New Developments in Monte Carlo Device Simulation

Umberto Ravaioli

Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

Abstract
Monte Carlo particle methods have a unique role in semiconductor device simulation, since they allow one to solve the Boltzmann equation statistically, with the inclusion of many physical details which cannot yet be included completely in other approaches. Of the main limiting drawbacks of the technique, memory requirements and computational costs are much alleviated by the increasing power of computers. The problems due to statistical noise can in some applications be corrected by *ad hoc* techniques. This paper briefly reviews the state-of-the-art of Monte Carlo device simulation and elaborates on the future applications of the method.

I. Introduction

Monte Carlo particle simulation has been a powerful tool for the investigation of transport in semiconductors for well over twenty years. The evolution of Monte Carlo methods has been directly influenced by advances in computers, which have made possible the implementation of more and more refined physical models [1,2]. Originally, applications were limited to tracking the evolution of a single particle, to obtain steady-state time-averages of transport parameters. The single particle Monte Carlo technique is adequate to study bulk semiconductor properties under uniform field conditions, yielding as typical results the distribution function, average velocity and energy, valley occupation percentage, and velocity-field characteristics. The single particle model is also adequate to obtain local information in device structures for which a potential distribution is approximately known.

Transient and selfconsistent simulations were implemented when the memory of computer was increased to allow the simultaneous tracking of thousand of particles. The so-called ensemble Monte Carlo technique has then made possible a whole new range of selfconsistent applications which have required the inclusion of methods for the local evaluation of electronic forces, e.g. Poisson's equation [3]. The early simulation models have treated the bandstructure with an analytical parabolic or non-parabolic approximation. The energy range for hot electron analysis has been considerably extended with the introduction of algorithms which implement numerically a complete bandstructure of the semiconductor material [2,4]. Initially limited to bulk material and single particle applications, the full bandstructure Monte Carlo has been extended in recent years to ensemble selfconsistent applications and can be now run fairly efficiently on workstations [4,5]. While hardware improvements are making Monte Carlo applications more realistic, many efforts have been devoted to overcome the natural limitations of the technique, to optimize the algorithms and to take advantage of new hardware capabilities to introduce more advanced physical models.

The flow-chart of a selfconsistent Monte Carlo device simulation is relatively simple. The method uses a time-dependent approach, which besides providing transient results

may also be run until a steady-state is achieved. The iteration oscillates between a block which utilizes the information on charge density to evaluate the electrical forces in space, and a block which tracks the particle movement within a given timestep. The frequency of forces update is chosen as a trade-off between accuracy of the physics which requires very frequent force recalculation, and overall efficiency.

The particle movement is divided into two distinct parts: free flight under the influence of the electrical forces, and scattering events that terminate the flights. The length of the free flight trajectory is determined statistically, by relating the total scattering probability rate to a pseudo-random number picked from a uniform sequence generated by the computer. Once the flight is terminated, random number techniques are again used to select the type of scattering, according to the relative rate strength of the various mechanisms at that particular energy, and to determine the final state after the scattering event.

The particles are treated as classical objects obeying Newtonian mechanics during the free flights, and the scattering events are assumed to be instantaneous. Just a few lines of code are necessary to evaluate the momentum evolution during a timestep, using the classical law of accelerated motion. In an analytical band formulation, the energy at the end of the timestep is directly computed by evaluation of a simple formula. In the full bandstructure formulation, since the energy values are available only on a 3-D grid in the Brillouin zone of momentum space, the energy at the end of the timestep has to be evaluated by interpolation. This process is one of the major bottlenecks in the simulation, and innovative gridding approaches have been recently applied to implement faster energy evaluation techniques [6].

The weights of the different parts of the code, in terms of the overall CPU time, vary according to the implementation and running conditions. In most applications, the force evaluation by solving the Poisson's equation represents only a few percents of the computation, when it is solved at typical time intervals of 10 fs. However, in some cases (e.g. high doping) the time between two Poisson solutions must be reduced to 1 fs or less, to avoid spurious plasma oscillations of the particle gas, and in 2-D or 3-D simulations with a large number of grid points, such numerical solutions may carry a considerable weight. When forces are evaluated bu a full molecular dynamics approach, the computational cost for this may be dominant, although there are multipole techniques which can be utilized to dramatically reduce the CPU time without sacrificing precision [7,8].

