Full-Band Monte Carlo Investigation of Hot Carrier Trends in the Scaling of Metal-Oxide-Semiconductor Field-Effect Transistors

Amanda Duncan, Umberto Ravaioli, *Senior Member IEEE,* and Jürgen Jakumeit

***Abstract*—A full-band Monte Carlo (MC) device simulator has been used to study the effects of device scaling on hot electrons in different types of n-channel metal-oxide-semiconductor field-effect transistor (MOSFET) structures. Simulated devices include a conventional MOSFET with a single source/drain implant, a lightly-doped drain (LDD) MOSFET, a silicon-on-insulator (SOI) MOSFET, and a MOSFET built on an epitaxial layer on top of a heavily-doped ground plane. Different scaling techniques have been applied to the devices, to analyze the effects on the electric field and on the energy distributions of the electrons, as well as on drain, substrate, and gate currents. The results provide a physical basis for understanding the overall behavior of impact ionization and gate oxide injection and how they relate to scaling. The observed nonlocality of transport phenomena and the nontrivial relationship between electric fields and transport parameters indicate that simpler models cannot adequately predict hot carrier behavior at the channel lengths studied (sub-0.3-μm). In addition, our results suggest that below 0.15 μm, the established device configurations (e.g., LDD) that are successful at suppressing the hot carrier population for longer channel lengths, become less useful and their cost-effectiveness for future circuit applications needs to be reevaluated.**

I. Introduction

THE metal-oxide-semiconductor field-effect transistor (MOSFET) has undergone many changes in the last decade in response to the constant demand for increased speed, decreased power, and increased packing density. Channel lengths and supply voltages have been decreased to meet these requirements. Although the applied bias voltages have dropped, small channel devices are generally characterized by the presence of high electric fields and high energy (or “hot”) electrons that create reliability problems for the transistor. Carriers may acquire enough energy to be injected into the gate oxide, and damage of the insulator may follow [1]–[3]. Hot electrons also undergo impact ionization, leading to the presence of a substrate current which can trigger parasitic bipolar action [4].

While there has been a high degree of success in scaling devices down into the submicron regime, using somewhat empirical guidelines, it is not well understood what the actual limits of MOS technology are. One of the key issues is to control the effects of hot carriers, by limiting bias voltages and adopting structural changes, in a way that yields satisfactory transistor performance and lifetime. Drain biases for standard MOSFET’s are expected to be lowered to about 1.8–1.5 V, and to 1.0 V for low-power portable applications, as the gate lengths approach 0.1 μm or go even below. In this deep submicron range, it is difficult to formulate simple scaling rules by drawing from intuition and prior experience, since the details of the carrier transport are dominated by nonlinear effects.

Until now, it has been possible to routinely conduct accelerated-stress testing, from which information is gathered to improve the process flow. However, the cost of implementation of new technologies is rising very sharply, and it is more and more desirable to rely on computational tools to assess the performance of new structures before fabrication is attempted. Device simulation therefore can be used to reduce the number of design iterations that are needed to achieve acceptable transistor performance and reliability.

The purpose of this work is to study the effects of different scaling techniques on hot carrier behavior in various kinds of n-channel MOSFET structures. In order to resolve hot carrier effects, one needs to use a simulation approach that can give the energy distribution of carriers and can account for the details of the semiconductor band structure at high energies. Monte Carlo (MC) methods solve statistically the Boltzmann transport equation and provide directly a sample of the carrier distribution function by simulating the history of a large number of particles [5], [6]. While MC approaches have the drawback of requiring large computational resources, it is relatively straightforward to include the bandstructure by using numerical tables of arbitrary accuracy [7], [8].

The reader should keep in mind that it is not our intention to directly compare the several device types we consider here. Many different aspects, besides the physical features we examine, must be considered when specific structures or technologies are adopted for application in extreme device scaling. The several reference devices we start from, to formulate our scaling examples, are chosen to be reasonable examples on

Manuscript received June 9, 1997; revised September 22, 1997. The review of this paper was arranged by Editor A. H. Marshak. This work was supported by the Joint Services Electronics Program under Grant N00014-90-J-1270, by the Semiconductor Research Corporation under Contract 96-SJ-406, and by fellowships from the U.S. Department of Energy Computational Science Graduate Fellowship Program and the Intel Foundation.

A. Duncan was with the University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA. She is now with Intel Corporation, Hillsboro, OR 97124-6497 USA.

U. Ravaioli is with the Coordinated Science Laboratory and with the Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA.

J. Jakumeit is with II Physikalisches Institut, Universität zu Köln, D-50937 Köln, Germany.

Publisher Item Identifier S 0018-9383(98)02306-5.

their own, but do not necessarily represent direct competitors for the same application, and they are certainly not optimized for such a purpose. A comparison in this terms would end up being largely unfair. A number of choices and tradeoffs had to be made in the overall physical model and the limitations of the assumptions should always be kept in mind for a correct interpretation of the results. Also, we focus mainly on interesting hot carrier features, while other important aspects like transconductance or threshold control, to name a few, must be carefully considered for technology assessment and design.

II. Monte Carlo Model

Since our goal is to examine a fairly large parameter space in the attempt to achieve a general understanding of scaling trends in the deep submicron regime, a tradeoff is necessary between physical model detail and computational efficiency. The numerical bandstructure included in the MC simulator used here is limited to the first two branches of the conduction band, since they cover most of the energy range of practical interest. The bandstructure is calculated using the empirical pseudopotential model of Cohen and Bergstresser [9], with form factors taken from [10]. The scattering model is limited to energy dependent scattering rates. A more exact scattering model should include anisotropy of the scattering rates and of the deformation potentials on which they are based, over the equal energy shells of momentum space. However, such an approach would require enormous tables for transition rates between individual momentum states. To-date it has found application only in very detailed bulk transport calculations [11], [12], and it is not deemed necessary nor practical for device simulation. The scattering mechanisms considered here include intervalley acoustic phonon scattering, $f-$ and g-type X-X intervalley phonon scattering, X-L intervalley phonon scattering, ionized impurity scattering, and impact ionization. We use Ridley's statistical screening for the ionized impurity scattering calculation [13], [14], and Kane's model for impact ionization [15]. The total scattering rate is adjusted at high energies so that it follows the total density of states [1]. Room temperature is assumed for all simulations.

