Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects

Massimo V. Fischetti and Steven E. Laux
IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598
(Received 12 May 1988)

The physics of electron transport in Si and GaAs is investigated with use of 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 by using the band structure obtained from empirical-pseudopotential calculations. (2) The electron-phonon, electron-impurity, and electron-electron scattering rates are computed in a way consistent with the full band structure of the solid, thus accounting for density-of-states and matrix-element effects more accurately than previous transport formulations. (3) The long-range carrier-carrier interaction and space-charge effects are included by coupling the Monte Carlo simulation to a self-consistent two-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 submicrometer 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 electron-electron interaction in these ultrasmall structures. In the systems considered, the inclusion of the full band structure has the effect of reducing the amount of velocity overshoot via electron transfer to upper conduction valleys, particularly at large biases and low temperatures. The reasonableness of the physical picture is supported by the close agreement of the results of the simulation to available experimental data.

I. INTRODUCTION

If we look at the way electronic transport in solids is currently treated in the context of electron devices, we find a very wide spectrum of approaches. At the "engineering" end of the spectrum, device designers often favor the simplicity, flexibility, and fast computing times of the standard drift-diffusion (DD) equations,[1] despite their limited range of validity. At the opposite end, we find many attempts to formulate a rigorous theory of quantum transport which transcends the semiclassical Boltzmann transport equation (BTE).[2]

Moving from the traditional DD models in the direction of a more complete description of the physics of transport, we find the so-called "hydrodynamic" or "energy-transport" models,[3] employing higher moments of the BTE. This improvement of the DD model accounts for modest amounts of carrier heating and nonlocal phenomena of importance in some of the smallest devices currently mass produced. For the submicrometer devices of the next generations, now under development, this approach might also be inadequate, and the Monte Carlo (MC) technique to solve the BTE has gained increasing popularity.[4–6] Despite its appreciable computational cost, its simplicity of implementation and its relatively complete description of semiclassical transport have rendered MC simulations appealing and successful. Still, the physics of "state-of-the-art" MC device simulations (based almost exclusively on the models presented in Ref. 5) might not be accurate enough to handle the submicrometer (<0.25 μm) devices now experimentally available. We can classify very crudely the problems which affect present MC simulations into three groups.

Problems in the first group stem from the quantum size phenomena occurring in devices (or regions of devices) of a scale length comparable to the de Broglie wavelength of the carriers: Quantum wells, inversion and accumulation layers, heterostructures, and tunneling phenomena are a few common examples. These quantum effects are relatively well understood and have indeed been coupled to semiclassical MC simulations in the past.[7] If we are willing to understate the issue, we can say that they do not constitute a major deviation from the BTE and the corresponding MC simulations, apart from numerical difficulties.[8] Still, they constitute an important element in some devices.

A second class of problems is related to the intrinsic limitations of the BTE at high electric fields, high carrier energies, and very short lifetimes.[2] A quantitative estimate of the relevance of these problems to devices which will be manufactured in the foreseeable future is still lacking. Actually, the issue is still controversial.[9]

Finally, there is a third class of problems to which little attention has been paid: Even within a semiclassical formulation, the physical models employed in present MC device simulations can be and must be improved. It might suffice to recall the rather simple band structure commonly used which is appropriate only in low-field ($\lesssim 10^5$ V/cm) and low-energy ($\lesssim 0.2$–0.5 eV) situations, but which is often extended outside this range. In our opinion, the gap between present MC models and formulations of quantum transport beyond the BTE is very wide for nearly all devices of current technological importance. In other words, we must learn and test more

semiclassical physics before it will become essential to treat quantum-transport issues in real devices (quantum size and tunneling effects apart, as we said above). We believe this is a necessary prerequisite for establishing the real limitations of the BTE.

Working from these motivations and limiting our attention to the case of electron transport only, we believe that a major problem with the present MC device simulations is their aforementioned inability to model correctly the transport of carriers whose kinetic energy is so large that a parabolic (or a conventional first-order $\mathbf{k}\cdot\mathbf{p}$ nonparabolic approximation) description of the semiconductor conduction band appears unjustified. This problem has immediate relevance to practical issues arising in devices already in advanced development, such as quarter-micrometer Si field-effect transistors (FET): impact ionization coefficients in Si (Ref. 10) and GaAs (Ref. 11), substrate currents in metal-oxide-semiconductor field-effect transistors (MOSFET's),[12] and injection into the gate insulator of Si MOSFET's (Refs. 13 and 14) (and the associate SiO_2 reliability problems) or into the $Al_xGa_{1-x}As$ layer of modulation-doped III-V-compound FET's (MODFET's), to mention the most important.

The strong role of the band structure and the importance that the correct density of states (DOS) has on the electron-lattice scattering rates have been already pointed out by Shichijo and Hess[11] and Tang and Hess.[10] Also Al-Omar and Krusius have recently improved the description of the band structure.[15] Essentially, we have extended these works, including the band-structure effects, not only on the kinematics, but also on the dynamics of carrier transport by computing the electron scattering rates in a way consistent with the full band structure of the lattice. We also include the short-range carrier-carrier interaction in an ensemble MC simulation. Furthermore, by moving the ensemble of particles in the electrostatic field obtained by solving self-consistently the Poisson equation in two dimensions, we account for the long-range Coulomb interaction together with space-charge effects. The self-consistency is retained up to frequencies large enough to account for the plasma oscillations, particularly important in highly doped semiconductor regions. Thus, we are able to handle small structures at high biases with a degree of confidence greater than ever before. We must again stress that quantum effects are ignored in the present work and that some degree of uncertainty obviously affects our predictions. To "limit the damage" as far as quantum size effects are concerned, we have so far focused our work on high-bias situations where channel-quantization effects are, we hope, of minor importance. Regarding high-field quantum effects, the good agreement of the predictions of our physical model with experimental data suggests that, as already argued, the range of validity of the BTE may have been underestimated in the past.[16]

This paper is organized as follows. In Sec. II we describe how the semiconductor band structure is embedded into our MC program kinematically (i.e., how electrons move during free flights) and dynamically (i.e., how scattering rates are computed and final states are selected). Particular attention is paid to the electron-electron interaction, as this is a notoriously troublesome issue in MC simulations.[5] In Sec. III we discuss the procedure we have employed to "calibrate" the few adjustable scattering parameters of our Monte Carlo models and present results relative to homogeneous, steady-state, bulk transport in Si and GaAs. A description of the numerical technique employed to couple the particle model to the electrostatic field is given in Sec. IV and some details on the actual program are presented in Sec. V. Finally, in Sec. VI we present the results of the simulation of small Si MOSFET's, highlighting some of the main features of electron transport in structures as small as 60 nm, such as the effect of the higher conduction bands on the electron energies and velocities at large biases and 77 K.

II. THE MONTE CARLO MODEL

The Monte Carlo technique we employ to treat electron transport is conceptually identical to the "state-of-the-art" technique described in the excellent reviews by Price[4] and Jacoboni and Reggiani.[5] 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[10,11] in simple cases. 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 $\mathbf{k}$ 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 electron states after collisions. We now discuss these issues in turn.

A. Band structure

The empirical pseudopotentials given by Cohen and Bergstresser[7] 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 a reliable description of the DOS. We show in Figs. 1 and 2 the band structure and DOS of Si and GaAs we have used. In Fig. 1(b) we compare the parabolic DOS of Si to the pseudopotential DOS to show the different magnitude and shape at high electron energies.

For the purpose of our MC simulation, we have generated a mesh of 916 $\mathbf{k}$ points in the $\frac{1}{48}$ 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. 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 an electron with an arbitrary wave vector $\mathbf{k}$ in band ν, we should first translate $\mathbf{k}$ 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 electron group velocity in order to obtain its correct orientation over the entire BZ. This last operation, performed over and over during the run, requires an excessive central-processing-unit (CPU) time. As a general rule, an increase in storage requirement can usually be traded off for an enhanced program speed. In this circumstance, we have found that storing the band-structure information over about 41 000 points in the entire BZ (and a few points outside for interpolation requirements) increased significantly the speed of the program by avoiding these symmetry transformations. Therefore, given a wave vector $\mathbf{k}$, we find the associated energy in band ν by first finding the eight corners $\{\mathbf{k}_\lambda\}$ $(\lambda=1,2,\ldots,8)$ of the cubic element of side length $l=0.05(2\pi/a)$ in the BZ to which $\mathbf{k}$ belongs, expanding the energy quadratically around each corner, i.e.,

$$E_{\nu,\lambda}(\mathbf{k})=E_\nu(\mathbf{k}_\lambda)+\frac{\partial E_\nu(\mathbf{k}_\lambda)}{\partial k_i}(k_i-k_{i,\lambda}) + \frac{1}{2}\frac{\partial^2 E_\nu(\mathbf{k}_\lambda)}{\partial k_i \partial k_j}(k_i-k_{i,\lambda})(k_j-k_{j,\lambda}) , \quad (1)$$

where sums over identical indices must be performed, and finally adding up the contributions from each corner with the appropriate weights,

$$E_\nu(\mathbf{k})=\sum_{\lambda=1}^{8} W_\lambda E_{\nu,\lambda}(\mathbf{k}) , \quad (2)$$

where the weights are given by

$$W_\lambda=\left[1-\frac{k_x-k_{x,\lambda}}{l}\right]\left[1-\frac{k_y-k_{y,\lambda}}{l}\right]\times\left[1-\frac{k_z-k_{z,\lambda}}{l}\right] .$$

The velocity at $\mathbf{k}$ is obtained in a similar way by interpolating linearly around each corner. This interpolation scheme, exact for parabolic bands, was found to be the best compromise between accuracy and simplicity.

Much more complicated is the task of inverting the

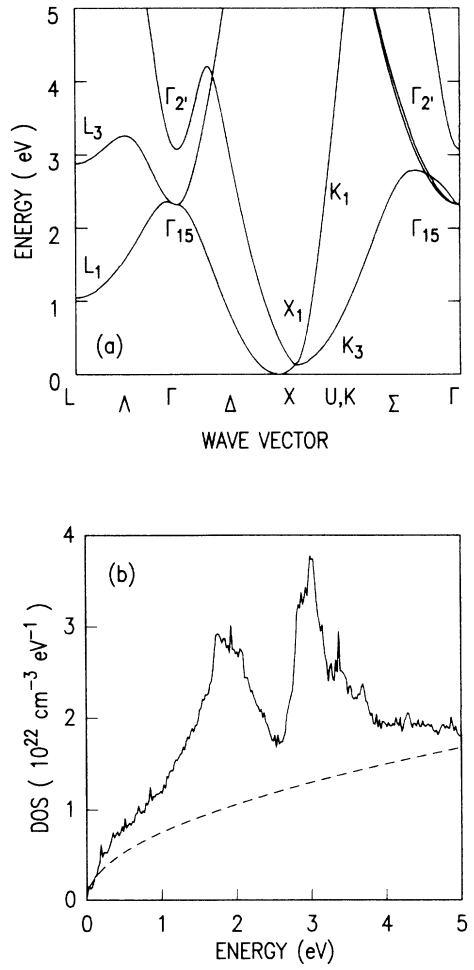

FIG. 1. (a) Band structure and (b) density of states for Si obtained from the empirical-pseudopotential calculation. The dashed line in (b) corresponds to the density of states obtained considering six ellipsoidal parabolic valleys. A spin-degeneracy factor 2 has been included in both the pseudopotential and the parabolic density of states.

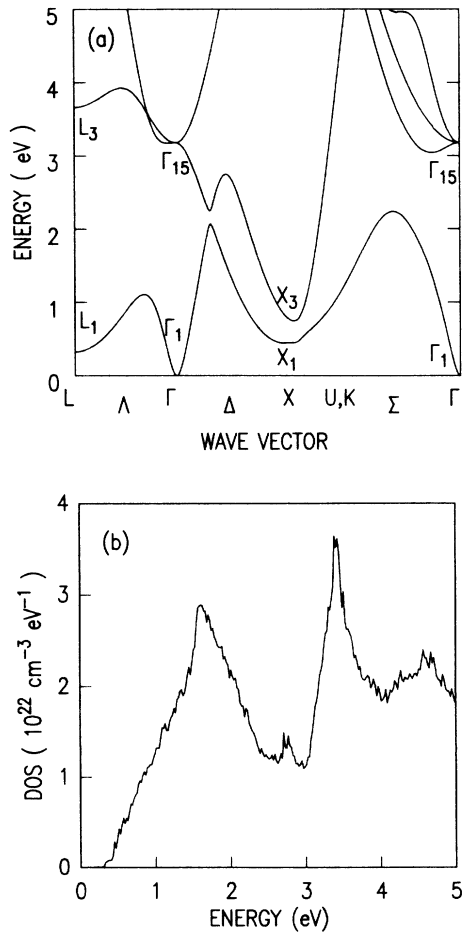

FIG. 2. (a) Band structure and (b) density of states for GaAs obtained from the empirical-pseudopotential calculation.

dispersion $E_\nu(\mathbf{k})$. This must be done very frequently in the simulation, since we must find a "new" wave vector $\mathbf{k}'$ corresponding to the final energy E' every time we have to select a final electron state after any collision event. The algorithm we use consists of searching through the first BZ for the cubes which intersect the constant-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 $l/2$ 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 equienergy surface. In each of them a particular $\mathbf{k}$ is chosen along the principal directions of the cube (edges, side diagonals, and cube diagonal) by inverting Eq. (2) up to third order, the coefficients needed to invert Eq. (2) having been previously stored in a look-up table. This guarantees that the selected $\mathbf{k}$ vectors will correspond to the desired energy E' within an average variance of 4 meV. Of course, simply selecting the central wave vector in each of the small cubes, as done by Hess and co-workers,[10,11] would improve efficiency. However, we found that the average error would be unacceptably large, up to several tens of meV in some regions of the BZ.

B. 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.[5] However, when a numerical representation of the bands is used, the higher efficiency of the self-scattering algorithm vanishes since the equations of motion

$$\frac{d\mathbf{r}}{dt}=\frac{1}{\hbar}\nabla_{\mathbf{k}}E_\nu(\mathbf{k})\ , \tag{3a}$$

$$\frac{d\mathbf{k}}{dt}=\frac{e}{\hbar}\nabla_{\mathbf{r}}\phi(\mathbf{r})=-\frac{e\mathbf{F}(\mathbf{r})}{\hbar} \tag{3b}$$

[where e is the magnitude of the electron charge, $\phi(\mathbf{r})$ is the electrostatic potential, and $\mathbf{F}$ the electric field at the electron position $\mathbf{r}$] cannot be integrated analytically. 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} for the numerical integration of Eqs. (3a) and (3b) over a free flight. A second-order Runge-Kutta scheme[18] was selected as a good compromise between efficiency (lower-order schemes may require a smaller number of interpolations per time step) and accuracy. This scheme provides excellent numerical stability over simulation times exceeding tens of psec for time steps of the order of 10^{-16} sec at the higher field gradients present in our simulations. Using time steps of this magnitude, an accuracy of better than 1 μeV per free flight is guaranteed in the worst case. The magnitude of the scattering rates poses an additional constraint on the highest possible value of Δt_{bal}, as discussed in Sec. IV B.

