Solving Boltzmann Transport Equation without Monte-Carlo Algorithms – New Methods for Industrial TCAD Applications

B. Meinerzhagen*, A. T. Pham*, S.-M. Hong†, and C. Jungemann†
*BST, TU Braunschweig, 38023 Braunschweig, Germany
Email: b.meinerzhagen@tu-bs.de
†EIT4, Universität der Bundeswehr München, 85577 Neubiberg, Germany

***Abstract*— The Drift-Diffusion model is still by far the most frequently used numerical device model in industry today. One important reason for this success is the robust numerical implementation of this model providing CPU efficient DC, AC, transient, and noise simulations with high accuracy and high convergence reliability. On the other hand, many of todays design applications vary strain, crystal and channel orientation, material composition, and the carrier confinement. Such applications certainly require the solution of the Boltzmann Transport Equation in order to be predictive. It will be demonstrated in this paper that with new alternative discretization and solution methods avoiding the Monte-Carlo algorithm many of the favorable numerical properties of the traditional Drift-Diffusion model can be transferred to numerical device models that include the solution of the Boltzmann Transport Equation.**

I. Introduction

For the design of advanced field effect transistors (FETs) with channel length of 30 nm and less the classical Drift-Diffusion (DD) device model is still widely applied though it is well known that for such devices the DD model cannot provide physically accurate solutions [1]. One reason for this is certainly that alternative more accurate models are typically based on the solution of Boltzmann's Transport Equation (BTE) and therefore much slower. However, another very important reason is that the BTE is typically solved with a Monte Carlo (MC) algorithm. This implies that many of the favorable numerical properties of the DD model like an easy control of numerical accuracy and simulation time, the availability of true DC solutions and the Jacobian of the complete discrete equation system allowing simultaneous Newton iterations as well as AC and noise analysis are lost.

II. The MC CPU-Time accuracy trade off

In order to show that already a closer look on numerical accuracy can easily identify important problems that are hard to solve with a MC algorithm, the accurate calculation of the linear response at equilibrium for a NMOSFET in weak or strong inversion is studied first. This is an easy task for a DD TCAD model if the usual solution methods are used. However, for a MC algorithm the situation is completely different. For a self-consistent multi particle solution of BTE and Poisson's Equation (PE) with the MC algorithm, where all particles have equal weights, the following equation holds for the CPU-time T_{CPU} necessary to reach a relative numerical error r for the DC current I_D with a probability of 95% [2]:

$$T_{CPU} = \frac{\alpha_{cost}\; 8U_T\; Q_{tot}}{r^2\; V_D\; I_D} \tag{1}$$

U_T is the thermal voltage, Q_{tot} is the total electron charge within the simulated device and V_D is the drain to source voltage, which should be smaller than 1 mV in order to suppress nonlinear transport effects in nanoscale FETs. Moreover, α_{cost} is the ratio of the CPU time and the simulated time divided by the number of simulated particles. Thus, α_{cost} depends on the hardware available for the simulation and can be easily determined by running the MC algorithm for a short time. For one 1 GHz CPU $\alpha_{cost} \approx 2.5 \cdot 10^{10}$ holds. First of all and most important this formula shows clearly that $r \sim 1/\sqrt{T_{CPU}}$, which is valid for all MC simulations. Moreover, it becomes clear that T_{CPU} can vary easily by 10 orders of magnitude and more depending on the drain current and the simulation accuracy r required which is not the case for DD simulations, where thanks to the availability of Newton's method CPU-time is only a weak function of the required numerical accuracy. Finally, T_{CPU} becomes extremely large if the linear response must be calculated with high accuracy in the subthreshold region as for example for magnetotransport measurements. The relative magnetoresistive response may be as small as 10^{-3} as shown below so that r should be 10^{-5} in order to calculate the magnetoresistive response with an accuracy of 1%. For such a problem a MC solution is basically impossible. In case single particle or weighted particle MC algorithms are used formula (1) is not exactly valid any more but the basic trends described above remain still the same.

