LE MATEMATICHE
Vol. LXIV (2009) – Fasc. II, pp. 57–69

ENHANCEMENT OF HIGH-ENERGY DISTRIBUTION TAIL IN MONTE CARLO SEMICONDUCTOR SIMULATIONS USING A VARIANCE REDUCTION SCHEME

VINCENZA DI STEFANO

The Multicomb variance reduction technique has been introduced in the Direct Monte Carlo Simulation for submicrometric semiconductor devices. The method has been implemented in bulk silicon. The simulations show that the statistical variance of hot electrons is reduced with some computational cost. The method is efficient and easy to implement in existing device simulators.

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. Statistical enhancement (also called variance reduction) was introduced in Monte Carlo simulation partly to compensate for a smaller number of simulated particles as compared to a larger number of actual particles in a device, especially in

Entrato in redazione: 20 ottobre 2009

AMS 2000 Subject Classification: 65C05, 82D37
Keywords: Monte Carlo device simulation, particle comb, variance reduction, statistical enhancement

sparsely populated regions of phase space. 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 [7] . In the second approach the scattering and the particle weights are modified [5]. The third class of variance-reduction methods, called population control, includes the variable-weight techniques, where a statistical weight factor is assigned to each particle, while the MC simulation remains unchanged [1, 6, 9]. The goal of such method 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. In this paper we shall study this algorithm for the bulk silicon, discussing its limitations.
The plan of the paper is the following: in section II we introduce the Direct Simulation Monte Carlo. In section III the *Multicomb* algorithm is formulated and simulation results are shown for bulk silicon in section IV. Conclusions are drawn in section V.

2. The Direct Simulation Monte Carlo

The Monte Carlo method produces a statistical solution of the Boltzmann transport equation, which is an integro-differential equation which describes the time evolution and the variation in the phase space of the unknown distribution function $f(t,x,k)$

$$\left[\frac{\partial}{\partial t} + v(k)\cdot\nabla_x - \frac{q}{\hbar}E(t,x)\cdot\nabla_k\right] f(t,x,k) = (Qf)(t,x,k)\,. \tag{2.1}$$

The solution $f(t,x,k)$ represents the probability density of finding an electron at time t, in the position x, with the wave-vector k.
In the *quasi parabolic* approximation the kinetic energy $\varepsilon(k)$ of an electron satisfies the relation

$$\varepsilon(k)\,[1+\alpha\varepsilon(k)] = \frac{\hbar^2|k|^2}{2m^\star}\,, \qquad k\in\Omega\,, \tag{2.2}$$

and the electron (group) velocity is given by

$$v(k) = \frac{1}{\hbar}\nabla_k\,\varepsilon(k) = \frac{\hbar k}{m^\star[1+2\alpha\varepsilon(k)]}\,. \tag{2.3}$$

In the previous equations q is the absolute value of the electron charge, $m^\star$ is the effective electron mass and equals 0.32 m_e in silicon, α is the nonparabolicity factor and $\hbar$ denotes Planck's constant divided by 2π. The domain Ω is called first Brillouin zone, which is a characteristic of each material. In silicon this zone is formed by six equivalent ellipsoidal valleys along the axis of the frame of reference at about 0.85 (in the units $\frac{2\pi}{a}$ where a is the lattice constant) from the zone center.
The electric field is defined as

$$E(t,x) = -\nabla_x \Phi(t,x)\,. \tag{2.4}$$

The electric potential Φ is related to the solution f by the Poisson equation

$$\varepsilon\,\Delta_x \Phi(t,x) = q\,[n(t,x) - N_D(x)]\,, \tag{2.5}$$

where the electron density is given by

$$n(t,x) = \int_\Omega f(t,x,k)\,dk\,. \tag{2.6}$$

Here N_D denotes the donor density, and ε is the permittivity. The linear scattering collision operator has the form

$$(Q\,f)(t,x,k) = \int_\Omega S(k',k)\,f(t,x,k')\,dk' - \lambda(k)\,f(t,x,k)\,,$$

