Pergamon

PII: S0038-1101(96)00198-0

Solid-State Electronics Vol. 41, No. 4, pp. 599–605, 1997

0038-1101/97 $17.00 + 0.00

ANALYSIS OF VARIANCE-REDUCTION SCHEMES FOR ENSEMBLE MONTE CARLO SIMULATION OF SEMICONDUCTOR DEVICES

A. PACELLI[1] and U. RAVAIOLI[2]

[1]Dipartimento di Elettronica e Informazione, Politecnico di Milano and CEQSE-CNR, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy

[2]Beckman Institute, University of Illinois at Urbana-Champaign, 405 N. Mathews Avenue, Urbana, IL 61801, U.S.A.

(Received 8 March 1996; in revised form 18 June 1996)

Abstract—A detailed discussion of variance-reduction techniques for Monte Carlo device simulation is presented. The use of variable statistical particle weights is examined from the point of view of numerical accuracy and physical consistency. An improved splitting/gathering technique is presented, which operates on the entire ensemble of particles rather than on particles located in one region of phase space. Application of the method to self-consistent ensemble simulation is also discussed, showing that the use of variable-weight techniques is limited to cases where the effect of short-range Coulomb interaction is negligible. Examples are given for simple space-homogeneous and one-dimensional cases, and for a submicron *n*-MOSFET device. © 1997 Elsevier Science Ltd. All rights reserved

1. INTRODUCTION

The use of the Monte Carlo (MC) technique for simulation of semiconductor devices allows a high degree of flexibility and accuracy in the modeling of hot-carrier and non-local effects. The main drawback of the MC method is the high cost in terms of computational resources. This is mainly due to the large statistical fluctuations, i.e. the high variance of some estimated quantities. An example is the calculation of the gate current of a MOSFET, where the physical mechanism of charge injection into the gate oxide is the emission of hot carriers across the Si/SiO_2 barrier. Since the injection of one carrier is a very rare event, a long simulation time is necessary to obtain a good estimate of the gate current.

Numerical techniques have been introduced to improve the computational efficiency by reducing the variance of the estimators, while conserving the same expected values. Three main approaches have been proposed in the literature. The simplest methods are based on the concept of repetition of individual particle trajectories, and are inherently suitable for one-particle simulations[1–5]. In the second approach the ensemble MC simulation is generalized assigning arbitrary probabilities to particle trajectories[6]. The resulting method, known as Weighted Ensemble Monte Carlo, requires some modifications to the simulation algorithm. In addition the theoretical treatment assumes the electric field to be assigned[6,7], and is not easily extended to self-consistent simulations, where Coulomb interaction between carriers must be accounted for. The third class of variance-reduction methods includes the variable-weight techniques, where a statistical weight factor is assigned to each particle, while the MC simulation remains unchanged[8–10]. The goal of such methods is the manipulation of the weights, in order to improve the sampling of phase space. For example, a large number of particles with small weight can be used in regions of phase space where accurate results are desired.

One of the main advantages of the variable-weight technique is the ease of implementation and integration with existing simulation codes. Although this method has been widely used in the past, some of its basic features have not been clearly addressed in the existing literature. The purpose of this work is to give a unified description of the technique as applied to ensemble MC simulation, discussing its limitations and presenting improvements. The organization of the paper is as follows. In Section 2 we define and classify the variable-weight methods, introducing a novel scheme that processes the entire set of particles, instead of working locally in phase space. In Section 3, the issues arising in the application to self-consistent simulations are briefly discussed. Section 4 gives some simulation results. The limits of validity of the method are illustrated, using the one-dimensional simulation of a high-density electron gas as an example. Finally we present results for a 0.5-μm *n*-MOSFET device, discussing specific techniques for dealing with the calculation of the gate-injection current.

