# A Hybrid Device Simulator that Combines Monte Carlo and Drift-Diffusion Analysis

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

***Abstract*—A hybrid simulator suitable for modeling small semiconductor devices has been developed in which Monte Carlo and drift-diffusion models are combined. In critical device regions, the position-dependent coefficients of an extended drift-diffusion equation are extracted from a Monte Carlo simulation. Criteria for identifying these regions are described. Additional features which make the code more efficient are presented. First, a free-flight time calculation method using a new self-scattering algorithm is described. It allows for an efficient reduction of self-scattering events. Second, a unique Monte Carlo-Poisson coupling scheme has been developed which converges faster than all presently known schemes. It exploits the so-called Monte Carlo-drift diffusion coupling technique, which also forms the basis of the hybrid method. The simulator has been used to model submicron MOSFET's with gate lengths down to $0.15$ $\mu$m. In addition to the non-local effects occurring in these devices, the performance of the hybrid simulation method is analyzed.**

## I. Introduction

BECAUSE OF its efficiency, the drift-diffusion transport model has been extensively used in many device simulation programs developed to date. As long as devices are sufficiently large, this transport model along with a static mobility model has been known to yield accurate results. However, with the continuous decrease of device dimensions, this situation has changed; the electric field in the active region of a submicron device is often very high and undergoes rapid variations over distances comparable to the carrier's mean free path.

The applicability of a mobility model that depends on local quantities such as electric field, driving force or the carrier's mean energy becomes questionable. Artificial effects, e.g., spurious velocity overshoot in simple $n^+nn^+$ diodes, may then be the consequence. Actually, the local distribution function has to be known in order to specify mobility correctly under these conditions. However, this is beyond the scope of any drift-diffusion or energy-transport model.

On the other hand, tools based on the Monte Carlo technique, which directly solve the Boltzmann transport equation (BTE) in a stochastic manner, do not have this limitation. The distribution function is in principle known at any position within the device. The Monte Carlo technique, although physically very accurate, has some other limitations when applied to global device simulation. It may become very inefficient, especially in device regions where the electric field is low or retarding, whereas drift-diffusion based device models perform quite satisfactorily. Significant benefits therefore can be expected from combining both methods in such a way as to retain the computational efficiency of the drift-diffusion method as well as the physical accuracy of the Monte Carlo technique.

Several attempts have been published on this matter [2]–[4]. Park *et al.* [5] have proposed the so called hybrid technique in order to study non-local transport in small bipolar devices. The emitter-base potential barrier drastically reduces the efficiency of the Monte Carlo technique. Therefore, Monte Carlo is applied only to those regions where it is mandatory; the remaining regions are described by the drift-diffusion equations with local transport coefficients. From the Monte Carlo data, a position-dependent mobility is extracted for the simulated regions and is used in a drift-diffusion equation which is applied globally to model the entire device.

On the basis of the BTE, Bandyopadhyay *et al.* [1] have developed a rigorous method for extracting mobilities and diffusion coefficients required by the hybrid technique.

This paper is organized a follows. In Section II we describe the basic ideas of the so-called Monte Carlo-drift-diffusion coupling technique, which is the theoretical background of the hybrid approach. In Section III we focus on the Monte Carlo part of the hybrid simulation program; for the drift-diffusion part, the well established MINIMOS program [6] is used. Several algorithms have been developed to improve the performance of the Monte Carlo code. Self-scattering, a numerical device for free-flight time calculation, is used to set up a piecewise linear total scattering rate. In this way the amount of wasteful self-scattering events is considerably reduced. In very small devices, the carrier distribution has to be computed self-consistently with the electrostatic potential given by Poisson's equation. By means of a novel iteration scheme, which is feasible only in conjunction with the Monte Carlo-drift-diffusion coupling technique, this task can be performed in a very efficient manner. Furthermore, in Section III interface and boundary conditions necessary when linking the different transport models are discussed. To identify the critical device regions, physics-based criteria are presented. Finally, in Section IV we present results of the simulation of small silicon MOSFET's, highlighting both the physical effects inside the devices and the performance of the self-consistent iteration scheme.

Manuscript received February 18, 1993; revised June 25, 1993. This work was supported by Austrian Industries–AMS International, Unterpremstätten, Austria; Digital Equipment Corporation, Hudson, NY; Siemens Corporation, Munich, Germany; and Sony Corporation, Atsugi, Japan. This paper was recommended by Associate Editor D. Scharfetter.

The authors are with the Institute for Microelectronics, Technical University of Vienna, A-1040 Vienna, Austria.

IEEE Log Number 9212368.

0278–0070/94$04.00 © 1994 IEEE

## II. Theory

If in a portion of a device the drift-diffusion model becomes invalid, the carrier's mean velocity $\mathbf{v}(\mathbf{r})$ is reproduced incorrectly. However, from the conventional semiconductor equations [7] a subset consisting of Poisson's and continuity equations still remains valid. One can now ask how a velocity profile $\mathbf{v}(\mathbf{r})$ originating from another source, e.g., from a Monte Carlo calculation, can be inserted into these remaining equations. One straight forward way would be to relate the correct $\mathbf{v}(\mathbf{r})$ to the current density by

$$\mathbf{j} = -en(\mathbf{r})\mathbf{v}(\mathbf{r}), \qquad (1)$$

where (1) is then used in a continuity equation.

In [1] it is argued that during an iterative solution of (1), together with continuity and Poisson's equations, the electric field inside the device may change with each successive iteration and the Monte Carlo simulation has to be rerun each time to compute the new $\mathbf{v}(\mathbf{r})$. In that case, it is advantageous to recast (1) in the form of a drift-diffusion equation, involving certain coefficients. These coefficients are expected to change very little with the electric field unless the distribution function changes drastically. Consequently, the Monte Carlo simulation does not need to be rerun so often and convergence can be achieved faster [1].

### 2.1. Basic Equations

To find the above mentioned current equation we consider the time independent BTE for electrons. Generation-recombination of electron-hole pairs is neglected at this stage. From the BTE one can directly derive the moment equation of first order,

$$e\left(E_j + \frac{1}{en}\frac{\partial(n < \hbar k_j \cdot v_k >)}{\partial r_k}\right) = -\left(\frac{dp_j}{dt}\right)_c, \qquad (2)$$

where the distribution function implied by the averages is still treated as an unknown function. Summation over repeated indices is assumed. Here, the right-hand side represents the average momentum loss rate which explicitly reads as

$$\left(\frac{dp_j}{dt}\right)_c = < \int (\hbar k_j - \hbar k'_j) S(\mathbf{k}, \mathbf{k}') d\mathbf{k}' > . \qquad (3)$$

In (2) and (3), $E_j$ denotes the electric field, $k_i$ and $v_j$ the electron wave vector and group velocity, respectively, and $S(\mathbf{k}, \mathbf{k}')$ is the differential scattering rate.