Self-consistency is obtained by solving the Poisson equation to update the electron forces during the simulation. We use a finite differences discretization approach over a nonuniform rectangular mesh. Since only electrons are simulated as particles, for simplicity holes in the substrate are assumed to be in quasi-equilibrium, and are accounted for by a nonlinear term in the Poisson equation, requiring the use of a Newton-iteration method for the solution. The long-range electron-electron interaction (electron-plasmon scattering) is included implicitly by solving the Poisson equation at closely spaced time intervals (typically, 1 fs or less). This is not a problem, because our MC implementation is based on an ensemble constant-time technique [16] which optimizes the evaluation of particle flight trajectories by using a very short time step. We do not include a separate scattering rate for short-range interaction (electron-electron scattering) since the procedure is very computationally intensive and would lead to prohibitive execution times in our attempt to examine a wide range of devices and bias conditions. In addition, it is quite difficult to include exactly the effect with a separate electron-electron scattering rate in a rigorously physical way for a highly nonuniform structure solved self-consistently, without risking to double count part of the coulombian effect when Poisson equation is solved at the same time. A molecular dynamics approach [5] for evaluation of the forces is ideally a more natural way to include the entire coulombian interaction, but the overall computational cost involved remains too challenging for device applications. Some controversy still exists on the actual importance of the effect on channel transport [17]–[19]. The lack of a component of the overall electron-electron interaction will certainly have an influence on the high energy tail of the carrier distribution at energies at and above the level corresponding to the voltage bias. The estimation of gate currents depend on the detailed knowledge of such tails. Therefore, gate current results should be taken with caution and with full understanding of the underlying assumption that in our calculations the electron-electron scattering is accounted for to the extent that coulombian interaction is resolved by the very fine mesh that we adopt for Poisson equation, with the electron density assigned on the mesh points by a two-dimensional (2-D) cloud-in-cell scheme [5].

In the channel, we still adopt a fully classical model without explicit introduction of quantization effects. We do not expect quantum corrections to be significant for the present investigation focusing on hot carrier effects, with biases well above threshold.

A variance reduction technique is used to improve the statistics of the high energy tails [20] for the simulated particles. Each MC particle is assigned a weight which corresponds to the number of electrons that the particle represents, and the weights of the particles are adjusted to sample underrepresented regions of phase space. The use of variance reduction techniques extends the resolution of the simulated particle energy distribution by several orders of magnitudes, virtually eliminating the statistical noise in most of the practically relevant energy range.

The substrate currents and the current injected into the oxide are good indicators of reliability. The estimation of substrate current I_{sub} is accomplished by using

$$I_{\text{sub}} = e \cdot \int \int_0^\infty f\, d(E,x) W_{\text{ii}}(E)\, dE\, dx \tag{1}$$

where e is the electron charge, x is the channel coordinate, $fd(E,x)$ is the energy distribution of electrons integrated over the channel cross section, and $W_{\text{ii}}(E)$ is the impact ionization rate. Evaluation done by counting the number of actual impact ionization events in the MC simulation leads to similar results but fairly long simulation times are necessary to smooth out the noise.

An estimation of the gate current I_{inj} can be accomplished by using

$$I_{\text{inj}} = e \cdot \iint_0^\infty \frac{fd(E,x)}{n(x)} P_{\text{trans}}(E,x) \cdot P_{\text{hit}}(x)\, dE\, dx \tag{2}$$

where $n(x)$ is the carrier density, $P_{\text{trans}}(E, x)$ is the probability to cross the oxide interface, and $P_{\text{hit}}(x)$ is the current density of electrons hitting the oxide interface, which is obtained from the MC simulation. The term $P_{\text{trans}}(E, x)$ includes tunneling, image lowering of the barrier, and a probability of being backscattered due to collisions in the oxide.

This gate current estimator is simplified, in the sense that it uses an energy distribution of probability. It is also formulated for an ideal interface, using a standard approach of quantum mechanics, where the transmission coefficients can be evaluated with a classical WKB scheme or a more precise Airy function expansion approach. Similar estimation of gate currents can be obtained by counting the particles crossing the oxide interface, but extremely long simulation times are required to collect the statistics, even when the variance reduction is employed, and the same limitations of the physical model would apply.

It is very difficult to capture the physics of what happens at the interface. There are process dependent parameters that should be considered (namely, the roughness of the surface) and the actual oxide bandstructure, which is only known approximately and in an average sense. In addition, transport of injected carriers in the oxide should be carefully considered. Ultimately, a detailed atomic level model would be necessary to really produce a complete theoretical understanding of the oxide injection of carriers. In more practical terms, careful measurements for the behavior of a specific technology are necessary to achieve a quantitative calibration of $P_{\text{trans}}(E, x)$ in the equation above. These considerations on the physical model are in addition to the fact that there may be some uncertainty in the actual shape of the high energy tail of the distribution function when the electron-electron interaction is not completely accounted for. Therefore, one should take the results presented here for their trends, rather than as absolute data, in the absence of accurate experimental verification. The scaling of gate currents should be reasonably correct, but we expect that the current levels predicted by the quantum mechanical model will always overestimate the injection levels from silicon to oxide because of the idealized nature of the model.

The simulation software was designed to run with reasonable efficiency on average computer workstations. All the simulations reported here were executed on Hewlett-Packard 735 and C-100 workstations. A typical simulation, with 10 000 particles and a nonuniform grid with 90×66 nodes for the Poisson equation, requires approximately 30 MB of RAM. On the Hewlett-Packard 735 workstation, about 2500 self-consistent iterations per hour are executed, where one iteration normally corresponds to a time step of 1 fs or less.

III. Summary of Simulated Structures

In this work we have examined four basic MOSFET structures. The first device is the simple "conventional MOSFET" which is schematically shown in Fig. 1(a). This structure is useful for more than a reference, since an interest in the simplest form of the MOSFET has been revived by the need to reduce as much as possible the device footprint in extreme scaling applications. The rationale behind this is that with the reduction of biases it might be unnecessary to use specialized geometries that are designed to control hot electron generation, while involving more fabrication steps and larger contact areas. One should keep in mind, however, that threshold control is a serious problem to be resolved for ultra-small structures of this kind. We assume that an arsenic implant is used to create the source and drain regions. The doping profiles for the source and drain are modeled with hemi-Gaussian profiles, where the doping peaks at the Si/SiO$_2$ interface and decreases with distance into the substrate. Lateral diffusion of the source and drain regions (also modeled with hemi-Gaussian profiles, in the horizontal direction) provides a 0.05-μm gate-to-source and gate-to-drain overlap for all channel lengths. The doping profile in the channel is formed by two boron implants in addition to the background doping. A punchthrough implant is modeled by a full-Gaussian profile in the vertical direction, with its peak at the source and drain junction depth. Finally, a threshold-adjust implant is implemented by another vertical hemi-Gaussian profile with its peak at the Si/SiO2_2 interface. The junction depths for the contact diffusions are: 0.1-μm (0.3-μm gate); 0.05-μm (0.15-μm gate); 0.025-μm (0.075-μm gate).

