SIMULATION OF SEMICONDUCTOR DEVICES AND PROCESSES Vol. 3
Edited by G. Baccarani, M. Rudan - Bologna (Italy) September 26-28, 1988 - Tecnoprint

MONTE CARLO SIMULATION OF SUBMICRON Si MOSFETs

Massimo V. Fischetti and Steven E. Laux

IBM Research Division
T. J. Watson Research Center
P.O. Box 218, Yorktown Heights, New York 10598

SUMMARY

Electronic transport in Si devices is investigated using a Monte Carlo technique which improves the 'state-of-the art' treatment of high-energy carrier dynamics: 1. The semiconductor is modeled beyond the effective-mass approximation using the band structure obtained from empirical-pseudopotential calculations. 2. The carrier-phonon, carrier-impurity, and carrier-carrier scattering rates are computed in a way consistent with the full band-structure of the solid. 3. The long-range carrier-carrier interaction and space-charge effects are included by coupling the Monte Carlo simulation to a self-consistent 2-dimensional Poisson solution updated at a frequency large enough to resolve the plasma oscillations in highly-doped regions. The technique is employed to study experimental submicron Si field-effect-transistors with channel lengths as small as 60 nm operating at 77 and 300 K. Velocity overshoot and highly nonlocal, off-equilibrium phenomena are investigated together with the role of carrier-carrier interaction in these ultra-small structures. In the systems considered, the inclusion of the full band structure can have dramatic effects on electron transport, by reducing the amount of velocity overshoot via transfer to upper conduction valleys, in agreement with available experimental data.

INTRODUCTION

In recent years the Monte Carlo (MC) technique to treat electronic transport in semiconductor devices has been gaining increasing popularity (Hockney and Eastwood, 1981; Moglestue, 1986; Hesto *et al.*, 1985). Despite its extensive computational requirements, the main reasons for this success can be ascribed to 1. its simplicity of implementation, 2. the complete physical picture used to describe electronic transport, which allows the treatment of the non-local, off-equilibrium, and hot-carrier effects which can occur in small devices, and, 3. the possibility of investigating the behavior of the devices at the level of a 'computer microscope'. Sophisticated MC device-simulators have been developed which include quantum effects in inversion layers (Chu-Hao *et al.*, 1985; Wang and Hess, 1985; Ravaioli and Ferry, 1986) and quantum wells (Yokoyama and Hess, 1986; Artaki and Hess, 1987), also coupled self-consistently to a 2-dimensional solution of the Poisson equation (Tomizawa and Hashizume, 1988). Still, little attempts have been made to provide an accurate description of the kinematics and dynamics of hot carriers – usually a simple parabolic or first-order $\mathbf{k}\cdot\mathbf{p}$ approximation to the semiconductor band structure is employed in MC device simulators. Hess and co-workers have pointed out in their pioneering work that hot-carrier behavior is dominated by band-structure effects, both directly (*i.e.*, kinematically, through group velocities and energy-wavevector dispersion) and indirectly (*i.e.*, dynamically, through the effect of the density of states, DOS, on the carrier scattering rates) (Shichijo and Hess, 1981; Tang and Hess, 1983a, 1983b). Considering the importance of these effects on reliability and performance of submicron devices, we believe the description of the semiconductor band-structure used in a MC device simulator must be as accurate as possible.

In this paper we describe a 2-dimensional self-consistent MC-Poisson device simulation program we have developed in the attempt to improve the treatment of hot-carrier effects. (Laux and

Fischetti, 1988; Fischetti and Laux, 1988). The MC portion of the program employs the band-structure of silicon obtained from empirical-pseudopotential calculations to describe the carrier kinematics and dynamics. In addition, we have applied the program to the simulation of realistic devices in which the high dopant concentrations make it necessary to look carefully at the Coulomb coupling between carriers both at short range (direct carrier-carrier interaction) and at long range (plasma effects).

THE MONTE CARLO MODEL

The Monte Carlo technique we employ to treat electronic transport is conceptually identical to the 'state-of-the-art' technique described in the excellent reviews by Price (1979) and Jacoboni and Reggiani (1983). From a computational point of view, the inclusion of the full band structure has been already implemented in a MC simulation by the Urbana group in simple cases (Tang and Hess, 1983a, 1983b). However, there are some significant differences between our program and the program developed by Hess and co-workers: 1. A better interpolation accuracy which arises from employing a finer mesh of **k**-points in the first Brillouin Zone (BZ), 2. a different evaluation of the scattering rates, and 3. a different algorithm (and a correspondingly different physics) for the selection of the final particle states after collisions. We now discuss these issues in turn.

Band Structure

The empirical pseudopotentials given by Cohen and Bergstresser (1966) represent a reliable representation of the excitation spectrum of the semiconductors of technological interest. Unlike pseudopotentials designed to describe the total energy as a function of atomic coordinates, they fit experimental transport data and provide an reliable description of the DOS. We show in Fig. 1 the band structure and DOS of Si we have used.

For the purpose of our MC simulation, we have generated a mesh of 916 **k**-points in the irreducible wedge of the BZ, spaced by $0.05(2\pi/a)$, a being the lattice constant. At these points we have computed the energy $E_\nu(\mathbf{k})$, gradients $\partial E_\nu(\mathbf{k})/\partial k_i$, and second derivatives $\partial^2 E_\nu(\mathbf{k})/\partial k_i \partial k_j$, where $i,j = x,y,z$, and the index ν runs over the first five conduction bands and the first three valence bands. These values are then stored in a look-up table. In principle, this is all that is needed to recover information over the entire BZ, thanks to its symmetry. In practice, during a MC run, given a particle with an arbitrary wavevector **k** in band ν, we should first translate **k** into the first BZ, then rotate it into the irreducible wedge, in order to obtain its energy $E_\nu(\mathbf{k})$ and group velocity $\nabla_{\mathbf{k}} E_\nu(\mathbf{k})/\hbar$, $\hbar$ being the reduced Planck's constant. In performing this operation we must also store the symmetry transformation involved in the mapping, so that the inverse transformation can be applied to the group velocity in order to obtain its correct orientation over the entire BZ. These last operations can be bypassed by storing the band structure information over about 41,000 points in

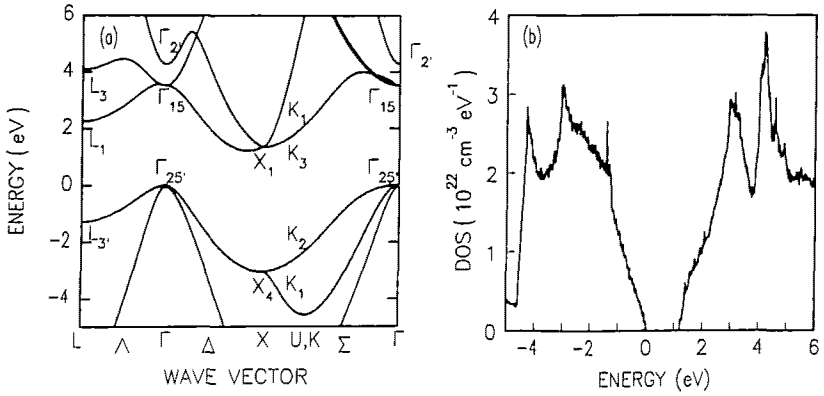

Figure 1. *Band-structure (a) and density of states (b) for Si obtained from the empirical pseudopotential calculation. No spin-orbit interaction is included, but the spin-orbit valence band is shifted by 0.045 eV to match the experimental splitting.*

the entire BZ (and a few points outside for interpolation requirements), thus increasing the speed of the program. Therefore, given a wave-vector $\mathbf{k}$, we find the associated energy in band ν by first finding the 8 corners $\{\mathbf{k}_\lambda\}$ ($\lambda = 1, 8$) of the cubic element of side length $\ell = 0.05\ (2\pi/a)$ in the BZ to which $\mathbf{k}$ belongs, expanding the energy quadratically around each corner, and finally adding up the contributions from each corner with appropriate weights. The velocity at $\mathbf{k}$ is obtained in a similar way by interpolating linearly around each corner.

The much more complicated task of inverting the dispersion $E_\nu(\mathbf{k})$ – as needed after each collision when a wavevectors $\mathbf{k}'$ corresponding to the final energy E' must be found – is handled by searching through the first BZ for the cubes which intersect the equi-energy surface $E_\nu(\mathbf{k}) = E'$ over all bands in which E' might be found. This is done by generating two meshes in the BZ, the first one (which we shall call the *coarse mesh*) using cubes with sides of length 4ℓ, the second one (*fine mesh*) using cubes of sides $\ell/2$ long. For every cube in each mesh we store the maximum and minimum energies spanned. A search is done first over the coarse mesh, thus reducing the volume of $\mathbf{k}$ -space over which the second search (over the fine mesh) must be done. We then perform a search over the chosen subset of the fine mesh and find all cubes in the fine mesh which intersect the desired equi-energy surface. In each of them a particular $\mathbf{k}$ is chosen by inverting the direct energy-interpolation up to third order. This guarantees that the selected k-vectors will correspond to the desired energy E' within an average variance of a few meV.

Carrier free-flight

In MC simulations employing analytic approximations of the bands, a significant amount of CPU-time is saved by using the so-called 'self-scattering' algorithm to determine the duration of a carrier free-flight and compute its final position (Jacoboni and Reggiani, 1983, and references therein). However, when a numerical representation of the bands is used, the higher efficiency of the self-scattering algorithm vanishes since we cannot integrate analytically the equations of motions:

$$(1a) \qquad \frac{d\mathbf{r}}{dt} = \frac{1}{\hbar}\nabla_{\mathbf{k}} E_\nu(\mathbf{k})$$

$$(1b) \qquad \frac{d\mathbf{k}}{dt} = \mp \frac{e}{\hbar}\nabla_{\mathbf{r}}\phi(\mathbf{r}) = \pm \frac{e\mathbf{F}(\mathbf{r})}{\hbar} ,$$

where e is the magnitude of the electron charge, $\phi(\mathbf{r})$ is the electrostatic potential, and $\mathbf{F}$ the electric field at the particle position $\mathbf{r}$, and the upper and lower signs refer to holes and electrons respectively. Furthermore, the simulation of transient phenomena with the inclusion of carrier-carrier interactions requires the presence of a synchronous ensemble. For these reasons, we have decided to use a prefixed time-step, Δt_{bal} , and a second order Runge-Kutta scheme for the numerical integration of Eqns. (1a) and (1b) over a free-flight. Satisfactory numerical stability and accuracy are obtained using a time step of the order of 10^{-16} sec . The magnitude of the scattering rates poses an additional constraint on the highest possible value of Δt_{bal} , as discussed below.

