INVITED PAPER *Special Issue on TCAD for Semiconductor Industries*

# Efficient Full-Band Monte Carlo Simulation of Silicon Devices

**Christoph JUNGEMANN**[†a)], **Stefan KEITH**[†], **Martin BARTELS**[†], *and* **Bernd MEINERZHAGEN**[†], *Nonmembers*

**SUMMARY** The full-band Monte Carlo technique is currently the most accurate device simulation method, but its usefulness is limited because it is very CPU intensive. This work describes efficient algorithms in detail, which raise the efficiency of the full-band Monte Carlo method to a level where it becomes applicable in the device design process beyond exemplary simulations. The $\boldsymbol{k}$-space is discretized with a nonuniform tetrahedral grid, which minimizes the discretization error of the linear energy interpolation and memory requirements. A consistent discretization of the inverse mass tensor is utilized to formulate efficient transport parameter estimators. Particle scattering is modeled in such a way that a very fast rejection technique can be used for the generation of the final state eliminating the main cause of the inefficiency of full-band Monte Carlo simulations. The developed full-band Monte Carlo simulator is highly efficient. For example, in conjunction with the nonself-consistent simulation technique CPU times of a few CPU minutes per bias point are achieved for substrate current calculations. Self-consistent calculations of the drain current of a 60 nm-NMOSFET take about a few CPU hours demonstrating the feasibility of full-band Monte Carlo simulations.

***key words:*** *silicon, full-band Monte Carlo, microscopic relaxation time, velocity overshoot, impact ionization, drift-diffusion, deep submicron NMOSFET*

## 1. Introduction

The ongoing minimization of silicon devices has led to an increase in all kinds of hot-electron effects [1]–[3]. This in turn stimulated the development of silicon device modeling, and the full-band Monte Carlo (FB-MC) model is currently regarded as one of the most accurate simulation methods within the framework of semi-classical device physics [4]–[8]. For example, the first satisfactory microscopic impact ionization (II) model for electrons in silicon was based on the FB-MC model [9]. Application of similar models made it possible to calculate the substrate current of NMOSFETs with unprecedented accuracy [10], [11]. This was achieved without fitting any parameters on device level for various MOS technologies demonstrating the predictive capabilities of FB-MC device simulations [12]. Another important feature of the FB-MC model is that it captures the full anisotropy of the band structure, which plays a significant role in the quasi-ballistic transport in deep submicron MOSFETs. In the case of holes the warped band structure makes it often difficult to apply simpler modeling approaches at all. Especially for PMOSFETs with strained SiGe or Si layers, FB-MC is currently the only simulation method that can be expected to be predictive [13], [14].

Despite its considerable success the FB-MC method is still not widely used and most times it is only applied to exemplary simulations. This is due to the prevailing notion that FB-MC simulations are still as prohibitively CPU intensive as they were 10 years ago [5]. But in the meantime faster computers and new algorithms have reduced the CPU time substantially [15]–[17]. Moreover, in the case of standard industrial applications it is often possible to employ approximations without sacrificing accuracy too much. Electron-electron scattering, for example, as a two particle interaction is extremely CPU intensive [5], but so far only for sub-band-gap impact ionization or gate currents of MOSFETs at low drain voltages a significant influence of this effect could be shown [18]–[20]. Another cause of the huge CPU times is the self-consistent solution of the FB-MC model and Poisson's equation. The high doping concentrations of modern devices require time steps for the forward-Euler scheme which are of the order of 0.1 fs [21]. Since the times between scattering events are much longer than that, the MC method becomes increasingly inefficient at short time steps. This is avoided by nonself-consistent MC simulations, where the MC simulation is based on a fixed electric field calculated with a classical numerical device simulator based on the drift-diffusion (DD) or hydrodynamic model [22]. In Ref. [11] it is shown that this approximation does not distort the hot-electron distribution. Despite neglecting electron-electron scattering and self-consistency it was possible to calculate the substrate current of deep submicron devices over up to 8 orders of magnitude in good agreement with experimental results [10], [11]. The corresponding CPU times of the FB-MC model were as low as a CPU minute and of the same magnitude as the CPU time used by the numerical device simulator for the calculation of the electric field [11]. Thus, the application of dedicated FB-MC models is feasible in the design process of deep submicron devices and useful beyond exemplary simulations.

Manuscript received October 5, 1998.
Manuscript revised November 19, 1998.

[†] The authors are with the University of Bremen, Postfach 330 440, Kufsteiner Straße, 28334 Bremen, Germany.
a) E-mail: junge@item.uni-bremen.de

In this work a detailed description of efficient FB-MC methods is given, and the paper is organized as follows. In Sect. 2 a method for the generation of nonuniform tetrahedral grids is presented, which minimizes the discretization error and the memory requirements, and in Sect. 3 a consistent discretization of the inverse mass tensor is given. A fast algorithm for particle scattering in FB structures is described in Sect. 4 and efficient estimators for transport parameters are given in Sect. 5. In Sect. 6 the important properties of the device simulator are presented. In Sect. 7 the efficiency and usefulness of the FB-MC model are demonstrated by application to various examples including an NMOSFET with a metallurgical channel length of 60 nm.

## 2. *K*-Space Grid

Four conduction bands and three valence bands are calculated with the nonlocal empirical pseudopotential method [23]. The band energy is piecewise linearly interpolated in the $\boldsymbol{k}$-space which is partitioned into tetrahedra [15], [16], [24]. The energy $\varepsilon$ as a function of the $\boldsymbol{k}$-vector within the $i$-th tetrahedron reads:

$$\varepsilon(\boldsymbol{k}) = \varepsilon_i + \boldsymbol{v}_i^{\mathrm{T}} \hbar (\boldsymbol{k} - \boldsymbol{k}_i) \ , \tag{1}$$

where $\hbar$ denotes Planck's constant divided by $2\pi$ and the constants $\varepsilon_i$, $\boldsymbol{v}_i$ and $\boldsymbol{k}_i$ are calculated on condition that the energy interpolation must reproduce the exact band energy at the four corners of the tetrahedron. Since two contiguous tetrahedra share three vertices, the energy interpolation is continuous. The irreducible wedge is discretized with nonuniform tetrahedra, which completely fill the wedge and exactly resolve its shape.

For each energy band a nonuniform tetrahedral mesh is generated with an adaptive method somewhat similar to the ones described in Refs. [15], [25]. Each grid is created starting with the basic tetrahedra of the irreducible wedge. Subsequently, the grid is refined by splitting edges of the tetrahedra. A new node is inserted in the middle of an edge, and all tetrahedra, which contain this particular edge, are split. The edges to be split are selected according to the following rules. The error in the linear interpolation of the energy must be less than 2% of the energy at the center of the edge or less than 50 $\mu$eV, whatever is larger. Furthermore, the maximum length of an edge of a tetrahedron must be less than one tenth of the distance from $\Gamma$ to $X$ in the Brillouin zone. Among the edges violating these conditions the longest one is marked for splitting. Before the marked edge is split, it is checked, whether the volume quality $Q_{\mathrm{V}}$ of one of the tetrahedra containing this edge is less than 5%. If this is the case, the longest edge among all edges of the tetrahedra containing the marked edge is split instead of the marked one. This ensures convergence of the algorithm and reduces the loss in volume quality. $Q_{\mathrm{V}}$ is defined by:

$$Q_{\mathrm{V}} = 6\sqrt{2}\frac{V_{\mathrm{tet}}}{h_{\mathrm{max}}^3} \ , \tag{2}$$

where $V_{\mathrm{tet}}$ is the volume of the tetrahedron and $h_{\mathrm{max}}$ its maximum edge length [25]. The refinement is stopped when all edges satisfy the above criteria.

