldsymbol{K}_0(y), \boldsymbol{R}_0(y), y)\, dy\right) \\ &\cdot \theta_\Omega(\boldsymbol{K}_0(t), \boldsymbol{R}_0(t)). \end{aligned} \tag{21}$$

Initial conditions for the $\boldsymbol{k}$-space trajectories are given first by $\boldsymbol{K}_2(0) = \boldsymbol{k}_i$ and then by the after-scattering states $\boldsymbol{K}_1(t_2) = \boldsymbol{k}_2^a$ and $\boldsymbol{K}_0(t_1) = \boldsymbol{k}_1^a$. The real space trajectory starts initially from $\boldsymbol{R}_2(0) = \boldsymbol{r}_i$ and is continuous at the time of scattering: $\boldsymbol{R}_2(t_2) = \boldsymbol{R}_1(t_2)$, $\boldsymbol{R}_1(t_1) = \boldsymbol{R}_0(t_1)$.

The iteration term (21) describes the contribution of all particles that undergo two scattering events when they propagate

from time zero to t. At time t, we find the particles on their third free flight path. Analogously, $f_\Omega^{(0)}$ represents the contribution of all particles that propagate without scattering from zero to t, $f_\Omega^{(1)}$ the contribution of all particles that propagate with one scattering event, and so forth.

The next task is to separate the integrand in (21) into a probability density p and a random variable ψ according to (6). To accomplish this task the integrand is augmented in two steps. First, the term $\exp(-\int_{t_1}^{t})$, which represents the probability that the particle drifts without scattering from t_1 to t, is expressed as an integral over the corresponding probability density. For the sake of brevity the time and position-dependence of the scattering rates is not written explicitly.

$$\exp\left(-\int_{t_1}^{t}\lambda(\boldsymbol{K}_0(y))\,dy\right) = \int_t^\infty \lambda(\boldsymbol{K}_0(t_0))\exp\left(-\int_{t_1}^{t_0}\lambda(\boldsymbol{K}_0(y))\,dy\right)\,dt_0. \quad (22)$$

Second, products of the form $\exp(-\int)S$ in (21) are multiplied by $\lambda\lambda^{-1}$. These changes yield the following expression for the second iteration term:

$$\begin{aligned} f_\Omega^{(2)} = \int_0^t dt_2 \int_{t_2}^t dt_1 \int_t^\infty dt_0 \int d\boldsymbol{k}_2^a \\ \cdot \int d\boldsymbol{k}_1^a \int d\boldsymbol{k}_i \int d\boldsymbol{r}_i \{f_i(\boldsymbol{k}_i, \boldsymbol{r}_i)\} \\ \cdot \left\{\lambda(\boldsymbol{K}_2(t_2))\exp\left(-\int_0^{t_2}\lambda(\boldsymbol{K}_2(y))\,dy\right)\right\} \\ \cdot \left\{\frac{S(\boldsymbol{K}_2(t_2), \boldsymbol{k}_2^a)}{\lambda(\boldsymbol{K}_2(t_2))}\right\} \\ \cdot \left\{\lambda(\boldsymbol{K}_1(t_1))\exp\left(-\int_{t_2}^{t_1}\lambda(\boldsymbol{K}_1(y))\,dy\right)\right\} \\ \cdot \left\{\frac{S(\boldsymbol{K}_1(t_1), \boldsymbol{k}_1^a)}{\lambda(\boldsymbol{K}_1(t_1))}\right\} \\ \cdot \left\{\lambda(\boldsymbol{K}_0(t_0))\exp\left(-\int_{t_1}^{t_0}\lambda(\boldsymbol{K}_0(y))\,dy\right)\right\} \\ \cdot \theta_\Omega(\boldsymbol{K}_0(t), \boldsymbol{R}_0(t)). \end{aligned} \quad (23)$$

Probability densities assigned to elementary events, such as generation of an initial state, of a free flight time or an after scattering state are enclosed in curly brackets for easier recognition. Note that $\lambda^{-1}S$ is the normalized distribution of the after-scattering states, since $\int \lambda^{-1}(\boldsymbol{k})S(\boldsymbol{k}, \boldsymbol{k}')d\boldsymbol{k}' = 1$ for all $\boldsymbol{k}$. The free flight time distributions are normalized on semi-infinite time intervals. For instance, for t_1 we get

$$\int_{t_2}^\infty p(t_1)\,dt_1 = \int_{t_2}^\infty \lambda(\boldsymbol{K}_1(t_1)) \cdot\exp\left(-\int_{t_2}^{t_1}\lambda(\boldsymbol{K}_1(y))\,dy\right)\,dt_1 = 1. \quad (24)$$

None of the distributions for t_0, t_1, t_2 is normalized in the time intervals given in (23) which reflects the simple fact that $f_\Omega^{(2)}$ does not represent the whole solution f_Ω but only a partial contribution.

When the multiple integrals of the iteration terms are evaluated by MC integration the well-known EMC algorithm is recovered. The separation of the integrand into p and ψ follows quite naturally from (23).

$$x = (\boldsymbol{k}_i, \boldsymbol{r}_i, t_2, \boldsymbol{k}_2^a, t_1, \boldsymbol{k}_1^a, t_0) \quad (25)$$

$$p = \{f_i\}\left\{\lambda\exp\left(-\int\right)\right\}\{\lambda^{-1}S\}\left\{\lambda\exp\left(-\int\right)\right\} \cdot \{\lambda^{-1}S\}\left\{\lambda\exp\left(-\int\right)\right\} \quad (26)$$

$$\psi = \theta_\Omega(\boldsymbol{K}_0(t), \boldsymbol{R}_0(t)). \quad (27)$$

To evaluate (23) by MC integration one has to generate N realizations of the multidimensional variable x, which are referred to as numerical trajectories. The factors in (26) denote conditional probability densities, except f_i, which is unconditional. Therefore, one would first generate a phase space point $(\boldsymbol{k}_i, \boldsymbol{r}_i)$ from the initial distribution f_i, then choose t_2 from the free flight time distribution, select $\boldsymbol{k}_2^a$ with density $\lambda^{-1}S$, and so forth. Finally, at time t the indicator function θ_Ω, which plays the role of ψ in (6), needs to be evaluated. The result will be simply one or zero. Doing so for N trajectories corresponds to counting the number of particles found in Ω at time t.

The generated times are of ascending order, $0 < t_2 < t_1 < t_0$, which means that the trajectory is followed forward in time. With a forward algorithm it is only possible to evaluate an average of the distribution function over some subdomain, but not the exact value in a given point.