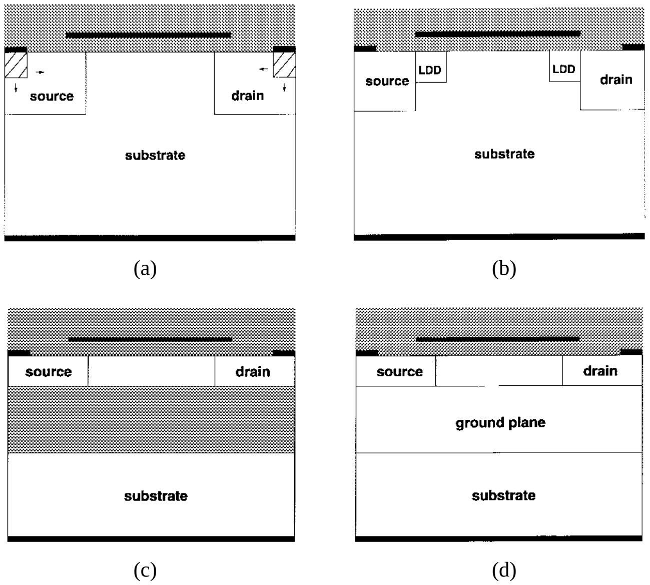


Fig. 1. Schematic diagram of simulated devices: (a) conventional MOSFET, (b) LDD MOSFET, (c) SOI MOSFET, and (d) EPI MOSFET. Black regions correspond to contacts, and grey areas represent oxide.

For simplicity, complete ionization of dopants is assumed, although deep in the contacts we limit the effective ionized doping at a maximum value of approximately 3×10^{19} cm^{-3}, with the intent to reduce the number of simulated particles. Beyond the boundary where this value is attained by the doping profiles, the regions are treated as ideal injecting contacts. Careful tests have been conducted in all cases, to ensure that no appreciable voltage drop is neglected in the contact regions.

The second device that we analyze is the lightly-doped drain (LDD) MOSFET, shown in Fig. 1(b). This type of structure has been very important for control of hot carriers, by reducing the fields at the drain-channel junction with the addition of a transition region of lower doping than in the contact. The lengthened drain contact provides a softer acceleration profile in the high field region and more room for energy relaxation by scattering, effectively reducing the high energy electron tail. This improvement comes at the cost of a somewhat reduced drain current drive capability, larger device area and additional processing steps. In each of the LDD devices studied here, the gate completely overlaps the LDD regions, reducing the possibility of external channel pinchoff. In some cases, the gate also overlaps part of the heavily-doped source and drain regions so that the total overlap between gate and source and gate and drain was 0.05 μm for all devices. The same contact junction depths as above are used for the highly doped regions. The additional LDD regions have junction depth 0.05-μm (0.3-μm gate); 0.025-μm (0.15-μm gate); 0.0125-μm (0.075-μm gate).

The third device that we consider is a fully depleted silicon-on-insulator (SOI) MOSFET shown in Fig. 1(c). The MOS structure is formed on a very thin epitaxial layer of silicon which is grown on top of a thick layer of substrate oxide. In the simulations we assume that the source and junction depths are equal to the width of the top layer of silicon. Note that in the results below, values of a *substrate* current are shown also for the SOI device. This current is simply the result of applying (1) to the carrier distributions obtained from the MC simulation, as an indicator of overall impact ionization processes. Obviously, holes would not be collected by the substrate, because of the thick substrate oxide barrier. Since holes would mainly be collected by the source or recombine, they actually cause a slight change in the drain current. This effect does not affect significantly the results, since the drain current is much larger. For uniformity, we decided to plot the carrier generation by impact ionization for SOI devices together with the substrate current results for the remaining structures. One should also keep in mind that this is only one of the many possible SOI structures. The results presented here should not be interpreted as a general evaluation of SOI technology performance.

Finally, we examine the epitaxial (EPI) MOSFET [21], shown in Fig. 1(d). The structure consists of an almost intrinsic epitaxial layer of silicon in which the source and drain are formed. A heavily p-doped "ground plane" is added, which acts as a punchthrough stopper and controls the threshold voltage. The source and drain junction depths are equal to the thickness of the epitaxial layer. The same contact junction depths are used for the SOI and EPI examples 0.02- μm (0.3-μm gate); 0.014-μm (0.15-μm gate); 0.01-μm (0.075-μm gate).

IV. Scaling Criteria

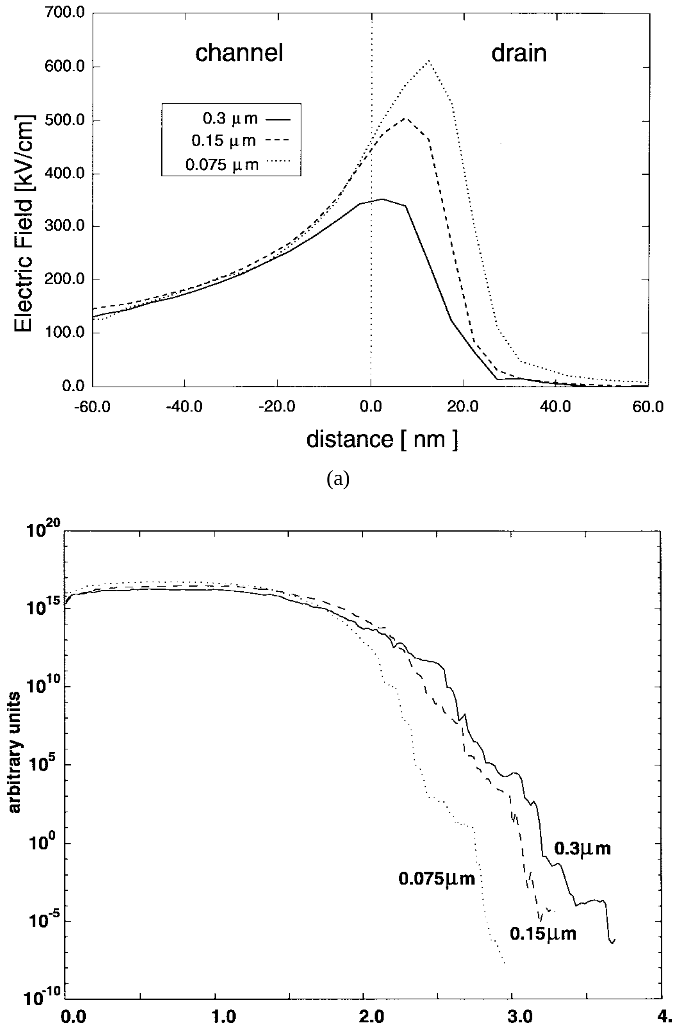
In this work, particular versions of the well-known "constant bias scaling" and "quasi-constant bias scaling" techniques have been applied to the four n-channel MOSFET structures introduced above. In both techniques, starting from a reference structure, the channel length is scaled by a factor $1/\lambda$. For the oxide thickness we choose to scale as $1/\sqrt{\lambda}$, which is a fairly conservative approach. Here, λ is simply a chosen parameter, larger than 1. The junction depth is scaled by $1/\lambda$ but when the structure contains a thin epitaxial layer of silicon, the thicknesses of the epitaxial layer and the junction depth are both scaled by $1/\sqrt{\lambda}$. The junction depth is conventionally defined as the distance from the top interface at which the donor doping is equal to the background acceptor doping. Unless otherwise noted, the length of the source and drain regions and the overlap between the gate and the source and between the gate and the drain are kept constant at 0.2 and 0.05 μm, respectively. The background doping and the peak values of the doping profiles in the region between the source and drain are increased by a factor of λ. The peak value of the doping used to evaluate the profiles in the source and in the drain regions is kept constant at 2×10^{20} cm^{-3}, but the extent of the lateral diffusion under the gate is modified so that the donor doping at the end of the overlap region is always equal to the acceptor concentration at the interface in the channel region. Since the acceptor concentration is raised with scaling, the donor doping in the overlap region is increased.