II. Statistical noise

Statistical noise due to the randomness of the events and discreteness of the particles always affects Monte Carlo simulation results. In many cases, it is sufficient to increase the number of particles and average in time the ensemble averages, to improve the accuracy of the collected statistics for various observables. However, there are cases where a mere increase of the number of samples is not practical. This is typically true in the case of effects which depend on high energy tails of the distribution function. An example is the injection of carriers in the oxide of a MOSFET structure. Steady-state parameters like potential and carrier distribution in the device are not very much affected by these rare events, but the evaluation of gate currents is exclusively determined by them. It is necessary to assign different weights to particles in different energy ranges in order to emphasize the statistics of high energy tails, while preserving the overall physical charge for force evaluation.

If the transient behavior is of interest, it is not possible to perform time averages and only an increase of the size of the ensemble would improve the statistics, but again,

this is seldom practical. Small-signal parameters of microwave devices can be determined for instance by Fourier transform of the transient currents. Deterministic models like drift-diffusion, which is not affected by statistical noise, are frequently used. They are also applied to study large-signal transient response of digital circuits. The Monte Carlo technique can be applied when drift-diffusion fails, provided that integrated quantities, rather than instantaneous ones, are considered. This is done by extending an idea first introduced by Hockney and Eastwood [2], where one keeps track of the total charge or current which transits through contacts. When the transient simulation is carried through steady-state, it is possible to identify two contributions to the cumulative charge, associated to the transient and to the steady-state regime [9]. Since the noise fluctuations are not very large for the time-integrated charge, it is possible to precisely fit in time the transient contribution with a polynomial or a combination of exponentials. This procedure yields very smooth curves for the transient currents, which can now be Fourier transformed to yield the intrinsic small-signal parameters.

III. Optimization

The core of the Monte Carlo algorithm determines the times of flight t by solving the integral equation $-lnr = \int_0^t \lambda(t)dt'$, where r is a uniform random number between 0 and 1, and $\lambda(t)$ is the total scattering rate which changes in time as the particle momentum and energy vary under the influence of the fields. The choice of an appropriate solution is quite important for an efficient algorithm. An important concept is the *self-scattering*, a fictitious event which does not affect the electron state when selected. A self-scattering rate can be adjusted as convenient to facilitate the evaluation of the flight time, since it does not affect the statical properties of the process. The simplest approach is to add a self-scattering rate which makes the total rate constant, so that the integral above can be trivially determined.

A comparison of various techniques can be found in [10]. Self-scattering rates can be fixed as a function of energy, or can be dynamically adjusted in time along the particle trajectory. The goal of optimization of this process is to reduce as much as possible the amount of self-scattering while maintaining but still retaining it to guarantee that the solution of the integral is always statistically *correct*, rather than introducing approximations. The constant time technique [10,11] offers a good trade-off for self-consistent applications. The simulation time is divided into small constant steps for all the particles, and a self-scattering rate is adjusted to make the total rate constant in time within that interval. The integral is solved by adding up trivial contributions, until the equality is satisfied.

As mentioned earlier, the determination of the energy at the end of a timestep is a time consuming operation in full bandstructure calculations, which can create major bottlenecks. Optimization can be achieved by using a tetrahedral mesh in momentum space, instead of a regular grid, arranged so that nodes of a tetrahedron are positioned on adjacent energy isosurfaces [6]. With these grids, a linear expression can be used to determine directly the energy for given momentum coordinates, with precision controllable by adapting the grid locally. While the construction of this algorithm involves a considerable initial development cost, this technique promises to be a breakthrough which should make the full band approaches not more expensive than analytical band algorithms.

IV. Supercomputation

The advent of supercomputers has offered new opportunities to improve the performance of Monte Carlo codes and increase the size of the problems to be solved. The particle transport has an inherent parallel behavior which can be exploited, but the main obstacle is the fact that the histories of different particles can be very different due to the randomness of the processes, and such a difference can be quite emphasized in the case of sharply nonuniform structures.

Vectorization techniques can be very effective for bulk analysis, since it is sufficient to follow particles from one scattering event to another in parallel. The particle histories are equalized by sacrificing synchronism, which is not necessary for averaging. Synchronism must be maintained for self-consistent algorithms, because forces must be evaluated at specified intervals. Applications using an ensemble constant time approach for the flight time evaluation, yield a vectorization speed-up between 3-5 on a CRAY Y-MP supercomputer [11], where the maximum possible speed-up is about 10. Reports by several groups indicate similar speed-up for different Monte Carlo implementations. Comparisons in this area are extremely difficult, since when the efficiency of a code is improved, the achievable vectorization speed-up tends to decrease.