2. VARIABLE-WEIGHT SCHEMES

In a variable-weight MC simulation, a specific weight is assigned to each particle. The weight is proportional to the number of physical carriers (electrons or holes) that a single computational particle represents. Estimators are then computed as weighted averages, where each particle gives a contribution proportional to its weight. Most variance-reduction schemes are based on the concept of *splitting* a particle with weight W into N particles with weights w_i, conserving the total weight:

$$W = \sum_{i=1}^{N} w_i.$$

Clearly, each splitting increases the computation and storage requirements for the simulation, and some mechanism is needed to limit the total number of particles. However, the removal of particles from the simulation can potentially alter the mean and/or increase the variance of the results. For example, one might think of replacing a set of N particles with a single particle, carrying the total weight and the average wave vector of the original set. Although this method conserves the total momentum of the ensemble of particles, it is easy to see that the total energy is not conserved, and thus unphysical effects may occur. In the following, we will only consider stochastic methods, where the weights are changed according to well-defined probability distributions, while the physical state variables remain unchanged.

A basic distinction can be made between single-particle and many-particle schemes for particle removal. In a single-particle scheme, the weights are processed independently from each other. For example, a particle can be removed with probability $1 - 1/N$, but if it is retained, its weight is increased by a factor N[8]. The expected value of the particle weight is unchanged, and on average only one every N particles is retained in the simulation. In a many-particle scheme, instead, a set of N particles is transformed into another set of M particles, where M can be a deterministic or random variable. An example of a many-particle scheme is the *comb method*, which is used in neutron-transport simulations[11]. In this scheme, M particles are generated with equal weight W/M, with states selected from the N original ones. In the simplest approach, the state assignment is performed randomly and independently for each particle. If the N initial particles have weights w_i, the probability of a new particle being assigned the state of particle i is:

$$p_i = \frac{w_i}{\sum_{j=1}^{N} w_j}. \tag{1}$$

The new particles can thus be expected to be distributed similarly to the original ensemble. A more efficient method is obtained by further correlating the assignment probabilities of different particles, in order to improve the statistics of the resulting ensemble. For the special case of $M = 1$, we obtain the so-called *gathering* scheme, where only one particle retains the total weight of the initial ensemble[9].

The goal of a variable-weight scheme is to find the optimum assignment of weights to the simulated particles, in order to minimize the variance of the estimators. A common approach to this problem is to partition phase space into discrete domains, or *bins*. For each phase-space bin, a reference weight is set. Particles that have a weight larger than the reference weight are split, while those with a smaller weight are removed, either with a single-particle or a many-particle procedure. This approach requires some adaptive algorithms to determine the reference weights during the simulation[8,9]. An alternative strategy is to apply a comb-type algorithm to each phase-space bin. Given the set of N particles contained in a bin, another set of M particles with equal weights is stochastically generated. With this simple and efficient scheme, the occupation of each bin is raised or lowered to any desired number of computational particles (M can be larger or smaller than N), and the particle weights are equalized.

We introduce here a third method, that attempts to heuristically distinguish "significant" from "redundant" particles. For example, we can tag as redundant all the particles that lie in phase-space bins where much larger particles are also present. We can reasonably assume that such small particles give a negligible contribution to the estimators, since the weighted averages are dominated by the larger (significant) particles. Once the set of all redundant particles has been identified, we can perform a gathering on its elements. A particle is marked as redundant when its weight is smaller than the weight of the largest particle in the same bin by a factor R_r, usually in the range 10^{-2}–10^{-1}. The reference weights are adaptively controlled by the weight of the largest particle in each bin, and no monitoring of the weights is necessary. Situations may also occur when no redundant particles are available, for example when all particles have the same weight. In such cases, this scheme must be supported by another "backup" method.

