4-4

Phase Space Multiple Refresh: A Versatile Statistical Enhancement Method for Monte Carlo Device Simulation

Chr. Jungemann, S. Decker[†], R. Thoma[‡], W. L. Engl[§], and H. Goto

Fujitsu Limited, Kawasaki, 211-88, Japan (Company Mail No.: C832)
Tel.: +81 44 754 2401, Fax.: +81 44 754 2575, E-mail: junge@ulsim.ed.fujitsu.co.jp
[†]University of Bremen, Kufsteiner Strasse, Postfach 33 04 40, 28334 Bremen, Germany
[‡]Motorola Semiconducteurs SA, Avenue du General Eisenhower, 31023 Toulouse Cedex, France
[§]RWTH Aachen, Kopernikusstr. 16, 52074 Aachen, Germany

In this abstract we describe a new Multiple Refresh (MR) technique for Monte Carlo (MC) device simulation which is an improved version of a method developed for bulk simulations [1]. The purpose of this method is to enhance statistics of rare events (e.g. impact ionization, oxide injection) without increasing the number of common events (i.e. low energetic particles). The MR technique allows to control *directly* the stochastic noise of the MC simulation in predefined regions of phase space by maintaining a given number of particles in these regions. To this end the MC simulation is interrupted at fixed times when the particle ensemble is synchronous (e.g. before the update of the Poisson equation) and the state of the subensemble in each region is checked. In the previous version of the MR technique the subensemble in each region was *always* replaced (refreshed) by an ensemble of uniformly weighted particles of the desired number (cf. [1]). This refresh conserves *exactly* the particle charge in each region in contrast to all other methods which only conserve the expected value of charge [2,3]. Exact conservation of particle charge is an important prerequisite for the stochastic stability of a statistical enhancement method. In the case of other averages of the subensembles (e.g. velocity, energy) it can be proved that their *expected values* with respect to the stochastic MR technique are conserved [4]. Exact conservation of these averages themself is not possible since for example a reduction in particle number always results in a loss of information. Although the MR technique in general reduces the stochastic noise, it also introduces with each refresh a small portion of noise due to its stochastic nature (like any other method using random numbers [2,3]). Therefore in the new MR technique this stochastic noise is minimized by refreshing a subensemble only when it is *necessary*. This is decided by two criteria. The first one controls the particle number and a subensemble is refreshed when its particle number N deviates more than a factor r_{par} from the desired number of particles N_{des} [5]:

$$N > r_{par} \cdot N_{des} \qquad or \qquad N < \frac{N_{des}}{r_{par}} . \tag{1}$$

This criterion alone does not guarantee an efficient simulation. Between two refreshs particles of different regions may mix and as a result a region may contain a subensemble consisting of particles with non uniform statistical weight. In this case the statistic is dominated by the particles with the highest weight and cpu-time is wasted by simulating particles with low weight in the same region. The variance (noise) of an average quantity of the subensemble is proportional to the sum over the squares of the particle weights. The minimum (optimum) value of this sum is found for a uniformly weighted particle subensemble [4]. Thus a refresh of a subensemble is performed whenever this sum deviates by more than a factor r_{squ} from its minimum value to ensure efficiency (the second criterion):

$$\sum_{i=1}^{N} w_i^2 > \frac{r_{squ}}{N} , \tag{2}$$

where w_i is the statistical weight of the particles normalized to one within a region ($\sum_{i=1}^{N} w_i = 1$). These two criteria allow a stable and efficient use of the MR technique and make its application to phase space possible (in contrast to [1]). Good results are obtained when for the parameters r_{par} and r_{squ} a value of 2 is used independently from the position of the region in phase space.

The MR technique is easy to implement into a MC program since only the particle states at certain times have to be known for its application. In other respects this method is independent from the rest of the MC program. Furthermore the refreshs of the different regions of the phase space are independent, allowing parallelization of the method *without any overhead*. Since the refreshs at different times are independent, the partitioning of phase space and the number of particles per region can be changed during the simulation. This makes adaptive partitioning of phase space possible and easy.

The new MR technique has been implemented into the parallel Full Band Ensemble Monte Carlo device simulation program Falcon [6]. It takes only a few percent of the total cpu-time spent for a simulation.

In Fig. 1 the electron density in the drain region of a 0.5μm-LDD-NMOSFET is shown calculated with the MR technique. This made it possible to calculate the low densities ($10^{10}/cm^3$) in the vicinity of the drain junction, needed to explore the limitations of hydrodynamic models [7]. Comparison with a simulation performed for the same cpu-time without MR (Fig. 1) reveals the dramatic improvement of statistics over many orders of magnitude. Thus without MR the investigations in [7] would have been impossible since a reduction of noise by orders of magnitude without MR results in an increase of cpu-time by orders of magnitude. In Fig. 2 the electron impact ionization generation rate

distribution for a 0.16μm-NMOSFET with $V_{gate} = 1.0V$ and $V_{drain} = 1.5V$ is shown. Similar simulations have been used for a comparison of MC and Lucky Electron models [8]. Due to the low drain voltage impact ionization events are very seldom and the necessary low noise to signal ratio over ten orders of magnitude can be obtained only with phase space MR. In Fig. 3 the currents of electrons and holes with certain energies are shown which hit the Si/SiO_2-interface in the 0.16μm-NMOSFET for $V_{gate} = 1.0V$ and $V_{drain} = 2.7V$. These simulations are not only important for the calculation of gate current but also for the modeling of device degradation. Again simulations performed for the same cpu-time without MR fail to achieve the necessary low noise level by orders of magnitude. Without MR no electrons with an energy of $3eV$ hit the interface, making it impossible to calculate oxide injection. In Fig. 4 the corresponding energy distribution functions integrated over the whole simulation domain for electrons and holes are shown. Furthermore the excellent agreement between simulations with and without MR for common events (Figs. 1, 3 and 4) clearly demonstrates the correctness of the MR technique (for a mathematical proof see [4]).