where

$$\lambda(k) = \int_\Omega S(k,k')\,dk' \tag{2.7}$$

is the total scattering rate.
The main scattering mechanisms in silicon, at room temperature, are due to electron-phonon interactions (acoustic and optical). The transition rate from a state k to a state k' is modeled as [2]

$$\begin{aligned} S(k,k') \;=\; & K_0\,\delta(\varepsilon(k') - \varepsilon(k)) + \sum_{i=1}^{6} K_i \times \\ & \left[\delta(\varepsilon(k') - \varepsilon(k) + \hbar\omega_i)(\mathsf{n}_{q_i} + 1) + \delta(\varepsilon(k') - \varepsilon(k) - \hbar\omega_i)\,\mathsf{n}_{q_i}\right], \end{aligned} \tag{2.8}$$

where $\hbar\omega_i$ is a phonon energy. According to Bose-Einstein statistics, the phonon equilibrium distribution is given by

$$\mathsf{n}_{q_i} = \frac{1}{\exp(\hbar\omega_i/k_B T_L) - 1}\,,$$

where T_L is the lattice temperature. The function

$$K_0 = \frac{k_B T_L \Xi_d^2}{4\,\pi^2 \hbar\,\rho\, v_s^2}$$

represents the intravalley elastic scattering transition rate, where Ξ_d is the acoustic-phonon deformation potential, ρ is the silicon mass density and v_s denotes the sound velocity of the longitudinal acoustic mode. The inelastic scattering rates have the form

$$K_i = \frac{Z_f (D_t K_i)^2}{8\,\pi^2 \rho\,\omega_i}, \qquad i = 1, \ldots, 6,$$

where $D_t K_i$ is the deformation potential for the i-th optical phonon, and Z_f is the number of final equivalent valleys for the considered inter-valley scattering. The Monte Carlo method for evolving a solution of the Boltzmann transport equation consists in recreating the history evolution of electrons in time and space inside the crystal, subject to the action of external and self-consistent electric field and of the given scattering mechanisms [2, 8]. The simulation starts with one or more electrons in given initial conditions with momentum $\hbar k$ and position x. During the *free flight* (i.e. the time between two collisions) particles move according to Newton's equations of motion

$$\frac{dx}{dt} = \frac{1}{\hbar}\nabla_k \varepsilon(k) \tag{2.9}$$

$$\hbar\frac{dk}{dt} = -q\mathsf{E}(t,x) \tag{2.10}$$

The equations (2.9),(2.10),(2.5) are solved with a stable numerical scheme by using an appropriate time step Δt [4]. Then a scattering mechanism is chosen randomly as responsible for the end of the free flight, according to the relative probabilities of all possible scattering mechanisms. From the differential cross section of this mechanism (eq. (2.7)) a new k state after scattering is randomly chosen as initial state of the new free flight. After the collision the electron can remain in the same valley (intravalley scattering) or be drawn in another equivalent valley (intervalley scattering).

The electrons can scatter by themselves, with the impurities and the lattice. The electron-electron interaction is considered in the framework of the mean field approximation through the Poisson equation. This is reasonable since we consider the case of low doping and therefore we can neglect the short range collisions between electrons.

3. The algorithm

In this paper we have chosen the population control technique because of its simplicity and user-friendliness in the implementation. In particular, we have used the so called *Multicomb* algorithm which has been introduced, in the beginning, in the field of neutral particle transport.
Let us suppose to have N initial particles with momentum $\hbar k_i$, energy ε_i and weight w_i. At time zero we choose

$$w_i = \frac{1}{N} \quad .$$

Then we divide the energy space in K bins (or stat-boxes), whose extension is

$$\Delta\varepsilon = \frac{\varepsilon_M}{K}$$

where ε_M is the estimated max energy reached by all particles during the run. So, the energy space is partitioned into $K+1$ bins, i.e.

$$[0, \Delta\varepsilon[, ..., [(K-1)\Delta\varepsilon, K\Delta\varepsilon[, [K\Delta\varepsilon, +\infty[.$$

For each j-th bin we can define :