Scattering rates

The approach we have chosen to compute the scattering rates emphasizes the role of the band structure and of the DOS, in the same spirit as Tang and Hess (1983a, 1983b). However, we have extended this approach down to very low carrier energies by using a finer discretization of the BZ. This might be of some importance in Si around the X symmetry-point, where the first and second conduction band behave quite differently from a parabolic-band representation at energies around 130 meV. Therefore, the different kinematics (via the group velocities) and dynamics (via the different DOS) of these regions will be well represented by our approach.

i) Electron-phonon scattering

The nonpolar scattering rate, $1/\tau_{\eta,\nu}(\mathbf{k})$, between a particle of wave-vector $\mathbf{k}$ in the ν-th band and a phonon of type (acoustic or optical) and polarization (transverse or longitudinal) η has been calculated from the Golden Rule expression:

$$(2) \qquad \frac{1}{\tau_{\eta,\nu}(\mathbf{k})} = \sum_{\nu',\mathbf{q}} \frac{\pi}{\rho\omega_{\eta,\mathbf{q}}} \Delta_{\eta,\nu'}(\mathbf{q})^2 \mid \mathcal{I}(\nu,\nu';\mathbf{k},\mathbf{k}') \mid^2 \delta(E_\nu - E'_{\nu'} \mp \hbar\omega_{\eta,\mathbf{q}})(n_{\eta,\mathbf{q}} + \frac{1}{2} \pm \frac{1}{2}).$$

In this equation the upper and lower signs correspond to emission and absorption of a phonon, respectively, while ρ is the density of the semiconductor, $\Delta_{\eta,\nu'}(\mathbf{q})$ is a coupling constant, $\omega_{\eta,\mathbf{q}}$ is the frequency of the phonon of type η and wave-vector $\mathbf{q}$, $\mathbf{k}' = \mathbf{k} \mp \mathbf{q} + \mathbf{G}$ is the final particle wave-vector which is mapped into the first BZ by adding a vector $\mathbf{G}$ of the reciprocal lattice. Also, $\mathcal{I}$ is the overlap integral, $E_\nu = E_\nu(\mathbf{k})$, $E'_{\nu'} = E_{\nu'}(\mathbf{k} \mp \mathbf{q})$, and $n_{\eta,\mathbf{q}}$ is the phonon occupation number at the lattice temperature T . The sum in Eq. (2) extends over all bands ν' and over all phonon wave-vectors $\mathbf{q}$ in the first BZ. This implies that for some of the phonon wavevectors $\mathbf{q}$, a nonzero $\mathbf{G}$ is required to bring $\mathbf{k}'$ into the first BZ. This corresponds to the inclusion of Umklapp processes. The *fine mesh* previously defined is used to discretize the zone. A numerical algorithm proposed by Gilat and Raubenheimer (1966) is used to perform the integration over the energy-conserving delta function. The results are then stored in a look-up table, together with the rate-gradients, and interpolated as done for the energy and velocity.

In the absence of better information about the deformation potentials, the nonpolar carrier-phonon coupling constant has been approximated by the isotropic expression $\Delta_\eta q$ for longitudinal acoustic (LA) and transverse acoustic (TA), or $(\Delta K_{op})_\eta$ for longitudinal optical (LO) and transverse optical (TO) phonons. The overlap integral has been approximated by the rigid-ion expression (Ziman, 1974) in the numerical range, ignoring the band-index dependence. The acoustic phonon dispersion has been approximated by:

$$(3) \qquad \hbar\omega_{\eta,\mathbf{q}} = \begin{cases} \hbar\omega_{\eta,max}\left[1 - \cos\left(\frac{qa}{4}\right)\right]^{1/2} & (q \leq 2\pi/a) \\ \hbar\omega_{\eta,\max} & (q > 2\pi/a) . \end{cases}$$

Here $\omega_{\eta,max} = 4c_\eta/a$, where c_η is the sound velocity with polarization η. This expression underestimates the phonon energy at small q, $\hbar\omega_{\eta,\mathbf{q}} \simeq \hbar c_\eta q$, by a factor $2^{1/2}$, but it provides a very good approximation of the zone-edge energies, as obtained from experimental spectra (Nilsson and Nelin, 1972), more important in high-energy and high-field transport. As a consequence, at small q the scattering rates will be overestimated by the same factor $2^{1/2}$. This will be compensated by smaller coupling constants Δ_η . The dispersion of the optical phonons has been ignored.

ii) Carrier-impurity scattering

The scattering rate, $1/\tau_{\mathrm{imp},\nu}(\mathbf{k})$, for the collision suffered by a carrier of wave-vector $\mathbf{k}$ in the ν-th band in the screened Coulomb field of an ionized dopant has been computed starting from the Brooks-Herring formula (see Ridley, 1977, and references therein) corrected for the band-structure effects:

$$(4) \qquad \frac{1}{\tau_{BH,\nu}(\mathbf{k})} = \frac{N_{dop} Z^2 e^4}{4\pi^2 \hbar \varepsilon^2} \sum_{\nu',\mathbf{k}',\mathbf{G}} \frac{\mid \mathcal{I}(\nu,\nu';\mathbf{k},\mathbf{k}') \mid^2}{\left[\beta_s^2 + \mid \mathbf{k} - \mathbf{k}' + \mathbf{G} \mid^2\right]^2} \delta\left[E_\nu(\mathbf{k}) - E_{\nu'}(\mathbf{k}')\right] ,$$

where ε is the static dielectric constant, N_{dop} is the concentration of ionized dopants, eZ their charge, and the screening parameter β_s has been obtained in the Debye approximation:

$$(5) \qquad \beta_s(\mathbf{r}, t) = \left[\frac{e^2 n_p(\mathbf{r}, t)}{\varepsilon k_B T_p(\mathbf{r}, t)} \right]^{1/2},$$

where k_B is the Boltzmann constant. In Eq. (4) and in the following, the sum over the vectors of the reciprocal lattice, $\mathbf{G}$, will be restricted to the particular $\mathbf{G}$ needed to map $\mathbf{k}-\mathbf{k}' + \mathbf{G}$ into the first BZ. The simulation of a synchronous ensemble of particles gives us the possibility of estimating the local average particle density n_p as a function of time and position and the average particle energy, $\overline{E}$, which is then converted to an effective particle temperature $T_p = 2\overline{E}/3k_B$. These values are then used to compute the screening parameter β_s in a self-consistent way. Finally, Ridley's *statistical screening* model (Ridley, 1977) is employed to compute the final rate:

$$(6) \qquad \frac{1}{\tau_{imp,\nu}(\mathbf{k})} = \frac{v_g(\mathbf{k})}{d} \left[1 - \exp\left(- \frac{d}{v_g(\mathbf{k}) \tau_{BH,\nu}(\mathbf{k})} \right) \right],$$

where $v_g(\mathbf{k})$ is the group velocity and $d = (2\pi N_{dop})^{-1/3}$ is the average distance between the ions. This rate must be calculated during the Monte Carlo simulation, since its dependence on many variables (n_{el}, T_{el}, N_{dop}) with a wide dynamic range would imply an unmanageable size for a look-up table and too many CPU-time-consuming interpolations.

iii) Carrier-carrier scattering

The short-range carrier-carrier interaction has often been found difficult to treat in Monte Carlo simulations (Jacoboni and Reggiani, 1983; Lugli and Ferry, 1985a). The total rate, $1/\tau_{ee,\nu}(\mathbf{k})$, for the screened collision suffered at time t by a particle at position $\mathbf{r}$ and of wave-vector $\mathbf{k}$ in the ν-th band with any other particle in the system can be obtained in the Born approximation as:

$$(7) \qquad \frac{1}{\tau_{ee,\nu}(\mathbf{k}, \mathbf{r}, t)} = \frac{e^2}{8\pi^5 \hbar v^2} \sum_{\substack{\nu', \mu, \mu', \\ \mathbf{k}', \mathbf{p}, \mathbf{p}'}} \frac{| \mathcal{I}(\mu,\mu'; \mathbf{p}, \mathbf{p}') |^2 \; | \mathcal{I}(\nu,\nu'; \mathbf{k}, \mathbf{k}') |^2}{\left[\beta_s^2 + | \mathbf{k}-\mathbf{k}' + \mathbf{G} |^2 \right]^2} \delta(E_{tot}) \delta_{\mathbf{K}} f(\mathbf{r}, \mathbf{p}, t) .$$

The sum extends over all final states $\mathbf{k}'$, over the distribution $f(\mathbf{r}, \mathbf{p}, t)$ of 'partner' particles at $\mathbf{r}$ with wavevector $\mathbf{p}$ at time t, over the possible final states $\mathbf{p}'$ of the partners, and over all possible bands, as allowed by conservation of momentum, $\mathbf{K} = \mathbf{k} + \mathbf{p} - \mathbf{k}' - \mathbf{p}' + \mathbf{G} = 0$, and total energy, $E_{tot} = E_\nu(\mathbf{k}) + E_\mu(\mathbf{p}) - E_{\nu'}(\mathbf{k}') - E_{\mu'}(\mathbf{p}') = 0$. $\mathbf{G}$ is the vector of the reciprocal lattice – nonzero for Umklapp processes – such that $\mathbf{k}-\mathbf{k}' + \mathbf{G}$ is in the first BZ. The recognized difficulty is the presence of the distribution function f, an unknown, in the expression for the scattering rate, which renders the BTE nonlinear. Self consistent methods of various types have been proposed in the past. The review article by Jacoboni and Reggiani (1983) gives a detailed account of this issue.

