Hybrid CA/Monte Carlo Modeling of Charge Transport in Semiconductors

M. Saraniti*, S. M. Goodnick** and S. J. Wigger**

*Department of Electrical and Computer Engineering,
Chicago, IL 60616-3793, USA, saraniti@iit.edu
**Department of Electrical Engineering,
Arizona State University, Tempe, AZ 85287-5706, USA, Shela.Wigger@asu.edu

ABSTRACT

We report on the modeling of ultra-small MOS devices using a newly developed full band device simulator. The simulation tool is based on a novel approach, featuring a hybrid Monte-Carlo/Cellular Automata simulation engine self-consistently coupled with a 2D and 3D multi-grid Poisson solver.

Keywords: Device simulation, Cellular Automata, Monte Carlo, Full-Band, 3D Poisson Solver, Multi-grid Method.

1 INTRODUCTION

The Ensemble Monte Carlo (EMC) method [1] has been employed now for over thirty years to simulate semi-classical carrier transport in semiconductor materials and devices. The limitations of the approach have been its large computational burden, particularly when the full band structure is incorporated into the physical description of the system [2][3].

In order to reduce the computational demands of particle-based simulation, the cellular automaton (CA) approach was developed in the context of semiconductor device simulation [4]. Within the CA framework, both **k**-space and real space are represented on a discrete numerical grid, which simplifies the description of scattering and the particle motion in real and momentum space. This technique was successfully demonstrated using an analytical, non-parabolic band model [4], where significant speed-up was observed compared to more traditional EMC methods.

This early work on CA methods utilized simplified, non-parabolic band models to represent the energy dispersion and scattering rates, whereas state of the art particle simulation techniques have increasingly moved towards full-band models [2][3]. For this reason, we have developed a full-band CA based simulator.

This simulation tool is based on the representation of the first Brillouin zone (BZ) of the crystal onto a non-uniform mesh, over which the transition table for the scattering probability for every initial state to every final state is generated and stored. This leads to considerable simplification in the final state selection after scattering, which typically is a time-consuming process in full-band EMC simulation.

Because the full **k**-space is modeled, fully anisotropic scattering rates may be incorporated without loss of performance, although at the cost of large memory usage to store the transition table. Of course, a trade-off exists in the full-band CA approach between energy resolution in the momentum space and the dimension of the pre-computed transition table. In fact, contrary to the EMC approach, the nature of the CA algorithm makes any correction of the energy of a carrier after scattering impossible. To ensure acceptable energy conservation one is then forced to reduce the grid spacing, consequently increasing the number of cells of BZ, and the resulting dimension of the transition table. Good results have been obtained for electrons in Si [5] with tables about 1GB large, but the approach needs to be modified to ensure energy conservation, particularly when bipolar devices are simulated.

2 HYBRID CA/MC METHOD

A new approach has been recently proposed [6],[7] as an alternative to the memory demanding CA method and to the CPU intensive EMC. The new "hybrid" method is based on a full-band representation of the momentum space, and models the carrier dynamics by using a fast but memory demanding CA algorithm in those regions of the space where most of the interactions occurs, while the slow but memory-efficient MC method is used elsewhere. Small EMC scattering tables can in fact be used whenever the carrier population is reduced, such as in the high energy regions of BZ, and in those regions where the scattering probability is small, for example close to the minima of the first conduction band and the maxima of the valence band, as shown in Fig.1.

The low performance of the EMC algorithm will then not significantly affect the overall performance of the simulator because of the small number of scattering events to be simulated in these regions. At the same time, a considerable amount of memory is saved, allowing a finer spacing in the regions where the scattering probability and/or the carrier population is higher. The resulting simulation code is highly scalable in terms of RAM requirement and demonstrates considerable speedup as more memory is allocated for the algorithm.