III. Direct Solution of the BTE

The solution methods for the BTE that avoid the MC algorithm, which will be addressed as direct BTE solution methods throughout the rest of this paper, are typically based on the expansion of the distribution function in spherical harmonics in the 3D k-space. This has a long history and so many scientists have contributed to this art that we can cite only a few pioneering papers [3]–[6]. The biggest disadvantage of this method is the high memory requirement if a 2D real space is considered. Therefore until recently only low order expansions were possible so that the accuracy of these BTE solution methods was still inferior compared to the MC alternative. The situation has completely changed in the last couple of years due to the availability of random access memories of 100 GByte and more, which make higher order expansions with very little residual error possible and allow direct self-consistent solution of the BTE and PE (BTE-PE system) [7] thanks to the existence of efficient linear

978-1-4244-7700-5/10/$26.00 ©2010 IEEE

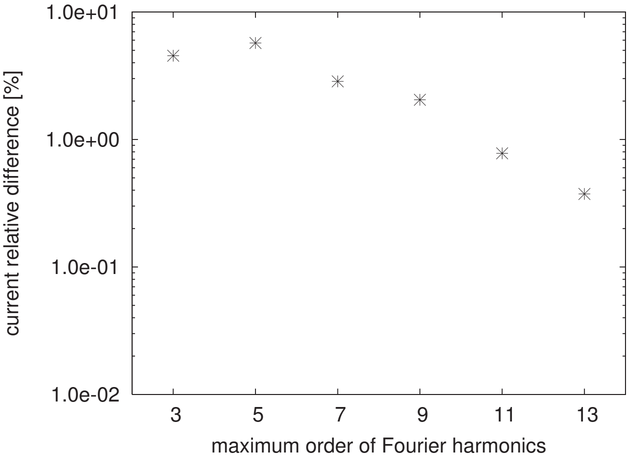

Fig. 1. Relative error of the DC drain current evaluated with lower order Fourier expansions in comparison to the current resulting for maximum order 15 for a Ge PMOS double gate transistor with 16 nm channel length (reprint from [9]).

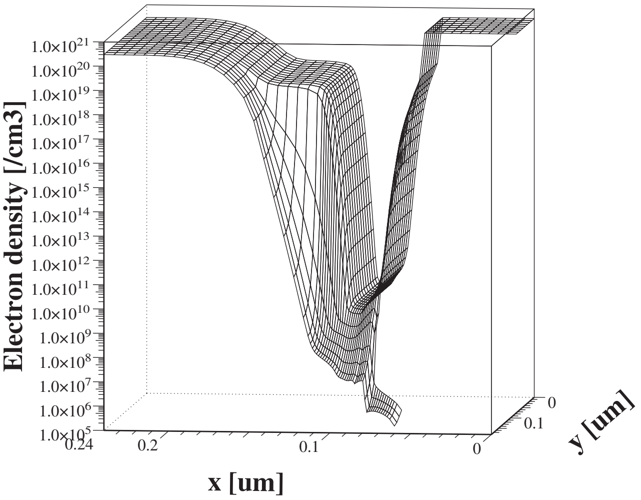

Fig. 2. Electron density in a SiGe HBT varying over 16 orders of magnitude on a coarse grid calculated with the maximum entropy dissipation scheme (reprint from [7]).