The following considerations help in finding a solution to the problem (Wang and Hess, 1985; Artaki and Hess, 1987): At time t during an ensemble MC simulation, the distribution function at a given position at time $t - \Delta t_{buf}$ is known, at least within the statistical uncertainty caused by the finite number of particles in the simulation. For a given particle at $\mathbf{r}$ at time t we can search for all particles within a distance R from it, obtaining a statistical estimate of the function f to compute the rate (7). Opposite requirements work in putting constraints on the value of R: It must be small enough so that variations of density, average energy, and other other averaged quantities are negligible and a homogeneous situation exists within the distance R from the given particle. In this case, a sufficiently 'local' statistical sampling of the function $f(\mathbf{r}, \mathbf{p}, t)$ can be obtained. On the other hand, if the distance R is too short, the number of 'partners' within the distance R will be too small to provide a meaningful statistical sample of f. A further constraint arises when the MC particle simulation is coupled self-consistently to the space-charge: since the long-range Coulomb carrier-carrier interaction is already accounted for by the self-consistent scheme, a value of R larger than the spacing, Δx, of the mesh used to solve Poisson equation, would result in double-counting the

long-range coupling. We have opted for a compromise, by choosing $R \approx \beta_s^{-1}$. In order to minimize the number of operations needed to identify the partners within the distance R, a standard technique described by Hockney and Eastwood (1981) is used to construct a list of pointers to partners lying within a set of squares centered on each particle. The total area A of this set of squares is chosen to be as close as possible to the area of the screening circle. Finally, the probability of finding partners within the screening circle must be rescaled to account for the fact that the probability of finding $\mathcal{N}_{2D}$ partners within a screening length in the 2-dimensional space used in the simulation is different from what would be obtained in a 3-dimensional space, $\mathcal{N}_{3D}$. The 3-dimensional number of partners is related to the 2-dimensional value by $\mathcal{N}_{3D} = (4\pi s\, \mathcal{N}_{2D}) / (3\beta_s^3 A)$, where s is the scale factor determining how many carriers per unit length in the third dimension each simulated particle represents. This factor is determined at the beginning of the simulations, as we shall see below. Both the short-range inter-particle scattering and the long-range Coulomb interaction mediated by the self-consistent particle-Poisson coupling (discussed below) are correctly accounted for only in the limit of a large number of particles, or, equivalently, of small s-factors.

We are now ready to treat the short-range carrier-carrier collisions: When the scattering rate is needed for a particle at position $\mathbf{r}$ of wavevector $\mathbf{k}$ in band ν, one among its neighbors within the screening circle centered at $\mathbf{r}$ is selected randomly (Matulionis *et al.*, 1975). Denoting by $\mathbf{p}$ the wavevector of this partner and by μ its band, we evaluate the scattering rate for this pair as:

$$(8)\quad \frac{1}{\tau_{ee,\nu,\mu}(\mathbf{k},\mathbf{p})} = \frac{\mathcal{N}_{3D}}{(4/3)\pi\beta_s^{-3}} \frac{e^2}{8\pi^5\hbar\varepsilon^2} \sum_{\substack{\nu',\mu',\\ \mathbf{k}',\mathbf{p}'}} \frac{|\mathcal{I}(\mu,\mu';\mathbf{p},\mathbf{p}')|^2\, |\mathcal{I}(\nu,\nu';\mathbf{k},\mathbf{k}')|^2}{\left[\beta_s^2 + |\mathbf{k}-\mathbf{k}'+\mathbf{G}|^2\right]^2} \delta(E_{tot})\delta_{\mathbf{K}},$$

by using a trivial, but time consuming, extension of the technique employed to compute the carrier-phonon rates.

iv) Impact ionization

A simple Keldysh formula (Keldysh, 1965) is used to derive the rate for impact ionization for a carrier of energy E:

$$(9)\quad \frac{1}{\tau_{ii}(E)} = \begin{cases} 0 & (E \le E_{th}) \\ \dfrac{P}{\tau_{op}(E_{th})} \left(\dfrac{E - E_{th}}{E_{th}} \right)^2 & (E > E_{th}) \end{cases}$$

where E_{th} is a threshold energy and $1/\tau_{op}(E_{th})$ is the carrier/optical-phonon scattering rate averaged over all wave-vectors corresponding to the threshold energy E_{th}. Finally, P is a coefficient which we consider merely a fitting parameter.

v) Degeneracy effects

Degeneracy effects are important in heavily doped regions (Lugli and Ferry, 1985b). In these regions the large charge density yields very low fields and negligible carrier heating. The strong carrier-carrier interaction is also very efficient in distributing energy among the carriers and driving them towards a Fermi-Dirac distribution (Lugli and Ferry, 1983; Brunetti *et al.*, 1985). Therefore, we approximate the distribution function f as:

$$(10)\quad f_{\text{app}}(E,\mathbf{r},t) \simeq \frac{1}{1 + \exp\left[\dfrac{E - E_F(\mathbf{r},t)}{k_B T_{el}(\mathbf{r},t)} \right]},$$

where $E_F(\mathbf{r},t)$ is the Fermi level at position $\mathbf{r}$ and time t obtained self-consistently from the local particle density during the simulation. Any collision process is then rejected if the final state, $\mathbf{k}'$, and band index ν' selected after the collision is such that

$$1 - f_{\text{app}}[E_{\nu'}(\mathbf{k}'), \mathbf{r}, t] \leq \xi , \tag{11}$$

where ξ is a random number in $[0,1]$. Thus, we account correctly for degeneracy in heavily doped regions, even when the carriers are slightly heated. Major errors are made in regions where the carriers are hot and largely off-equilibrium. However, in these regions the densities are usually low and degeneracy plays an insignificant role.

Selection of final states

After a collision process involving a particle in the initial state $(\mathbf{k}, \nu)$, its final state, $(\mathbf{k}', \nu')$, is selected with a technique similar to the one employed to compute the scattering rates. First, all cubes centered around the $\mathbf{k}$-points in the fine BZ-mesh are scanned to select those which intersect the equi-energy surface at the desired final energy E'. This is done searching over the coarse mesh first, over the fine mesh afterwards. Once the energy-conserving cubes are found, each one centered around a vector $\mathbf{k}_{m'}$, each cube is assigned a weight given by its DOS, $\mathcal{D}_{\nu'}(E', \mathbf{k}_{m'})$, as given by the Gilat-Raubenheimer algorithm (Gilat and Raubenheimer, 1966), the associated overlap integral, $\mathcal{I}(\nu,\nu'\mathbf{k}, \mathbf{k}_{m'})$, and the squared matrix element $|\mathcal{M}(\mathbf{q}_m)|^2$, where $\mathbf{q}_m = \mathbf{k} - \mathbf{k}_{m'} + \mathbf{G}$. A random vector, $\mathbf{k}_{\bar{m}'}$ is then selected, with probability given by its weight. The rejection technique is employed for this random selection. The final-band index and k-vector are then known. A final step is necessary, as the energy associated with this wavevector can differ from E' by as much as a few tens of meV for the mesh-size we have used. Therefore, a correction is made within the selected small cube to adjust the final state, as explained above.

In the case of impact ionization a much simpler approximation is made, consistent with the simplicity of Eq. (9). The recoil particle is assigned an energy $E_\nu(\mathbf{k}) - E_{gap}$, where E_{gap} is the band-gap of the semiconductor, and a random wavevector at this energy is selected. The generated particle is placed at the bottom of the band.

HOMOGENEOUS TRANSPORT

Considering our present inability to describe the variations of the carrier-phonon matrix elements Δ_η over the various bands in the BZ, we have taken a very simple approach and treated these constants as empirical properties of the material. A first reason for doing so stems from our belief that band-structure effects play a dominant role at high-energies. A second justification is that some of the best experimental determinations of the deformation potentials have been obtained in the past from low-field, homogeneous transport data in the bulk semiconductor fitted to Monte Carlo simulations (Canali *et al.*, 1975; Jacoboni *et al.*, 1977). Because of the different band-structure we employ, we expect possible differences from previous work, even at low fields. Therefore, we shall follow the same 'fitting' path of the past, but paying attention also to high-field and high-energy situations. Our guideline is 'simplicity'. We look for the simplest possible set of values which match experimental data. In search for this simplicity, we shall make many crude approximations. We now discuss electron and hole transport in turn. It is important to stress that these coupling constants are determined *solely* by bulk, steady-state transport data. Therefore, the device-modeling results we shall present in following sections are to be considered *transport-parameter free*. In the following two subsections impurity and carrier-carrier scattering are ignored.

Electrons

The simplest possible choice we can make is a unique acoustic Bardeen-like deformation potential (Bardeen and Shockley, 1950), $\Delta_{ac}(\mathbf{q}) = \Delta_{LA}q = \Delta_{TA}q$ for both LA and TA phonons and a unique nonpolar-optical Harrison-like deformation potential (Harrison, 1956),

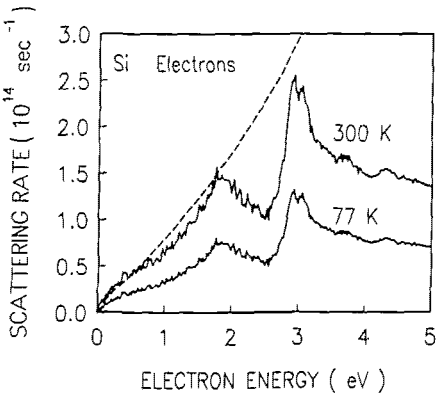

Figure 2. *Total electron-phonon scattering rate for Si at room and liquid-nitrogen temperature. The dashed line corresponds to the total electron-phonon scattering rate given by Jacoboni and Reggiani (1983).*

$\Delta_{op}(\mathbf{q}) = (\Delta K)_{op}$, for both LO and TO phonons. Also, considering the small energy difference of the two optical modes, we shall consider LO phonons only.

With this set of assumptions, we adjust the values of Δ_{ac} and $(\Delta K)_{op}$ to match the experimental velocity-field characteristics at 300 K, and the low-field results of previous Monte Carlo simulations (Canali *et al.*, 1975; Jacoboni *et al.*, 1977; Jacoboni and Reggiani, 1983). We immediately run into troubles at fields exceeding 10^4 V/cm. We must complicate our picture slightly by allowing the deformation potentials to take different values in the second and higher bands. We also use the impact-ionization parameters P and E_{th} to fit the experimental ionization coefficients (Lee *et al.*, 1964; Crowell and Sze, 1966; van Overstraeten and DeMan, 1970) and probability of injection into SiO_2 (Ning *et al.*, 1976), exactly as done by Tang and Hess (1983b).