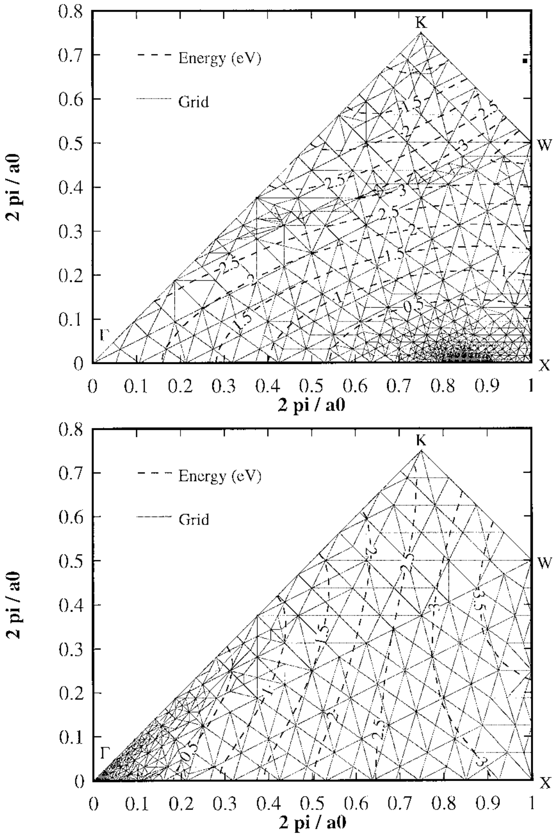


**Fig. 1** Nonuniform adaptive tetrahedral mesh of the first conduction band (upper graph) and of the first valence band (lower graph) in the $k_x, k_y$-plane for silicon.

In Fig. 1 the grid of the first conduction and valence band in the $k_x, k_y$-plane are shown, which are extremely dense in the vicinity of the band minima to allow transport calculations at temperatures as low as 50 K. The shortest edge of all tetrahedra is 0.0003 times the distance from $\Gamma$ to $X$. The average volume quality of the tetrahedra is 0.26.

The energy is symmetric in the $LUWK$-surface of the irreducible wedge with respect to the $LW$-line (cf. Fig. 2). Particles changing the Brillouin zone via this surface appear in the adjacent wedge at a position, which is on the opposite side of the $LW$-line compared to the initial position. To avoid a discontinuity in the energy interpolation the $\boldsymbol{k}$-space grid must have the same symmetry properties as the band structure (Fig. 2).

This $\boldsymbol{k}$-space discretization has several advantages [15], [16] compared to other methods [5], [26]. The shape of the irreducible wedge is exactly resolved by the nonuniform tetrahedral grid. The nonuniform grid

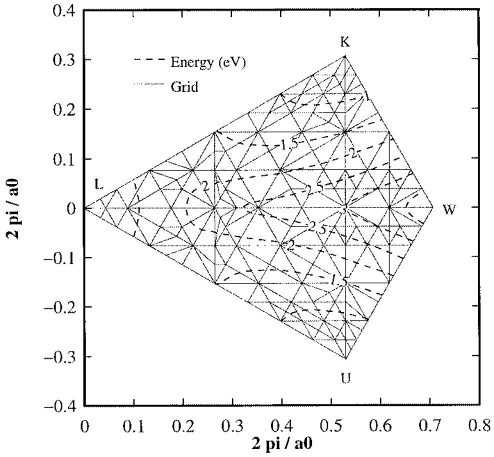


**Fig. 2** Symmetric grid of the *LUWK*-plane of the first conduction band of silicon.

can be locally refined without introducing an excessive number of grid nodes. In contrast to higher order schemes [5], [26] the linear interpolation allows to calculate the density of states ($DOS$) without any further approximations, and the $DOS$ of a single tetrahedron is a piecewise parabolic function of the energy. Since the particle velocity within a tetrahedron and the electric field within a primitive of the real space mesh are constant, the integration of the equations of motion can be performed exactly, even in the case of a nonzero magnetic field. The resultant formulas are linear in time and can be inverted easily. Thus, the solution of the equations of motion is even simpler than the one in the case of an analytical band structure which is at least quadratic in time. This is one of the reasons why the FB-MC models presented in Refs. [16], [17] are at least as efficient as analytical band models.

## 3. Mass Tensor Calculation

The inverse mass tensor is required for hydrodynamic models [27] and for the evaluation of various transport coefficients [28]. It is the derivative of the group velocity:

$$\underline{\underline{m}}^{-1} = \frac{1}{\hbar}\nabla_{\boldsymbol{k}}\boldsymbol{v}^{\mathrm{T}} \ , \tag{3}$$

which in turn is the derivative of the band energy:

$$\boldsymbol{v} = \frac{1}{\hbar}\nabla_{\boldsymbol{k}}\varepsilon \ . \tag{4}$$

Due to the linear interpolation (Eq. (1)) the velocity is constant within a tetrahedron and discontinuous between the tetrahedra. The inverse mass tensor is, therefore, Dirac-function-like. To circumvent the problem of sampling such a kind of quantity the inverse mass tensor is averaged over the volume of a tetrahedron. Based on the theorem of Gauß it can be shown that the integral of the inverse mass tensor $\underline{\underline{m}}^{-1}$ over the volume $V_{\mathrm{tet}}$ of a tetrahedron equals the integral of the group velocity $\boldsymbol{v}$ over the surface $S_{\mathrm{tet}}$ of the tetrahedron:

$$\frac{1}{V_{\mathrm{tet}}}\int_{V_{\mathrm{tet}}} \underline{\underline{m}}^{-1}\,\mathrm{d}V = \frac{1}{\hbar V_{\mathrm{tet}}}\oint_{S_{\mathrm{tet}}} \boldsymbol{v}\,\mathrm{d}\boldsymbol{S}^{\mathrm{T}} = \underline{\underline{m}}_{\mathrm{tet}}^{-1} \ . \tag{5}$$

Since the velocity is discontinuous at the surface of the tetrahedron, the surface integral is evaluated with the arithmetical mean of the velocities on both sides of the interface:

$$\oint_{S_{\mathrm{tet}}} \boldsymbol{v}\,\mathrm{d}\boldsymbol{S}^{\mathrm{T}} = \sum_{j=1}^{4} \frac{\boldsymbol{v}_j + \boldsymbol{v}_0}{2}\boldsymbol{A}_j^{\mathrm{T}} \ , \tag{6}$$

where $\boldsymbol{v}_j$ is the velocity in the $j$-th neighboring tetrahedron, $\boldsymbol{A}_j$ is the outwardly orientated surface area vector and $\boldsymbol{v}_0$ is the velocity of the tetrahedron for which the inverse mass tensor is calculated. The sum extends over all four adjacent tetrahedra. This yields for each tetrahedron a symmetric inverse mass tensor consistent with the nonuniformly discretized band structure.