We now take as the defining relation for the mobility

$$\mu_{ij}\frac{1}{e}\left(\frac{dp_j}{dt}\right)_c = < v_i >, \qquad (4)$$

which holds in the presence of any distribution function being physically meaningful.

If we insert (4) in the first moment equation (2) and if we interpret the second moment as energy tensor, $w_{jk} = (1/2) < \hbar k_j \cdot v_k >$, then we end up with a current equation in the form

$$j_i = en\mu_{ij}E_j + \mu_{ij}\frac{\partial 2 w_{jk} n}{\partial r_k}. \qquad (5)$$

Due to the similarity of (5) and the conventional drift-diffusion equation, both equations comprise a drift and a diffusive term at the right-hand side, we consider equation (5) as an extended drift-diffusion equation.

In what follows we illustrate the basic ideas of coupling the Monte Carlo method with the rigorous current equation (5). In the next subsection we discuss how (5) can be approximated in order to obtain a current equation better suitable for numerical implementation.

To find the required coefficients $\mu_{ij}$ and $w_{ij}$ Bandyopadhyay et al. [1] suggest calculating the three lowest order moments of the distribution function, namely $n(\mathbf{r})$, $<v_i>$ and $< \hbar k_i \cdot v_j >$, and use them in the defining relations[1]

$$\mu_{ij}\left(E_j + \frac{1}{en} \cdot \frac{\partial(n < \hbar k_j \cdot v_k >)}{\partial r_k}\right) = - < v_i >, \qquad (6)$$

$$w_{jk} = \frac{1}{2} < \hbar k_j \cdot v_k > . \qquad (7)$$

Equation (6) can be found immediately by comparing (2) and (4).

If the Monte Carlo technique is used to calculate the moments then the resulting coefficients $\mu_{ij}$ and $w_{ij}$ can be interpreted as a link between the Monte Carlo and the extended drift-diffusion models.

The derivation of the coupling coefficients from the moments outlined above shows the generality of this method. All the physical models affecting the distribution function which are accounted for in the Monte Carlo simulator, e.g., band-structure models or scattering processes, directly influence the moments and thus the coupling coefficients. In addition to the approximations inherent to the Monte Carlo-model, no further approximation is introduced when the Monte Carlo-model is coupled with (5).

Under conditions far from thermal equilibrium, (5) together with Monte Carlo generated space-dependent coefficients simply reproduce the Monte Carlo current-density. Approaching thermal equilibrium, the energy tensor becomes independent of space and is solely determined by the lattice temperature, and $\mu$ reverts to the low-field mobility $\mu_0$. In this manner the conventional drift-diffusion equation is recovered. The set of semiconductor equations incorporating the extended drift-diffusion equation (5) is therefore capable of describing high-energy transport as well as low-field transport in very small devices.

### 2.1. Approximation of the Basic Equations

Theoretically, additional approximations are not needed since the tensor-like coefficients in (5) can be directly extracted from a Monte Carlo simulation. Practically, due to the anisotropy of the coefficients it is difficult to numerically solve a continuity equation including (5). Isotropic mobility and temperature are therefore desirable.

In this work we substitute the Monte Carlo-drift-diffusion coupling coefficients, $\mu_{ij}$ and $w_{ij}$, by scalar quantities as

[1] In [1] a slightly different notation is used. The coefficients there are related to those used here by $D_{ij} = (2/e)\mu_{ik} \cdot w_{kj}$ and $\xi'_i = (2/e)\partial w_{ij}/\partial x_j$.

follows:

$$\mu = e\sqrt{\frac{< \mathbf{v} >^2}{(d\mathbf{p}/dt)_c^2}}, \tag{8}$$

$$U_T = \frac{2}{e} \cdot \frac{1}{d} \sum_{i=1}^{d} w_{ii}. \tag{9}$$

Here $U_T$ is the thermal voltage and $d$ is the number of considered space dimensions. With these assumptions (5) becomes

$$j_i = en\mu\Big(E_i + \frac{1}{n}\frac{\partial nU_T}{\partial r_i}\Big). \tag{10}$$

Another possible approximation of the energy tensor is based on random wave vector and velocity components that are defined as

$$\Delta k_i = k_i - < k_i >, \quad \Delta v_i = v_i - < v_i > . \tag{11}$$

The energy tensor can then be separated in two terms,

$$w_{ij} = \frac{\hbar}{2} < k_i >< v_j > + \frac{\hbar}{2} < \Delta k_i \cdot \Delta v_j > . \tag{12}$$

By analogy with the hydrodynamic transport model an isotropic temperature voltage can be defined by means of the random components,

$$U_T' = \frac{\hbar}{e} \cdot \frac{1}{3} \mathrm{Tr} < \Delta k_i \cdot \Delta v_j > . \tag{13}$$

With this definition one obtains from (5) a different current equation,

$$j_i = en\mu\Big(E_i + C_i + \frac{1}{n}\frac{\partial nU_T'}{\partial r_i}\Big), \tag{14}$$

in which an additional convective term occurs,

$$C_i = \frac{1}{e} < v_j > \frac{\partial < \hbar k_i >}{\partial r_j}. \tag{15}$$

(14) could be employed in the Monte Carlo-drift-diffusion coupling technique as well, however within this approximation one would have to extract the convective term as a third coupling coefficient from the Monte Carlo simulation.

## III. The Simulation Program

The simulator implementing the hybrid technique is composed of two parts, the conventional MINIMOS program to solve the semiconductor equations, and a single-particle Monte Carlo program to provide the Monte Carlo-drift-diffusion coupling coefficients. A simulation is controlled by MINIMOS and all required input data are supplied by its specific user-interface. In the rest of this section, we focus on the Monte Carlo module, in particular on newly developed algorithms used herein and on the underlying physical model.

### *3.1. The Semiconductor Model*

The Monte Carlo technique applied to charge transport in semiconductors relies on the simulation of individual electron trajectories, each consisting alternately of a drift governed by the electron's $\epsilon(\mathbf{k})$ relation followed by a scattering event.

The band-model of silicon used here accounts for the six equivalent minima of the conduction band near the X-points. Anisotropy is treated by the Herring-Vogt transformation, while nonparabolicity is described by the band-form function

$$\gamma(\epsilon) = \epsilon(1 + \alpha\epsilon), \tag{16}$$

which is defined as $\gamma = \hbar^2\mathbf{k}^{*2}/2m_d$. Here, $\epsilon$ denotes the electron energy, $\alpha$ the non-parabolicity factor, $\mathbf{k}^*$ the wave vector in the Herring-Vogt transformed space and $m_d$ the density-of-states effective mass. The scattering mechanisms included in our model are acoustic intravalley scattering in the elastic approximation, intervalley phonon scattering of both f- and g-type [8], scattering by ionized impurities and surface roughness scattering, the latter being of particular importance in MOSFET's. With the coupling constants and effective masses given in [8] the lattice mobility of bulk silicon can be reproduced very well.