equation solvers [8]. Moreover even the multi subband BTE in inversion layers can now be solved by a direct method based on a Fourier expansion of the distribution function in the 2D k-space. Today already a simulator exists that solves the 1D Schrödinger Equation (SE), PE and the multi subband (MSB) BTE (BTE-PE-SE system) self-consistently for FET devices without using MC algorithms [9]. In order to show that simulators allowing harmonic expansions of the distribution function of variable order are really necessary even for fundamental unknowns in Fig. 1 the dependence of the drain current in an advanced Ge PMOSFET on the maximum Fourier expansion order is shown. The simulator solving the BTE-PE-SE system described in [9] was used for this study. Please note that a too low expansion order may introduce an error of nearly 10% even for the drain current. For direct BTE solution methods the box integration method can be used for the space discretization, which comparable to the DD model case allows to make the discretized transport model charge conservative [7]. This implies that for the discretized model Kirchhoff's current law holds for the terminal currents as soon as the transport equation is solved with sufficient accuracy. One key features of the classical DD TCAD model is the availability of the Scharfetter-Gummel discretization scheme for the particle current densities which makes numerically stable CPU efficient solutions on sparse space grids possible. Comparable discretization schemes have meanwhile been developed for the direct solution of the BTE as well and Fig. 2 shows the electron density in a SiGe HBT calculated with the maximum entropy dissipation scheme [10] and the H-transform [5] as described in [7], where the density varies smoothly over 16 orders of magnitude on a coarse space grid without showing any indications for instability. More details about the discretization methods of the BTE on which the direct solution methods are based can be found in [7], [9].

IV. Accuracy and CPU-Time for the direct methods

After discretization of the BTE within the k-space based on spherical or Fourier harmonics and in the real space based on the box integration method the discrete BTE can be solved without substantial problems even for two space dimensions due to the availability of efficient linear solvers [8] and large random access memories. Even if the BTE is nonlinear because the Pauli principle is considered, the Jacobian of the discrete BTE is readily available and Newton's method can be used to establish a solution of the BTE. A very important solution method for the DD TCAD model is Gummel's nonlinear relaxations scheme. A similar scheme can be applied for the direct solution of the BTE-PE or BTE-SE-PE systems by solving each equation successively one after the other and considering the coupling of the model equations iteratively by repeating this solution loop until convergence is established. Fig. 3 shows by the dashed line for a typical PMOS case the relative drain current error r as a function of the number of nonlinear Gummel type relaxations for the BTE-SE-PE system. In the log/linear plot shown in Fig. 3 an upper bound for this error is given by the straight dotted line. Since this dotted line represents a decaying exponential function and each relaxation step consumes exactly the same CPU-time it is clear that for this Gummel relaxation method $r < C_1 \cdot \exp(-C_2 \cdot T_{CPU})$ with positive constants C_1 and C_2 holds for the error r established after a total CPU-time T_{CPU}. This shows that for the direct solution methods the error improves at least exponentially with CPU-time. Therefore a high accuracy can be easily achieved in contrast to the alternative MC solution method where this is often impossible due to the square root dependence on CPU-time. The second very important solution algorithm for the DD TCAD model is the simultaneous Newton algorithm considering the complete coupling between all equations. This solution method has already been demonstrated for the BTE-PE system [7]. For the BTE-PE-SE system the consideration of the coupling between the equations is more difficult since

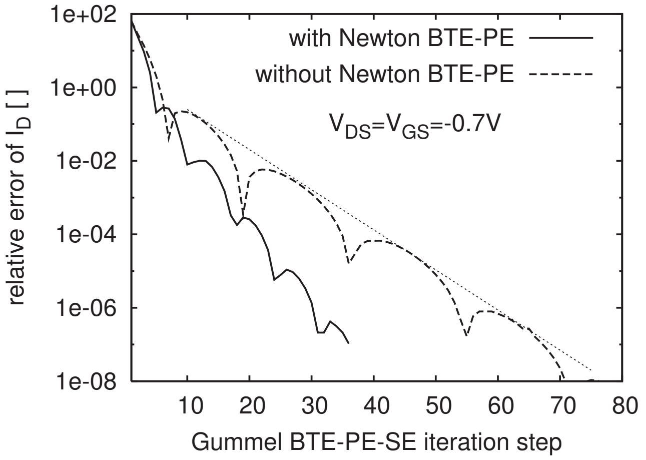

Fig. 3. Relative error of DC drain current versus number of relaxation loops for the BTE-PE-SE system and a Si Double Gate PMOSFET with 16 nm channel length. The dashed line refers to the Gummel type relaxation scheme and the drawn line refers to the relaxation method where the coupling between PE and the BTE is considered by one incomplete simultaneous Newton step as part of the relaxation loop.