Parallelization can also be very advantageous for bulk calculations, since the particles are substantially independent. For more complicated self-consistent models, performance depends on the actual architecture of the hardware and on the strategy used to balance the load on the processors. Work in this area is still largely experimental, reflecting the immaturity of parallel computers and compilers. Parallelization is particularly appealing for 3-D simulation, because realistic applications require a very large number of particles [12]. Balancing of the load between processors is very important, because for massively parallel applications even a small percentage of non-parallelizable code may make the computation inefficient.

The future evolution of parallel architectures will have an important influence on Monte Carlo applications. This is particularly true for the most advanced applications which require the storage of large tables. The most memory intensive model is at the top of the hierarchy, where tables for the full bandstructure and momentum-dependent scattering rates must be stored and need to be equally accessible by all the processors. The information is only read by the processors during the simulation, since these tables are not changed. Therefore, for efficiency, such tables should reside in a shared memory region which can be quickly accessed by all nodes with uniform times, rather than being distributed in local memory areas appended to the processors or being completely copied in each of these areas, to limit storage requirements. The remaining compelling memory requirements are related to particle attributes, like position and momentum, which are continuously updated, and position dependent data (charge, fields) for large grids, particularly in 3-D. The main issue is to efficiently handle the communication between blocks of distributed memory. A logical storage scheme, in the case of distributed memory, is to map particles and grid nodes of domain subregions onto separate processors, trying to balance the number of particles per processor, adaptively throughout the simulation. As particles cross boundaries between subregions, they should be reassigned to new processors. For the determination of charge on grid nodes and the subsequent solution of Poisson equation, a small amount of communication is necessary between processors corresponding to physically contiguous regions. Much more challenging is the implementation of molecular dynamics, where in principle all processors need to communicate with each other. The optimal solution for a Monte Carlo code would be to have only shared memory available. In such a way it would be possible to load particles on fixed processors, regardless of position. Provided that each single processor can address a large shared memory necessary for the applications, still great challenges remain in designing system software and compilers for such a system.

V. Force Evaluation

The numerical solution of Poisson's equation only provides forces in a quasistatic approximation. In some applications this is not sufficient. It has been shown that to simulate fast phenomena associated with the transport of carriers generated by femtosecond laser pulses, the full Lorentz force must be evaluated [13], where the electric and magnetic fields should be obtained by solving the time dependent Maxwell's curl equations (note that we refer here the fields are generated by the fast moving charge particles, not to the laser radiation which is absorbed by the sample) Implementation of the algorithm can be very suitable to parallel computation and actually cheaper than solving the Poisson's equation for multi-dimensional simulations [12].

It is not clear at which frequency range the inclusion of the magnetic field begins to be necessary. In the simulation of general microwave devices, it should be possible to simply substitute Poisson's equation with the time-dependent wave equation for the retarded potential, which has the same space-dependent terms, to account for the displacement currents (in 1-D the displacement current contribution can be integrated and applied as additional boundary condition to Poisson's equation). However, in the THz regime the wavelength of the electromagnetic field in the doped semiconductor layers can be comparable or smaller than the dimensions of the active regions,

In other cases, the solution of Poisson's equation on discrete points may not be accurate enough to resolve the coulomb interaction between charges. Reduction of the mesh size may not improve much the situation, because the number of simulated carriers is fixed. A molecular dynamics can be used to calculate the force acting on a particle by adding the coulomb potential due to all the other charges. This approach would automatically include the electron-electron interaction effects, which can be very important in the case of high concentrations. The major computational obstacle is in the fact that the full molecular dynamics evaluation of the forces involves a number of operations of order N^2, where N is the number of particles. In order to develop practical algorithms, it should be possible to apply multipole techniques, which have been extensively used to calculate molecule configurations, ionic systems [7] and capacitances in complicated VLSI interconnect layouts [8], to name a few applications. With accurate calibration, multipole algorithms only require a number of operations of order N. The idea is to consider particle-particle forces only within an appropriate neighborhood, and treat interactions from longer range particles through interpolated forces on a mesh with increasing coarseness at farther distances.

In many applications it is common to simulate only particles in a relatively small cell of a periodic structure, or a sample of a larger device. When the Molecular Dynamics method is implemented, the interactions with charges in other regions, which are not simulated, cannot be neglected. If one assumes that the simulated geometric sample is one element of a periodic structure, every particle in the simulated cell corresponds to a "replica" in each of the other cells of the periodic domain. The replicas of a given particle constitute then a "lattice" of charge, and special techniques must be applied to get a convergent sum of the interactions keeping down the computer time requirements [14]. For simulation of a bulk material, one has to consider 3-D lattices of replicas, 2-D and 1-D lattices for 1-D and 2-D device simulations, respectively, and in the case of a complete 3-D device simulation no replica has to be taken into account. Although more accurate than Poisson's equation, a molecular dynamics algorithm is still providing forces in an electrostatic approximation.

