RC 21795 (7/18/00)
Physical Sciences

IBM Research Report

Long-range Coulomb interactions in small Si devices. PartI: Performance and reliability

M. V. Fischetti and S. E. Laux

IBM Research Division
T.J. Watson Research Center
Yorktown Heights, New York

LIMITED DISTRIBUTION NOTICE

This report has been submitted for publication outside of IBM and will probably be copyrighted if accepted for publication. It has been issued as a Research Report for early dissemination of its contents. In view of the transfer of copyright to the outside publisher, its distribution outside of IBM prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filled only by reprints or legally obtained copies of the article (e.g., payment of royalties).

IBM Research Division
Almaden · T.J. Watson · Tokyo · Zurich

Long-range Coulomb interactions in small Si devices.
Part I: Performance and reliability

M. V. Fischetti and S. E. Laux
IBM Research Division, Thomas J. Watson Research Center
P. O. Box 218, Yorktown Heights, NY 10598, USA
(July 20, 2000)

In the ever smaller silicon metal-oxide-semiconductor field-effect transistors of the present technology, electrons in the conductive channel are subject to increasingly stronger long-range Coulomb interactions with high-density electron gases present in the source, drain, and gate regions. We first discuss how two-dimensional, self-consistent full-band Monte Carlo/Poisson simulations can be tailored to reproduce correctly the semiclassical behavior of a high-density electron gas. We then employ these simulations to show that for devices with channel lengths shorter than about 40 nm and oxides thinner than about 2.5 nm, the long-range Coulomb interactions cause a significant reduction of the electron velocity, and so a degradation of the performance of the devices. In addition, the strong 'thermalization' of the hot-electron energy distribution induced by Coulomb interactions has an effect on the expected reliability of the transistors.

I. INTRODUCTION

Conventional scaling rules[1] for small silicon metal-oxide-semiconductor field-effect transistors (MOSFETs) demand increasing doping concentrations in the substrate, drain, and source regions and a reduction of the insulator thickness. This implies that, as devices are scaled, electrons in the conducting channel of Si n-MOSFETs are in ever closer proximity to the high-density electron gases present in the source and drain regions – separated from each other by as little as tens of nanometers – and in the polycristalline Si gate – separated from the channel by as little as 1.5 nm of SiO_2 (Ref. 2). It follows that long-range Coulomb interactions between the channel and the heavily-doped regions are expected to acquire increasing importance as devices are scaled to smaller dimensions. Short-range electron-electron interactions have been shown to play a non-negligible role in determining the energy distributions of the hot channel electrons, particularly near the channel/drain junction and at low applied bias.[3–5] While these effects have an impact on the 'secondary' characteristics of the devices (such as substrate currents and, possibly, reliability), they do not affect the raw performance (such as mobility, current drive, and transconductance). On the contrary, the purpose of this paper and of its companion[6,7] is to show that long-range interactions are expected to degrade (with respect to more naïve expectations) the transconductance and electron mobility of small devices: For short channels, the interaction between the heavily-doped source/drain regions and the channel, while not directing subtracting momentum from the electron gas, induces a very strong and fast 'thermalization' of the energy distributions of the hot electrons in the channel. Electrons in the stronger high-energy tails suffer more frequent collisions with momentum dissipating scatterers (phonons, ionized impurities, band-to-band ionization), thus losing velocity. This results in a reduced drain current and transconductance. In structures with thin oxides, the interaction between the gate and the channel, on the contrary, causes a direct transfer of momentum from the channel to the gate, which also affects the performance of the device in obvious ways.

We proceed as follows: Since our results are derived using *two-dimensional* and *semiclassical* self-consistent Monte Carlo/Poisson simulations[8], in Sec. II we discuss the expected behavior of a homogeneous semiclassical high-density electron gas, we argue that quantum corrections, while not negligible, are relatively small, and show in Sec. III that two-dimensional (2D) simulations can be tailored so to reproduce the three-dimensional (3D) results. We also discuss how at high electron densities the electron kinetic-energy distributions, while exhibiting non-thermal high-energy tails, are actually consistent with thermodynamic equilibrium. In Sec. IV we briefly justify the applicability of 2D, semiclassical simulations to the highly inhomogeneous situations encountered in Si n-MOSFETs, and show that long-range Coulomb interactions degrade the 'effective' electron velocity in the channel of scaled devices at high source-to-drain bias. We also discuss the impact of Coulomb interactions on the expected hot-electron reliability. We hopefully render the discussion more agile by discussing separately in three appendices some algebraically cumbersome but necessary details about the normalization of the plasmon field (Appendix A), about the derivation of the unscreened and screened Green's function for the Poisson equation in the semiconductor-insulator-semiconductor geometry (Appendix B), and about the scattering rates for short-range Coulomb interactions between 2D electrons in the substrate (both in the 2D and bulk models) and bulk electrons in the gate (Appendix C).

Despite our conclusions, reached in Sec. II, that semiclassical and 2D simulations can mimic satisfactorily the 'correct' behavior of a high-density electron gas, it is desirable to confirm our results with a three-dimensional, quantum model. This is done in a companion paper[6] investigating the effect of quantized interface plasmons in an MOS geometry on the electron mobility in Si inversion layers. While the 2D, semiclassical simulations employed here permit the study of off-equilibrium, highly inhomogeneous, and high-bias situations, Ref. 6 must be

confined to the study of the homogeneous linear regime, but it complements the results obtained here by avoiding some possible shortcomings of 2D Monte Carlo/Poisson simulations by accounting for: *i)* the quantization of the plasmon field, with the associated quantized energy exchanged in long-range Coulomb interactions, *ii)* the Bose-Einstein, as opposite to Maxwell-Boltzmann, thermal population of plasmons, and *iii)* the possible difference between 3D and 2D models concerning the length-scale over which plasma oscillations decay and time-scale over which the electronic momentum relaxes.

II. HOMOGENEOUS ELECTRON GAS

A. The three-dimensional semiclassical electron gas

Consider a homogeneous electron gas of density $n = N/\Omega$, where N is the number of electrons and Ω the volume occupied by the gas. Ignoring Coulomb interactions, the total energy W per electron is purely the Fermi kinetic energy K_F:

$$W \ = \ K_F \ = \ \frac{2}{n} \int \frac{d\mathbf{k}}{(2\pi)^3} \ E_0(\mathbf{k}) \ f_{FD}[E_0(\mathbf{k})] \ , \quad (1)$$

where $f_{FD}(E) = \{1+\exp[(E_0-E_F)/(k_BT)]\}^{-1}$, is the equilibrium Fermi function at temperature T, k_B is the Boltzmann constant, $E_0(\mathbf{k})$ is the electron dispersion, and the Fermi energy E_F is determined so that

$$n \ = \ 2 \int \frac{d\mathbf{k}}{(2\pi)^3} \ f_{FD}[E_0(\mathbf{k})] \ . \quad (2)$$

In the simplest possible picture, the Hartee model, the spatial distribution of both electrons and background positive charge (donors), required by charge neutrality, is assumed to be homogenous and uniform. In this model the Coulomb interaction between electrons and donors vanishes. The first correction to this model consists in what Kittel calls the 'modified Hartree' approximation[9]. In this model one assumes that the electrons still constitute a uniform distribution of charge, but they are now embedded in a lattice of background, positively charged, pointlike donors. Customarily, one thinks of the donor potential as statically screened by the electrons according to the Debye-Hückel theory, resulting in a Yukawa-like potential $e^{-\beta_s r}/(4\pi\epsilon_s^\infty r)$, where $\beta_s = [(e^2/\epsilon_s^\infty)(\partial n/\partial E_F)]^{1/2}$ is the screening parameter, ϵ_s^∞ being the static (valence) permittivity of Si, and e the magnitude of the electron charge. However, Mahan[10], employing quantum-mechanical variational calculations, has shown that this picture is correct only at moderate densities ($\sim 10^{19}$ cm^{-3} or below in Si), but at higher densities the assumption of a uniform electron distribution gives better agreement with the variational results. This is because the Yukawa-model overcounts the electrons: On the one hand, they are assumed to completely screen the donors, and so to be fully localized around each dopant ion. On the other hand, they are assumed to be uniformly distributed.

In order to compute the Coulomb energy of this model system, in principle one should follow a procedure similar to what is required to compute the Madelung energy in crystals and derive the positions of the donors by minimizing their total energy. Here we assume the impurities frozen in a cubic lattice, and consider spherical volumes of radius $r_0 = [3/(4\pi n)]^{1/3}$ around each donor. Thus, the total energy per dopant (*i.e.*, in a volume $(4/3)\pi r_0^3$ around each impurity) of the electron gas and background pointlike donors will be lowered by the electron-impurity Coulomb attraction,

$$\delta W_{ed} \ \approx \ -\frac{e^2}{(4/3)\pi r_0^3} \int_0^{r_0} dr \frac{r}{\epsilon_s^\infty} \ = \ -\frac{3}{2}\frac{e^2}{4\pi\epsilon_s^\infty r_0}$$
$$= \ -\frac{3}{2}\left(\frac{4\pi}{3}\right)^{1/3} \frac{e^2 n^{1/3}}{4\pi\epsilon_s^\infty} \approx -2.418 \ \kappa \ n^{1/3} \ , \quad (3)$$

where $\kappa = e^2/(4\pi\epsilon_s^\infty)$, and raised by the Coulomb repulsion of the uniform electron distribution with itself within an elemental spherical volume ,

$$\delta W_{ee} \ \approx \ \frac{e^2}{[(4/3)\pi r_0^3]^2} \int_0^{r_0} dr \frac{4}{3}\pi r^3 \ 4\pi r^2 \ \frac{1}{4\pi\epsilon_s^\infty r}$$
$$= \frac{3}{5}\frac{e^2}{4\pi\epsilon_s^\infty r_0} \ = \frac{3}{5}\left(\frac{4\pi}{3}\right)^{1/3} \frac{e^2 n^{1/3}}{4\pi\epsilon_s^\infty} \ \approx \ 0.967 \ \kappa \ n^{1/3} \ . \quad (4)$$

Note that the mutual donor-donor repulsion vanishes in this model, since the impurities in different volume elements are viewed as fully screened by the surrounding electrons when frozen in their minimum-energy configuration. Therefore, the total energy per impurity atom due to the Coulomb interaction will be:

$$\delta W_C \ = \ \delta W_{ed} \ + \ \delta W_{ee} \ \approx \ -1.451 \ \kappa \ n^{1/3} \ . \quad (5)$$

Depending on the configuration of the impurities, one may actually obtain for the 'Madelung coefficient' 1.451 above a slightly different value of 1.444 for close-packed fcc, hcp, or bcc lattices[10,11]. Here, we shall not worry about these numerically small differences.

At densities low enough for the standard Debye-Hückel screening to be valid, an expression given by Lanyon and Tuft[12] may more appropriately replace Eq. (5),

$$\delta W_C \ = \ -\frac{3}{4} \ \kappa \ \beta_s \ . \quad (6)$$

As discussed by Mahan[10], the numerical values provided by Eqns. (5) and (6) are very similar in the range of densities of interest here, since the three characteristic lengths one can consider, the inverse wavevector at the Fermi surface k_F, the screening length β_s^{-1}, and the average donor separations r_0, are not too different in Si.

In the presence of random motion, induced by the Fermi kinetic energy and, at nonzero temperatures, by the electron-phonon interactions, the electrons will move around their 'frozen' uniform configuration and set-up a fluctuating potential whose spatial Fourier components are identified with plasmons. We shall discuss these 'potential fluctuations' below in Sec. III B. Here we are interested in the net result of this thermal 'agitation' : The total energy δW_C will redistribute into a potential energy,

δU_C, and a kinetic energy component δK_C. The virial theorem[13] can be used to estimate the time-average of the kinetic energy δK_C once δW_C is known. Since for N particles interacting via a central potential $V(r)$, a function only of the interparticle distance r,

$$< \delta K_C > = \frac{1}{2} < \sum_{i \neq j} \nabla_i V(\mathbf{r}_i - \mathbf{r}_j) \cdot \mathbf{r}_i > , \quad (7)$$

where the brackets $< ... >$ denote time-average, for a purely Coulomb system

$$< \delta K_C > = -\frac{1}{2} < \delta U_C > , \quad (8)$$

so that (dropping the time-average brackets here and in the following, since we shall deal exclusively with time- or ensemble-average, equivalent by the ergodic theorem)

$$\delta K_C \approx 1.451 \, \kappa \, n^{1/3} . \quad (9)$$

The final quantity to consider is the shift of the chemical potential, $\delta\mu$, (in principle different from the Fermi energy E_F for interacting particles, as discussed below) induced by the Coulomb interactions among the N electrons and the N_D donors. At zero temperature or, more generally, in degenerate situations, $\delta\mu$ gives the electron contribution to the narrowing of the energy gap. In general the chemical potential μ of a grand canonical ensemble is defined as:[14]

$$\mu = \frac{1}{\Omega} \frac{\partial F}{\partial n} , \quad (10)$$

at constant temperature and volume Ω, where $F = W_{tot} - TS$ is the (Helmholtz) free energy of the system, W_{tot} being its total energy and S its entropy. Therefore, at zero temperature the change of the chemical potential caused by the Coulomb interactions will be:

$$\delta\mu = \frac{1}{\Omega} \frac{\partial \delta W_{C,tot}}{\partial n} , \quad (11)$$

where $\delta W_{C,tot} = N_D \delta W_C$ is the total Coulomb energy of the system. Keeping the volume Ω constant means that the identity $N_D = N$ should be used only after having taken the derivative. Thus

$$\delta\mu = \frac{\partial}{\partial N}(N_D \, \delta W_C) = -0.481 \, \kappa \, n^{1/3} . \quad (12)$$

Ignoring band-tailing and renormalization of the electron dispersion, since the total number of particles remains constant when accounting for Coulomb interactions, this shift of the chemical potential must be accompanied by an equal downward shift of the reference energy, that is, the bottom of the conduction band, $\delta E_{CB} = \delta\mu$.

B. Quantum corrections

The modified Hartree model considered so far ignores the effect of the electron spin: Since Pauli's principle prevents electrons with parallel spins from getting too close to each other, their mutual Coulomb repulsion will be reduced by an amount known as exchange energy, δW_x. Following Haas[15] and Mahan[10], for Si the exchange energy per particle is given by

$$\delta W_x = -\frac{3}{4} \frac{e^2 \Lambda k_F}{4\pi^2 \epsilon_s^{\infty}} \approx -0.3861 \, \kappa \, n^{1/3} , \quad (13)$$

where $\Lambda = (m_l/m_t)^{1/3} \, (\tan^{-1} \delta/\delta)$, $\delta = [(m_l/m_t) - 1]^{1/2}$, $m_{t/l}$ denoting the transverse/longitudinal mass in the six ellipsoidal valleys of the Si conduction band. This second correction amounts to the Hartree-Fock results. The difference between this result and the energy of the ground state one would obtain performing an 'exact' calculation is called 'correlation' energy: The electrons are not uniformly distributed, but the energy of the system is lowered when accounting for a correlation of electron wavefunctions (or positions) which minimizes their mutual Coulomb repulsion. In the range of densities considered here[16], an expression for the correlation energy, δW_{corr} has been given by Gell-Mann and Brueckner[17],

$$\delta W_{corr} = -0.096 + 0.00622 \ln r_s \quad (14)$$

while Nozières and Pines[18] give

$$\delta W_{corr} = -0.115 + 0.0031 \ln r_s , \quad (15)$$

where the unit of energy is the Rydberg (≈ 32.1 meV in Si) and r_s is the average separation r_0 in units of the Bohr radius (≈ 1.92 nm in Si), so that $r_s \approx 3.231 \, (n/10^{18})^{-1/3}$, where the electron density n is measured in cm^{-3}. Re-expressing Eq. (5) in atomic units, we have

$$\delta W_C = -\frac{1.798}{r_s} , \quad (16)$$

while the exchange term, Eq. (13), is

$$\delta W_x = -\frac{0.479}{r_s} , \quad (17)$$

so we see that the correlation energy can be neglected, while ignoring the exchange term, Eq. (13), amounts to underestimating the Coulomb energy by about 26%. Correspondingly, the kinetic energy of the system will be modified only by a small term, usually called 'correlation kinetic energy'[19].

Finally, note that these corrections to the total energy of the electron gas at high densities are some of the causes of the well-known band-gap narrowing in semiconductors: We have already noticed above how the Coulomb interactions in a classical three-dimensional electron gas cause a shift $\delta E_{CB} = -0.481\kappa n^{1/3}$ of the bottom of the conduction band. Mahan[10] has evaluated the shifts of the chemical potential $\delta\mu_x = [e^2/(4\pi^2\epsilon_s^{\infty})] k_F \Lambda \approx -0.515 \, \kappa \, n^{1/3}$ and $\delta\mu_{corr}$ associated to exchange and correlation effects, and so the corresponding shifts δE_{CB}. When accounting also for the contribution to the Coulomb energies due to the (minority) hole in the valence band (hole-donor and hole-electron interaction), Mahan[10] has accounted for all of the contributions to the band-gap narrowing, with the exception of band-tailing, discussed below. Clearly, the semiclassical model errs in ignoring exchange and correlation. As

for its small contribution to the total energy, correlation can be safely ignored. Not so for exchange terms: Since the shift $\delta\mu$ due to the electron-donor and the hole-donor interaction cancel, exchange is now a dominat contribution.

C. Two-dimensional approximation

In a numerical simulation employing a two-dimensional solution of Poisson equation, we typically encounter a situation complementary to what described above. Now electrons are idealized as 'superparticles' actually represented by line-charges of statistical weight s (i.e., of charge density $-es$ per unit length), while the donors are described as a uniform background distribution of positive charge ('jellium'). Again, we can estimate the Coulomb energy of the system by first assuming that the electrons are distributed in a square lattice configuration and are assigned the minimum kinetic energy compatible with Pauli's principle (that is, their momenta are chosen from a Fermi-Dirac distribution). Thus, a snapshot of the electron-jellium/donor system in this initial lowest-energy configuration shows the electrons sitting at the sites of a regular cubic lattice, and the total energy of the system is given by the electron Fermi energy plus the Coulomb energy we are about to discuss. Employing mesh-elements of side-length Δx to discretize the Poisson equation, each element being centered around each particle, obviously must have $n = s/\Delta x^2$. In the spirit of the discussion above, we consider an average interparticle separation R such that $\pi R^2 = \Delta x^2$. Therefore, in each area element there is one 'particle' with an associated potential $\phi^{(2D)}(r)$ which vanishes for $r > R$, while

$$\phi^{(2D)}(r) \;=\; \frac{es}{2\pi\epsilon_s^\infty}\; ln\left(\frac{r}{R}\right) \;, \tag{18}$$

for $r \leq R$. Since Monte Carlo simulations treat electrons in 3 dimensions as far as their motion is concerned, the Fermi kinetic energy is given by Eq. (1) above. The change of total energy per unit length due to the Coulomb interaction between the particles and the background positive charge will be:

$$\begin{aligned}\delta W_{ed}^{(2D)} \;&=\; \int_0^R dr 2\pi r\; \frac{es}{\pi R^2}\; es\; \frac{1}{2\pi\epsilon_s^\infty}\; ln\left(\frac{r}{R}\right) \\ &=\; \frac{e^2 s^2}{\pi R^2 \epsilon_s^\infty}\int_0^R dr\; r\; ln\left(\frac{r}{R}\right) \;=\; -\kappa\; s^2 \;,\end{aligned} \tag{19}$$

while the jellium-jellium repulsive interaction gives a contribution

$$\begin{aligned}\delta W_{dd}^{(2D)} \;&=\; -\left(\frac{es}{\pi R^2}\right)^2 \int_0^R dr\; 2\pi r\; \pi r^2\; \frac{1}{2\pi\epsilon_s^\infty}\; ln\left(\frac{r}{R}\right) \\ &=\; \frac{1}{4}\;\kappa\; s^2 \;,\end{aligned} \tag{20}$$

so that the change of total energy *per electron* due to the Coulomb interaction is

$$\delta W_C^{(2D)} \;=\; -\frac{3}{4}\;\kappa\; s \;. \tag{21}$$

By requiring

$$s \;=\; \frac{6}{5}\; r_0^{-1} \;, \tag{22}$$

and

$$\Delta x \;=\; \left(\frac{8\pi}{5}\right)^{1/2}\; r_0 \;\approx\; 2.242\; r_0 \;, \tag{23}$$

we recover the result of Eq. (5). In an actual simulation the mesh employed to discretize the Poisson equation obviously cannot reproduce the 'spherical' arrangement consider so far. When employing a uniform square mesh, the conditions given by Eqns. (23) and (22) may be satisfied approximately by chosing a mesh spacing $\Delta x \sim n^{-1/3}$ and a 'statistical weight' $s \sim n^{1/3}$.

Note one major difference between two models: In the three-dimensional gas we have assumed pointlike donors embedded in a uniform distribution of electrons. On the other hand, the actual two-dimensional simulation embeds line-charges into the positive background jellium. Even if the expressions for the total energy, Eqns. (5) and (21) are made to agree numerically with the proper choice of s and Δx, a crucial difference remains: When evaluating the shift of the chemical potential, as in Eq. (11), we must keep in mind that in the two-dimensional simulation the energy $\delta W_C^{(2D)}$ is now the energy *per electron*, so that Eq. (12) is replaced by:

$$\delta\mu \;=\; \frac{\partial}{\partial N}(N\delta W_C^{(2D)}) \;=\; -1.935\;\kappa\; n^{1/3} \;. \tag{24}$$

In order to relate the results of the present discussion to the band-gap narrowing effect, we must account for the energy change of the minority holes due to their interactions with the electrons and the donors. Often, when simulating electron transport with the MC algorithm, hole transport is treated by a simpler drift-diffusion approximation which regards holes as a fluid. In our homogeneous case, the Coulomb energy of this uniform jellium of holes $\delta W_{C,h}^{(2D)}$, is due to the sum of the repulsive hole-jellium interaction, $\delta W_{hd}^{(2D)}$, which is numerically equal to the jellium-jellium term $\delta W_{dd}^{(2D)}$, and the attractive electron-hole term, $\delta W_{he}^{(2D)}$, which is numerically equal to $\delta W_{ed}^{(2D)}$. Thus, the Coulomb energy per minority hole will be exactly of the form given by Eq (21). At the low temperatures at which both the electron and the hole gases are degenerate, the shift of the hole chemical potential will be

$$\delta\mu_h \;=\; \frac{\partial}{\partial N_h}(N_h \delta W_{C,h}^{(2D)}) \;=\; -1.451\;\kappa\; n^{1/3} \;, \tag{25}$$

and the total narrowing of the band-gap will be given by

$$\delta E_G \;=\; \delta\mu \;+\; \delta\mu_h \;=\; -3.386\;\kappa\; n^{1/3} \;. \tag{26}$$

At higher temperatures in non-degenerate situations, instead,

$$\delta E_G \;=\; \delta E_C^{(2D)} \;+\; \delta E_{C,h}^{(2D)} \;=\; -2.902\;\kappa\; n^{1/3} \;. \tag{27}$$

If we now allow for the thermal agitation of the pointlike electrons, once again we must consider how much of

the total energy $\delta W_C^{(2D)}$ is converted into kinetic energy. In this case, however, a simple use of the virial theorem is less than obvious. Indeed, quantitatively more precise statements can be made only by analyzing the algorithm employed to assign the charge of the pointlike superparticles to the mesh nodes, and on the way the forces acting on the particles are interpolated from the mesh to the position of the superparticles, and by the way the Coulomb interactions are handled in practice. Typical complications we have in mind are the role played by the electron-phonon interaction and by the short-range electron-electron and electron-impurity interactions. The resulting picture is so intricate that a validation of the algorithm is best left to computer experiments. Numerical results will be presented in the next section.