## 4. Efficient Particle Scattering

One of the most CPU intensive parts of an FB-MC program can be particle scattering. In this work impact ionization and scattering with phonons and impurities are taken into account. Phonon scattering is modeled similar to Ref. [6] but without adding new processes to the model of Ref. [29] as described in Refs. [30], [31]. This model reproduces experimental results over a wide range of temperatures and electric fields [31] and the resultant distribution functions agree well with the outcome of other FB-MC models [30]. Impact ionization is described with energy dependent scattering rates, which are similar to the models presented in Refs. [32], [33], and experimental data of the quantum yield and impact ionization coefficient are well reproduced [30]. In addition, the model has been extensively verified on device level [12]. The impurity scattering model is described in Ref. [31], which includes an empirical adjustment to obtain the correct doping dependent mobility. The variable Γ-scheme is employed to reduce the number of self-scattering events [34].

In contrast to the other scattering models the transition rate of the impurity scattering model strongly depends on the momentum transfer. Since the efficiency of the method for particle scattering mentioned below degrades in the case of an anisotropic transition rate, an approximation is employed for impurity scattering based on the microscopic relaxation time. The tensor valued microscopic relaxation time $\underline{\tau}$ is defined by [28], [35]:

$$\boldsymbol{v}_b^w(\boldsymbol{k}) = \frac{\Omega}{(2\pi)^3} \sum_{b',w'} \int_{\text{wedge}} W_{b',b}^{w',w}(\boldsymbol{k}', \boldsymbol{k}) \times \left[ \underline{\underline{\tau}}_b^w(\boldsymbol{k}) \boldsymbol{v}_b^w(\boldsymbol{k}) - \underline{\underline{\tau}}_{b'}^{w'}(\boldsymbol{k}') \boldsymbol{v}_{b'}^{w'}(\boldsymbol{k}') \right] \mathrm{d}^3k' \ . \quad (7)$$

$\Omega$ is the system volume, $b$ is the energy band index, $w$ is the index of the irreducible wedge, and $W$ is the transition rate. Due to the employed scattering models the tensor valued microscopic relaxation time here reduces to an energy dependent scalar. For the sake of clarity and simplicity in the following the energy band index and the wedge index will be dropped. The anisotropic transition rate $W$ is approximated by an isotropic rate $\widehat{W}$ consisting of a normalized energy conserving Dirac-function multiplied with the inverse of the microscopic relaxation time [36], [37]:

$$\widehat{W}(\boldsymbol{k}', \boldsymbol{k}) = \frac{1}{\tau(\varepsilon(\boldsymbol{k}))} \times \frac{\delta(\varepsilon(\boldsymbol{k}') - \varepsilon(\boldsymbol{k}))}{\frac{\Omega}{(2\pi)^3} \int \delta(\varepsilon(\boldsymbol{k}'') - \varepsilon(\boldsymbol{k}))\, \mathrm{d}^3k''} \ . \quad (8)$$

This approximation *exactly* reproduces the low-field mobility and the comparison of simulations with and without this approximation yielded no significant difference. The advantages of this approximation are the exclusively energy dependent transition rate[†] and a reduced scattering rate, which improves the efficiency of the FB-MC simulation.

The above mentioned scattering models are all based on a transition rate of the following form:

$$W(\boldsymbol{k}', \boldsymbol{k}) = C(\varepsilon(\boldsymbol{k}))\delta(\varepsilon(\boldsymbol{k}') - \varepsilon(\boldsymbol{k}) - \varepsilon_{\text{trans}}) \ , \quad (9)$$

where $\varepsilon_{\text{trans}}$ is the energy gain/loss of the transition (eg: phonon energy) and the function $C$ depends on the scattering process[††]. Thus, the probability density of the states after scattering is uniform on an equienergy surface in a given band and wedge. This allows to employ very efficient methods for the selection of the final state in $\boldsymbol{k}$-space, which is often the most CPU intensive part of an FB-MC program.

The selection of the final state in $\boldsymbol{k}$-space consists of three parts, the determination of the final band and wedge, the selection of the final tetrahedron and the generation of the final $\boldsymbol{k}$-vector within the tetrahedron. The final band and wedge are determined with the usual methods [29]. The probability of the $i$-th tetrahedron in the final wedge and band is:

$$\begin{aligned} P_i(\varepsilon_{\text{final}}) &= \frac{\int_{V_{\text{tet},i}} W(\boldsymbol{k}', \boldsymbol{k})\, \mathrm{d}^3k'}{\int_{\text{wedge}} W(\boldsymbol{k}', \boldsymbol{k})\, \mathrm{d}^3k'} \\ &= \frac{\int_{V_{\text{tet},i}} \delta(\varepsilon_{\text{final}} - \varepsilon(\boldsymbol{k}'))\, \mathrm{d}^3k'}{\int_{\text{wedge}} \delta(\varepsilon_{\text{final}} - \varepsilon(\boldsymbol{k}'))\, \mathrm{d}^3k'} \\ &= \frac{DOS_i(\varepsilon_{\text{final}})}{\sum_{j \in \text{wedge}} DOS_j(\varepsilon_{\text{final}})} \ . \end{aligned} \quad (10)$$

The final energy of the particle $\varepsilon_{\text{final}}$ is the sum of the initial energy $\varepsilon(\boldsymbol{k})$ and the energy transfer $\varepsilon_{\text{trans}}$. The $DOS$ of a tetrahedron reads:

$$DOS_i(\varepsilon) = \frac{1}{(2\pi)^3\hbar} \frac{A_i(\varepsilon)}{\|\boldsymbol{v}_i\|} \ . \quad (11)$$

$A_i$ is the area of the intersection of the equienergy surface and the $i$-th tetrahedron. Since the group velocity is constant within a tetrahedron, the equienergy surface is a plane and the intersections are either triangles or quadrangles depending on the final energy [24]. In the case that the final energy is outside of the tetrahedron the area $A$ is treated as zero. The evaluation of the denominator of Eq. (10) is very CPU intensive, because the number of tetrahedra can be as large as 22000. To reduce this number lists are employed [5]. The possible final energies are divided into intervals of 10 meV width. For each energy interval a list of all tetrahedra is set up, which overlap with the energy interval. This reduces the number of final tetrahedra considerably, but the number is still too high. To avoid the evaluation of the sum over all tetrahedra of the corresponding list the rejection technique is employed [17]. Instead of using the direct method with the probabilities given by Eq. (10) the final tetrahedron $i$ is chosen from the corresponding list with a uniform probability. But this tetrahedron is only accepted, if the following condition is true:

$$DOS_i(\varepsilon_{\text{final}}) \geq r \max_j \left[DOS_j(\varepsilon_{\text{final}})\right] \ , \quad (12)$$

where $r$ is a uniformly distributed random number between 0 and 1. If this test fails, the chosen tetrahedron is rejected, and the whole procedure is repeated until a tetrahedron is accepted. The search for the maximum in Eq. (12) still extends over all tetrahedra. But for the rejection technique only an upper bound of the $DOS$ is required. Here, upper bounds ($\overline{DOS}$) are used, which are upper bounds of the maximum $DOS$ for consecutive energy intervals of 1 meV width. This list of the

[†] The reader should keep in mind that the microscopic relaxation time is calculated with the momentum-dependent transition rate (7) and that the approximation (8) is correct within the first order of the Cramers-Moyal expansion of the scattering integral [38].