VI. Hybrid techniques

Monte Carlo techniques have found many useful applications beyond full self-consistent simulation of devices. Assuming that the *correct* solution of the Boltzmann equation is obtained, Monte Carlo simulations are often used to parameterize other models. Field dependent mobility and diffusion coefficient are used in drift-diffusion applications, for instance, and a number of other parameters are extracted to calibrate hydrodynamic and energy transport models. Although the Monte Carlo results should not always be trusted as completely exact, the potential problems in these schemes are more due to the fact that parameters, obtained for a bulk with uniform field, are often employed for nonuniform field condition, which can be considerably different.

The Monte Carlo approach is also used as a *postprocessor* using the potential profile obtained by a drift-diffusion or hydrodynamic approach. This is certainly a valid approach for large devices, where not only a drift-diffusion solution is acceptable, but also a full self-consistent Monte Carlo solution would be impractical. By tracking particles simulated with Monte Carlo in the fixed potential, it is possible to evaluate high energy effects (injection into oxide, impact ionization) which are not well account for in simpler models. The advantage of a postprocessor is in the fact that one can use very efficient vector and parallel algorithms, since the tracked particles are essentially independent in nonself-consistent simulations.

A recent application of Monte Carlo simulation involves the calculation of tables for a scattering matrix technique [15]. The Monte Carlo procedures provides the possible outgoing momentum values, with associated probabilities, for a given incoming momentum into a thin slab of the device. By partitioning the device into slabs and matching the solutions at the interfaces with the momentum scattering tables, it is possible to get a solution which provides all the information of a self-consistent Monte Carlo, but with smooth solutions without the noise. Because of this, rare events should be easier to observe directly. Multi-dimensional applications are also possible with this technique.

VII. Improved Physical Models

A shortcoming of Monte Carlo models is due to the unavailability of many constants which are necessary to determine the scattering rates. A typical example is represented by deformation potentials of phonon scatterings. The usual procedure is to choose a set of deformation potentials which fit experimental data for steady-state velocity-field characteristic curves. Unfortunately, many slightly different sets can be found which provide a reasonable fit. Measurements of some important parameters, like the intervalley deformation potential for transitions between Γ and L valleys in GaAs, have been attempted, but the data reported by various groups are too contradictory to resolve the uncertainty.

The solution is to formulate new models which are based on first principles and rely less on parameter fitting. Electron-phonon scattering is usually treated with a simplistic dispersion relation, and is defined on arbitrary partitions of the Brillouin zone centered around energy minima (valleys). At high fields, the validity of this picture is questionable. A complete phonon model with accurate dispersion relations is needed, but the computational complexity is formidable. A consistent approach should treat both the bandstructure and the electron-phonon interaction on the same footage.

Electron-phonon matrix elements of the true Hamiltonian are equivalent to the matrix elements of a pseudo-Hamiltonian, as long as certain conditions on the true potential are met. Several calculations have been performed for semiconductors using local and nonlocal empirical pseudopotentials. These are assumed to be the sum of spherical potentials which

move rigidly with the atoms [16,17]. Another approach uses the *ab initio* pseudopotentials, which avoid self-consistency problems of the empirical approach by approximating the change in charge density when the atoms are displaced [18]. Computations using the Harris functional approach have provided the deformation potentials for phonon scattering in Si, throughout the Brillouin zone. A significant result is the evidence of the variation of deformation potentials with initial and final state wavevector. The total deformation potential exhibits a high degree of dispersion, especially for transitions away from the valley minima [18]. These results have significant implications for improving the predictive power of high-field Monte Carlo simulations.

VIII. Conclusions

Monte Carlo techniques for device simulation are undergoing dramatic developments due to the recent evolution of available computational platforms. Full bandstructure applications are already practical for use on top of the line workstations. New optimization techniques for the determination of momentum space trajectories, new approaches for electron-phonon interaction which remove much of the uncertainties of current models, and emerging applications on parallel architectures, should contribute to provide, in the next few years, accurate and efficient simulators with sufficient predictive capability in the high field transport regime to meet the needs of CAD designers of new generations of integrated devices. Monte Carlo techniques have already found a very important role as tools for the calibration of simpler models, and as postprocessing complements of conventional simulators, to quickly assess the importance of hot electron phenomena affecting device reliability.