2) Weighted EMC Method: In the WEMC method, instead of the physical densities in (23), which follow in a natural way from the kernel, arbitrary densities are used for numerical trajectory construction. The ratio of the physical density over the numerical density determines the weight of the numerical trajectory. The basic ideas can be explained by rewriting the random variable x, the density p, and the dependent random variable $\psi(x)$ for the nth iteration term in a formal way as

$$x = (x_0, x_1, \cdots, x_n) \quad (28)$$

$$p = f_0(x_0)K(x_0, x_1)\cdots K(x_{n-1}, x_n) \quad (29)$$

$$\psi = \theta_\Omega \quad (30)$$

where K stands for the kernel of the conjugate equation. One can now choose an arbitrary initial distribution p_0 and arbitrary transition probabilities P for numerical trajectory construction. Since the product $p\psi$ has to remain unchanged, the random variable ψ has to compensate for the changes in the density p.

$$p = p_0(x_0)P(x_0, x_1)\cdots P(x_{n-1}, x_n) \quad (31)$$

$$\psi = \frac{f_0(x_0)K(x_0, x_1)\cdots K(x_{n-1}, x_n)}{p_0(x_0)P(x_0, x_1)\cdots P(x_{n-1}, x_n)}\,\theta_\Omega. \quad (32)$$

The numerical initial distribution p_0 and the numerical transition probability P have to be nonzero where the physical counterparts are nonzero, i.e., $p_0(x_0) \neq 0$ if $f_0(x_0) \neq 0$ and $P(x_i, x_j) \neq 0$ if $K(x_i, x_j) \neq 0$. Furthermore, only normalized densities are considered, $\int p_0(x_0)\,dx_0 = 1$ and $\int P(x_i, x_j)\,dx_j = 1$ for all x_i.

Consequently, whenever in the process of numerical trajectory construction a random variable is selected from a numerical density rather than from a physical density, the weight of the trajectory changes by the ratio of the two densities.

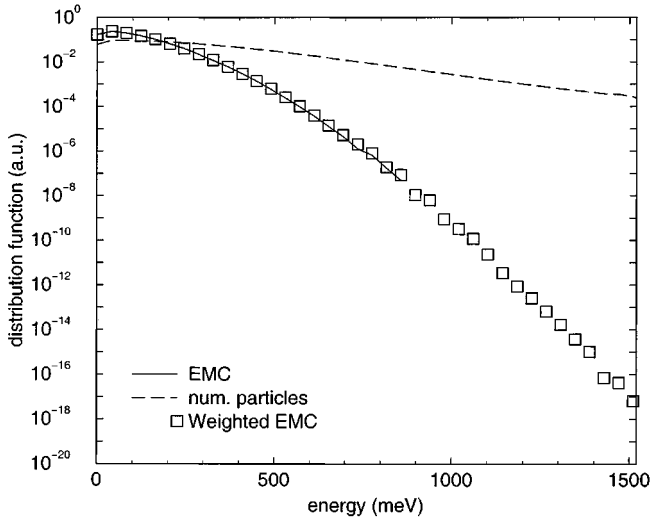


Fig. 1. Electron energy distribution functions obtained by the EMC and WEMC algorithms for $E = 30$ kV/cm.

As an example, the WEMC method is applied to compute the energy distribution of electrons in Si. The used semiconductor model accounts for the phonon spectrum described in [6] and for an analytical, nonparabolic band-structure characterized by $m_n^* = 0.32m_0$, $\alpha = 0.5\ \mathrm{eV}^{-1}$. To increase the probability for electrons to gain energy and thus to populate the high energy tail, the probability for phonon absorption has been increased at the expense of phonon emission. Fig. 1 shows the result of a simulation of $4 \cdot 10^7$ electrons at $E = 30$ kV/cm for $t = 1$ ps. The initial distribution is a Maxwellian at lattice temperature, chosen as $T_L = 300$K. For the given particle number, the EMC method can resolve not more than seven decades of the energy distribution, while the WEMC method gives reasonable accurate results within 17 decades. However, it is to note that the variance of the WEMC method increases with increasing evolution time. The dashed line in Fig. 1 represents the distribution of the endpoints of the numerical trajectories.

3) MCB Method: In the previous sections, forward algorithms were formally derived from the Neumann series of the conjugate equation. If the Neumann series of the integral form of the BE is approached by the MC method, backward algorithms will be obtained.

As an instructive example, we consider the term $f^{(2)}$ of the Neumann series of (13).

$$\begin{aligned} f^{(2)}(\boldsymbol{k}, \boldsymbol{r}, t) = \int_0^t dt_1 \int d\boldsymbol{k}_1 \int_0^{t_1} dt_2 \int d\boldsymbol{k}_2 \\ \cdot f_i(\boldsymbol{K}_2(0), \boldsymbol{R}_2(0)) \\ \cdot \exp\left(-\int_0^{t_2} \lambda(\boldsymbol{K}_2(y), \boldsymbol{R}_2(y), y)\, dy\right) \\ \cdot S(\boldsymbol{k}_2, \boldsymbol{K}_1(t_2), \boldsymbol{R}_1(t_2), t_2) \\ \cdot \exp\left(-\int_{t_2}^{t_1} \lambda(\boldsymbol{K}_1(y), \boldsymbol{R}_1(y), y)\, dy\right) \\ \cdot S(\boldsymbol{k}_1, \boldsymbol{K}_0(t_1), \boldsymbol{R}_0(t_1), t_1) \\ \cdot \exp\left(-\int_{t_1}^{t} \lambda(\boldsymbol{K}_0(y), \boldsymbol{R}_0(y), y)\, dy\right). \end{aligned} \quad (33)$$

Final conditions for the $\boldsymbol{k}$-space trajectories are given first by $\boldsymbol{K}_0(t) = \boldsymbol{k}$ and then by the before-scattering states $\boldsymbol{K}_1(t_1) = \boldsymbol{k}_1$ and $\boldsymbol{K}_2(t_2) = \boldsymbol{k}_2$. The real space trajectory ends at final time t in the given point $\boldsymbol{R}_0(t) = \boldsymbol{r}$ and is continuous at the time of scattering: $\boldsymbol{R}_1(t_1) = \boldsymbol{R}_0(t_1)$, $\boldsymbol{R}_2(t_2) = \boldsymbol{R}_1(t_2)$.