One final comment concerns the role played by electron-phonon collisions. As the Coulomb energy, Eq. (5), is converted into kinetic energy, Eq. (9), the electron phonon-interaction will attempt to damp this additional kinetic energy. Instead of being accelerated by the potential fluctuations associated to the Coulomb interactions, electrons will scatter and lose energy (in inelastic electron-phonon collisions) and be redirected in randomized directions. Thus, we expect that as long as the potential fluctuation responsible for the Coulomb heating are of sufficiently low amplitude and at a sufficiently low frequency, the Coulomb kinetic energy will not reach the value expected from Eq. (9). This is a restatement of the well-known collisional damping of plasma oscillations. Indeed, a fraction $n'/(3n) \approx 0.018\ r_s^{2/3}$ of the electronic degrees of freedom of the $3n$ degrees of freedom of the electron gas are associated with plasma collective modes[20], which are known to be damped whenever their frequency $\omega_P = [e^2 n/(m_P \epsilon_s^\infty)]^{1/2}$ is lower than the momentum relaxation time, τ_m. (In the expression for the plasma frequency we have used an effective mass m_P which, in an isotropic and nonparabolic band structure, will be the conductivity effective mass averaged over the entire electron distribution). Heuristically, we may describe this effect by damping the kinetic energy (9) as follows:

$$\delta K_C^{(ep)} \approx 1.451 \frac{e^2 n^{1/3}}{4\pi\epsilon_s^\infty} \frac{1}{[1+1/(\omega_P \tau_m)^2]^{1/2}} . \qquad (28)$$

III. 2D MC/POISSON SIMULATION OF THE HOMOGENEOUS ELECTRON GAS

In order to establish how well 2D semiclassical simulations can reproduce the expected behavior described above, in this section we present results from semiclassical 2D Monte Carlo/Poisson simulations of a homogeneous electron gas. We show that the model-system behaves in a sufficiently accurate way, being able to reproduce the expected change in kinetic energy and band-gap narrowing effects, and showing the correct dielectric response. In addition, the potential exhibits the plasma oscillations of approximately the correct amplitude at the expected frequency.

We have simulated a homogeneous gas of electrons in Si at the dopant and electron concentrations of 10^{19}, 3×10^{19}, 6×10^{19}, and 10^{20} cm^{-3} using the self-consistent Monte Carlo/Poisson algorithm described elsewhere[8]. We have employed a square mesh of uniform spacing $\Delta x = \Delta y = n^{-1/3}$ and a statistical weight $s = n^{1/3}$, consistently with the discussion of Sec. II C. At the lowest density (and perhaps also at $n = 3\times 10^{19}$ cm^{-3}), the validity of the Debye-Hückel theory in the nondegenerate limit would suggest instead $\Delta x = \Delta y \sim 2\beta_s^{-1}$, and $s \approx (1/2)\beta_s$. However, at these low densities Coulomb effects are small anyway and we shall accept the slight underestimation of δK_C implied by the former choice. At higher densities, as we have noted elsewhere[8], there are competing requirements for the selection of the mesh spacing: First, we must approximate correctly δK_C, as just stated. Secondly, we must damp plasma oscillations at short wavelengths, by requiring $\Delta x \sim q_{LD}^{-1}$, where q_{LD} is the Landau-damping parameter defined by $E_0(q_{LD} + k_F) = \hbar\omega_P + E_F$. Finally, we must avoid double-counting the mutual Coulomb interactions at short-range (treated as MC scattering processes) and at long range (handled by the self-consistency between the MC scheme and the Poisson solver), by now requiring $\Delta x \sim \beta_s^{-1}$. Thus, we have reached the conclusion that a good compromise among these requirements is obtained when setting $\Delta x \approx 2\beta_s^{-1}$ in the nondegenerate situations, and $\Delta x \approx q_{LD}^{-1}$ in degenerate cases[8]. It is intriguing that for Si in the range $n > 3\times 10^{19}$ cm^{-3}, the three characteristic lengths associated with the requirements above, respectively r_0, q_{LD}^{-1}, and $2\beta_s^{-1}$, are numerically very similar: For example, at 300 K, for $n = 10^{20}$ cm^{-3}, $q_{LD}^{-1} \approx$ 1.27 nm, $2\beta_s^{-1} \approx 1.32$ nm, while $r_0 \approx 1.33$ nm, so that, for all practical purposes, for electrons in Si and at the densities where Coulomb effects are important, the 'compromise' actually amounts to a satisfactory simultaneous fullfillment of the three requirements. Finally, the use of $\Delta x = n^{-1/3}$ instead of any of the other length-scales is dictated by the attempt to approximate a spherical 3D geometry with a square 2D mesh. 'Small' factors of '2' and 'π', definitely important when cubed, cannot be resolved heuristically. As we said above, the results of computer simulations are our only guide at this level of precision.

We must comment about the model chosen to treat the short-range interactions among electrons and between electrons and ionized impurities. We have already remarked[8] that Coulomb interactions at distances shorter than the mesh spacing are suppressed by the numerical coupling between the MC algorithm and the solution of Poisson equation on the mesh. Therefore, the short-range electron-electron and electron-impurity collisions have to be treated as additional scattering processes in the MC model. In the context of a homogeneous electron gas at steady-state the short-range electron-electron scattering is not expected to affect the results: Energy and momentum are simply re-distributed among the superparticles. The effect of this scattering process should manifest itself as a reduced relaxation time towards equilibrium. Therefore, the choice of the particular electron-electron scattering model, here taken to be given by antisymmetrized

Coulomb scattering screened by the dynamic dielectric function[5], is not critical, since the system is already at equilibrium . On the contrary, the choice of the electron-impurity scattering model is more troublesome. Note that we are forced to change abruptly the physical picture, reverting now to pointlike donors embedded in a plane-wave representation of the electrons. In this model, a statically-screened Coulomb potential (Yukawa-like) is usually employed to compute the electron-impurity scattering rate in the Born approximation. This is commonly known as the Brooks-Herring (BH) model[21], which leads to a matrix element

$$M^{(BH)}(\mathbf{k}',\mathbf{k}) \;=\; \frac{e^2}{\epsilon_s^\infty}\,\frac{1}{q^2+\beta_s^2}\;, \qquad (29)$$

with $q = |\mathbf{k} - \mathbf{k}'|$, (possibly replacing β_s with a momentum-dependent expression $\beta_s(q) = \beta_s G(q)$ given by Eq. (46) below, as in Ref. 5) for an electron transition from an intial state at wave vector $\mathbf{k}$ to a final wave vector $\mathbf{k}'$. For a dispersion described by ellipsoidal (with masses m_t and m_l along the principal axes of the ellipsoids) and nonparabolic (with nonparabolicity parameter α) equi-energy surfaces, the scattering rate for an electron of wave vector $\mathbf{k}$ is:

$$\frac{1}{\tau^{(BH)}(\mathbf{k})} \;=\; \frac{N_D\gamma(\mathbf{k})e^4}{2^{3/2}\pi\epsilon_s^{\infty 2}m_d^{1/2}E_\beta^2}\,\frac{1\;-\;2\;\alpha\;E_0(\mathbf{k})}{1\;+4\gamma(\mathbf{k})/E_\beta}\;, \qquad (30)$$

where $\gamma(\mathbf{k}) = (\hbar^2/2)(k_t^2/m_t + k_l^2/m_l)$ is the parabolic dispersion, k_t and k_l being the components of the k-vector along the transverse and longitudinal axes of the ellipsoid, respectively, $m_d = (m_l m_t^2)^{1/3}$ is the density-of-states effective mass, $E_\beta = \hbar^2\beta_s^2/(2m_d)$, and N_D is the concentration of ionized impurities. In the nondegenerate limit this is indeed the appropriate model. But at higher densities, the proper interaction potential to consider is the bare (*i.e.*, unscreened) Coulomb potential truncated at the average separation r_0 (or $n^{-1/3}$ for square meshes). This is the well-known Conwell-Weisskopf (CW) model[22] whose matrix element will be[23]:

$$M^{(CW)}(\mathbf{k}',\mathbf{k}) \;=\; \frac{e^2}{\epsilon_s^\infty}\,\frac{1-\cos(r_0 q)}{q^2}\;, \qquad (31)$$

which, for ellipsoidal, nonparabolic bands, results in the scattering rate

$$\frac{1}{\tau^{(CW)}(\mathbf{k})} \;=\; \frac{3m_d e^4}{(2\pi\epsilon_s^\infty)^2\hbar^3}\,\mathcal{I}(2kr_0)\;, \qquad (32)$$

having recalled that $N_D r_0^3 = 3/(4\pi)$. The function $\mathcal{I}(x)$ is defined by:

$$\mathcal{I}(x) \;=\; \frac{1}{x}\int_{xy}^{x} dt\,\frac{(1\;-\;\cos t)^3}{t^3}\;, \qquad (33)$$

where in the lower integration limit $y = (E/E_r)/[1 + (E/E_r)^2]^{1/2}$ with $E_r = e^2/(8\pi\epsilon_s^\infty r_0)$. This results from truncating the impact parameter to a maximum value r_0. A much smaller momentum relaxation time is expected from the BH model. This will enter Eq. (28) and reduce the kinetic energy. Therefore, in order to avoid this unphysical damping (caused by incorrectly double-counting the electron-donor interaction), the BH model should be used only in nondegenerate situations, while the CW model, while far from prefect, is preferable at high donor densities.

We have considered regions covered by a 100×100-mesh, which resulted in the simulation of transport of about 10,000 ‘superparticles’. In order to ensure an accurate integration of the equations of motion during the free flights between collisions and an accurate resolution of the plasma oscillations , we have employed very small time steps (10^{-17} s) for the Runge-Kutta integration in free-flight, for the interval between successive solutions of the Poisson equations, and for the time interval between successive checks for scattering.[8]

In order to assign boundary conditions for the numerical Poisson solver and to mimic Ohmic contacts, our computer program requires the definition of a ‘contact’ specified by the value of the electrostatic potential, ϕ_c, and the carrier distribution in momentum space. Conventional equilibrium boundary conditions are chosen. Thus, the contacts will absorb – by removing them – particles ‘hitting’ them, while whenever the requirement of charge neutrality in the mesh-elements next to the contacts asks for the addition of negative charge, the contacts will inject superparticles with energy distributed according to an unperturbed Fermi-Dirac distribution at the lattice temperature and with Fermi energy μ_0 fixed by the condition Eq. (2) with $\mu = \mu_0 + e\phi_c$. A similar distribution is assumed at the beginning of the simulation. It follows that these boundary and initial conditions effectively amount to an initial condition for the total energy of the system, now described by a grand canonical distribution: If we set the zero of total energy at the chemical potential of the particle reservoir (*i.e.*, $\mu = 0$), the kinetic energy of the superparticles will be simply K_F, their potential energy set by the bottom of the conduction band, $E_{CB}^{(0)} = -\mu_0$, and so the total energy per electron will be $-\mu_0 + K_F$. In absence of electron-phonon scattering, the total energy is conserved. Thus, as the electrons acquire the Coulomb kinetic energy δK_C, the potential energy must change by an opposite amount. It is convenient to ignore, for the time being, the renormalization of the electron dispersion caused by the Coulomb interactions, which is related to the fluctuations of the electrostatic potential discussed below in Sec. III B. So, obviously, in this approximation, also made by Mahan[10], the bottom of the conduction band must shift down by the amount $\delta E_{CB} = -\delta K_C$. (In the simulations discussed here the contact was defined as a very short line covering only 2 mesh elements at the center of one side of the square semiconductor region.) Indeed, in all cases we found that the *time-averaged* total energy was ‘almost’ conserved, phonon-losses causing the main deviation from ideality, so that, if renormalization effects were really negligible, $\delta E_{CB} \sim -\delta K_C$. The drop $\delta W_C = -\delta K_C$ in total energy per particle, expected from the preceeding discussions, would require $\delta E_{CB} = -2\;\delta K_C$, but cannot be seen by virtue of the initial conditions we have enforced. Differently stated, instead of placing particles at random positions, we should have ‘closed’ the system (described,

in this case, not by a grand canonical but by a canonical distribution) and should have chosen the lowest energy configuration corresponding to superparticles distributed at the sites of the regular square lattice discussed above. The extra potential energy assigned by our initial conditions and by our choice of an open system has to remain in the system, if energy is to be conserved.

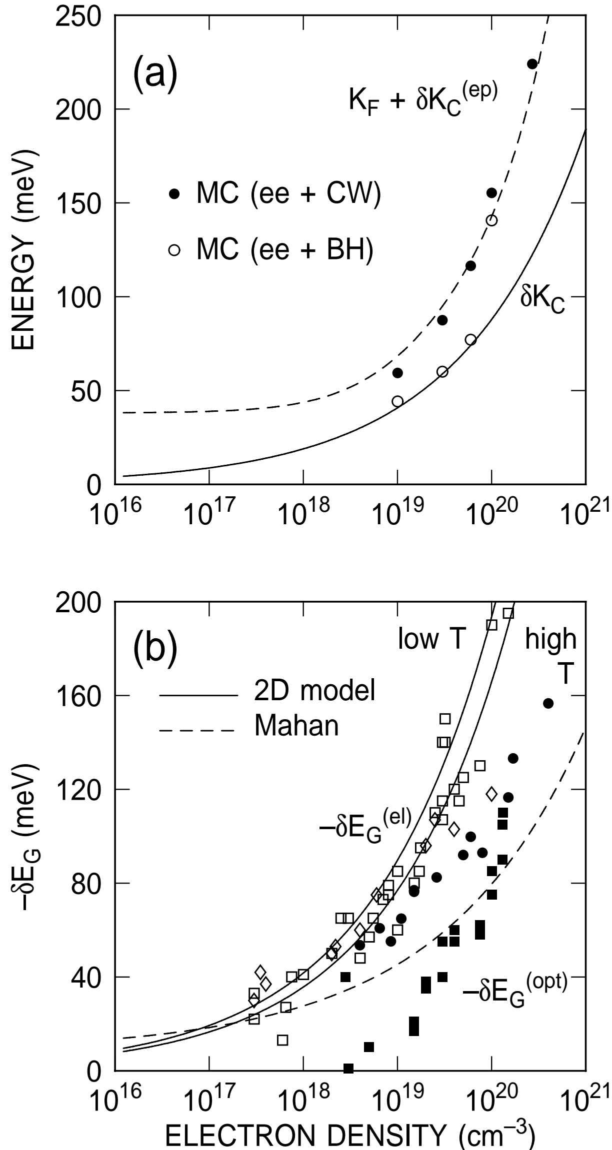


FIG. 1. **(a)**: Kinetic-energy shift δK_C calculated from Eq. (9) (solid line), total kinetic energy $K_F + \delta K_C$ calculated from Eqns. (9) and (28) (dashed line), and total kinetic energy obtained from simulations including the short-range electron-impurity scattering with the Conwell-Weisskopf (CW, dots) or Brooks-Herring (BH, circles) models. The result at the highest electron density has been extracted from the simulation of the source and drain regions of the devices described in Sec. IV. **(b)**: Negative of the bandgap narrowing expected from Eqns. (26) (solid line labeled 'low-T') and (27) (solid line labeled 'high-T'), from Mahan's theory (Ref. 10, dashed line), and experimental data for the 'apparent electrical gap' (open symbols) and optical gap (solid symbols) from Refs. 24–26. Note that the agreement between experimental electrical data and our model is actually an accidental result, as explained in the text.

A. Kinetic energy and band-gap narrowing

Figure 1 shows the results of our simulations. In the top panel, (a), the solid line indicates the density dependence of the additional kinetic energy δK_C due to Coulomb effects evaluated from Eq. (9), the dashed line the total kinetic energy, $K_F + \delta K_C^{(ep)}$, with $\delta K_C^{(ep)}$ now corrected for collisional damping, Eq. (28) with $\tau_m \approx 10^{-14}$ s. The dots are the numerical results obtained by accounting for short-range electron-impurity scattering in the Conwell-Weisskopf approximation, the open circles using the Brooks-Herring model. As expected, the BH model behaves as an additional damping term, although the effect is not very large.

The solid line in Fig. 1(b) shows the negative of the band-gap narrowing δE_G, evaluated from Eqns. (26) and (27). It is interesting to compare this result with experimental data relative to the band-gap narrowing effect. We show with open symbols data relative to the 'apparent electrical gap', with solid symbols data relative to what is known as the 'optical gap'. The data have been taken from the collections of Ref. 24 (open diamond and optical data), Ref. 25 (open squares) and Ref. 26 (circles). The theoretical result of Mahan[10] is also indicated by the dashed line. Note that this is a typical example of a better theory (Mahan's) yielding worse agreement with data than a worse theory (ours). The agreement between the experimental data for the apparent electrical gap and our value from Eqns. (26) and (27) is indeed the result of shear luck: The difference between Eq. (24) and Eq. (12) is indeed crucial and is simply the result of having switched the roles of point-like and jellium-like charge between the electrons and the donors. In addition, quantum corrections are significant in the scale of Fig. 1. From a shamelessly practical perspective, though, we seem to be able to account in a quantitatively correct way for the all-important bandgap-narrowing effect.

B. Potential fluctuations and dielectric response

So far we have presented only global energetic considerations giving information on temporal, spatial, and ensemble averaged energy of the system, as in Eqns. (5), (8), and (9). However, as stated above, degeneracy and collisions induce a thermal agitation of the electrons in the gas, which, in turn, causes the potential to fluctuate. These dynamic and spatially-varying phenomena are also correlated to the momentum distribution of the electrons.

Fluctuations of the electron density and the self-consistently associated fluctuations of the potential (which are induced by and also induce the density-fuctuations) are eigen-modes of the system. Their temporal behavior is derived from their dispersion ω_q which is set by the vanishing of the (macroscopic) electric displacement field, and so of the dielectric function (*i.e.*, $\epsilon_s(q, \omega_q) = 0$). In the long-wavelength limit, the density and potential fluctuations oscillate at the plasma frequency, $\omega_P = [e^2 n/(m_P \epsilon_s^{\infty})]^{1/2}$. As far as their spatial dependence is concerned, Appendix A presents

the canonical decomposition of the system into these collective degrees of freedom (plasmons) of various wavelengths, also leading to their quantization. From these results, as well as from additional semiclassical arguments, we can derive qualitative estimates for the expected amplitude of these fluctuations. From the canonical quantization of Appendix A, the root-mean-square (rms) amplitude $\mid \phi_{\mathbf{q}} \mid$ of the field associated to plasmons of mode $\mathbf{q}$ is:

$$\mid \phi_{\mathbf{q}} \mid = \frac{\mid a_q \mid}{q} = \left(\frac{\hbar\omega_P}{2\epsilon_s^\infty}\right)^{1/2} \frac{1}{q} \sqrt{1+2n_{\mathbf{q}}} \; . \qquad (34)$$

Accounting for all modes up to the cut-off wave vector q_c, separating long-wavelength collective modes from short-wavelength single-particle excitations[20], we get for the total amplitude:

$$< \phi >_{QM} = \left\{ \int_{q<q_c} \frac{d\mathbf{q}}{(2\pi)^3} \mid \phi_{\mathbf{q}} \mid^2 \right\}^{1/2} = \left(\frac{\hbar\omega_P q_c}{4\epsilon_s^\infty \pi^2}\right)^{1/2} \sqrt{1+2n_P} \; , \qquad (35)$$

having assumed dispersionless plasmons (*i.e.*, $\omega_{\mathbf{q}} = \omega_P$).

Semiclassically, as in the quantum case, the potential has the form

$$\phi_{\mathbf{q}}(\mathbf{r},t) \approx \frac{a}{q} \sin(\mathbf{q}\cdot\mathbf{r} - \omega_P t) \; . \qquad (36)$$

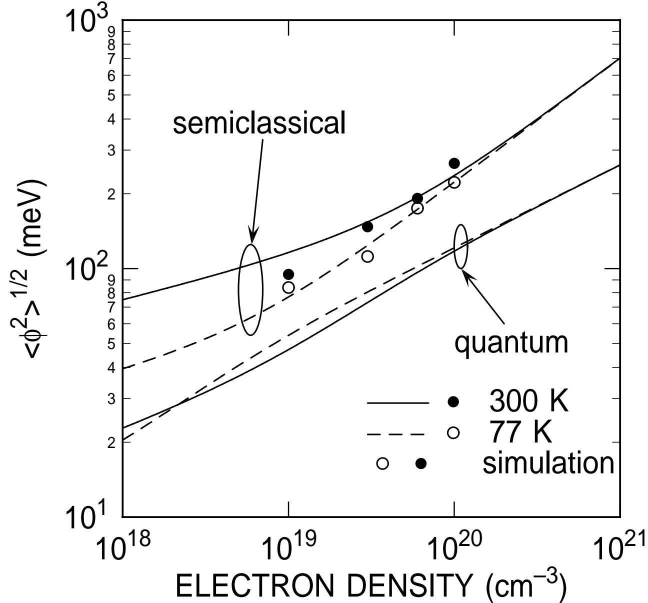
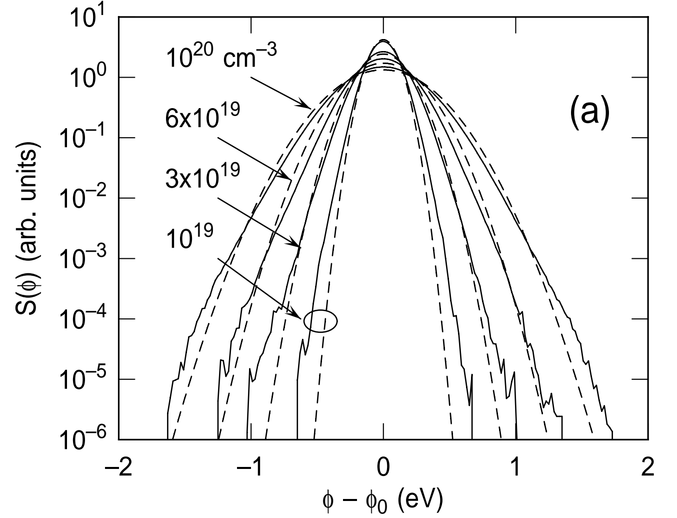


FIG. 2. Root-mean-square amplitude of the (plasma) fluctuations of the potential as a function of the density of a homogeneous electron gas calculated theoretically quantum-mechanically from Eq. (35) or semiclassically from Eq. (41), at 300 K (solid lines) and 77 K (dashed lines). The symbols are the results of self-consistent two-dimensional semiclassical Monte Carlo/Poisson simulations using mesh spacing, number of superparticles, and time steps optimized to yield the 'correct' results.