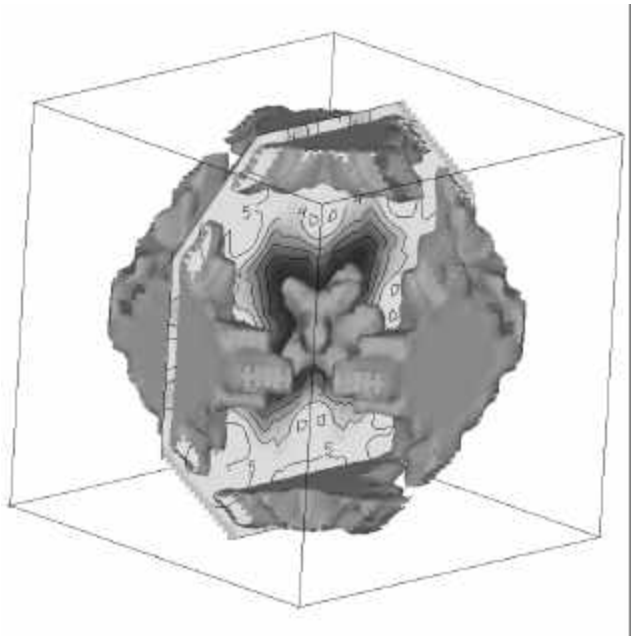

Figure 1. Representation of regions of BZ where the EMC algorithm is used, for the first valence band of Si. The EMC scattering selection process is forced in the central low energy, warped region and within the six gray regions close to the band edge, where the energy is higher than 2.5eV. The contour plot slice shows the value of the scattering rate.

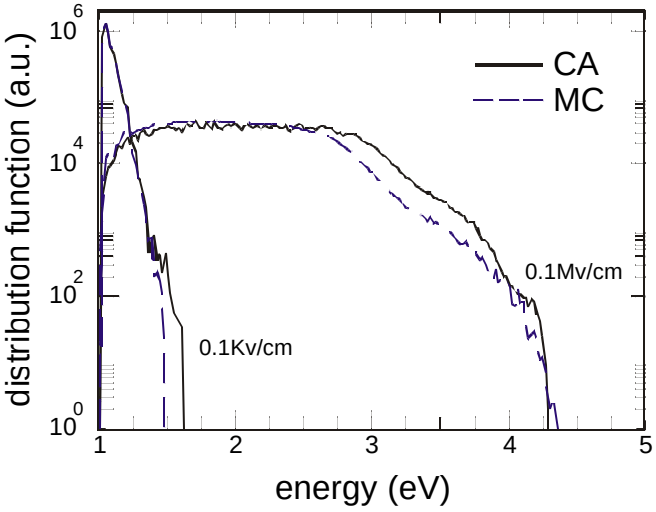

Figure 2 Energy distribution function as computed by the Monte Carlo method (dashed line) and using the Hybrid approach (full line).

The comparison of the electron distribution function as computed with the hybrid approach and with a standard full-band EMC is shown in Fig. 2 for two values of the field strength.

In spite of the sensitivity of the energy distribution on the transport model, the agreement of the two approaches is really good. A difference is noticeable in the high-energy tail of each curve. The higher values given by the hybrid approach are due to its smaller energy resolution in those regions of BZ that are far from the minima of the conduction band. This smaller resolution produces a spurious diffusion (i.e. a numerical heating) of electrons toward higher energies regions. The effect is not present in the section above 4eV of the distribution at 0.1MV/cm, because the numerical heating is compensated by the cooling effect of impact ionization. The small discrepancy between the EMC and the hybrid result can be easily reduced by grid refinement.

3 DEVICE SIMULATION RESULTS

The CA/MC algorithm is coupled self-consistently with a 2D and 3D multi-grid Poisson solver, to account for the spatial distribution of the electric field and charge concentration. The repeated solution of the Poisson equation requires a fast and efficient solver, and solvers based on the multi-grid method have already been shown to be one of the fastest available [9]. In order to demonstrate the applicability of the hybrid approach for modeling 3D devices, an n-channel MOSFET structure with a 50nm gate length is simulated. The schematic layout of this ultra-small structure is shown in Fig. 3, and corresponds to a process currently under development at Arizona State University.

Both 2D and 3D representations of the layout have been simulated, in order to investigate effects caused by the increase in dimensionality. These effects are still under investigation, and full I-V characteristics have been only obtained for the 2D case, at present.

The 2D layout is represented over an 85x75, rectangular non-uniform mesh. The source and drain n^+-regions have a doping concentration of $1x10^{20}$ cm^{-3}, while the p- region is doped with an acceptor concentration of $1x10^{19}$. The electron population is represented by $4x10^4$ super-particles[10], while $8x10^4$ super-particles are used to simulate holes in the device. The I-V results of the 2D simulation are presented in Fig. 4, and show compatible values with the high doping concentration and the short gate length.