In the adaptive gathering scheme, redundant particles are collected from the entire simulation domain. We call this method *non-local*, because the gathering can be performed on particles lying in different bins. On the contrary, all many-particle schemes presented in the literature are *local*, i.e. the gathering of particles always affects particles lying in the same phase-space region[9,10]. A non-local scheme offers significant advantages, mainly in the reduction of the statistical noise due to the gathering itself. The latter is due to the random change of weight of the particles, which adds fluctuations to the

estimators. Consider for example an estimator A whose dominant contribution comes from only one particle with weight w_i:

$$A \approx a_i w_i. \tag{2}$$

Let us assume that this particle can suffer a change in weight by a multiplication factor M, from w_i to Mw_i (M need not be an integer). Therefore, there must be a probability $1 - 1/M$ that the particle be removed. From eqn (2), the final variance of the estimator will be $(M-1)A^2$. This suggests that, as a general rule, M should be close to one, i.e. the variance-reduction algorithm should be "smooth" rather than aggressive. For local schemes, this requires a large number of phase-space bins, with average weights very close to each other. In the non-local algorithm, instead, a large set of redundant particles is partitioned into sub-sets, based on the weights of the particles rather than on their location in phase space. The sub-sets are used as gathering sets, thus ensuring that only particles with similar weights are gathered. This is not always possible with a local scheme, because particles within a single phase-space bin can have largely different weights. If we denote with R_g the ratio between the maximum and minimum weight in the N_g-particle gathering set, the maximum possible value for the multiplication factor M will be:

$$M_{\max} \approx R_g N_g,$$

in the case when the total weight of the ensemble is assigned to the smallest particle. In a typical application to device simulation, we can use a value of R_g very close to unity. This limitation also ensures that, after a gathering, the retained particle will still be smaller than other particles in the same bin, so that the gathering algorithm does not modify its own parameters. The associated stability condition can be written as:

$$N_g R_g R_r < 1. \tag{3}$$

Since the event of the smallest particle being assigned the weight of the entire set is the most unlikely, eqn (3) can be regarded as fairly conservative.

3. SELF-CONSISTENT SIMULATIONS

So far, we have assumed that the transport properties of the particles were independent of their weight. This is the case if a fixed-field simulation is performed, i.e. if the electric field is computed by some independent means and treated as an input to the MC algorithm. However, a full self-consistent simulation is necessary for transient and frequency-response analysis, and may also be convenient for the steady-state analysis of small devices, where non-local transport effects dominate. In these cases, particles are assumed to carry a charge proportional to their weight. Poisson's equation is solved at short time steps, so that inter-particle forces can be adequately resolved. Since the mass of a computational particle scales with the same factor as the charge, the equations of motion for a given electric field are independent of the weight. On the other hand, particles with large weights affect the electrostatic potential more than particles with small weights.

It is well known that the Monte Carlo method solves the Boltzmann transport equation (BTE) for a system of physical particles, in the limit of a large number of computational particles[12]. We define the time-dependent one-particle distribution function $f_1(\vec{r}, \vec{k}, t)$ as the expected value of the density of particles in the position/wave vector phase space. (For simplicity, we include band index and spin within the $\vec{k}$-space coordinate.) To fully describe the statistical ensemble of particles, it would be necessary to know also the two-particle distribution $f_2(\vec{r}_1, \vec{k}_1, \vec{r}_2, \vec{k}_2, t)$, that accounts for correlations between pairs of particles; the three-particle distribution f_3; and so on. If the short-range correlations between particles can be neglected, it can be assumed that each charge carrier experiences only the average effect of all other carriers. We then write the Vlasov or Boltzmann transport equation (BTE), including the effect of external scatterings processes, by truncating the sequence of distribution functions to first order:

$$\frac{\partial f_1}{\partial t} + \vec{v}_g \nabla_r f_1 - \frac{e \nabla_r \phi}{\hbar} \nabla_k f_1 = \left(\frac{\partial f_1}{\partial t}\right)_{\text{scat}}, \tag{4}$$

where e is the particle charge, and the electrostatic potential ϕ is computed self-consistently from f_1 through Poisson's equation:

$$-\nabla(\epsilon \nabla \phi) = \rho(\vec{r}) = e \int d^3\vec{k} f_1(\vec{r}, \vec{k}, t). \tag{5}$$