As in the forward case, the integrand of (33) is augmented in two steps. The probability $\exp(-\int_0^{t_2})$ is expressed as an integral over the corresponding density.

$$\begin{aligned} &\exp\left(-\int_0^{t_2} \lambda(\boldsymbol{K}_2(y))\, dy\right) \\ &\quad = \int_{-\infty}^{0} \lambda(\boldsymbol{K}_2(t_3)) \exp\left(-\int_{t_3}^{t_2} \lambda(\boldsymbol{K}_2(y))\, dy\right) dt_3. \end{aligned} \quad (34)$$

Second, the normalization of the distribution of the before-scattering states has to be introduced.

$$\lambda^*(\boldsymbol{k}) = \int S(\boldsymbol{k}', \boldsymbol{k})\, d\boldsymbol{k}' \quad (35)$$

From (35), it follows that $\int \lambda^*(\boldsymbol{k})^{-1} S(\boldsymbol{k}', \boldsymbol{k})\, d\boldsymbol{k}' = 1$ for all $\boldsymbol{k}$. Products of the form $S\exp(-\int)$ in (33) are augmented using both λ and λ^*, as shown in the following expression:

$$\begin{aligned} f^{(2)}(\boldsymbol{k}, \boldsymbol{r}, t) = \int_0^t dt_1 \int_0^{t_1} dt_2 \int_{-\infty}^{0} dt_3 \\ \cdot \int d\boldsymbol{k}_1 \int d\boldsymbol{k}_2 f_i(\boldsymbol{K}_2(0), \boldsymbol{R}_2(0)) \\ \cdot \left\{\lambda(\boldsymbol{K}_2(t_3)) \exp\left(-\int_{t_3}^{t_2} \lambda(\boldsymbol{K}_2(y))\, dy\right)\right\} \\ \cdot \frac{\lambda^*(\boldsymbol{K}_1(t_2))}{\lambda(\boldsymbol{K}_1(t_2))} \left\{\frac{S(\boldsymbol{k}_2, \boldsymbol{K}_1(t_2))}{\lambda^*(\boldsymbol{K}_1(t_2))}\right\} \\ \cdot \left\{\lambda(\boldsymbol{K}_1(t_2)) \exp\left(-\int_{t_2}^{t_1} \lambda(\boldsymbol{K}_1(y))\, dy\right)\right\} \\ \cdot \frac{\lambda^*(\boldsymbol{K}_0(t_1))}{\lambda(\boldsymbol{K}_0(t_1))} \left\{\frac{S(\boldsymbol{k}_1, \boldsymbol{K}_0(t_1))}{\lambda^*(\boldsymbol{K}_0(t_1))}\right\} \\ \cdot \left\{\lambda(\boldsymbol{K}_0(t_1)) \exp\left(-\int_{t_1}^{t} \lambda(\boldsymbol{K}_0(y))\, dy\right)\right\}. \end{aligned} \quad (36)$$

Writing the position and time-dependence of the scattering rates has been omitted for the sake of brevity.

To evaluate (36) by MC integration the integrand is separated into p and ψ.

$$x = (t_1, \boldsymbol{k}_1, t_2, \boldsymbol{k}_2, t_3) \quad (37)$$

$$\begin{aligned} p = &\left\{\lambda \exp\left(-\int\right)\right\} \{S/\lambda^*\} \left\{\lambda \exp\left(-\int\right)\right\} \\ &\cdot \{S/\lambda^*\} \left\{\lambda \exp\left(-\int\right)\right\} \end{aligned} \quad (38)$$

$$\psi = \frac{\lambda^*(t_1)}{\lambda(t_1)} \frac{\lambda^*(t_2)}{\lambda(t_2)} f_i. \quad (39)$$

N realizations of the multidimensional variable x have to be generated. Since $\boldsymbol{k}$, $\boldsymbol{r}$, t are given, the construction of the numerical trajectory starts at this point by choosing a random variable t_1, which obviously is less than t. The next random variable to be chosen is $\boldsymbol{k}_1$, a before-scattering state, and so forth. The selected times are of descending order, $t < t_1 < t_2 < t_3$, which

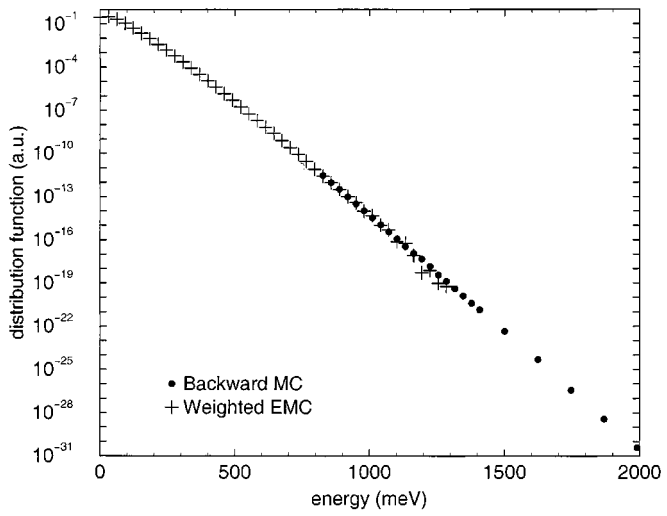


Fig. 2. Electron energy distribution functions obtained by the WEMC and MCB algorithms for $E = 10$ kV/cm.

means that the numerical trajectory is followed back in time. During trajectory construction the product $\prod \lambda^*(t_i)/\lambda(t_i)$ has to be recorded. At time zero, the product is multiplied by the initial distribution f_i evaluated at the reached phase space point to give the random variable ψ. After construction of N numerical trajectories, the sample mean of ψ is formed.

$$f(\boldsymbol{k}, \boldsymbol{r}, t) \approx \frac{1}{N} \sum_{j=1}^{N} \psi_j. \tag{40}$$

The MCB method allows to evaluate the distribution function at given phase space points with desired accuracy.

In analogy with the forward case event biasing can be applied leading to weighted backward algorithms.

Fig. 2 compares the energy distributions of electrons in Si as computed by the MCB and the WEMC methods. Conditions assumed are $E = 10$ kV/cm and $t = 3$ ps. The initial distribution and the number of particles for the WEMC simulation are as in Fig. 1. The MCB method is used to evaluate the energy distribution at discrete points above 800 meV. The statistical uncertainty of the result is controlled by the number of numerical trajectories starting from each point. In the simulation 10^7 backward trajectories are computed for each point. Using the MCB method the high energy tail is obtained with high precision, as shown in Fig. 2. The depicted range of 30 decades is out of reach even for the here considered variant of WEMC method.

D. MC Algorithms for Steady-State Carrier Transport

In the steady state the Boltzmann equation is given by

$$(\boldsymbol{v}(\boldsymbol{k}) \cdot \boldsymbol{\nabla}_{\boldsymbol{r}} + \boldsymbol{F}(\boldsymbol{r}) \cdot \boldsymbol{\nabla}_{\boldsymbol{k}}) f(\boldsymbol{k}, \boldsymbol{r}) = Q[f](\boldsymbol{k}, \boldsymbol{r}). \tag{41}$$

The applied field and all material properties are independent of time rendering the system time invariant. Equation (41) describes an open system that exchanges particles with one or more reservoirs. The latter provide a stationary boundary condition for the distribution function.

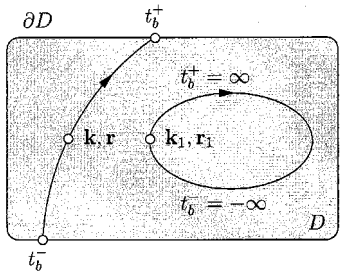


Fig. 3. Illustration of the functions $t_b^-(\boldsymbol{k}, \boldsymbol{r})$ and $t_b^+(\boldsymbol{k}, \boldsymbol{r})$ which give the time at a trajectory's entry point and exit point, respectively. If $\boldsymbol{k}_1, \boldsymbol{r}_1$ is the initial point of a closed trajectory, the times are $t_b^\pm(\boldsymbol{k}_1, \boldsymbol{r}_1) = \pm\infty$.