In constant bias scaling, the applied biases are kept constant as the device is scaled. This approach is used in practice only over limited ranges of channel lengths, for integrated families that share the same type of bias sources. When applied in general, constant bias scaling provides a useful "worst case" scenario for the understanding of the limits of high energy transport. In quasi-constant bias scaling, the biases are scaled by a factor $1/\sqrt{\lambda}$.

Of course, in practical device design there are no fixed scaling rules and many different variations are implemented routinely in new families of integrated devices. In particular, as better quality oxides can be grown at lower cost, many of the newer commercial devices are built with much thinner gate oxides than it was ever thought possible just a few years ago. In our simulation work we have adhered to the well defined scaling rules outlined above in order to arrive at systematic analysis of the trends. However, one should keep in mind that the simulation approach is general and any scaling strategy could be implemented within the same framework.

V. Simulation Results

In the MOSFET simulations illustrated below, we use as reference the following bias values: source bias $V_s = 0$ V; backgate bias $V_b = 0$ V; drain bias $V_d = 3$ V; gate bias $V_g = 3$ V. This condition is applied to all simulations in the case of constant bias scaling. For quasi-constant bias scaling, the above bias condition is applied to the devices with channel length 0.3 μm as a starting point, and when the channel length is scaled down by a factor $1/\lambda$, bias voltages are scaled by $1/\sqrt{\lambda}$. For all the MOSFET structures examined here, we consider three gate lengths: 0.3, 0.15, and 0.075 μm, corresponding to two consecutive scaling steps with $\lambda = 2$. For the three devices, starting from the 0.3-μm channel, we have oxide thicknesses of 9.9, 7.0, and 5.0 nm. While we note, again, that this is a conservative choice, more aggressive scaling could be readily used by the simulator. For all cases,



(a)

(b)

Fig. 2. (a) Distribution of the electric field parallel to the interface in the region close to the channel/drain boundary and (b) electron distributions in energy at the drain/channel boundary for the conventional MOSFET with constant bias scaling. The solid line is for a channel length of 0.3 μm, the dashed line for a channel length of 0.15 μm, and the dotted line for a channel length of 0.075 μm.

we verified that no significant punchthrough takes place in the structures.

In general, the electric field peaks just inside the lateral diffusion region of the drain. The average energy peak is located past the peak of the electric field, indicating that nonlocal effects take place at small channel lengths. With quasi-constant bias scaling, the peak electric field still increases with decreasing channel length while there is a decrease of the peak average energy (and associated energy distributions).

Simulations under constant-bias scaling provide a useful limit case to understand transport behavior, although for practical applications progressively lower bias voltages would normally be used as gate lengths are shrunk. Fig. 2(a) shows the distibution of the field parallel to the interface close to the channel/drain boundary, which is indicated as zero reference in the plots, for the three conventional MOSFET structures under constant-bias scaling. Fig. 2(b) shows the energy distributions of electrons at the channel-drain boundary, taking into account all the particles in the entire cross section. At low energies, where one can find most of the electrons, the curve for the shortest device is noticeably higher, because of the greater density of electrons in the channel. Interestingly, the high energy tails of the distributions do not follow the increase of the electric field with decreasing of size at this location, the shortest device having the smallest energy tail.

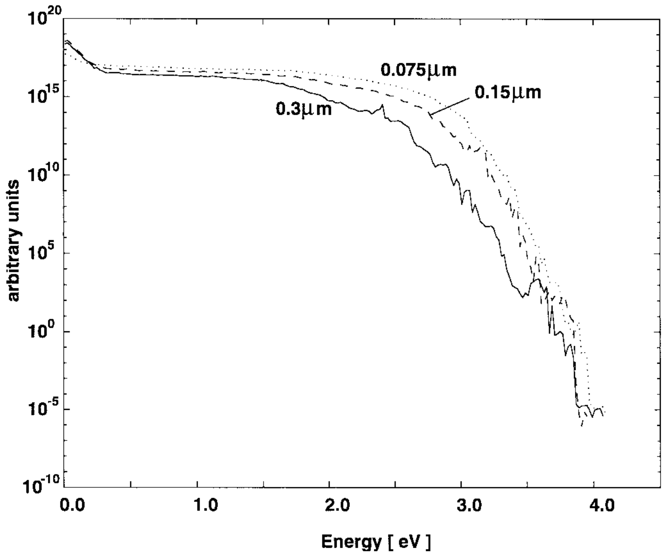


Fig. 3. Electron distributions in energy at the locations with largest high energy tails for the conventional MOSFET with constant bias scaling. The solid line is for a channel length of 0.3 μm, the dashed line for a channel length of 0.15 μm, and the dotted line for a channel length of 0.075 μm.

The widest tails for the energy distributions are found inside the drain, past the point of maximum field and the point of maximum average energy. The energy distribution of the 0.075-μm device attains its widest tail about 27.5 nm past the channel-drain boundary, the 0.15-μm device at 22.5 nm, and the 0.3-μm device at 17.5 nm. The energy distributions at the locations with the widest high energy tails for the three devices are plotted in Fig. 3. The smallest device has the largest number of high energy electrons, as one would have expected, but they appear deeper inside the drain. The voltage drop in the drain is the reason for this behavior. As the MOSFET is scaled, the channel resistance drops while the resistances of source and drain remain approximately constant. A large portion of the potential drop occurs across the region of lateral diffusion in the drain, which extends from the end of the channel to the end of the gate-to-drain overlap. The potential drop across the drain is approximately 0.36 V for the 0.3-μm device, and increases to 0.7 V in the 0.15-μm device, and 1.1 V in the 0.075-μm device. The larger the potential drop across the drain, the smaller the potential at the channel-drain boundary for the same drain bias, and the less energy the electrons can gain from the electric field when traveling from the source to the end of the channel.