The parameters we employ are: $\Delta_{LA} = \Delta_{TA} = 1.2$ eV (band 1), 1.7 eV (higher bands); $(\Delta K)_{op} = 1.75\times10^8$ eV/cm (band 1), 2.10×10^8 eV/cm (higher bands) $E_{th} = 1.2$ eV, $P/\tau_{op}(E_{th}) = 10^{11}$ sec^{-1}. We show in Fig. 2 the total electron-phonon scattering rate as a function of electron energy obtained integrating numerically our anisotropic rates over all directions. The strong role played by the density of final states is clearly evident comparing Fig. 1(b) to Fig. 2. The low-energy rates at 300 K resemble closely the magnitude of the rates used in previous Monte Carlo work (Brunetti *et al.*, 1981), shown by the dashed line. In Fig. 3(a) we show the drift velocities vs. electric field at 300 K and 77 K. Barely visible in the figure is a region of negative differential mobility at

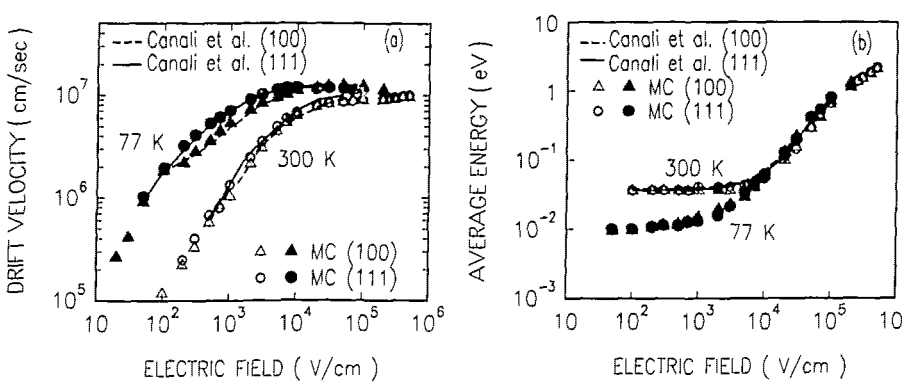

Figure 3. *(a) Experimental and simulated electron drift velocity in Si as a function of electric field along two crystallographic directions at 77 and 300 K. (b) Simulated average electron energy for Si at 77 and 300 K as a function of electric field along two crystallographic directions. Results of previous Monte Carlo simulations at 300 K are shown for comparison.*

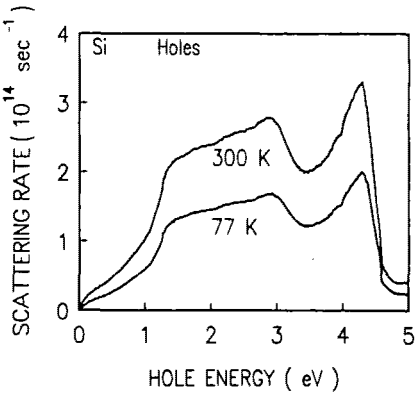

Figure 4. *Total hole-phonon scattering rate for Si at 77K and 300 K.*

77 K at high fields ($\gtrsim 3\times10^4$ V/cm), as a few carriers begin to transfer into the L-valley at about 1 eV. The average electron energy as a function of electric field, shown in Fig. 3(b), is slightly lower than that obtained by 'parabolic' Monte Carlo simulations at high-fields. More about this effect and the role played by L-valley transfer in small devices will be said below.

Holes

In the case of hole transport in Si the band structure is even more complicated due to the warping and strong nonparabolicity of the bands. The calibration of deformation potentials parameters has been done in the same spirit as in the approach taken for electrons. After fitting experimental drift velocity vs. field curves (Ottaviani *et al.*, 1975) and ionization-coefficients (Grant, 1973; Lee *et al.*, 1964), the parameters we employ are: $\Delta_{ac} = 3.5$ eV, $(\Delta K)_{op} = 6.0\times10^8$ eV/cm, $E_{th} = 1.21$ eV, $P/\tau_{op}(E_{th}) = 9.0\times10^{14}$ sec^{-1} . In Figs. 4 and 5, we show the hole-phonon scattering rates, hole drift-velocity and average energy, at 77 K and 300 K, obtained with the parameters listed above.

SPACE-CHARGE EFFECTS

At present, the self-consistent coupling of the particle (MC) picture to the Poisson equation can be handled with almost standard techniques, such as those described by the pioneers of this approach (Hockney and Eastwood, 1981; Moglestue, 1986). For completeness, we shall outline the

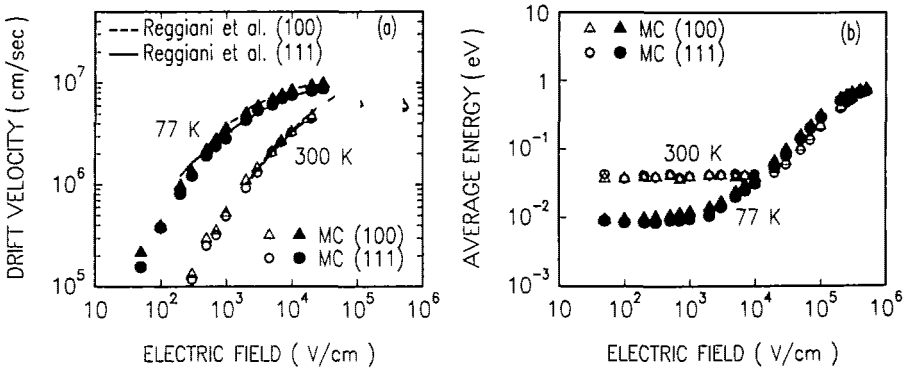

Figure 5. *(a) Experimental and simulated hole drift velocity as a function of electric field along two crystallographic directions in Si at 77 K and 300 K. (b) Simulated hole average energy in Si at 77 K and 300 K as a function of electric field along the two crystallographic directions.*

procedure we have followed, emphasizing the few instances which deviate from the Hockney and Eastwood prescription.

Poisson Equation

The mesh which describes the 2-dimensional cross section of the device is a tensor product, nonuniform, finite difference mesh with no terminating lines. Mesh sizes have typically been 100 × 50 mesh lines in the x- and y-axis directions, respectively. (Here, the source-to-drain direction is the x-axis direction). The Poisson equation is solved on this mesh taking into account the instantaneous particle density, the ionized dopant density, potential boundary condition, and a piecewise constant dielectric constant. Holes (for electron transport) or electrons (in the case of hole transport) are included in the zero-current (constant quasi-Fermi level) approximation, so that the depletion region in the substrate of a n- (or p-) MOSFET is self-consistently and automatically included in the calculation. Therefore, the Poisson equation to be solved (for the case of electron transport, to fix the ideas) is:

$$(12) \qquad -\nabla_{\mathbf{r}} \cdot (\epsilon \nabla_{\mathbf{r}} \phi) = e\left[N_V \mathscr{F}_{1/2}\left(\frac{E_V - E_F}{k_B T} \right) - n_{el} + N_D^+ - N_A^- \right],$$

where $\mathscr{F}_{1/2}$ is the Fermi-Dirac integral of order one half, E_F is the hole quasi-Fermi level, $N_V \mathscr{F}_{1/2}$ is the hole-concentration, E_V the energy of the top of the valence band, and N_D^+ and N_A^- are the concentrations of the ionized donors and acceptors, respectively. To calculate the fraction of ionized donors, quasi-equilibrium is assumed with the local electron density. This permits a local electron quasi-Fermi level to be defined (assuming parabolic bands for simplicity) which establishes the donor occupancy according to Fermi-Dirac statistics. A similar treatment is embraced to determine the ionized acceptor from the local hole density. A Newton-Raphson method is used to solve the nonlinear system of equations, together with a damping scheme proposed by Bank and Rose (1980). A polynomial pre-conditioned conjugate gradient technique is used to solve the resultant linearized matrix equations (Johnson and Micchelli, 1983).

The standard cloud-in-a-cell (CIC) algorithm (Hockney and Eastwood, 1981) is employed to assign the particle charge density $\pm e n_p(\mathbf{r})$ to the mesh-nodes and to interpolate the mesh-forces acting on the particles, with two extensions. These extensions are necessary because the standard CIC method applies to a mesh with uniform spacing in the axis directions and for a constant dielectric constant. Unfortunately, these conditions are inappropriate for our MOSFET simulations. Our two extensions are discussed elsewhere (Fischetti and Laux, 1988) and are only summarized here. The standard CIC is extended to nonuniform meshes, and is quite straightforward. However, the correctness of this extension is uncertain since the particle self-forces are no longer zero. To guard against errors, the Poisson mesh is only graded where the particle densities are negligible. The second extension to the CIC method involves accommodating the spatially-varying dielectric constant. A centered-difference approximation to $\mathbf{F}$ is inappropriate, since the perpendicular component of $\mathbf{F}$ is spatially discontinuous across interfaces of varying dielectric constant. This is remedied by calculating the electric flux density $\mathbf{D}$ ($= \epsilon \mathbf{F}$) directly in this case, as it remains continuous. Finally, the local electric field is obtained from the electric flux vector.

Monte Carlo-Poisson coupling

We are now ready to describe the structure of the self-consistent coupling between the Monte Carlo particle model and the solution of the Poisson equation.

1. *Setting up the problem*

At the beginning of the simulation, the grid, doping profiles, and contacts are defined for the device under investigation. We also need to specify the initial particle locations in real and **k**-space. We shall now give a few details of the procedure we followed to start the simulation.

The grid is chosen empirically, refining the mesh spacings in regions where high gradients of carrier concentrations and electrostatic potential are expected. Doping profiles of various forms

(constant, gaussian, or empirical) can be introduced. Ohmic contacts must be separated from active regions of the device by a distance sufficient to ensure that an equilibrium condition (*i.e.*, thermal carriers and charge neutrality) exists in their immediate neighborhood. Some experimentation is normally required to guarantee the fulfillment of this condition. Whenever the local particle density at the contact drops below the known equilibrium value (not necessarily spatially uniform, to accommodate nonuniform doping profiles), carriers are 'injected' into the device with k-vectors selected randomly according to the local Fermi-Dirac distribution at the lattice temperature.