[††] The presented methods can be extended to more general transition rates. A function $C$ depending on the initial and final momentum can be included by using the rejection technique [29], although this is not advisable in the case of impurity scattering due to the extremely anisotropic scattering rate. It is also possible to include a momentum dependent energy transfer by using a generalized version of the below described rejection technique for the selection of the final tetrahedron. But these extensions of the presented algorithm degrade the performance of the FB-MC program. Using the phonon system of Ref. [5] with inelastic acousitc phonons in conjunction with the extended version of the algorithm described in this work as, for example, done in Ref. [39] increases the CPU time of device simulations by about a factor of 4.

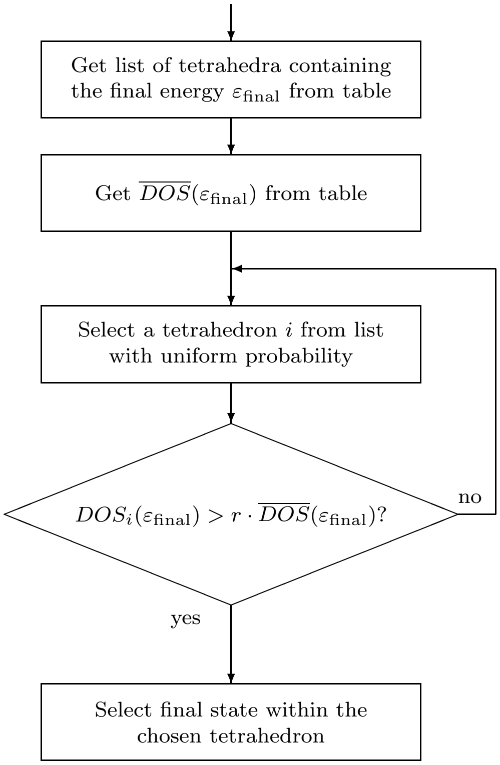

Get list of tetrahedra containing the final energy $\varepsilon_{\rm final}$ from table
Get $\overline{DOS}(\varepsilon_{\rm final})$ from table
Select a tetrahedron $i$ from list with uniform probability
$DOS_i(\varepsilon_{\rm final}) > r \cdot \overline{DOS}(\varepsilon_{\rm final})$?
no
yes
Select final state within the chosen tetrahedron

**Fig. 3** Rejection technique for the selection of the final state.

$\overline{DOS}$ is built only once in the initialization phase of the FB-MC program. Thus, the CPU intensive search for an upper bound during the simulation is avoided. The flowchart of this method is shown in Fig. 3. The average number of rejection steps is 13 for electrons and 18 for holes. This number is significantly smaller than the number of final tetrahedra, and therefore, the rejection method is much faster than the direct method. Since the number of rejection steps depends on the ratio of the maximum $\overline{DOS}$ and the average $DOS$, it does not depend on the total number of tetrahedra. Therefore, an increase of the total number of tetrahedra due to, for example, a grid refinement does not necessarily result in an increase of CPU time as long as the $DOS$ ratio is not increased.

After a tetrahedron has been chosen the final state within the tetrahedron is selected. In the case that the intersection is a quadrangle the equienergy surface is divided into two triangles. One of the two triangles is randomly selected with a probability proportional to the area of the triangle. The state within the triangle is selected with the usual direct method to generate uniformly distributed random points in a triangle.

The memory requirement of the FB structure together with the lists for particle scattering is less than 30 MB. The small size of the FB data is important for simulations of SiGe, where the memory requirement is 70 MB for each SiGe alloy due to the three times larger wedge.

## 5. Efficient Estimators

Many experiments, which are used for the calibration of the FB-MC model, are performed under equilibrium conditions or close to it, where standard MC methods are notoriously inefficient. Moreover, hydrodynamic models require transport coefficients, which are experimentally inaccessible and thus have to be calculated with the FB-MC model. Since these coefficients are needed for a wide range of parameters, efficient estimators are required to reduce the CPU times. With the above given microscopic relaxation time (Eq. (7)) estimators can be constructed which are efficient even in the case of equilibrium. The estimator for the drift mobility in the case of an arbitrary electric field reads [40]:

$$\underline{\underline{\mu}} = q \left\langle \tau \underline{\underline{m}}^{-1} + \frac{\mathrm{d}\tau}{\mathrm{d}\varepsilon} \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}} \right\rangle , \tag{13}$$

where $\langle\,\rangle$ denotes the expected value of the distribution function and $q$ is electron charge. This estimator is readily extended to the case of a nonzero magnetic field and can be used for the calculation of the Hall factor [41]. The mobility of the energy current is given by:

$$\underline{\underline{\mu}}_{\varepsilon} = q \left\langle \tau \varepsilon \underline{\underline{m}}^{-1} + \frac{\mathrm{d}\tau}{\mathrm{d}\varepsilon} \varepsilon \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}} + \tau \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}} \right\rangle . \tag{14}$$

The corresponding diffusion constants for a homogeneous system are:

$$\underline{\underline{D}} = \left\langle \tau \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}} \right\rangle , \tag{15}$$

and

$$\underline{\underline{D}}_{\varepsilon} = \left\langle \tau \varepsilon \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}} \right\rangle . \tag{16}$$

In the case of equilibrium an MC simulation can be avoided, because the distribution function is known, and the transport coefficients are evaluated by numerical integration, which is much faster than an MC simulation.

The statistics are either sampled with before-scattering statistics [29] or at the end of each time step of the self-consistent solution of the FB-MC model and Poisson's equation. The statistics of rare events are enhanced with population control methods [42], [43]. To ensure the reliability of the results and to minimize the CPU times the convergence of the statistics is estimated with the methods reported in Ref. [44].

## 6. Device Simulation

The device is discretized with a tensor-product grid. The electrostatic potential is either calculated self-consistently [21] or is imported from a numerical device simulator [22]. In the self-consistent case Poisson's equation is discretized with the box integration

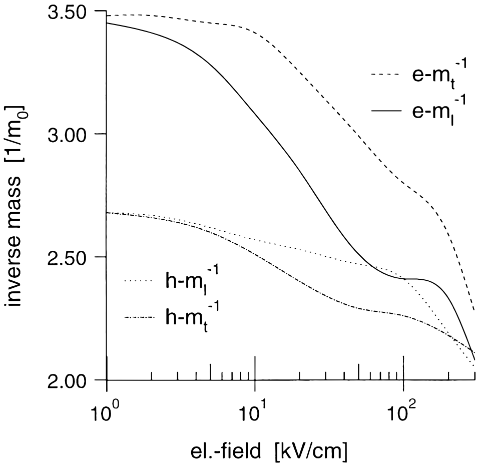


**Fig. 4** Expected values of the longitudinal and transversal inverse mass for electrons and holes in undoped silicon at room temperature as a function of the electric field in ⟨100⟩-direction.