It is interesting to look at the drain current as well as the substrate current and total gate injection current to obtain an idea of the tradeoffs between reliability and performance associated with each scaling technique. Fig. 4 shows the drain current (I_d) as a function of channel length with constant

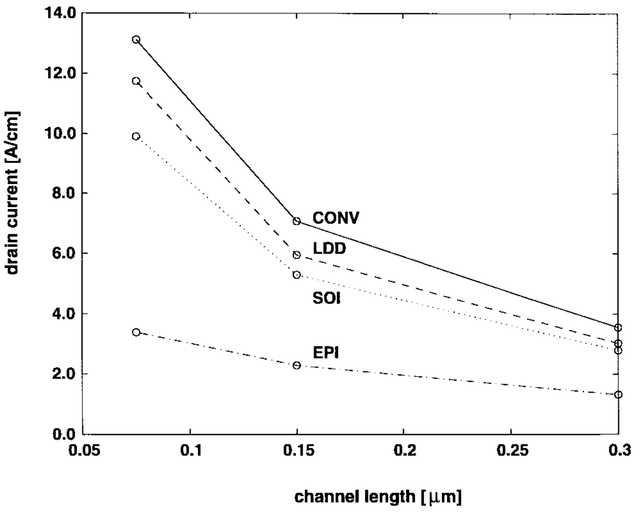


Fig. 4. Drain current as a function of channel length with constant bias scaling for the conventional (solid line), LDD (dashed line), SOI (dotted line), and EPI (dashed-dotted) MOSFET's.

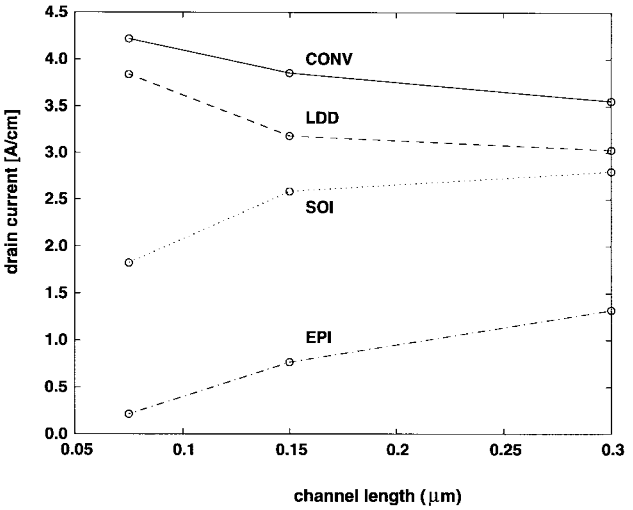


Fig. 5. Drain current as a function of channel length with quasi-constant bias scaling for the conventional (solid line), LDD (dashed line), SOI (dotted line), and EPI (dashed-dotted) MOSFET's.

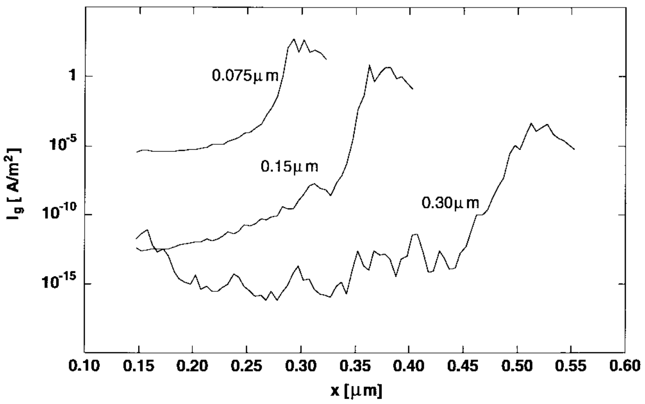


Fig. 6. Electron injection current density in the conventional MOSFET with constant bias scaling. The solid line is for a channel length of 0.3 μm, the dashed line for a channel length of 0.15 μm, and the dotted line for a channel length of 0.075 μm.

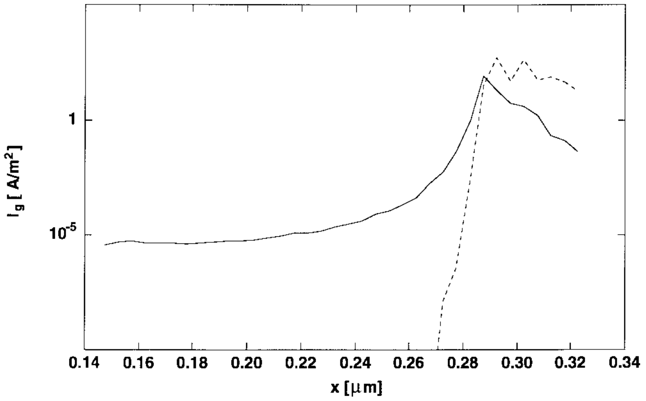


Fig. 7. Electron injection current density for the 0.075-μm conventional MOSFET due to tunneling (solid line) and thermionic emission (dashed line) for $V_g = V_d = 3$ V.

bias scaling applied to the four types of MOSFET's discussed above. The price in performance (in addition to the price in area) for adding the LDD regions to the conventional device can be appreciated. Although I_d increases rapidly for both the conventional and the LDD devices, the LDD device always has a lower drain current than the conventional MOSFET with the same channel length. The drain current increases more slowly with constant bias scaling in the EPI and SOI devices than in the conventional and LDD MOSFET's, and the drain current actually decreases with quasi-constant bias scaling, as shown in Fig. 5.

As the drain bias is scaled, the only way to increase I_d is to decrease the effective channel resistance. In quasi-constant bias scaling, there are several parameters that change, which affect the channel resistance (including channel length, gate voltage, threshold voltage and oxide capacitance), and indeed the general effect is that the resistance is decreased. The drain and source resistances remain nearly the same in the conventional device, decrease slightly in the LDD and SOI devices, and increase slightly in the EPI MOSFET, with scaling. The SOI and EPI devices have such small junction depths that the source and drain resistances account for a significant fraction of the resistance between the source and drain contacts. The reduction of the channel resistance is not enough to make up for the reduction of V_d when quasi-constant bias scaling is applied, so the drain current decreases. For the conventional and LDD devices, the drain current still rises, although at a much slower rate than when the bias is held constant.