To initialize the particle distribution, the particle locations can be obtained from a previous solution (at a different bias or temperature, for instance) or from a standard Drift-Diffusion-solution of the device and using the resulting particle-density as the probability distribution to place a predetermined number of particles in the device. As a crude third alternative, particles can be distributed according to the doping profiles. In any case, the total charge in the simulated region is known. This fixes the charge, *es*, associated with each simulated particle per unit length. The factor *s* represents the number of real electrons per simulated particle per unit length as defined above. The initial wavevector assigned to each particle is either obtained from a previous solution or chosen randomly with a probability-distribution given by the local Fermi function. Once the initial distribution of particles in real and **k**-space is obtained, the Poisson equation is solved to initiate the transient evolution with a consistent electric field solution.

A characteristic problem of MC device simulations originates from the fact that in most practical cases the particle density exhibits a wide dynamic range, since contacts may be degenerately doped. Typically, the program must be able to handle carrier concentrations ranging from a few 10^{16} cm^{-3} to 10^{20} cm^{-3} or more. This would result in a large number of particles in the highly-doped (and usually, but not always, uninteresting) contact regions. To avoid spending excessive CPU-time simulating these carriers, we have treated low-energy electrons (typically below 50 meV) with a first-order nonparabolic approximation to the band structure – thus bypassing time-consuming interpolations while still retaining the basic features of the full band structure at higher energies – and have used the usual technique for enhancing rare events in MC simulations (Phillips and Price, 1977; Sangiorgi *et al.*, 1988) by increasing the particle population in the active regions of the devices (such as inversion channels in MOSFETs). Particular care has been taken to handle correctly the short-range carrier-carrier interaction between particles with different statistical weights (Fischetti and Laux, 1988).

2. *Time evolution*

Given the initial particle locations and field configuration in the simulated region, the particles are moved in free-flight for a time Δt_{bal}. At the end of this step, particles which leave the region at a contact are tallied as positive current at that contact. Particles hitting interfaces (such as the $Si\text{-}SiO_2$ interface) are specularly reflected, diffused elastically or inelastically, depending on the physical model chosen. The results we shall present below are obtained by using a mixture of specular reflections and elastic diffusions (each occurring with probability 0.5 at every 'hit'). We shall discuss below the expected limitations of this approach. Particles will also be injected from the contacts, and tallied as negative current in response to charge-neutrality considerations, as explained above.

At a time interval Δt_{sc} the number of particles undergoing any collision and the type of collisions is selected in a conventional way. To account properly for the total scattering probability, the scattering time step Δt_{sc} must be chosen in such a way that $\Delta t_{sc} \ll \tau_{tot,max}$, where $1/\tau_{tot,max}$ is the maximum scattering rate occurring in the simulation. Obviously, the free-flight-time Δt_{bal} must be less than or equal to Δt_{sc}.

The Poisson equation is periodically solved, to update the electric field corresponding to the new positions of the particles. The frequency at which this update is performed is a critical issue. An electron gas of density n_{el} can develop plasma oscillations (Pines, 1963) of angular frequency, ω_p, which, in the effective-mass approximation, is given by:

$$\omega_p = \left(\frac{e^2 n_{el}}{\varepsilon m_{eff}} \right)^{1/2} , \qquad (13)$$

where m_{eff} is the electron effective mass, corresponding to small-q collective excitations arising from the long-range Coulomb interaction. In highly doped regions, the frequency of the plasma oscillations may approach the typical frequency at which free-flights and scattering-checks are performed. To quantify the ideas, in our simulation we used $\Delta t_{bal} \simeq 0.2$ fsec, while in the contact regions (where the electron concentration can be as high as 2.0×10^{20} cm^{-3} in the accumulation layers under the gate contact), we have $\omega_p^{-1} \simeq 2.4$ fsec. According to the Nyquist theorem, the field configuration must be updated at least every $0.5\omega_p^{-1} \approx 1.0$ fsec, or else an 'undersampling' of the plasma modes occurs. Such an undersampling results in catastrophic instabilities. To overcome this 'plasma-catastrophe', we could smoothen the potential by averaging it over a suitable time interval (Throngnumchai *et al.*, 1986) or spatial region (Venturi *et al.*, 1988). However, in so doing we would loose the possibility of accounting for plasma losses by hot carriers. Therefore, we have chosen to update the electric field at a very high-frequency by using a time, Δt_{Pois}, between successive Poisson updates such that $\Delta t_{Pois} \lesssim 1/(5\omega_p)$ (= 0.2 to 0.4 fsec in our simulation, *i.e.* every one or two ballistic steps). The higher CPU-time spent in solving Eq. (12) very often is the price paid for the additional physics added to the model.

It is not sufficient to resolve the plasma oscillations in the time-domain, since other conditions concerning the real-space resolution of the simulation must be satisfied: The discussion of the considerations involved in the choice of the Poisson-mesh size, of the 'screening' radius R, and on the s-factor defined above is given by Fischetti and Laux (1988), together with a discussion on the effect of these variables on the treatment of various plasma effects in the highly doped regions.

Finally, average quantities are computed and dumped onto a mass-storage device. Particle positions, their trajectories, types of collisions, average energies, velocities, densities, currents, and other possible quantities of interest can be viewed with an interactive graphics program developed for this application.

THE PROGRAM

In typical applications, an ensemble of 5,000 to 10,000 particles is employed, with a statistical-enhancing factor $M = 10$. The Poisson mesh consists typically of 100×50 nodes. The time steps used were (as mentioned above): $\Delta t_{bal} = 2 \times 10^{-16}$ sec, $\Delta t_{sc} = 2 \times 10^{-16}$ sec (T = 300K) and 10^{-15} sec (T = 77K), $\Delta t_{Pois} = 4 \times 10^{-16}$ sec. For the small devices we have simulated, steady state was obtained after 0.4 psec (60 nm channel length) to 2 psec (0.25 μm channel length). The simulations have been continued for about 3 to 5 psec after the end of the transient in order to gather accurate steady-state solution statistics.

The program, written in VS/FORTRAN, runs on an IBM model 3090/600E computer with vector facilities. In many cases, standard algorithms have been modified for vectorization purposes. Typically, the program spends 50% to 70% of its total CPU-time in the vector hardware. The size of the memory region required by the look-up tables for the band-structure and scattering rates over the entire BZ makes it necessary to employ the extended architecture. Region sizes of 400 Mbytes are normally required.

CPU-times are typically quite large. These times are attributable to the large number of interpolations over the BZ, the high frequency of Poisson solutions to resolve the plasma oscillations, and, more important, to the extremely costly evaluations of Eq. (4) and particularly Eq. (8). The program requires 1 to 6 CPU-sec to simulate one particle for one psec when the short-range electron-electron scattering is turned off. This time increases to 10 or even 40 CPU-sec (depending on many parameters, such as the various time steps, the particle density, etc.) when this interaction is included. Thus, a typical bias point requires CPU times on the order of tens of hours. For comparison, a simulation using parabolic bands and updating the field at a frequency ten times smaller requires CPU-times 20 to 100 times shorter.

RESULTS

In this section we present the results of simulations we have performed on exploratory short n-channel Si-MOSFETs (Sai-Halasz *et al.*, 1987). New results which emerge from the simulations are: 1. The quasi-ballistic nature of electron transport in devices having 60 nm effective channel-length at 77 K. 2. The strong role played by the electron-electron interaction in these conditions. 3. The strong velocity-overshoot predicted theoretically and observed experimentally. 4. The dramatic role that band-structure effects play at high biases and low temperatures. 5. The much different behavior of holes in p-channel devices.

The devices we studied have an effective channel length ranging from a little over 0.25 μm down to 60 $\pm$ 5 nm. (Sai-Halasz *et al.*, 1987). Direct-write electron-beam lithography was used. Degenerate source/drain double-implants (arsenic and/or antimony) have a peak concentration of 1.5×10^{20} cm^{-3}. A deep channel implant was used to prevent punch-through and to minimize the degradation of the channel mobility by maintaining a low impurity concentration in the channel. Finally, the thickness of the gate oxide was 4.5 nm and a (100)-oriented Si substrate was employed. Gaussian profiles matching the implant conditions were employed in the simulation. The devices with channel length smaller than 0.1 μm are designed for operation at 77 K with a reduced supply voltage (0.8 V) and 0.6 V applied to the substrate contact. This substrate bias was employed for the simulation at liquid-nitrogen temperature, while the contact was grounded in the 300 K experiments and simulations.

Before discussing these issues in some detail, we wish to spend a few words on the limitations of our model and how they might affect the results. Our concerns focus on the absence of 2D-quantization in the channel, the poor treatment of interface scattering, the crude approximation used to handle impact ionization, and general open issues about high-field transport.

Quantization in the channel and interface scattering should affect strongly the field-effect mobility at low drain fields (*i.e.* at low source/drain bias, V_{DS}). It is well known that the saturated velocity in long channels is much smaller than in the bulk of the semiconductor (Fang and Fowler, 1970; Modelli and Manzini, 1988). Unless proper account is taken of the 2D-features of electron transport, of the correct scattering with interfacial impurities in the gate insulator (Stern and Howard, 1967; Ning and Sah, 1972; Manzini, 1985), and of the roughness of the Si-SiO_2 interface (Park *et al.*, 1983), no agreement can be expected from the model. Even authors who have accounted for these features have met some difficulties, with theoretical analysis predicting mobilities higher than those observed experimentally (Hess and Sah, 1974; Basu, 1977, 1978). For these reasons, we have concentrated our attention on the high- V_{DS} characteristics, corresponding to average source/drain fields in excess of 10^5 V/cm, so that the carriers are sufficiently hot over a large fraction of the channel to be correctly described by their bulk transport dynamics and kinematics. Hot carriers will be significantly displaced from the Si-SiO_2 interface, so that interface scattering has, hopefully, a minor effect. Of course, at the source-end of the channel 2D-effects always dominate. In very short channels, electrons do not spend enough time in the channel to thermalize by the drain-end, even when the short-range electron-electron interaction is included. Thus, some 'memory-effect' might carry information of the 2D-configuration from the source to the drain-end. At present, we lack any information on the importance of this effect in the shortest devices we have considered.