The computational burden of simulating the 3D structure is much greater due to the larger number of real space grid points, and consequently, to the larger number of simulated electrons and holes. The simulation time is also much longer because of both the extra time needed to solve Poisson's equation on a larger grid, and the time required

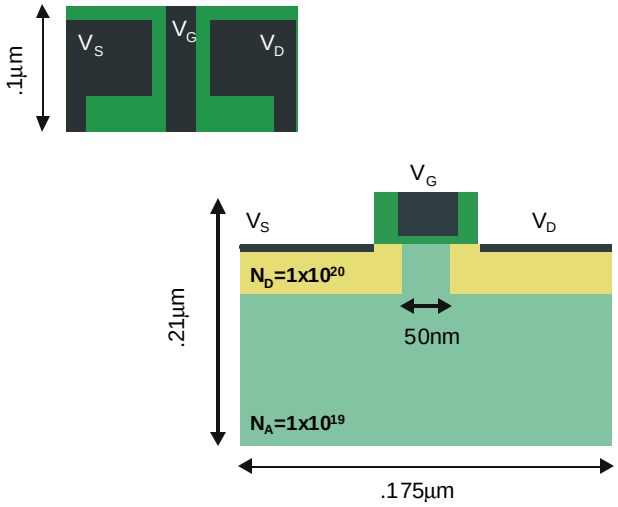

Figure 3. Schematic layout of 50nm n-channel MOSFET, used for both 2D and 3D simulations. The oxide thickness is 3nm.

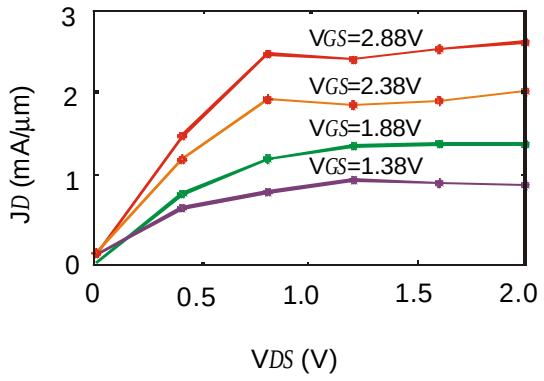

Figure 4. I-V characteristics for MOSFET structure using the 2D Poisson solver.

for the simulation of the dynamics of a larger number of carriers, both in real- and **k**-space.

An irregularly spaced 85x80x10 grid is used in the 3D simulation, with a maximum value of 1.5 for the mesh expansion ratio. The relatively low expansion ratio allows the preservation over the geometry of an almost constant order of the approximated Laplace operator in Poisson's equation [11], and maximizes the self-consistency of the algorithm. The time interval between subsequent solutions of Poisson's equation can then be fixed at $1x10^{-15}$ seconds, a value close to the inverse of the plasma frequency [10], which is a limit imposed by the physics of the system rather than by numerical instability. Since the convergence rate of multi-grid Poisson solvers scales linearly with the number of grid points [9], the CPU time required by the Poisson solver is increased approximately by a factor of 10. The number of simulated electrons and holes in the 3D layout is $5x10^5$ and $1.5x10^6$, respectively; larger numbers would lead to unrealistic simulation times. The simulated population results in a number of particles per cell smaller than the in the 2D case, and in a consequently rougher profile of the charge concentration, which slows the Poisson solver down. It should be noted that the usual trade-off between the number of simulated carriers (and consequent time spent to simulate their dynamics) and the smoothness of the charge concentration (and time consequently saved by a faster convergence of the Poisson solver) is exacerbated by the three-dimensionality of the problem.