The injection current density into the oxide estimated by the simulator is shown in Fig. 6 as a function of channel coordinate for the three conventional MOSFET's. Again, the numerical values of the currents presented here are the results of the ideal quantum mechanical model, and should be taken as a qualitative reference, since the model is likely to overestimate the rate of electrons that successfully transfer from silicon to the oxide. Although the high energy tail for the smallest device is smaller than the high energy tail for the largest device at all points along the channel, the injection current density is larger for the smallest device across its entire length. Even though the electrons have less energy in the 0.075-μm channel than the electrons in the 0.3-μm channel, the

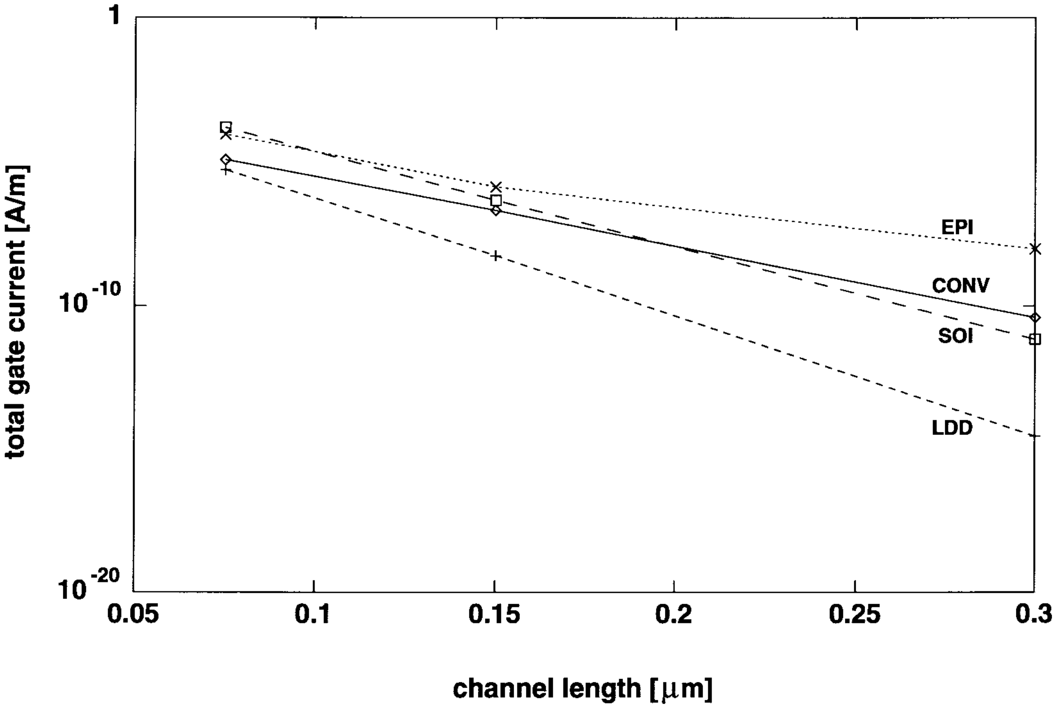


Fig. 8. Gate injection current as a function of channel length with constant bias scaling for the conventional (solid line), LDD (dashed line), SOI (dotted line), and EPI (dashed-dotted) MOSFET's.

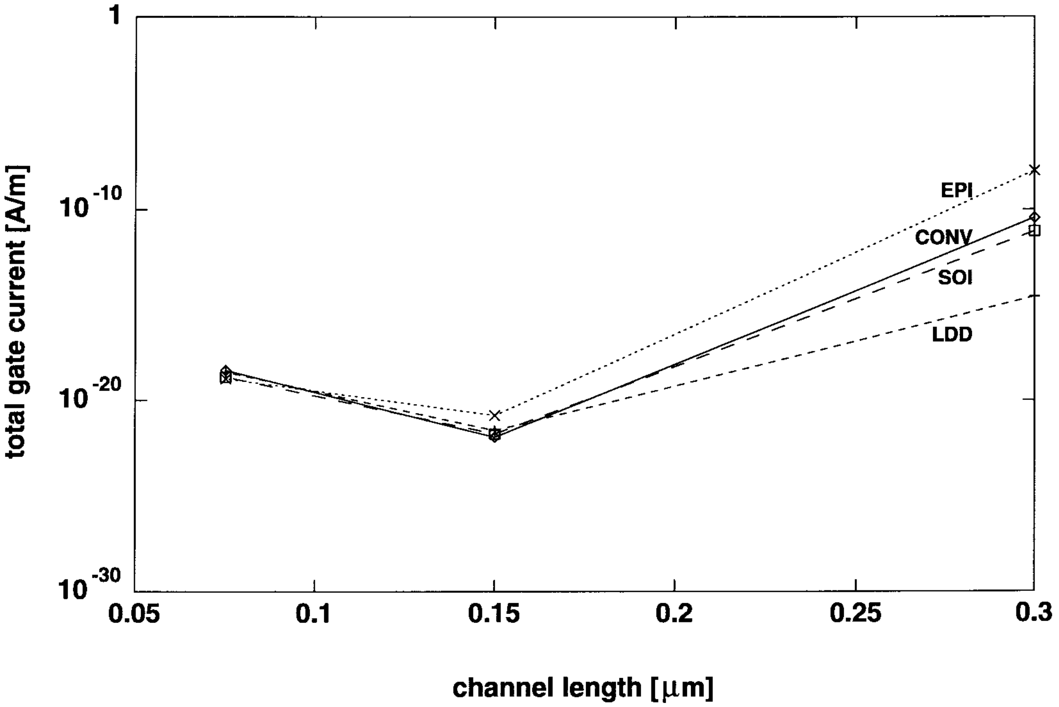


Fig. 9. Gate injection current as a function of channel length with quasi-constant bias scaling for the conventional (solid line), LDD (dashed line), SOI (dotted line), and EPI (dashed-dotted) MOSFET's.

thinner oxide in the smallest device causes its injection current density to be larger. Fig. 7 shows the contribution of tunneling and thermionic emission to the injection current. Thermionic emission dominates at the drain, but tunneling contributes to the injection current all along the channel.

Fig. 8 shows the total current injected into the oxide I_{inj} as a function of channel length with constant bias scaling in the four devices. The injection current is a function of the heating of the electrons as they travel along the channel, the number of electrons that flow down the channel and hit the interface, and the field normal to the interface. In constant bias scaling, these factors cause the injection current to increase in all the devices, reflecting the longitudinal field increase due to channel length reduction. At a channel length of 0.3 μm,