A final consideration concerning the integration of the equations of motion (3) is the possibility of band crossing, which occurs, for example, at the X symmetry point in Fig. 1(a), or approaching the Γ point from the X point. Such an occurrence is dealt with by imposing continuity of the electron group velocity in a free flight. At highly degenerate points of zero group velocity (such as the Γ point), the continuity of the second derivative (effective mass) is imposed.

C. 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 Refs. 10 and 11. However, we have extended this approach down to very low electron energies (at least in the case of Si) 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 bands behave quite differently from a parabolic-band representation at energies around 130 meV. In GaAs, a similar difference can be expected at the X valley whose minimum is not exactly at the X symmetry point but close to $0.9(2\pi/a)$ along the symmetry line Δ (Ref. 19). Therefore, the different kinematics (via the group velocities) and dynamics (via the different DOS) of these regions will be well represented by our approach.

We should stress that even accounting for the "exact" band structure does not free us from serious difficulties. The choice of the matrix elements for the electron-phonon coupling,[20] the determination of the screening length for the Coulomb electron-electron[21] and electron–ionized-impurity collisions,[22] and the complex structure of the impact-ionization double matrix element,[23] among other issues, remain unresolved problems in our work. We have bypassed these difficulties by adopting various empirical, but system-independent, approaches, as we will discuss.

1. *Electron-phonon scattering*

The nonpolar scattering rate, $1/\tau_{\eta,\nu}(\mathbf{k})$, between an electron of wave vector $\mathbf{k}$ in the νth band and a phonon of type (acoustic or optical) and polarization (transverse or longitudinal) η has been calculated from the Fermi golden-rule expression:[20]

$$\frac{1}{\tau_{\eta,\nu}(\mathbf{k})}=\sum_{\nu',\mathbf{q}}\frac{\pi}{\rho\omega_{\eta,\mathbf{q}}}\Delta_{\eta,\nu'}(\mathbf{q})^2\,|\,\mathcal{I}(\nu,\nu';\mathbf{k},\mathbf{k}')\,|^2 \times\delta(E_\nu-E'_{\nu'}\mp\hbar\omega_{\eta,\mathbf{q}})(n_{\eta,\mathbf{q}}+\tfrac{1}{2}\pm\tfrac{1}{2})\ , \tag{4}$$

while in a polar semiconductor the rate for the polar collisions with longitudinal-optical phonons, $1/\tau_{\text{LO}}(\mathbf{k})$—usually restricted to the Γ valley—is given by

$$\frac{1}{\tau_{\mathrm{LO}}(\mathbf{k})}=\frac{2\pi}{\hbar}\sum_{\mathbf{q}}\frac{e^2F^2}{q^2}\,|\,\mathcal{I}(\mathbf{k},\mathbf{k}')\,|^2\,\delta(E-E'\mp\hbar\omega_{\mathrm{LO}})\times(n_{\mathrm{LO}}+\tfrac{1}{2}\pm\tfrac{1}{2})\,. \tag{5}$$

In these formulas 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 electron 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 polar coupling constant F is given by the usual Fröhlich expression:[24]

$$F^2=\frac{\hbar\omega_{\mathrm{LO}}}{4}\left[\frac{1}{\epsilon_\infty}-\frac{1}{\epsilon_0}\right],$$

where ϵ_∞ and ϵ_0 are the optical and static dielectric functions, respectively. The sum in Eq. (4) 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 wave vectors $\mathbf{q}$, 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.

The numerical integration over the energy-conserving δ function is done using an algorithm proposed by Gilat and Raubenheimer.[25] First, we select all "final" cubes centered around points $\mathbf{k}'_m$ in the fine mesh which intersect the surfaces $E_{\nu'}(\mathbf{k}')=E'$, for all bands ν'. In each cube the equienergy surface is approximated by the surface formed by slicing the cube with the plane normal to $\nabla_{\mathbf{k}}E_{\nu'}(\mathbf{k}'_m)$, displaced from the cube center by the amount $[E_{\nu'}(\mathbf{k}'_m)-E']/|\nabla_{\mathbf{k}}E_{\nu'}(\mathbf{k}'_m)|$ along the direction of the gradient. The area of this surface is proportional to the DOS at energy E' in band ν' in the cube, $\mathcal{D}_{\nu'}(E',\mathbf{k}'_m)$. Here and in the followng a factor of 2 is included to account for spin degeneracy. The nonpolar matrix element $\Delta_{\eta,\nu'}(\mathbf{q}_m)$, where $\mathbf{q}_m=\pm(\mathbf{k}-\mathbf{k}'_m+\mathbf{G})$, is then evaluated in the approximation described below and the scattering rate in Eq. (4) is obtained as

$$\sum_{\nu',m}{}'\frac{\pi}{\rho\omega_{\eta,\mathbf{q}_m}}\,|\,\Delta_{\eta,\nu'}(\mathbf{q}_m)\,|^2\,|\,\mathcal{I}(\nu,\nu';\mathbf{k},\mathbf{k}'_m)\,|^2\times\mathcal{D}_{\nu'}(E',\mathbf{k}'_m)(n_{\eta,\mathbf{q}_m}+\tfrac{1}{2}\pm\tfrac{1}{2})\,, \tag{6}$$

the primed sum over the cube indices m meaning that only energy-conserving vectors $\mathbf{k}'_m$ in the BZ must be considered. The result for each process (phonon type, emission or absorption, and total electron-phonon scattering rate) is then stored in a look-up table, together with the rate gradients. During the Monte Carlo simulation an interpolation is made, as done for the energy and velocity. A similar procedure is used for the polar rate of Eq. (5), which can be extended outside the Γ valley, if needed.

In principle, the matrix element $\Delta_{\eta,\nu'}(\mathbf{q})$ and the overlap integral $\mathcal{I}(\nu,\nu';\mathbf{k},\mathbf{k}')$ can be obtained from the pseudopotential theory in the long-wavelength (small-q) limit. This, however, would not help us in the opposite limit which dominates the carrier transport at high energies. In the absence of better information about the deformation potentials, the nonpolar electron-phonon matrix element has been approximated by an isotropic coupling constant $\Delta_\eta q$ for longitudinal-acoustic (LA) and transverse-acoustic (TA), or $(\Delta K_{\mathrm{op}})_\eta$ for longitudinal-optical (LO) and transverse-optical (TO) phonons. The overlap integral has been approximated by the rigid-ion expression[20] in the numerical range, ignoring the band-index dependence.

The acoustic phonon dispersion has been approximated by

$$\hbar\omega_{\eta,\mathbf{q}}=\begin{cases}\hbar\omega_{\eta,\max}\left[1-\cos\left[\dfrac{qa}{4}\right]\right]^{1/2}, & q\le 2\pi/a\\ \hbar\omega_{\eta,\max}, & q>2\pi/a\,.\end{cases} \tag{7}$$

The maximum phonon frequency $\omega_{\eta,\max}$ has been chosen to be $4c_\eta/a$, c_η being the sound velocity with polarization η. This expression underestimates the phonon energy at small q, $\hbar\omega_{\eta,\mathbf{q}}\sim\hbar c_\eta q$, by a factor of $\sqrt{2}$, but it provides a very good approximation of the zone-edge energies, as obtained from the spectra of Ref. 26, 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 of $\sqrt{2}$. This will be compensated by smaller coupling constants Δ_η, as discussed in Sec. III. The dispersion of the optical phonons has been ignored, as implied in Eq. (5), and their energy has been taken from Refs. 5 (Si) and 27 (GaAs).

Before discussing specific values of the various parameters, we shall complete the discussion of the Monte Carlo technique by examining the other scattering processes and the selection of the final state after collision.

2. *Electron-impurity scattering*

The scattering rate $1/\tau_{\mathrm{imp},\nu}(\mathbf{k})$ for the collision suffered by an electron of wave vector $\mathbf{k}$ in the νth band in the screened Coulomb field of an ionized dopant has been computed starting from the Brooks-Herring (BH) formula[28] corrected for the band-structure effects:

$$\frac{1}{\tau_{\mathrm{BH},\nu}(\mathbf{k})}=\frac{N_{\mathrm{dop}}Z^2e^4}{4\pi^2\hbar\epsilon^2}\sum_{\nu',\mathbf{k}',\mathbf{G}}\frac{|\,\mathcal{I}(\nu,\nu';\mathbf{k},\mathbf{k}')\,|^2}{(\beta_s^2+|\,\mathbf{k}-\mathbf{k}'+\mathbf{G}\,|^2)^2}\times\delta(E_\nu(\mathbf{k})-E_{\nu'}(\mathbf{k}'))\,, \tag{8}$$

where ϵ is the static dielectric, N_{dop} is the concentration of ionized dopants, eZ their charge, and the screening parameter β_s has been obtained in the Debye approximation:[29]

$$\beta_s(\mathbf{r},t)=\left[\frac{e^2n_{\mathrm{el}}(\mathbf{r},t)}{\epsilon k_BT_{\mathrm{el}}(\mathbf{r},t)}\right]^{1/2}, \tag{9}$$

where k_B is the Boltzmann constant. In Eq. (8) 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 electron density n_{el} as a function of time and position and the average electron energy $\bar{E}$, which is then converted to an effective temperature $T_{el}=2\bar{E}/3k_B$. These values are then used to compute the screening parameter β_s in a self-consistent way.

Ridley's *statistical screening* model is then employed to compute the final rate:[30]

$$\frac{1}{\tau_{\text{imp},\nu}(\mathbf{k})}=\frac{v_g(\mathbf{k})}{d}\left[1-\exp\left[-\frac{d}{v_g(\mathbf{k})\tau_{\text{BH},\nu}(\mathbf{k})}\right]\right], \quad (10)$$

where $v_g(\mathbf{k})$ is the electron group velocity and $d=(2\pi N_{\text{dop}})^{-1/3}$ is the average distance between the ions.

The rate (10) must be calculated during the Monte Carlo simulation, since its dependence on many variables $(n_{el},T_{el},N_{\text{dop}})$ with a wide dynamic range would imply an unmanageable size for a look-up table and too many CPU-time-consuming interpolations.

3. Electron-electron scattering

As is well known, the short-range electron-electron interaction has often been found difficult to treat in Monte Carlo simulations. The golden-rule expression for the total rate, $1/\tau_{ee,\nu}(\mathbf{k})$, for the screened collision suffered at time t by an electron at position $\mathbf{r}$ and of wave vector $\mathbf{k}$ in the νth band with any other electron in the system can be obtained in the Born approximation as

$$\frac{1}{\tau_{ee,\nu}(\mathbf{k},\mathbf{r},t)}=\frac{e^2}{8\pi^5\hbar\epsilon^2}\sum_{\nu',\mu,\mu',\mathbf{k}',\mathbf{p},\mathbf{p}'}\frac{|\mathcal{I}(\mu,\mu';\mathbf{p},\mathbf{p}')|^2\,|\mathcal{I}(\nu,\nu';\mathbf{k},\mathbf{k}')|^2}{(\beta_s^2+|\mathbf{k}-\mathbf{k}'+\mathbf{G}|^2)^2}\delta(E_{\text{tot}})\delta_{\mathbf{K}}f(\mathbf{r},\mathbf{p},t)\ . \quad (11)$$

The sum extends over all final states $\mathbf{k}'$, over the distribution $f(\mathbf{r},\mathbf{p},t)$ of "partner" electrons at $\mathbf{r}$ with wave vector $\mathbf{p}$ at time t, over the possible final states $\mathbf{p}'$ of the partners, and over all possible bands, as allowed by conservation of total energy,

$$E_{\text{tot}}=E_\nu(\mathbf{k})+E_\mu(\mathbf{p})-E_{\nu'}(\mathbf{k}')-E_{\mu'}(\mathbf{p}')=0\ ,$$

and momentum,

$$\mathbf{K}=\mathbf{k}+\mathbf{p}-\mathbf{k}'-\mathbf{p}'+\mathbf{G}=0\ .$$

Here $\mathbf{G}$ 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 paper by Jacoboni and Reggiani gives a detailed account of this issue.[5]

The following considerations help in finding a solution to the problem: At time t during an ensemble MC simulation, the distribution function at a given position at time $t-\Delta t_{\text{bal}}$ 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. In this way we can get a statistical estimate of the function f to compute the rate (11). 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 averaged quantities are negligible and a homogeneous situation exists within the distance R from the given particle. Thus, the statistical sample of the function $f(\mathbf{r},\mathbf{p},t)$ obtained by looking at all other particles in this region can be considered to be "local" in a sufficiently accurate way. On the other hand, if the distance R is too small, the necessity of simulating a finite number of particles will render the number of partners within the distance R too small to provide a meaningful statistical sample of the distribution function. A further constraint arises when the MC particle simulation is coupled self-consistently to the space charge: since the long-range Coulomb electron-electron 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 the Poisson equation would result in double-counting the long-range coupling.[31] Unfortunately, there is no clear cutoff for the long-range interaction handled by the Poisson equation, since it depends on the local mesh size (for a nonuniform mesh) and on the algorithm chosen to map the discrete charges onto the mesh modes. Thus, any choice of R will result in some amount of "double-counting" in some regions and in some underestimation of the short-range interaction in other regions. We shall clarify this issue in Sec. IV B and argue that this is not a serious problem in our case. Finally, a numerical bottleneck is given by the number of operations to be performed in the search for neighbors: a straightforward search would grow as the square of the number of particles in the ensemble, n_{prt}. Typically, $n_{\text{prt}}\sim 10^4$, too large for the simple search to be practical.

The solution to this last problem is given by a technique discussed by Hockney and Eastwood.[6] During a MC simulation, a list is ultimately constructed containing, for every particle, pointers to the indices of the neighbors within the prefixed distance R. The list is updated at every free-flight. The fast updating of this list requires the layout of a uniform mesh (called the *chaining mesh*) for fast location of the particles. The net effect is that the number of operations required to ascertain neighbors within a distance R grows only as n_{prt}. We limit R to less than about 20 nm. At this distance (corresponding to the screening length at densities of the order of 10^{17} cm^{-3}) we turn off the short-range electron-electron interaction. This is justified since the Coulomb

interaction at these large distances is already accounted for by the self-consistent particle-Poisson coupling. Whenever the carrier-carrier scattering rate has to be computed for a particle, the neighbor list is used to look for those particular partners closer than the local screening length from the particle under consideration. The resulting "screening circle" is actually a set of squares (typically of 2-nm-long sides) whose area A is as close as possible to the area of the screening circle, determined in the self-consistent way described in the context of the electron-impurity scattering [see Eq. (9)]. In the limit of large $n_{\rm prt}$, this provides a good sampling of the local distribution function in regions where the electron density and energy do not exhibit large gradients. If large gradients occur, the function f is not sampled strictly in a local way and the determination of the screening parameter itself is affected by errors. At present, we cannot estimate quantitatively the effect of this approximation. In our simulations, problems might arise in the junction regions, but not in the "active" regions of the devices, such as the channels of FET's.