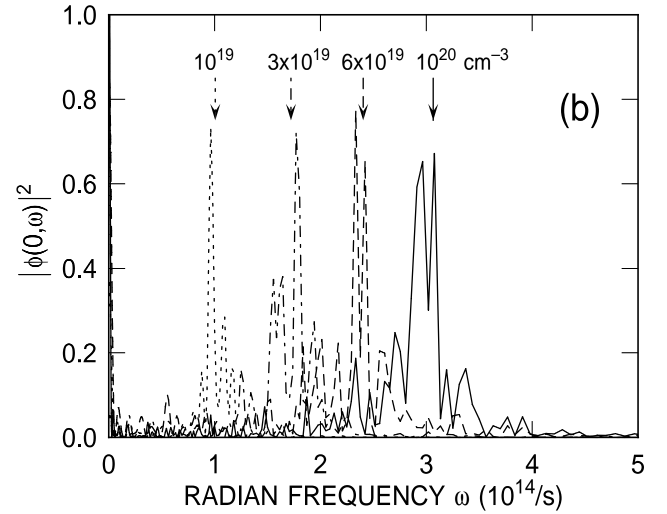


FIG. 3. **(a)**: Probability distribution for the potential due to the space-dependent fluctuations caused by the Coulomb interactions at 300 K. The solid lines are obtained from the two-dimensional self-consistent MC/Poisson simulations, while the dashed line is the empirical Gaussian fit from Eq. (42). **(b)**: Power spectrum $|\phi(\mathbf{q}=0,\omega)|^2$ of the time-dependent fluctuations of the space-averaged potential, showing that the spatial average of the potential oscillates at the plasma frequency. The arrows indicated the expected positions of the peaks. A small red-shift is caused by non-parabolicity of the band-structure, resulting in a heavier conductivity mass.

The time-averaged electrostatic (self)energy W_q over a volume Ω associated with mode q will be:

$$< W_q > = \frac{1}{2}\epsilon_s^\infty \left\langle \int_\Omega d\mathbf{r} \; a^2 \; \sin^2(\mathbf{q}\cdot\mathbf{r} - \omega_P t) \right\rangle = \frac{1}{4}\Omega \; \epsilon_s^\infty \; a^2 \; . \qquad (37)$$

As in Eq. (35), accounting for all modes up the Debye wave vector β_s, which is the semiclassical cut-off for Landau damping, the total energy of all modes will be:

$$W = \int_{q<\beta_s} \frac{d\mathbf{q}}{(2\pi)^3} < W_q > = \frac{\epsilon_s^\infty a^2 \Omega}{8\pi^2} \frac{\beta_s^3}{3} \; . \qquad (38)$$

For harmonic oscillators, the time-averaged kinetic energy equals the time-averaged electrostatic energy. Thus. Eq. (38) must be equal to the kinetic energy δK_C (per particle) given by Eq. (9), multiplied by the total number of particles in a volume Ω:

$$W \ = \ n \ \Omega \ \delta K_C \ . \tag{39}$$

This implies

$$a^2 \ = \ 1.451 \ \frac{6\pi e^2 n^{4/3}}{\epsilon_s^{\infty 2} \beta_s^3} \ , \tag{40}$$

so that the total amplitude of the potential fluctuations will be

$$\begin{aligned} < \phi >_{CL} &= \left\{ \int_{q<\beta_s} \frac{d\mathbf{q}}{(2\pi)^3} \mid \phi_{\mathbf{q}} \mid^2 \right\}^{1/2} \\ &= \left(1.451 \times \frac{3}{\pi} \right)^{1/2} \ \frac{e \ n^{2/3}}{\epsilon_s^{\infty} \beta_s} \ . \end{aligned} \tag{41}$$

The lines in Fig. 2 shows the quantum-mechanical, Eq. (35), and semiclassical, Eq. (41), rms amplitudes of the potential fluctuations (or of the 'plasmon field') at 300 K and 77 K. We have chosen $q_c \approx \beta_s$ also for the evaluation of the quantum fluctuations. The quantum mechanical fluctuations are weaker since, roughly speaking, the electron wavefunctions spread over some distance, unlike the point-like semiclassical picture, thus reducing the spatial variation of the self-consistent field.

In order to compare these expectations with the results of the simulations, we first show, in Fig. 3(a), the distribution of the potential fluctuations, that is, the probability $S(\phi)$ of finding the potential an amount $\delta\phi = \phi - \phi_0$ away from the global spatial and temporal average, ϕ_0, at any given time and location. Empirically, we find that this fluctuation-distribution S can be approximated by the Gaussian form:

$$S(\phi) \ = \ \frac{1}{2\pi^{1/2}\eta} \ \exp\left\{ - \left[\frac{\phi - \phi_0}{\eta} \right]^2 \right\} \ . \tag{42}$$

The magnitude of the variance, $\eta/\sqrt{2}$, of the distribution $S(\phi)$ (*i.e.*, the rms amplitude of the potential fluctuations) compares favorably with the semiclassical theoretical expectations, as shown by the symbols in Fig. 2 (See note 27).

Figure 3(b) shows that the spatial average (that is, the long-wavelength components) of the potential oscillates in time at the plasma frequency, moving up and down as the total charge is moved to the simulated region from the contact and vice-versa, charge neutrality being maintained only in a time-averaged way. Note that the 'plasma peaks' are broadened by collisions (with phonons and impurities), the broadening increasing with density, since the momentum relaxation time decreases with increasing electron (*i.e.*, plasma) energy. 'Side bands' at frequencies lower than the bulk plasma frequency are possibly due to the finite size of our simulation, since the frequency of plasmons decreases at the boundaries (Si-vacuum) of the simulated region.

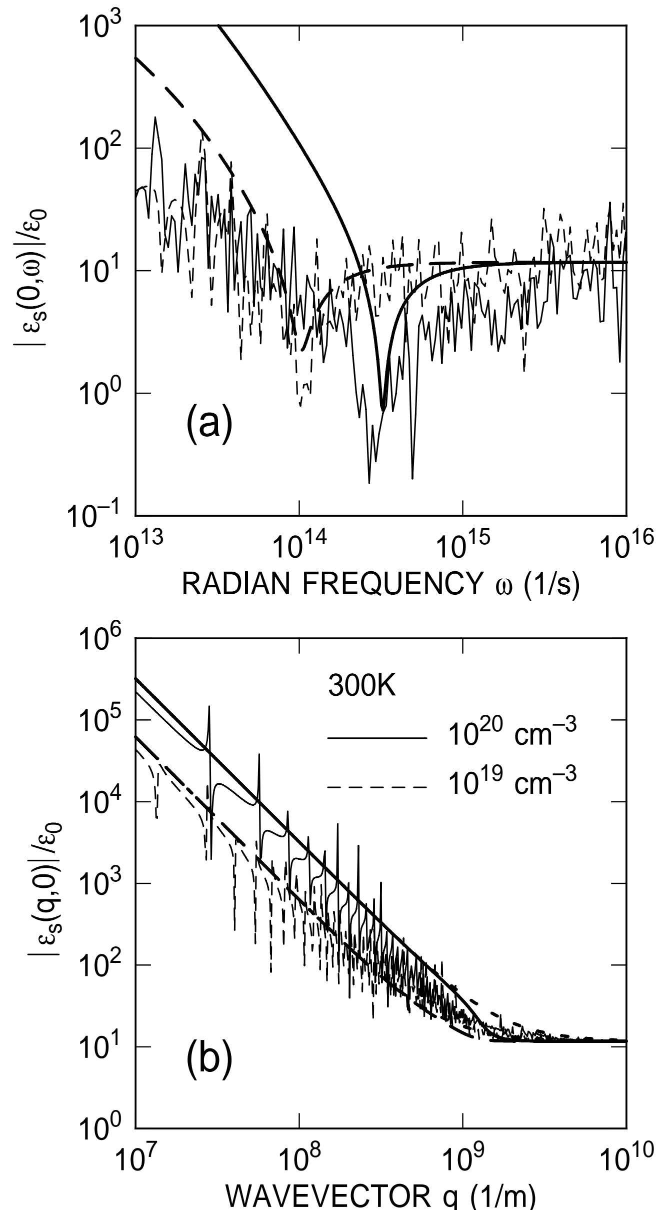


FIG. 4. **(a)**: Frequency dependence of the long-wavelength dielectric function obtained from the simulation (thin, noisy lines) compared wiith the dielectric function of a damped plasma, Eq. (44), (thick, smooth lines) at the two indicated densities. **(b)**: Wavelength dependence of the static dielectric functions (thin lines) compared with the Thomas-Fermi expression, Eq. (45) (thick solid lines) or with the more common approximation of setting $G(q) = 1$ in Eq. (45) (dashed thick lines). Aliasing effects, visible as resonances, are caused by the finite size of the simulated region and of the mesh elements.

The relation between the potential and density fluctuations permits an estimation of the (longitudinal) dielectric function, $\epsilon(\mathbf{q}, \omega)$, via the dielectric susceptibility

$$\chi(\mathbf{q}, \omega) \ = - \ \frac{e \ \delta n(\mathbf{q}, \omega)}{q^2 \ \delta\phi(\mathbf{q}, \omega)} \ , \tag{43}$$

as it follows from the fact that the charge fluctuations are exclusively due to a *polarization* charge. This equation is valid at q small enough for mesh-size effects to be negliglible ($q < 2\pi/\Delta x$ where Δx is the size of a mesh

element) while also large enough to avoid artifacts due to the finite size of the simulated region ($q > 2\pi/(N_{el}^{1/2}\Delta x)$, where N_{el} is the total number of mesh-elements employed, $\approx 10^4$ here). In Fig. 4(a) we show the frequency depencence of the long-wavelength ($q = 2\pi/(N_{el}^{1/2}\Delta x)$) limit of the magnitude of the dielectric function, labeled $\epsilon_s(0,\omega)$ for convienence, compared with the dielectric response of a classical electron gas damped by a 'viscous' term $1/\tau_m$:

$$\epsilon_P(0,\omega) \ = \ \epsilon_s^\infty \left(1 \ - \ \frac{\omega_P^2}{\omega^2 + i\omega/\tau_m} \right) \ . \qquad (44)$$

The 'plasma zero' (or the 'plasmon pole' of $\epsilon_s(\mathbf{q},\omega)^{-1}$) is clearly reproduced by the simulation. Indeed, since as seen in Fig. 3(b), most of the oscillations occur at the plasma frequency, the simulations yield more reliable (that is, with better stochastic noise) results at the plasma frequency. As we move away from ω_P, our sampling ability decreases, as reflected in the noise of the data. Note that in the limit $\omega \to 0$ the results of the simulation do not diverge, simply because we have evaluated ϵ at a small but nonzero q. In the bottom panel of the same figure we compare the low-frequency limit $\epsilon_s(q,0)$ with the conventional self-consistent-field result

$$\epsilon_{SCF}(q,0) \ = \ \epsilon_s^\infty \left(1 \ + \ \frac{\beta_s^2 \ G(q)}{q^2} \right) \ , \qquad (45)$$

where $G(q)$ is the function:

$$G(q) \ = \ \frac{1}{\sqrt{\pi}\xi F_{-1/2}(\eta_F)} \int_{-\infty}^{\infty} dx \ln \left| \frac{x+\xi}{x-\xi} \right| \times \frac{x}{1+\exp(x^2-\eta_F)} \ , \qquad (46)$$

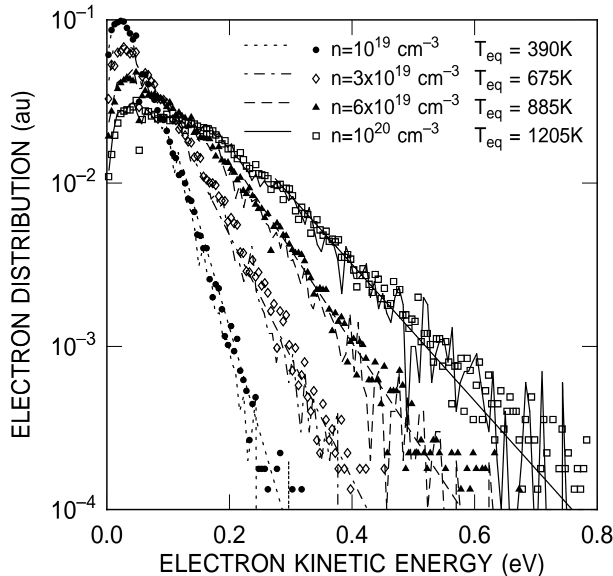


FIG. 5. Electron energy distributions obtained from the simulation of 10,000 electrons in homogeneous Si regions at the donor concentrations indicated. The lines refer to distributions obtained ignoring the short-range electron-electron interactions, the symbols those obtained by accounting for this additional process. The straight lines are associated with exponentials at the indicated 'equivalent' temperatures.

where $\xi = (\beta\hbar^2 q^2)/(8m_d)$, $\eta_F = \mu_0/(k_B T)$, and F_ν is the Fermi integral of order ν. Note that the simulations are meaningful only for q smaller than the size of the mesh element $\Delta x \sim \beta_s^{-1}$. At even smaller distances short-range collisions – treated with the theoretical screening length Eq. (45) – take over. A small amount of underscreening is observed, consistently with the larger thermal velocities of the electrons, caused by Coulomb effects.

C. Kinetic energy distributions and statistical mechanics

So far we have established that Coulomb interactions yield an average kinetic energy per particle larger than its 'thermal' value, Eq. (1). This extra kinetic energy, δK_C, can be estimated using the semiclassical modified Hartree model, Eq. (9), since exchange and correlation effects can be approximately ignored. In addition, δK_C can be well approximated by 2D MC/Poisson self consistent simulations, when a proper choice of mesh size and number of particles is made.

It is clear that this 'extra' kinetic energy must be reflected in a modification of the kinetic-energy distribution of the electrons, as also implied by the presence of the potential fluctuations $S(\phi)$ resulting from the plasma oscillations. Figure 5 shows indeed that the distribution in kinetic energy of the electrons are consistent with larger average kinetic energy, $K_F + \delta K_C$, as shown in the top panel of Fig. 1. This extra kinetic energy appears as a stronger high-energy tail of exponential form (that is, 'thermalized' by the strong interparticle energy-exchange caused by the Coulomb interactions), but at a temperature higher than the lattice temperature, quantitatively consistent with the magnitude of δK_C.

These results beg the question: Are these high-energy tails, with the obvious appearance of being at odds with thermal equilibrium at the lattice temperature, compatible with basic expectations from statistical mechanics and thermodynamics? Should not the high-energy tails of the equilibrium electron energy distributions always exhibit a Boltzmann-like behavior $\sim \exp[-K/(k_B T)]$ at large electron kinetic energy K?

We should first remark that an 'equilibrium' distribution in kinetic energy should not be expected for several reasons: First, as stated by Landau and Lifshitz[28], for a classical system of total energy $W(p,q) = K(p) + U(q)$ (where p and q are the generalized canonical momenta and coordinates, K and U the kinetic and potential energy, respectvely) the (Gibbs) canonical probability distribution dw_{qp} in phase space, $dw_{qp} \propto dpdq \exp[-W(p,q)/(k_B T)]$ yields the probability-distribution of the momenta (by integrating over the coordinates), $dw_p \propto dp \exp[-K(p)/(k_B T)]$, only in the case of classical particles, not for Fermions obeying Fermi-Dirac statistics. In our simulations degeneracy enters in several ways: Contacts inject electrons with a Fermi-Dirac distribution, and the same statistics is used to accept or reject collision processes according to the occupation of the final state.[30] Secondly, for Fermions (non spin-polarized) at equilibrium at temperature T,

Landau[29] has shown that the probability-distribution of the single-particle momentum p, $n(p)$, obeys the expected Fermi-Dirac law

$$n(p) \ = \ \frac{1}{1+\exp\{[\ w(p)-\mu\ ]/(k_BT)\}} \ , \qquad (47)$$

provided the energy $w(p)$ is interpreted as the energy of a 'quasi-particle'. Thus, $w(p)$ is the amount by which the total energy of the system changes when adding an excitation of momentum p. For non-interacting electrons one obviously has $w(p) = E_0(p/\hbar)$, where $E_0(k)$ is the unperturbed dispersion in the conduction band. For interacting particles $w(p)$ is formally given by the solution of the Dyson equation $w(p) = E_0(p)+\Sigma(p, w(p)/\hbar)$, $\Sigma(p, \omega)$ being the self-energy resulting from all interactions. Only after having renormalized the dispersion of the particles one can recover the 'free' dynamics and the equilibrium properties following from Eq. (47). A different behavior should be expected for the undressed particles. For example, notice that even at zero temperature the function $n(p)$ does not have to be of the form $n(p) = \theta(p - p_F)$ where p_F is the momentum at the Fermi surface, valid only for noninteracting Fermions. This counterintuitive form has been discussed by Pines[31]. Detailed quantum Monte Carlo simulations of the electron gas[19,32,33] at equilibrium at zero temperature have actually allowed the numerical evaluation of the function $n(p)$, typically fitted by power-law tails of the asymptotic form p^{-8} for $p >> p_F$.

We can now see how the 'apparently non-equilibrium' distributions in kinetic energy shown in Fig. 5 emerge from an equilibrium distribution function in *total* energy, using Fig. 6 as a visual reference. Here we follow a heuristic argument similar to the one employed by Kane[34] to calculate another contribution to the band-gap narrowing of heavily-doped semiconductors, namely, the Gaussian tails of the density of states in the gap induced by the random spatial fluctuation of the potential caused by the random positions of the dopants. We should keep in mind that while in this case the potential fluctuations are static, since they are due to the randomly distributed dopants, in our case they are dynamic, since they are due to the (self-consistent) random electronic motion, and only self-consistent schemes, such as the simulations described above, can provide a quantitatively correct picture. Therefore, the the simple picture we shall obtain will only be 'suggestive' of the basic physics.

Let us describe the electron gas by a global equilibrium Fermi-Dirac distribution of the form (47), and consider the time-averaged fluctuating potential described by the distribution $S(\phi)$, centered around its space- and time-averaged value ϕ_0. Then, the distribution in kinetic energy, $n(K)$, can be obtained by convoluting the equilibrium Fermi-Dirac distribution as a function of the *total* energy $w = K + \phi$ with the fluctuating potential:

$$\begin{aligned} n(K) \ &= \ \int_{-\infty}^{\infty} dw \int_{-\infty}^{\infty} d\phi \ f_{FD}(w) \\ &\qquad \times \rho_0(w-\phi)\ S(\phi)\ \delta(w-\phi-K) \\ &= \ \rho_0(K) \int_{-\infty}^{\infty} dw \ f_{FD}(w)\ S(w-K) \ , \end{aligned} \qquad (48)$$

where $\rho_0(K)$ is the density of states (DOS) at kinetic energy K, here taken from empirical pseudopotential calculation for Si, $f_{FD}(w) = \{1+\exp[(w-\mu)]/(k_BT)\}^{-1}$, and the global Fermi potential μ must be determined from the condition

$$n \ = \ \int_0^{\infty} dK \ n(K) \ = \ \int_{-\infty}^{\infty} dw \ f_{FD}(w)\ \rho(w) \ . \qquad (49)$$

The 'effective' (or 'renormalized') density of states $\rho(w)$ at the total energy w is defined as

$$\rho(w) \ = \ \int_{-\infty}^{w} d\phi \ S(\phi)\ \rho_0(w-\phi) \ , \qquad (50)$$

where ρ_0 is the unperturbed DOS. Let's embrace the Gaussian approximation Eq. (42) for $S(\phi)$. It is interesting to see the qualitative features of the effective density of states in the case of six parabolic valleys described by a DOS effective mass m_d:

$$\rho_0(K) \ = \ 6\ \frac{(2m_d)^{3/2}}{2\pi^2\hbar^3}\ K^{1/2} \ . \qquad (51)$$

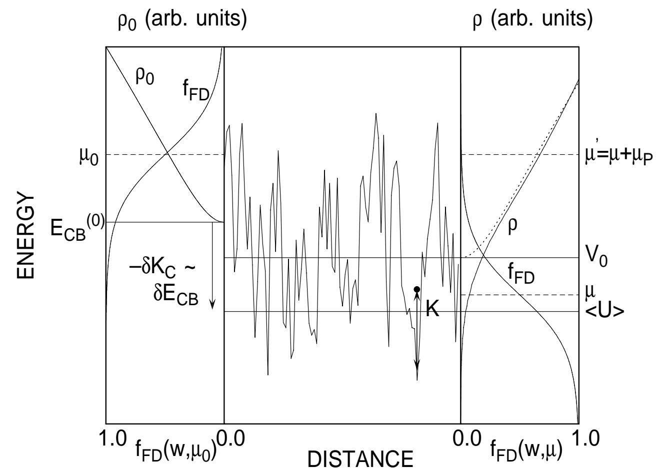


FIG. 6. Schematic diagram illustrating graphically the origin of the Coulomb kinetic energy δK_C and of the associated high-energy tails. If the gas is assumed to reach a global equilibrium population $f_{FD}(w)$ in total energy $w = K + \phi$, shown by the curve at right, the presence of potential fluctuations cause electrons even at moderate total energies to populate regions of high kinetic energy K. The dot represents an average thermal electron of total energy of just about the Fermi level, which will contribute to the large kinetic-energy tails, since it is located at a position where the fluctuating potential energy takes a large negative value. The zero of the ordinate scale taken as the reference energy is the Fermi level of the electron gas. As explained in the text (see Eqns. (59) and (60)) the Fermi energy μ is related to the equilibrium chemical potential of the reservoirs μ_0 via $\mu_0 = \mu + \mu_P$, where μ is the 'single particle' Fermi energy E_F and μ_P originates from the collective plasma oscillations. Also shown are the renormalized DOS, ρ, at right, compared to the unperturbed DOS (solid curve at left, shifted dotted curve at right). Note the shift $\delta E_{CB} = -\delta K_C$, of the bottom of the conduction band, as required by conservation of total energy, and the equality of the chemical potentials of the nointeracting gas in the contact, μ_0, and of the interacting gas in the simulated region, μ'.

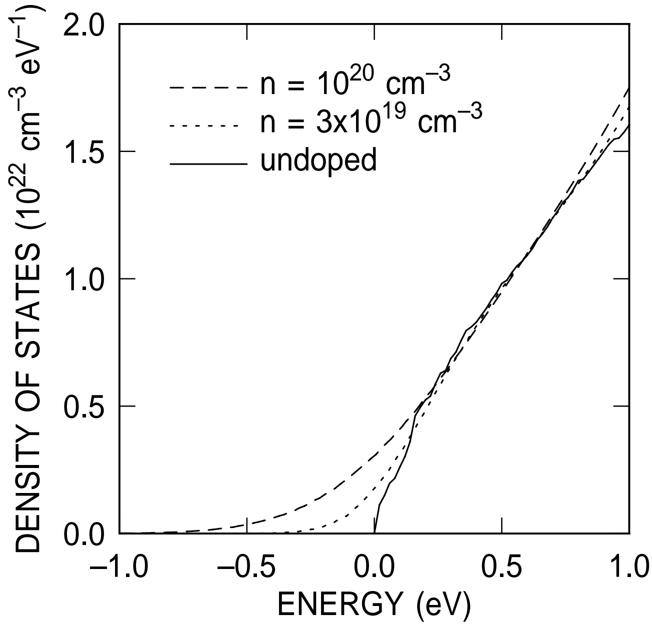


FIG. 7. Effective density of states calculated from Eq. 50 with the potential fluctuations shown in Fig. 3

.