On the other side, if realistic V_{DS} are to be used, the maximum energies gained by the carriers are still below 2.5 eV for channel lengths smaller than 0.25 μm. We feel fairly confident that the band-structure effects and our 'fitting' approach employed to determine the scattering rates reproduce very well the main features of electron transport at these energies. However, we must still keep in mind the uncertainties surrounding the theoretical formulations of transport in this regime. From all these considerations, our results must be viewed cautiously – our approach improves significantly the 'state-of-the-art', but more work and additional experimental verification are needed to bolster our confidence.

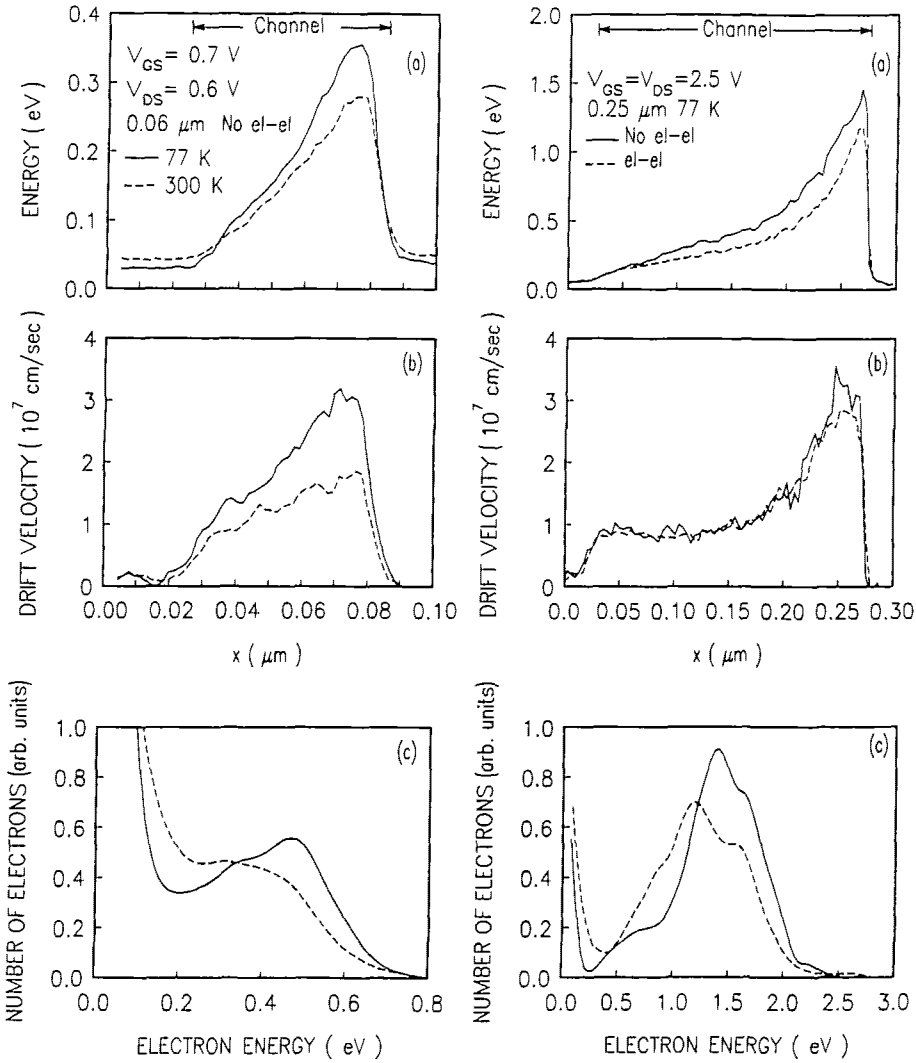

Figure 6. **(left column)** *Average electron energy (a) and x-directed drift velocity (b) at the distance of 1 nm from the Si-SiO_2 interface along the n-channel of a Si MOSFET having an effective channel length of 60 nm. The smoothed electron energy distributions at the drain-end of the channel (x=0.085 μm in (a) and (b)) are also shown in (c).*

Figure 7. **(right column)** *Electron average energy (a) and x-directed drift velocity (b) profiles along the channel 1 nm away from the Si-SiO_2 interface for a device having an effective channel length of 0.25 μm at 77 K at the bias conditions indicated in the figure with and without the inclusion of short-range electron-electron scattering. Electron energy-distribution at the drain-end of the channel (x=0.275 μm in (a) and (b)) are shown in (c). Features related to DOS effects at about 1 eV (onset of the L-valleys) and 1.7 eV (see Fig. 1(b)) can be observed.*

Quasi-ballistic transport

In the simulation, the metallurgical channel of the smallest devices we consider was assumed to be 43 nm long, which yielded an effective channel length of about 60 nm. The effective channel length was estimated by plotting the electron quasi-Fermi level, ϕ_n , from source to drain at the Si-SiO_2 interface, and estimating the breakpoints in ϕ_n at the ends of the channel. This corresponds

roughly to the positions at which the electron density stops following the doping profiles in the source and drain implanted regions as one moves from these regions into the channel. We start by presenting results of the simulation performed at V_{DS} = 0.6 V, gate voltage V_{GS} = 0.7 V, *i.e.* about 0.50 V above the 77 K-threshold voltage of the devices. Results of runs performed without the short-range electron-electron interaction are discussed first.

In Fig. 6(a) we show the average energy of the electrons along the channel, 1 nm away from the Si-SiO_2 interface at 77 K and room temperature. Fig. 6(b) shows the velocity-profile along the channel. The low-temperature results indicate that the electrons can reach, on average, as much as 0.35 eV, which is a very significant fraction of the total voltage applied between the source and drain contact. This suggests that very few collisions occur along the high-field region of the channel, as indicated clearly by energy distributions at the metallurgical junction at the drain-end, *i.e.*, just inside the drain-contact. The electron energy-distribution (shown at two temperatures in Fig. 6(c)) indicate that the highest energies are reached just inside the drain contact. Very pronounced off-equilibrium features are seen: a peak at about 0.5 eV at low-temperature (the lower average-energy seen in Fig. 6(a) results from the large number of thermal carriers in the drain) and the absence of cooler carriers. Most of the collisions are in the form of LO-phonon emission (intervalley, both *g* and *f*-scattering) and interface scattering, but occur mostly in the first half of the channel. Once the electrons enter the pinched-off region, their high velocity (shown in Fig. 6(b)) and the overshoot regime result in mean-free-paths exceeding 20 nm. Thus, the average electron undergoes at most two phonon-collisions in the high-field region. The room temperature behavior is less dramatic, since the shorter relaxation lengths prevent the carriers from flowing quasi-ballistically along the channel.

Short-range e-e collisions

The inclusion of the short-range electron-electron interaction has a very strong effect on the details of the electron-energy distributions, as shown in Fig. 7 for a device having a 0.25μm-long channel. The partial randomization of the electron trajectories results in lower mean-free-paths for phonon emissions along the channel. This yields lower average energies, lower velocities approaching the drain region, and energy distributions shifted to lower energies, as seen in Figs. 7(a), (b), and (c) respectively.

Velocity overshoot

The average drift velocities along the channel shown in Figs. 6(b) and 7(b) indicate that a significant overshoot occurs near the drain end of the device, even at room temperature (Shahidi *et al.*, 1988; Sai-Halasz *et al.*, 1988). A direct comparison with the experimental data can be made by looking at the small-signal transconductance, g_m, as a function of channel length in the saturated region. This is illustrated in Fig. 8. For the shorter devices at 77 K, the 'effective' velocity, $v_{eff} = g_m/C_{ox}$ (C_{ox} being the oxide capacitance) extracted from the *extrinsic* transconductance (*i.e.* not corrected for the contact resistance, amounting only to a 5 to 10% correction in any case) is about 1.2×10^7 cm/sec. (Sai-Halasz *et al.*, 1988). This effective velocity actually represents a lower bound to the actual average electron-velocity along the channel (Laux and Fischetti, 1988). Its value, very close to the saturated bulk drift-velocity at 77 K in the (100) crystallographic direction, indicates clearly the presence of velocity overshoot in the experimental data, even ignoring series-resistance corrections. The value of g_m obtained from the simulation agrees within an error better that a few percent with the extrinsic experimental value. This does not prove that the actual velocity distribution is, in reality, as shown in Fig. 6(b). Nevertheless, it proves that the model can predict the macroscopic behavior of these small devices. A simpler DD-model with realistic values for the electron mobility and saturated velocity is obviously unable to yield velocities larger than 1.2×10^7 cm/sec and would underestimate the transconductance of the device by about 30%. The room temperature simulations also predict overshoot along a significant fraction of the channel. But in this case both the simulated and the experimental transconductance (once more in very good agreement) imply a value of v_{eff} which is smaller than the bulk saturated value.