Density fluctuations in a two-dimensional simulation are of some concern. The probability of finding partners within the screening circle must be rescaled to account for the fact that, given that $\mathcal{N}_{\rm 2D}$ partner particles are found within the screening circle in two dimensions, the number of partner electrons $\mathcal{N}_{\rm 3D}$ within the screening sphere in three dimensions is given by

$$\mathcal{N}_{\rm 3D}=\frac{4\pi s}{3\beta_s^3 A}\mathcal{N}_{\rm 2D}\ , \tag{12}$$

where s is the scale factor determining how many electrons 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. If too few particles are used in the ensemble, large density fluctuations will result in the simulation. In fact, the factor s will be very large in this case and fluctuations of $\mathcal{N}_{\rm 2D}$ will result in wildly amplified nonphysical fluctuations of $\mathcal{N}_{\rm 3D}$. Therefore, both the short-range interparticle 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 large $n_{\rm prt}$, or, equivalently, of small s factors. In practice, a reasonable compromise is reached when $s<\beta_{s,\rm min}$, where $\beta_{s,\rm min}$ is the minimum screening parameter which can be found in the particular simulation to be performed ($\approx 10^8$ m^{-1}, while $s\simeq 5\times 10^7$ m^{-1} in the simulations we have performed to date).

We are now ready to treat the short-range electron-electron collisions: When the scattering rate is needed for an electron at position $\mathbf{r}$ of wave vector $\mathbf{k}$ in band ν, one among its neighbors within the screening circle centered at $\mathbf{r}$ is selected randomly. Denoting by $\mathbf{p}$ the wave vector of this partner and by μ its band, we evaluate the scattering rate for this pair as

$$\frac{1}{\tau_{ee,\nu,\mu}(\mathbf{k},\mathbf{p})}=\frac{\mathcal{N}_{\rm 3D}}{\frac{4}{3}\pi\beta_s^{-3}}\frac{e^2}{8\pi^5\hbar\epsilon^2}\sum_{\nu',\mu',\mathbf{k}',\mathbf{p}'}\frac{|\mathcal{J}(\mu,\mu';\mathbf{p},\mathbf{p}')|^2\,|\mathcal{J}(\nu,\nu';\mathbf{k},\mathbf{k}')|^2}{(\beta_s^2+|\mathbf{k}-\mathbf{k}'+\mathbf{G}|^2)^2}\delta(E_{\rm tot})\delta_{\mathbf{K}}\ , \tag{13}$$

by using a trivial (but numerically tedious, time consuming, and quite complicated because of the double search over final states) extension of the technique employed to compute the electron-phonon rates. As discussed by Matulionis *et al.*,[32] if the ensemble is large enough the rate given by Eq. (13) is a good statistical approximation of the full rate (11), since all neighbors will be sampled randomly during the simulation.

During the evaluation of Eq. (13), all the possible pairs $(\mathbf{k}',\mathbf{p}')$ which satisfy energy and momentum conservation are stored, so that the selection of the final states after collision can be made more efficiently.

One extra numerical "trick" is used to speed up the computation. Equation (13) should be evaluated for every particle at every time step. But only a small fraction of particles will actually suffer electron-electron collisions. Therefore we use a sort of self-scattering technique: an upper bound to the rate given by Eq. (13) is evaluated by using a proper multiple of the rate obtained in the parabolic-band approximation. Thus, we need to evaluate the fully numeric rate (13) only for particles which are selected by this "parabolic" scattering.

4. *Impact ionization*

A simple Keldysh formula is used to derive the rate for impact ionization for an electron of energy E (Ref. 33):

$$\frac{1}{\tau_{ii}(E)}=\begin{cases}0, & E\le E_{\rm th}\\ \dfrac{P}{\tau_{\rm op}(E_{\rm th})}\left[\dfrac{E-E_{\rm th}}{E_{\rm th}}\right]^2, & E>E_{\rm th}\end{cases} \tag{14}$$

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

This is a very simple formula, inconsistent with the band-structure approach we have followed. In particular, we do not believe that long standing issues—such as (1) whether a soft threshold (i.e., low P, low $E_{\rm th}$), or a hard threshold (large P, large $E_{\rm th}$), is a better description, or (2), the controversial issue of the orientation dependence of the ionization coefficients in GaAs (Refs. 11 and 34)—can be resolved with this simple approximation. A soft threshold seems to be more reasonable, but this must be coupled to realistic anisotropic thresholds and DOS available to the recoil and generated particles. In Sec. III B we shall mention a failure of Eq. (14) in the case of GaAs. For the time being we shall manage to fit empirically the Si and GaAs ionization coefficients as best as we can and avoid the simulation of phenomena depending crucially on the ionization rate.

5. *Degeneracy effects*

Lugli and Ferry have proposed a Monte Carlo Technique to account for degeneracy effects in Monte Carlo simulations.[35] Their algorithm has been implemented for homogeneous, steady-state situations. However, in space- and time-dependent simulations, tabulating the distribution function at various locations and times requires an unmanageable amount of storage. Therefore we looked for a simpler, albeit approximate, method. Degeneracy effects are important in heavily doped regions. In these regions the large charge density yields very low fields and negligible carrier heating. The strong electron-electron interaction is also very efficient in distributing energy among the carriers and driving them towards a Fermi-Dirac distribution.[36] Therefore, we approximate the distribution function f as

$$f_{\rm app}(E,\mathbf{r},t)\simeq\frac{1}{1+\exp\left[\dfrac{E-E_F(\mathbf{r},t)}{k_B T_{\rm el}(\mathbf{r},t)}\right]}\ , \tag{15}$$

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

$$1-f_{\rm app}[E_{\nu'}(\mathbf{k}'),\mathbf{r},t]\le\xi\ , \tag{16}$$

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.

D. 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 equienergy 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 $\mathcal{D}_{\nu'}(E',\mathbf{k}'_m)$, 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 $\mathbf{q}_m=\mathbf{k}-\mathbf{k}'_m+\mathbf{G}$. This is obtained from the $\mathbf{q}$ dependence of Eqs. (4) and (5), or, in the case of impurity scattering, Eq. (8), and electron-electron interaction, Eq. (13), from the dependence on $\mathbf{k}-\mathbf{k}'+\mathbf{G}$. In the case of carrier-carrier collisions, the second overlap integral can be evaluated at once, since to every $(\mathbf{k}'_m,\nu')$ there corresponds an associated $(\mathbf{p}'_m,\mu')$ uniquely defined by energy and momentum conservation. These pairs, as we mentioned above, are stored during the evaluation of Eq. (13), so that no further time is spent in a search for $(\mathbf{p}'_m,\mu')$. A random vector $\mathbf{k}'_{\bar{m}}$ is then selected, with probability given by its weight. The rejection technique[5] is employed for this random section. The final-band index and k vector are then known. A final step is necessary, as the energy associated with this wave vector 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 in Sec. II A.

In the case of impact ionization a much simpler approximation is made, consistent with the simplicity of Eq. (14). The recoil electron is assigned an energy $E_\nu(\mathbf{k})-E_{\rm gap}$, where $E_{\rm gap}$ is the band gap of the semiconductor, and a random wave vector at this energy is selected. This accounts very poorly for DOS effects. Moreover, no account is made for the matrix-element and overlap-integral effects. The generated particle is placed at the bottom of the conduction band.

III. HOMOGENEOUS TRANSPORT

The electron-phonon coupling constants entering the rates (4) could be obtained from first-principles calculations. But at high energy above the minimum of the first conduction band, we lack self-consistent-pseudopotential or even simpler tight-binding estimates of dilation coefficients and deformation potentials (even if we were willing to take this concept seriously much above the band minima). Considering our present inability to describe the variations of the electron-phonon matrix elements Δ_η over the various bands in the BZ, we take a very empirical approach and treat 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 transport data fitted to Monte Carlo simulations. 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 seach for this simplicity, we shall make many crude approximations. We now discuss Si and GaAs in turn. In the following two subsections impurity and electron-electron scattering are ignored.

It is important to stress that these coupling constants are determined *uniquely* by bulk, steady-state transport data. Therefore, the device-modeling results we shall present in the following sections are to be considered transport-parameter-free.

A. Silicon

The simplest possible choice we can make is a unique acoustic Bardeen-like deformation potential,[37] $\Delta_{\rm ac}(\mathbf{q})=\Delta_{\rm LA}q=\Delta_{\rm TA}q$, for both LA and TA phonons and a

unique nonpolar-optical Harrison-like deformation potential,[38] $\Delta_{op}(\mathbf{q})=(\Delta K)_{op}$, for both LO and TO phonons. Also, considering the small energy difference of the two optical models, we shall consider LO phonons only.

This simple choice violates some elementary considerations, such as the vanishing of Δ_{TA} at the bottom of the Γ valley, the anisotropy at the bottom of the valleys along the symmetry line Δ, and effects related to other selection rules which would be active in various regions of the BZ (Ref. 5). In order to account for these effects, local in $\mathbf{k}$ space, we should arbitrarily define the boundary of the valleys, such as the maximum energy at which the nonsphericity of the Γ valley could be ignored, or the maximum energy at which a single electron transverse mass in the X valleys can be used. This would increase the number of available parameters, but we doubt that a more meaningful description would be obtained.

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.[39] 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 and the probability of emission into SiO_2 (Ref. 40), exactly as done in Ref. 41.

After many laborious attempts, we found a possible set of parameters which reproduces the desired results:

$$\Delta_{LA}=\Delta_{TA}=\begin{cases}1.2\text{ eV} & \text{(band 1)}\\ 1.7\text{ eV} & \text{(higher bands)}\end{cases}, \tag{17a}$$

$$(\Delta K)_{op}=\begin{cases}1.75\times10^8\text{ eV/cm} & \text{(band 1)}\\ 2.10\times10^8\text{ eV/cm} & \text{(higher bands)}\end{cases}, \tag{17b}$$

$$E_{th}=1.2\text{ eV}, \tag{17c}$$

$$\frac{P}{\tau_{op}(E_{th})}=10^{11}\text{ sec}^{-1}. \tag{17d}$$

It must be stressed that this set is by no means uniquely determined. Many other sets were found which yielded the desired agreement with the velocity-field curves and the ionization-coefficients data. The information which restricts tremendously the range of possible parameters is the injection into SiO_2. Here, the controversial issue of image-force barrier lowering at the Si-SiO_2 interface,[42] the absence of tunneling in the simulation, and the lack of confidence on the use of the Keldysh formula render the fitting procedure somewhat uncertain. Therefore, we expect that additional experimental results on the shape and magnitude of the high-energy tails of the electron distributions at the Si-SiO_2 interface from work now in progress[43] will help us in a more accurate determination of the parameters.

Despite these words of caution, we are confident that electron transport up to energies of about 2 eV is described very well by our model. Our skepticism is confined to the description of transport in the range above 3 eV, where impact ionization plays a dominant role and we have the strong doubts expressed above on modeling injection into the SiO_2, which remains the only available experimental information we can use to test the model. As a low-field internal check, we simulated the velocity-field characteristics at 77 K without adjusting the parameters (17), obtaining a very good agreement with the experimental data.[5,39]

We show in Fig. 3 the total electron-phonon scattering rate as a function of electron energy obtained integrating numerically our anisotropic rates over all directions:

$$\frac{1}{\tau_{el-ph}(E)}=\frac{1}{\mathcal{D}(E)}\sum_{\mu,\mathbf{k},\eta}\frac{1}{\tau_{\mu,\eta}(\mathbf{k})}\delta(E_\mu(\mathbf{k})-E),$$

where $\mathcal{D}(E)$ is the density of (initial) states at energy E. The strong role played by the density of final states is clearly evident comparing Fig. 1(b) to Fig. 3. The low-energy rates at 300 K resemble closely the magnitude of the rates used in previous Monte Carlo work, shown by the dashed line. In Fig. 4 we show the drift velocities versus electric field at 300 and 77 K. Barely visible in the figure is a region of negative differential mobility at 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. 5, is slightly lower than that obtained by "parabolic" Monte Carlo simulations at high fields.[5,39] More about this effect and the role played by L-valley transfer in small devices will be said below. Figure 6 shows the 300-K electron mean free path at various fields and in Fig. 7 we present the 300-K ionization coefficient obtained from our model compared to experimental data.[44]

We wish to stress here that we do not have the freedom to vary the six intervalley deformation potentials nor the energies of the phonons assisting the processes. These values are uniquely fixed, respectively, by the choice of the only two constants we have, the optical and acoustic deformation potentials, and from the dispersion (7) evalu-

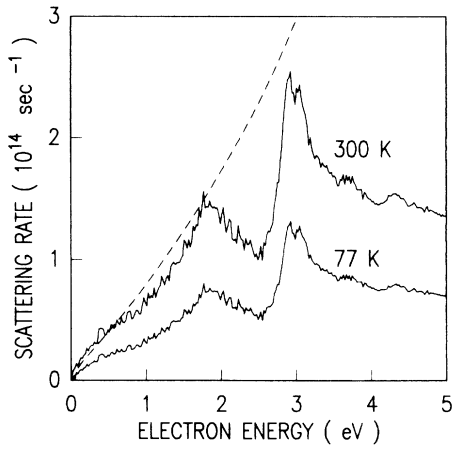

FIG. 3. Total electron-phonon scattering rate for Si at room and liquid-nitrogen temperature. The rate is plotted as a function of electron energy by integrating the anisotropic rates given by Eq. (4) of the text over all directions in the Brillouin zone. The dashed line corresponds to the total electron-phonon scattering rate given by Ref 5.

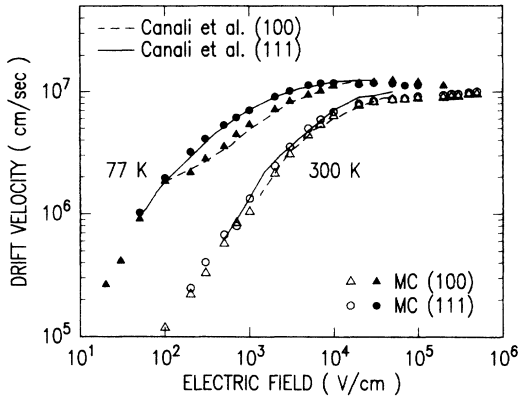

FIG. 4. Experimental and simulated electron drift velocity as a function of electric field along two crystallographic directions in Si at room and liquid-nitrogen temperature.

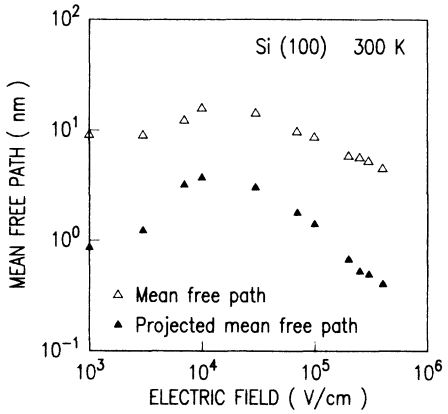

FIG. 6. Simulated electron mean free path for Si at room temperature for a field along the (100) crystallographic direction. The average mean free path and its projection along the direction of the field are shown.

ated at various **q** vectors connecting the different valleys. We do not wish to revive any controversy about these constants.[5] All we want to say is that our choice in (17) satisfies a wealth of experimental data, within the approximations we have employed to compute the scattering rates.