In order to implement the hybrid technique, a generalized mobility given by (6) has to be evaluated. This step can cause difficulties, as in the diffusion term spatial derivatives of Monte Carlo generated quantities are needed. Because of the noise associated with such quantities, this procedure is expected to be inaccurate.

To circumvent this problem we directly employ the defining relation (4). The required term $(dp_j/dt)_c$ can be calculated by the Monte Carlo method without a need for spatial differentiation.

For band-structures with either spherical or ellipsoidal energy surfaces, the average momentum loss rate given by (3) can be expressed in terms of an energy-dependent momentum relaxation rate $\tau_m^{-1}(\epsilon)$ as

$$\left(\frac{dp_i}{dt}\right)_c = < \hbar k_i \cdot \tau_m^{-1}(\epsilon) >, \tag{17}$$

where $\tau_m^{-1}(\epsilon)$ is related to the differential scattering rate $S(\mathbf{k}, \mathbf{k}')$ by

$$\tau_m^{-1}(\epsilon) = \int \left(1 - \frac{k'}{k}\cos\theta\right) S(\mathbf{k}, \mathbf{k}')d\mathbf{k}'. \tag{18}$$

Considering the scattering mechanisms included in our model, the momentum relaxation rate can be written as a superposition which implies independence of all scattering processes:

$$\tau_m^{-1}(\epsilon) = \lambda_{\mathrm{int}}(\epsilon) + \lambda_{\mathrm{ac}}(\epsilon) + \lambda_{\mathrm{surf}}^{\mathrm{tot}}(\epsilon) + \tau_{\mathrm{m}}^{\mathrm{ion}}(\epsilon)^{-1}. \tag{19}$$

Here we have exploited the fact that, for isotropic scattering mechanisms, $\tau_m^{-1}(\epsilon) = \lambda(\epsilon)$, where $\lambda(\epsilon)$ is the total scattering rate. In (19), the isotropic mechanisms are intervalley phonon scattering $\lambda_{\mathrm{int}}(\epsilon)$, acoustic intravalley scattering $\lambda_{\mathrm{ac}}(\epsilon)$, and surface roughness scattering $\lambda_{\mathrm{surf}}(\epsilon)$. The latter is a two-dimensional scattering process which is treated isotropic in the plane parallel to the *Si/SiO*$_2$ interface [9].

Ionized-impurity scattering, which is strongly anisotropic, requires a distinction to be made between $\lambda_{\text{ion}}(\epsilon)$ and $\tau_{\text{m}}^{\text{ion}}(\epsilon)^{-1}$. In the Brooks-Herring model one obtains

$$\lambda_{\text{ion}}(\epsilon) = C_{\text{ion}}(\epsilon)\cdot\frac{1}{E_\beta^2(1+b)}, \tag{20}$$

$$\tau_{\text{m}}^{\text{ion}}(\epsilon)^{-1} = C_{\text{ion}}(\epsilon)\cdot\frac{1}{8\gamma^2}\cdot\left(\ln(1+b)-\frac{1}{1+b}\right), \tag{21}$$

where

$$C_{\text{ion}}(\epsilon) = \frac{N_I e^4}{2^{3/2}\pi(\epsilon_0\epsilon_r)^2\sqrt{m_d}}\cdot\sqrt{\gamma}\cdot(1+2\alpha\epsilon), \tag{22}$$

$$E_\beta = \frac{\hbar^2\beta_s^2}{2m_d},\quad b=\frac{4\gamma}{E_\beta}. \tag{23}$$

To get realistic impurity scattering in the heavily doped source and drain regions, Fermi–Dirac statistics has to be used in the reciprocal screening length $\beta_s$ in (23) [10],

$$\beta_s^2 = \frac{e^2 n(\mathbf{r})}{\epsilon_0\epsilon_r k_B T(\mathbf{r})}\,\frac{\mathcal{F}_{-1/2}(\eta)}{\mathcal{F}_{1/2}(\eta)} \tag{24}$$

To compute the final impurity scattering rate we use the formula suggested in [11], thus avoiding the high values of the scattering rate at low energies typically found for the Brooks–Herring formula.

### 3.2. Boundary Conditions

In order to save computer time, the Monte Carlo simulation should be restricted to that region where the drift-diffusion model is in error. In that region, the accelerating field causing elevated carrier energies is usually high, so that the Monte Carlo technique is expected to perform quite efficiently. This fact is exploited in the so-called "regional" Monte Carlo approach [12]. Regions with low fields and, in particular, those with retarding fields which make Monte Carlo methods very inefficient are excluded. Following this approach, a severe problem arises from the treatment of the boundaries. In fact, the electric field there is significant and, in order to model carriers entering and leaving the analyzed region, a certain distribution function has to be assumed. A special guess, e.g., a displaced Maxwellian [2], [12], seems to be too crude to perform accurate device analysis.

A solution to this problem is to extend the Monte Carlo domain in regions where the electric field nearly vanishes, implying that the carriers there obey an equilibrium distribution [13], [14].

To clarify how we treat the boundary conditions in the hybrid approach, let us consider the MOSFET sketched in Fig. 1. There, the whole simulation domain is labeled $D_0$. Furthermore, let $D_1$ be a subdomain of $D_0$ and $D_2$ a subdomain of $D_1$. The boundaries of $D_1$ and $D_2$ are referred to as $\partial D_1$ and $\partial D_2$, respectively.

Physically, $D_1$ denotes that window to which Monte Carlo is restricted; and in $D_2$ the Monte Carlo-drift-diffusion coupling technique described above is applied. $D_2$ is considered as that subdomain in which hot carrier effects are prevalent and hence the drift-diffusion model is invalid.

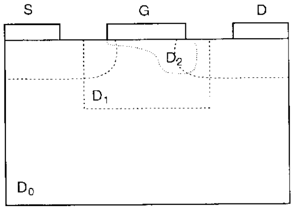


Fig. 1. Placement of subdomains $D_1$ and $D_2$ in the overall simulation domain $D_0$ for a MOSFET. $D_0$: overall simulation domain, described by the extended DD-equation, $D_1$: domain for MC-simulation, $D_2$: application of the MC-DD-coupling technique.

Carrier transport in the overall simulation domain $D_0$ is described by the extended drift-diffusion equation (10). The solution of (10) in conjunction with continuity and Poisson's equations is unique in $D_0$. In particular, neither of the boundaries $\partial D_1$ and $\partial D_2$ has an influence on the solution process. From the standpoint of transport physics, both the zero- and first-order moments, $n(\mathbf{r})$ and $\mathbf{j}(\mathbf{r})$, are continuous at either boundary $\partial D_1$ and $\partial D_2$. Therefore, no special care has to be taken of fulfilling interface conditions for $n(\mathbf{r})$ and $\mathbf{j}(\mathbf{r})$.