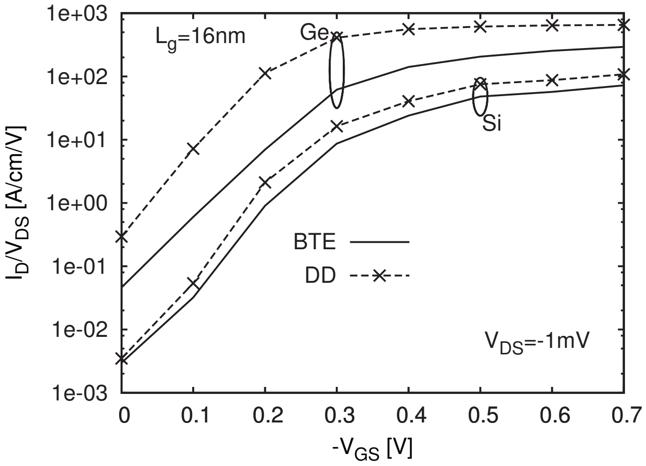

Fig. 4. Characteristic of channel self conductance evaluated based on the solution of the BTE-PE-SE system and formula (2) for two short channel double gate PMOSFETs. Two cases (unstrained Si ((001) surface/[110] channel) and uniaxially stressed Ge ((110) surface/[110] channel, the stress direction is parallel to the channel direction)) are considered. V_{DS} = 1mV.

SE is an eigenvalue problem which must be solved in order to establish the coupling between PE and BTE. First order perturbation theory can be used to evaluate this coupling in a linearized manner. Based on this idea an efficient incomplete simultaneous Newton method for the BTE-PE-SE was realized and reported in [11]. In Fig. 3 it can be seen by the drawn line that even this incomplete Newton scheme leads to a substantial improvement of convergence speed if it is combined with the Gummel like relaxation scheme.

V. Accuracy of calibrated DD

One successful idea to improve the accuracy of classical TCAD DD or hydrodynamic (HD) device models was the hierarchical modeling approach. The basis of this method is the extraction of the DD and HD transport parameters under homogeneous material and field conditions using a physically more accurate model based on the BTE and the application of these parameters in the TCAD models using a table model approach [2]. Of course such a calibration is at least partly possible as well using special test structures and experimental data. However, for very advanced FET and bipolar devices with channel lengths in the order of about 40 nm or less or base thicknesses in the order of 10 nm calibration does no longer lead to models with sufficient accuracy. This is demonstrated next for the DD model and and the linear response of ultrashort FETs at equilibrium. For simplicity 2D FETs that are homogeneous in width direction are considered. It can be shown that for the linear response drain self-conductance at equilibrium the following equation is an exact result following from the DD ansatz [1] even in the case of a degenerate particle gas:

$$g_D^{DD} = qW \left\{ \int \left[N_{\text{inv}}(x) \mu_{\text{eff}}(x) \right]^{-1} dx \right\}^{-1} \qquad (2)$$

The integration is performed along the channel from source to drain. N_{inv} is the inversion charge per unit area and μ_{eff} the effective mobility. This formula can be perfectly calibrated with respect to the BTE-PE-SE system if $\mu_{\text{eff}}(x)$ and $N_{\text{inv}}(x)$ are evaluated locally based on the results of the BTE-PE-SE system solver using the Kubo-Greenwood formula [1]. Fig. 4 compares the results of the calibrated formula (2) and the self-conductance at equilibrium resulting from the solution of the BTE-PE-SE system for a double gate PMOSFET with 16nm channel length and unstrained Si with (001) interface orientation or uniaxially strained Ge with (011) interface orientation as channel material, respectively. It can be seen that in the Si case the difference is typically larger than a factor of two and reaches an order of magnitude in the Ge case. Though the calibration of the DD model must be called ideal in this case large differences between both approaches are observable for these advanced devices. These differences are due to the DD ansatz itself, which is no longer valid for such devices [1]. Therefore, calibration of TCAD DD and HD models does in general not lead to accurate models for devices with such small dimensions.

