RC 21796 (7/18/00)
Physical Sciences

IBM Research Report

Long-range Coulomb interactions in small Si devices. PartII: Effective electron mobility in thin-oxide structures

M. V. Fischetti

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

Long-range Coulomb interactions in small Si devices. Part II: Effective electron mobility in thin-oxide structures

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

In metal-oxide-semiconductor structures with poly-crystalline Si gates, electrons in the inverted channel of the substrate scatter with electrons in the gate via long-range Coulomb interactions. For thin oxides, these interactions can cause a significant transfer of momentum from the channel to the gate, thus reducing the effective mobility of the two-dimensional electron gas in the substrate. We present calculations of the dispersion of the interface plasmons in poly-Si/SiO_2/Si structures, comparing the results obtained in the long-wavelength limit to those obtained using the random-phase approximation. Employing the former model, we compute the effect of plasmon scattering on the effective electron mobility in Si inversion layers. We find a significant reduction of the mobility for oxides thinner than about 3 nm.

I. INTRODUCTION

Electron-electron Coulomb interactions usually have a small effect on the electron mobility. They do not contribute directly to any momentum loss, but simply alter the carrier distribution function, thus triggering, indirectly, more collisions with momentum-dissipating scatterers, mainly phonons. However, in metal-oxide-semiconductor (MOS) structures, or, more generally, in 'layered' structures in which electrons are separated into different layers (or 'channels'), net momentum transfer can occur via Coulomb interactions among electrons in different layers resulting in a direct effect on the electron mobility in the layers. Indirect effects caused by long-range Coulomb interactions between electrons in the channel and those in the source and drain of small devices have been considered elsewhere[1], together with a semiclassical study of these channel-gate interactions and their effect on electron transport under high bias. Here we are interested in the mobility of electrons in the inversion layer of MOS structures: The gate material is typically another (poly-crystalline) Si layer, the thickness of the insulating layer (typically SiO_2) is at present as small as 1.5 nm (Ref. 2), so that electrons in the channel of the Si substrate can interact very strongly with the electrons in the depletion layer of the gate. Since the electrons in the gate are almost 'at rest', while the channel electrons drift from the source to the drain under the action of the applied bias, we can expect a 'gate drag', electrons in the channel losing momentum to those in the gate. The purpose of this paper is to show that this is indeed the case in technologically relevant structures, and that the reduction of the thickness of the gate insulator in small Si n-channel MOS field-effect transistors (n-MOSFETs) does indeed cause a reduction of the electron mobility in the channel.[3]

Originally considered by Pogrebinskii[4] and Price[5], the mutual drag between a two-dimensional electron gas (2DEG) and a 3DEG in compound-semiconductor systems has been investigated theoretically early on by Boiko and Sirenko[6] and by Leikhtman and Solomon[7], at electron densities sufficiently low to justify the use of static Coulomb interactions. The case of two interacting 2DEGs has been considered by Gramila and co-workers[8], by Zheng and MacDonald[9], who have employed the zero-temperature approximation for the dielectric response, by Jauho and Smith[10], who have also treated Coulomb effects in the static-screening approximation, and by Flensberg *et al.*[11] using many-body techniques. The review paper by Rojo[12] provides a more comprehensive account of work done on this subject. Jacoboni and Price[13] have employed coupled Monte-Carlo/molecular dynamics methods to investigate the inter-layer momentum transfer caused by short-range interactions between 3D electrons in two adjacent Si channels, finding that significant energy transfer-rates are possible. What differentiates these studies from the present one is not so much the different system under investigation (the interaction between a 2DEG – the inverted Si substrate – and a 3DEG – the depleted Si gate), since a coupled 2D/3D system has been considered before[6,7,14]. Rather, it is the different final result we seek (the calculation of the electron mobility in real devices, rather than the temperature depedence of the momentum-transfer rates or the relations between Coulomb interactions, disorder and localization[15]); and, most notably, the fact that none of the previous studies has accounted for the long-range, plasmon contribution to the drag, a notable exception being the work by Flensberg and Hu[16]. Indeed, Coulomb interactions come in two categories: Short-range, single particle effects for wavelengths shorter than a cutoff value $\sim 1/q_c$, and long-range, collective phenomena associated with plasmons.[17] At large carrier densities short-range processes weaken because of dielectric screening. On the contrary, long-range effects, being 'screening' themselves, become dominant, as also discussed by Pines[18]. Therefore, we should expect that even stronger effects may be caused by the long range process we consider here. Indeed, this is what we find, as shown in Sec. III E below. Our results are qualitatively similar to those of Ref. 16, but are relative to a different system and have been obtained using a different – albeit ultimately equivalent – physical scheme: At the large electron densities and temperature of interest, plasmon effects dominate over short-range inter-electronic Coulomb interactions.[19] Neglecting the latter will – hopefully – result in a clearer physical picture.

In MOS geometries, long-range Coulomb effects are characterized by quantized interface modes, called 'interface plasmons'. Thus, in this paper our goal is to characterize the interface plasmons in realistic geometries and evaluate the effect of scattering of channel electrons by gate/insulator modes on the mobility of the electrons themselves. In addition to pure interface plasmon modes, long-range electronic oscillations can couple to the long-range field of optical phonons present in the polar insulator. Scattering with the phonon component of these modes has been considered before[20], but its effect on the mobility appears to be small. Therefore, we consider only scattering with pure plasmon modes.