It should be noted that eqn (4) is non-linear, due to the dependence of ϕ on f_1, as opposed to the linear BTE that holds, apart from degeneracy effects, in the case of an assigned field profile. The truncation of the sequence implies the so-called *collisionless plasma* approximation. The term "collision" is used here in the sense of interaction between charged particles, and does not refer to other scattering processes, e.g. electron–phonon interaction. Quantitatively, low collisionality is achieved if the second-order correction to the BTE is small with respect to the scattering term [the right-hand side of eqn (4)].

In the case of variable-weight MC, the particles are not identical and the usual concepts of statistical mechanics do not strictly apply. Nevertheless, we can still define in phase space a one-particle distribution function as the expected value of the *weight density*, as opposed to the number density:

$$f_1(\vec{r}, \vec{k}, t) = \langle F(\vec{r}, \vec{k}, t) \rangle \tag{6}$$

where $F(\vec{r}, \vec{k}, t)$ represents the generalized density of particles for a particular configuration of the ensemble:

$$F(\vec{r}, \vec{k}, t) = \sum_{i=1}^{N} w_i \delta(\vec{r} - \vec{r}_i(t))\delta(\vec{k} - \vec{k}_i(t)). \quad (7)$$

Again, in the limit of vanishing particle weights, short-range interactions average out, and we should get the correct solution of the BTE. Instead of discussing this property formally, we will give examples of simulations in a collisional and collisionless regime, and a physical analysis of the behavior of the system in the different cases. As will be shown in Section 4, the low-collisionality constraint is much more demanding for the case of variable-weight simulations than for the case of constant weights.

It has been observed by Laux and Fischetti[13,14] that there may be situations where collisional phenomena affect charge transport properties. In other terms, the finite amount of charge carried by individual particles can have a physical relevance. In these cases, the ensemble MC method can still give a solution for the semiclassical many-body problem, but the granularity of the physical system must be properly accounted for. For example, the weight assigned to the particles cannot be indefinitely small, but must be physically related to the electron charge and possibly to the screening length[13]. This, of course, would rule out any possibility of assigning arbitrary values to the particle weights, as we have assumed so far. However, it is outside the scope of this work to deal with such a subtle matter. For example, it is still an open question if, and how, the full Coulomb dynamics of the three-dimensional system can be effectively modeled by means of a two-dimensional simulation. In the following, we will limit ourselves to the discussion of a self-consistent system described by eqn (4). We will implicitly assume that collisional effects are either negligible, or can be introduced as an additional scattering mechanism, with all the limitations that this approach implies[14,15].

4. EXAMPLES

We present first an example of the adaptive non-local method as compared to a local scheme, for the case of a space-homogeneous simulation. We consider an ensemble of non-interacting electrons in silicon at 300 K, with an applied electric field of 300 kV/cm. The transport model incorporates two conduction bands for electron kinematics, where the band structure of silicon has been computed with the local empirical pseudopotential method[16]. The scattering processes included in the simulation are inelastic acoustic and optical phonons, and impact ionization[2,17]. Since space dependence has not been considered, phase space has been simply partitioned into 0.1 eV energy intervals. The same number of particles (5000) has been used for the two simulations. For the local simulation, the comb technique has been applied to each of the energy bins. For the non-local scheme, values of $R_r = 0.1$, $R_g = 1.5$ and $N_g = 6$ have been adopted. The stability criterion of eqn (3) is met, since $N_g R_g R_r = 0.9$. For this and the following simulations, splitting is performed on a one-particle basis, i.e. the largest particle in each bin is split into two identical particles with half the weight. Whenever storage for new computational particles is needed, gathering is performed on a subset of N_g redundant particles, thus making memory available for $N_g - 1$ additional particles. The gathering/splitting process is repeated until all the energy bins contain about the same number of particles. As mentioned before, the algorithm falls back to a local gathering scheme when no redundant particles are available. However, the largest part of the gatherings is still performed in the non-local mode. Figure 1 compares the relative error in the distribution function, as obtained from a constant-weight simulation and with the two variance-reduction schemes. According to a standard method for variance estimation[18], the fluctuation has been averaged on a large number of sub-histories, each with a duration of 100 fs. Both methods perform satisfactorily in reducing the statistical fluctuations, and achieve a nearly uniform population of the bins. With respect to the constant-weight result, both schemes increase the variance at low energy, in order to keep the error at high energies within an acceptable range. The non-local scheme obtains a slightly better accuracy at low energies, due to the lower changes of weight during the gathering process.