VI. Hard problems for MC algorithms

Finally, the excellent numerical properties of the direct BTE solution methods are demonstrated for some examples which cannot be simulated using MC algorithms. As already pointed out in section II one such example is the magnetoresistive effect especially if this effect is small. For the direct BTE-PE-SE device solver presented in [9] such simulations with high numerical accuracy are no problem as shown in Fig. 5. Please note that for a small magnetic field the relative magnetoresistive response can be well below 0.1%. Another notoriously difficult simulation example even for larger device dimensions are SOI MOSFETs with high substrate doping levels. In order to describe the kink effect in these devices correctly, impact ionization (II) must be modeled with high accuracy. Therefore, the DD model yields incorrect results

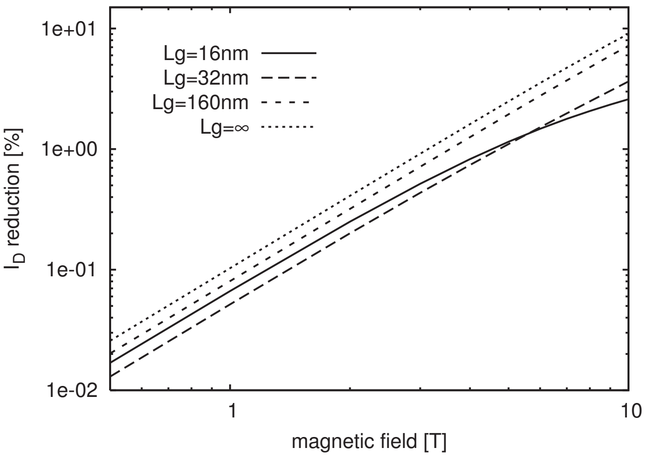

Fig. 5. Relative reduction of drain current due to the magnetoresistive effect as a function of magnetic field for biaxially strained Si double gate PMOSFETs with 10nm body thickness. VGS = 0.7V. VDS = 1mV for finite gate length ($L_g \neq \infty$) and 1V/cm lateral electric field for the homogenous channel case ($L_g = \infty$). T = 300K. (reprint from [9]).

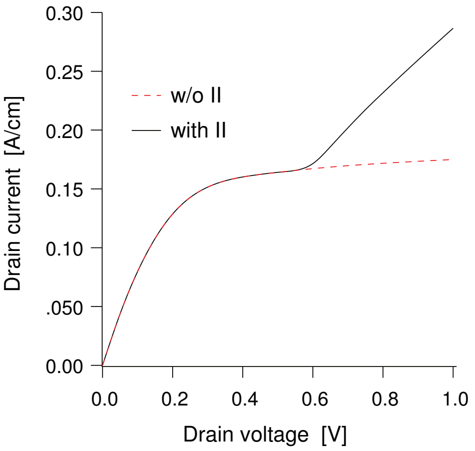

Fig. 6. Output characteristics of the SOI MOSFET for V_{GS} = 1.0 V with and without impact ionization (reprint from [7]).

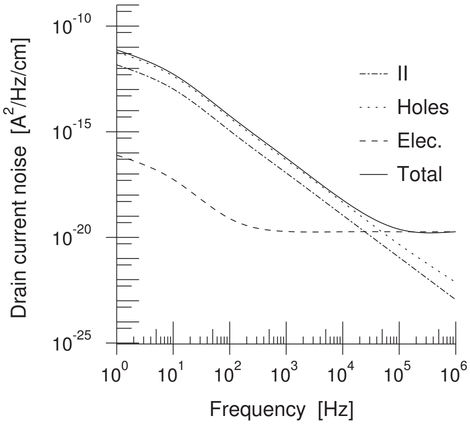

Fig. 7. Spectral intensity of the drain current fluctuations for the SOI device and V_{GS} = 1.0 V and V_{DS} = 1.0 V (reprint from [7]).