Acknowledgments - This work has been supported by the Joint Services Program (grant N00014-90-J1270).

References

[1] C. Jacoboni and P.Lugli, *The Monte Carlo Method for Semiconductor Device Simulation*, Springer-Verlag, Vienna, 1989.

[2] K. Hess, ed., *Monte Carlo Device Simulation: Full Band and Beyond*, Kluwer Academic Publishers, 1991.

[3] R.W.Hockney and J.W.Eastwood, *Computer Simulation using Particles*, McGraw Hill, New York, 1981.

[4] M.V. Fischetti and S.E. Laux, "Monte Carlo Simulation of Transport in Technologically Significant Semiconductors of the Diamond and Zinc-Blende Structures-Part II: Submicrometer MOSFET's," IEEE Trans. Elect. Devices, vol. 38, pp. 650-660, 1991.

[5] C.H. Lee and U. Ravaioli, "Monte Carlo Simulation of Si Devices", Proc. VPAD '93, pp. 36-39, Nara, Japan, May 14-15, 1993.

[6] M.R. Pinto, J. Bude and C.S. Rafferty, "Simulation of ULSI Silicon MOSFETs", Proc. VPAD '93, pp. 22-25, Nara, Japan, May 14-15, 1993.

[7] Z.A. Rycerz, "Calculation of the Coulomb Interactions in Condensed Matter Simulation", Molecular Simulation, vol. 9, pp. 327-349, 1992.

[8] K. Nabors, M. Kamon and J. White, "Multipole accelerated 3-D Interconnect Analysis", Proc. VPAD '93, pp. 72-75, Nara, Japan, May 14-15, 1993.

[9] M.B. Patil and U. Ravaioli, "Transient Simulation of Semiconductor Devices Using the Monte-Carlo Method", Solid-State Electronics, vol. 34, pp. 1029-1034, 1991.

[10] R.M. Yorston, "Free-Flight Time Generation in the Monte Carlo Simulation of Carrier Transport in Semiconductors," J. of Comp. Phys., vol. 64, pp. 177-194, 1986.

[11] U. Ravaioli, "Vectorization of Monte Carlo Algorithms for Semiconductor Simulation", Ch. 9 in *Monte Carlo Device Simulation: Full Band and Beyond*, K. Hess, ed., Kluwer Academic Publishers, 1991.

[12] S. Pennathur, U.A. Ranawake, V.K. Tripathi, P. Lenders and S.M. Goodnick, "PMC-3D: A Parallelized 3D Monte Carlo Simulator for Electronic and Electro-optic Devices", these Proceedings.

[13] K.M. Connolly, R.O. Grondin, R.P. Joshi, and S.M. El-Ghazaly, "Numerical Modeling of Ultrafast Electrical Waveform Generation and Characterization", Proc. of SPIE Symposium on Ultrafast Laser Probe Phenomena in Bulk and Microstructure Semiconductors III, vol. 1282, pp. 172-181, 1990.

[14] D.K. Ferry, A.M. Kriman and M.J. Kann, "Molecular Dynamics Extensions of Monte Carlo Simulation in Semicondcutor Device Modeling", Computer Physics Communications, vol. 67, pp. 119-134, 1991

[15] M.A. Alam, M.A. Stettler and M.S. Lundstrom, "Formulation of the Boltzmann Equation in Terms of Scattering Matrices", Solid-State Electronics, vol. 36, pp. 263-271, 1992.

[16] M.V. Fischetti and J. Higman, "Theory and Calculation of the Deformation Potential Electron-Phonon Scattering Rates in Semiconductors, Ch. 5 in *Monte Carlo Device Simulation: Full Band and Beyond*, K. Hess, ed., Kluwer Academic Publishers, 1991.

[17] T. Kunikiyo, Y. Kamakura, M. Yamaji, H. Mizuno, M. Takenaka, K. Taniguchi and C. Hamaguchi, "Adjustable Parameter Free Monte Carlo Simulation for Electron Transport in Silicon Including Full Band Structure", Proc. VPAD '93, pp. 42-43, Nara, Japan, May 14-15, 1993.

[18] P.D. Yoder, V.D. Natoli and R.M. Martin, "*Ab Initio* Analysis of the Electron-Phonon Interaction in Silicon", J. of Appl. Physics, vol. 73, pp. 4378-4383, 1993.