- the number of particles N_j in the bin, such that

$$N = \sum_{j=1,K+1} N_j$$

- the total weight, as the sum of the weights of the particles which are in the j-th bin

$$W_j = \sum_i w_i \quad , j = 1, .., K+1 \quad .$$

Let's suppose the stationary regime is reached. Then we decide to run the enhancement algorithm each Δt_{enh}, which we shall call enhancement time step. Let be N_b the number of non-empty bins. Then we define

$$M = \text{int}\left[\frac{N}{Nb}\right] \quad , \quad rest = N - M \times Nb$$

where M is the target particle number in each bin (of course, by definition $M \leq N$), and *rest* is the division rest. In order to have constant particle number N during the simulation, we make this assumption :

$$M_1 = M + rest, \quad M_2 = M, ..., M_{Nb} = M$$

where M_1 is the target particle number in the first bin and so on.
The enhancement algorithm applied to each non-empty bin is called *simple comb*. Let us consider the j-th non-empty bin: this is the *simple comb* algorithm

1. we construct a comb of length W_j with M equally spaced teeth. The position of the m-th tooth is given by

$$t_m = (\xi + m - 1)\frac{W_j}{M_j}$$

where ξ is an uniform random number in [0, 1).

2. Place the weights w_i consecutively on a line segment, obtaining N_j bins

$$[0, w_1],]w_1, w_2],]w_2, w_3], ...,]w_{N_j-1}, w_{N_j}]$$

3. Now we "comb" this line segment with the previous one, obtained in 1).
do $m = 1, .., M_j$

if $t_m \in]w_{i-1}, w_i]$, for some i=1, ..,N_j
a copy of the i-th particle from the old distribution is added to the new distribution, with assigned weight

$$w_i' = \frac{W_j}{M_j}$$

else no copy.

enddo

The algorithm maintains constant, in the j-th bin, the total weight W_j and particle number N_j. It is possible to prove that [1]: i) the *simple comb* preserves on average the individual weights of the pre-combed particles; ii) the distribution of particles with identical weights produced by the comb gives the smaller variance than any distribution with unequal weights. The application of different *simple comb* to all non-empty bins is called *Multicomb* method. Consequently, during the simulation, the *Multicomb* maintains constant (by default) the total particles number N, and the sum of the overall weights

$$W = \sum_{j=1}^{K+1} W_j \quad .$$

4. Results

We have tested the *Multicomb* algorithm in a bulk silicon semiconductor, doped to a density of 10^{15}cm^{-3}, using a total target ensemble size of 11000 computational electrons, in which a constant electric field has been frozen. In this case the particles are independent and we can apply the previous variance reduction technique, developed under the hypothesis of independent particles.

To determine the electron energy distribution (EED) function, we have discretized the whole energy space in a system of concentric shells with increasing radius

$$\rho_i = ih, \quad i = 1, ..., K, \quad h = \frac{\varepsilon_M}{K}$$

(in our case ε_M= 2 eV and K = 200) and by counting the number of particles which are in the corresponding shells, i.e.

$$\begin{aligned} f_1 &= \{\#\text{particles} : \varepsilon < \rho_1\} \\ f_i &= \{\#\text{particles} : \rho_{i-1} \leq \varepsilon < \rho_i\} \\ f_{N_\varepsilon+1} &= \{\#\text{particles} : \varepsilon \geq \varepsilon_M\} \quad . \end{aligned}$$

The confidence interval for each point of the EED is obtained by means of the Central Limit Theorem in the following way. Let us suppose to make N_r independent runs (repetitions). Then we obtain f_{ij} quantities (i=1,...,K , j=1,...,N_r). For simplicity we shall call $\xi_j = f_{ij}$ omitting the index i. The confidence interval is evaluated as

$$\bar{\xi} \pm S_\xi$$

with

$$\bar{\xi} = \frac{1}{N_r} \sum_{j=1}^{N_r} \xi_j$$

$$S_\xi = 3 \sqrt{\frac{1}{N_r} \left(\frac{1}{N_r} \sum_{j=1}^{N_r} \xi_j^2 - \left[\frac{1}{N_r} \sum_{j=1}^{N_r} \xi_j \right]^2 \right)}$$

where the factor 3 corresponds to a 99.7 % confidence level. In our case, each repetition had a duration of 20 ps, but data were discarded during the 5 ps initial transient.