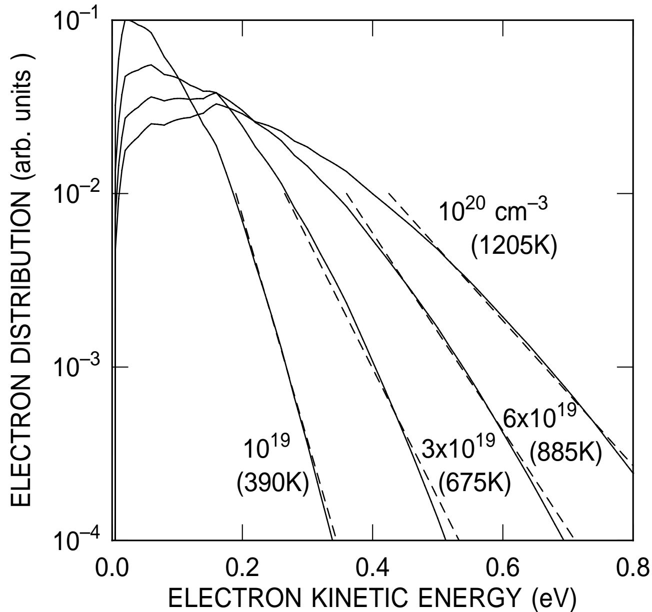


FIG. 8. Kinetic energy distribution calculated from potential fluctutations, Eq. (48). The dashed lines are Boltzmann factors at the same 'Coulomb' temperatures shown in Fig. 5.

In this case, similarly to what found by Kane[34],

$$\rho(w) \;=\; 6\;\frac{(2m_d)^{3/2}}{2\pi^2\hbar^3}\;\frac{1}{\pi^{1/2}\eta}\;\Gamma\left(\frac{3}{2}\right)\;\left(\frac{\eta}{2}\right)^{3/4} \times\;\exp\left(-\frac{w^2}{2\eta^2}\right)\;D_{-3/2}\left(-\frac{2^{1/2}w}{\eta}\right)\;, \qquad (52)$$

where the 'parabolic cylinder function' $D_\nu(z)$ is defined as[35]

$$D_\nu(z) \;=\; \frac{\exp(-z^2/4)}{\Gamma(-\nu)}\;\int_0^\infty\;dx\;e^{-zx-x^2/2}\;x^{-\nu-1}\;, \qquad (53)$$

for $\nu < 0$. At large positive energies ($w >> \eta$; $w > 0$),

$$\rho(w)\;\rightarrow\;\rho_0(w)\;\left(1\;-\;\frac{\eta^2}{16w^2}\right)\;, \qquad (54)$$

that is, the DOS approaches the unperturbed DOS. At large and negative energies ($-w >> \eta$; $w < 0$),

$$\rho(w)\;\rightarrow\;\frac{3m_d^{3/2}}{4\pi^2\hbar^3}\;\eta^{1/2}\;\left(\frac{\eta}{-w}\right)^{3/2}\;e^{-w^2/\eta^2}\;, \qquad (55)$$

which exhibits a Gaussian tail in the gap of the semiconductor. Figure 7 shows the resulting effective density of states $\rho(w)$ obtained using in Eq. (50) the fluctuations shown in Fig. 3 and the unperturbed DOS $\rho_0(w)$ obtained from the pseudopotential band-structure used in our simulations. Fig. 8 shows the kinetic energy distributions similarly obtained from Eq. (48). A comparison with Fig. 5 shows a striking similarity. Particularly noticeable are the high-energy tails described by the same large effective temperature observed in the 'nonthermal' tails obtained from MC simulations. The main difference is that the distributions of Fig. 8 are slightly 'hotter', because of the heuristic nature of our arguments (a dynamic self-consistent convolution should replace the static convolution (48)) and due to the fact that we have ignored electron-phonon collisions. We have already accounted qualitatively for this phonon damping in Eq. (28). Differently stated, the 'exact' Fermi-Dirac global distribution can only be obtained by employing the 'real' total energy $w(p) = E_0(p)+\phi+\Sigma_{sr}(p, w(p)/\hbar)$, where $\Sigma_{sr}(p, \omega)$ is the electron self-energy due to all short-range interactions accounted for by the Monte Carlo algorithm, electron-phonon and short-range Coulomb scattering among electrons and between electrons and ionized donors. For large p this does indeed correspond to quasiparticle excitations, since for $q > \beta_s$ the effective DOS in Fig. 7 can be thought of as extracted from a renormalized dispersion $w(p)$ with a positive effective mass (*i.e.*, $(\partial^2 w(p)/\partial p^2) > 0$). For small p, instead, the dressed excitations are collective modes. Finally, note that the average potential ϕ_0, in the absence of electron-phonon interactions, adjusts itself so that the average shift of the potential energy $< \delta U >$ of the ensemble compensates for the additional kinetic energy acquired by the system, that is

$$< \delta U > = \;\frac{1}{n}\;\int_{-\infty}^{\infty}\;dw\;f_{FD}(w) \times\;\int_{-\infty}^{w}\;d\phi\;\phi\;\rho_0(w-\phi)\;S(\phi)\;=\;-\;\delta K_C\;. \qquad (56)$$

This relation replaces the expression we have employed in Sec. III, $\delta E_{CB} = -\delta K_C$, when ignoring the renormalization of the electron dispersion associated with the fluctuations of the potential. Loosely speaking, one may identify $< U >$ with the renormalized bottom of the conduction band. Since also the minority carriers (holes) will see their potential energy lowered by an amount $< \delta U_h > \approx < \delta U >$, the fluctuations induce a renormalization of the band-gap consistent with the results we have obtained above, Eq. (27) when ignoring renormalization effects. An additional feature, not accounted for

by the simple analysis of Sec. II C, is the presence of quasi-Gaussian band-tailing.

Fig. 8 represents the main result of this section: It shows that global thermodynamical equilibrium at the lattice temperature does indeed produce the high-energy tails observed in the electron distributions in kinetic energy. In a way, the kinetic-energy distributions shown in Fig. 5 are misleading. In a homogeneous electron gas, there are no electrons at the very high *total* energies which one could infer from a cursory look at the figure. The 'hottest' electrons are actually sitting in troughs of the potential fluctuations, at total energies fully consistent with thermal equilibrium. In the inhomogeneous situations of Si devices the situation is different, since the interaction between global-equilibrium electrons in the heavily-doped contacts and the off-equilibrium electrons drifting in the channel gives raise to interesting effects, as discussed below.

Several comments are in order. First, note that the Fermi potential μ entering Eqns. (48) and (49) via $f_{FD}(w)$ is *not* the chemical potential of the gas. Indeed, the chemical potential μ' is the change of the total free energy of the system as we change the number of particles. As we add one particle, we change the total free energy in two ways: We modify the single-particle total kinetic energy, as in a non-interacting system, but we also enhance the fluctuations of the potential. To see how these two contributions arise, let us express the total energy W_{tot} of the system as $W_{tot} = \Omega < w >$, where the time- and ensemble-averaged average energy density $< w >$ is

$$< w > = \int_{-\infty}^{+\infty} dw\ w\ f_{FD}(w) \int_{-\infty}^{w} d\phi\ \rho_0(w-\phi)\ S(\phi)\ . \tag{57}$$

Similarly, the entropy density of the gas, $< \sigma >$, will be:

$$< \sigma > = -\, k_B \int_{-\infty}^{+\infty} dw\ f_{FD}(w)\ \ln[f_{FD}(w)] \times \int_{-\infty}^{w} d\phi\ \rho_0(w-\phi)\ S(\phi)\ . \tag{58}$$

The chemical potential of the gas is, by Eq. (10):

$$\mu' \ = \ \frac{\partial < w >}{\partial n} \ - \ T\ \frac{\partial < \sigma >}{\partial n}\ , \tag{59}$$

at constant T and Ω. Now notice that $< w >$ and $< \sigma >$ depend on the density n via the Fermi potential μ, as in the usual case of non-interacting particles, *and* via the potential fluctuations. Since $[\partial(< w > + T < \sigma >)/\partial\mu]/(\partial n/\partial\mu) \ = \mu$, (as it can be verified numerically or analytically in the nondegenerate, $\mu/(k_BT) << -1$, or in the low-temperature, $1/(k_BT) \to \infty$ limits) we can write μ' as $\mu + \mu_P$ where:

$$\mu_P \ = \ \int_{-\infty}^{\infty} dw\ f_{FD}(w)[\ w \ + \ k_BT \ln f_{FD}(w)\ ] \times \int_{-\infty}^{w} d\phi\ \rho_0(w-\phi)\ \frac{\partial S(\phi)}{\partial n}\ . \tag{60}$$

This term is a reflection of the fact that the energy of the interacting system results not only from the sum of the single-particle energy, but also from the energy of collective excitations. Stated differently, in a non-interacting system at zero temperature the minimum energetic cost of adding a particle is just the single-particle Fermi energy μ. But in a system of interacting particles, adding a particle at the single-particle Fermi energy μ (that is, in the lowest available empty state) will also change the energy of all of the other particles in the system via the long-range forces among particles. Thus, μ_P may be viewed as the chemical potential of the 'collective modes'.

Secondly, in the zero-temperature limit, assuming the Gaussian fluctuations Eq. (42), we obtain

$$n(K) \ = \ \frac{1}{2}\ \rho_0(K)\ \left[\ 1 - \Phi\left(\frac{K-\mu}{2\eta}\right)\ \right]\ , \tag{61}$$

(where $\Phi(x) = 2\pi^{-1/2}\int_0^x e^{-t^2}dt$) which for $K >> \eta$ exhibits a Gaussian high-energy tail:

$$n(K) \ \to \ \pi^{1/2}\ \left(\frac{\eta}{K}\right)\ \rho_0(K)\ e^{-(K/\eta)^2}\ . \tag{62}$$

On the contrary, as anticipated above, particles obeying Boltzmann statistics will not exhibit the same 'hot' tails: Consider Eq. (48), but assume for simplicity the parabolic DOS, Eq. (51), approximate the function S of Fig. 3 with the Gaussian (42), and substitute the Fermi-Dirac distribution f_{FD} with the Boltzmann factor $f_B(w) \propto\ \exp[-w/(k_BT)]$. Then for $K >> \mu$:

$$n(K) \ \to \ \rho_0(K)\ e^{\mu/(k_BT)+\eta^2/(2k_BT)^2}\ e^{-K/(k_BT)}\ , \tag{63}$$

showing that indeed the high-K behavior is still Boltzmann-like at the lattice temperature. This is consistent with Landau and Lifshitz' observation we have mentioned above[28] and with the fact that at low carrier densities, when Boltzmann statistics is a good approximation, no Coulomb heating is observed.

IV. SIMULATION OF SMALL SILICON MOSFETS

A. Qualitative picture

In the previous sections we have analized the behavior of a homogeneous electron gas at high density. It is now easy to understand how the electron gases present in the heavily-doped source (S), drain (D), and gate (G) regions of small Si MOSFETs affect the transport properties of electrons in the channel: As discussed in Sec. III B, the long-range Coulomb interactions in these regions cause the potential to fluctuate. Fluctuations in the source and drain diffusions obviously do not vanish abruptly at the S/channel or D/channel n^+-p junctions. Rather, they penetrate into the lower-density channel. Viewed from the channel, the density and potential fluctuations in the S and D-regions appear as oscillating dipoles. Thus we expect the fluctuations to decay rather slowly (as a power-law) with distance from the S and D

regions. In Fig. 9 we show an extreme situation: We have simulated a high-density Si region ($N_D = n = 10^{20}$ cm^{-3}), as in Sec. III, exemplifying the drain of a MOSFET, and placed it adjacent to a lower-density region ($N_D = n = 10^{17}$ cm^{-3}), representing the channel. The addition of an oxide layer separating the two regions mimics, instead, the effect of potential fluctuations extending from the heavily-doped gate into the channel. The figure shows the rms amplitude of the potential fluctuations along the $n^+ - n$ junction. The presence of an oxide layer forbids the creation of rather large 'dipoles' as electrons in the n^+ region cannot 'spill over' into the n-region. Thicker oxides (10 vs 3 nm) damp the fluctuations more effectively. By removing the insulator, we allow electrons to oscillate across the junction, creating larger dipoles. This results in large fluctuations, even larger than the homogeneous n^+ case, across the junction. In addition, we see that the fluctuations decay as a function of distance z from the n^+-region as z^{-s}, where the power s is between 1 and 2. In 3D, integrating the $1/z^3$ potential of many dipoles uniformly distributed in $-\infty < z < 0$, we would expect $s = 2$. Purely in 2D, we would expect $s = 1$, which indeed occurs at large distances away from the n^+ region, as illustrated in Fig. 9.

The situation illustrated in Fig. 9 is 'extreme' in the sense that the fluctuations of the surface potential of a MOSFET are strongly 'clamped' by the boundary conditions at the gate, so that the results of Fig. 9 largely overemphasize our point. Nevertheless, potential fluctuations caused by Coulomb effects in the S and D-regions penetrate into the channel. Their main effect is to cause a fast termalization of the electron distribution function. Consider again an extreme case: For very short channels, of length comparable to the electron mean free path in

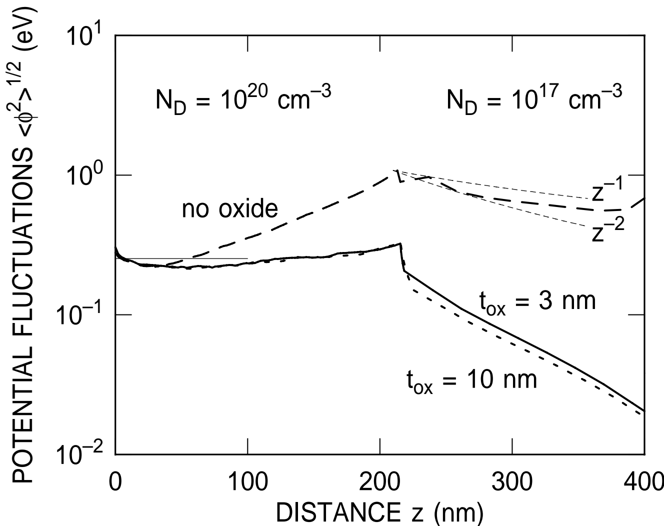


FIG. 9. Simulated rms amplitude of the potential fluctuations in a $n^= - n$ junction, without (dashed line) an intervening SiO_2 layer, and with an insulating layer 3 nm (solid line) and 10 nm-thick (dotted line). In the absence of the insulator, oscillations of electrons from the n^+ region across the junction cause large dipoles responsible for large fluctuations. Viewed from the n-region, the fluctuations decay approximately as the field of an ensemble of dipoles.

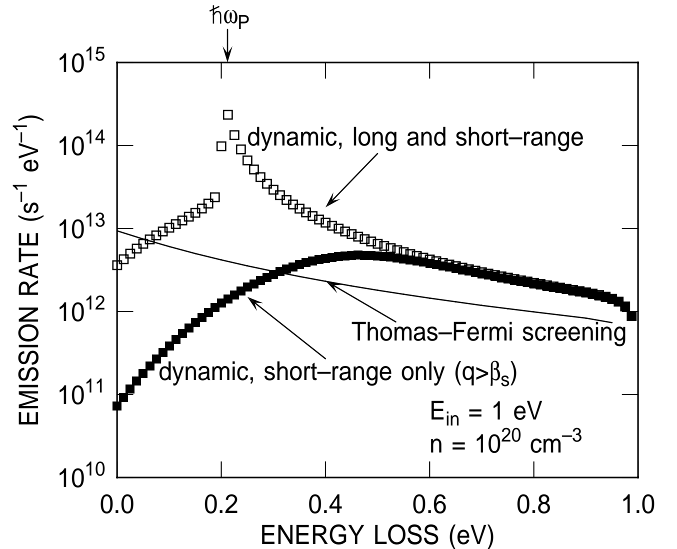


FIG. 10. Differential scattering rate at which an electron in the channel with an inital energy $E_{in} = 1$ eV loses energy between $\hbar\omega$ and $\hbar\omega + d(\hbar\omega)$ by a Coulomb excitation in the heavily-doped (10^{20} cm^{-3}) drain of a Si MOSFET. The rate has been computed employing a spherical, parabolic dispersion with an isotropic electron effective mass $m^* = 0.32 m_0$, an electron temperature $T_e = 300$ K (see Eq. (2) of Hess *et al.*[36]) and using either the static, long-wavelength Thomas-Fermi screening parameter β_s (solid line), or dynamic, wavelength-dependent screening in the high-temperature limit (open squares). The solid squares have been obtained by employing a cut-off $q > \beta_s$ to isolate the short-range contribution of the total Coulomb rate. Note the dominant contribution of long-range interactions for energy losses around the plasma energy $\hbar\omega_P$.

absence of Coulomb interactions, we would expect ballistic transport. This would result in a very narrow energy distribution function along the channel, with width controlled by the width of the electron distribution in the source. Very strong thermalization effects due to the channel/S and channel/D interactions would attempt to drive the kinetic energy distribution towards an equilibrium with the non-thermal tails observed in the n^+ regions, illustrated in Fig. 5.

This semiclassical argument can be rephrased in more conventional quantum-mechanical arguments. Recently, Hess and co-workers[36] have analyzed the electron-electron interactions in the channel of a Si MOSFET by decomposing the total scattering rate of an electron of energy E, $1/\tau_{ee}(E)$, into its 'spectral' representation $d(\hbar\omega)/\tau(E,\hbar\omega)$, representing the rate at which the electron scatters with any other electron while losing an amount of energy between $\hbar\omega$ and $\hbar\omega + d(\hbar\omega)$. (A similar decomposition relative to 'absorption processes' must also be considered in order to recover the total scattering rate). When considering scattering with the low-density gas present in the channel and using static, long-wavelength screening, the rate $1/\tau_{ee}(E)$ is indeed very small, as claimed by Hess *et al.*[36]. However, let us consider the interaction between a hot electron in the channel, of a typical initial energy $E = E_{in} = 1$eV, and the electron gas in the drain. We plot in Fig. 10 the differential emission rate, $1/\tau(E,\hbar\omega)$, as calculated ac-

cording to Eq. (2) (and plotted in Fig. 1) of Ref. 36. The solid line shows the results of using Thomas-Fermi screening, the open squares those obtained by using the dynamic, wavelength-dependent dielectric function given by Fetter and Walecka[37], as also used in Ref. 5. Finally, the solid squares are obtained from the same model, but restricting the integration over the variable y in Eq. (2) of Ref. 36 to momentum-transfers q larger than the Thomas-Fermi screening parameter β_s in the drain. This isolates the short-wavelength component of the interaction. The main feature to notice is the huge effect of dynamic screening: Very large emission rates occur for energy transfer at the plasma energy of the drain, $\hbar\omega_P$. These obviously correspond to the emission of a plasmon in the drain by the hot electron in the channel. These are indeed long-range effects, active over a large fraction of the channel of short devices, since the short-range (short-wavelength, large momentum transfer) component of the process (solid symbols) is two orders of magnitude smaller for energy losses $\approx \hbar\omega_P$. Clearly, the physical picture illustrated in Fig. 10 is fully equivalent to what is implied by Fig. 9: Either figure shows the remarkable strength of the long-range interaction between an electron in the channel and the high-density gas in the drain. A more stringent parallel between Figs. 9 and 10 is drawn in Fig. 11. Here the total (emission+absorption) electron-electron scattering rate is computed for two values of the intial electron energy by restricting the integration over momentum transfers q to increasingly larger upper bounds q_{max}. Thus, the dependence of the scattering rate on q_{max} gives an indication of dependence of the strength of the interaction on the distance $2\pi/q_{max}$. For small values of q_{max}, the scattering rate grows as q^s_{max}, where $1 < s < 2$, corresponding to a decay with distance z going as z^{-s}, as found in Fig. 9.

A comparison between Fig. 12 and Fig. 13 directly illustrates the end result of these interactions. Figure 12, adapted from Ref. 3, shows the kinetic-energy distribution of electrons along the channel of a relatively long (effective channel length $L_{eff} \approx 180$ nm) Si-on-insulator (SOI) MOSFET. At the chosen location, approximately at the channel/drain junction, the potential is about 1 V below the Fermi energy of the source. Therefore, 'ballistic' electrons can acquire a maximum kinetic energy, K_{max}, of about 1eV. The solid line, obtained using simulations which suppress all inter-electronic Coulomb interactions, shows three components: At energies of the order of a few tens of eV one sees the thermal population of electrons in the drain. At intermediate energies ($0.25 < E < 1.0$ eV), hot electrons in the channel are characterized by a smaller slope (higher effective temperature), determined by the field along the channel. Finally, the distribution appears to be truncated at the applied bias, exhibiting an exponential tail at higher energies. This tail, dropping at the lattice temperature, has been the subject of several investigations[38–41], and we do not discuss it here. When all inter-electron Coulomb interactions are included (symbols), the distribution exhibits a significant thermalization. This is observed only in the vicinity (relative to the channel length) of the channel-drain junction and the magnitude of the effect is confined to a relative size of 10^{-7} or so. When considering a device with a much smaller channel length ($L_{eff} \approx 65$ nm), the situation changes dramatically. Figure 13 shows the kinetic-energy distributions in a Si MOSFET at three locations

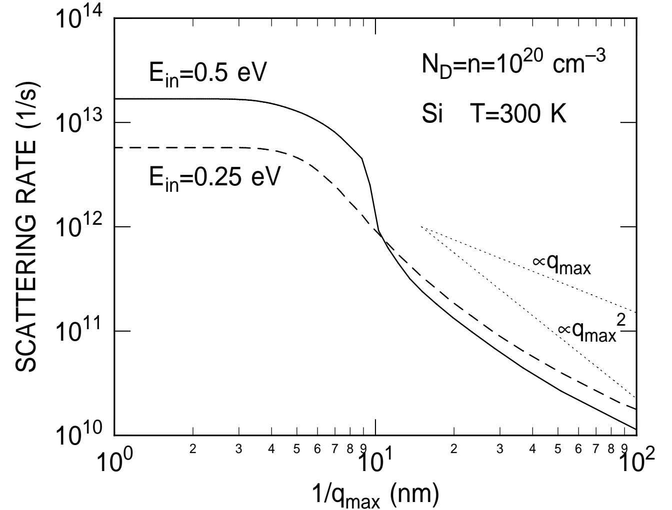


FIG. 11. Total electron-electron scattering rate computed as in Fig. 10 for two values of intial electron energy but accounting for both energy emission and absorption and restricting the integration to momentum transfers q smaller than a maximum value q_{max}. The decay of the rate at large distance $\propto 1/q_{max}$ closely resembles the decay obtained from semiclassical, 2D simulations as shown in Fig.9.

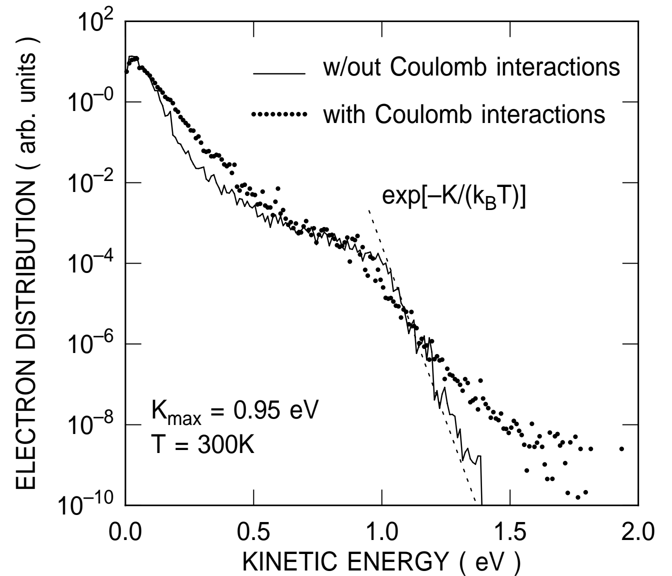


FIG. 12. Computed distribution in kinetic energy of electrons in a Si SOI MOSFET ($L_{eff} \sim 180$ nm) with a source-drain bias of 1 V at a location along the channel for which the maximum energy a ballistic carrier might acquire is expected to be about 0.95 eV. The solid line has been computed ignoring all Coulomb interactions and shows the well-known 'thermal' tail above K_{max}. The break at K_{max} disappears and the high-energy tail grows when accounting for long- and short-range Coulomb interactions (dots).