The integral form of (41) is derived by using the same approach as in the transient case. The analysis is carried out for a phase space trajectory $\boldsymbol{K}(\tau)$, $\boldsymbol{R}(\tau)$ with the initial conditions $\boldsymbol{K}(0) = \boldsymbol{k}$, $\boldsymbol{R}(0) = \boldsymbol{r}$, where $\boldsymbol{k}, \boldsymbol{r}$ is a given point at which the value of f is sought. In the steady state, the lower bound of the time integration in (12) is chosen as that time at which the considered trajectory enters the simulation domain. This treatment ensures that the given boundary distribution f_b appears in the free term. The resulting integral equation is posed in the 6-D phase space. The real space coordinate is restricted to the simulation domain D.

$$f(\boldsymbol{k}, \boldsymbol{r}) = \int d\boldsymbol{k}' \int d\boldsymbol{r}' f(\boldsymbol{k}', \boldsymbol{r}') K(\boldsymbol{k}', \boldsymbol{r}', \boldsymbol{k}, \boldsymbol{r}) + f_0(\boldsymbol{k}, \boldsymbol{r}). \tag{42}$$

For the kernel K and the free term f_0 the following expressions are found:

$$\begin{aligned} K(\boldsymbol{k}', \boldsymbol{r}', \boldsymbol{k}, \boldsymbol{r}) &= \int_{t_b^-(\boldsymbol{k},\boldsymbol{r})}^{0} dt' S(\boldsymbol{k}', \boldsymbol{K}(t'), \boldsymbol{r}') \\ &\quad \cdot \exp\left(- \int_{t'}^{0} \lambda(\boldsymbol{K}(y), \boldsymbol{R}(y))\, dy \right) \\ &\quad \cdot \delta(\boldsymbol{r}' - \boldsymbol{R}(t')) \theta_D(\boldsymbol{r}') \end{aligned} \tag{43}$$

$$\begin{aligned} f_0(\boldsymbol{k}, \boldsymbol{r}) = {} & f_b(\boldsymbol{K}(t_b^-), \boldsymbol{R}(t_b^-)) \\ & \cdot \exp\left(- \int_{t_b^-(\boldsymbol{k},\boldsymbol{r})}^{0} \lambda(\boldsymbol{K}(y), \boldsymbol{R}(y))\, dy \right). \end{aligned} \tag{44}$$

Since the time integral, which originates from the solution of the differential equation (11), cannot stay in the integral equation (42), it has to be assigned to the kernel (43).

Two functions $t_b^-(\boldsymbol{k}, \boldsymbol{r})$ and $t_b^+(\boldsymbol{k}, \boldsymbol{r})$ are introduced, denoting the times when the trajectory enters, respectively leaves, the simulation domain (see Fig. 3). If the considered point is the initial point of a closed trajectory, the times are $t_b^\pm = \pm\infty$. This means that a particle moving on such a trajectory has been scattered onto this trajectory at some time between $(-\infty, 0)$. The particle can leave the closed trajectory again only by scattering at some time between $(0, \infty)$.

For the formal derivation of the steady-state forward algorithms, the conjugate equation is required. This equation results from (42) by applying similar steps as in the transient case.

$$g(\boldsymbol{k}', \boldsymbol{r}') = \int d\boldsymbol{k}_a \int_0^{t_b^+(\boldsymbol{k}_a,\boldsymbol{r}')} d\tau S(\boldsymbol{k}', \boldsymbol{k}_a, \boldsymbol{r}') \cdot \exp\left(-\int_0^\tau \lambda(\boldsymbol{K}(y), \boldsymbol{R}(y))\, dy\right) g(\boldsymbol{K}(\tau), \boldsymbol{R}(\tau)) + g_0(\boldsymbol{k}', \boldsymbol{r}').$$

New algorithms for the steady state, such as the weighted and the backward OPMC algorithms, are currently under investigation are not yet reported in the literature.

1) OPMC Algorithm: The formal derivation of the OPMC algorithm is based on the Neumann series of the conjugate equation. Due to the boundary condition, the transformations involved in the derivation of the iteration terms are more complicated than those involved in the pure initial value problem. One result of such a calculation is that the initial points of the trajectories have to be generated with the velocity-weighted boundary distribution, $-v_\perp H(-v_\perp) f_b$, where $v_\perp = \boldsymbol{v} \cdot \boldsymbol{n}$ is the component of the velocity normal to the boundary. The total incident flux appears as a normalization constant.

$$\Gamma_D = -\oint_{\partial D} d\sigma(\boldsymbol{r}) \int_{v_\perp(\boldsymbol{k})<0} d\boldsymbol{k} v_\perp(\boldsymbol{k}) f_b(\boldsymbol{k}, \boldsymbol{r}). \tag{45}$$

Assume we are interested in the average of f over some subdomain Ω.

$$f_\Omega = \int d\boldsymbol{k} \int d\boldsymbol{r} f(\boldsymbol{k}, \boldsymbol{r}) \theta_\Omega(\boldsymbol{k}, \boldsymbol{r}). \tag{46}$$

For example, Ω might denote a cell of the spatial mesh. The iteration terms can be formulated in two slightly different ways leading to the before-scattering and the time recording methods for average recording, given by (47) and (48), respectively.

$$f_\Omega = \Gamma_D \sum_{j=1}^{\infty} \frac{\theta_\Omega(\boldsymbol{k}_{bj}, \boldsymbol{r}_j)}{\lambda(\boldsymbol{k}_{bj}, \boldsymbol{r}_j)} \tag{47}$$

$$f_\Omega = \Gamma_D \sum_{j=1}^{\infty} \int_0^{t_j} \theta_\Omega(\boldsymbol{K}_j(\tau), \boldsymbol{R}_j(\tau))\, d\tau. \tag{48}$$

In the simulation, an initial state is generated from the velocity-weighted boundary distribution [55], [65], and the trajectory is followed through the device. When an absorbing boundary is encountered another initial state is generated and a new trajectory is followed. In (47) $\boldsymbol{k}_{bj}$ denotes the before-scattering state and $\boldsymbol{r}_j$ the position of the particle when the jth scattering event takes place. In (48) t_j stands for the duration and $\boldsymbol{K}_j$, $\boldsymbol{R}_j$ for the phase space trajectory of the jth free flight. The sum (48) simply gives the time the particle has spent in Ω. The counter j is incremented whenever a scattering event is processed, regardless of the trajectory count. This is possible since in the OPMC algorithm all trajectories have equal weight.