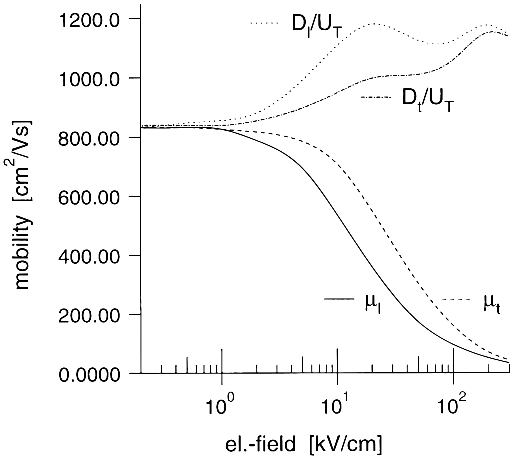


**Fig. 5** Longitudinal and transversal mobility and diffusion constant (divided by $U_T = k_\mathrm{B}T/q$, $T = 300\,\mathrm{K}$) for electrons in silicon with a doping of $10^{17}\,\mathrm{cm}^{-3}$ at room temperature as a function of the electric field in ⟨100⟩-direction.

method and the resultant band matrix is diagonal dominant [45]. Thus, a pivot search is not necessary and the band structure is preserved during the LU-decomposition. This considerably reduces the memory requirements and the CPU times. Since electrons and holes are simulated, Poisson's equation is linear and the LU-decomposition can be done at set-up. During the simulation only the backward substitution has to be performed, which consumes just about one percent of the total CPU time of a self-consistent MC device simulation [46]. Compared to iterative methods (eg: SLOR, Fast-Fourier, ... [21]) this method has three advantages: (i) it is accurate, (ii) it is much simpler to program (about 40 lines of code), (iii) it is faster.

Scattering at the rough Si/SiO$_2$-interface is treated with the usual mix of reflective and diffusive scattering [47]. The reflective scattering is calculated under the condition of parallel momentum and energy conservation. In contrast to electrons this does not necessarily lead to specular reflection in the case of holes because of the warped band structure. This reduces the hole mobility in addition to surface-roughness scattering [14]. Good agreement with experiment [48] is obtained for electrons and holes [13], [49].

## 7. Results

In Fig. 4 the expected value of the inverse mass tensor based on Eq. (5) is shown for electrons and holes in undoped silicon at room temperature as a function of the electric field. The anisotropy and dependence on the electric field are more pronounced in the case of electrons. Electron mobilities and diffusion constants are shown in Fig. 5 for a doping concentration of $10^{17}\,\mathrm{cm}^{-3}$ at room temperature. At low fields the Einstein relation is reproduced with less than 1% error demonstrating the high quality of the mass tensor discretization.

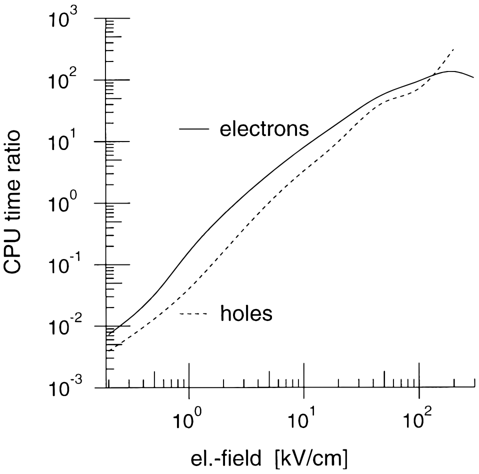


**Fig. 6** CPU time ratio of longitudinal mobility simulations based on the drift mobility and drift velocity for electrons and holes in undoped silicon at room temperature as a function of the electric field in ⟨100⟩-direction.

The longitudinal drift mobility, for example, can be either estimated by averaging the drift velocity ($\mu_\mathrm{l} = \left\langle \boldsymbol{v}^\mathrm{T} \right\rangle \boldsymbol{E}/\boldsymbol{E}^2$, where $\boldsymbol{E}$ is the vector of the electric field) or by sampling the mobility with Eq. (13) ($\mu_\mathrm{l} = \boldsymbol{E}^\mathrm{T} \underline{\underline{\mu}} \boldsymbol{E}/\boldsymbol{E}^2$). In Fig. 6 the ratios of the CPU times to achieve the same simulation error with both estimators are shown for the cases of electrons and holes in undoped silicon at room temperature. In the case of low electric fields the mobility estimator is more efficient, because it is based on even moments of the distribution function. Whereas the estimator employing the drift velocity is based on odd moments. Similar results are obtained for other transport parameters of the hydrodynamic model. Thus, it is possible to sample transport parameters in the "warm"-electron regime, where standard estimators are inefficient. The CPU

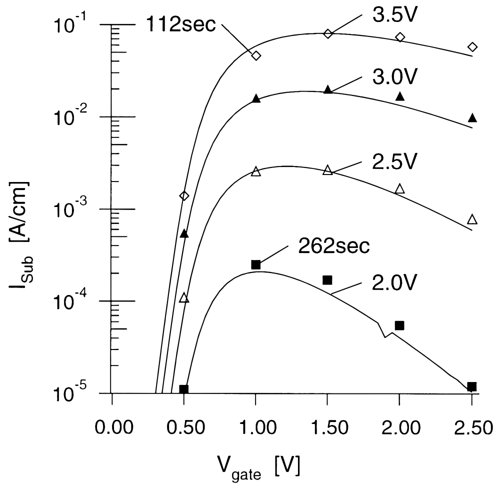
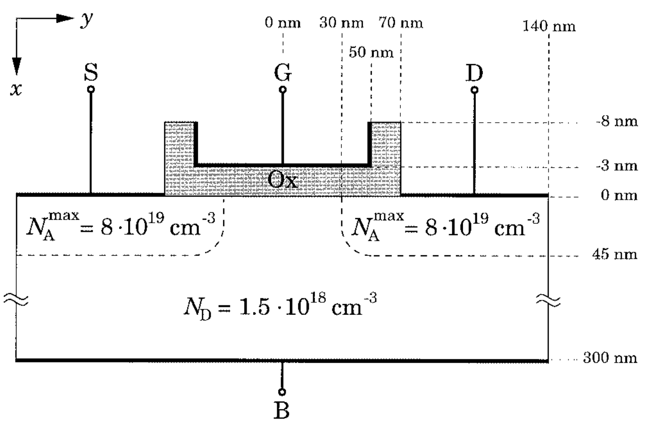


**Fig. 7** Substrate currents for a 0.18 $\mu$m-NMOSFET for different drain voltages at room temperature (lines: experiment, symbols: FB-MC). The CPU times are for a 60 MHz SUN SUPERSPARC.



y
x
0 nm
30 nm
70 nm
50 nm
140 nm
S
G
D
Ox
-8 nm
-3 nm
0 nm
45 nm
300 nm
$N_A^{max} = 8 \cdot 10^{19}$ cm$^{-3}$
$N_A^{max} = 8 \cdot 10^{19}$ cm$^{-3}$
$N_D = 1.5 \cdot 10^{18}$ cm$^{-3}$
B

**Fig. 8** NMOSFET with a metallurgical channel length of 60 nm.

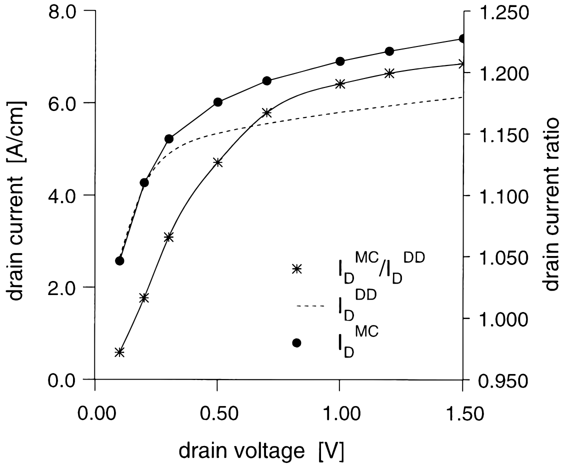


**Fig. 9** Drain current of the 60 nm-NMOSFET for a gate voltage of 1.5 V based on the FB-MC and DD model.

times for sampling the longitudinal mobility (and simultaneously all the other transport parameters) on a 300 MHz SUN ULTRA 2 with an error of ±1% with a probability of 95.45% are 123 s for an electric field of 0.1 kV/cm and 98 s for 300 kV/cm. These low CPU times are a prerequisite for the parameter extraction of hydrodynamic models, especially in the case of SiGe, because thousands of simulations must be performed for different electric fields, lattice temperatures and Ge-contents [50].