In the figures 1,2 we plot the EED functions with the confidence interval, obtained with the *Multicomb* algorithm and without (*Unenhanced*), where an electric field of 48.000 and 120.000 V/cm have been frozen. The accuracy of the Multicomb method is verified by the fact that its errors fall within the error of the unenhanced method (note that the error bars appear asymmetrical about the mean because the vertical axis has a logarithmic scale).

In the figure 1 the maximum energy reached during the unenhanced simulation is 0.44 eV, whereas that with the Multicomb is 0.52 eV, obtaining an enhancement of the tail.

The Relative Error

$$\mathrm{RE} = \frac{S_\xi}{\bar{\xi}}$$

is plotted in the figures 3,4 showing that the Multicomb method gives a lower error at high energies (i.e. ≥ 0.2 eV for an electric field of 48.000 V/cm, ≥ 1.2 eV for an electric field of 120.000 V/cm). Although the relative error provides useful information about the performance of the methods, it does not take into account the CPU time used. In general, the variance reduced methods pay the price of a greater CPU time consumption, with respect to the unehnanced ones because more control instruction must be added in the code. For example, for an electric field of 48.000 V/cm, the CPU time for Multicomb has been 12100 sec and that for the unenhanced 5200 sec. A useful parameter for comparing the performance of the two methods is the *Figure of Merit*, that takes into account the relative error as well as the CPU time.
We define the Figure of Merit (FoM), for the EED, as

$$\mathrm{FoM} = \frac{1}{(\mathrm{RE})^2\, T_{CPU}}$$

where T_{CPU} is the total CPU time. Since the Relative Error is proportional to $1/\sqrt{Nr}$ and the time T_{CPU}, that takes to run Nr runs, is proportional to Nr, then the FoM is independent of Nr. Because the FoM is inversely proportional to the total CPU time, a method which approaches a given level of error faster will have a higher figure of merit.
In the figures 5, 6 we plot the FoM. From these figures is evident that for low energies there is an additional cost in the Variance Reduction, whereas at high energies the Variance Reduction code is better than the normal code.
Another way to compare the efficiency is to run the unenhanced code for a longer simulation time T_{sim} (i.e. with a longer CPU time), to have the same maximum energy (or the same EED tail) of the *Multicomb*.
We have the following table, obtained with an electric field of 48.000 V/cm

CPU times

Method	Unenh	Unenh	Multi
T_{sim} (ps)	20	140	20
Max ene (eV)	0.44	0.52	0.52
T_{CPU} (sec)	5200	36400	12100

where Unenh and Multi indicate respectively the Unenhanced and Multicomb methods, Max ene is the maximum energy reached during the run, and T_{sim} is the total simulation time.

In order to reach the same max energy, the unenhanced algorithm must run for a longer simulation time (T_{sim}= 140 ps), consuming 36400 sec of CPU, which means a factor 3 with respect to the Multicomb algorithm, whose CPU time is 12100 sec. Similar results have been obtained with a higher electric field. These results have been obtained with opteron dual core processors.

5. Conclusions

We have presented a Variance Reduction method for Monte Carlo simulations in bulk silicon, in the class of population control techniques. This algorithm provides the population of high energy electrons and give better information about the high energy distribution. The cost associated with Multicomb has been estimated as a factor 3 respect to the normal algorithm.

Acknowledgments

The work has been supported by ”Progetto Giovani Ricercatori GNFM 2009” and ”Progetti di ricerca di Ateneo”, University of Catania.