[1] R. Thoma et al., *J. Appl. Phys.* **69**, p. 2300, 1991.
[2] S. E. Laux et al., *IBM J. RES. DEVELOP.* **34**, p. 466, 1990
[3] K. Shigeta et al., in *Simulation of Semiconductor Devices and Processes* **6**, Springer-Verlag, Wien, p. 384, 1995.
[4] Chr. Jungemann, *thesis*, RWTH Aachen, 1995.
[5] H.-J. Peifer, *thesis*, RWTH Aachen, 1992.
[6] Chr. Jungemann et al., in *Simulation of Semiconductor Devices and Processes* **6**, Springer-Verlag, Wien, 1995.
[7] I. Bork et al., *NUPAD Tech. Dig.* **5**, p. 63, 1994.
[8] Chr. Jungemann et al., *ICVC Tech. Dig.* **4**, p. 211, 1995.

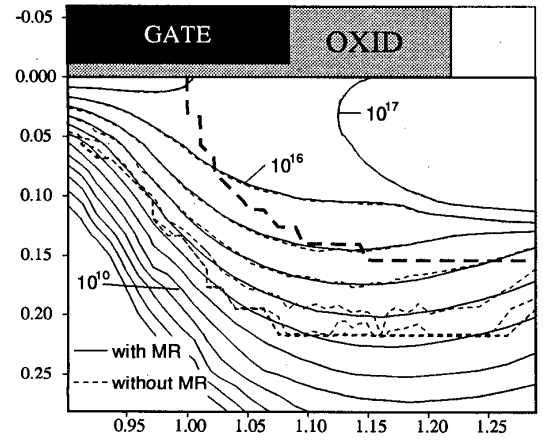

Fig. 1: Electron density (cm^{-3}) in the vicinity of the drain junction of a 0.5μm-LDD-NMOSFET with $V_{gate} = 1.5V$ and $V_{drain} = 5.0V$.

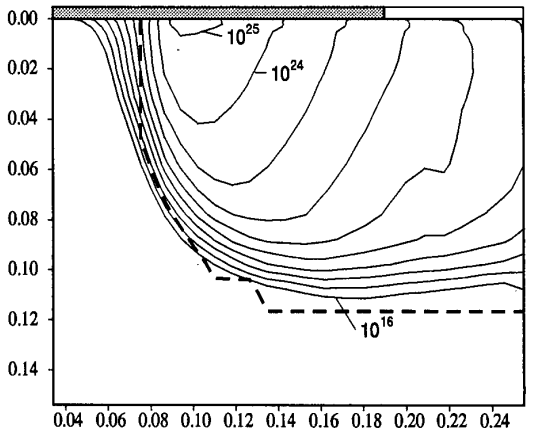

Fig. 2: Electron impact ionization generation rate distribution ($1/scm^3$) for a 0.16μm-NMOSFET with $V_{gate} = 1.0V$ and $V_{drain} = 1.5V$.

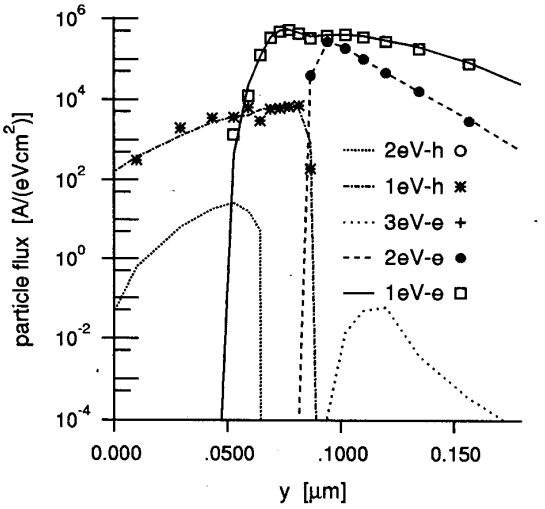

Fig. 3: Currents of electron and holes with certain energies hitting the Si/SiO_2-interface of a 0.16μm-NMOSFET with $V_{gate} = 1.0V$ and $V_{drain} = 2.7V$ (lines: with MR, symbols without MR).

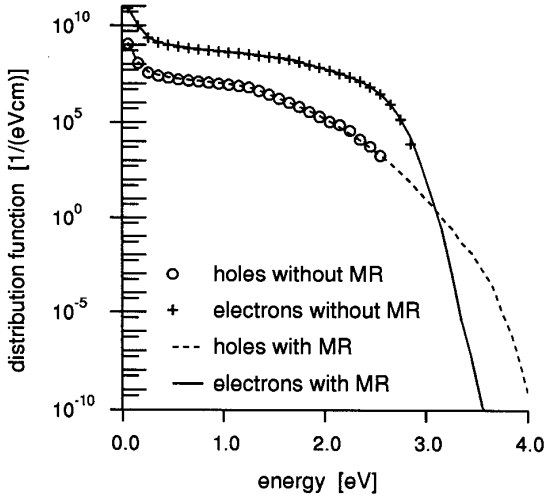

Fig. 4: Electron and hole energy distribution function integrated over the whole simulation domain for a 0.16μm-NMOSFET with $V_{gate} = 1.0V$ and $V_{drain} = 2.7V$.