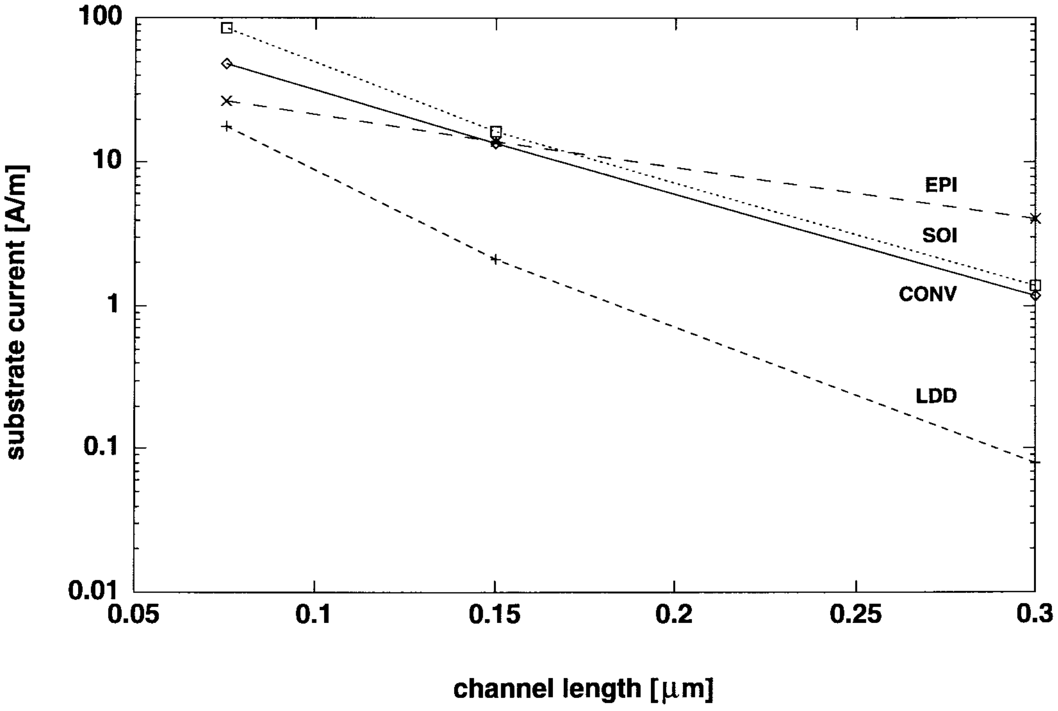


Fig. 10. Substrate current $I_{\rm sub}$ as a function of channel length with constant bias scaling for the conventional (solid line), LDD (dashed line), SOI (dotted line), and EPI (dashed-dotted) MOSFET's. Injection current as a function of channel length with quasi-constant bias scaling for the conventional (solid line), LDD (dashed line), SOI (dotted line), and EPI (dashed-dotted) MOSFET's.

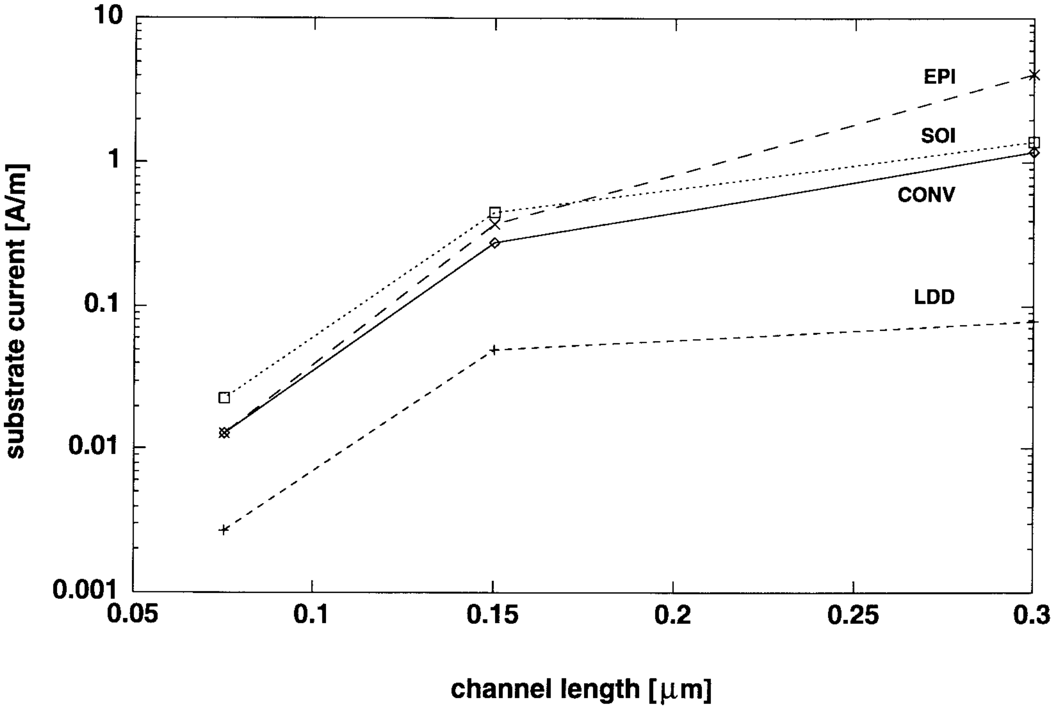


Fig. 11. Substrate current $I_{\rm sub}$ as a function of channel length with quasi-constant bias scaling for the conventional (solid line), LDD (dashed line), SOI (dotted line), and EPI (dashed-dotted) MOSFET's.

the EPI MOSFET has the highest injection current, followed by the conventional, SOI, and LDD devices, in that order. The devices are ranked in the same order when the average energy peak is considered. The injection currents for the four devices at a channel length of 0.075 μm are much closer to one another than at 0. μm.

Fig. 9 shows the trends for the total oxide injection current $I_{\rm inj}$ as a function of channel length, estimated with quasi-constant bias scaling. The reduction of the bias causes the injection current to drop in all cases, when scaling from 0.3 to 0.15 μm. The current appears to increase again when scaling from 0.15 to 0.075 μm, because of an increase in tunneling on the source side of the channel, due to the corresponding scaling of the oxide thickness. This kind of behavior can be expected when gate thickness and voltages are scaled according to certain ratios. The tunneling component of the gate current, that we observe within the limits of our model, is predominant for the shortest devices because voltage scaling reduces the electron high energy tails with less thermionic emission. If the tunneling becomes sufficiently strong with oxide scaling, the injection current goes back up originating the valley in the plot. The details of the device structure do not seem to be very important in this regime, leading to an injection current which is approximately equal for all devices for the thinnest oxides we examined. A more aggressive oxide scaling than used here would cause an even larger tunneling contribution, so that one should expect a reduction, or even the disappearance, of the valley in the current curves. Some advanced commercial devices are based already on a much more aggressive oxide scaling to reduce even further gate bias voltages. Obviously, there must be a finite limit to practical oxide thickness to maintain a reasonable control of the tunneling current, which should be carefully considered in such cases to ensure device reliability. With extremely thin oxides the physical models would require an additional level of sophistication, since the presence of any traps in the oxide would greatly influence tunneling processes.

The substrate current $I_{\rm sub}$ as a function of channel length, obtained with constant bias scaling, is shown in Fig. 10. (Again, for the SOI device, the "substrate current" is defined as the current flow due to the holes generated by impact ionization.) The rate of impact ionization events depends on the heating of the carriers and the rate at which the carriers flow through the channel. Most impact ionization events occur where the high energy tail is the largest, near the interface and inside the drain. At a channel length of 0.3 μm, the different structures simulated show an amount of substrate current which approximately reflects the respective peak average energies for the bias considered. When the devices are scaled, $I_{\rm sub}$ increases for all devices but not as quickly for the EPI device. The same trend is observed when considering drain current, peak electric field and average energy.