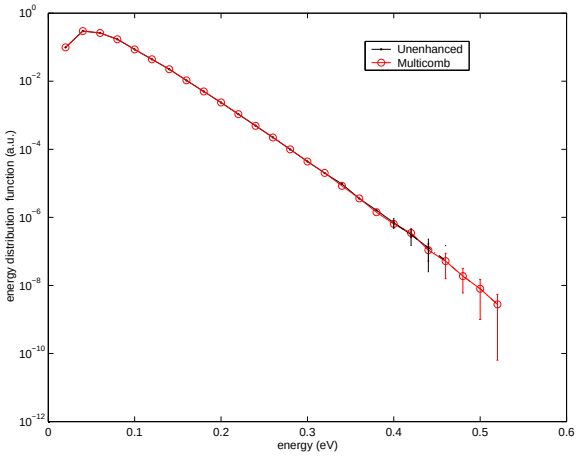

Figure 1: The energy distribution with error bar, for the unenhanced and multicomb methods, averaged over 100 simulations of 20 ps and using 11000 electrons, obtained with an electric field of 48000 V/cm.

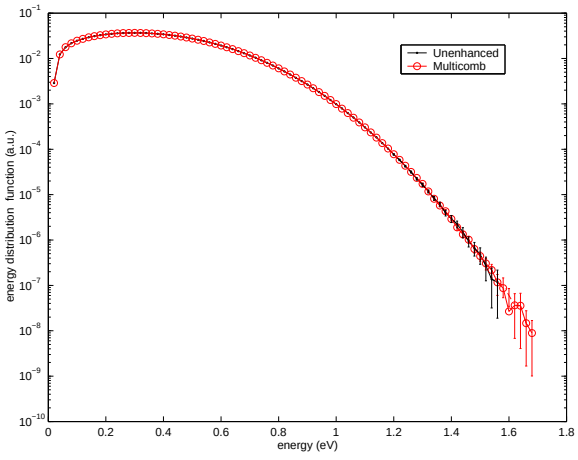

Figure 2: The energy distribution with error bar, for the unenhanced and multicomb methods, averaged over 100 simulations of 20 ps and using 11000 electrons, obtained with an electric field of 120000 V/cm.

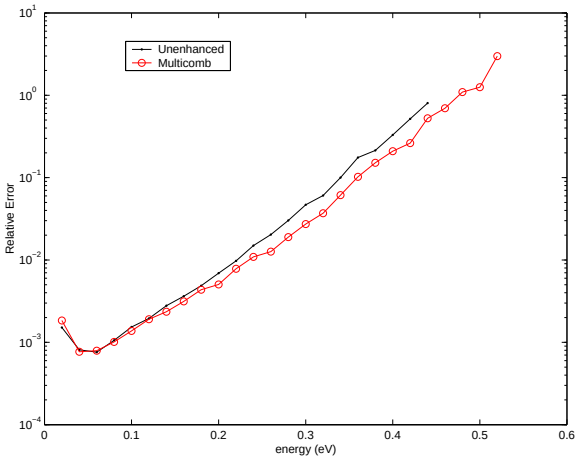

Figure 3: Relative error for the energy distribution versus energy for the unenhanced and multicomb methods, averaged over 100 simulations of 20 ps and using 11000 electrons, obtained with an electric field of 48000 V/cm

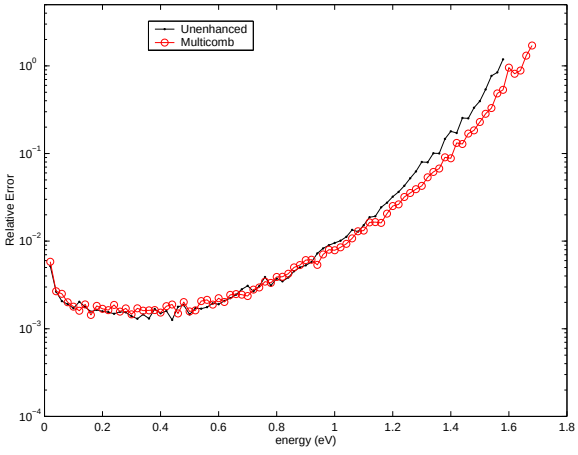

Figure 4: Relative error for the energy distribution versus energy for the unenhanced and multicomb methods, averaged over 100 simulations of 20 ps and using 11000 electrons, obtained with an electric field of 120000 V/cm