We begin the analysis of space-dependent results by discussing the one-dimensional simulation of a heavily doped slab of silicon at 300 K. The slab is 0.5 μm thick, n-doped with a donor concentration of

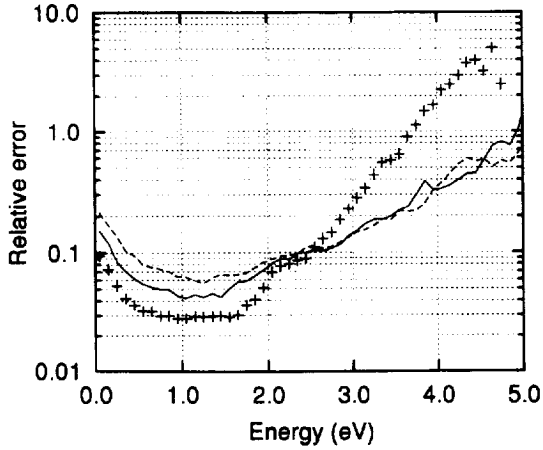

Fig. 1. Simulation results for electrons in bulk silicon with an applied field of 300 kV/cm. Crosses: constant-weight simulation. Solid line: non-local gathering. Dashed line: local comb method.

$10^{19}\,cm^{-3}$. For simplicity, degeneracy effects are neglected in the model. Boundary conditions for both electrostatic potential and particle concentration simulate the effect of ohmic contacts at the two faces of the slab. Self-consistency is obtained by solving the one-dimensional Poisson equation at time steps of 2 fs. The use of the cloud-in-cell scheme on a uniform mesh ensures that no unphysical self-force effects occur[12].

As a general rule, a high-density electron gas can be classified as collisional or collisionless, based on the ratio of the average kinetic and interaction energies, or equivalently on the ratio of the Debye length to interparticle spacing[12]. In our case, we used a number of particles ranging from 1000 to 10 000, with an average interparticle spacing of 0.5–0.05 nm. The Debye length for a concentration of $10^{19}\,cm^{-3}$ in silicon is about 1.3 nm. For the case of 1000 particles, the Coulomb interaction energy thus accounts for a relevant fraction of the total energy of the ensemble. We can expect that in the case of 10 000 particles, with about 26 particles within a screening range, the system will be approximately collisionless.

We performed simulations of this system both with constant weights and with a variable-weight scheme. In the latter case, the non-local gathering technique was used, with parameters similar to those of the previous example. Phase space was partitioned on the basis of energy and position in real space. Non-local gathering was performed on the entire set of bins. The energy range was partitioned into 0.1 eV intervals. The width of the space bins, instead, was computed automatically by a simple algorithm in order to maintain a minimum number of particles (5 in this case) in each space/energy bin. Assuming for simplicity that particles have unit weight in the constant-weight simulations, we fixed an upper bound of one for the particle weights in the variable-weight simulations. This precaution avoids excessive fluctuations of charge and ensures a comparable granularity of the two systems.