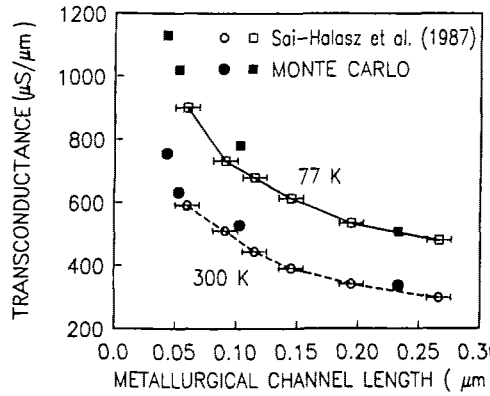

Figure 8. *Experimental and simulated small-signal transconductance as a function of the channel-length. The metallurgical channel length – about 17 nm shorter than the 'effective' channel length defined in the text – has been used for a direct comparison with the experimental data (Sai-Halasz et al., 1988). The experimental data were obtained at a gate bias of 0.6 V above threshold and* V_{DS} = 0.8 *V. The estimated error in the determination of the metallurgical channel is indicated by the horizontal error bars. The simulated values are obtained by taking the difference between the drain currents at* V_{GS} = 1.0 *V and 0.7 V for the 0.06 μm device (0.043 μm metallurgical length) with* V_{DS} = 0.6 *V, at* V_{GS} = 1.0 *V and 0.8 V for the 0.07 μm device (0.053 μm metallurgical length) with* V_{DS} = 0.6 *V, and at* V_{GS} = 1.0 *V and 0.8 V for the 0.12 μm and 0.25 μm devices with* V_{DS} = 1.0 *V. Both the experimental and the simulated values are plotted 'as measured', without correcting for the series-resistance in the source and drain contacts. This amounts to a 5 to 10% increase of the transconductance in both cases at the smallest channel lengths.*

Band-structure effects

A device having a 0.25μm channel length has been simulated using our model including the full band-structure of Si. The results have been compared with those obtained by simulating the same device with a more conventional model employing a parabolic approximation to the conduction band. We have looked for a 'worst-case', but relevant, situation: a relatively high-bias (V_{DS} = 2.5 V , V_{GS} = 2.5 V) at low temperature (77 K) in a device short enough to exhibit strong non-equilibrium effects. The rather large mean-free-path allows the electrons to become hot enough, so that a significant region of the BZ is populated and a good idea of the kinematic and dynamic effects of the band-structure can be obtained. The electron-electron interaction has been suppressed in these runs, as well as first-order nonparabolicity corrections to the band, in order to reproduce the modeling configuration employed in recent simulations (Tomizawa *et al.*, 1988). While the terminal-currents obtained from the two models are virtually identical, the band-structure effects on the internal behavior of the device are indeed dramatic. The parabolic model appears to overestimate consistently the average energies by a large factor, as high as 2, along the channel (Fig. 9(a)). A similar situation was already hinted by the high-field, homogeneous results of Fig. 3(b). Similar results apply to the electron drift-velocity, shown in Fig. 9(b): the parabolic model deviates from the full-band-structure model already in low-field portion of the channel, and it exhibits velocities in excess of 6×10^7 cm/sec at the high field ($\simeq 3\times10^5$ V/cm) present in the pinched-off region. Fig. 9(c) illustrates the electron energy distribution at the drain-end of the channel, stressing, if still necessary, the enormous difference between the two models. It should be noted that at 77 K we have employed in the 'parabolic-band' approximation the set of scattering parameters given by Brunetti *et al.* (1981) and Jacoboni and Reggiani (1983). Those given by Canali *et al.* (1975) provide a much smaller coupling constant for the intervalley g-scattering with LO-phonons and yield even larger discrepancies at low temperature.

The origin of the large difference can be understood looking at the structure of the Si conduction band shown in Fig. 1. At the bias considered and approaching the drain-end of the channel, a significant fraction of the carriers appears to be very close to the L symmetry-point, or even in the Γ-valley. For electrons to reach energies in excess of 1 eV in the 'correct' band structure, regions

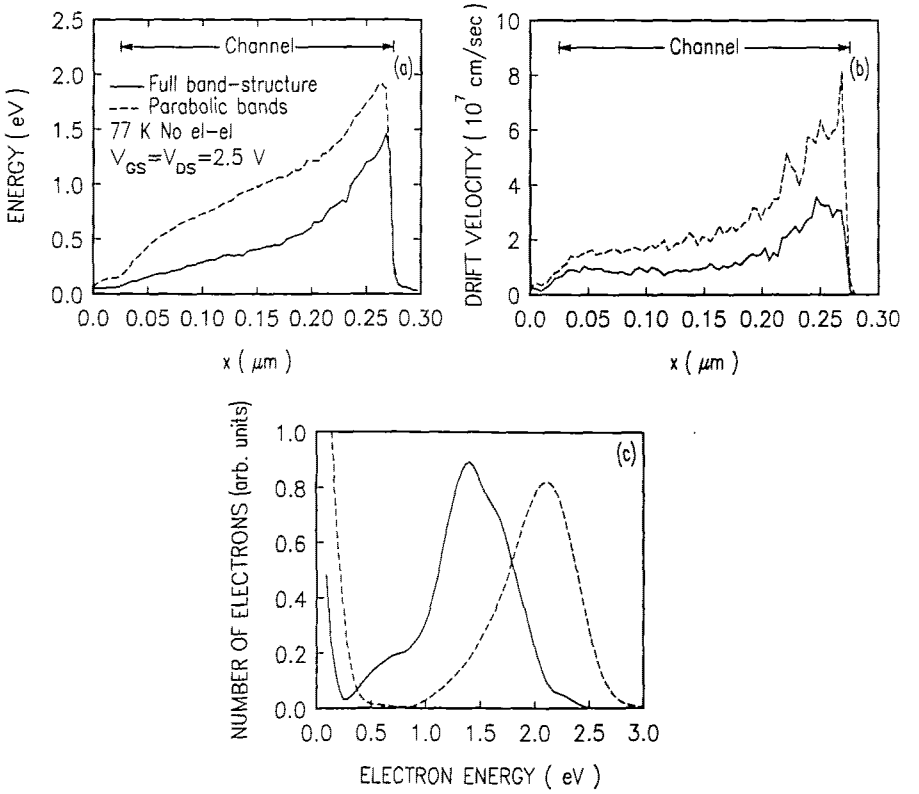

Figure 9. *Electron average energy (a) and x-directed drift velocity (b) at 77 K along the channel 1 nm away from the Si-SiO_2 interface for the device and bias conditions of Fig. 7 obtained from a model including the full band-structure and from a model employing a parabolic-band approximation. In (c) we show the electron energy distribution at the drain-end of the channel (x=0.275 μm in (a) and (b)).*

of the zone having low group velocities (or even hole-like dispersion) must be populated. This has the effect of slowing down the carriers even in the absence of scattering. As an example, electrons accelerated from the band-minimum towards the Γ-point have to climb through a 'crest' of zero group velocity. Therefore, the electron mean-free-path is reduced and more phonon-emissions occur. On the other side, a parabolic-band approximation yields unlimited group velocities, thus missing altogether the important kinematical effects we just discussed. Indirect dynamic effects further worsen the picture, since the higher velocities imply longer mean-free-paths and even smaller energy-loss rates.

Admittedly, we chose the worst case. Indeed, in the opposite limit of higher temperatures the picture appears less dramatic. In particular, the introduction of nonparabolic corrections (Jacoboni *et al.*, 1975), though quite unjustified at high energies, damps the velocities very effectively and increases the scattering rates. However, it does not change the picture qualitatively as far as average electron energies and energy-distributions are concerned, as shown by Fischetti and Laux (1988).

p-channels

As expected, holes in p-MOSFETs exhibit a much more moderate behavior. In Fig. 10 we show the average energies and drift velocities in a 0.25 μm p-MOSFET in a configuration mirroring Fig. 7. Both quantities are significantly lower than the corresponding quantities in the n-channel. Velocity overshoot, however, is observed in the pinched-off region also in this case. Compared to the n-channel case, lower currents and transconductances are obtained for the p-channel devices (about

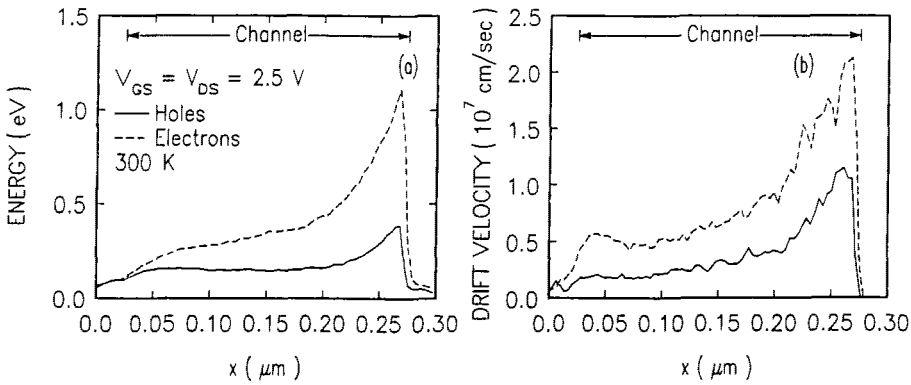

Figure 10. *Hole average energy (a) and x-directed drift velocity (b) at 300 K along the channel 1 nm away from the Si-SiO_2 interface for a 0.25 μm long device. Interparticle collisions are excluded in this case. The electron energy and velocity for a specularly biased n-channel device at 300 K are also shown here for a direct comparison between the electrons and holes behavior.*

a factor 2 lower than in n-channels for the 0.25 μm device, about 20 % lower for the 0.06 μm device at 300 K).

CONCLUSIONS

From the work we have presented it is clear that there is still room for improvement in the semiclassical description of electron transport. The introduction of the full band-structure of the semiconductor, the calculation of scattering rates consistent with the DOS, and the inclusion of short-range and long-range electron-electron interaction are factors which play a major role in controlling the microscopic behavior of short devices. We have shown that these effects can have dramatic consequences in *realistic situations* in submicron Si devices. Band-structure effects, in particular, can be important even at low-fields and have dramatic effects at low temperatures and high biases. Coulomb screening, high-energy transport, quantum-size effects, and interface scattering have been either crudely approximated or ignored in our model. Their effect on our results remains to be determined.

ACKNOWLEDGMENTS

We are indebted to R. Car for the routine to set up and invert the matrix for the pseudopotential band-structure calculations. We wish to thank P. J. Price and C. Jacoboni for suggestions on how to handle the short-range electron-electron interaction, and G. Sai-Halasz and M. Wordeman for providing the transconductance data of their short devices prior to publication. J. Tang and A. Mayo made contributions in the early stages of the implementation of the Poisson-Monte Carlo coupling we have presented. We also acknowledge the assistance received from the IBM T. J. Watson Computing Systems staff.

REFERENCES

Artaki M. and K. Hess (1987). Transient and steady-state electron transport in $GaAs/Al_xGa_{1-x}As$ heterojunctions at low temperatures: The effect of electron-electron interactions. *Phys. Rev. B*, **37**, 2933-2945.

Bank R. E. and D. J. Rose (1980). Parameter selection for Newton-like methods applicable to nonlinear partial differential equations. *SIAM J. Numer. Anal.*, **17**, 806-822.

Bardeen J. and W. Shockley (1950). Deformation Potentials and Mobilities in Non-Polar Crystals. *Phys. Rev.*, **80**, 72-80.