The remaining question concerns the interface conditions for the coefficients in (10), $\mu(\mathbf{r})$ and $U_T(\mathbf{r})$. Outside $D_2$, the electric field is so moderate that conventional drift-diffusion analysis suffices. Crossing the boundary $\partial D_2$, $\mu$ and $U_T$ continuously evolve from their near-equilibrium values in the region $D_0\backslash D_2$ to their off-equilibrium values in $D_2$. In a practical hybrid simulation, this continuous transition at $\partial D_2$ is ensured by the fact that in the moderate-field limit for a given material any local mobility-model has to coincide with the non-locally defined one after (4).

In the rest of this section we discuss the criteria for domain boundary placement. The Monte Carlo domain $D_1$ is chosen such as to include parts of the heavily doped source and drain regions where the electric field nearly vanishes and carriers are thermally distributed. Since there is an overlap area, $D_1\backslash D_2$, where particle trajectories are already examined but the final transport description is still done by drift-diffusion , the choice of any boundary distribution at $\partial D_1$ is by far less stringent than it would be in the original "regional" approach. Such a choice of boundaries, while essential for the accuracy, degrades the efficiency of the Monte Carlo code. Therefore, to overcome the difficulties associated with the retarding potential barrier which occurs at the source-channel transition, we have implemented a particle split algorithm [15], such that the number of particles injected in the channel is increased and the statistics in the domain of interest, $D_2$, is enhanced.

To identify the critical region $D_2$ we look at the Monte Carlo-drift-diffusion coupling coefficients in their scalar approximation, (8) and (9).

As in $D_1$ the spatial distribution of $U_T$ is known, it would be straightforward to identify $D_2$ by comparing $U_T$ with its equilibrium value $U_{T0}$. $D_2$ shall include those points $\mathbf{r}$ where

$$U_T(\mathbf{r}) \geq U_{T0}\cdot f_1 \tag{25}$$

holds. Here $f_1$ is a factor slightly above 1, we typically use $f_1 = 1.05$. With criterion (25) we certainly identify a region

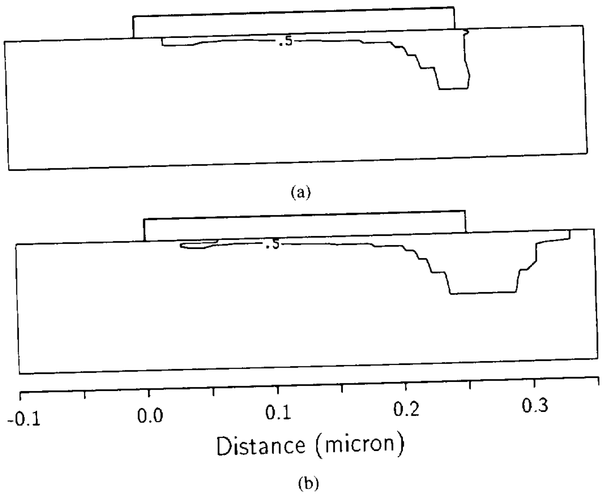


Fig. 2. Comparison of different criteria to identify the off-equilibrium region $D_2$, demonstrated on a quarter-micron MOSFET. The critical region consists of those points where (a) the average momentum loss rate is in excess of 2 kV/cm, and (b) $U_T$ is by 5% larger than $U_{T0}$.

where carriers are more or less far away from equilibrium conditions. However, due to the peculiarities of the Monte Carlo method this argument needs to be refined. $U_T$, being proportional to the second order moment, depends solely on the symmetric part of the distribution function. On the other hand, the terms entering the expression for the mobility, namely $< \mathbf{v} >$ and $(dp/dt)_c$, depend on the antisymmetric part of $f(\mathbf{k},\mathbf{r})$. This part however, disappears when thermal equilibrium is approached. As a consequence, $U_T$ can be very well reproduced by the Monte Carlo method both in the equilibrium and non-equilibrium regimes, whereas $\mu$ will be very noisy in conditions near equilibrium. We therefore use different criteria for $\mu$ and $U_T$, leading to different regions $D_2$. For $U_T$ criterion (25) is natural. For $\mu$ we check whether its constituent terms are large enough to avoid division of too small quantities in (8). In the criterion determining the mobility window we account for the denominator in (8), namely the average momentum-loss rate, and compare it to a certain value $F_{\min}$,

$$\left(\frac{dp}{dt}\right)_c \geq e \cdot F_{\min}. \tag{26}$$

In practical simulations for $F_{\min}$ a range of 1.5kV/cm to 2kV/cm has turned out to be useful.

Fig. 2 depicts those regions in the cross section of a quarter-micron MOSFET that are identified by (25) and (26). Obviously, the temperature criterion, (25), yields a larger area than the momentum loss criterion (26). The reasons for this are twofold. First, with the assumed parameters ($f_1 = 1.05$ in (25), $F_{\min} = 2$ kV/cm in (26)) criterion (25) more sensitively detects off-equilibrium conditions than criterion (26). Second, in the area near the drain edge two distributions coexist, one of hot carriers arriving from the channel and another of cold carriers residing in the drain. This mixture exhibits an increased $U_T$ detected by (25), but a drift component which is too low to fulfill (26).

### 3.3. The Self-Consistent Iteration Scheme

For very short devices, an increasing fraction of the electron population is in non-equilibrium conditions. As Monte Carlo treats the average motion of those electrons in a way which differs substantially from a standard drift-diffusion model, the resultant distribution of mobile charge in real space will also differ. Realistic results can therefore only be expected by applying some sort of self-consistent technique. The standard technique described in [16] couples the BTE solved by the Monte Carlo method with a linear Poisson equation. This method, though straightforward, may lead to stability problems. Improvements are obtained by a non-linear coupling scheme proposed in [14].

On the basis of the Monte Carlo-drift-diffusion coupling method presented in this work a novel self-consistent solution strategy can be investigated. Let us consider the following set of equations,

$$\operatorname{div}(\epsilon \text{ grad } \psi) = e(n - p - N_C), \tag{27}$$

$$\operatorname{div} \mathbf{j} = 0, \tag{28}$$

$$\mathbf{j} = en\mu\left(-\text{grad } \psi + \frac{1}{n}\text{grad}(nU_T)\right). \tag{29}$$

The extended drift-diffusion-equation (29) corresponds to (5) if scalar coefficients are inserted instead of tensor ones.