The energy distributions obtained for the entire slab are shown in Fig. 2 for different numbers of particles. We found the constant-weight simulation result to be independent of the number of particles. In fact, the single-particle energy distribution is still very close to a Maxwell–Boltzmann distribution, although a non-negligible part of the total energy is stored as electrostatic potential energy. In the variable-weight simulation, instead, the high-energy tail clearly shows a higher temperature, tending to the lattice temperature as the degree of collisionality decreases. This result is independent of the time step adopted for the solution of the Poisson solution, which in any case is much shorter than the time step of about 25 fs dictated by the Nyquist criterion for the sampling of the plasma oscillations[13]. The interpretation of this artifact is simple: computational particles carry a total energy proportional to their

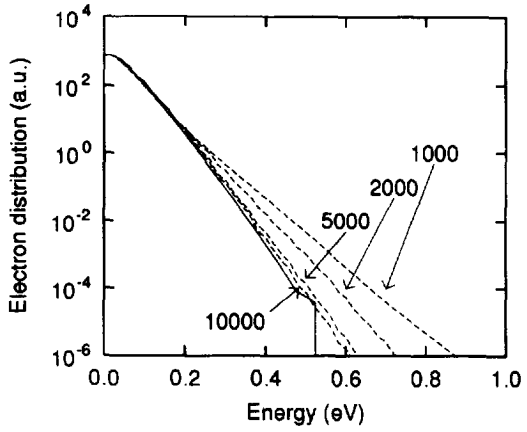

Fig. 2. Simulation of a homogeneous slab of silicon with a concentration of carriers of $10^{19}\,cm^{-3}$, for various numbers of particles. Solid line: simulation with constant weights. Dashed lines: variable-weight simulations with 1000, 2000, 5000 and 10 000 particles.

weight, so that particles with large weight transfer large energies to the electron plasma. Therefore, in all simulations the field fluctuations are dominated by the unit-weight particles. At thermal equilibrium, these large particles emit and absorb equal amounts of energy with respect to the plasma. On the other hand, particles with small weights, which mostly lie at high energies, represent fewer physical carriers, and cannot excite the plasma. They can only *absorb* energy when coupled to the collective modes, and this causes the unphysical deviation from the Maxwellian tail. The effect disappears when the energy exchange with plasma modes becomes weak as compared to electron–phonon interaction, which is independent of the granularity of the simulation [cf. the discussion of eqn (4)]. This behavior is not unique to the gathering scheme, but arises from the use of variable-weight particles. We also remark that this problem is not due to a biasing of the estimators, since the latter are exactly the same for all simulations. Rather, it is due to the fact that the variable-weight simulation cannot account properly for all the many-body effects occurring in the physical system.

To illustrate a practical application to device simulation, we consider the calculation of the gate current in a simple *n*-MOSFET test structure. The oxide thickness is 100 Å, the metallurgical channel length is 0.5 μm, source and drain peak dopings are $5 \times 10^{19}\,cm^{-3}$, and the substrate *p*-doping is $5 \times 10^{16}\,cm^{-3}$. Bias voltages are 4 and 3 V for the drain and *n*-polysilicon gate, respectively. A nonlinear two-dimensional Poisson solver has been employed, including holes in the simulation self-consistently under the assumption of a constant hole quasi-Fermi level. Poisson's equation is solved every 1.0 fs of physical time, in order to avoid instabilities in the high-doping regions[8]. For purposes of illustration, we performed a full self-consistent

simulation, although a fixed-field simulation would be more appropriate for this problem. Since incomplete ionization of dopants is included in the model, the concentration of mobile carriers at the contacts is about $10^{19}\,cm^{-3}$. The low-collisionality criterion forces us to use a number of particles much larger than 15 000, in order to keep many particles within a screening circle of radius πL_s^2, where the screening length $L_s = 1.3$ nm. Drawing a balance between numerical accuracy and computational cost, we used a minimum of 80 000 particles, allowing for 80 000 more to resolve the high-energy tails. We adopted the same limitation of weights as in the one-dimensional case, in order to retain the small fluctuations of the constant-weight simulation. Figure 3 shows the distribution of electrons as a function of X (lateral) position along the channel and of kinetic energy, averaged along the Y (vertical) direction. The "bumps" in the drain region are due to the energy dependence of the scattering rate, that makes hot electrons at different energies relax with different rates.