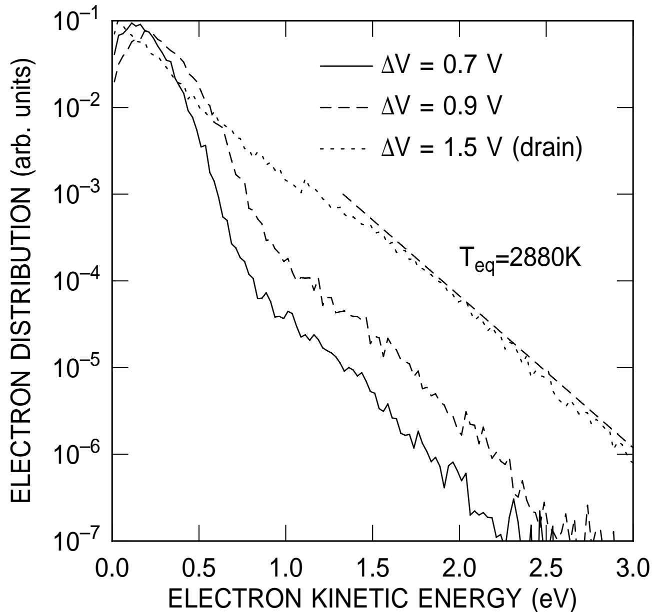


FIG. 13. Calculated electron energy distributions at three locations along the 65 nm-long channel of a Si MOSFET. The device is driven by a drain-to-source bias of 1.5 V, with 1.5 V also applied to the gate across a 2.8 nm-thick oxide. Distributions are shown at two locations along the channel at which the potential drops ΔV are 0.7 and 0.9 V, respectively, and at a location deeper inside the drain implant. The peak doping concentration in the drain is about 2.6×10^{20} As atoms/cm^3.

along the channel: As in Fig. 12 the distributions show break-points at the energy corresponding the potential drop from the source at that location (0.7, 0.9 and 1.5 V in the figure). But, unlike what is seen when neglecting Coulomb effects, at energies above the break-point the distributions inherit the 'Coulomb temperature' (see Fig. 5) of the heavily-doped drain ($N_D = n \approx 2.6 \times 10^{20}$ cm^{-3}). Note how even at mid-channel, in this device only 30 nm away from the drain and source, the distribution (solid line) is affected at the 10^{-3} level. By placing carriers at high energies, the long-range S/channel and D/channel Coulomb interactions do not remove momentum *directly* from electron gas, but *indirectly* enhance momentum-dissipating process, mainly electron-phonon collisions and impact ionizations. The G/channel interactions, instead, can *directly* cause a transfer of momentum from the electrons in the channel to those in the gate. This 'gate drag' is also expected to depress the effective electron velocity in the channel. A quantitative analysis of the importance of these effects is presented in the next section.

B. Transconductance

Using our 2D, self-consistent Monte Carlo/Poisson program[8], we have simulated a set of n-channel MOSFETs. The 'template' device, described in Ref. 42, has a metallurgical channel length, L_{met}, of about 50 nm (corresponding to an effective channel length, L_{eff} of ≈ 65 nm), S and D peak doping concentrations of about 2.6×10^{20} As atoms/cm^3, and an SiO_2 gate insulator of thickness $t_{ox} \approx 2.8$ nm. This template-device has been scaled to larger ($L_{met} \approx 100$ nm, $t_{ox} = 5.6$ nm) and smaller dimensions, down to a (probably unrealistic) 11.8 nm/0.7 nm device. While reducing the physical dimensions (junction depth, position of the peak of the retrograde substrate doping, oxide thickness, gate length) according to conventional scaling[1], we have also increased the p-type substrate doping. The peak doping levels in the S, D, and G regions, however, are already at the technological limit in the template device and have been kept unchanged.

The transport model we have employed has been amply described before. Simulations have been performed either by accounting for quantization in the channel, as in Ref. 43, or by treating all electrons according to a bulk, 3D model. The major difference observed concerns only the obvious difference in gate capacitance, and so threshold voltage and transconductance. All results given below either refer to simulations which account for quantization effects, or, when a bulk 3D model (numerically faster) was used, the transconductance has been corrected for the different gate capacitance. Refs. 5 and 8 (Appendix A) describe the electron-phonon and short-range Coulomb scattering models. We should add that the discussion of Sec. II C has prompted us to employ the 'conventional' screened Brooks-Herring model to handle scattering between electrons and ionized impurities only in the low-density channel region, but the unscreened Conwell-Weisskopf model in the high-density S, D, and G regions. In addition, Ref. 43 describes transport in inversion layers. As described in more detail in our companion paper[6], two major deviations from the model of Ref. 43 should be noted: 1. Results from Ref. 44 have prompted the use of different anisotropic intravalley deformation potentials for scattering with acoustic phonons and also intervalley deformation potentials, now from Ref. 46. The discouraging results of Ref. 43 regarding the phonon-limited mobility are now much improved[6,45]. 2. Short-range electron-electron scattering in inversion layers is now included improving on the model by Lee and Galbraith[47,48] by accounting for the anti-symmetrized matrix element, dynamic multi-subband screening, and the Green's function for the Poisson equation appropriate to the MOS system under study. As described in Appendix B, this accounts not only for polarization charges at the interfaces, but also for screening by the near-by gate. Details about the calculation of these processes is given in Appendix C. This appendix also presents the models used to account for the short-range interactions between channel and gate electrons. Their implementation in the Monte Carlo program is described in Refs. 5 and 8. Finally, we have accounted for short-range scattering between channel and gate electrons, but have neglected scattering with remote ionized impurities in the gate depletion layer, as discussed in Ref. 6. Scattering with interface roughness, a scattering process whose numerical implementation (as well as physical understanding) still leaves much to be desired, has been treated with the same formalism and parameters discussed in Ref. 43 (with channel-quantization), or – when using a bulk transport model in the channel – with an empirical mixiture of specular/diffusive reflections at the Si-SiO_2

interface, with a 75% fraction of specular scattering. This was found to provide the best agreement with both experimental data on the 'template' device and with the 'less wrong' 2D-model.

In all simulations the number of (super)particles and the mesh spacing in the heavily-doped S, D, and G regions has been chosen following the discussion in Sec. II C. The poly-crystalline Si gate has been approximated by a region of bulk Si at the appropriate doping. All short-range electron-electron processes have been treated by calculating the scattering rates described in Appendix C, but truncating the integrations over the momentum-transfers q at short wavelengths, that is, $q > \beta_s$.

Finally, comparison has been made between 'full Coulomb' simulations, 'metal gate' devices, and 'no Coulomb' results. By 'metal gate' we mean simulations performed by removing the Si gate from the simulation region and placing the gate contact (that is, the constant-potential boundary condition specifying the gate contact) directly on the 'upper' surface of the SiO_2 layer. 'No Coulomb' simulations have been performed by first simulating the devices using a number of superparticles much larger (by a factor of 10 or more) than dictated by the discusion of Sec. II C. This already damps the plasma oscillations almost completely. Having obtained a steady-state result, the potential has been smoothened by taking a time average over a sufficiently long simulation time (> 1 ps), and running the Monte Carlo simulation without both the short-range electron-electron scattering processes and the self-consistent solution of the Poisson equation. The transconductance, g_m, has been calculated by computing the drain current I_{ds} at the maximum value of drain-source bias, $V_{ds} = 1$ V for two values of gate bias, V_{gs} (= 0.75 and 1.0 V above threshold), and taking the 'numerical derivative' $\Delta I_{ds}/\Delta V_{gs}$. In all cases we have verified that this procedure yields a value very close to the maximum transconductance.

Figure 14 presents our results: The computed transconductance is plotted as a function of metallurgical channel length of the devices. We show results obtained with the 'full Coulomb' model (solid circles), with the 'metal gate' devices (open squares) (thus separating the effect of the S/channel and D/channel interactions from the G/channel 'drag') and without any Coulomb interactions (open circles): While for relatively large

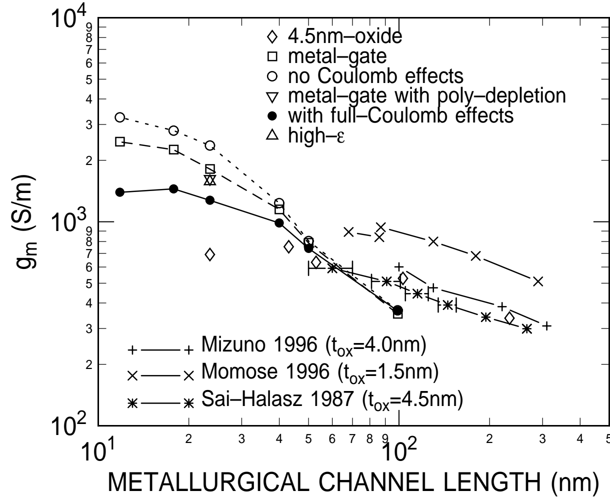


FIG. 14. Room temperature transconductance obtained from two-dimensional Monte Carlo simulations of n-MOSFETs scaled from a 'nominal' 50 nm channel-length/2.8 nm-thick oxide to smaller (11.8/0.7 nm) and larger (100/5.6 nm) devices. Results obtained accounting for Coulomb interactions with the S/D and G regions (dots, 'full Coulomb') are compared to results obtained by ignoring the channel-gate interaction (open squares, 'metal gate'), and all long- and short-range Coulomb interaction (circles, 'no Coulomb'). Results for a 23.5/1.4 nm device with a high-dielectric-constant insulator (triangle), and simulation results for 4.5 nm-oxide (open diamonds) devices are also shown. The calculated results are no more than 10% accurate, because of numerical noise and since g_m is evaluated from the difference in calculated drain current at G/S biases of 1.0 V and 0.75 V. above threshold. Finally, comparison is made with some published experimental data.

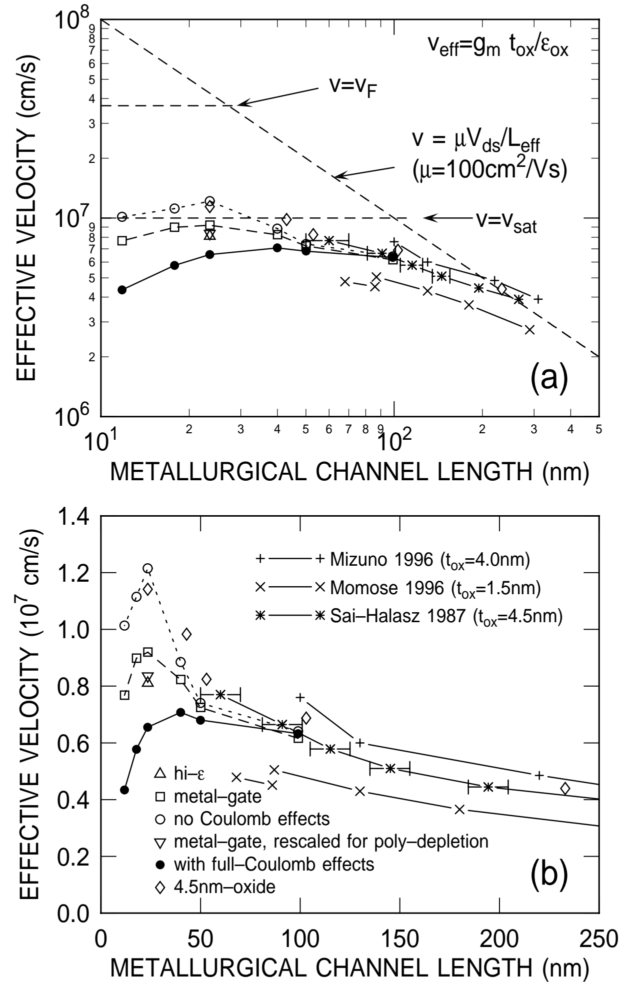


FIG. 15. Room temperature effective electron velocity, g_m/C_g, obtained from the transcondutance results shown in Fig. 14. In (a) the dashed lines represent simple theoretical expectations in various limits: The long-device limit ($v = \mu V_{ds}/L_{eff}$), the limit in which the performance is limited by saturation velocity (v_{sat}), and the ballistic limit with velocity controlled by the Fermi velocity in the source ($v = v_F$), as suggested by Datta *et al.*[49].

devices ($L_{met} \geq 40$ nm, $t_{ox} \geq 2.1$ nm) Coulomb effects appear to be small, in smaller devices they depress the performance of the transistors by as much as a factor of 2 or more at the shortest dimensions. We have also performed simulations with a gate insulator having the dielectric constant of Si (open triangle labeled 'high-ϵ'). Finally, note that since the 'metal gate' results benefit from the absence of the potential drop in the gate depletion layer, we have artificially rescaled the transconductance obtained in this case (inverted triangle) in order to make a fair comparison with other results. We also show our older simulation results[50] (no channel/G interactions) for devices having a 4.5 nm-thick oxide, as well as some experimental data from Refs. 52 and 2.

While the results of Fig. 14 already show a clear trend, the fact that we have scaled the channel length while simultaneously reducing the oxide thickness (which acts as an obvious transconductance booster) somewhat hides the physical picture. Therefore, it is convenient to plot the simulated 'effective' electron velocity, defined as transconductance, g_m, divided by gate capacitance, C_g. Figure 15 shows this quantity as a function of device dimension, both in a log-log scale (a) and in a linear scale (b). In Fig.15(a) we also show some of our 'expectations': The dashed line represents the most naive expectations for the electron velocity: In the conventional picture of very long devices, the electron velocity is controlled by the mobility and the longitudinal (source to drain) electric field. As the channel length shrinks, the velocity increases linearly without limit. Clearly, velocity saturation sets in, thus limiting this picture. However, accounting for velocity overshoot, one expects to see effective velocities exceeding the saturated value, v_{sat}. Ultimately, since transport close to the source/channel junction controls the device behavior, one may expect as an 'ultimate performance limit', to see all electrons streaming at the velocity they possess (the Fermi velocity in the source, v_F) as they are injected into the channel.[49]

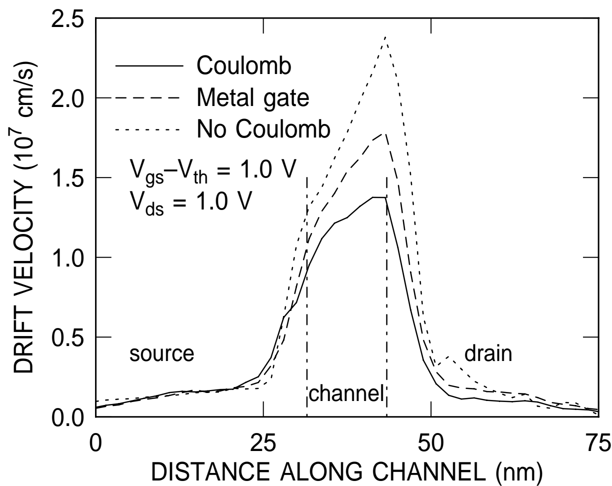


FIG. 16. Electron drift velocity along the channel of the smallest of the devices considered in Fig. 15 obtained from simulations perfomed at 300 K including all Coulomb interactions (solid line labeled 'Coulomb'), by suppressing the channel-gate interaction (dashed line labeled 'Metal gate'), and by suppressing all Coulomb interactions (dotted line).

A maximum velocity is obtained for a channel length between about 40 nm and 23.5 nm. At the minimum simulated dimensions (11.8 nm channel length, t_{ox} = 0.7 nm), the device is probably 'unrealistic' and semiclassical transport is likely to be inapplicable. Yet, within the theoretical and practical assumptions, its performance lies well below any 'ballistic limit'[49]. Coulomb interactions depress the electron velocity by more than a factor of 2, S/D-channel and G-channel interactions sharing the blame almost equally[51]. More conventional scattering mechanisms (surface-roughness and scattering with acceptors in the channel) are responsible for the remaining performance degradation seen in the 'no Coulomb' results at lengths below 23.5 nm. While Fig. 15 shows the 'effective velocity' obtained from the simulated transconductance, perhaps an indirect measure of the physical carrier velocity, these results are directly confirmed in Fig. 16 which shows the 'real' electron velocity along the channel of the smallest device. The velocity has been obtained from the Monte Carlo simulations by considering the component of the electron velocity parallel to the Si/SiO_2 interface and taking its average, weighted by the electron density, along the direction perpendicular to the interface. Consistently with the results shown in Fig. 15, both the peak velocity as well as the velocity in the proximity of the source/channel junction – in whose neighborhood the drain current is fixed – appears to be strongly depressed in equal measure by channel/gate and channel/source-channel/drain interactions.

C. Reliability

With the scaling of Si metal-oxide-semiconductor field-effect-transistors (MOSFETs), and the associated reduction of the supply voltage, it was hoped that a significant reduction of hot-electron degradation effects would take place. These hopes have been proven unwarranted both experimentally[54,55] and theoretically[39,3]: Even in relatively 'large' devices, effects caused by electrons at energies above the maximum applied bias have been observed and predicted, such as the presence of substrate currents at source-to-drain biases below the Si band-gap[54] and of gate currents at biases below the Si-SiO_2 barrier[55]. Short-range electron-electron interactions have been identified as the cause of carriers at energies above the supplied voltage via a redistribution of kinetic energy among carriers.[39,3] It is clear that long-range Coulomb interactions and the induced high-energy tails in the kinetic energy distribution of electrons in the channel constitute another reason why concerns regarding hot-carrier reliability should not be alleviated by the smaller dimensions and applied bias of scaled Si MOSFETs. Here we concentrate on the 'template', 50 nm-long device of the previous section. For this device we have performed four different types of simulations: 1. 'Full Coulomb' simulations, now employing the Monte Carlo algorithm to handle both electron and hole transport (with the parameters given in Ref. 56), and accounting fully for Coulomb effects. The use of a Monte Carlo model to treat hole transport allows us to account also for the 'impact-ionization feedback' process[57]. 2. A similar set of self-consistent simulations, but now treating hole transport with a sim-

pler drift-diffusion approximation *and* suppressing hole-initiated impact ionization. This allows us to estimate directly the role of the impact-ionization feedback process. 3. Non self-consistent simulations using a Monte Carlo scheme to treat electron and hole transport, as in 1. above, but now employing a 'frozen' field (obtained from a suitable time average of the results from the self-consistent runs), and suppressing the short-range Coulomb processes. 4. Finally, non self-consistent runs with drift-diffusion holes.

Figure 17 shows the electron energy distributions – obtained from the four simulation conditions just described – at a gate-source and drain-source bias of 1.5 V and at a lattice temperature of 300 K. The distributions have been gathered over a region surrounding the surface channel/drain junction. The non self-consistent results show the expected thermal fall-off at an energy given roughly by the applied bias. The effect of the impact-ionization feedback is clearly seen in this case. The introduction of the Coulomb interactions results in the expected build-up of high-energy tails at the temperature $T_C \approx 2900$K corresponding to the total kinetic energy (Fermi + Coulomb, $K_F + \delta K_C$) at the large electron density ($\approx 2.6 \times 10^{20}$ cm^{-3}) present in the drain region. Results of simulations performed accounting for the long-range Coulomb interactions (via the Poisson/Monte Carlo self-consistency), but suppressing the short-range interparticle collisions, show that the latter ones are not the dominant effect, unlike what is found in devices with longer (>150nm) channels. In any event, Coulomb interactions, short- and long-range, dominate over the impact-ionization feedback in enhancing the high-energy tails of the electron distributions[58]. The short-range interparticle collisions have already been shown to be more important than impact-ionization feedback[59]. Here we find that for very small devices even more important are long-range collective modes excited in the drain and penetrating well into the channel.

To further assess the importance of Coulomb effects, we have employed the degradation data of Stathis and DiMaria[60] to estimate the generation of defects in the oxide: During the Monte Carlo simulation, we have gathered the energy distribution of electrons hitting the Si-SiO_2 interface and of those emerging into the gate after tunneling across the gate insulator, as described in Ref. 5. The probability of generating defects at either interface was estimated using the model of Ref. 60 expressing the generation rate as a function of electron kinetic energy at either interface. An example of the results is shown in Fig. 18: The rate at which interface states are generated along the Si-SiO_2 interface is shown to be significantly higher, particularly at the drain-side of the channel, when accounting for Coulomb effects.

V. CONCLUSIONS

We have shown that long-range Coulomb interactions in aggressively scaled Si *n*-MOSFETs have a significant impact on the performance and reliability of the devices. Our results have been obtained using two-dimensional, semiclassical self-consistent Monte Carlo/Poisson simulatons. We have argued that, when properly implemented, these simulations can reproduce the semiclassical behavior of a homogeneous electron gas: Coulomb energy, dielectric response, band-gap narrowing, and amplitude of the (plasma) potential fluctuations obtained from the simulations compare favorably with theoretical expectations. Moreover, we have argued that quantum corrections are relatively small, albeit not negligible.

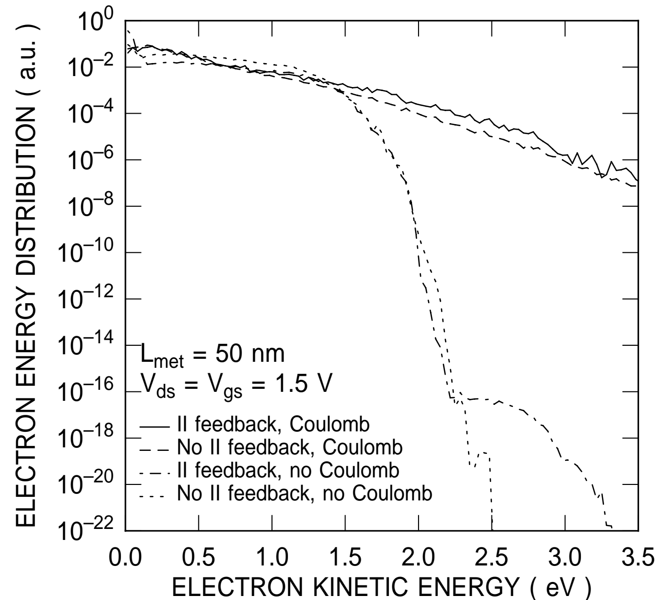


FIG. 17. Calculated electron energy distribution integrated over a region close to the channel/drain junction of a 65 nm-long channel Si MOSFET. The 'colder' distributions have been obtained from non self-consistent simulations neglecting Coulomb interactions, while the 'hotter' distributions show the effect of the (mainly long-range) Coulomb processes. The small effect of impact-ionization feedback is also shown.