In each cycle of the self-consistent iteration loop, a Monte Carlo simulation has to be performed where the potential is taken from the previous cycle, and the distributions of $\mu$ and $U_T$ serve as result. These coefficients, given by Monte Carlo just in the critical device region (see Fig. 1), are then extended analytically over the rest of the simulation domain. With the $\mu$ and $U_T$ profiles assembled in such a way, the coupled set of equations, (27)–(29), is solved. With the updated potential the iteration cycle is repeated until the change in the potential is sufficiently small. Fig. 4 shows a flowchart of this algorithm. The initial potential distribution, $\psi^{(0)}$, is generated by a standard drift-diffusion simulation.

This new approach to self-consistency is expected to yield a high convergence rate. Generally, the carrier concentration is rather sensitive to small changes in the potential, due to the roughly exponential dependence. On the other hand, the potential strongly depends on the space-charge density, which is determined by the carrier concentration. A procedure that iteratively calculates $\psi$ from $n$ (solving Poisson's equation) and then $n$ from $\psi$ (solving the transport problem) will suffer from this strong coupling and thus will exhibit a slow convergence rate.

When we now consider the coefficients used in our iteration scheme, we find that they depend on some kind of first order moments, $< \mathbf{v}(\mathbf{k}) >, < \hbar\mathbf{k} \cdot \tau_m^{-1}(\epsilon) >$, and on the second order moment, $<\hbar k_i \cdot v_j(\mathbf{k})>$, but that they definitely do not depend on the zero-order moment, $n(\mathbf{r})$. The critical potential dependence as it exists for $n(\mathbf{r})$ is thus removed from the coupling coefficients. Consequently, there is just a quite moderate coupling between the iterated quantities, $\psi$ and $(\mu, U_T)$ (see Fig. 4). This explains the extremely low number of required iterations, as it will be demonstrated in the next section.

### *3.4. Self-Scattering*

In a Monte Carlo simulation, a stochastic sequence of free-flight times must be generated according to a given probability distribution. Following the direct technique [8] the free-flight time $t_f$ is determined by an integral equation,

$$\int_0^{t_f} \lambda(\mathbf{k}_{(t)}, \mathbf{r}_{(t)})dt = -\ln(r), \tag{30}$$

where $r$ is a random number evenly distributed between 0 and 1. Usually, $\lambda(\mathbf{k}, \mathbf{r})$ is a rather complex function of its arguments. Furthermore, (30) along with the equations of motion form a coupled system of equations, which has to be solved simultaneously. For these reasons, a direct analytic solution to (30) is in most cases prohibited. This problem can be significantly eased by the introduction of a virtual scattering mechanism called self-scattering [17]. In the so-called constant-$\Gamma$ technique the self-scattering rate is chosen so that $\lambda$ becomes independent from $\mathbf{k}$ and $\mathbf{r}$. Then (30) simplifies to a first-order algebraic equation

$$\Gamma \cdot t_f + \ln(r) = 0. \tag{31}$$

Although the solution for $t_f$ is quite trivial, this technique has computational drawbacks. Since the total scattering rate is kept artificially high, a high percentage of self-scattering events occurs. Improvements of the constant-$\Gamma$ technique are the piecewise-constant $\Gamma$ technique [13] and those techniques that try to optimize a constant $\Gamma$ level with respect to the current particle's state [18], [19]. All the methods outlined above share the assumption that self-scattering can only be used to simplify (30) into (31), at least in predefined energy intervals. An equation of intermediate complexity, yet analytically solvable, shall be derived in what follows.

Our Ansatz starts with the equation of motion in $\mathbf{k}$ space, which is solved analytically under the assumption that the electric field is constant within a mesh-cell,

$$\mathbf{k}(t) = \mathbf{k}_0 - \xi \cdot t. \tag{32}$$

Here $\xi$ is related to the electric field, $\xi = e\mathbf{E}/\hbar$, and $\mathbf{k}_0$ denotes the wave vector at the beginning of the free flight. Now we deviate from the constant $\Gamma$ and allow a linear dependence from $k^2$,

$$\Gamma(\mathbf{k}) = ak^2 + b. \tag{33}$$

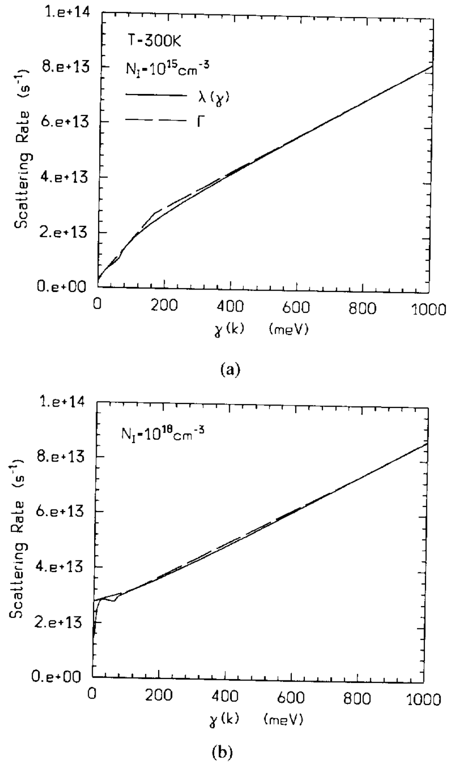
Since $\mathbf{k}(t)$ evolves linearly in time, $\Gamma(\mathbf{k}(t))$ has a quadratic time dependence. If (32) and (33) are inserted in (30) we end up with a third-order algebraic equation in $t_f$,

$$t_f^3 - 3T_0t_f^2 + St_f - T = 0. \tag{34}$$

Here the coefficients are defined as

$$T_0 = \frac{(\mathbf{k}_0 \cdot \xi)}{\xi^2}, \quad S = \frac{3(ak_0^2 + b)}{a\xi^2}, \quad T = -\ln(r)\frac{3}{a\xi^2}. \tag{35}$$

From (34) the free-flight time can be obtained analytically, since for third-order algebraic equations a closed solution always exists. Due to the inclusion of impurity and surface scattering the parameters $a$ and $b$ have to be chosen space-dependent.



(a)

(b)

Fig. 3. Use of self-scattering to set up a piecewise linearenvelope function ($\Gamma_{(\gamma)}$)for the physical scattering rate($\lambda_{(\gamma)}$), for different doping levels.The argument of thefunctions is $\gamma$, defined as $\gamma = \hbar^2\mathbf{k}^2/2m_d$.

Fig. 3 illustrates the small size of the area bounded by the curves $\lambda(\gamma)$ and $\Gamma(\gamma)$ which is obtained by means of the linear Ansatz (33). Since that area can be considered as a measure of the amount of self-scatterings that will occur, an efficient suppression of self-scattering can be expected. It should furthermore be noted that an assumption of a maximum energy during the simulation, as required in the piecewise-constant $\Gamma$ technique, is not needed in the approach presented here.