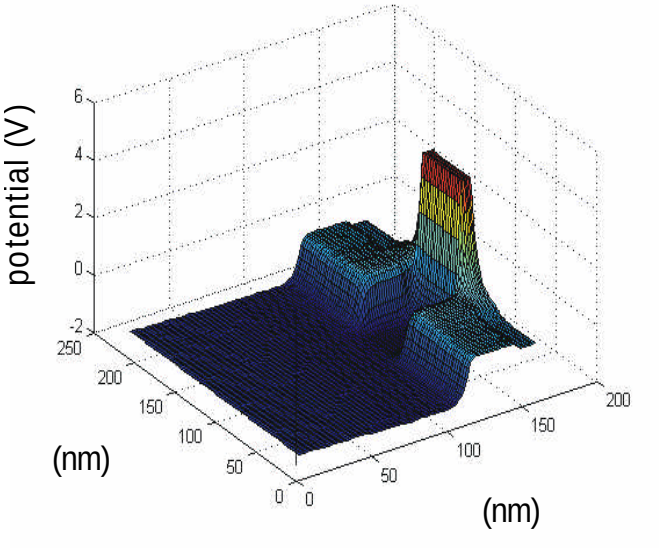

Figure 5. Cross-section of potential profile for MOSFET structure using the 3D Poisson solver. The potentials on the source, drain and gate are V_s=0V, V_d=0V and V_g=2.88V, respectively.

4 FUTURE WORK

As a natural evolution of the hybrid approach, a parallel version of the algorithm is being implemented in such a way that the full computational load (both CPU and RAM) will be shared between concurrent processes.

Within the proposed "hybrid" parallel framework, the carrier ensemble is initially distributed between processes, and a complete time-step is simulated. Fine load balancing can be obtained by periodically redistributing carriers from the slower processors to the less loaded ones. The particle-based nature of the algorithm easily allows this ultra-fine balancing, because of the small computational load due to each single carrier.

The efficiency of this approach will be further improved by performing a decomposition of the computational domain, resulting in a scattering table shared within the processor pool. Given the highly non-local nature of the interactions in the discrete momentum space, cells will be sorted as a function of their energy, and the decomposition will occur in the energy rather than in the momentum domain. This adjustment will allow the implementation of larger scattering tables, so reducing the regions in momentum space to be modeled by the slow EMC approach. The carrier migration between processes is expected to generate a small computational overhead with respect to the speed-up due to the higher fraction of CA-modeled states.

The hardware currently being used to implement the parallel version of the hybrid approach is a cluster of 4 networked dual-processor workstations, equipped with 1GB of RAM per processor. The inter-process communication software is based on the standard Message Passing Interface (MPI)[8].

REFERENCES

[1] C. Jacoboni, and L. Reggiani," The Monte Carlo method for solution of charge transport in semiconductors with applications to covalent materials", Rev. Mod. Phys. **55** (3), 645 (1983).

[2] 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, **38** (19), 9721 (1988)

[3] Karl Hess, *Monte Carlo Device Simulation: Full Band and Beyond*, Kluwer Academic Publishers Group, Dordrecht, The Netherlands (1991).

[4] K.Kometer, G. Zandler, and P.Vogl, " Lattice-gas cellular-automaton method for semiclassical transport in semiconductors", Phys. Rev. B **46** (3), 1382 (1992).

[5], M. Saraniti, S. J. Wigger and S.M. Goodnick, "Full-Band cellular automata for modeling transport in sub-micrometer devices", in Proc. of MSM99, S.Juan (PR),415, (1999).

[6] M. Saraniti, and S.M. Goodnick, "Hybrid Full-Band Cellular Automaton/Monte Carlo Approach for Fast Simulation of Semiconductor Devices, submitted for publication in IEEE Trans. on El. Dev.

[7] S.J. Wigger, M. Saraniti, and S.M. Goodnick, "Full Band CA/Monte Carlo Modeling of Ultrasmall {FET}'s", in publication in Superlattices and Microstructures, (2000).

[8] M.Snir, S.Otto, S. Hauss-Lederman, D.Walker, and J.Dongarra, *MPI-The Complete Reference*, 2nd ed., The MIT Press, Cambridge, MA, (1998).

[9] W. Hackbusch, *Multi-grid Methods and Applications.* Berlin: Springer-Verlag, 1985.

[10]R.W.Hockney and J.W.Eastwood, *Computer Simulation Using Particles*, Adam Hilger, Bristol (1988)

[11] E.Kàlnay de Rivas, "On the Use of Nonuniform Grids in Finite-Difference Equations",J. of Comp.Phys. **10**,202 (1972).