Fig. 11 shows the trends for the substrate current $I_{\rm sub}$ when quasi-constant bias scaling is used. The current drops with scaling for all types of devices, but particularly for the EPI device, due to rapid reduction of drain current with channel length. The LDD device still presents the lowest substrate current, but the advantage with respect to the conventional MOSFET structure is much less dramatic at small gate lengths, particularly if the drain current penalty is also factored in. When scaling even further, one has to expect that the doping of the LDD region would have to be quite high, to the point where there would be no advantage at all with respect to the conventional structure. The additional processing steps and the larger area required by LDD devices should also be considered when making a global comparison between the various structures. We explain the large comparative values of the estimates for gate and substrate currents despite lower drain currents in the EPI structures, by observing that the simulations indicate fairly high average energies for this case.

VI. Conclusions

A full-band, ensemble MC simulator has been used to study the trends of hot carrier behavior with scaling, examining several MOSFET structures. Simulations with the constant bias and quasi-constant bias scaling rules provide a clear illustration of the role of hot electron effects when the gate length is reduced into the deep submicron regime. Substrate and gate currents have been estimated, based on the carrier distribution in phase space determined by the MC simulations, showing that as the gate length decreases, there is less advantage is using an LDD MOSFET structure with respect to the simple conventional MOSFET for hot carrier control.

References

[1] J. Y. Tang and K. Hess, "Theory of hot electron emission from silicon into silicon dioxide," *J. Appl. Phys.*, vol. 54, pp. 5145–5151, 1983.

[2] T. H. Ning and H. N. Yu, "Optically-induced injection of hot electrons into $\rm SiO_2$," *J. Appl. Phys.*, vol. 45, pp. 5373–5378, 1974.

[3] ——, "Hot-electron emission from silicon into silicon dioxide," *Solid State Electron.*, vol. 21, pp. 273–282, 1978.

[4] H. B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*. Reading, MA: Addison-Wesley, 1990.

[5] R. W. Hockney and J. W. Eastwood, *Computer Simulation Using Particles*. New York: McGraw-Hill, 1981.

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

[7] 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*, vol. 38, p. 9721, 1988.

[8] K. Hess, Ed., *Monte Carlo Device Simulation: Full Band and Beyond*. Norwood, MA: Kluwer, 1991.

[9] M. L. Cohen and T. K. Bergstresser, "Band structures and pseudopotential form factors for fourteen semiconductors of the diamond and zinc-blende structures," *Phys. Rev.*, vol. 141, pp. 789–796, 1996.

[10] J. R. Chelikowsky and M. L. Cohen, "Nonlocal pseudopotential calculations for the electronic structure of eleven diamond and zinc-blende semiconductors," *Phys. Rev.*, vol. 14, pp. 556–582, 1976.

[11] P. D. Yoder, V. D. Natoli, and R. M. Martin, "*Ab Initio* analysis of the electron-phonon interaction in silicon," *J. Appl. Phys.*, vol. 73, pp. 4378–4383, 1993.

[12] T. Kunikiyo, M. Takenaka, Y. Kamakura, M, Yamaji, H. Mizuno, M. Morifuji, K. Taniguchi, and C. Hamaguchi, "A Monte Carlo simulation of anisotropic electron transport in silicon including full band structure and anisotropic impact ionization model," *J. Appl. Phys.*, vol. 75, p. 297–312, 1994.

[13] B. K. Ridley, "Reconciliation of the Conwell-Weisskopf and Brooks-Herring formulae for charged-impurity scattering in semiconductors: Third-body interference," *J. Phys. C*, vol. 10, pp. 1589–1593, 1977.

[14] T. G. Van de Roer and F. P. Widdershoven, "Ionized impurity scattering in Monte Carlo calculations," *J. Appl. Phys.*, vol. 59, pp. 813–815, 1986.

[15] E. O. Kane, "Electron scattering by pair production in silicon," *Phys. Rev.*, vol. 159, no. 3, pp. 624–631, 1967.

[16] U. Ravaioli, "Vectorization of Monte Carlo algorithms for semiconductor simulation," in *Monte Carlo Device Simulation: Full Band and Beyond*, K. Hess, Ed. Norwood, MA: Kluwer, 1991, ch. 9.

[17] J. Bude and M. Mastrapasqua, "Impact ionization and distribution functions in submicron NMOSFET technologies," *IEEE Electron Device Lett.*, vol. 16, pp. 439–441, Oct. 1995.

[18] M. Fischetti, S. Laux, and E. Crabbe, "Understanding hot-electron transport in silicon devices: Is there a shortcut?," *J. Appl. Phys.*, vol. 78, no. 2, pp. 1058–1087, 1995.

[19] R. B. Hulfachor, K. W. Kim, M. A. Littlejohn, and C. M. Osburn, "Comparative analysis of hot electron injection and induced device degradation in scaled 0.1 micrometer SOI n-MOSFET's using Monte Carlo simulation," *IEEE Electron Device Lett.*, vol. 17, pp. 53–55, Feb. 1996.

[20] A. Pacelli and U. Ravaioli, "Analysis of variance reduction schemes for ensemble Monte Carlo simulation of semiconductor devices," *Solid-State Electron.*, vol. 41, pp. 599–605, 1997.

[21] C. Fiegna, H. Iwai, T. Wada, M. Saito, E. Sangiorgi, and B. Riccó, "Scaling the MOS transistor below 0.1 μm: Methodology, device structures, and technology requirements," *IEEE Trans. Electron Devices*, vol. 41, pp. 941–950, June 1994.

Amanda Duncan received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana-Champaign in 1991, 1993, and 1996, respectively. Her Ph.D. thesis topic was on Monte Carlo simulation of silicon devices.

She is currently a Senior Design Engineer with Intel Corporation, Portland Technology Development, Portland, OR. Her interests include circuit design, silicon devices, and TCAD.

Umberto Ravaioli (M'88–SM'93) received the Laurea Dr. degree in electronics engineering and the Laurea Dr. degree in physics from the University of Bologna, Bolgna, Italy, in 1980 and 1982, respectively, and the Ph.D. degree in electrical engineering from Arizona State University, Tempe, in 1986.

Since 1986, he has been with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, where he is now a Professor. His current research interests are in the area of Monte Carlo device simulation, nanoelectronics, and computer-based education.

Jürgen Jakumeit was born in Cologne, Germany, on April 27, 1965. He received the Diploma and Ph.D. degree from the University of Cologne in 1990 and 1994, respectively. His research concerned the theoretical and experimental investigation of quantum size and interference effects in semiconductor structures. During two half-year stays at the University of Michigan, Ann Arbor, and the University of Illinois, Urbana-Champaign, he became engaged in the field of device simulation. Currently, he works as a Post-doctoral Fellow at the University of Cologne on the simulation of silicon MOSFET's and the investigation of quantum interference effects.