An implementation of the new algorithm using three different linear segments has shown that the amount of self-scatterings typically lies below 10%.

## IV. Results and Discussion

In this section, we apply the hybrid technique described in the previous sections to the simulation of $n$-channel MOS-devices. Gate mask length ranges from 0.75 $\mu$m down to 0.15 $\mu$m. Table I summarizes the characteristic parameters of the devices under investigation. $\text{L}_{\text{eff}}$ denotes the distance between the vertical pn-junctions, $r_j$ is the junction depth and $\text{t}_{\text{ox}}$ is the oxide thickness. The threshold voltage, $V_{\text{TH}}$, is determined at room temperature for $V_{DS} = 0.05\text{V}$. For devices A and B, doping profiles are modeled with process parameters similar to those described in [20], [21].

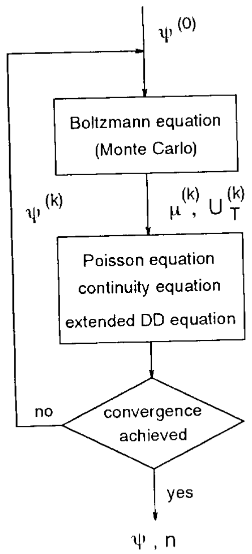


Fig. 4. Flowchart of the self-consistent iteration scheme.

TABLE I
CHARACTERISTIC PARAMETERS OF THE SIMULATED $n$-CHANNEL MOSFETS

| Device | $L_G$ ($\mu$m) | $L_{\text{eff}}$ ($\mu$m) | $r_j$ ($\mu$m) | $t_{\text{ox}}$ (nm) | $V_{\text{DS}}$ (V) | $U_{\text{TH}}$ (V) |
|---|---|---|---|---|---|---|
| A | 0.15 | 0.121 | 0.039 | 5 | 2.0 | 0.24 |
| B | 0.25 | 0.168 | 0.125 | 5 | 2.5 | 0.26 |
| C | 0.75 | 0.604 | 0.181 | 15 | 5.0 | 0.70 |

Throughout the simulations, we use $\alpha = 0.7\text{eV}^{-1}$, $\Delta \cdot L = 0.3\ \text{nm}^2$ (surface roughness parameter, see [9]) and $N = 5 \cdot 10^7$ as the number of scattering events to be calculated during a single Monte Carlo simulation.

Normalized averages are assigned to the grid nodes by a convolution method, with weighting functions as described in [22]. The lateral and vertical widths of the weighting function are related to the device's gate length by $L_x = L_G/10$ and $L_y = L_G/50$, respectively.

Quantities such as drain current, drift velocity and carrier concentrations plotted in this section are obtained from a solution to the extended semiconductor equations, (27)–(29).

Fig. 5 shows the evolution of the drain current with the number of iterations for device B. An iteration number of zero corresponds to the initial solution determined by a standard drift-diffusion simulation. The drastic increase in the drain current after the first iteration can be attributed to velocity overshoot. Subsequent iterations cause the impact of velocity overshoot on $I_D$ to be reduced so that the final stationary value of $I_D$ lies slightly above $I_D^{(0)}$. In Fig. 6(a) and (c) the evolution of $I_D$ is shown for the 0.75 $\mu$m device at different gate biases. In Fig. 6(c), where $V_{GS} = V_{DS}$, an overshoot of $I_D$ can be observed at the beginning of the iteration. Fig. 6(b) and (d) show the relative norms of the increments of carrier concentration and electrostatic potential as a function of the number of iterations. The norms first decrease rapidly but are then limited due to the statistical noise inherent in the Monte Carlo method. The relative norms of the $\psi$-increments typically taper off below $10^{-3}$. In all the simulations we have performed, an iteration count no larger than five was required in order to obtain the final drain current. Any systematic transient in the relative norms also dies out within this iteration number. With the new iteration scheme therefore, the number of costly Monte Carlo-Poisson iterations is considerably reduced compared to other coupling schemes reported in the literature [14], [16].

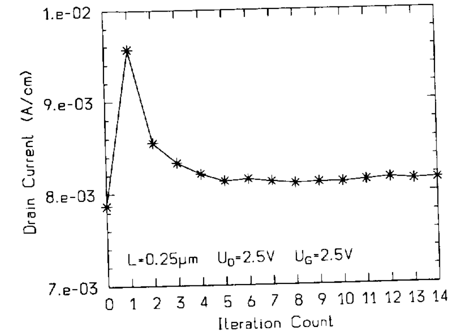


Fig. 5. Convergence rate of the self-consistent coupling scheme: drain current as a function of the number of iterations for device B at a high gate bias.

In order to figure out the differences between self-consistent Monte Carlo and self-consistent drift-diffusion simulations we have plotted in Fig. 7 the potential and the electric field occurring at the *Si/SiO*$_2$ interfaces of devices A and B. In the Monte Carlo case, in the high field region the potential profile becomes smoother (Fig. 7(a) and (c)) so that a significant lower lateral electric field is predicted (Fig. 7(b) and (d)). This effect comes from a reduced space charge density in that area, as is indicated by the reduced carrier concentration shown in Fig. 9(b). When using standard drift-diffusion simulations for such small devices, one should be aware first of the tendency to overestimate the maximum electric field by some 10 % (e.g., 39% in device A, 35% in device B). Second, in order to predict hot carrier-induced phenomena such as impact ionization the maximum electric field is not as significant as in long-channel devices. Due to the narrowness of the field peak, its capability of producing damage is decreased.

In Fig. 8 we compare the surface mobility obtained from a local model with that from a Monte Carlo simulation. At the beginning and at the end of the depicted lateral distance mobility is mainly determined by the high doping levels in that sections.

In the high-field region where the extended semiconductor equations massively reproduce velocity-overshoot (see Fig. 9(a)), the non-local mobility (solid line) exceeds the local one (dashed line). In the example shown in Fig. 8 the absolute minima differ by 37%. The local mobility recovers from its minimum to the same extent as the electric field decreases, whereas the non-local mobility, which is degraded due to a hot distribution function, recovers with some delay, since cooling has to take place.