The oxide-injection current has been computed assuming conservation of the crystal momentum parallel to the Si/SiO_2 interface[19,14]. The tunneling probability has been evaluated by a numerical integration in the WKB approximation, assuming a trapezoidal barrier with image-force lowering on both sides. Since the resulting transmission coefficient is strongly momentum-dependent, the partitioning of phase space has been optimized in order to reduce the computation time. Particles are assigned to phase-space bins according to their (a) position, (b) energy and (c) momentum parallel to the interface. Thus, the variance-reduction algorithm always maintains a large number of carriers in states with a small parallel momentum, resulting in a high injection probability. Again, the width of the energy bin is 0.1 eV, while the real-space partitioning is chosen in such a way to allocate a sufficient number of particles to each bin. With such a multi-dimensional partitioning, several thousand bins

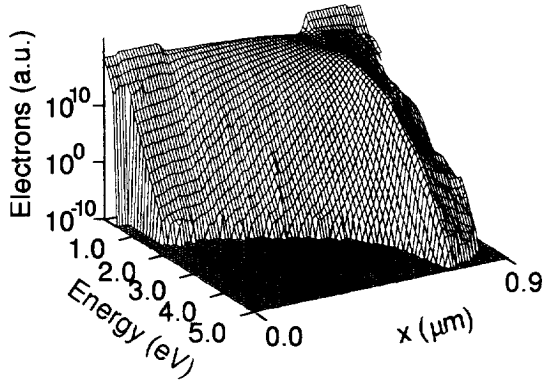

Fig. 3. Position/energy distribution of electrons along the channel of a 0.5 μm n-MOSFET, as computed from MC simulation.

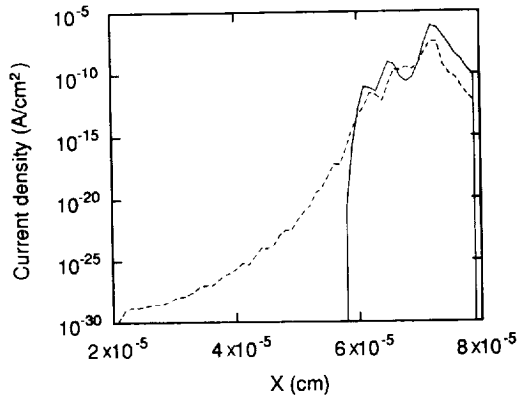

Fig. 4. Oxide current density for a 0.5 μm MOSFET, as computed from the MC simulation. Solid line: thermionic current. Dashed line: tunneling current.

are needed in order to ensure a uniform sampling of the device.

Figure 4 shows the position dependence of the oxide-injection current, separating the thermionic-emission and tunneling components. The abrupt step in the thermionic current around $X = 580$ nm is due to the finite dynamic range of weights. The soft increase of the tunneling component reflects the smooth dependence of the transmission probability on electron energy. Thermionic emission only depends on conservation of energy and crystal momentum parallel to the interface, so that a sharp increase occurs as the energy supplied by the field approaches the barrier height. The dip around $X = 680$ nm occurs exactly at the end of the gate electrode, and is not due to statistical noise. In fact, a retarding potential of 1 V is applied between the drain and gate electrodes. The resulting electric field increases the effective oxide barrier and pushes electrons away from the interface. As hot electrons proceed further into the drain electrode, the vertical field decreases and the possibility of oxide injection rises again.

In all the examples shown, the use of the splitting/gathering procedure takes only a very small fraction of the total processing time. This is due to the simplicity of the scheme and to the efficiency of the non-local sweep of the ensemble of particles. The method is thus applicable with no significant penalty even when the number of particles and phase-space bins is very high, as in the case of the MOSFET simulation.