The normalization constant needs not be evaluated from the theoretical definition (45). The normalization of f is given by the total number of particles or by the amount of mobile charge in the device. For example, for $\Omega \to D$ one obtains $f_D = N = \Gamma_D \sum_1^\infty \lambda^{-1}(\boldsymbol{k}_{bj}, \boldsymbol{r}_j)$, which gives Γ_D as function of N and the recorded sum. The finite sum recorded during the simulation is an unbiased estimate of the total time the particle path is followed.

As in the transient case event biasing can be used in the steady state as well leading to weighted OPMC algorithms.

2) One-Particle Backward MC Algorithm: Approaching the Neumann series of (42) by the MC method yields steady-state backward algorithms. Two algorithms are found. The first one allows to evaluate the distribution function at given points and is basically identical with the transient backward algorithm (Section III-C-3) with the only difference that the numerical trajectory is followed in a variable time interval $(t_b^-(\boldsymbol{k}, \boldsymbol{r}), 0)$ rather than in a predefined one.

Let us consider the problem of injection of channel hot carriers into the gate oxide. Using a backward algorithm, carriers are launched at the semiconductor/oxide interface only at energies above the relevant energy threshold. In other words, only the rare events are simulated. Each high energetic carrier is followed back in time until it reaches an equilibrium region such as source or drain, where the distribution function is known. Since each trajectory is of different duration, a steady-state formulation employing a variable time $t_b^-(\boldsymbol{k}, \boldsymbol{r})$ and a boundary distribution appears appropriate.

The second algorithm can be viewed as the backward version of OPMC algorithm (Section III-D-1). N Trajectories are constructed starting from an absorbing boundary. The weight of the particle changes by λ^*/λ at each scattering event, however, the weight remains undeterminated with respect to a scaling factor. Both the before-scattering and the time-recording method for average recording are available, yet the current weight of the particle has to be taken into account. The particle weights are finally determined when the trajectory terminates at an injecting boundary, where the boundary distribution f_b evaluated at the reached phase space point gives the scaling factor.

E. Small-Signal MC Algorithms

Understanding the MC method as a versatile tool to solve integral equations enables its application to a class of problems which are not accessible by purely physically based, imitative MC methods. In electrical engineering the linear small-signal analysis of nonlinear systems plays an important role. Whether the linearized system is analyzed in the frequency or time domain is just a matter of convenience since the system responses obtained in either domain are linked by the Fourier transform.

At present, linear small-signal analysis of semiconductor devices by the MC method is beyond the state of the art. However, recently progress has been made in performing MC small signal analysis of bulk carrier transport [15]. Choosing a formulation in the time domain, a small perturbation $\boldsymbol{E}_1$ is superimposed to a stationary field $\boldsymbol{E}_s$. The stationary distribution function f_s will thus be perturbed by some small quantity f_1.

$$\boldsymbol{E}(t) = \boldsymbol{E}_s + \boldsymbol{E}_1(t) \tag{49}$$

$$f(\boldsymbol{k}, t) = f_s(\boldsymbol{k}) + f_1(\boldsymbol{k}, t) \tag{50}$$

Inserting this Ansatz into the transient BE and retaining only first order perturbation terms yield a Boltzmann-like equation for f_1 which is linear in the perturbation $\boldsymbol{E}_1$, as follows:

$$\frac{\partial f_1(\boldsymbol{k},t)}{\partial t} + \frac{q}{\hbar}\boldsymbol{E}_s \cdot \nabla f_1(\boldsymbol{k},t) = Q[f_1](\boldsymbol{k},t) - \frac{q}{\hbar}\boldsymbol{E}_1(t)\cdot\nabla f_s(\boldsymbol{k}). \tag{51}$$

Compared with the common BE, (51) has an additional term on the right-hand side, which contains f_s, the solution of the stationary BE. The integral form of (51) is found by the method pointed out in Section III-B. Assuming an impulselike excitation $\boldsymbol{E}_1(t) = \delta(t)\boldsymbol{E}_{\rm im}$ results in the following integral equation for the impulse response f_1 [15]:

$$f_1(\boldsymbol{k},t) = \int_0^t dt' \int d\boldsymbol{k}' f_1(\boldsymbol{k}',t')S(\boldsymbol{k}',\boldsymbol{K}(t')) \cdot \exp\left(-\int_{t'}^t \lambda(\boldsymbol{K}(y)\,dy\right) + G(\boldsymbol{K}(0)) \cdot \exp\left(-\int_0^t \lambda(\boldsymbol{K}(y)\,dy\right) \tag{52}$$

$$G(\boldsymbol{k}) = -\frac{q}{\hbar}\boldsymbol{E}_{\rm im}\cdot\nabla f_s(\boldsymbol{k}). \tag{53}$$

The free term of (52) is formally equivalent to (16), the free term of the BE. The only difference is that G takes on also negative values, and can therefore not be interpreted as an initial distribution. In [15] G is expressed as a difference of two positive functions, $G = G^+ - G^-$, an Ansatz which decomposes (52) into two common BE for the unknowns f_1^+ and f_1^-. The initial conditions of this BE are s $f_1^{\pm}(\boldsymbol{k},0) = G^{\pm}(\boldsymbol{k}) \geq 0$. In this way, the impulse response is understood in terms of the concurrent evolution of two carrier ensembles.

Using different methods to generate the initial distributions of the two ensembles gives rise to a variety of MC algorithms. Both existing and new MC algorithms are obtained in a unified way, and a transparent, physical interpretation of the algorithms is supported.

For electrons in Si the impulse response of mean energy and mean velocity has been calculated. Fig. 4 shows the response of the differential energy $\langle\epsilon\rangle_1/E_{\rm im}$ in the time domain for different field strengths. The frequency-dependent differential velocity obtained by a Fourier transform of the impulse response $\langle v\rangle_1/E_{\rm im}$ is plotted in Figs. 5 and 6. The low frequency limit of the real part tends to the static differential mobility $\partial\langle v\rangle_s/\partial E_s$.

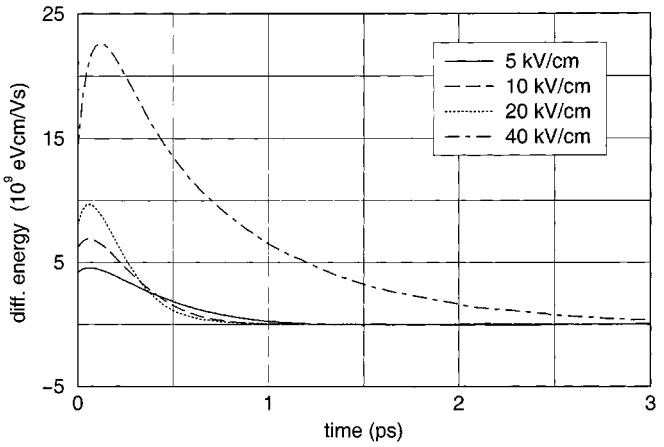


Fig. 4. Impulse response of the differential energy.