It is instructive to compare the electron-phonon transport parameters implied by our choice in Eq. (17) with those employed in parabolic-band simulations. In Table I we show such a comparison. The intravalley and intervalley acoustic deformation potentials obtained from Eq. (17) account for both TA- and LA-phonon-assisted transitions. We should also recall that the $\sqrt{2}$ error we make on the small-q phonon dispersion of Eq. (7) has the effect of increasing the value of the small-q matrix element for acoustic transitions, so that our low-q acoustic coupling constants are depressed by a factor $2^{1/4}$. The intervalley coupling constants and phonon energies have been obtained by accounting for the q dependence of acoustic transition and for the phonon wave vectors needed in the f processes and the g processes. Note that intervalley and intravalley transitions are not distinct processes in our model, since the "valleys" themselves are ill-defined **k**-space regions. Apart from an apparent "switch" in the f_2 and g_1 intervalley deformation potentials and phonon energies, our constants appear consistently lower, with the exception of the g_2 intervalley transition. This is strictly due to band-structure effects, even at very low energies. Electrons can flow from one valley along the (100) direction to an equivalent valley in another BZ without the assistance of phonons (Bloch oscillations), the crossing between bands 1 and 2 occurring only along the symmetry lines from the symmetry point X to the symmetry point W. Thus, lower rates are needed to maintain the carriers in a low-energy and low-velocity regime in a homogeneous situation. Also, the ellipsoidal shape of the equienergy surfaces around the energy minima along the 6 Δ directions seems to be a poor representation of the bands. These surfaces, when cut by a plane normal to the symmetry line Δ, are not spherical but noticeably "square-looking" above 50 meV, implying that the transverse electron effective mass is higher as we move away from the Δ minimum along the (110) transverse direction

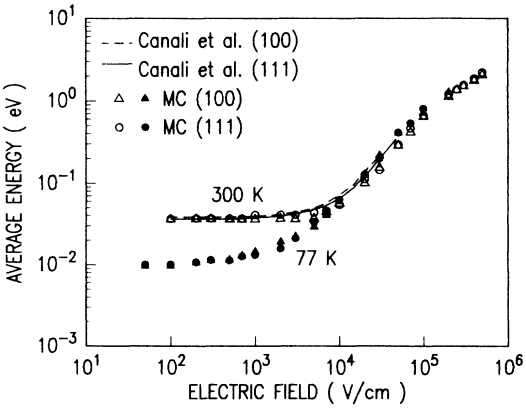

FIG. 5. Simulated average electron energy for Si at room and liquid-nitrogen temperature as a function of electric field along two crystallographic directions. Results of previous Monte Carlo simulations at 300 K are shown for comparison.

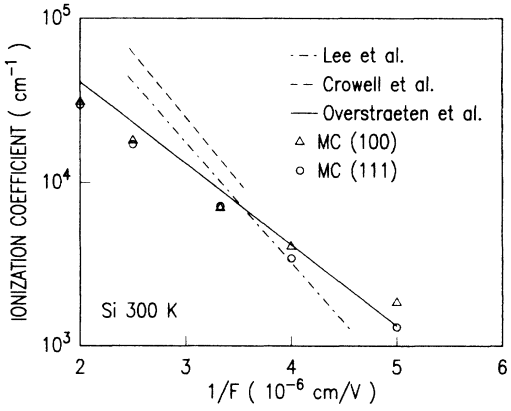

FIG. 7. Experimental and simulated impact-ionization coefficient for Si at room temperature for a field along the (100) and (111) crystallographic directions.

TABLE I. Transport parameters in Si.

Quantity and units	This work	Previous works
$\hbar\omega_{LA,max}$ (meV)	22.1	17.5[a]
$\hbar\omega_{TA,max}$ (meV)	44.3	48.4[a]
$\hbar\omega_{TO}$ (meV)		59.7[a]
$\hbar\omega_{LO}$ (meV)	62.0	62.6[a]
$\Delta_{ac,X}$[b] (eV)	2.4 (band 1) 3.4	9.0[c,d]
(ΔK_{op}) (10^8 eV/cm)	1.75 (band 1) 2.1	
$(\Delta K)_{f,1}$ (TA)[e] (10^8 eV/cm)	0.30(14.7)	0.15(18.1),[d] 0.3(18.2)[c]
$(\Delta K)_{f,2}$ (LA) (10^8 eV/cm)	0.30(7.2)	3.4(43.1),[d] 2.0(47.4)[c]
$(\Delta K)_{f,3}$ (LO,TO) (10^8 eV/cm)	1.75(62.0)	4.0(54.2),[d] 2.0(59.0)[c]
$(\Delta K)_{g,1}$ (TA) (10^8 eV/cm)	1.18(44.3)	0.5(12.1),[d] 0.5(12.1)[c]
$(\Delta K)_{g,2}$ (LA) (10^8 eV/cm)	1.18(22.1)	0.8(18.1),[d] 0.8(18.5)[c]
$(\Delta K)_{g,3}$ (LO) (10^8 eV/cm)	1.75(62.0)	3.0(60.3),[d] 11.0(62.0)[c]

[a]G. Nilsson and G. Nelin, Phys. Rev. B **6**, 3777 (1972).
[b]$\Delta_{ac,X}=\Delta_{TA}+\Delta_{LA}$.
[c]C. Canali *et al.*, Phys. Rev. B **12**, 2265 (1975).
[d]C. Jacoboni and L. Reggiani, Rev. Mod. Phys. **55**, 645 (1983).
[e]In parentheses is the energy of the corresponding phonons (meV).

then the value of $0.193m_{el}$, where m_{el} is the free-electron mass, found along the (100) transverse direction. The higher DOS at low energies shown in Fig. 1(b) is due to this effect and to the presence of the second conduction band at about 130 meV, as already noted by Tang and Hess.[10] This also contributes to an increase in the rates, even at moderate energy. Thus, a straightforward comparison of the various parameters is rendered obscure by these band-structure considerations. At even higher energies, no similarities of the two models can be expected.

We would like to stress that *the ability to model electron transport in Si with only four adjustable parameters for all possible scattering processes is a very remarkable result. In our opinion, it emphasizes the fact that a better description of the band structure is a fundamental ingredient for the understanding of transport, even at relatively low fields.*

B. Gallium arsenide

Considerations similar to those for Si apply to GaAs. In this case, the band structure is even more important and some uncertainty remains in the literature about parameters such as the effective masses in higher valleys, energy splitting $\Delta_{\Gamma,L}$ and $\Delta_{\Gamma,X}$ and nonparabolicity corrections. In Table II we list the band-structure parameters obtained from the pseudopotential calculations compared to typical values employed by different authors. An extensive review of the various published parameters is outside the scope of this work, but it would reveal clearly that confusion persists in this matter. The nonparabolicity parameter α reported in Table II depends on the energy at which the "exact" bands are matched. At low energies (up to about 0.4 eV), a low value is reported. This is still larger than the one usually employed, but it fits the quantum oscillations observed in ballistic devices.[45] Larger values (due to higher-order nonparabolic corrections lumped into a single first-order parameter) are obtained at larger energies. At these energies, however, the Γ valley becomes appreciably nonspherical and the notion of a nonparabolic correction is dubious anyway.

Of course, the determination of the electron-phonon coupling parameters depends on the band structure selected. The calibration of these parameters has been done in the same spirit as in the approach taken for Si. Only two parameters are treated as adjustable quantities: the acoustic-phonon deformation potential Δ_{ac}, and the optical deformation potential $(\Delta K)_{op}$. The polar coupling with LO phonons in the Γ valley is not a source of con-

TABLE II. Band-structure parameters for GaAs.

Symbol and units	This work	Previous works
m_Γ (m_{el})	0.063	0.063,[a] 0.069[b]
$m_{L,l}$ (m_{el})	1.538	1.473[b]
$m_{L,t}$ (m_{el})	0.127	0.12[b]
$m_{X,l}$ (m_{el})	1.987	1.58[b]
$m_{X,t}$ (m_{el})	0.229	0.24[b]
$\Delta E_{\Gamma,L}$ (eV)	0.323	0.33,[a] 0.29[b]
$\Delta E_{\Gamma,X}$[c] (eV)	0.457	0.522,[a] 0.48[b]
α (eV^{-1})	−0.834[d] −1.158[e]	−0.61,[a] −0.67[b]

[a]M. A. Littlejohn *et al.*, J. Appl. Phys. **48**, 4587 (1977).
[b]T. Wang and K. Hess, J. Appl. Phys. **57**, 5336 (1985).
[c]Minimum at $0.9(2\pi/a)$ along the symmetry line Δ in our model, at the symmetry point X in previous Monte Carlo works.
[d]Low-energy value (0.3 eV).
[e]High-energy value (0.7 eV).

cern, as the Fröhlich coupling constant can be considered well known.[27] As for Si, the acoustic deformation potential has been allowed to vary above some energy threshold (0.3 eV), in order to fit the velocity-field characteristics.

The fitting procedure is expected to exhibit a wider flexibility than we had in the silicon case. No experimental information is available in the 3-eV range and the ionization coefficients are still somewhat controversial. Therefore, the set of parameters below is even less "unique." After fitting drift velocity versus field curves[46] and ionization coefficients,[34,47] the parameters we employ are

$$\Delta_{ac}=\begin{cases}7.0\text{ eV}, & E<0.3\text{ eV in }\Gamma\\ 5.0\text{ eV}, & \text{otherwise}\end{cases}, \tag{18a}$$

$$(\Delta K)_{op}=\begin{cases}0.0, & E<0.3\text{ eV in }\Gamma\\ 2.1\times10^{8}\text{ eV/cm}, & \text{otherwise}\end{cases}, \tag{18b}$$

$$E_{th}=1.7\text{ eV}, \tag{18c}$$

$$\frac{P}{\tau_{op}(E_{th})}=2.5\times10^{15}\text{ sec}^{-1}. \tag{18d}$$

The impact-ionization coefficients seem to indicate a hard threshold, opposite to what is found in Si, consistently with the numerical results of Ref. 11. But there is an important caveat with regard to this model Contrary to the theoretical findings of Ref. 11, our anisotropic electron-phonon scattering rates yield an anisotropic ionization coefficient. Unfortunately, this orientation dependence is in complete disagreement with the experimental data.[34] We believe that the oversimplification implied by the Keldysh formula is at the origin of this problem, the anisotropy of the threshold energy being an important factor for which we did not account. Thus, instead of settling the controversy, as we had hoped, we end up confusing the issue even more. Obviously, we must say that *further work is necessary.*

In Table III we compare our electron-phonon coupling constants to those used elsewhere. The intervalley deformation potentials we have tabulated include both optical- and acoustic-phonon transitions. The values listed in Table III are obtained by adding the two contributions as follows:

$$(\Delta K)_{i\to j}=[\,|\,\Delta_{ac}\mathbf{q}_{i\to j}\,|^{2}+(\Delta K)^{2}_{op}]^{1/2},$$

to facilitate the comparison with the total coupling constant for intervalley transitions from valley i to valley j assisted by phonons of wave vectors $\mathbf{q}_{i\to j}$. In Table III, the intervalley phonon energies we use are given by an average of the acoustic and optical intervalley phonons, weighted by the respective coupling constants squared. We show in Fig. 8 the total electron-phonon scattering rates at 300 K.

Because of the very narrow Γ valley, the numerical procedure becomes too inaccurate at low electron energy (too few cubes in the BZ to describe the Γ valley). Therefore, in the case of GaAs we have simulated the electrons in a first-order nonparabolic band approximation for energies below 0.3 eV, as done by Shichijo and Hess.[11] The transport parameters, in this case, are those used by Littlejohn *et al.*[27] Inelastic acoustic-phonon scattering, the usual Fröhlich scattering, overlap integral including the *s-p*-wave mixing as done by Fawcett *et al.*,[48] and the nonparabolicity parameter of -0.834 eV^{-1} shown in Table II have been employed. All intervalley processes, however, have been treated within our full-band-structure scheme. In Figs. 9–12 we show the drift velocity, average energy, mean free path, and valley populations as functions of the electric field at 300 K.

As is the case for Si, we can reproduce the low-field transport with only three electron-phonon coupling constants. This is in contrast to a Monte Carlo simulation using parabolic or first-order nonparabolic bands, where ten or more scattering parameters must be adjusted. However, our conclusion in the case of GaAs is undoubtedly weaker, because of the problems we encountered with the ionization coefficients, and must be restricted to the collision processes above 0.3 eV.

TABLE III. Transport parameters in GaAs.

Symbol and units	This work	Previous work
$\hbar\omega_{LA,max}$ (meV)	24.3	
$\hbar\omega_{LO}$ (meV)	35.36	35.36[a]
$\Delta_{ac,\Gamma}$ (eV)	7 (below 0.3 eV) 5 (above 0.3 eV)	7.0[a]
$\Delta_{ac,L}$ (eV)	5.0	9.0[a]
$\Delta_{ac,X}$ (eV)	5.0	9.27[a]
$(\Delta K)_{\Gamma,L}$[b] (10^8 eV/cm)	5.2(23.6)	6.5[c](27.8)[a]
$(\Delta K)_{\Gamma,x}$ (10^8 eV/cm)	5.9(24.2)	10.0(29.9)[a]
$(\Delta K)_{L,X}$ (10^8 eV/cm)	5.2(23.6)	5.0(27.8)[a]
$(\Delta K)_{L,L}$ (10^8 eV/cm)	5.9(24.2)	10.0(29.0)[a]
$(\Delta K)_{X,X}$ (10^8 eV/cm)	2.1(35.2)	7.0(29.9)[a]

[a]M. A. Littlejohn *et al.*, J. Appl. Phys. **48**, 4587 (1977).

[b]In parentheses are the phonon energies in meV. Acoustic and optical phonons are included in intervalley scattering in this work.

[c]J. Shah *et al.*, Phys. Rev. Lett. **59**, 2222 (1987).

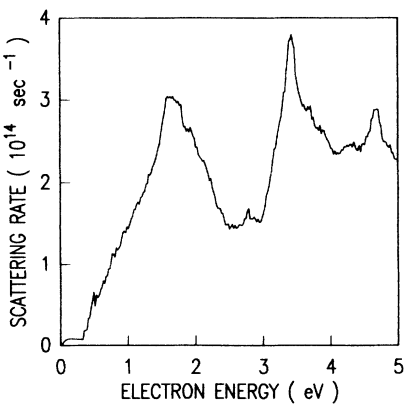

FIG. 8. Total electron-phonon scattering rate for GaAs at 300 K. The polar scattering with LO phonons below 0.3 eV has been computed in a spherical, nonparabolic-band approximation for the Γ valley, as explained in the text.