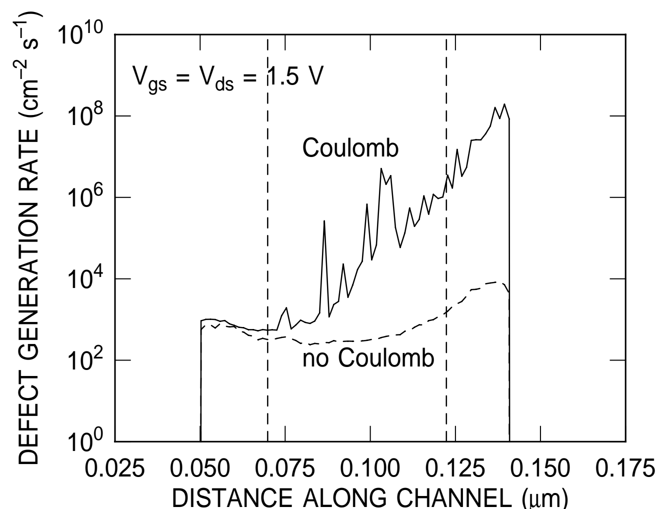


FIG. 18. Generation rate of defects at the Si-SiO_2 interaface calculated using the data of Ref. 60. The results obtained accounting for (long- and short-range) Coulomb interactions are compared with those obtained ignoring Coulomb effects.

Moving to the highly inhomogeneous situations of MOSFETs, we have shown that the plasma-fluctuations present in high-electron-density regions (source, drain, and gate) penetrate into the channel of the device to distances of the order of tens of nanometers. In short-channel devices, this is a significant fraction of the entire channel, so that transport is affected significantly: By 'thermalizing' the electron distribution to the elevated S/D effective temperature over the entire length of the channel, momentum dissipating processes are enhanced by the interactions between the electrons in the channel and those in the (close by) source and drain. Similarly, plasma fluctuations in the gate strongly couple with the channel for small thicknesses of the gate insulator and trigger a transfer of momentum (and so velocity) from the conduction electrons to the electrons in the gate. The net result is that for devices shorter than about 40 nm and oxides thinner than 2.5-3 nm the effective electron velocity can be depressed by as much as a factor of 2. In absolute terms, for all but the smallest dimensions considered, the transconductance of the devices continues to improve as device length (and oxide thickness) shrinks, but not at the pace expected when neglecting the long-range Coulomb interactions. Finally, we have also shown that hot-electron reliability is expected to be strongly influenced by Coulomb interactions.

Conceptually, the use of 2D Monte Carlo/Poisson simulations allows the study of inhomogeneous, off-equilibrium situations. Reservations remain about the range of validity of these simulations: The 'semiclassical' nature of the simulations should render them sufficiently accurate in the limit of strong injection and large plasma excitations we have considered here. Concerns about the use of two dimensional simulations should be mitigated by the slow (power-law) spatial decay of the fluctuations (seen in Fig. 9) and by the remarkable agreement between the the results of Figs. 9 and 11. Additional support to the conclusions reached here is provided by the results of the companion paper[6]. There, results obtained using three-dimensional, quantum mechanical calculations are presented focussing only on the role played by MOS interface-plasmons on reducing the effective mobility of electrons in the channel of thin-oxide structures. These calculations complement those presented here, since they are restricted only to the linear, near-equilibrium, homogenous regime, but account for the quantum-mechanical nature of the plasmons. The qualitative agreement between the two sets of results confirms our major conclusions: The performace of scaled Si devices should not be expected to improve at the pace we have been witnessing during the past decades.

ACKNOWLEDGMENTS

We thank D. K. Ferry and K. Hess for discussions related to ref. 36 and clarifications regarding Fig. 1 of this reference.

APPENDIX A: CANONICAL QUANTIZATION OF THE BULK PLASMON FIELD

In this appendix we describe the canonical quantization procedure for plasmons in a homogeneous, bulk electron gas. We compare the resulting normalization of the plasmon field with the result of simpler semiclassical energetic arguments.

Following Ngai and Economou[61], the Lagrangian density of the electron gas can be written as:

$$\mathcal{L}(\mathbf{r}) \;=\; \frac{1}{2}\, n\, m\, \pi(\mathbf{r})\cdot\pi(\mathbf{r}) \;-\; \frac{1}{2}\delta\rho(\mathbf{r})\,\phi(\mathbf{r})\;, \tag{A1}$$

where $en = \rho$ is the uniform charge density, m the electron mass, $\phi(\mathbf{r})$ is the potential, and $\pi(\mathbf{r})$ is the momentum-density

$$\pi(\mathbf{r}) \;=\; \frac{\partial \mathbf{u}(\mathbf{r})}{\partial t}\;. \tag{A2}$$

The displacement field $\mathbf{u}(\mathbf{r})$ is related to the charge fluctuations $\delta\rho(\mathbf{r})$ via

$$\frac{\delta\rho(\mathbf{r})}{\rho} \;=\; -\nabla\cdot\mathbf{u}(\mathbf{r})\;. \tag{A3}$$

Thus, Poisson equation can be written as:

$$\nabla^2\phi(\mathbf{r}) \;=\; \frac{\delta\rho(\mathbf{r})}{\epsilon} \;=\; -\frac{\rho}{\epsilon}\nabla\cdot\mathbf{u}(\mathbf{r}), \tag{A4}$$

where ϵ is the permittivity. Using the Fourier decompositions

$$\phi(\mathbf{r}) \;=\; \sum_{\mathbf{q}}\;\phi_{\mathbf{q}}\;e^{i\mathbf{q}\cdot\mathbf{r}}\;, \tag{A5}$$

and

$$\mathbf{u}(\mathbf{r}) \;=\; \sum_{\mathbf{q}}\;\mathbf{u}_{\mathbf{q}}\;e^{i\mathbf{q}\cdot\mathbf{r}}\;, \tag{A6}$$

Eq. (A4) implies

$$-\sum_{\mathbf{q}}\;q^2\phi_{\mathbf{q}}e^{i\mathbf{q}\cdot\mathbf{r}} \;=\; -i\frac{\rho}{\epsilon}\;\sum_{\mathbf{q}}\;\mathbf{q}\cdot\mathbf{u}_{\mathbf{q}}\;e^{i\mathbf{q}\cdot\mathbf{r}} \tag{A7}$$

Since plasmons are longitudinal oscillations, $\mathbf{q}\cdot\mathbf{u}_{\mathbf{q}} = qu_q$, so that $\phi_q = -i\rho u_q/(\epsilon q)$. The kinetic energy can be written as:

$$T \;=\; \frac{1}{2}\int d\mathbf{r}\; n\; m\; \pi(\mathbf{r})\cdot\pi(\mathbf{r}) \;=\; \frac{1}{2}\;n\;m\;\sum_{\mathbf{q}}\;\dot{\mathbf{u}}_{\mathbf{q}}\cdot\dot{\mathbf{u}}_{-\mathbf{q}}\;, \tag{A8}$$

where the dots indicate derivatives with respect to time. The potential energy is

$$U \;=\; \frac{1}{2}\int d\mathbf{r}\;\delta\rho(\mathbf{r})\phi(\mathbf{r}) \;=\; -\frac{1}{2}\;\frac{e^2n^2}{\epsilon}\sum_{\mathbf{q}}\mathbf{u}_{\mathbf{q}}\cdot\mathbf{u}_{-\mathbf{q}}\;. \tag{A9}$$

Defining the canonical conjugate momentum

$$\pi_{\mathbf{q}} \; = \; \frac{\partial L}{\partial \dot{\mathbf{u}}_{\mathbf{q}}} \; = \; \frac{\partial T}{\partial \dot{\mathbf{u}}_{\mathbf{q}}} \; = \; n \; m \; \dot{\mathbf{u}}_{-\mathbf{q}} \; , \tag{A10}$$

the Hamiltonian becomes

$$H \; = \; \frac{1}{2} \sum_{\mathbf{q}} \left\{ \frac{1}{enm} \; \pi_{\mathbf{q}} \cdot \pi_{-\mathbf{q}} \; + \; \frac{e^2 n^2}{\epsilon} \; \mathbf{u}_{\mathbf{q}} \cdot \mathbf{u}_{-\mathbf{q}} \right\} \; . \tag{A11}$$

Quantization can be now performed imposing the canonical commuation rules (CCR)

$$[u_{i,\mathbf{q}}, \pi_{j,\mathbf{q}'}] \; = \; i\hbar \; \delta_{ij} \; \delta_{\mathbf{q}\mathbf{q}'} \; . \tag{A12}$$

Following a standard procedure, we rewrite the Hamiltonian as

$$H \; = \; \frac{1}{2} \; \sum_{\mathbf{q}} \; [\; P_{\mathbf{q}} P_{-\mathbf{q}} \; + \; Q_{\mathbf{q}} Q_{-\mathbf{q}} \;] \; , \tag{A13}$$

having set:

$$P_{\mathbf{q}} \; = \; \frac{\pi_{\mathbf{q}}}{(enm)^{1/2}} \; , \tag{A14}$$

$$Q_{\mathbf{q}} \; = \; \left(\frac{e^2 n^2}{\epsilon} \right)^{1/2} \; u_{\mathbf{q}} \; . \tag{A15}$$

Defining now:

$$c^{\dagger}_{\mathbf{q}} \; = \; \frac{1}{(2\hbar\omega_P)^{1/2}} \; (Q_{-\mathbf{q}} - iP_{\mathbf{q}}) \; , \tag{A16}$$

$$c_{\mathbf{q}} \; = \; \frac{1}{(2\hbar\omega_P)^{1/2}} \; (Q_{\mathbf{q}} + iP_{-\mathbf{q}}) \; , \tag{A17}$$

where we have defined the plasma frequency $\omega_P = [(en)/(\epsilon m)]^{1/2}$, the Hamiltonian Eq. (A11) becomes

$$H \; = \; \frac{\hbar\omega_P}{2} \sum_{\mathbf{q}} \left\{ c^{\dagger}_{\mathbf{q}} c_{\mathbf{q}} \; + \; c_{\mathbf{q}} c^{\dagger}_{\mathbf{q}} \right\} \; . \tag{A18}$$

Using the definitions (A14)-(A17) and the CCR (A12), the only nonvanishing commutator for the operators $c^{\dagger}_{\mathbf{q}}$ and $c_{\mathbf{q}}$ is

$$[c^{\dagger}_{\mathbf{q}}, c_{\mathbf{q}}] \; = \; -\frac{1}{\omega_P} \left(\frac{en}{\epsilon m} \right)^{1/2} \; = \; -1 \; . \tag{A19}$$

This shows that the operators $c^{\dagger}_{\mathbf{q}}$ and $c_{\mathbf{q}}$ behave like raising (creation) and lowering (annihilation) operators, so that the Hamiltonian Eq. (A18) becomes a collection of harmonic oscillators

$$H \; = \; \hbar\omega_P \sum_{\mathbf{q}} \left\{ c^{\dagger}_{\mathbf{q}} c_{\mathbf{q}} \; + \frac{1}{2} \right\} \; , \tag{A20}$$

while the displacement field and the potential can be expressed in terms of the creation and annihilation operators

$$u_{\mathbf{q}} \; = \; \left(\frac{\hbar\omega_P \epsilon}{2e^2 n^2} \right)^{1/2} \; (c^{\dagger}_{-\mathbf{q}} + c_{\mathbf{q}}) \; , \tag{A21}$$

$$\phi_{\mathbf{q}} \; = \; -\frac{i}{q} \left(\frac{\hbar\omega_P}{2\epsilon} \right)^{1/2} \; (c^{\dagger}_{-\mathbf{q}} + c_{\mathbf{q}}) \; . \tag{A22}$$

The root-mean-square (rms) amplitude of the field associated with the Fock state $\mid n_{\mathbf{q}} >= (n_{\mathbf{q}}! \;)^{-1/2} \; (c^{\dagger}_{\mathbf{q}})^{n_{\mathbf{q}}} \mid 0 >$ (where $\mid 0 >$ is the vacuum) containing $n_{\mathbf{q}}$ plasmons of wave vector $\mathbf{q}$ can be obtained from the expectation value:

$$< n_{\mathbf{q}} \mid \phi^{\dagger}_{-\mathbf{q}} \phi_{\mathbf{q}} \mid n_{\mathbf{q}} > \; = \; \frac{\hbar\omega_P}{2\epsilon} \; \frac{1}{q^2} \; (1 + 2n_{\mathbf{q}}) \; . \tag{A23}$$

Equation (A23) is our final result: Setting $n_{\mathbf{q}} = 0$ above, it shows that plasmon zero-point motion at wave vector $\mathbf{q}$ is associated with a field of rms amplitude

$$\mid \phi_{\mathbf{q}} \mid \; = \; \left(\frac{\hbar\omega_P}{2\epsilon} \right)^{1/2} \; \frac{1}{q} \; . \tag{A24}$$

Can we obtain this result without going through the procedure of canonical quantization? Stern and Ferrel[62] have provided the following semiclassical argument, based on energetic considerations: Let us write the magnitude of the plasmon field at wave vector $\mathbf{q}$ as

$$E_{q,\omega}(\mathbf{r}, t) \; = \; a_q \; \cos(\mathbf{q} \cdot \mathbf{r} - \omega_P t) \; , \tag{A25}$$

so that the electrostatic potential is

$$\phi_{q,\omega}(\mathbf{r}, t) \; = \; - \; \frac{a_q}{q} \; \sin(\mathbf{q} \cdot \mathbf{r} - \omega_P t) \; . \tag{A26}$$

For harmonic oscillators such as plasmons, the time-averaged kinetic energy, $< T >$, is equal to the time-averaged potential energy, $< U >$. Thus, the total energy, W, associated with the zero-point plasmon field over a volume Ω, *including* its own self-energy (which explains the absence of the factor 1/2 in front of the integral below), will be:

$$\begin{aligned} W \; &= < T > + < U > \; = \; 2 < U > \\ &= \; 2 \; \left\langle \int_{\Omega} d\mathbf{r} \; \phi^*_{q,\omega}(\mathbf{r}) \; \rho_{q,\omega}(\mathbf{r}) \right\rangle \; = \; \epsilon \mid a_q \mid^2 \Omega \; , \end{aligned} \tag{A27}$$

where ϵ must be understood to be ϵ^{∞} in our case, since $\rho_{q,\omega}(\mathbf{r})$ is just a microscopic polarization charge, so that only the background (*i.e.*, valence) dielectric response should be considered. Quantum mechanically, this energy must represent the zero-point plasmon energy within the volume Ω, so that $W \; = \; \frac{1}{2} \; \hbar\omega_P \Omega$, which implies

$$\mid \phi_{\mathbf{q}} \mid \; = \; \frac{\mid a_q \mid}{q} \; = \; \left(\frac{\hbar\omega_P}{2\epsilon} \right)^{1/2} \; \frac{1}{q} \; , \tag{A28}$$

in agreement with Eq. (A24).

Another example is given by plasmons at the interface between a dielectric with permittivity ϵ_1 for $z < 0$ and a semiconductor with a model dielectric function $\epsilon_2(\omega) = \epsilon_2^{\infty}(1 - \omega^2_{P,2}/\omega^2)$ for $z \geq 0$. The electrostatic potential associated with an interface plasmon of frequency ω_{SP} has the form

$$\phi_{Q,\omega_{SP}}(z, \mathbf{R}, t) \; = \; a_Q \; e^{-Q|z|} \; \cos(\mathbf{Q} \cdot \mathbf{R} - \omega_{SP} t) \; . \tag{A29}$$

Here upper-case letters denote the two-dimensional vectors on the plane of the insulator-semiconductor interface, while z is the coordinate of the axis normal to this plane. The continuity of the 'in plane' electric field and of the 'normal' displacement field requires $\epsilon_1 + \epsilon_2(\omega_{SP}) = 0$, which implies an interface plasma frequency lower than its bulk value $\omega_{P,2}$, $\omega_{SP} = \omega_{P,2}/(1+\epsilon_1/\epsilon_2^\infty)^{1/2}$. The total energy associated with the field Eq. (A29) over an area A will be, as in (A27) above,

$$W \ = | a_Q |^2 \ \frac{\epsilon_1 + \epsilon_2^\infty}{Q} \ A \ . \tag{A30}$$

Setting $W = (1/2)\hbar\omega_{SP}A$, we have

$$| \phi_{\mathbf{Q}} | = \left[\frac{\hbar\omega_{SP}}{2(\epsilon_1 + \epsilon_2^\infty)} \right]^{1/2} \frac{1}{Q^{1/2}} \ , \tag{A31}$$

in agreement with the result obtained by Ferrel[63].

APPENDIX B: GREEN'S FUNCTION FOR THE MOS GEOMETRY

In this appendix we derive the Green's function for the Poisson equation in the MOS geometry described by a medium of permittivity ϵ_g for $z < 0$ (gate), an insulator with permittivity ϵ_i for $0 < z < t$, and a semiconducting substrate with ϵ_s for $z > t$. We consider first the 'bare' Green's function, *i.e.* the case of constant (valence-electron only) permittivities. Next, we shall consider the Green's function screened by bulk free (conduction) electrons in the gate and substrate.

i) Bare Green's function. We must solve the equation

$$\frac{d^2 G_Q^{(j)}(z,z')}{dz^2} \ - Q^2 G_Q^{(j)}(z,z') \ = \ \delta(z-z') \ , \tag{B1}$$

where $\mathbf{Q}$ is the two-dimensional wave vector in the plane of the interface, and $j = g$ when the source is in the gate ($z' < 0$), $j = s$ when the source is in the substrate ($z' > t$). The boundary conditions are:

$$\begin{cases} G_Q^{(j)}(z=0^-,z') & = & G_Q^{(j)}(z=0^+,z') \\ G_Q^{(j)}(z=t^-,z') & = & G_Q^{(j)}(t=0^+,z') \\ \epsilon_g^\infty \frac{dG_Q^{(j)}(z=0^-,z')}{dz} & = & \epsilon_i^\infty \frac{dG_Q^{(j)}(z=0^+,z')}{dz} \\ \epsilon_i^\infty \frac{dG_Q^{(j)}(z=t^-,z')}{dz} & = & \epsilon_s^\infty \frac{dG_Q^{(j)}(z=t^+,z')}{dz} \end{cases} \ , \tag{B2}$$

with the supplemental condition

$$\frac{dG_Q^{(j)}(z=z'^-,z')}{dz} \ - \ \frac{dG_Q^{(j)}(z=z'^+,z')}{dz} \ = \ -1 \ , \tag{B3}$$

expressing the presence of the delta-function at $z = z'$. We consider first the case of the source-term in the gate, that is, $z' < 0$, which applies to the case of the interaction between an electron in the channel (Si substrate) and one in the gate. For convenience, let us look for a general solution expressed in the form:

$$G_Q^{(g)}(z,z') = \ - \ \frac{1}{2Q} \times \begin{cases} a_Q \ e^{-Q|z-z'|} + b_Q \ e^{-Q|z+z'|} & (z<0) \\ c_Q \ e^{Q(z+z')} + d_Q \ e^{-Q(z-z')} & (0 \le z < t) \\ f_Q \ e^{-Q(z-z')} & (z \ge t) \end{cases} \ . \tag{B4}$$

Equation (B3) immediately implies $a_Q = 1$, while the set of Eqns. (B2) become:

$$\begin{cases} 1 + b_Q & = & c_Q + d_Q \\ c_Q \ e^{Qt} \ + \ d_Q \ e^{-Qt} & = & f_Q \ e^{-Qt} \\ 1 - b_Q & = & \frac{\epsilon_i^\infty}{\epsilon_g^\infty} \ (d_Q - c_Q) \\ c_Q \ e^{Qt} \ - d_Q \ e^{-Qt} & = & -\frac{\epsilon_s^\infty}{\epsilon_i^\infty} \ f_Q \ e^{-Qt} \end{cases} \ . \tag{B5}$$

Solving this system and inserting the results into Eq. (B4) we find:

$$G_Q^{(g)}(z,z') \ = \ - \ \frac{1}{2Q} \times \begin{cases} e^{-Q|z'-z|} \ + \ \frac{\mathcal{H}_Q}{\mathcal{C}_Q} e^{-Q|z'+z|} & (z<0) \\ \frac{2\epsilon_g^\infty}{\mathcal{C}_Q} \left[(\epsilon_i^\infty - \epsilon_s^\infty) e^{-Qt} e^{Q(z'+z)} \ + \ (\epsilon_i^\infty + \epsilon_s^\infty) e^{Qt} e^{Q(z'-z)} \right] & (0 \le z < t) \\ \frac{4\epsilon_g^\infty \epsilon_i^\infty}{\mathcal{C}_Q} e^{Qt} e^{Q(z'-z)} & (z \ge t) \end{cases} \ , \tag{B6}$$

with

$$\mathcal{C}_Q \ = \ e^{Qt}(\epsilon_s^\infty + \epsilon_i^\infty)(\epsilon_g^\infty + \epsilon_i^\infty) \ - \ e^{-Qt}(\epsilon_s^\infty - \epsilon_i^\infty)(\epsilon_g^\infty - \epsilon_i^\infty) \ , \tag{B7}$$

and

$$\mathcal{H}_Q \ = \ e^{Qt}(\epsilon_s^\infty + \epsilon_i^\infty)(\epsilon_g^\infty - \epsilon_i^\infty) - \ e^{-Qt}(\epsilon_s^\infty - \epsilon_i^\infty)(\epsilon_g^\infty + \epsilon_i^\infty) \ . \tag{B8}$$

When the source term is in the substrate ($z' > t$), the case of interest when dealing with short-range interactions among electrons in the channel or with the dielectric response of the 2DEG, following the same procedure we find:

$$G_Q^{(s)}(z,z') \ = \ -\frac{1}{2Q} \times \begin{cases} \frac{4\epsilon_s^\infty \epsilon_i^\infty}{\mathcal{C}_Q} e^{Qt} e^{-Q(z'-z)} & (z<0) \\ \frac{2\epsilon_s^\infty}{\mathcal{C}_Q} e^{Qt} \left[(\epsilon_i^\infty - \epsilon_g^\infty) e^{Q(z'+z)} \ + \ (\epsilon_i^\infty + \epsilon_g^\infty) e^{Q(z'-z)} \right] & (0 \le z < t) \\ e^{-Q|z'-z|} \ + \ \frac{\mathcal{E}_Q}{\mathcal{C}_Q} e^{2Qt} e^{-Q|z'+z|} & (z \ge t) \end{cases} \ , \tag{B9}$$

with

$$\mathcal{E}_Q \ = \ e^{Qt}(\epsilon_g^\infty + \epsilon_i^\infty)(\epsilon_s^\infty - \epsilon_i^\infty) \ - \ e^{-Qt}(\epsilon_s^\infty + \epsilon_i^\infty)(\epsilon_g^\infty - \epsilon_i^\infty) \ . \tag{B10}$$

ii) Screened Green's function. We now consider the Green's function screened by bulk free electrons in the gate and in the substrate. Assuming their response is isotropic and homogeneous, the constant permittivities ϵ_g^∞ and ϵ_s^∞ are replaced by $\epsilon_g^\infty[1+\beta_g(q,\omega)/q^2]$ and $\epsilon_s^\infty[1+\beta_s(q,\omega)/q^2]$, respectively, $\beta_j(q,\omega)$ (for $j = g,\ s$) being the dynamic and wavelength-dependent screening parameter in the random-phase-approximation (RPA). We can proceed in two alternative ways. The first approach consists in recalling that $eG_Q^{(j)}(z,z')\exp[i\mathbf{Q}\cdot(\mathbf{R}-\mathbf{R}')]/\epsilon_j^\infty$, as derived above, is the bare potential due to a charge e at $(\mathbf{R}', z')$. Proceeding as usual, we consider the potential due to the 'external' charge $e\delta^{(2)}(\mathbf{R}-\mathbf{R}')\delta(z-z')$ *plus* the potential caused by the polarization charge induced by $G_Q^{(j)}(z,z')\exp[i\mathbf{Q}\cdot(\mathbf{R}-\mathbf{R}')]$ itself. Thus we derive for the Green's function $\tilde{G}_Q^{(j)}(z,z')$ screened by free electrons in region j' ($=g,\ s$) the integral equation:

$$\tilde{G}_Q^{(j)}(z,z') \ = \ G_Q^{(j)}(z,z') \ + \sum_{q_z,j',j''} \beta_{j'}^2(Q_z,\omega) \int_{\Omega_{j'}} dz'' \int_{\Omega_{j''}} dz''' \ G_Q^{(j')}(z,z'') \ e^{iq_z(z''-z''')} \ \tilde{G}_Q^{(j)}(z''',z') \ , \tag{B11}$$

where Ω_j is the range of z associated with region j and $Q_z = (Q^2+q_z^2)^{1/2}$. This integral equation is hard to solve and may require the use of the Hartee approximation (that is, lowest-order in screening replacing $\tilde{G}_Q^{(j)}(z''',z')$ with $G_Q^{(j)}(z''',z')$ inside the integral) and/or ignoring the frequency and wavelength dependence of the screening parameter $\beta_j^2(Q_z,\omega)$ (that is, by restricting the analysis to the Thomas-Fermi or Debye-Hückel limits). The former approximation mistreats the small-Q region, where screening is strong, the latter the high-Q region. A preferable approach is to solve – albeit with additional approximations – the fully-screened problem

$$\nabla_{\mathbf{r}} \cdot \left[\int d\mathbf{r}'' \epsilon(\mathbf{r},\mathbf{r}'') \ \nabla_{\mathbf{r}''} \tilde{G}^{(j)}(\mathbf{r}'',\mathbf{r}') \right] \ = \ \delta^{(3)}(\mathbf{r}-\mathbf{r}') \ . \tag{B12}$$