Basu B. K. (1977). High-Field drift velocity of silicon inversion layers – A Monte Carlo calculation. *J. Appl. Phys.*, **48**, 350-353.

Basu B. K. (1978). Monte carlo calculation of hot electrons drift velocity in silicon (100)-inversion layer by including three subbands. *Solid State Comm.*, **27**, 657-660.

Brunetti R. *et al.* (1981). Diffusion Coefficients of Electrons in Silicon. *J. Appl. Phys.*, **52**, 6713-6722.

Brunetti R. *et al.* (1985). Effects of Interparticle Collisions on Energy Relaxation of Carriers in Semiconductors. *Physica B+C*, **134B**, 369-373.

Canali C. *et al.* (1975). Electron drift velocity in silicon. *Phys. Rev. B*, **12**, 2265-2283.

Chu-Hao *et al.* (1985). Monte Carlo Study of Two-Dimensional Electron Gas Transport in Si-MOS Devices. *Solid-State Electron.*, **28**, 733-740.

Cohen M. L. and T. K. Bergstresser (1966). Band Structures and Pseudopotential Form Factors for Fourteen Semiconductors of the Diamond and Zinc-blend Structures. *Phys. Rev.*, **141**, 789-796.

Crowell C. R. and S. M. Sze (1966). Temperature Dependence of Avalanche Multiplication in Semiconductors. *Appl. Phys. Lett.*, **9**, 242-244.

Fang F. F. and A. B. Fowler (1970). Hot Electron Effects and Saturation Velocities in Silicon Inversion Layers. *J. Appl. Phys.*, **41**, 1825-1831.

Fischetti, M. V. and S. E. Laux (1988). Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects. *Phys. Rev. B* (in press).

Gilat G. and L. J. Raubenheimer (1966). Accurate Method for Calculating Frequency-Distribution Functions in Solids. *Phys. Rev.* **144**, 390-395.

Grant, W. N. (1973). Electron and Hole Ionization Rates in Epitaxial Silicon at High Electric Fields. *Solid State Electron*, **16**, 1189-1203.

Harrison W. A. (1956). Scattering of Electrons by Lattice Vibrations in Nonpolar Crystals. *Phys. Rev.*, **104**, 1281-1290.

Hess K. and C. T. Sah (1974). Hot carriers in silicon surface inversion layers. *J. Appl. Phys.*, **45**, 1254-1257.

Hesto P. *et al.* (1985). Monte Carlo modelling of Semicondutor Device, in *Nasecode IV, Proceeding of the Fourth International Conference on the Numerical Analysis of Semiconductor Devices and Integrated Circuits*, edited by J. J. H. Miller, Boole Press, Dublin, p. 315-319.

Hockney R. W. and J. W. Eastwood (1981). *Computer Simulation Using Particles*, McGraw-Hill, New York.

Keldysh L. V. (1965). Concerning the Theory of Impact Ionization in Semiconductors. *Sov. Phys. JETP*, **21**, 1135-1144.

Jacoboni C., R. Minder, and G. Maini (1975). Effect of band non-parabolicity on electron drift velocity in silicon above room temperature. *J. Phys. Chem. Solids*, **36**, 1129-1133.

Jacoboni C. *et al.* (1977). A review of some charge transport properties of silicon. *Solid-State Electron.*, **20**, 77-89.

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

Johnson O. G., C. A. Micchelli, and G. Paul (1983). Polynomial preconditioners for conjugate gradient calculations. *SIAM J. Numer. Anal.*, **20**, 362-376.

Laux S. E. and M. V. Fischetti (1988). Monte Carlo Simulation of Submicron Si MOSFETs at 77 and 300K. *IEEE Electron Device Lett.* (in press).

Lee C. A. *et al.* (1964). Ionization Rates of Holes and Electrons in Silicon. *Phys. Rev.*, **134**, A761-A773.

Lugli P. and D. K. Ferry (1983). Effect of Electron-Electron Scattering on Monte Carlo Studies of Transport in Submicron Semiconductor Devices. *Physica B+C*, **117&118B**, 251-253.

Lugli P. and D. K. Ferry (1985a). Effect of Electron-Electron and Electron-Plasmon Interactions on Hot Carrier Transport in Semiconductors. *Physica B*, **129**, 532-536.

Lugli P. and D. K. Ferry (1985b). Degeneracy in the Ensemble Monte Carlo Method for High-Field Transport in Semiconductors. *IEEE Trans. Electron Devices*, **ED-32**, 2431-2437.

Manzini S.(1985). Effects of Coulomb scattering in n-type silicon inversion layers. *J. Appl. Phys.*, **57**, 411-414.

Matulionis A., J. Pozela, and A. Reklatis (1975). Monte Carlo treatment of electron-electron collisions. *Solid State Comm.*, **16**, 1133-1137.

Moglestue C. (1986). A Self-Consistent Monte Carlo Particle Model to Analyze Semiconductor Microcomponents of any Geometry. *IEEE Trans. Computer-Aided Design*, **CAD-5**, 326-345.

Modelli A. and S. Manzini (1988). High-field drift velocity of electrons in silicon inversion layers. *Solid-State Electron.*, **31**, 99-104.

Nilsson G. and G. Nelin (1972). Study of the Homology between Silicon and Germanium by Thermal Neutron Spectrometry. *Phys. Rev. B*, **6**, 3777-3786.

Ning T. H. and C. T. Sah (1972). Theory of Scattering of Electrons in a Nondegenerate-Semiconductor-Surface Inversion Layer by Surface-Oxide Charges. *Phys. Rev. B*, **6**, 4605-4613.

Ning T. H., C. M. Osburn, and H. N. Yu (1976). Emission probability of hot electrons from silicon into silicon dioxide. *J. Appl. Phys.*, **48**, 286-293.

Ottaviani G. *et al.* (1975). Hole drift velocity in silicon. *Phys. Rev. B* **12**, 3318-3329.

Park Y.-J., T.-W. Tang, and D. H. Navon (1983). Monte Carlo Surface Scattering Simulation in MOSFET Structures. *IEEE Trans. Electron Devices*, **ED-30**, 1110-1116.

Phillips A. and P. J. Price (1977). Monte Carlo calculations on hot electron energy tails. *Appl. Phys. Lett.*, **30**, 528-530.

Pines D. (1963). *Elementary excitations in Solids*, Benjamin, New York.

Price P. J. (1979). Monte Carlo Calculation of Electron Transport in Solids. *Semiconductors and Semimetals*, **14**, 249-308.

Ravaioli U. and D. K. Ferry (1986). Monte Carlo study of the quasi two-dimensional electron gas in the high electron mobility transistor. *Superlattices and Microstructures*, **2**, 75-78.

Ridley B. K. (1977). Reconciliation of the Conwell-Weisskopf and Brooks-Herring formulae for charged-impurity scattering in semiconductors: Third-body interference. *J. Phys. C*, **10**, 1589-1593.

Sai-Halasz G. A. *et al.* (1987). Design and Experimental Technology for 0.1-μm Gate-Length Low-Temperature Operation FETs. *IEEE Electron Device Lett.*, **EDL-8**, 463-466.

Sai-Halasz G. A., *et al.* (1988). High Transconductance and Velocity Overshoot in nMOS Devices at the 0.1-μm Gate-Length Level. *IEEE Electron Device Lett.*, (in press).

Sangiorgi E., B. Riccó, and F. Venturi (1988). MOS^2 : An Efficient *MO*nte Carlo *S*imulator for *MOS* devices. *IEEE Trans. Computer-Aided Design*, **CAD-7**, 259-271.

Shahidi G. S., D. A. Antoniadis, and H. I. Smith (1988). Electron Velocity Overshoot at Room and Liquid Nitrogen Temperatures in Silicon Inversion Layers. *IEEE Electron Device Lett.*, **ED-9**, 94-96

Shichijo H. and K. Hess (1981). Band-structure-dependent transport and impact ionization in GaAs. *Phys. Rev. B*, **23**, 4197-4207.

Stern F. and W. E. Howard (1967). Properties of Semiconductor Surface Inversion Layers in the Electric Quantum Limit. *Phys. Rev.*, **163**, 816-835.

Tang J. Y. and K. Hess (1983a). Impact ionization of electrons in silicon (steady state). *J. Appl. Phys.*, **54**, 5139-5144.

Tang J. Y. and K. Hess (1983b). Theory of hot electron emission from silicon into silicon dioxide. *J. Appl. Phys.*, **54**, 5145-5151.

Throngnumchai K., K. Asada, and T. Sugano (1986). Modeling of 0.1-μm MOSFET on SOI Structure Using Monte Carlo Simulation Technique. *IEEE Trans. Electron Devices*, **ED-33**, 1005-1011.

Tomizawa M., K. Yokoyama, and A. Yoshii (1988). Nonstationary Carrier Dynamics in Quarter-Micron Si MOSFETs. *IEEE Trans. Computer-Aided Design*, **CAD-7**, 254-258.

Tomizawa K. and N. Hashizume (1988). Ensemble Monte Carlo Simulation of an $Al_xGa_{1-x}As$/GaAs Heterostructure MIS-Like FET. *IEEE Trans. Elec. Devices*, **35**, 849.

van Overstraeten R. and H. DeMan (1970). Measurements of the Ionization Rates in Diffused Silicon p-n Junctions. *Solid-State Electron.*, **13**, 583-608.

Venturi F., R. K. Smith, E. Sangiorgi, M. R. Pinto, and B. Riccó (1988). *A Self-Consistent Monte Carlo Simulator for Deep Submicron MOSFETs*, presented at the Workshop on Numerical Modeling of Process and Devices for Integrated Circuits, NUPAD II, San Diego, Ca.

Wang T. and K. Hess (1985). Calculation of the electron velocity distribution in high electron mobility transistors using an ensemble Monte Carlo method. *J. Appl. Phys.*, **57**, 5336-5339.

Yokoyama K. and K. Hess (1986). Calculation of warm electron transport in AlGaAs/GaAs single heterostructures using a Monte Carlo method. *J. Appl. Phys.*, **59**, 3798-3802.

Ziman J. M. (1974). *Electrons and Phonons*, Oxford University Press, Oxford.