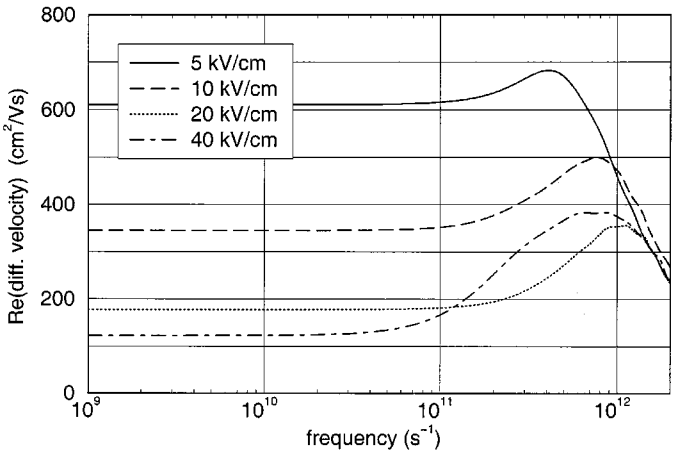


Fig. 5. Real part of the differential velocity.

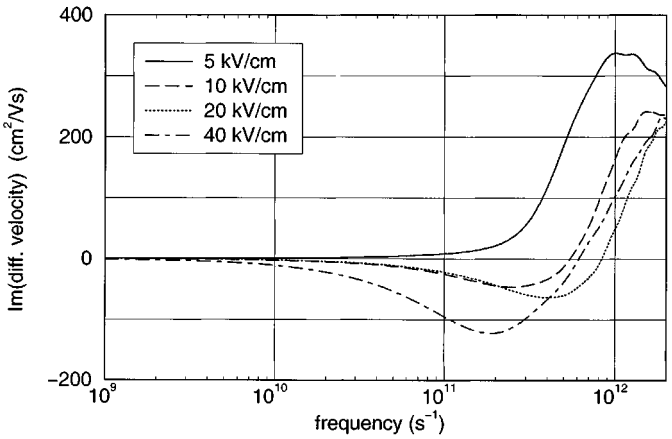


Fig. 6. Imaginary part of the differential velocity.

IV. Open Problems and Outlook

The MC method has traditionally been used to study hot carrier effects in semiconductors and semiconductor devices. The involved physical processes have been studied extensively and sophisticated models have been developed. However, when using the MC method for TCAD purposes, additional requirements arise. It should be possible to analyze a device in all operating conditions with one numerical method. Using different methods, such as moment equations for conditions close to equilibrium and the MC method for conditions far from equilibrium, has severe drawbacks. First, one had to maintain both macroscopic and microscopic models, such as mobility and scattering models, and to ensure that both models give consistent results in conditions where they can be applied simultaneously. Especially in view of the ever increasing complexity of the considered effects, development and maintenance of two kinds of models appears very uneconomical. Second, using two completely different numerical methods in a

simulator makes the software more complex and increases the number of parameters that control the simulation. The criterion for switching between the methods will remain empirical, and some discontinuity in the result at the switching point is likely to appear.

To date, the two standard algorithms that are widely used for MC device simulation on a semiclassical level are the ensemble and the one-particle algorithms. An inherent peculiarity of these algorithms is that the computational efficiency depends on the physical conditions in the device. In particular they may become prohibitively time consuming if regions with retarding electric field are included. Such a situation occurs in any semiconductor device in which transport is controlled by one or more energy barriers. On the other hand, these algorithms let the simulated particle spend too much time in highly doped contact regions where no events of interest occur and simply the equilibrium distribution is recovered.

From the BE represented as an integral equation, generalized MC algorithms can be derived, as has been demonstrated by the WEMC and the MCB algorithms. Although these algorithms have the potential to solve various problems occurring in device simulation, they have not yet been applied to realistic structures. The great freedom in choosing density functions for numerical trajectory construction opens a wide field for the development of new algorithms. Backward and forward algorithms can be combined. For example, at an artificial boundary located on top of an energy barrier backward and forward trajectories can be linked allowing a direct control of the number of numerical particles moving across the barrier.

With the WEMC method, the variance of the particle weights increases with time. Hence a robust, weighted algorithm will be likely to include some kind of weight control technique possibly realized by one of the established variable-weight methods [58].

Furthermore, an MC algorithm for linear small signal analysis of semiconductor devices is desirable. From our experience with small signal algorithms for bulk semiconductors (Section III-E) we conclude that an algorithm in the time domain will be more efficient than one in the frequency domain. In the former case, the Fourier transform is carried out as a postprocessing step in order to obtain the admittance matrix of the device.

A class of problems not considered in this work are related to the nonlinear BE (Section II-B). MC methods for integral equations with a polynomial nonlinearity do exist. Their applicability to the semiconductor transport problem has to be investigated.

REFERENCES

[1] C. Jacoboni, P. Poli, and L. Rota, "A new Monte Carlo technique for the solution of the Boltzmann transport equation," *Solid-State Electron.*, vol. 31, pp. 5523–5526, 1988.

[2] M. Nedjalkov and P. Vitanov, "Iteration approach for solving the Boltzmann equation with the Monte Carlo method," *Solid-State Electron.*, vol. 32, pp. 893–896, 1989.

[3] R. W. Hockney and J. M. Eastwood, *Computer Simulation Using Particles*. Bristol/Philadelphia, PA: Adam Hilger, 1988.

[4] C. Moglestue, "Monte Carlo particle modeling of small semiconductor devices," *Comput. Meth. Appl. Mechan Eng.*, vol. 30, pp. 173–208, 1982.

[5] P. J. Price, *IBM J. Res. Develop.*, 1970, vol. 14, p. 12.

[6] C. Jacoboni and P. Lugli, *The Monte Carlo Method for Semiconductor Device Simulation*. Berlin, Germany: Springer, 1989.

[7] W. Fawcett, A. S. Boardman, and S. Swain, "Monte Carlo determination of electron transport properties in gallium arsenide," *J. Phys. Chem. Solids*, vol. 31, pp. 1963–1990, 1970.

[8] G. Baccarani, C. Jacoboni, and A. M. Mazzone, "Current transport in narrow-base transistors," *Solid-State Electron.*, vol. 20, pp. 5–10, 1977.

[9] A. Reklaitis, "The calculation of electron transient response in semiconductors by the Monte Carlo technique," *Phys. Lett.*, vol. 88A, pp. 367–370, 1982.

[10] L. Reggiani, Ed., *Hot-Electron Transport in Semiconductors*. Berlin, Germany: Springer-Verlag, 1985, vol. 58 of Topics in Applied Physics.

[11] P. A. Lebwohl, "Monte Carlo simulation of response of a semiconductor to periodic perturbations," *J. Appl. Phys.*, vol. 44, no. 4, pp. 1744–1752, Apr. 1973.