Substrate current is an important monitor for MOSFET degradation [2], [3]. Substrate current simulations are used during the device design process to minimize the hot-carrier degradation and thus to reduce the number of test-wafer cycles. Therefore, the substrate current model must be as reliable as possible. The best available model is currently the FB-MC model, and it has been verified for CMOS technologies covering the range from 256 kBit to 1 GBit DRAM generation [12]. In Fig. 7 experimental data and results of nonself-consistent FB-MC simulations are shown for a 0.18 $\mu$m-NMOSFET, and good agreement is obtained [17]. The simulations are converged within ±10% with a probability of 95.45%. The CPU times are a few CPU minutes on a 60 MHz SUN SUPERSPARC and comparable to the CPU time per bias point of the hydrodynamic model (90 CPU seconds). This example clearly demonstrates that FB-MC substrate current simulations are feasible in the device design process.

In the case of these nonself-consistent simulations only about 10% of the CPU time is spent for particle scattering. The largest fraction of the CPU time is used for the integration of the equations of motion, because the very strong electric fields in the channel of the MOSFET result in frequent changes of the tetrahedra in $\boldsymbol{k}$-space.

An NMOSFET with a metallurgical channel length of 60 nm (Fig. 8), a metal gate and an oxide thickness of 3 nm has been simulated with the FB-MC model and with the drift-diffusion model of the in-house simulator Galene III. $I_{off}$ at 1.5 V drain bias is 67 nA/cm. The drain current of the self-consistent simulation for a drain bias of 1.5 V is converged within ±2% with a probability of 95.45% in about 14 CPU hours on a 300 MHz SUN ULTRA 2 (10 ps, 87000 particles). In Fig. 9 the drain current versus drain voltage is shown for both models. The effective channel mobilities under equilibrium conditions are almost the same in both cases, and at a drain voltage of 0.1 V the FB-MC model reproduces the DD result within 3%. Although both models reproduce the same saturation velocity under homogeneous conditions, the drain current of the FB-MC model exceeds the one of the DD model by more than 20% for a drain voltage of 1.5 V. This is due to the quasi-ballistic transport, and in Fig. 10 the effec-

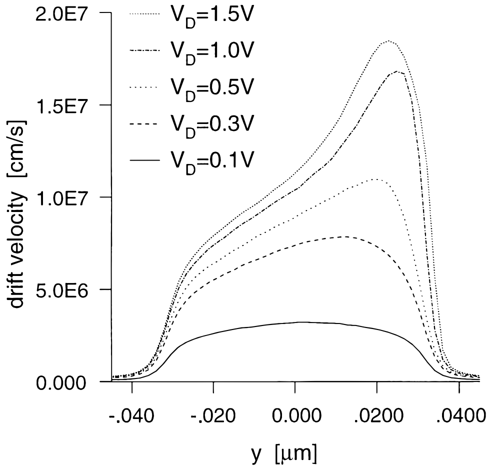


**Fig. 10** Effective drift velocities of the 60 nm-NMOSFET for a gate voltage of 1.5 V based on the FB-MC model.

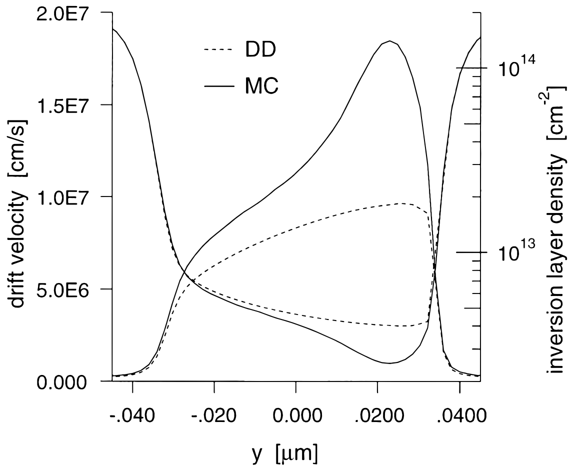


**Fig. 12** Effective drift velocities and inversion layer densities of the 60 nm-NMOSFET for a gate and drain voltage of 1.5 V based on the FB-MC and DD model.

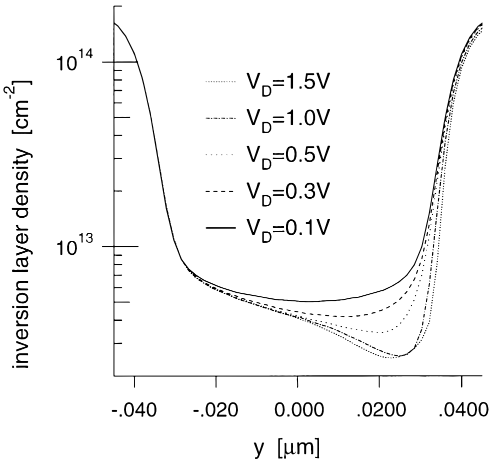


**Fig. 11** Inversion layer densities of the 60 nm-NMOSFET for a gate voltage of 1.5 V based on the FB-MC model.

tive drift velocities in the channel region are shown, which are defined as the quotient of the drain current and the inversion layer charge density. With increasing drain voltage a strong velocity overshoot appears at the end of the channel. But this velocity overshoot is not the cause of the higher drain current of the FB-MC model, because current continuity reduces the inversion layer particle density accordingly (Fig. 11) [51]. The difference in the drain current is due to the velocity overshoot ($v_{FB-MC} > v_{DD}$) at the beginning of the channel (Fig. 12). This velocity overshoot increases the drain current, because the inversion layer density is nearly the same at the beginning of the channel for both models. Since this velocity overshoot is smaller than the peak velocity overshoot, 1.2 compared to 1.8, the increase in drain current compared to the DD model is only 1.2 and not 1.8.

## 8. Conclusions

With the adaptive nonuniform tetrahedral grid in $\boldsymbol{k}$-space the problem of the discretization error of the FB structure has been solved without excessive memory requirements. The linear energy interpolation and the presented method for particle scattering enhance the efficiency of the FB-MC model beyond the level of analytical MC models. This makes it possible not only to generate with the FB-MC model the vast amount of data required for hydrodynamic models but also to facilitate the higher modeling accuracy of the FB-MC model directly during the device design process. In the case of nanoscale MOSFETs the use of the FB-MC model is not only mandatory for hot-electron problems like impact ionization, but becomes also more and more important for the accurate calculation of the drain current as well due to the quasi-ballistic transport and the anisotropic band structure.

## Acknowledgement

The authors thank S. Yamaguchi and H. Goto of Fujitsu Ltd, Kawasaki for many discussions and support. Partial funding by the Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie under contract number 01 M2416C is gratefully acknowledged.