As for Fig. 9(a), the velocity profile from the local mobility model (dashed line) is clearly bounded by the bulk saturation velocity ($v_{\text{sat}} = 10^7$ cm/s). The reasons why the carrier

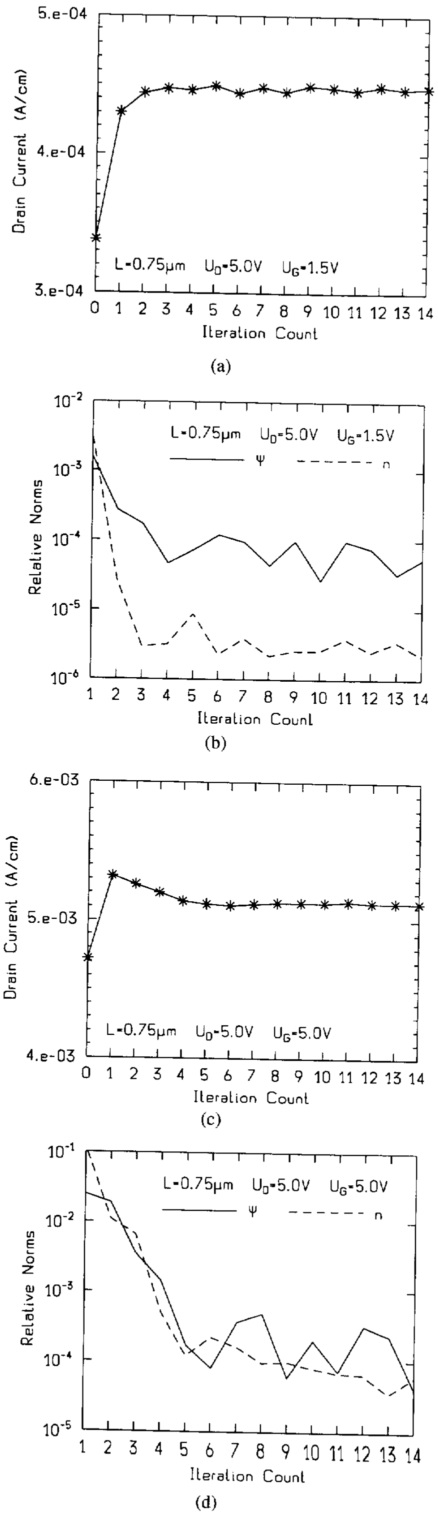


Fig. 6. Convergence rate of the self-consistent coupling scheme: drain currents and relative norms of the $n$- and $\psi$-increments as a function of the number of iterations for device C at different gate biases.

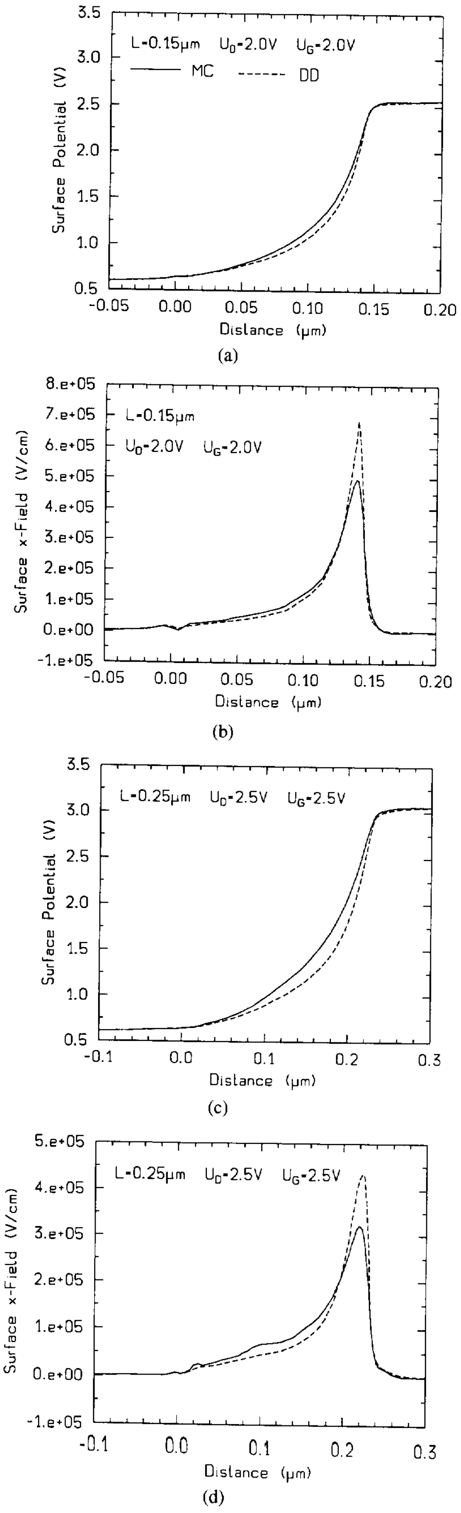


Fig. 7. Comparison of self-consistent DD (dashed line) and self-consistent MC (solid line)results for devices A and B. (a) and (c): surface potential; (b) and (d): lateral electric field at the surface.

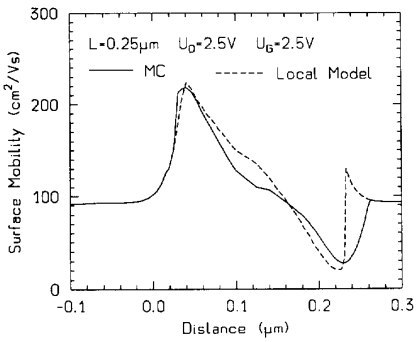
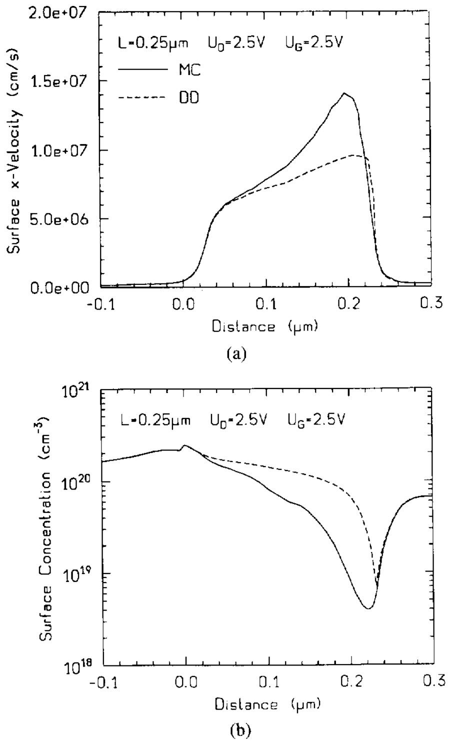


Fig. 8. Surface mobilities in device B for $V_G = V_D = 2.5$ V. Solid line: nonlocal mobility from a MC simulation. Dashed line: local (analytical) mobility model.



(a)

(b)

Fig. 9. Comparison of MC- (solid line) and DD-results (dashed line) for device B. (a) average velocity at the surface; (b) surface concentration of electrons.

concentrations in the non-local case (Fig. 9(b)) are lower than in the local case are twofold. First, due to the introduction of carrier heating, which is absent in the standard drift-diffusion model, the inversion layer broadens and thus the surface concentration lowers. The second contribution comes from the continuity of the total current, which is controlled by the situation at the source-sided part of the channel. Velocity overshoot in this region is not very pronounced. In the pinch-off region, the massive overshoot in the velocity is then compensated by an undershoot in the carrier concentration.