Let us write:

$$\tilde{G}^{(j)}(\mathbf{r},\mathbf{r}') \ = \ \sum_{\mathbf{Q}} \ e^{i\mathbf{Q}\cdot(\mathbf{R}-\mathbf{R}')} \ \tilde{G}_Q^{(j)}(z,z') \ , \tag{B13}$$

(understanding that in the following we must take the real part of the complex exponentials) and let's make the first major approximation of neglecting higher-order terms expressing the effect of polarization charges in region j on the potential in region $j' \neq j$. Thus, we retain self-consistent screening within each region j', but assume linear screening across regions. Then, for $z \in \Omega_{j'}$ and $z' \in \Omega_j$:

$$\nabla_{\mathbf{r}} \cdot \left[\int d\mathbf{r}'' \epsilon(\mathbf{r},\mathbf{r}'') \ \nabla_{\mathbf{r}''} \tilde{G}^{(j)}(\mathbf{r}'',\mathbf{r}') \right] \ = $$

$$\sum_{\mathbf{Q}q_z} \int_{\Omega_{j'}} dz'' \ e^{i\mathbf{Q}\cdot(\mathbf{R}-\mathbf{R}')} \ \ e^{iq_z(z-z'')} \ \epsilon_{j'}(Q_z,\omega) \left[-Q^2 \tilde{G}_Q^{(j)}(z'',z') \ + \ iq_z \frac{d\tilde{G}_Q^{(j)}(z'',z')}{dz''} \right] \ . \tag{B14}$$

Solving Eq. (B12) amounts to setting the term within the square brackets in Eq. (B14) above equal to $\delta_{jj'}\delta(z''-z')/\epsilon_{j'}(Q_z,\omega)$. In order to reduce the problem to a form similar to the unscreened case, we make the second major approximation of replacing in Eq. (B14) the q_z-dependence of $\epsilon_{j'}(Q_z,\omega)$ with an average value which, by analogy with the following, we take to be $\tilde{\epsilon}_{j'}(Q,\omega)$ defined by Eqns. (B18) and (B20) below. Integrating by parts the second term inside the integral over z'' and noticing that the 'interface' terms cancel by the boundary conditions (B16) below, thanks to our approximation, solving Eq. (B11) is now reduced to finding a solution of the equation

$$\frac{d^2\tilde{G}_Q^{(j)}(z,z')}{dz^2} \ - Q^2 \tilde{G}_Q^{(j)}(z,z') \ = \ \frac{1}{\tilde{\epsilon}_j(Q,\omega)} \ \delta(z-z') \ , \tag{B15}$$

which is identical to Eq. (B1) with the exception of an overall multiplicative factor $1/\tilde{\epsilon}_j(Q,\omega)$ and, more important, by the different boundary conditions, still expressing continuity across the interfaces of the parallel component of the electric field and the normal component of the electric displacement field:

$$\begin{cases} \tilde{G}_Q^{(j)}(z=0^-,z') \; = \; \tilde{G}_Q^{(j)}(z=0^+,z') \\ \tilde{G}_Q^{(j)}(z=t^-,z') \; = \; \tilde{G}_Q^{(j)}(t=0^+,z') \\ \sum_{q_z} \int_{-\infty}^{0} dz'' e^{iq_z(z-z'')} \;\; \epsilon_g(Q_z;\omega) \frac{d\tilde{G}_Q^{(j)}(z'',z')}{dz''} \Big|_{z\to 0^-} \; = \; \epsilon_i(\omega) \frac{d\tilde{G}_Q^{(j)}(z=0^+,z')}{dz} \\ \epsilon_i(\omega) \frac{d\tilde{G}_Q^{(j)}(z=t^-,z')}{dz} \; = \; \sum_{q_z} \int_t^{\infty} dz'' e^{iq_z(z-z'')} \;\; \epsilon_s(Q_z;\omega) \frac{d\tilde{G}_Q^{(j)}(z'',z')}{dz''} \Big|_{z\to t^+} \end{cases} , \tag{B16}$$

having assumed a frequency dependence also for the insulator dielectric function. Typically, this may be due to the TO phonons in polar insulators, such as SiO_2. The solution is formally very similar to Eq. (B6). When the source-term is in the gate ($z' < 0$), in analogy with the unscreened case we define the Green's function as the solution of Eq. (B15), but multiplied by the 'constant' $\tilde{\epsilon}_j(Q,\omega)$. Thus we have:

$$\tilde{G}_Q^{(g)}(z,z') \; = \; -\frac{1}{2Q} \times \begin{cases} e^{-Q|z'-z|} \; + \; \frac{\tilde{\mathcal{H}}_Q}{\tilde{\mathcal{C}}_Q} e^{-Q|z'+z|} & (z<0) \\ \frac{\tilde{\epsilon}_g(Q,\omega)+\widehat{\epsilon}_g(Q,\omega;z')}{\tilde{\mathcal{C}}_Q} \left\{ [\epsilon_i(\omega)-\tilde{\epsilon}_s(Q,\omega)] e^{-Qt} e^{Q(z'+z)} + [\epsilon_i(\omega)+\tilde{\epsilon}_s(Q,\omega)] e^{Qt} e^{Q(z'-z)} \right\} & (0 \le z < t) \\ \frac{2[\tilde{\epsilon}_g(Q,\omega)+\widehat{\epsilon}_g(Q,\omega;z')] \;\; \epsilon_i(\omega)}{\tilde{\mathcal{C}}_Q} e^{Qt} e^{Q(z'-z)} & (z \ge t) \end{cases} , \tag{B17}$$

where we have defined 'effective' (*i.e.*, averaged over the normal direction) dielectric functions

$$\tilde{\epsilon}_j(Q,\omega) \; = \; \frac{1}{\pi} \int_{-\infty}^{\infty} dq_z \; \frac{Q}{Q_z^2} \; \epsilon_j(Q_z,\omega) \; , \tag{B18}$$

$$\widehat{\epsilon}_j(Q,\omega;z') \; = \; \frac{1}{\pi} \int_{-\infty}^{\infty} dq_z \; \frac{2q_z e^{Q|z'|} \sin(q_z|z'|) - Q}{Q_z^2} \; \epsilon_j(Q_z,\omega) \; . \tag{B19}$$

Note that in the long-wavelength limit ($\epsilon(q,\omega)$ independent of q) these equations imply $\tilde{\epsilon}_j(Q,\omega), \widehat{\epsilon}_j(Q,\omega;z') \to \epsilon_j(\omega)$. On the contrary, in the Thomas-Fermi limit ($\beta_j(q;\omega) \to$ constant), we have

$$\tilde{\epsilon}_j(Q,\omega) \; , \widehat{\epsilon}_j(Q,\omega) \; \to \; \epsilon_j(\sqrt{2}Q,\omega) \; . \tag{B20}$$

In order to simplify the evaluation of the q_z-averaged dielectric functions $\tilde{\epsilon}_j$ and $\widehat{\epsilon}_j$, we retain this approximations even when going beyond the Thomas-Fermi limit, namely, when employing the Fetter-Walecka[37] expression for the dielectric functions.[64]

Similarly, when the source is in the substrate ($z' > t$), we have in complete analogy with Eq. (B9):

$$\tilde{G}_Q^{(s)}(z,z') \; = \; -\frac{1}{2Q} \times \begin{cases} \frac{2[\tilde{\epsilon}_s(Q,\omega)+\widehat{\epsilon}_s(Q,\omega;z')]\epsilon_i(\omega)}{\mathcal{C}_Q} e^{Qt} e^{-Q(z'-z)} & (z<0) \\ \frac{\tilde{\epsilon}_s(Q,\omega)+\widehat{\epsilon}_s(Q,\omega;z')}{\tilde{\mathcal{C}}_Q} e^{Qt} \;\; \left\{ [\epsilon_i(\omega)-\tilde{\epsilon}_g(Q,\omega)] \; e^{Q(z'+z)} + [\epsilon_i(\omega)+\tilde{\epsilon}_g(Q,\omega)] e^{Q(z'-z)} \right\} & (0 \le z < t) \\ e^{-Q|z'-z|} \; + \; \frac{\tilde{\mathcal{E}}_Q}{\tilde{\mathcal{C}}_Q} e^{2Qt} e^{-Q|z'+z|} & (z \ge t) \end{cases} , \tag{B21}$$

The quantities $\tilde{\mathcal{E}}_Q$ $\tilde{\mathcal{H}}_Q$ and $\tilde{\mathcal{C}}_Q$ are identical to those defined in Eqns. (B7) and (B8), with the dielectric constants ϵ_j^{∞} replaced by the 'normal' averages $\tilde{\epsilon}_j(Q,\omega)$. Note, finally, how electrons do screen the source term even when this is on the other side of the insulator. This effect weakens, as expected, as the insulator thickness increases.

APPENDIX C: SHORT-RANGE ELECTRON-ELECTRON INTERACTION

In this appendix we present the short-range electron-electron scattering rates we have employed in our simulatons. Since transport in Si inversion layers is modeled by representing low-energy electrons as quantized, 2D carriers, and high-energy electrons as bulk particles, as described in Ref. 43, the co-existence of 2D and 3D model-electrons in the channel implies that we must consider several types of 'pairs' of scattering particles: 2D electrons in the channel scattering among themselves, 2D electrons in the channel scattering with bulk electrons in the gate, bulk electrons in the channel scattering with bulk gate-electrons, etc.

Details about the implementation of short-range electron-electron scattering in Monte Carlo simulations are given in Refs. 8 and 5. In essence, for each 'primary' electron, a 'partner' is selected among those particles within a screening length from the primary particle. The rate between these two electrons is computed using the local density. This procedure amounts to a stochastic evaluation of the integration over the local distribution in k-space of partner electrons. Note that the scattering probability employed in the Monte Carlo code is computed using only one half of the 'correct' scattering rates (as given in Refs. 8 and 5 and below), in order to account for the fact that two particles scatter at each event. Note also that, as described in the text, in order to avoid double counting the long-range Coulomb interactions, all short-range scattering rates are obtained by restricting the integration over the matrix element to values of the momentum transfer larger that the screening length, so that we account only for scattering at 'impact parameters' – or inter-particle distances - smaller than the mesh-spacing used to solve the discretized Poisson equation. This effectively suppresses the long-range plasmon contribution to the interaction rate, since it precludes the squared Coulomb matrix element in Eqns. (C5), (C14), and (C16) from reaching its small-Q poles along the dispersion of the interface plasmons, expressed mainly by the vanishing of the factor $\tilde{\mathcal{C}}_Q^2$, defined in Appendix B (cf. Eq. (20) of Ref. 6).

Scattering among bulk electrons in the channel and among those in the gate is treated as in bulk Si (*i.e.*, using the bulk Green's function of the Poisson equation). Indeed, scattering between two bulk electrons in the substrate occurs mainly in the 'pinched-off' region of the channel, where quantization effects are weak and the carriers are sufficiently far from the interface to warrant a bulk description of their properties, while the choice of scattering model to treat the collision between two electrons in the gate is largely immaterial, since the gate remains very close to thermal equilibrium. In addition, scattering between a pair of electrons in the channel described by different transport models (one electron as a bulk electron, the other one as a 2D electron), is a relatively rare event and we (mis)treat it by converting the 2D electron to a bulk electron, as described in Ref. 43, letting the pair scatter as in a bulk process, and converting one or both of the final bulk states to 2D-states whenever the energy of the final states is below the threshold identifying the boundary between 2D and bulk transport models in the channel. This has been typically set by the energy of the bottom of the 10-th unprimed subband. On the contrary, scattering between different 'types' of pairs is strongly affected by the presence of the interfaces and by screening at either side of the insulator, so these processes are treated using the Green's function for the MOS geometry, as follows.

i) Interaction among 2D-electrons in the channel. We follow Lee and Galbraith[47,48], but extend their result using dynamic screening and wave-function (anti)symmetrization. Let us consider an electron of two-dimensional wave vector $\mathbf{K}$ in subband μ interacting with an electron of wave vector $\mathbf{P}$ in subband ν. Ignoring dielectric screening and spin for now, the Coulomb matrix element for a transition $|\mathbf{K}\mu> \to |\mathbf{K}'\mu'>$, $|\mathbf{P}\nu> \to |\mathbf{P}'\nu'>$ will be:

$$< \mathbf{K}'\mu';\mathbf{P}'\nu'|V|\mathbf{K}\mu;\mathbf{P}\nu > \\ = \frac{e^2}{\epsilon_s^\infty}\, \mathcal{G}^{(s)}_{Q;\mu\mu',\nu\nu'}\ \delta^{(2)}(\mathbf{K}+\mathbf{P}-\mathbf{K}'-\mathbf{P}')\ , \qquad \text{(C1)}$$

where $Q = |\mathbf{K}-\mathbf{K}'|$ and $\mathcal{G}^{(s)}_{Q;\mu\mu',\nu\nu'}$ is the 'form-factor' (see Eq. (A10) of Ref. 6):

$$\mathcal{G}^{(s)}_{Q;\lambda\lambda';\mu\mu'} \\ = \int dz \int dz'\ \zeta_\lambda(z)\ \zeta_{\lambda'}(z)\ G_Q^{(s)}(z,z')\ \zeta_\mu(z')\ \zeta_{\mu'}(z')\ . \quad \text{(C2)}$$

$G_Q^{(s)}(z,z')$ being the Green's function of the MOS system given by Eq. (B9) of Appendix B. Acccounting for screening by the 2DEG we must replace $\mathcal{G}^{(s)}_{Q;\mu\mu',\nu\nu'}$ with the screened quantity $\tilde{\mathcal{G}}^{(s)}_{Q,\omega;\mu\mu',\nu\nu'}$ solution of the system

$$\tilde{\mathcal{G}}^{(s)}_{Q,\omega;\mu\mu',\nu\nu'} = \mathcal{G}^{(s)}_{Q;\mu\mu',\nu\nu'} + \\ \sum_{\lambda\lambda'} \mathcal{G}^{(s)}_{Q;\mu\mu',\lambda\lambda'}\ \beta_{\lambda\lambda'}(Q,\omega)\ \tilde{\mathcal{G}}^{(s)}_{Q,\omega;\lambda\lambda',\nu\nu'}\ , \qquad \text{(C3)}$$

where $\hbar\omega = E_\mu(\mathbf{K}) - E_{\mu'}(\mathbf{K}')$ is the energy exchanged in the interaction and the other symbols are as in Appendix A of Ref. 6. In order to account for screening by the bulk electrons in the gate and in the substrate (often present in our MC simulations in which bulk and two-dimensional electrons may artificially coexist at the same location) we must replace the 'bare' Green's function $G_Q^{(s)}(z,z')$ in Eq. (C2) with its 'bulk-screend' expression, Eq. (B21). This assumes that the various components of the electron gas in the substrate (bulk and quantized electrons) screen independently, so that the self-consistency of screening is lost and we are reduced to a linear screening approximation. Obviously, double-counting the screening effects of the channel must be avoided. Therefore, in Eq. (C3) screening is limited to those electrons actually treated as 2D carriers in the simulation, while the bulk dielectric function of the substrate accounts only for those electrons represented as bulk particles at the selected location. Their distribution is assumed to be characterized by an effective temperature consistent with their average kinetic energy, a density and Fermi level obtained from integrating the RPA expression only above the cut-off energy separating the bulk from the 2D transport models, so to obtain the density of bulk electrons in the channel at the selected location.

Accounting now for the possibility of scattering between pairs of electrons in the triplet and single spin-states, we must replace the squared amplitude of the matrix element (C3) with its symmetrized expression:

$$\left| \frac{\tilde{\mathcal{G}}^{(s)}_{Q,\omega;\mu\mu',\nu\nu'}}{\tilde{\epsilon}_s(Q,\omega)} \right|^2 \to |V_{Q,\omega;\mu\mu',\nu\nu'}|^2 = \\ \left| \frac{\tilde{\mathcal{G}}^{(s)}_{Q_d,\omega_d;\mu\mu',\nu\nu'}}{\tilde{\epsilon}_s(Q_d,\omega_d)} \right|^2 + \left| \frac{\tilde{\mathcal{G}}^{(s)}_{Q_x,\omega_x;\mu\nu',\nu\mu'}}{\tilde{\epsilon}_s(Q_x,\omega_x)} \right|^2 - \\ \frac{\tilde{\mathcal{G}}^{(s)}_{Q_d,\omega_d;\mu\mu',\nu\nu'}\ \tilde{\mathcal{G}}^{(s)}_{Q_x,\omega_x;\mu\nu',\nu\mu'}}{\tilde{\epsilon}_s(Q_d,\omega_d)\tilde{\epsilon}_s(Q_x,\omega_x)}\ , \qquad \text{(C4)}$$

where $Q_d = Q = |\mathbf{K} - \mathbf{K}'| = |\mathbf{P}' - \mathbf{P}|$, $Q_x = |\mathbf{K} - \mathbf{P}'| = |\mathbf{K} - \mathbf{P} - \mathbf{Q}|$, $\hbar\omega_d = E_\mu(\mathbf{K}) - E_{\mu'}(\mathbf{K}')$, and $\hbar\omega_x = E_\mu(\mathbf{K}) - E_{\nu'}(\mathbf{P}')$ are the 'direct' and 'exchange' momentum and enegy transfers, respectively. For intervalley processes, the appropriate valley separation $\Delta\mathbf{K}$ must be added to the transfer wave vectors and the Bloch overlap factors at the valley bottoms – evaluated using the $\mathbf{k}\cdot\mathbf{p}$ approximation as in Appendix B of Ref. 3 – must multiply each term of Eq. (C4). Finally, for a given 'primary' electron in the channel, we evaluate the scattering rate by selecting one 'partner' electron stochastically from those – of concentration n_s – which are within a screening length from the primary electron. Using the Born approximation we have:

$$\frac{1}{\tau^{(2D-2D)}(\mathbf{K}\mu, \mathbf{P}\nu)} = \frac{e^4 n_s}{4\pi\hbar} \sum_{\mu'\nu'} \int_0^\pi d\phi\ Q_0(\phi) \times |V_{Q_0(\phi),\omega(\phi);\mu\mu',\nu\nu'}|^2 \left|\frac{dE_{tot}}{dQ}\right|^{-1}_{Q=Q_0(\phi)} , \quad \text{(C5)}$$

where $Q_0(\phi)$ is a solution of $E_{tot}(\mathbf{K}, \mathbf{P}, \phi, Q_0) = 0$, $E_{tot} = E_\mu(\mathbf{K}) + E_\nu(\mathbf{P}) - E_{\mu'}(\mathbf{K}') - E_{\nu'}(\mathbf{P}')$ being the total energy-transfer, expressed as a function of wave vector transfer $\mathbf{Q}$ in polar coordinates (*i.e.*, of the magnitude Q of the transferred wave vector and of the angle ϕ between $\mathbf{Q}$ and the x-axis). Finally, $\hbar\omega(\phi)$ is the energy transfer expressed as a function of $Q_0(\phi)$ and ϕ. Using ellipsoidal nonparabolic dispersions for the subband structure, so that each subband μ is described by in-plane masses $m_{\mu x}$ and $m_{\mu y}$ along the x and y axes, respectively, and by a nonparabolicity parameter α (= - 0.5 eV^{-1}), after rather laborious algebra we find for E_{tot} the following expression of fourth degree in Q:

$$\begin{aligned} E_{tot}(\mathbf{K}, \mathbf{P}, \phi, Q) &= \Delta E_{\mu\mu'\nu\nu'}(\mathbf{K}, \mathbf{P}) \\ &+ 2Q\,[N_{\mu'}(\mathbf{K}) H_{\mu'}(\mathbf{K}, \phi) - N_{\nu'}(\mathbf{P}) H_{\nu'}(\mathbf{P}, \phi)] \\ &- Q^2\{N_{\mu'}(\mathbf{K}) L_{\mu'}(\phi) + N_{\nu'}(\mathbf{K}) L_{\nu'}(\phi) + \\ &\qquad 4\alpha[H^2_{\mu'}(\mathbf{K}, \phi) + H^2_{\nu'}(\mathbf{P}, \phi)]\} \\ &+4\,\alpha\,Q^3[H_{\mu'}(\mathbf{K}, \phi) L_{\mu'}(\phi) - H_{\nu'}(\mathbf{P}, \phi) L_{\nu'}(\phi)] \\ &-\alpha\,Q^4[L^2_{\mu'}(\phi) + L^2_{\nu'}(\phi)] , \end{aligned} \quad \text{(C6)}$$

where:

$$\Delta E_{\mu\mu'\nu\nu'}(\mathbf{K}, \mathbf{P}) = E_\mu(\mathbf{K}) + E_\nu(\mathbf{P}) - E_{\mu'}(\mathbf{K}) + E_{\nu'}(\mathbf{P}) , \quad \text{(C7)}$$

$$H_\lambda(\mathbf{K}, \phi) = \frac{\hbar^2}{2} \left(\frac{K_x}{m_{\lambda x}} \cos\phi + \frac{K_y}{m_{\lambda y}} \sin\phi \right) , \quad \text{(C8)}$$

$$L_\lambda(\phi) = \frac{\hbar^2}{2} \left(\frac{\cos^2\phi}{m_{\lambda x}} + \frac{\sin^2\phi}{m_{\lambda y}} \right) , \quad \text{(C9)}$$

$$N_\lambda(\mathbf{K}) = 1 + 2\alpha\gamma_\lambda(\mathbf{K}) + 2\alpha < E^{(0)}_\lambda - V >_\lambda , \quad \text{(C10)}$$

where $< ... >_\mu$ denotes the expectation value in subband μ, $\gamma_\lambda(\mathbf{K}) = (\hbar^2/2)[K_x^2/m_{\lambda x} + K_y^2/m_{\lambda y}]$ is the parabolic dispersion, and $V(z)$ is the confining potential in the channel. Standard methods[65] can be used to solve the quartic algebraic equation $E_{tot} = 0$, retaining only positive and real solutions $Q_0(\phi)$. Finally, the high-temperature limit of Fetter and Walecka, as described in Ref. 43 (Eq. (73)) is employed to approximate the intra- and inter-subband two-dimensional screening parameter $\beta_{\lambda\lambda'}(Q, \omega)$.