### References

[1] T.H. Ning, P.W. Cook, R.H. Dennard, C.M. Osburn, S.E. Schuster, and H.-N. Yu, "1 $\mu$m MOSFET VLSI technology: Part IV—Hot-electron design constrains," IEEE Trans. Electron Devices, vol.26, pp.346–353, 1979.

[2] E. Takeda and N. Suzuki, "An empirical model for device degradation due to hot-carrier injection," IEEE Electron Device Lett., vol.4, pp.111–113, 1983.

[3] C. Hu, S.C. Tam, F.-C. Hsu, P.-K. Ko, T.-Y. Chan, and K.W. Terrill, "Hot-electron induced MOSFET degradation—Model, monitor, and improvement," IEEE Trans. Electron Devices, vol.32, pp.375–385, 1985.

[4] J.Y. Tang, H. Shichijo, K. Hess, and G.J. Iafrate, "Band-structure dependent impact ionization in silicon and gallium arsenide," Journal de Physique, vol.42, pp.63–69, 1981.

[5] 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, 1988.

[6] K. Hess, ed., "Monte Carlo Device Simulation: Full Band and Beyond," Kluwer, Boston, 1991.

[7] A. Abramo, L. Baudry, R. Brunetti, R. Castagne, M. Charef, F. Dessenne, P. Dollfus, R. Dutton, W.L. Engl, R. Fauquembergue, C. Fiegna, M.V. Fischetti, S. Galdin, N. Goldsman, M. Hackel, C. Hamaguchi, K. Hess, K. Hennacy, P. Hesto, J.M. Higman, T. Iizuka, C. Jungemann, Y. Kamakura, H. Kosina, T. Kunukiyo, S.E. Laux, H. Lin, C. Maziar, H. Mizuno, H.J. Peifer, S. Ramaswamy, N. Sano, P.G. Scrobohaci, S. Selberherr, M. Takenaka, T. Tang, J.L. Thobel, R. Thoma, K. Tomizawa, M. Tomizawa, T. Vogelsang, S. Wang, X. Wang, C. Yao, P.D. Yoder, and A. Yoshii, "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.

[8] M.V. Fischetti and S.E. Laux, "Monte Carlo simulation of electron transport in Si: The first 20 years," Proc. ESSDERC, Bologna, vol.26, pp.813–820, 1996.

[9] E. Cartier, M.V. Fischetti, E.A. Eklund, and F.R. McFeely, "Impact ionization in silicon," Appl. Phys. Lett., vol.62, pp.3339–3341, 1993.

[10] J.D. Bude and M. Mastrapasqua, "Impact ionization and distribution functions in sub-micron nMOSFET technologies," IEEE Electron Device Lett., vol.16, pp.439–441, 1995.

[11] C. Jungemann, S. Yamaguchi, and H. Goto, "On the accuracy and efficiency of substrate current calculations for sub-$\mu$m n-MOSFET's," IEEE Electron Device Lett., vol.17, pp.464–466, 1996.

[12] C. Jungemann, B. Meinerzhagen, S. Decker, S. Keith, S. Yamaguchi, and H. Goto, "Is physically sound and predictive modeling of NMOS substrate currents possible?" Solid-State Electron., vol.42, pp.647–655, 1998.

[13] S. Keith, C. Jungemann, and B. Meinerzhagen, "Full band Monte Carlo device simulation of 0.1–0.5 $\mu$m strained-Si P-MOSFETs," Proc. ESSDERC, Bordeaux, 1998, pp.312–315.

[14] C. Jungemann, S. Keith, and B. Meinerzhagen, "Full-band Monte Carlo simulation of a 0.12 $\mu$m-Si-PMOSFET with and without a strained SiGe-channel," Tech. Dig. IEDM, San Francisco (USA), pp.897–900, 1998.

[15] R. K. Smith and J. Bude, "Highly efficient full band Monte Carlo simulations," Proc. International Workshop on Computational Electronics, Leeds, Aug., pp.224–230, 1993.

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

[17] C. Jungemann, S. Yamaguchi, and H. Goto, "Efficient full band Monte Carlo hot carrier simulation for silicon devices," Proc. ESSDERC, Bologna, 1996, vol.26, pp.821–824.

[18] J.E. Chung, M.-C. Jeng, J.E. Moon, P.-K. Ko, and C. Hu, "Low-voltage hot-electron currents and degradation in deep-submicrometer MOSFET's," IEEE Trans. Electron Devices, vol.37, pp.1651–1657, 1990.

[19] M.V. Fischetti and S.E. Laux, "Monte Carlo study of sub-band-gap impact ionization in small silicon field-effect transistors," IEDM, 1995, pp.305–308.

[20] B. Fischer, A. Ghetti, L. Selmi, R. Bez, and E. Sangiorgi, "Bias and temperature dependence of homogeneous hot-electron injection from silicon into silicon dioxide at low voltages," IEEE Trans. Electron Devices, vol.44, pp.288–296, 1997.

[21] R.W. Hockney and J.W. Eastwood, "Computer Simulation Using Particles," Institute of Physics Publishing, Bristol, Philadelphia, 1988.

[22] J.M. Higman, K. Hess, C.G. Hwang, and R.W. Dutton, "Coupled Monte Carlo-drift diffusion analysis of hot-electron effects in MOSFET's," IEEE Trans. Electron Devices, vol.36, pp.930–937, 1989.

[23] M.M. Rieger and P. Vogl, "Electronic-band parameters in strained $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ alloys on $\mathrm{Si}_{1-y}\mathrm{Ge}_y$ substrates," Phys. Rev. B, vol.48, pp.14276–14287, 1993.

[24] G. Lehmann and M. Taut, "On the numerical calculation of the density of states and related properties," Phys. Status Solidi B, vol.54, pp.469–477, 1972.

[25] E.X. Wang, M.D. Giles, S. Yu, F.A. Leon, A. Hiroki, and S. Odanaka, "Recursive M-tree method for 3-D adaptive tetrahedral mesh refinement and its application to Brillouin zone discretization," Proc. SISPAD, Tokyo, pp.67–68, 1996.

[26] G. Wiesenekker and E. J. Baerends, "Quadratic integration over the three-dimensional brillouin zone," J. Phys.: Condens. Matter, vol.3, pp.6721–6742, 1991.

[27] R. Thoma, A. Emunds, B. Meinerzhagen, H.J. Peifer, and W.L. Engl, "Hydrodynamic equations for semiconductors with nonparabolic bandstructures," IEEE Trans. Electron Devices, vol.38, pp.1343–1352, 1991.

[28] O. Madelung, "Introduction to Solid-State Theory," Springer, Berlin, 1978.

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

[30] C. Jungemann, S. Keith, F.M. Bufler, and B. Meinerzhagen, "Effects of band structure and phonon models on hot electron transport in silicon," Electrical Engineering, vol.79, pp.99–101, 1996.

[31] F.M. Bufler, "Full-band Monte Carlo simulation of electrons and holes in strained Si and SiGe," Dissertation, Universität Bremen, Bremen, pp.881–884, 1997.

[32] T. Kunikiyo, M. Takenaka, M. Morifuji, K. Taniguchi, and C. Hamaguchi, "A model of impact ionization due to the primary hole in silicon for a full band Monte Carlo simulation," J. Appl. Phys., vol.79, pp.7718–7725, 1996.