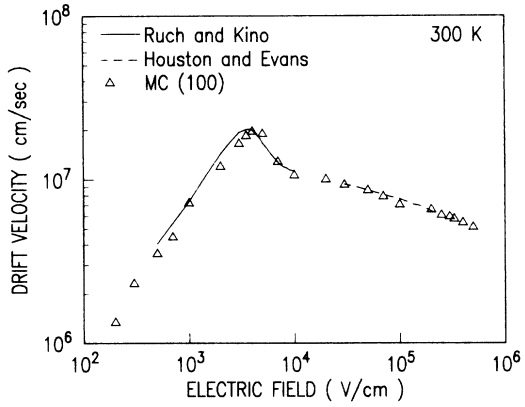

FIG. 9. Experimental and simulated electron drift velocity as a function of electric field along the (100) crystallographic direction in GaAs at 300 K.

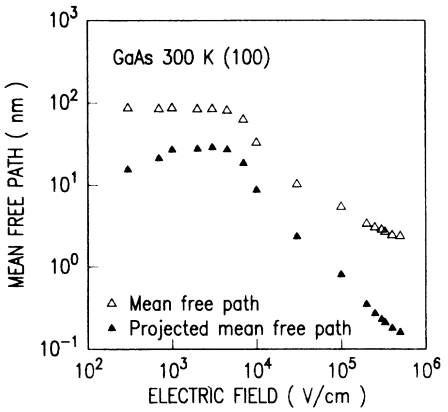

FIG. 11. Simulated electron mean free path and its projection along the direction along the field as a function of the field itself along the (100) crystallographic direction in GaAs at 300 K.

IV. SPACE-CHARGE EFFECTS

The ability to monitor the time evolution of the electric fields in the devices self-consistently with the particle motion is important in small devices. The carrier velocities are expected to be different from the "equilibrium" velocities implied by simpler DD models. This, in turn, implies a redistribution of the charge density in the device and, consequently, a different electric field configuration which feeds back into the particle motion.

At present, this self-consistent scheme can be handled with almost standard techniques, such as those described by the pioneers of this approach in Ref. 6. For completeness, we shall outline the procedure we have followed, emphasizing the few instances which deviate from the Hockney and Eastwood prescription.

A. Poisson equation

The mesh which describes the two-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 electron density, the ionized dopant density, potential boundary conditions, and a piecewise constant dielectric constant. Holes are included in the zero-current (constant–quasi-Fermi-level) approximation, so that the depletion region in the substrate of an n-type-channel metal-oxide-semiconductor field-effect transistor (MOSFET) is self-consistently and automatically included in the calculation. Therefore, the Poisson equation to be solved is

$$-\nabla_r\cdot(\epsilon\nabla_r\phi) = e[N_V\mathcal{F}_{1/2}((E_V-\phi_P)/k_BT)-n_{\rm el}+N_D^+-N_A^-]\ , \tag{19}$$

where $\mathcal{F}_{1/2}$ is the Fermi-Dirac integral of order one-half, ϕ_P is the hole quasi-Fermi-level, $N_V\mathcal{F}_{1/2}$ 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 frac-

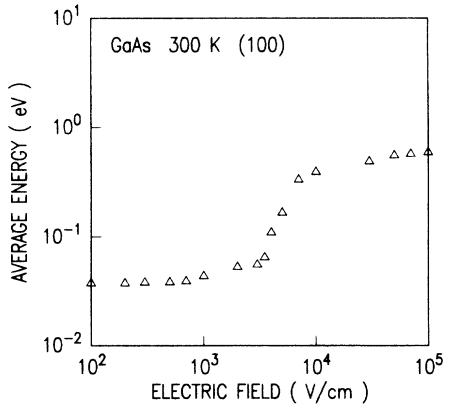

FIG. 10. Simulated electron average energy in GaAs at 300 K as a function of electric field along the (100) crystallographic direction.

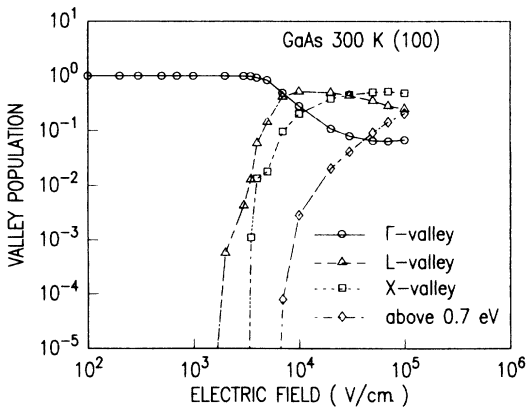

FIG. 12. Fraction of electrons in each valley and above an arbitrary cutoff of 0.7 eV as a function of the field along the (100) crystallographic direction in GaAs at 300 K.

tion of ionized donors, quasiequilibrium 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.[49] A polynomial preconditioned conjugate gradient technique is used to solve the resultant linearized matrix equations.[50]

The standard cloud-in-a-cell (CIC) algorithm[6] is employed to assign the particle charge density $en_{\rm el}(\mathbf{r})$ 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. As a first extension, we continue to use the standard CIC formula but substitute the local mesh spacings in the charge assignment process. Hence, where the mesh is refined, the particle charge is only spread over a small distance, and where the mesh is coarse, charge is spread over a greater distance. This contrasts with the "extended" CIC method employed by Tomizawa *et al.*,[51] where charges are always spread over the same distance, independent of the nonuniform mesh spacing. We shall briefly defer discussion of this concern until after the next paragraph.

Our second extension to the standard CIC method involves the manner in which electric field values are assigned to the mesh points, prior to CIC force interpolation at the particle location. We attempt to address the presence of a piecewise constant dielectric constant in our mesh as follows. Consider a finite-difference mesh with mesh lines located at $\{x_i\}$ along the x axis, and $\{y_j\}$ along the y axis. The dielectric constant in the rectangles determined by these lines is taken as constant and is denoted by $\epsilon_{i+1/2,j+1/2}=\epsilon(x,y)$, $x_i \le x \le x_{i+1}$ and $y_j \le y \le y_{j+1}$. Also, define the mesh spacings as $\Delta x_{i+1/2}=x_{i+1}-x_i$ and $\Delta y_{j+1/2}=y_{j+1}-y_j$. The electric field $\mathbf{F}=(F_x,F_y)$ at the four mesh nodes (i,j), $(i+1,j)$, $(i,j+1)$, and $(i+1,j+1)$ is required to interpolate the field for any particle inside this mesh rectangle. These field values are

$$-(F_x)_{i+l,j+m}=\frac{\epsilon_{i+l+1/2,j+1/2}\left[\dfrac{\phi_{i+l+1,j+m}-\phi_{i+l,j+m}}{\Delta x_{i+l+1/2}}\right]+\epsilon_{i+l-1/2,j+1/2}\left[\dfrac{\phi_{i+l,j+m}-\phi_{i+l-1,j+m}}{\Delta x_{i+l-1/2}}\right]}{2\epsilon_{i+1/2,j+1/2}}, \tag{20a}$$

$$-(F_y)_{i+l,j+m}=\frac{\epsilon_{i+1/2,j+m+1/2}\left[\dfrac{\phi_{i+l,j+m+1}-\phi_{i+l,j+m}}{\Delta y_{j+m+1/2}}\right]+\epsilon_{i+1/2,j+m-1/2}\left[\dfrac{\phi_{i+l,j+m}-\phi_{i+l,j+m-1}}{\Delta y_{j+m-1/2}}\right]}{2\epsilon_{i+1/2,j+1/2}}, \tag{20b}$$

for $l,m=0,1$. These formulas attempt to minimize the self-force on the particle when the mesh spacing is nonuniform, yet still reduce to a second-order accurate centered difference approximation when $\Delta x,\Delta y$ are both constant.[52] There are many other prescriptions concerning the force calculation and interpolation which need to be explored[6] and extended to the case of a nonuniform mesh spacing and spatially dependent dielectric constant.

Given the possibility of unphysical self-forces when nonuniform meshes are used with Eq. (20), we construct our Poisson mesh so that Δx and Δy are constant directly below the Si-SiO_2 interface to a depth of 10 nm. This forces the self-force to zero in the region where the majority of the particles in the channel reside. Questions concerning the force calculation due to the change in dielectric constant at the interface beyond those addressed by Eq. (20) remain for future examination, however.

B. 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 DD solution of the device and using the resulting electron 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 used in Eq. (12). Whatever starting configuration is chosen, a transient is simulated at first. Of course, this is physically meaningful only if the initial configuration itself is physically meaningful. In the simulations we have performed to date, we have generated initial particle locations according to the doping profiles. The very nonphysical nature of this initial "solution" yields very wild, nonphysical transients at the onset. The simulation time needed to reach a physical configuration can be clearly observed by monitoring the trend of all physical quantities (energies, velocities, densities, etc.) towards steady-states values.

The initial wave vector 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} 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 added three different features to our model.

(i) Particles with kinetic energies below some low threshold (typically 50 meV for Si, 0.3 eV for GaAs) are simulated in a first-order nonparabolic ($\alpha = -0.5$ eV^{-1} for Si, -0.834 eV^{-1} for GaAs) band approximation, as done in standard MC simulations.[39,27] Since most carriers will be quasithermal in the contacts, a significant amount of CPU time is saved in these regions given the much higher computational speed of these "analytic" particles. The energy thresholds we chose are such that the band-structure features previously discussed are well preserved.

(ii) A portion of the highly doped regions is cut out from the domain in which the particle motion is simulated, while still remaining in the electrostatic portion of the problem. This "cut" must be such that its boundary meets the requirement described above for the location of the Ohmic contacts. Thus, apart from some contact-resistance effects, there is no loss of accuracy in the simulation. If an estimate of the source and drain resistance is desired, this shortcut may be easily bypassed.

(iii) The usual technique for enhancing rare events in (MC) simulations[53] is employed in the active regions of the devices (such as inversion channels in MOSFET's). Particles within a predefined "statiscally enhanced region" of real space are assigned a different s factor, s_{mult}. When a particle enters this region, it is replicated $M-1$ times. Conversely, a particle leaving it will be kept in the simulation with probability M^{-1}. The charge weight of each particle in the statistically enhanced region is clearly $s_{\text{mult}} = s/M$. This technique has also been employed by Sangiorgi *et al.*[13] to enhance statistically regions of **k** space. Particular care must be taken to handle correctly the short-range electron-electron interaction between particles with different s factors: the number of neighbors in the screening circle is obtained by counting each particle outside the statistical-enhancement region M times. Equation (12) is modified accordingly by summing over the various s factors. This will ensure that the correct local density is employed to evalute Eq. (13). If a collision of particles with different s factor is selected, the state of the particle outside the statistically enhanced region (i.e., with larger s) is modified with probability M^{-1}. The ensures that the total energy of the ensemble is statistically conserved.

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}, as described in Sec. II B. At the end of this step, we must ensure that all particles remain within the allowed region. 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} a check is made to determine how many particles undergo collisions of any type. This is done in a very conventional way, by comparing, for every electron of wave vector $\mathbf{k}(t)$, the ratio $\Delta t_{\text{sc}}/\tau_{\text{tot}}[\mathbf{k}(t)]$ with a random number ξ in [0,1], $1/\tau_{\text{tot}}(\mathbf{k})$ being the total scattering rate. For every scattering electron, the type of scattering is selected according to the relative weight of the various processes included in the model and a new state $(\mathbf{k}', \nu')$ is selected, as explained in Sec. II D. To account properly for the total scattering probability, the scattering time step Δt_{sc} must be chosen in such a way that $\Delta t_{\text{sc}} \ll \tau_{\text{tot,max}}$, where $1/\tau_{\text{tot,max}}$ is the maximum scattering rate occurring in the simulation. This can be estimated at the beginning of the simulation from the bias condition, lattice temperature, device dimensions, and other input parameters. Obvious-

ly, the free-flight time $\Delta t_{\rm bal}$ must be less or equal than $\Delta t_{\rm sc}$.

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 $n_{\rm el}$ can develop plasma oscillations of angular frequency ω_p, which, in the effective-mass approximation, is given by

$$\omega_p = \left[\frac{e^2 n_{\rm el}}{\epsilon m_{\rm eff}} \right]^{1/2}, \tag{21}$$

where $m_{\rm eff}$ 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 $\Delta t_{\rm bal} \simeq 0.2$ 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, as follows: the thermal random motion of particles results in density fluctuations. Whenever a lower density region appears, the large fields due to the "exposed" dopants strongly drive the nearby carriers towards the low-density region. Unless a quick correction to the potential is made, this would result in even larger "density voids" in nearby regions. Huge, oscillating fields would result, yielding "explosions" of particles in real and **k** space. This problem becomes severe when realistically high concentrations are employed in the simulation, so that the associated plasma frequency becomes comparable to the frequency at which the particle positions are updated. To overcome this "plasma catastrophe," two choices are available. The first one consists of smoothing the potential in some way, either by performing a suitable average over a time interval longer than ω_p^{-1} (Ref. 54), or by using a smoothed charge density to solve the Poisson equation.[55] In the first case one loses the possibility of simulating fast time transients and in either case one misses some components of the long-range Coulomb interaction, such as the plasma losses of hot carriers. If any smoothing algorithm is chosen, this energy-loss mechanism must be included explicitly as an additional scattering mechanism.[31] Therefore, we have chosen to update the electric field at a very high frequency by using a time, $\Delta t_{\rm Poisson}$, between successive Poisson updates such that $\Delta t_{\rm Poisson} \lesssim (5\omega_p)^{-1}$ ($=0.2$–0.4 fsec in our simulation, i.e., every one or two ballistic steps). The higher CPU time spent in solving Eq. (19) 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. In the first place, as we stressed above in the context of the short-range Coulomb scattering, a large number of particles, $n_{\rm prt}$, must be included in the simulation to treat properly the plasma oscillations. We can state more precisely this condition in terms of the s factor: in addition to the condition $s < \beta_{s,\min}$ stated for the short-range interaction, we must have a number of particles large enough so that $s \lesssim \Delta x^{-1}$, where Δx is the mesh spacing, so that density fluctuations in each two-dimensional mesh element represent a correct sampling of the real three-dimensional fluctuations.[56] Secondly, we must face a rather complicated situation concerning the single-particle, collective-mode nature of the Coulomb excitations in degenerate regions. With reference to the well-known Fig. 13 (Ref. 57), plasmons of wave vector $q > \bar{q}$, where $\hbar\omega_p = \hbar\omega_+(\bar{q})$ [$\hbar\omega_\pm(\mathbf{q}) = E(\mathbf{q}\pm\mathbf{k}_F) - E_F$, and $\mathbf{k}_F$ is the Fermi wave vector along the direction of **q**] merge into the single-particle continuum and quickly decay into single-particle excitations via Landau damping. In the simulation, plasmon damping is controlled by the mesh spacing, since oscillations of $q > \Delta x^{-1}$ cannot be spatially resolved. By choosing $\Delta x \simeq \bar{q}^{-1}$ we would correctly damp the short-wavelength plasmons, but we would miss the single-particle excitations produced by their decay for $\Delta x^{-1} < q < \beta_s$. It is not easy to correct the situation in general, because more constraints effectively limit the choice of the mesh spacing: it cannot be too large in regions where large gradients of carrier concentration are anticipated, if an accurate solution of Poisson equation is to be obtained, nor can it be too small because of CPU-