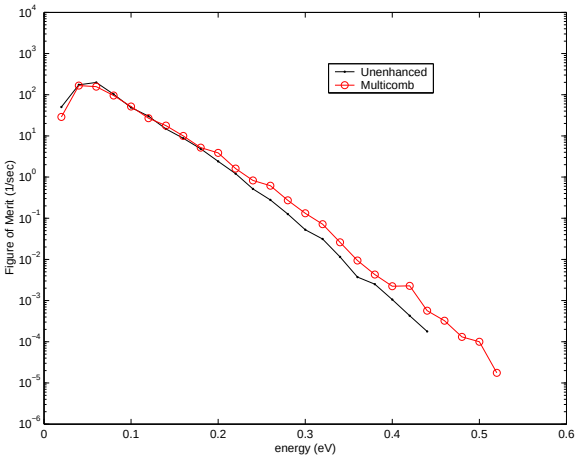

Figure 5: Figure of merit versus the energy for the unenhanced and multicomb methods, using 11000 electrons, obtained with an electric field of 48000 V/cm

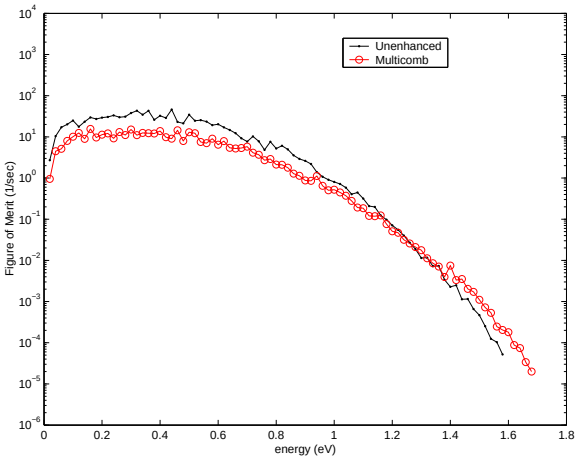

Figure 6: Figure of merit versus the energy for the unenhanced and multicomb methods, using 11000 electrons, obtained with an electric field of 120000 V/cm

REFERENCES

[1] M. G. Gray - T. E. Booth - T. J. T. Kwan - C. M. Snell, *A Multicomb variance reduction scheme for Monte Carlo semiconductor simulators*, IEEE Trans. Elec. Dev. 45 (4) (1998), 918–924.

[2] C. Jacoboni - L. Reggiani, *The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials*, Rev. Modern Phys. 55 (3) (1983), 645–705.

[3] P. A. Markowich - C. A. Ringhofer - C. Schmeiser, *Semiconductor equations*, Springer-Verlag, Vienna, 1990.

[4] O. Muscato - W. Wagner, *Time step truncation in direct simulation Monte Carlo for semiconductors*, COMPEL 24 (4) (2005), 1351–1366.

[5] M. Nedjalkov - H. Kosina - S. Selberherr, *The stationary Monte Carlo method for device simulation.II. Event biasing and variance estimation*, Journal of Applied Physics 93 (6) (2003), 3564–3571.

[6] A. Pacelli - U. Ravaioli, *Analysis of variance-reduction schemes for ensemble Monte Carlo simulation of semiconductor devices*, Solid- State Electron. 41 (4) (1997), 599–605, .

[7] Jr. A. Phillips - P. J. Price, *Monte Carlo calculations on hot electron energy tails*, Appl. Phys. Let. 30 (1977), 528–532.

[8] K. Tomizawa, *Numerical Simulation of Submicron Semiconductor Devices*, Arthech House, Boston, 1993.

[9] C. J. Wordelman - T. E. Booth - C. M. Snell, *Comparison of statistical enhancement methods for Monte Carlo semiconductor simulations*, Computer-Aided Design of Integrated Circuits and Systems, IEEE Transaction 17 (12) (1998), 1230–1235.

VINCENZA DI STEFANO
Dipartimento di Matematica e Informatica
Viale Andrea Doria 6 – 95125 Catania, Italy
e-mail: vdistefano@dmi.unict.it