[33] M.V. Fischetti, N. Sano, S.E. Laux, and K. Natori, "Full-band-structure theory of high-field transport and impact ionization of electrons and holes in Ge, Si, and GaAs," IEEE J. Tech. Comp. Aided Design, no.3, 1997.

[34] E. Sangiorgi, B. Riccò, and F. Venturi, "$\mathrm{MOS}^2$: An efficient monte carlo simulator for MOS devices," IEEE Trans. Comput.-Aided Des. Integrated Circuits & Syst., vol.7, pp.259–271, 1988.

[35] W. Brauer and H.W. Streitwolf, "Theoretische Grundlagen der Halbleiterphysik," Vieweg, Braunschweig, 2nd edition, 1977.

[36] S. Jallepalli, M. Rashed, W.-K. Shih, C.M. Maziar, and A.F. Tasch, Jr., "A full-band Monte Carlo model for hole transport in silicon," J. Appl. Phys., vol.81, pp.2250–2255, 1997.

[37] P. Graf, F.M. Bufler, B. Meinerzhagen, and C. Jungemann, "A comprehensive SiGe Monte Carlo model for transient 2D simulations of HBTs," IEDM Tech. Dig., pp.881–884, 1997.

[38] N.G. van Kampen, "Stochastic Process in Physics and Chemistry," North-Holland Publishing, Amsterdam, 1981.

[39] Y. Tagawa and Y. Awano, "Enhanced hole drift velocity in sub-0.1 $\mu$m Si devices caused by anisotropic velocity overshoot," Proc. IWCE, Osaka (Japan), pp.206–209, 1998.

[40] R. Stratton, "Diffusion of hot and cold electrons in semiconductor barriers," Phys. Rev., vol.126, pp.2002–2013, 1962.

[41] C. Jungemann, M. Bartels, S. Keith, and B. Meinerzhagen, "Efficient methods for Hall factor and transport coefficient evaluation for electrons and holes in si and sige based on a full-band structure," Proc. IWCE, Osaka (Japan), pp.104–107, 1998.

[42] C. Jungemann, S. Decker, R. Thoma, W.-L. Engl, and H. Goto, "Phase space multiple refresh: A general purpose statistical enhancement technique for Monte Carlo device simulation," IEEE J. Tech. Comp. Aided Design, no.2, 1997.

[43] M.G. Gray, T.E. Booth, T.J.T. Kwan, and C.M. Snell, "A multi-comb variance reduction scheme for Monte Carlo semiconductor simulators," IEEE Trans. Electron Devices, vol.45, pp.918–924, 1998.

[44] C. Jungemann, S. Yamaguchi, and H. Goto, "Convergence estimation for stationary ensemble Monte Carlo simulations," IEEE J. Tech. Comp. Aided Design, no.10, 1998.

[45] R.S. Varga, "Matrix Iterative Analysis," Series in Automatic Computation, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.

[46] H.-J. Peifer, "Monte-Carlo Simulation des Hochenergietransports von Elektronen in submikron MOS-Strukturen," Doctor thesis, RWTH Aachen, Aachen, Augustinus Buchhandlung, 1992.

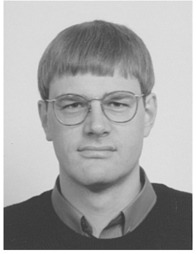
[47] E. Sangiorgi and M.R. Pinto, "A semi-empirical model of surface scattering for Monte Carlo simulation of silicon n-MOSFET's," IEEE Trans. Electron Devices, vol.39, pp.356–361, 1992.

[48] S. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the universality of inversion layer mobility in Si MOSFET's: Part I—Effects of substrate impurity concentration," IEEE Trans. Electron Devices, vol.41, pp.2357–2362, 1994.

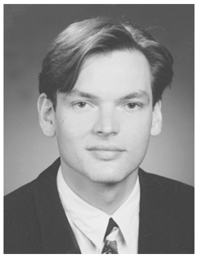
[49] S. Keith, F.M. Bufler, and B. Meinerzhagen, "Full band Monte-Carlo device simulation of an 0.1 $\mu$m n-channel MOSFET in strained silicon material," Proc. ESSDERC, Stuttgart, pp.200–203, Sept. 1997.

[50] B. Neinhüs, S. Decker, P. Graf, F.M. Bufler, and B. Meinerzhagen, "Consistent hydrodynamic and Monte-Carlo simulation of SiGe HBTs based on table models for the relaxation times," VLSI Design, vol.8, pp.387–391, 1998.

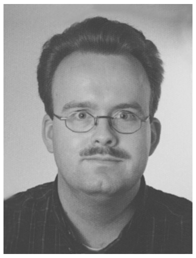
[51] M. Lundstrom, "Scattering theory of the short channel MOSFET," IEDM Tech. Dig., pp.387–390, 1996.



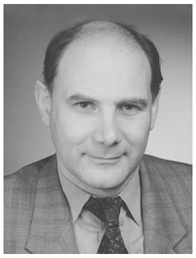
**Christoph Jungemann** received the Dipl.-Ing. degree and the Dr.-Ing. degree in electrical engineering in 1990 and 1995 from the RWTH Aachen (Technical University of Aachen), Aachen, Germany. From 1990 to 1995 he was a research and teaching assistant at the "Institut für Theoretische Elektrotechnik", RWTH Aachen. From 1995 until 1997 he was employed at the Research and Development facility of Fujitsu Limited, Kawasaki, Japan. Since 1997 he is a chief engineer at the "Institut für Theoretische Elektrotechnik and Mikroelektronik", University of Bremen. His main research interests are full-band Monte Carlo simulation of Si and SiGe devices, numerical device modeling and transport in inversion layers.



**Stefan Keith** received the Dipl.-Ing. degree in electrical engineering in 1994 from the RWTH Aachen (Technical University of Aachen), Aachen, Germany. Since then he was a research Assistant at the "Institut für Theoretische Elektrotechnik", RWTH Aachen, and moved in 1995 as a research and teaching Assistant to the "Institut für Theoretische Elektrotechnik and Mikroelektronik", University of Bremen. His current research interests are full-band Monte Carlo device simulation of Si-MOSFETs and SiGe based heterojunction devices.



**Martin Bartels** received his Dipl.-Ing. degree in electrical engineering in 1997 from the University of Bremen, Germany. Since then he is a research assistant at the "Institut für Theoretische Elektrotechnik und Mikroelektronik" (ITEM), University of Bremen. His work covers Monte Carlo simulation of the SiGe material system and related devices as well as numercial device and circuit simulation.



**Bernd Meinerzhagen** received the Dipl.-Ing. degree in electrical engineering in 1977, the Dipl.-Math. degree in mathematics in 1981, the Dr.-Ing. degree in electrical engineering in 1985 and the "venia legendi" in 1995 all from the RWTH Aachen (Technical University of Aachen), Aachen, Germany. From 1978 to 1986 as a research and teaching assistant at the RWTH Aachen he worked mainly on the development of numerical device modeling codes. In 1986 he joint AT&T Bell Laboratories in Allentown, PA, as a Member of Technical Staff, where he developed advanced numerical models for MOS substrate and gate currents. He went back to the RWTH Aachen in 1987 and became head of the research and development group for Silicon technology modeling and simulation (TCAD). In 1995 he was appointed Professor at the University of Bremen, where his current research interests include TCAD and the theory of electromagnetic fields and networks.