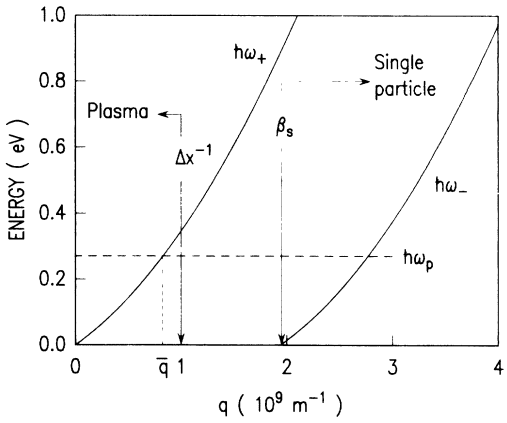

FIG. 13. Single-particle and plasma dispersion relation at zero lattice temperature for the degenerate situation occurring in the accumulation regions of source and drain under the gate contact ($n_{\rm el} = 2.0\times10^{20}$ cm^{-3}, $T_{\rm el} = 1335$ K). Parabolic bands are assumed for simplicity in this plot. Long-lived plasmons should be excited at wave vector q's to the left of the curve labeled $\hbar\omega_+$, single particles in the region between the curves $\hbar\omega_-$ and $\hbar\omega_+$. In the simulation, single particles are excited only for $q \gtrsim \beta_s$ to account for screening, collective modes for $q \lesssim \Delta x^{-1}$, Δx being the mesh spacing. In spatial regions in which $\Delta x^{-1} < \bar{q}$, long-lived plasmons are incorrectly ignored for $\Delta x^{-1} < q < \bar{q}$. If, instead, $\bar{q} < \Delta x^{-1}$, for $\bar{q} < q < \Delta x^{-1}$ plasmons are incorrectly left undamped. In our case, in the degenerate regions $\Delta x^{-1} \simeq \bar{q}$ and the only error made in the simulation is the absence of single-particle excitations via Landau damping for $\bar{q} < q < \beta_s$. At a finite lattice temperature the cutoff at $q = \bar{q}$ is smeared by the broadening of the curve $\hbar\omega_+$.

time and data-storage requirements. Therefore, in a typical simulation we might have regions in which $\Delta x^{-1} < \bar{q}$, so that we will ignore long-lived plasmons for $\Delta x^{-1} < q < \bar{q}$, and Landau damping for $q > \bar{q}$. On the contrary, we might have $\Delta x^{-1} > \bar{q}$ in other regions and we shall allow the excitation of incorrectly undamped plasma oscillations for $\bar{q} < q < \Delta x^{-1}$. These errors may be significant in the junction regions, but not in the highly doped regions (where $\Delta x^{-1} \simeq \bar{q}$ and the only error will be the absence of Landau-damped excitations for $\bar{q} < q < \beta_s$) or in the channel, where the energy of the plasmons is very small. The distinction between long-range and short-range Coulomb interaction and the way it is treated in our model is summarized in Fig. 13.

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. Only the size of the output files thus generated poses a constraint on the frequency at which this information is stored.

V. THE PROGRAM

In typical applications, an ensemble of 5000–10 000 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)

$$\Delta t_{\text{bal}} = 2 \times 10^{-16} \text{ sec} ,$$

$$\Delta t_{\text{sc}} = \begin{cases} 2 \times 10^{-16} \text{ sec}, & T = 300 \text{ K} \\ 10^{-15} \text{ sec}, & T = 77 \text{ K} , \end{cases}$$

$$\Delta t_{\text{Poisson}} = 4 \times 10^{-16} \text{ sec} .$$

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 3–5 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-processor facilities. In many cases, standard algorithms have been modified for vectorization purposes. Typically, the program spends 50–70 % 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. (Here 1 Mbyte $\equiv 10^6$ bytes and 1 byte $\equiv 2^3$ binary digits.)

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. (8) and particularly Eq. (13). The program requires 1–6 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 electron density, the size and dopant concentration of the contact regions, the statistical-enhancement ratio M, 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 20–100 times shorter.

VI. SUBMICROMETER Si MOSFET

In this section we present the results of simulations we have performed on exploratory short n-type-channel Si MOSFET's.[58,59] We shall not discuss issues strictly related to the device aspects of the simulation, but we will focus on the physical information extracted from the simulation and its comparison with the experimental data. New results which emerge from the simulations are as follows. (1) The quasiballistic nature of electron tranport 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.

The devices we studied have an effective channel length ranging from a little over 0.25 μm down to 60 ± 5 nm. Details about the process employed to manufacture the devices can be found in Ref. 58. It suffices to mention that direct-write electron-beam lithography was used. Degenerate source and 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 drain-to-source bias, V_{DS}). It is well known that the saturated velocity in long channels is much smaller than in the bulk of the semiconductor.[60] Unless proper account is taken of the 2D features of electron transport, of the correct scattering with interfacial impurities in the gate insulator,[61] and of the roughness of the Si-SiO_2 interface,[62] 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.[63] For these reasons, we have concentrated our attention on the high-V_{DS} characteristics, corresponding to average source-to-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.

A. Quasiballistic 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. 14(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. Figure 14(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. 14(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. 14(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. 14(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-

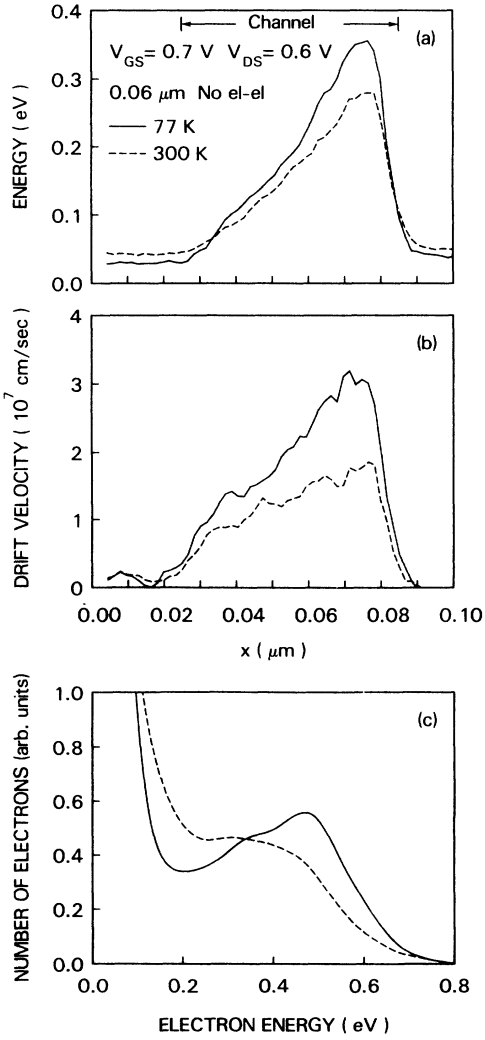

FIG. 14. (a) Average electron energy and (b) x-directed drift velocity at the distance of 1 nm from the Si-SiO_2 interface along the channel of a Si MOSFET having an effective channel length of 60 nm. The applied biases are $V_{DS}=0.6$ V, $V_{GS}=0.7$ V, and $V_{sub}=0.6$ V. The smoothed electron energy distributions at the drain end of the channel [$x=0.085$ μm in (a) and (b)] are shown in (c) for two ambient temperatures.

temperature behavior is less dramatic, since the shorter relaxation lengths prevent the carriers from flowing quasiballistically along the channel.

B. The role of the 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. 15. Despite the remarkable tendency of the energy distribution to be "smeared out" by the interparticle collisions, the effect is not a complete thermalization. Typically, the mean free path for the *e-e* collisions is of the order of 5 nm along the channel at 77 K. However, due to the almost coherent motion of the carriers, momentum relaxation does not occur in any significant amount outside the drain contact. On the contrary, some energy redistribution does occur, as shown by the high-energy tail observed in Fig. 15(c) which is affected significantly.

As the channel length increases, the electron-electron scattering becomes more effective in thermalizing particles and relaxing momentum. The partial randomization

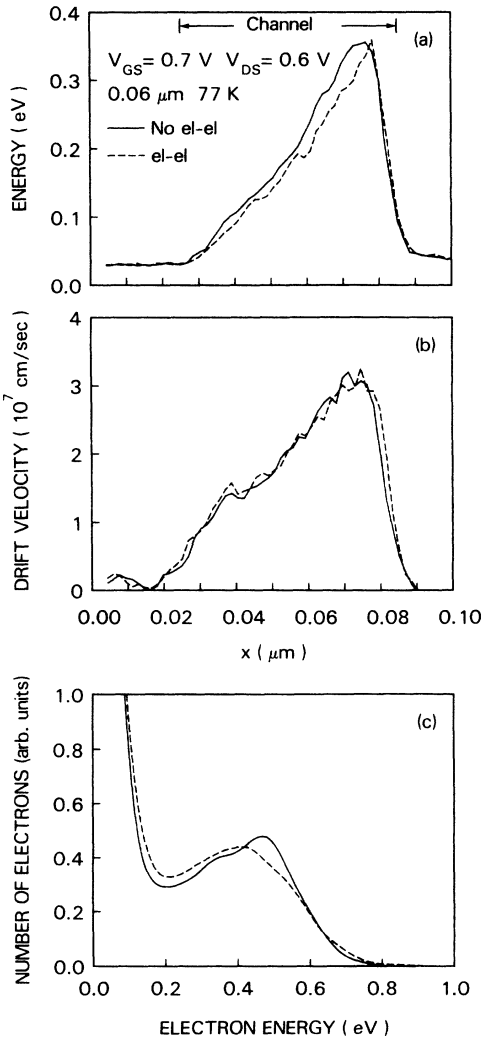

FIG. 15. (a) Average electron energy and (b) x-directed drift velocity profiles along the channel 1 nm away from the Si-SiO_2 interface for the same device and bias conditions described in Fig. 14 with and without the short-range electron-electron interaction. As expected, the difference is negligible. The electron energy distributions at the drain end of the channel are shown in (c). The "thermalization" effects of the Coulomb scattering are evident but not very pronounced due to the short time the electrons spend in the channel ($\sim$375 fsec).

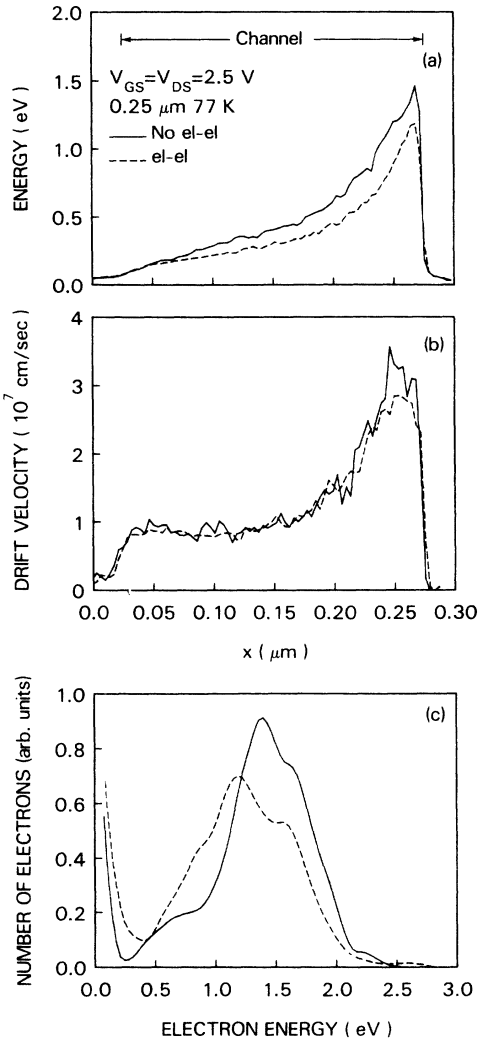

FIG. 16. (a) Electron average energy and (b) x-directed drift velocity profiles along the channel 1 nm away from the Si-SiO_2 interface for a device having an effective channel length of 0.25 μm at 77 K at the bias conditions indicated in the figure with and without the inclusion of short-range electron-electron scattering. Electron-energy distributions 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. Notice also the effects of the short-range electron-electron scattering in this longer device compared to the situation illustrated in Fig. 15—in the high-field present in the pinched-off region lower velocities result from the larger momentum-relaxation rate due to the short-range electron-electron scattering. These collisions have the effect of increasing the path length of the electrons and inducing enhanced energy-loss collisions. The result is a lower average energy (b) and an electron-energy distribution just inside the drain (c) shifted to lower energies and slightly broadened. The average electron transit time is about 2 psec in this case.

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. This is illustrated in Figs. 16(a), 16(b), and 16(c), respectively, for a device having a 0.25-μm-long channel.

C. Velocity overshoot

The average drift velocities along the channel shown in Figs. 14(b) and 15(b) indicate that a significant overshoot occurs near the drain end of the device, even at room temperature.[64,65] 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. 17. For the shorter devices at 77 K, the "effective" velocity, $v_{\text{eff}}=g_m/C_{\text{ox}}$ (C_{ox} being the oxide capacitance) extracted from the *extrinsic* transconductance[64] (i.e., not corrected for the contact resistance, amounting to a 5–10% correction in any case) is about 1.2×10^7 cm/sec. This effective velocity actually represents a lower bound to the actual average electron velocity along the channel.[59,64] 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 than a few percent with the extrinsic experimental value. This does not prove that the actual velocity distribution is, in reality, as shown in Fig. 14(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%.[64]

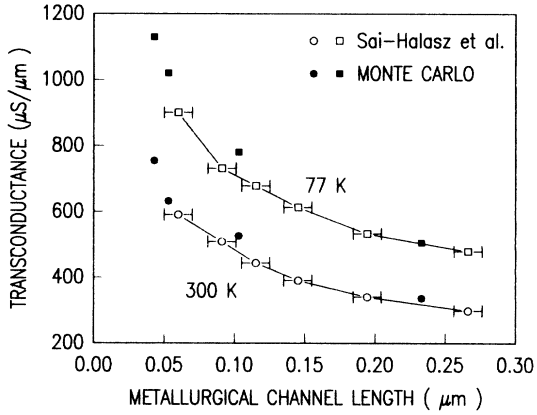

FIG. 17. Experimental and simulated small-signal transconductance as a function of the channel length. The metallurgical channel length—about 17 nm shorter than the "effective" channel length defined in the text—has been used for a direct comparison with the experimental data of Ref. 64. This difference is significant for the shortest devices. 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 $V_{GS}=1.0$ and 0.7 V for the 0.06-μm device (0.043 μm metallurgical length) with $V_{DS}=0.6$ V, at $V_{GS}=1.0$ and 0.8 V for the 0.07-μm device (0.053 μm metallurgical length) with $V_{DS}=0.6$ V, and at $V_{GS}=1.0$ and 0.8 V for the 0.12- 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 5–10% increase of the transconductance in both cases at the smallest channel lengths.

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. This can be understood by looking at Fig. 14(b): the amount of overshoot at 300 K is much smaller than at 77 K and it extends over a smaller fraction of the channel. The average drift velocity in the channel is thus below the saturated value also in the MC model. For this reason, DD models can be "stretched" to fit the experimental transconductance, provided low-field electron mobility is adjusted to fit the value of g_m on longer devices, and the bulk saturation velocity is used.[66] However, the "microscopic" pictures provided by MC and DD models are totally different: for a DD model to provide the correct g_m, the electron velocity must be at the saturated value over a very large fraction of the channel. Thus, the charge density and potential configuration of the device will differ greatly from the MC picture, despite the agreement of the two models as far as the macroscopic (or "terminal") characteristics are concerned.