[12] J. Zimmermann, Y. Leroy, and E. Constant, "Monte Carlo calculation of microwave and far-infrared hot-carrier mobility in n-Si: Efficiency of millimeter transit time oscillators," *J. Appl. Phys.*, vol. 49, pp. 3378–3383, 1978.

[13] P. Price, "On the calculation of differential mobility," *J. Appl. Phys.*, vol. 54, pp. 3616–3617, 1983.

[14] E. Starikov and P. Shiktorov, "New approach to calculation of the differential mobility spectrum of hot carriers: Direct modeling of the gradient of the distribution function by the Monte Carlo method," *Sov. Phys. Semiconduct.*, vol. 22, pp. 45–48, 1988.

[15] H. Kosina, M. Nedjalkov, and S. Selberherr, "A Monte-Carlo method for small signal analysis of the Boltzmann equation," *J. Appl. Phys.*, vol. 87, pp. 4308–4314, 2000.

[16] L. Reggiani *et al.*, "Modeling of small-signal response and electronic noise in semiconductor high-field transport," *Semiconduct. Sci. Technol.*, vol. 12, pp. 141–156, 1997.

[17] T. Kurosawa, "Monte Carlo calculation of hot electron problems," *J. Phys. Soc. Jpn.*, vol. 21, pp. 424–426, 1966.

[18] W. Fawcett and E. G. S. Paige, "Negative differential mobility of electrons in germanium: A Monte Carlo calculation of the distribution function, drift velocity and carrier population in the $\langle 111 \rangle$ and $\langle 100 \rangle$ minima," *J. Phys. C: Solid State Phys.*, vol. 4, pp. 1801–1821, 1971.

[19] C. Canali *et al.*, "Electron drift velocity in Silicon," *Phys. Rev. B*, vol. 12, pp. 2265–2284, Aug. 1975.

[20] C. Jacoboni, R. Minder, and G. Majni, "Effects of band nonparabolicity on electron drift velocity in Silicon above room temperature," *J. Phys. Chem. Solids*, vol. 36, pp. 1129–1133, 1975.

[21] H. Shichijo and K. Hess, "Band-structure-dependent transport and impact ionization in GaAs," *Phys. Rev. B*, vol. 23, pp. 4197–4207, Apr. 1981.

[22] M. V. Fischetti and S. E. Laux, "Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects," *Phys. Rev. B*, vol. 38, pp. 9721–9745, Nov. 1988.

[23] T. Kunikiyo *et al.*, "A Monte Carlo simulation of anisotropic electron transport in silicon including full band structure and anisotropic impact-ionization model," *J. Appl. Phys.*, vol. 75, pp. 297–312, 1994.

[24] P. D. Yoder, J. M. Higman, J. Bude, and K. Hess, "Monte Carlo simulation of hot electron transport in Si using a unified pseudopotential description of the crystal," *Semiconduct. Sci. Technol.*, vol. 7, pp. 357–359, 1992.

[25] J. Bude and R. K. Smith, "Phase-space simplex Monte Carlo for semiconductor transport," *Semiconduct. Sci. Technol.*, vol. 9, pp. 840–843, 1994.

[26] M. V. Fischetti and S. E. Laux, "Monte Carlo simulation of electron transport in Si: The first 20 years," in *Proc. 26th Eur. Solid State Device Research Conf.*, G. Baccarani and M. Rudan, Eds. Bologna, Italy, 1996, pp. 813–820.

[27] M. V. Fischetti, "Monte Carlo simulation of transport in technologically significant semiconductors of the diamond and zinc-blende structures—Part I: Homogeneous transport," *IEEE Trans. Electron Devices*, vol. 38, pp. 634–649, Mar. 1991.

[28] R. Brunetti *et al.*, "A many-band Silicon model for hot-electron transport at high energies," *Solid-State Electron.*, vol. 32, pp. 1663–1667, 1989.

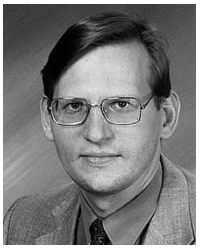
[29] T. Vogelsang and W. Hänsch, "A novel approach to include band-structure effects in a Monte Carlo simulation of electron transport in silicon," *J. Appl. Phys.*, vol. 70, pp. 1493–1499, Aug. 1991.

[30] A. Abramo *et al.*, "A numerical method to compute isotropic band models from anisotropic semiconductor band structures," *IEEE Trans. Computer-Aided Design*, vol. 12, pp. 1327–1335, 1993.

[31] A. Abramo *et al.*, "A comparison of numerical solutions of the Boltzmann transport equation for high-energy electron transport silicon," *IEEE Trans. Electron Devices*, vol. 41, pp. 1646–1654, 1994.

[32] D. Chattopadhyay and H. J. Queisser, "Electron scattering by ionized impurities in semiconductors," *Rev. Mod. Phys.*, vol. 53, no. 4, pp. 745–768, Oct. 1981.