## V. Conclusion

A hybrid simulation method based on the Monte Carlo-drift-diffusion coupling technique has been implemented in a two-dimensional device simulator. In the high-field region of a device, an extended drift-diffusion equation reproduces non-local transport phenomena, provided that the coefficients there are calculated correctly by the Monte Carlo method.

To impose safe interface conditions when connecting the drift-diffusion and the particle models, an overlap area is admitted, in which transport is still described by the drift-diffusion model but particle trajectories are also examined.

As a by-product of the Monte Carlo-drift-diffusion coupling technique, a self-consistent iteration scheme is obtained which has been shown to require a very low number of iterations. In each cycle of that scheme an update of the Monte Carlo-drift-diffusion coupling coefficients is performed.

Additionally, a self-scattering algorithm has been proposed, which takes advantage of a piecewise linear total scattering rate. Suppression of self-scattering has been shown to be very efficient. Application to submicron MOSFET's has demonstrated the applicability of the hybrid simulation method as well as the necessity of self-consistent simulation for such small devices.

## References

[1] S. Bandyopadhyay, M. Klausmeier-Brown, C. Maziar, S. Datta, and M. Lundstrom, "A rigorous technique to couple Monte Carlo and drift-diffusion models for computationally efficient device simulation," *IEEE Trans. Elect. Dev.*, vol. ED-34, pp. 392–399, Feb. 1987.

[2] D. Cheng, C. Hwang, and R. Dutton, "PISCES-MC: A multiwindow, multimethod 2-D device simulator," *IEEE Tran. Computer-Aided Design*, vol. CAD-7, pp. 1017–1026, Sept. 1988.

[3] J. Higman, K. Hess, C. Hwang, and R. Dutton, "Coupled Monte Carlo-drift diffusion analysis of hot-electron effects in MOSFET's," *IEEE Tran. Elect. Dev.*, vol. 36, pp. 930–937, May 1989.

[4] C. Hwang, D. Navon, and T. Tang, "Monte Carlo simulation of the GaAs permeable base transistor," *IEEE Trans. Elect. Dev.*, vol. ED-34, pp. 154–159, Feb. 1987.

[5] Y. Park, D. Navon, and T. Tang, "Monte Carlo simulation of bipolar transistors," *IEEE Trans. Elect. Dev.*, vol. ED-31, pp. 1724–1730, Dec. 1984.

[6] W. Hänsch and S. Selberherr, "MINIMOS 3: A MOSFET simulator that includes energy balance," *IEEE Trans. Elect. Dev.*, vol. ED-34, pp. 1074–1078, May 1987.

[7] S. Selberherr, *Analysis and Simulation of Semiconductor Devices*. Wein, Austria: Springer Verlag, 1984.

[8] C. Jacoboni and L. Reggiani, "The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials," *Rev. Mod. Phys.*, vol. 55, pp. 645–705, July 1983.

[9] Chu-Hao, J. Zimmermann, M. Charef, R. Fauquembergue, and E. Constant, "Monte Carlo study of two-dimensional electron gas transport in Si-MOS devices," *Solid-State Electron.*, vol. 28, no. 8, pp. 733–740, 1985.

[10] D. Chattopadhyay and H. Queisser, "Electron scattering by ionized impurities in semiconductors," *Rev. Mod. Phys.*, vol. 53, pp. 745–768, Oct. 1981.

[11] T. Van de Roer and F. Widdershoven, "Ionized impurity scattering in Monte Carlo calculations," *J. Appl. Phys.*, vol. 59, pp. 813–815, Feb. 1986.

[12] P. Nguyen, D. Navon, and T. Tang, "Boundary conditions in regional Monte Carlo device analysis," *IEEE Trans. Elect. Dev.*, vol. ED-32, pp. 783–787, Apr. 1985.

[13] E. Sangiorgi, B. Ricco, and F. Venturi, "$MOS^2$: An efficient Monte Carlo simulator for MOS devices," *IEEE Trans. Computer-Aided Design*, vol. CAD-7, pp. 259–271, Feb. 1988.

[14] F. Venturi, R. Smith, E. Sangiorgi, M. Pinto, and B. Ricco, "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.

[15] A. Phillips and P. Price, "Monte Carlo calculations on hot electron tails," *Appl. Phys. Lett.*, vol. 30, pp. 528–530, May 1977.

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

[17] H. Rees, "Calculation of distribution functions by exploiting the stability of the steady state," *J. Phys. Chem. Solids*, vol. 30, pp. 643–655, 1969.

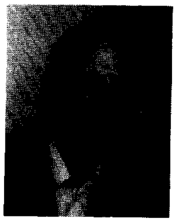
[18] K. Kato, "Hot-carrier simulation for MOSFET's using a high-speed Monte Carlo method," *IEEE Trans. Elect. Dev.*, vol. ED-35, pp. 1344–1350, Aug. 1988.

[19] C. Moglestue, "A self-consistent Monte Carlo particle model to analyze semiconductor microcomponents of any geometry," *IEEE Trans. Computer-Aided Design*, vol. CAD-5, pp. 326–354, Apr. 1986.

[20] G. Sai-Halasz and H. Harrison, "Device-grade ultra-shallow junction fabricated with Antimony," *IEEE Elect. Dev. Lett.*, vol. EDL-7, pp. 534–536, Sept. 1986.

[21] G. Sai-Halasz, M. Wordeman, K. Kern, E. Ganin, S. Rishton, D. Zicherman, H. Schmid, M. Polcari, H. Ng, P. Restle, T. Chang, and R. Dennard, "Design and experimental technology for 0.1-$\mu$m gate length low-temperature operation FET's," *IEEE Elect. Dev. Lett.*, vol. EDL-8, pp. 463–466, Oct. 1987.

[22] H. Kosina and S. Selberherr, "Analysis of filter techniques for Monte-Carlo device simulation," in *Simulation of Semiconductor Devices and Processes*, D. Aemmer, W. Fichtner, ed. Konstanz: Hartung–Gorre, pp. 251–256, Sept. 1991.



**Hans Kosina** (S'89–M'93) received the Diplomingenieur degree in electrical engineering and the Ph. D. degree from the Technical University of Vienna, Austria in 1987 and 1992, respectively.

He was with the *Institut für flexible Automation* for one year, then joined the *Institut für Mikroelektronik* at the Technical University of Vienna. Currently he is an Assistant Professor there, managing the device modeling group. His current interests include physics and technology of solid-state devices and integrated circuits.

**Siegfried Selberherr** (M'79–SM'84–F'93), for a photograph and a biography, please see page 81 of the January 1994 issue of this TRANSACTIONS.