Finally, the simple model we adopt for interface scattering has the net effect of yielding the maximum drift velocities at a distance of about 4–10 nm from the $Si\text{-}SiO_2$ interface, as illustrated by the velocity contour lines of Fig. 18(b).

The message we wish to convey is that situations might occur where even very simplistic models, such as drift diffusion, might provide a satisfactory description of the macroscopic characteristics of the device. This is particularly true in well-designed devices. In most cases, though, parameter adjustments will be necessary, such as the low-field mobility used in DD models, to obtain the results mentioned above. Much more important, however, is that the internal physical characterization of the device may bear little or no resemblance to "reality" in spite of the satisfactory macroscopic picture. This information is of vital importance for a basic understanding, for device design, and for modeling degradation processes associated with hot carriers.

D. Band-structure effects

One example of what we just said is clearly illustrated in Fig. 19. A device having a 0.25-μm 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 nonequilibrium 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.[51] 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. 19(a)]. A similar situation was already hinted at by the high-field, homogeneous results of Fig. 5.

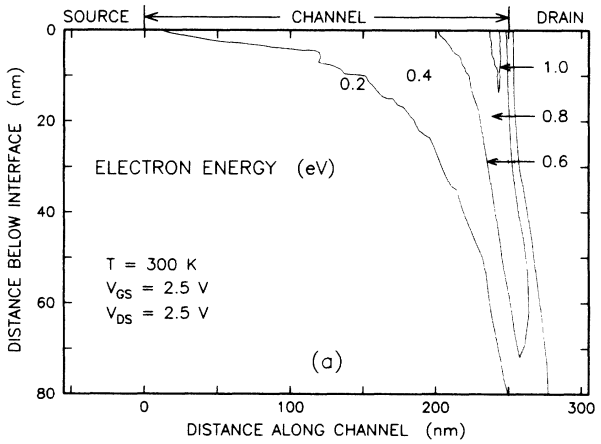

FIG. 18. Contour plot of (a) the electron energy and (b) x-directed drift velocity for the 0.25-μm-long device of Fig. 20 at 300 K. Note the different scales for the x and y axes we have employed for clarity. The highest velocities are reached about 5–10 nm away from the Si-SiO_2 interface. The Gaussian doping profiles in source and drain have steep gradients, decreasing from 10^{20} cm^{-3} to about 10^{17} cm^{-3} in 20 nm at about $x=0$ nm (source) and $x=250$ nm (drain). The metallurgical junctions are about 140 nm deep.

Similar results apply to the electron drift velocity, shown in Fig. 19(b): the parabolic model deviates from the full-band-structure model already in the 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. Figure 19(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 in Refs. 5 and 67. Those given by Canali *et al.*[41] provide a much smaller coupling constant for the intervalley g scattering with LO phonons

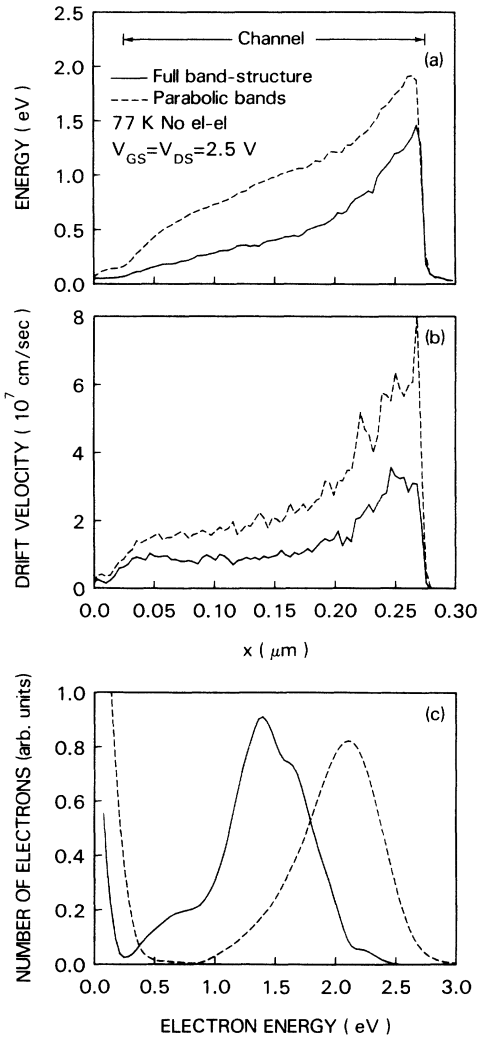

FIG. 19. (a) Electron average energy and (b) x-directed drift velocity at 77 K along the channel 1 nm away from the Si-SiO_2 interface for the device and bias conditions of Fig. 16 obtained from a model including the full band structure and from a model employing a parabolic-band approximation. In (a) the higher velocities obtained from the parabolic-band model imply longer inelastic electron mean free paths and a reduced energy relaxation rate. In (b) much lower velocities are seen in the numeric case, because of band-structure effects. In (c) we show the electron energy distribution at the drain end of the channel [$x=0.275$ μm in (a) and (b)].

and yield even larger discrepancies at low temperature.

To understand the origin of the large difference, we have plotted in Fig. 20 three "snapshots" of the electron population in the BZ at the source end of the channel (10 nm outside the source junction), at mid-channel, and at the drain end (10 nm outside the drain junction) in the full band-structure. We see how the electrons tend to fill almost uniformly the entire BZ close to the drain. A significant fraction of the carriers appears to be very close to the L symmetry point, one electron having even been scattered to the Γ valley. For electrons to reach energies in excess of 1 eV in the "correct" band structure, regions of the zone having low group velocities (or even holelike dispersion) must be populated. This has the effect of slowing down the carriers even in the absence of scattering. As an example, a look at Fig. 1 shows that 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

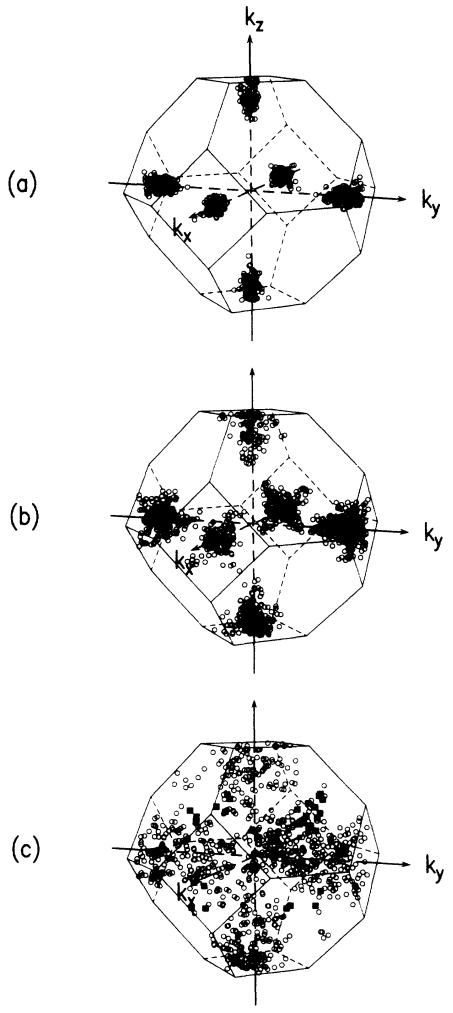

FIG. 20. Distributions of electrons in the BZ at various spatial locations along the channel of the 0.25-μm device of Fig. 19: at (a) $x=0.035$ μm, (b) $x=0.15$ μm, and (c) $x=0.265$ μm of Fig. 19(a). As they are accelerated by the drain field (along the $-k_x$ direction in this figure), the electrons shift in the k_x direction and expand around the six valleys along the symmetry line Δ, thus filling the entire BZ. Electrons within a distance $0.2(2\pi/a)$ from the L symmetry point are indicated by solid squares in (c), one electron in this sample (solid triangle at the zone center) being close to the Γ symmetry point. To help the visualization, not all electrons in the ensemble are plotted.

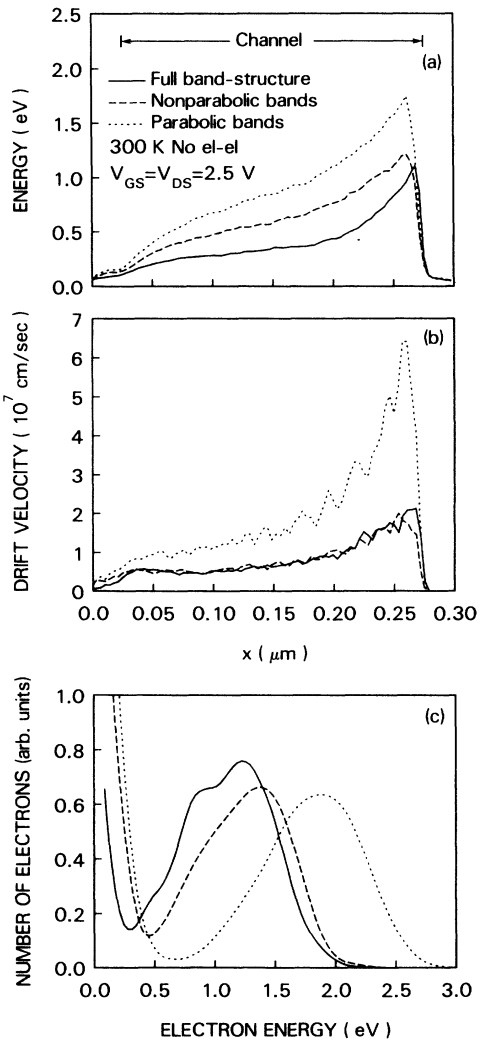

FIG. 21. (a) Electron average energy and (b) x-directed drift velocity profiles at 300 K along the channel 1 nm away from the Si-SiO_2 interface for the device and bias conditions of Fig. 19 obtained from a model including the full band structure and from models employing analytical approximations of the conduction band with and without first-order nonparabolic corrections. As expected, the higher temperature and the inclusion of nonparabolic corrections help to reduce the difference between the models. The sharper drop of the average energy in the drain in (a) and the smaller low-energy tails seen in (c) exhibited by the numeric model are due to its lower diffusion constant which prevent the cool electrons from "spreading" outside the drain region. The inclusion of nonparabolicity does not provide a good agreement with the numeric model for the average energies and energy distributions. The agreement between the full-band-structure model and the nonparabolic approximation in (b) is remarkably good because the nonparabolic scattering parameters of Ref. 5 are fitted to the experimental data on drift-velocity vs field curves.

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 a worst-case scenario. We can now look at the opposite limit, by increasing the temperature (so that the carriers will not be too hot and will sample a smaller region of the BZ) and by introducing the nonparabolic corrections to the approximated band structure. The introduction of a nonparabolicity parameter, though quite unjustified at these high energies, damps the velocities very effectively and increases the scattering rates, but it does not change the picture qualitatively as far as average electron energies and energy distributions are concerned. We show in Fig. 21 results of simulations performed at higher temperatures including the nonparabolic correction ($\alpha = -0.5$ eV^{-1}) compared to those obtained from the model based on the full band structure. The drift velocities resulting from these two models agree remarkably well. This has to be expected, since the nonparabolic correction has been shown to improve significantly the prediction of the approximated model for the high-temperature drift velocities.[68] The average energy, on the contrary, still exhibits values higher than those obtained by accounting for the full band structure. Therefore, despite the mild improvement obtained by the inclusion of first-order nonparabolic corrections, *band-structure effects (Figs. 19 and 21) and the short-range electron-electron interaction (Fig. 16) appear to be major factors in controlling even the gross features of the distribution of the hot electrons in the channel.*

VII. 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 submicrometer 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, and quantum size effects have been either crudely approximated or ignored in our model. Their effect on our results remains to be determined.

ACKNOWLEDGMENTS

We are particularly indebted to R. Car for having kindly given us the routine to set up and invert the matrix for the pseudopotential band-structure calculations. We wish to thank A. B. Fowler for many discussions, 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 appreciate the constant support of F. Stern and A. R. Williams, and the help received from J. Wells, W. G. Pope, A. Rossi, and other members of the IBM Thomas J. Watson Research Center Computing Systems staff during our long computer runs. The manuscript has been patiently read by M. Artaki, D. J. DiMaria, A. B. Fowler, A. Gnudi, J. Higman, P. J. Price, F. Stern, and A. R. Williams.

---

[1]See, for instance, S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed. (Wiley, New York, 1981), p. 50.

[2]It is impossible to review here the wealth of literature on the subject of quantum transport. We may refer to general papers on the various approaches. The rigorous derivation of transport equations is described by W. Kohn and J. M. Luttinger, Phys. Rev. **108**, 590 (1957) and L. Van Hove, Physica **XXI**, 517 (1955). Approaches based on the Feynman path-integral techniques stem from the work reviewed in R. P. Feynman and F. L. Vernon, Jr., Ann. Phys. (N.Y.) **24**, 118 (1963), reviewed by K. K. Thornber, Solid-State Electron. **21**, 259 (1978), and recently applied to high-field transport in insulators by M. V. Fischetti and D. J. DiMaria, Phys. Rev. Lett. **55**, 2475 (1985). The Green's-function techniques are described in L. P. Kadanoff and G. Baym, *Quantum Statistical Mechanics* (Benjamin, New York, 1962) and developed by A. P. Jauho and J. W. Wilkins, Phys. Rev. B **29**, 1919 (1984); W. Hänsch and G. D. Mahan, *ibid.* **28**, 1920 (1983); Sanjoy K. Sarker, *ibid.* **32**, 743 (1985); L. Reggiani, P. Lugli, and P. Jauho, *ibid.* **36**, 6602 (1987). The density-matrix approach has been recently investigated by R. Brunetti and C. Jacoboni, in *Proceedings of the 18th International Conference on the Physics of Semiconductors*, edited by O. Engström (World Scientific, Singapore, 1987), p. 1527, and in *Proceedings of the 5th International Conference on Hot Carriers in Semiconductors*, edited by J. Shah and G. Iafrate [Solid-State Electron. **31**, 527 (1988)]. The utility of a Wigner-function approach to tackle the simulation of resonant tunneling devices has been demonstrated by U. Ravaioli, M. A. Osman, W. Pötz, N. Klusksdahl, and D. K. Ferry, Physica B + C **134B**, 36 (1985) and applied by W. R. Frensley, Phys. Rev. Lett. **57**, 2853 (1986), while the resolvent superoperator technique has been used by J. R. Barker, J. Phys. C **6**, 2663 (1973).

[3]See the early descriptions by K. Blotekjaer, IEEE Trans. Electron Devices **ED-17**, 38 (1970), R. K. Cook and J. Frey, IEEE Trans. Electron Devices **ED-29**, 970 (1982). Recent examples of rigorous derivations and numerical implementation of two-dimensional hydrodynamic models are described by M. Rudan and F. Odeh, COMPEL **5**, 149 (1986) and A. Forghieri, R. Guerrieri, P. Ciampolini, A. Gnudi, M. Rudan, and G. Baccarani, IEEE Trans. Computer-Aided Design **CAD-7**, 231 (1988).