even if nonlocal II models are used. The situation is better for HD models concerning the modeling of II, but HD models suffer from a non physical diffusion of the channel carriers into the quasi-neutral substrate. This effect was discovered long ago [12] and is of no concern for MOSFETs with a bulk contact. For SOI, however, this effect leads to a non-physical charging of the isolated substrate with minority carriers and can even lead to a non-physical negative differential resistance of the output characteristic. Moreover, this device cannot be simulated based on the BTE and MC algorithms, since due to the charging of the substrate not only very small currents but as well effects with extremely different time constants have to be resolved, which is another situation, where MC algorithms lead to extremely large CPU-times. Finally, even to calculate a DC solution without a good initial solution is a real challenge where many DD and HD solvers fail. Figures 6 and 7 demonstrate, however, that the numerical solution algorithms of the BTE-PE solver described in [7] are already so mature that not only a DC solution for a SOI device with $2{\cdot}10^{17}$ cm^{-3} substrate doping and 500 nm channel length can be calculated but the kink effect with high accuracy and the low frequency drain current noise as well.

References

[1] C. Jungemann, T. Grasser, B. Neinhüs, and B. Meinerzhagen, “Failure of moments-based transport models in nanoscale devices near equilibrium,” *IEEE Trans. Electron Devices*, vol. 52, no. 11, pp. 2404–2408, 2005.

[2] C. Jungemann and B. Meinerzhagen, *Hierarchical Device Simulation: The Monte-Carlo Perspective*. Computational Microelectronics, Wien, New York: Springer, 2003.

[3] G. A. Baraff, “Maximum anisotropy approximation for calculating electron distributions; Application to high field transport in semiconductors,” *Phys. Rev.*, vol. 133, no. 1A, pp. A26–A33, 1964.

[4] N. Goldsman, L. Henrickson, and J. Frey, “A physics-based analytical/numerical solution to the Boltzmann transport equation for use in device simulation,” *Solid–State Electron.*, vol. 34, pp. 389–396, 1991.

[5] A. Gnudi, D. Ventura, G. Baccarani, and F. Odeh, “Two-Dimensional MOSFET Simulation by Means of a Multidimensional Spherical Harmonics Expansion of the Boltzmann Transport Equation,” *Solid–State Electron.*, vol. 36, pp. 575–581, 1993.

[6] K. Rahmat, J. White, and D. A. Antoniadis, “Simulation of semiconductor devices using a Galerkin/spherical harmonic expansion approach to solving the coulped Poisson-Boltzmann system,” *IEEE Trans. Computer–Aided Des.*, vol. 15, pp. 1181–1196, 1996.

[7] S. M. Hong and C. Jungemann, “A fully coupled scheme for a Boltzmann-Poisson equation solver based on a spherical harmonics expansion,” *J. Compu. Electr.*, vol. 8, pp. 225–241, 2009.

[8] M. Bollhöfer and Y. Saad, “ILUPACK — preconditioning software package.” release 1.1 available online at `www.math.tu-berlin.de/ilupack/`, December 2004.

[9] A. T. Pham, C. Jungemann, and B. Meinerzhagen, “On the numerical aspects of deterministic multisubband device simulations for strained double gate PMOSFETs,” *J. Compu. Electr.*, vol. 8, pp. 242–266, 2009.

[10] C. Ringhofer, “A mixed spectral - difference method for the steady state Boltzmann - Poisson system,” *SIAM J. Num. Ana.*, vol. 41, pp. 64–89, 2003.

[11] A. T. Pham, C. Jungemann, and B. Meinerzhagen, “A convergence enhancement method for deterministic multisubband device simulations of double gate PMOSFET,” in *Proc. SISPAD*, pp. 115–118, 2009.

[12] I. Bork, C. Jungemann, B. Meinerzhagen, and W. L. Engl, “Influence of heat flux on the accuracy of hydrodynamic models for ultrashort Si MOSFETs,” in *NUPAD Tech. Dig.*, vol. 5, (Honolulu), 1994.