ii) Interaction between 2D-electrons in the channel and 3D electrons in the gate. The Coulomb matrix element between a two-dimensional electron in the channel with 2D wave vector $\mathbf{K}$ in subband μ and a bulk electron of wave vector $\mathbf{p} = (\mathbf{P}, p_z)$ in the gate can be calculated in a similar way, obtaining:

$$< \mathbf{p}'; \mathbf{K}', \mu' | V | \mathbf{p}; \mathbf{K}, \mu > =$$

$$\delta^{(2)}(\mathbf{K} + \mathbf{P} - \mathbf{K}' - \mathbf{P}')\ \frac{2e^2\epsilon_i(\omega)}{Q\ \tilde{\mathcal{C}}_Q\ (Q - iq_z)}\ \tilde{\mathcal{F}}_{\mu\mu'}(Q, \omega) , \quad \text{(C11)}$$

where $\mathbf{Q} = \mathbf{K} - \mathbf{K}' = \mathbf{P}' - \mathbf{P}$ is the momentum-transfer on the plane of the interface, and $q_z = p'_z - p_z$ is the change in the normal component of the wave vector of the gate electron. This quantity is not conserved, because of both the 'fuzzy' nature of the normal-momentum of the 2D electron and the presence of the interface which breaks the symmetry under translations along the z-axis. Also, $\hbar\omega = E_\mu(\mathbf{K}) - E_{\mu'}(\mathbf{K}') = E(\mathbf{p}') - E(\mathbf{p})$ is the energy transfer, and $\tilde{\mathcal{F}}_{\mu\mu'}(Q, \omega)$ is the screened form-factor solution of the linear problem (using once more the notation and results of Appendix A of Ref. 6):

$$\tilde{\mathcal{F}}_{\mu\mu'}(Q, \omega) = \sum_{\lambda\lambda'} [\mathbf{1} - \mathbf{\Pi}(Q, \omega)]^{-1}_{\mu\mu',\lambda\lambda'}\ \mathcal{F}_{\lambda\lambda'}(Q) \quad \text{(C12)}$$

where the unscreened form factor is

$$\mathcal{F}_{\lambda\lambda'}(Q) = \int_t^\infty dz\ \zeta_\mu(z)\ e^{-Q(z-t)}\ \zeta_{\mu'}(z) . \quad \text{(C13)}$$

Thus, the scattering rate, assuming the gate electron is chosen stochastically from a population of gate electrons at density n_g, will be:

$$\frac{1}{\tau^{(2D-3Dg)}(\mathbf{K}\mu, \mathbf{p})} = \frac{e^4 n_g \epsilon_i(\omega)^2}{\hbar\pi^2} \times \sum_{\mu'} \int d\mathbf{p}'\ \frac{|\tilde{\mathcal{F}}_{\mu\mu'}(Q, \omega)|^2}{Q^2\ \tilde{\mathcal{C}}_Q^2\ Q_z^2}\ \delta(E_{tot}) , \quad \text{(C14)}$$

where $E_{tot} = E_\mu(\mathbf{K}) + E(\mathbf{p}) - E_{\mu'}(\mathbf{K}') + E(\mathbf{p}')$. Note that the matrix element is not (anti)symmetric, since we assume that electrons in the gate (substrate) remain in the gate (substrate), so no exchange can take place as they are viewed as distinguishable. For very thin insulators where tunneling becomes significant this approximation may break down. Note also that all components of the electron gas (2DEG, bulk substrate and bulk gate electrons) are active in screening the interaction: Screening by the 2DEG is accounted for via the 2DEG polarizability, $\Pi_{\lambda\lambda';\mu\mu'}(Q, \omega) = \mathcal{G}^{(s)}_{Q;\lambda\lambda';\mu\mu'}\ \beta_{\mu\mu'}(Q, \omega)$ in Eq. (C12). Screening by gate and substrate bulk electrons is implicit in the definition of $\tilde{\mathcal{C}}_Q$ arising from the screened Green's function. Here, the same comments given above after Eq. (C3) apply concerning the approximations used to handle 'mixed' bulk and 2D screening in the substrate. Finally, for large insulator thickness, t, the factor $\tilde{\mathcal{C}}_Q^{-1}$ decreases as e^{-Qt}. In the unscreened case, integrating over

Q gives the expected $1/t$ dependence of the interaction. In the screened case, the other factors in the matrix element vanish quickly for small (as $(Q/t)^2$) and large (as $1/Q^2$) values of Q, so that only Q-vectors in a limited range contribute and the interactions decreases exponentially at large t. In the range of electron concentration in the substrate of technological interest, screening by substrate electrons strongly weakens the strength of the short-range Coulomb interaction between gate and substrate. This should be compared to the long-range plasmon effects which, being screening themselves, grow in strength at larger electron concentrations.

The evaluation of the rate (C14) must be performed numerically with a trivial modification of the algorithm described in Ref. 5, since the dispersion $E(\mathbf{p})$ of bulk (gate) electrons is treated using the empirical psudopotential band structure.

iii) Interaction between 3D-electrons in the substrate and 3D electrons in the gate. The analysis of this process requires some care and approximations. We consider here a bulk electron in the substrate (with wave vector $\mathbf{k}$) scattering with the potential induced in the substrate by an electron of wave vector $\mathbf{p}$ in the gate (or, equivalently, the case of the gate electron scattering with the potential induced in the gate by the substrate electron). It would be tempting the compute the matrix element of this process by describing both electrons with plane waves $e^{i\mathbf{k}\cdot\mathbf{r}}$ (with $\mathbf{r}=(\mathbf{R},z)$) and $e^{i\mathbf{p}\cdot\mathbf{r}'}$ (with $\mathbf{r}'=(\mathbf{R}',z')$) nonvanishing only in the half-spaces $z>t$ and $z'<0$, respectively. However, there are difficulties with this view: In first place, the cross section for this interaction is not properly defined whenever k_z and p_z are nonzero: Since the problem is not invariant under translations along the z-axis, the z-component of the momentum is not conserved. Roughly speaking, this is due to the reflections suffered by electrons hitting the interfaces. Moreover, in our self-consistent Monte Carlo/Poisson simulations, relaxation of the z-component of the momentum can occur via two processes: Via reflections (specular and/or diffusive) at the interfaces and via dynamic fluctuations of the potential. Indeed, the mesh spacing along the z-direction is chosen to be much smaller than the screening length in order to resolve the potential itself with sufficient numerical accuracy. Thus, the Poisson/particle coupling already accounts (albeit semiclassically) for changes of the normal component of the electron velocity due to the Coulomb interaction across the insulator. In addition, while describing a gate electron by a plane wave is consistent with our assumption of a homogenous situation in the depletion layer of the gate, a similar assumption for the electron in the substrate is clearly at odds with the highly inhomogeneous situation in the substrate. Finally, bulk electrons are pictured as 'semiclassical' particles. In the absence of interfaces, electron-electron scattering is viewed as a homogeneous scattering problem, electrons being represented by plane waves traveling from infinitely far, interacting, and continuing along deflected path to infinity. Deviations from homogeneity, as always in the semiclassical Boltzmann picture, are assumed to be significant only over length scales much larger than the electron wavelengths, so that the homogeneous picture is adapted to this 'weakly' inhomogeneous case. In the presence of interfaces, clearly we wish to maintain a similar semiclassical view and keep track of the distance of the substrate electron from the gate, inhomogeneities along this direction being very strong over short distances.

Accordingly, we idealize bulk electrons in the gate as plane waves, as in *ii)* above, but a bulk electron in the substrate at a distance $d-t$ from the substrate/insulator interface is idealized by a wavefunction of the form $\delta(z-d)e^{i\mathbf{K}\cdot\mathbf{R}}/(2\pi)^2$. The matrix element for this process can be calculated trivially to be

$$< \mathbf{p}';\mathbf{K}'|V_d|\mathbf{p};\mathbf{K} > =$$
$$\delta^{(2)}(\mathbf{K}+\mathbf{P}-\mathbf{K}'-\mathbf{P}')\ \frac{2e^2\epsilon_i(\omega)}{Q\ \tilde{\mathcal{C}}_Q\ (Q-iq_z)}\ e^{-Q(d-t)}\ , \quad \text{(C15)}$$

where, smilarly to Eq. (C11), $\mathbf{Q}=\mathbf{K}-\mathbf{K}'=\mathbf{P}'-\mathbf{P}$ is the momentum-transfer on the plane of the interface, $q_z=p_z'-p_z$ is the change in the normal component of the wave vector of the gate electron (which is not conserved, because of lack of translational invariance along the z-axis), $\hbar\omega=E(\mathbf{K},k_z)-E(\mathbf{K}',k_z)=E(\mathbf{p}')-E(\mathbf{p})$ is the energy transfer. The scattering rate in the Born approximation will be:

$$\frac{1}{\tau^{(3Ds-3Dg)}(\mathbf{k},\mathbf{p})}$$
$$= \frac{e^4 n_g \epsilon_i(\omega)^2}{\hbar\ \pi^2}\int d\mathbf{p}'\ \frac{e^{-2Q(d-t)}}{Q^2\ \tilde{\mathcal{C}}_Q^2\ Q_z^2}\ \delta(E_{tot})\ , \qquad \text{(C16)}$$

where $E_{tot}=E(\mathbf{K},k_z)+E(\mathbf{p})-E(\mathbf{K}',k_z)+E(\mathbf{p}')$. It is interesting to note that in the limit $t\to 0$ for $q_z=0$ the squared matrix element takes the form

$$\delta^{(2)}(\mathbf{K}+\mathbf{P}-\mathbf{K}'-\mathbf{P}')\ \frac{e^4\ e^{-2Qd}}{2[\epsilon_g(Q,\omega)+\epsilon_s(Q,\omega)]^2Q^4}\ . \quad \text{(C17)}$$

In addition to the factor e^{-2Qd} influenced by the distance of the substrate-electron from the gate, the squared matrix element is approximately (for $\epsilon_g=\epsilon_s$) a factor 4 smaller than its bulk counterpart, since we are not letting the gate (substrate) electron 'spill over' into the substrate (gate), thus reducing the strength of the Coulomb energy. Also note in Eq. (C17) the effect of both gate- and substrate-screening. As in Eq. (C14), at the large electron concentrations in the substrate of practical interest the strength of the short-range electron-electron scattering across the insulator is expected to be dominated much weaker than the long-range (plasmon) interaction.

Once more, the rate (C16) must be evaluated numerically using the tabulated full-band dispersion $E(\mathbf{k})$. Finally, screening is treated as a fully bulk problem, employing the screened factor $\tilde{\mathcal{C}}_Q$, assuming that all electrons in the substrate are bulk particles.

[1] R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. Leblanc, IEEE J. Solid State Circuits **SC-9**, 256 (1974).

[2] H. S. Momose, E. Morifuji, T. Yashitomi, T. Ohguro, M. Saito, T. Morimoto, Y. Katsumata, and H. Iwai, Int. Electron Devices Meet., Tech. Dig. IEEE, 105 (1996).
[3] M. V. Fischetti and S. E. Laux, Int. Electron Devices Meet., Tech. Dig. IEEE, 305 (1995).
[4] P. A. Childs and C. C. C. Leung, Electron. Lett. **31**, 139 (1995); J. Appl. Phys. **79**, 222 (1996).
[5] M. V.Fischetti, S. E. Laux, and E. Crabbé, J. Appl. Phys. **78**, 1058 (1995).
[6] M. V. Fischetti, *'Long-range Coulomb interactions in small Si devices. Part II: Effective electron mobility in thin-oxide structures'*, submitted to J. Appl. Phys. (2000).
[7] These results have been already pre-announced in M. V. Fischetti and S. E. Laux, Appl. Phys. Lett. **76**, 2277 (2000).
[8] See, for instance, M. V. Fischetti and S. E. Laux, Phys. Rev. B **38**, 9721 (1988). This references discusses the problem of double-counting long-range and short-range Coulomb interactions and of the plasma oscillations. More numerical details are given in S E. Laux and M. V. Fischetti, in *Monte Carlo Device Simulation: Full Band and Beyond*, Karl Hess ed. (Kluwer, Boston, Massachusetts, 1991).
[9] C. Kittel, *Quantum Theory of Solids* (Wiley, New York, 1963), Chaps 5 and 6.
[10] G. D. Mahan, J. Appl. Phys. **51**, 2634 (1980)
[11] C. A. Sholl, Proc. Phys. Soc. (London) **92**, 434 (1067).
[12] H. P. D. Lnyon and R. A. Tuft, IEEE Trans. Electron-Device **ED-26**, 1014 (1979).
[13] H. Goldstein, *Classical Mechanics*, 2nd edition, (Addison-Wesley, Reading, Massachusetts, 1980), Chp. 3
[14] R. Kubo, *Statistical Mehanics* (North-Holland, New York, 1964).
[15] C. Haas, Phys. Rev. **125**, 1965 (1962).
[16] G. D. Mahan, *Many Particle Physics* (Plenum, New York, 1991).
[17] M. Gell-Mann and K. Brueckner, Phys. Rev. **106**, 364 (1957).
[18] P. Nozières and D. Pines, Phys. Rev. **111**, 442 (1958).
[19] G. Ortiz and P. Ballone, Phys. Rev. B **50**, 1391 (1994).
[20] See, for example, D. Bohm and D. Pines, Phys. Rev. **92**, 609 (1953).
[21] H. Brooks and C. Herring, Phys. Rev. **83**, 879 (1951).
[22] E. M. Conwell and V. F. Weisskopf, Phys. Rev. **77**, 388 (1950).
[23] Alternative expressions are given in the Monte Carlo literature for the CW scattering rate. For example, C. Jacoboni and L. Reggiani, Rev. Mod. Phys. **55**, 645 (983), actually use the bare Coulomb potential of the impurity extending it to infinity, but introduce a cut-off at r_0 for the maximum impact parameter, as we have done in Eqns. (32) and (33).
[24] J. A. del Alamo, S. Swirhun, and R. M. Swanson, Solid-State Electron. **28**, 47 (1985).
[25] R. J. Van Overstraeten and R. .P. Mertens, Solid-State Electron. **30**, 1077 (1987).
[26] J. Wagner and J. A. del Alamo, J. Appl. Phys. **63**, 425 (1988).
[27] One may wonder whether the finite size of the simulated system has anything to do with these fluctuations. The energy fluctuations due to the finite number of particles N_p used in the simulation of the system are of a Gaussian form $\approx \exp(-\Delta E^2/2\sigma^2)$, where $\sigma^2 \approx C_V k_B T^2$ and C_V is the heat capacity of the electron gas (see D. Zubarev, V. Morozov, and G. Röpke, *Statistical Mechanics of Nonequilibrium Processes, Volume 1: Basic Concepts, Kinetic Theory* (Akademie Verlag, Berlin, 1996)). Thus, $\sigma \approx k_B T[\pi^2 T/(2N_p T_F)]^{1/2}$, where T_F is the Fermi temperature $2K_F/3k_B$. In our case $N_p \approx 10^4$, large enough to render σ two orders of magnitude smaller than $k_B T$, and so negligible.
[28] L. D. Landau and E. M. Lifshitz, *Statistical Physics* (Addison-Wesley, Reading, Massachusetts, 1974), Chp. 3.
[29] L. D. Landau, Soviet Physics JEPT **3**, 920 (1957).
[30] We have performed simulations in which degeneracy was included also in the integration of the equations of motion during free flights. This was handled heuristically by modifying the force entering the discretized equation of motion $\Delta\mathbf{k} = -e\mathbf{F}[\mathbf{r}(t)]\Delta t/\hbar$ (where $\Delta\mathbf{k}$ is the increment of the electron wave vector over the integration time step Δt, and $\mathbf{F}[\mathbf{r}(t)]$ is the electric field at position $\mathbf{r}(t)$ at time t) as $\mathbf{F}[\mathbf{r}(t)] \to \mathbf{F}[\mathbf{r}(t)][1 - f_{FD}(\mathbf{k}+\Delta\mathbf{k})]$. Thus, particles feel the action of the driving force only when their 'final state' after a free flight leads them into an unoccupied **k**-state. Lacking any formal justification for this procedure, we simply report that even at low temperature (T=4 K) this correction makes a small difference in the overall result.
[31] D. Pines, *The Many-Body Problem* (Benjamin, Reading, MA, 1961), Fig. 1-1(b).
[32] B. Farid, V. Heine, G. E. Engel, and I. J. Robertson, Phys. Rev. B **48**, 11602 (1993).
[33] G. Senatore, S. Moroni, and D. M. Ceperley, in *Quantum Monte Carlo Methods in Physics and Chemistry*, M. P. Nightingale and C. Umrigar eds. (Kluwer, Boston, Massachusetts, 1999), p. 183.
[34] E. O. Kane, Solid-State Electron. **28**, 3 (1985).
[35] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals series and products* (Academic Press, New York, 1965), p. 1064.
[36] K. Hess, B. Tuttle, F. Register, and D. K. Ferry, Appl. Phys. Lett. **75**, 3147 (1999).
[37] A. L. Fetter and J. D. Walecka, *Quantum Theory of Many-Particle Systems* (McGraw-Hill, New York, 1971), p. 305.
[38] A. Lacaita, Appl. Phys. Lett. **59**, 1623 (1991).
[39] C. C. C. Leung and P. A. Child, Appl. Phys. Lett. **66**, 162 (1995).
[40] A. Abramo, C. Fiegna, and F. Venturi, IEDM Tech. Dig., 301 (1995).
[41] A. Pacelli and A. L. Lacaita, Semicond. Sci. Technol. **11**, 1642 (1996).
[42] M. Hargrove, S. Crowder, E. Nowak, R. Logan, L. K. Han, H. Ng, A. Ray, D. Sinitsky, P. Smeys, F. Guarin, J. Oberschmidt, E. Crabbé, D. Yee, and L. Su, Int. Electron Devices Meet., Tech. Dig. IEEE, 627 (1998).
[43] M. V. Fischetti and S. E. Laux, Phys. Rev. B **48**, 2244 (1993).
[44] M. V. Fischetti and S. E. Laux, J. Appl. Phys. **80**, 2234 (1996).
[45] `http://www.research.ibm.com/DAMOCLES/html_files /mueff.html`
[46] C. Canali, C. Jacoboni, F. Nava, G. Ottaviani, and A. Alberigi-Quaranta, Phys. Rev. B **12**, 2265 (1975).
[47] S.-C. Lee and I. Galbraith, Phys. Rev. B **55**, R16025 (1997).
[48] S.-C. Lee and I. Galbraith, Phys. Rev. B **59**, 15796 (1999).
[49] S. Datta, F. Assad, and M. S. Lundstrom, Superlattices and Microstructures, **23**, 771 (1998).
[50] S. E. Laux and M. V. Fischetti, IEEE Electron Device Lett. EDL-9, 467 (1987).
[51] A large momentum-transfer rate between electron layers

across a dielectric has been previously obtained by C. Jacoboni and P. J. Price, Solid-State Electron. **31**, 649 (1988).

[52] T. Mizuno and R. Ohba, Int. Electron Devices Meet., Tech. Dig. IEEE, 109 (1996).

[53] G. A. Sai-Halasz, M. R. Wordeman, D. P. Kern, E. Ganin, S. Rishton, D. S. Zickerman, H. Schmid, M. R. Polcari, H. Y. Ng, P. J. Restle, T. H. P. Chang, and R. H. Dennard, IEEE Electron Device Lett. **EDL-9**, 463 (1987).

[54] L. Machanda, R. H. Storz, R. H. Yau, K. F. Lee, and E. H. Westerwick, IEDM Tech. Dig., 994 (1992).

[55] S. Tam, F.-C. Hsu, C. Hu, R. S. Muller, and P. K. Ho, Electr. Dev. Lett. **ED-4**, 249 (1983).

[56] M. V. Fischetti, N. Sano, S. E. Laux, and K. Natori, IEEE J. Technol. Comp. Aided Design, `http://www.ieee.org/products/online/journal/tcad/accepted/fischetti-feb97` (1997).

[57] J. D. Bude, Sym. on VLSI Tech. Tech. Dig., 101 (1995).

[58] Another reasons why the impact-ionization feedback process plays a decreasing role in smaller devices lies in the fact that scaled-down devices must employ shallower source and drain junctions. Holes generated by impact ionization initiated by electrons in the channel will in turn ionize deeper in the substrate. In small devices, these secondary ionization process will occur mainly in the low-field (or even retrogade field, for retrogade doping) beyond the drain/body junction. Many of the generated (feedback) electrons will be more likely to drift within the substrate, rather than be swept back towards the Si-SiO_2 interface.

[59] J. D. Bude, T. Iizuka, and Y. Kamakura, IEDM Tech. Dig., 865 (1996).

[60] J. Stathis and D. J. DiMaria, IEDM Tech. Dig., 167 (1998)

[61] K. L. Ngai and E. N. Economou, Phys. Rev. B **4**, 2132 (1971).

[62] E. A. Stern and R. A. Ferrel, Phys. Rev. **120**, 130 (1960).

[63] R. A. Ferrel, Phys. Rev. **111**, 1214 (1958).

[64] That this is indeed the correct limit for the screened Green's function is most easily seen in the simpler bulk case: The bulk screened Green's function in spherical (momentum) coordinates is $\tilde{G}_q = -1/[q^2\epsilon(q,\omega)]$. In the Thomas-Fermi limit $\epsilon(q,\omega) \to \epsilon^\infty(1+\beta_s^2/q^2)$ and the Green's function $\tilde{G}_Q(z,z')$ in cylindrical coordinates becomes $-\exp(-K|z-z'|)/(2\epsilon^\infty K)$, where $K=(Q^2+\beta_s^2)^{1/2}$. In the short-wavelength limit ($Q >> \beta_s$) this expression tends to $-\exp(-Q|z-z'|)/[2Q\epsilon(\sqrt{2}Q)]$, which is identical to the average considered here and it is appropriate at the short-distances we are interested in when dealing with short-range Coulomb interactions. From a physical point of view, Eq. (B20) states that the electron gas responds equally to a wave $\exp(iQz)$ and to a decaying potential $\exp(-Q|z|)$. Indeed, formally, Eq. (B18) implies $\tilde{\epsilon_j}(Q,\omega) = \epsilon_j(\mathbf{Q}, q_z = iQ, \omega)$. This observation makes it tempting to assign Eq. (B20) a more general validity by defining the dielectric function on the complex plane $q \to z$ as $\epsilon_j(|z|,\omega)$. Unfortunately, such an assignment is arbitrary, since it does not constitute an analytic continuation.

[65] *Handbook of tables for Mathematics* (4-th Ed., CRC Press, Cleveland, Ohio, 1975), p. 132.