[4]P. J. Price, Semicond. Semimet. **14**, 249 (1979).

[5]C. Jacoboni and L. Reggiani, Rev. Mod. Phys. **55**, 645 (1983), and references therein.

[6]Again, it is impossible to review all of the applications of the MC technique to device simulation. A complete description is given in the book of R. W. Hockney and J. W. Eastwood, *Computer Simulation Using Particles* (McGraw-Hill, New York, 1981). A recent review is given by C. Moglestue, IEEE Trans. Computer-Aided Design **CAD-5**, 326 (1986). The pioneering activity of these authors has been followed most notably by, e.g., P. Hesto, J.-F. Pone, M. Mouis, J.-L. Pelouard, and R. Castagné, in *Nasecode IV, Proceedings of the Fourth International Conference on the Numerical Analysis of Semiconductor Devices and Integrated Circuits,* edited by J. J. H. Miller (Boole, Dublin, 1985), p. 315, for the ensemble Monte Carlo simulation of GaAs metal-semiconductor field-effect transistor (MESFET's); by e.g., Chu Hao, J. Zimmermann, M. Charef, R. Fauquembergue, and E. Constant, Solid-State Electron. **28**, 733 (1985) for the simulation of electron transport in the inversion layer of Si MOSFET's including quantization effects; by, e.g, T. Wang and K. Hess, J. Appl. Phys. **57**, 5336 (1985), who also include the short-range electron-electron interaction and by, e.g., U. Ravaioli and D. K. Ferry, IEEE Trans. Electron Devices **ED-33**, 677 (1986), who also account for quantization in the inversion layer of GaAs high-electron-mobility transistors (HEMT's).

[7]In addition to some of the publications of Ref. 6, quantum size effects have been also tackled by K. Yokoyama and K. Hess, J. Appl. Phys. **59**, 3798 (1986); M. Artaki and K. Hess, Phys. Rev. B **37**, 2933 (1987); M. A. R. Mudares and B. K. Ridley, J. Phys. C **19**, 3179 (1986); K. Tomizawa and N. Hashizume, IEEE Trans. Electron Devices **ED-35**, 849 (1988).

[8]The regime of "mesoscopic" phenomena is a notable exception to our oversimplification. This is discussed from the experimental point of view by S. Washburn and R. A. Webb, Adv. Phys. **35**, 375 (1986). Theoretical aspects are reviewed by P. A. Lee, A. D. Stone, and H. Fukuyama, Phys. Rev. B **35**, 1039 (1987), and references therein.

[9]It may suffice to recall the controversies on the effects of the so-called *intracollisional fields* [which, for parabolic bands are predicted to be significant by J. R. Barker, Solid-State Electron. **21**, 267 (1978) or negligible up to 10^7 V/cm by F. S. Kahn and J. W. Wilkins, Semicond. Sci. Technol. **1**, 113 (1985) and F. Beleznay, J. Phys. C **19**, L447 (1986)], or on the *collisional broadening effects* [first pointed out by F. Capasso, T. P. Pearsall, and K. K. Thornber, IEEE Electron Device Lett. **EDL-2**, 295 (1981), included in a semiclassical MC simulation by Y. C. Chang, D. Z.-Y. Ting, T. Y. Tang, and K. Hess, Appl. Phys. Lett. **42**, 76 (1983), later criticized by J. Lin and L. C. Chiu, *ibid.* **49**, 1802 (1986), and recently revived by K. Kim, B. A. Mason, and K. Hess, Phys. Rev. B **36**, 6547 (1987)].

[10]J. Y. Tang and K. Hess, J. Appl. Phys. **54**, 5139 (1983).

[11]H. Shichijo and K. Hess, Phys. Rev. B **23**, 4197 (1981).

[12]See, for instance, S. Tam, F.-C. Hsu, C. Hu, R. S. Muller, and P. K. Ko, IEEE Electron Device Lett. **EDL-4**, 249 (1983).

[13]E. Sangiorgi, B. Riccó, and F. Venturi, IEEE Trans. Computer-Aided Design, **CAD-7**, 259 (1988).

[14]E. Sangiorgi, M. Pinto, F. Venturi, and W. Fichnter, IEEE Electron Device Lett. **ED-9**, 13 (1988).

[15]A. Al-Omar and J. P. Krusius, J. Appl. Phys. **62**, 3825 (1987).

[16]Arguments supporting this conclusion have been given by R. E. Peierls, *Quantum Theory of Solids* (Clarendon, Oxford, 1960) and by L. D. Landau, as discussed in R. E. Peierls, Helv. Phys. Acta **7** (Suppl.), **24** (1934).

[17]M. L. Cohen and T. K. Bergstresser, Phys. Rev. **141**, 789 (1966).

[18]*CRC Handbook of Tables for Mathematics,* 4th ed., edited by R. C. Weast (CRC, Cleveland, 1975), p. 640.

[19]James R. Chelikowsky and Marvin L. Cohen, Phys. Rev. B **14**, 556 (1976); R. R. L. Zucca and Y. R. Shen, *ibid.* **1**, 2668 (1970).

[20]See, for example, J. M. Ziman, *Electrons and Phonons* (Oxford University Press, Oxford, 1974). We shall ignore here the effect of screening on the polar and nonpolar electron-phonon matrix element. A discussion of this effect on the acoustic deformation potential of GaAs has been given by P. J. Price, Phys. Rev. B **32**, 2643 (1985).

[21]Self-consistent MC schemes to account for screening in the electron-electron interaction have been proposed by P. Lugli and D. K. Ferry, Phys. Rev. Lett. **56**, 1295 (1986) with a molecular-dynamics technique and by M. A. Osman and D. K. Ferry, Phys. Rev. B **36**, 6018 (1987). We found these techniques very difficult to implement in space- and time-dependent situations.

[22]D. Chattopadhyay and H. J. Queisser, Rev. Mod. Phys. **53**, 745 (1981); G. Berthold and P. Kocevar, J. Phys. C **17**, 4981 (1984).

[23]E. O. Kane, Phys. Rev. **159**, 624 (1967).

[24]H. Fröhlich, Proc. R. Soc. London Ser. A **160**, 230 (1937); **172**, 94 (1239); Adv. Phys. **3**, 325 (1954).

[25]G. Gilat and L. J. Raubenheimer, Phys. Rev. **144**, 390 (1966).

[26]G. Nilsson and G. Nelin, Phys. Rev. B **6**, 3777 (1972).

[27]M. A. Littlejohn, J. R. Hauser, and T. H. Glisson, J. Appl. Phys. **48**, 4587 (1977).

[28]H. Brooks and C. Herring, Phys. Rev. **83**, 879 (1951).

[29]This actually corresponds to the Debye-Hückel screening in nondegenerate situations. In the degenerate case $k_B T_{el} \approx 2E_F/3$, so that our approximation corresponds to the more appropriate Thomas-Fermi screening, $\beta_s^2 = 3e^2 n_{el}/(2\epsilon E_F)$, which is close to the general expression $\beta_s^2 = e^2(\partial n_{el}/\partial E_F)/\epsilon$.

[30]B. K. Ridley, J. Phys. C **10**, 1589 (1977); T. G. Van de Roer and F. P. Widdershoven, J. Appl. Phys. **59**, 813 (1986).

[31]This distinction between short-range and long-range interaction corresponds to the distinction between the large-q single-particle and the small-q collective modes which an electron can excite, as discussed by D. Pines, *Elementary Excitations in Solids* (Benjamin, New York, 1963). This has been implemented in homogeneous and steady-state MC simulations by P. Lugli and D. K. Ferry, Physica B + C **129B**, 532 (1985). Doubts about the possibility of actually making this distinction at low densities have been raised recently by J. M. Rorrison and D. C. Herbert, J. Phys. C **19**, 3991 (1986).

[32]A. Matulionis, J. Pozela, and A. Reklatis, Solid State Commun. **16**, 1133 (1975).

[33]L. V. Keldysh, Zh. Eksp. Teor. Fiz. **48**, 1692 (1965) [Sov. Phys.—JETP **21**, 1135 (1965)].

[34]T. P. Pearsall, F. Capasso, R. E. Nahory, M. A. Pollack, and J. R. Chelikowsky, Solid State Electron. **21**, 297 (1978).

[35]P. Lugli and D. K. Ferry, IEEE Trans. Electron. Devices, **ED-32**, 2431 (1985).

[36]N. S. Wingreen, C. J. Stanton, and J. W. Wilkins, Phys. Rev. Lett. **57**, 1084 (1986); N. Takenaka, M. Inoue, and Y. Inuishi, J. Phys. Soc. Jpn. **47**, 861 (1979); P. Lugli and D. K. Ferry, Physica B + C **117&118B**, 251 (1983); R. Brunetti, C. Jacoboni, A. Matulionis, and V. Dienys, Physica B + C **134B**, 369 (1985).

[37]J. Bardeen and W. Shockley, Phys. Rev. **80**, 72 (1950).
[38]W. A. Harrison, Phys. Rev. **104**, 1281 (1956).
[39]C. Canali, C. Jacoboni, F. Nava, G. Ottaviani, and A. Alberigi-Quaranta, Phys. Rev. B **12**, 2265 (1975); C. Jacoboni, C. Canali, G. Ottaviani, and A. Alberigi-Quaranta, Solid State Electron. **20**, 77 (1977).
[40]T. H. Ning and H. N. Yu, J. Appl. Phys. **45**, 5373 (1974).
[41]J. Y. Tang and K. Hess, J. Appl. Phys. **54**, 5145 (1983).
[42]Z. A. Weinberg and A. Harstein, Solid State Commun. **20**, 179 91976); G. Binnig, N. Garcia, and H. Rhorer, Phys. Rev. B **30**, 4816 (1984).
[43]D. J. DiMaria (unpublished).
[44]C. A. Lee, R. A. Logan, R. L. Batdorf, J. J. Kleimack, and W. Wiegmann, Phys. Rev. **134**, A761 (1964); C. R. Crowell and S. M. Sze, Appl. Phys. Lett. **9**, 242 (1966); R. van Overstraeten and H. DeMan, Solid State Electron. **13**, 583 (1970).
[45]M. Heiblum, M. V. Fischetti, W. P. Dumke, D. J. Frank, I. M. Anderson, C. M. Knoedler, and L. Osterling, Phys. Rev. Lett. **58**, 816 (1987).
[46]J. G. Ruch and G. S. Kino, Phys. Rev. **174**, 921 (1968); P. A. Houston and A. G. R. Evans, Solid State Electron. **20**, 197 (1977).
[47]H. D. Law and C. A. Lee, Solid State Electron. **21**, 331 (1978); S. N. Shabde and C. Yeh, J. Appl. Phys. **41**, 4743 (1970).
[48]W. Fawcett, A. D. Boardman, and S. Swain, J. Phys. Chem. Solids **31**, 1963 (1970).
[49]R. E. Bank and D. J. Rose, SIAM J. Numer. Anal. **17**, 806 (1980).
[50]O. G. Johnson, C. A. Micchelli, and G. Paul, SIAM J. Numer. Anal. **20**, 362 (1983).
[51]M. Tomizawa, K. Yokoyama, and A. Yoshii, IEEE Trans. Computer-Aided Design **CAD-7**, 254 (1988).
[52]One of us (S.E.L.) is grateful to Anita Mayo for enlightening discussions resulting in the ideas embodied in Eq. (20).
[53]A. Phillips and P. J. Price, Appl. Phys. Lett. **30**, 528 (1977).
[54]K. Throngnumchai, K. Asada, and T. Sugano, IEEE Trans. Electron Devices **ED-33**, 1005 (1986).
[55]F. Venturi, R. K. Smith, E. Sangiorgi, M. R. Pinto, and B. Riccó (unpublished).
[56]These considerations address problems originating from the two-dimensional nature of the Poisson equation which we solve. Our first concern is to reproduce correctly the *classical* energy losses to collective modes in the degenerate regions. The condition $s \lesssim \Delta x^{-1}$ is necessary in order to obtain this result. Moreover, it yields the correct *classical* thermal population of the density fluctuations associated with these modes. What the two-dimensional nature of the Poisson solution prevents us from doing is to relax correctly the momentum of the carriers. Only a full three-dimensional simulation can account for momentum relaxation in the "missing" dimension. Furthermore, our treatment of the long-range Coulomb coupling is semiclassical. Hence, strictly speaking, our use of the work *plasmon* (quantized plasma oscillations) is inappropriate. In our approach, the amplitude of the density fluctuations corresponds to the Bose-Einstein population of plasmons in the quantum formulation.
[57]D. Bohm and D. Pines, Phys. Rev. **82**, 625 (1951); **85**, 338 (1952); **92**, 609 (1953).
[58]G. A. Sai-Halasz, M. R. Wordeman, D. P. Kern, E. Ganin, S. Rishton, D. S. Zicherman, H. Schmid, M. R. Polcari, H. Y. Ng, P. J. Restle, T. H. P. Chang, and R. H. Dennard, IEEE Electron Device Lett. **EDL-8**, 463 (1987).
[59]S. E. Laux and M. V. Fischetti, IEEE Electron Device Lett. **EDL-9**, 467 (1988).
[60]A saturated velocity of about 6×10^6 cm/sec at room temperature has been measured by F. F. Fang and A. B. Fowler, J. Appl. Phys. **41**, 1825 (1970) and more recently by A. Modelli and S. Manzini, Solid State Electron. **31**, 99 (1988). This later paper references other experimental works yielding different results.
[61]F. Stern and W. E. Howard, Phys. Rev. **163**, 816 (1967); T. H. Ning and C. T. Sah, Phys. Rev. B **6**, 4605 (1972); S. Manzini, J. Appl. Phys. **57**, 411 (1985).
[62]Y. Park, T. Tang, and D. H. Navon, IEEE Trans. Electron Devices **ED-30**, 1110 (1983).
[63]K. Hess and C. T. Sah, J. Appl. Phys. **45**, 1254 (1974); P. K. Basu, *ibid.* **48**, 350 (1977); Solid State Commun. **27**, 657 (1978).
[64]G. Sai-Halasz, M. R. Wordeman, S. Rishton, E. Ganin, and D. P. Kern, IEEE Electron Device Lett. **EDL-9**, 464 (1988).
[65]G. S. Shahidi, D. A. Antoniadis, and H. I. Smith, IEEE Electron Device Lett. **ED-9**, 94 (1988).
[66]A. Gnudi (private communication).
[67]L. Reggiani, in *Proceedings of the 15th International Conference on the Physics of Semiconductors* [J. Phys. Soc. Jpn. Suppl. A **49**, 317 (1980)]; R. Brunetti, C. Jacoboni, F. Nava, L. Reggiani, G. Bosman, and R. J. J. Zijlstra, J. Appl. Phys. **52**, 6713 (1981).
[68]C. Jacoboni, R. Minder, and G. Maini, J. Phys. Chem. Solids **36**, 1129 (1975).