5. CONCLUSIONS

We have presented a general discussion of variable-weight techniques for Ensemble Monte Carlo simulation of semiconductor devices. Existing methods have been classified and compared, and a novel scheme has been introduced, that generalizes the gathering method of Ref. [9] with the use of

non-local, adaptive gathering sets. Our analysis and computer experiments show that a variable-weight simulation requires a lower degree of collisionality with respect to a conventional MC simulation. This limits the use of such methods to non-self-consistent simulations, or to cases where collisional effects can be neglected. Application to device simulation has been illustrated in detail, discussing the practical case of a submicron *n*-MOSFET.

Acknowledgements—The authors wish to thank Amanda W. Duncan for implementing the MOSFET simulation code, and Thomas E. Booth for supplying information about the comb variance-reduction method. This work has been supported by the Volta–Badoni Fellowship of the Associazione Elettrotecnica ed Elettronica Italiana and the Italian Ministry of Scientific Research (A. P.) and by the Joint Services Electronics Program, grant N00014-96-1-0129, and the Semiconductor Research Corporation, contract 95-CS-816 (U.R.).

REFERENCES

1. Phillips, Jr. A. and Price, P. J., *Appl. Phys. Lett.*, 1977, **30,** 528.
2. Tang, J. Y. and Hess, K., *J. Appl. Phys.*, 1983, **54,** 5139.
3. Sangiorgi, E., Riccò, B. and Venturi, F., *IEEE Trans. Comput.-Aided Design Integrated Circuits*, 1988, **CAD-7,** 259.
4. Ranawake, U. A., Huster, C., Lester, P. M. and Goodnick, S. M., *IEEE Trans. Comput.-Aided Design Integrated Circuits*, 1994, **13,** 712.
5. Lee, C. H., Ravaioli, U., Hess, K., Mead, C. A. and Hasler, P., *IEEE Electron Device Lett.*, 1995, **16,** 360.
6. Rossi, F., Poli, P. and Jacoboni, C., *Semiconductor Science Technol.*, 1992, **7,** 1017.
7. Venturi, F., Sangiorgi, E., Luryi, S., Poli, P., Rota, L. and Jacoboni, C., *IEEE Trans. Electron Devices*, 1991, **ED-38,** 611.
8. Fischetti, M. V. and Laux, S. E., *Phys. Rev. B*, 1988, **38,** 9721.
9. Venturi, F., Smith, R. K., Sangiorgi, E., Pinto, M. R. and Riccò, B., *IEEE Trans. Comput.-Aided Design Integrated Circuits*, 1989, **CAD-8,** 360.
10. Liebig, D., Lugli, P., Vogl, P., Claassen, M. and Harth, W., *Microelectronic Engineering*, 1992, **19,** 127.
11. Booth, T. E., Private communication.
12. Hockney, R. W. and Eastwood, J. W., *Computer Simulation Using Particles*. Institute of Physics, Bristol-Philadelphia, 1987.
13. Laux, S. E. and Fischetti, M. V., in *Monte Carlo Device Simulation: Full Band and Beyond*, ed. K. Hess. Kluwer Academic Publishers, Dordrecht, 1991.
14. Fischetti, M. V., Laux, S. E. and Crabbé, E., *J. Appl. Phys.*, 1995, **78,** 1050.
15. Mansour, N. S., Janzou, S. H. and Brennan, K. F., *J. Appl. Phys.*, 1992, **72,** 5277.
16. Cohen, M. L. and Bergstresser, T. K., *Phys. Rev.*, 1966, **141,** 789.
17. Bude, J., Hess, K. and Iafrate, G. J., *Phys. Rev. B*, 1992, **45,** 10958.
18. Jacoboni, C. and Lugli, P., *The Monte Carlo Method for Semiconductor Device Simulation*. Springer-Verlag, Wien–New York, 1989.
19. Krieger, G. and Swanson, R. M., *J. Appl. Phys.*, 1981, **52,** 5710.