[33] H. Kosina and G. Kaiblinger-Grujin, "Ionized-impurity scattering of majority electrons in silicon," *Solid-State Electron.*, vol. 42, pp. 331–338, 1998.
[34] J. Y. Tang and K. Hess, "Impact ionization of electrons in silicon (steady state)," *J. Appl. Phys.*, vol. 54, pp. 5139–5144, Sept. 1983.
[35] N. Sano, T. Aoki, M. Tomizawa, and A. Yoshii, "Electron transport and impact ionization in Si," *Phys. Rev. B*, vol. 41, pp. 12 122–12 128, June 1990.
[36] W. N. Grant, "Electron and hole ionization rates in epitaxial Silicon at high electric fields," *Solid-State Electron.*, vol. 16, pp. 1189–1203, 1973.
[37] M. V. Fischetti, "Effect of the electron–plasmon interaction on the electron mobility in silicon," *Phys. Rev. B*, vol. 44, pp. 5527–5534, 1991.
[38] N. S. Mansour, K. Diff, and K. F. Brennan, "Effect of inclusion of self-consistently determined electron temperature on the electron–plasmon interaction," *J. Appl. Phys.*, vol. 80, pp. 5770–5774, 1996.
[39] N. Takenaka, M. Inoue, and Y. Inuishi, "Influence of inter-carrier scattering on hot electron distribution function in GaAs," *J. Phys. Soc. Jpn.*, vol. 47, pp. 861–868, 1979.
[40] S. Bosi and C. Jacoboni, "Monte Carlo high-field transport in degenerate GaAs," *J. Phys. C*, vol. C9, p. 315, 1976.
[41] P. Lugli and D. K. Ferry, "Degeneracy in the Ensemble Monte Carlo method for high-field transport in semiconductors," *IEEE Trans. Electron Devices*, vol. ED-32, pp. 2431–2437, Nov. 1985.
[42] S. E. Laux, "On particle-mesh coupling in Monte Carlo semiconductor device simulation," *IEEE Trans. Computer-Aided Design*, vol. 15, pp. 1266–1277, 1996.
[43] F. Venturi *et al.*, "A new coupling scheme for a self-consistent Poisson and Monte Carlo device simulator," in *Proc. Simulation of Semiconductor Devices and Processes*, G. Baccarani and M. Rudan, Eds. Bologna, Italy: Tecnoprint, Sept. 1988, pp. 383–386.
[44] S. Bandyopadhyay *et al.*, "A rigorous technique to couple Monte Carlo and drift-diffusion models for computationally efficient device simulation," *IEEE Trans. Electron Devices*, vol. ED-34, pp. 392–399, Feb. 1987.
[45] H. Kosina and S. Selberherr, "A hybrid device simulator that combines Monte Carlo and drift-diffusion analysis," *IEEE Trans. Computer-Aided Design*, vol. 13, pp. 201–210, 1994.
[46] T. Ando, A. B. Fowler, and F. Stern, "Electronic properties of two-dimensional systems," *Rev. Mod. Phys.*, vol. 54, pp. 437–672, 1982.
[47] K. Yokoyama and K. Hess, "Monte Carlo study of electronic transport in $Al_{1-x}Ga_xAs$/GaAs single-well heterostructures," *Phys. Rev. B*, vol. 33, pp. 5595–5606, Apr. 1986.
[48] M. Shirahata, K. Taniguchi, and C. Hamaguchi, "A self-consistent Monte Carlo simulation for two-dimensional electron transport in MOS inversion layer," *J. Appl. Phys.*, vol. 9, pp. 1447–1452, Sept. 1987.
[49] C. Jungemann, A. Emunds, and W. L. Engl, "Simulation of linear and nonlinear electron transport in homogeneous silicon inversion layers," *Solid-State Electron.*, vol. 36, pp. 1529–1540, 1993.
[50] M. V. Fischetti and S. E. Laux, "Monte Carlo study of electron transport in Silicon inversion layers," *Phys. Rev. B*, vol. 48, pp. 2244–2274, 1993.
[51] E. Sangiorgi, B. Ricco, and F. Venturi, "MOS^2: An efficient Monte Carlo simulator for MOS devices," *IEEE Trans. Computer-Aided Design*, vol. 7, pp. 259–271, Feb. 1988.
[52] F. Venturi *et al.*, "A general purpose device simulator coupling Poisson and Monte Carlo transport with applications to deep submicron MOSFET's," *IEEE Trans. Computer-Aided Design*, vol. 8, pp. 360–369, Apr. 1989.
[53] A. Phillips and P. J. Price, "Monte Carlo calculations on hot electron tails," *Appl. Phys. Lett.*, vol. 30, pp. 528–530, May 1977.
[54] T. Wang, K. Hess, and G. J. Iafrate, "Time-dependent ensemble Monte Carlo simulation for planar-doped GaAs structures," *J. Appl. Phys.*, vol. 58, pp. 857–861, 1985.
[55] S. E. Laux and M. V. Fischetti, "Numerical aspects and implementation of the DAMOCLES Monte Carlo device simulation program," in *Monte Carlo Device Simulation: Full Band and Beyond*, K. Hess, Ed. Boston, MA: Kluwer, 1991, pp. 1–26.
[56] M. J. Martín, T. González, D. Pardo, and J. E. Veiåzquez, "Monte Carlo analysis of a Schottky diode with an automatic space-variable charge algorithm," *Semicond. Sci. Technol.*, vol. 11, pp. 380–387, 1996.
[57] C. Jungemann *et al.*, "Phase space multiple refresh: A versatile statistical enhancement method for Monte Carlo device simulation," in *Proc. Conf. Simulation of Semiconductor Processes and Devices*, Tokyo, Japan, 1996, pp. 65–66.
[58] C. J. Wordelman, T. J. T. Kwan, and C. M. Snell, "Comparison of statistical enhancement methods for Monte Carlo semiconductor simulation," *IEEE Trans. Computer-Aided Design*, vol. 17, pp. 1230–1235, 1998.
[59] C. Jacoboni, "A new approach to Monte Carlo simulation," in *IEDM Tech. Dig.*, 1989, pp. 469–472.
[60] L. Rota, C. Jacoboni, and P. Poli, "Weighted ensemble Monte Carlo," *Solid-State Electron.*, vol. 32, pp. 1417–1421, 1989.
[61] R. Chambers, "The kinetic formulation of conduction problems," *Proc. Phys. Soc Lond.*, vol. A65, pp. 458–459, 1952.
[62] M. Nedjalkov and P. Vitanov, "Application of the iteration approach to the ensemble Monte Carlo technique," *Solid-State Electron.*, vol. 33, pp. 407–410, 1990.
[63] P. Vitanov and M. Nedjalkov, "Iteration approach for solving the inhomogeneous Boltzmann equation," *Int. J. Comput. Math. Electr. Electron. Eng.*, vol. 10, pp. 531–538, 1991.
[64] F. W. Byron and R. W. Fuller, *Mathematics of Classical and Quantum Physics*. New York: Dover, 1992.
[65] M. Lundstrom, "Fundamentals of Carrier Transport," in *Modular Series on Solid State Devices*. Reading, MA: Addison-Wesley, 1990, vol. 10.



Hans Kosina (S'89–M'93) received the Dipl.Ing. degree in electrical engineering in 1987 and the Ph.D. degree in technical sciences in 1992, both from the Technical University (TU) of Vienna, Vienna, Austria.

In 1988, he joined the Institute for Microelectronics, TU, where he is currently an Associate Professor. In March 1998, he received the "venia docendi" on microelectronics. His current research topics include modeling of carrier transport and quantum effects in semiconductor devices, development of novel Monte Carlo algorithms, and computer aided engineering in ULSI-technology.

Michail Nedjalkov was born in Sofia, Bulgaria, in 1956. He received the M.S. degree in semiconductor physics from Sofia University, Sofia, in 1981, and the Ph.D. degree from the Bulgarian Academy of Sciences (BAS) in 1990.

Since 1993, he has been with the Center for Informatics and Computer Technology, BAS. Currently, he is a Visiting Researcher at the Institute for Microelectronics,Technical University of Vienna, Vienna, Austria. His research interests include numerical theory and application of the Monte Carlo method, physics and modeling of Boltzmann, and quantum transport in semiconductors and devices.

Siegfried Selberherr (M'79–SM'84–F'93) was born in Klosterneuburg, Austria, in 1955. He received the Dipl. Ing. degree in electrical engineering and the Ph.D. degree in technical sciences from the Technical University (TU) of Vienna, Austria, in 1978 and 1981, respectively.

Since that time he has been with TU as Professor. He has held the "venia docendi" on computer-aided design since 1984. He was the Head of the Institute for Microelectronics, TU, until 1998, and has been Dean of the Faculty of Electrical Engineering and Electronics, TU, since 1999. His current research topics are modeling and simulation of problems for microelectronics engineering.