The paper is organized as follows: In Sec. II we calculate the dispersion of the interface plasmons (and also coupled plasmon/optical-oxide-phonon modes) for a general depleted-semiconductor/insulator/inverted-semiconductor structure. We compare the results obtained using the (numerically quite cumbersome) random-phase approximation (RPA) for the dielectric functions of the semiconductors to those obtained in the simpler long-wavelength limit in order to assess the limitations of the latter. In Sec. III we calculate the mobility of electrons in the inversion layer in the substrate accounting for scattering with the interface modes in addition to phonon and surface roughness. Finally, in Sec. IV we discuss the relevance of our results to the performance of small MOSFETs. One appendix deals briefly with the dielectric properties of a 2DEG.

II. INTERFACE PLASMONS IN MOS STRUCTURES

Interface plasma oscillations in MOS structures have been considered in the past so extensively as to render impossible a full review of the subject[21]. Here we follow mainly the work by Economou[22] and Ngai and Economou[23], extending their analysis to account for the dielectric response of the 2DEG, as studied by Dahl and Sham[24], as done, for example, by Eguiluz and co-workers[25]. In Sec. II A we present the problem, considering in detail the boundary conditions at the gate/insulator and insulator/2DEG interfaces in Secs. II B and II C, respectively. In Sec. II D we deal with the secular equation whose solutions yield the dispersion of the interface modes. In Secs. II E and II F, respectively, we treat the dielectric functions of the semiconductor(s) in the RPA and long-wavelength limits. Finally, in Sec. II G we present results for the dispersion of the plasmon and coupled plasmon/TO-phonon modes.

A. Boundary conditions in the non-retarded limit

The structure we consider here consists of a degenerately-doped n-type semiconductor in the half-space $z < 0$ (gate), an insulating layer for $0 < z < t_{ox}$, and a p-type semiconductor (in principle different from the gate material) filling the half-space $z > t_{ox}$ (substrate). We shall ignore inhomogeneities on the (x, y)-plane: For small devices of gate length L_g, this amounts to ignoring corrections to the dispersion of the plasmons for wave vectors $q < L_g^{-1}$, which usually represents a small region of phase space. We are interested in situations in which a positive bias is applied to the gate, bias sufficiently strong to induce an inversion layer in the substrate, which will be viewed as a 2DEG. We refer to this structure as a metal-oxide-semiconductor (MOS) structure in order to comform to standard nomenclature, although 'semiconductor-insulator-semiconductor' (SIS) would be a more proper label. Specifically, we are interested in cases in which the gate – consisting of polycrystalline Si (poly-Si, or 'poly' for short) – is idealized as crystalline bulk Si, the insulator is SiO_2, and the substrate contains a Si 2DEG. We denote by $\mathbf{R}$ the two-dimensional vector coordinate in the (x, y)-plane of the interfaces.

We look for transverse-magnetic solutions (TM- or p-waves) of Maxwell's equations in this geometry in the non-retarded limit. Thus, the 'usual' boundary conditions require that the components $E_\parallel(\mathbf{R}, z, t)$ of the electric field $\mathbf{E}(\mathbf{R}, z, t)$ on the plane of the interfaces be continuous across the two interfaces at $z = 0$ and $z = t_{ox}$, and similarly for the component $D_z(\mathbf{R}, z, t)$ of the displacement field $\mathbf{D}(\mathbf{R}, z, t)$ normal to the plane of the interfaces. Using the cylindrical symmetry of the problem we can expand the electrostatic potential at frequency ω as:

$$\phi(\mathbf{R}, z, t) \;=\; \sum_{\mathbf{Q}} \; \phi_{\mathbf{Q},\omega}(z) \; e^{i\mathbf{Q}\cdot\mathbf{R}} \; e^{i\omega t} \;, \qquad (1)$$

where $\mathbf{Q}$ is the two-dimensional wave vector and it must be understood that we shall take the real part of the complex exponentials here and in the following. Retardation effects are significant for $Q \leq \kappa^{1/2}\omega/c$ (where c is the speed of light and κ the dielectric constant of the medium), which, in our case is of the order of 10^4 cm^{-1}. Since we are ultimately interested in evaluating the electron mobility, we have to deal with wave vectors of the order of the Fermi wave vector of the 2DEG, which are of the order of 10^6 cm^{-1} to 10^7 cm^{-1} in the density-range of interest. Thus, we can ignore retardation effects and, employing the procedure and approximations described in Appendix B of Ref. 1, finding solutions of Mawell's equations reduces to finding the solution of the Laplace equation:

$$\frac{d^2\phi_{\mathbf{Q},\omega}(z)}{dz^2} - Q^2\phi_{\mathbf{Q},\omega}(z) \;=\; 0 \;. \qquad (2)$$

Since in the following we assume an isotropic dielectric response everywhere, by symmetry $\phi_{\mathbf{Q},\omega}(z)$ depends only on the magnitude Q of the wave vector. A general physically acceptable (*i.e.*, finite as $z \to \pm\infty$) solution of Eq. (2) is of the form

$$\phi_{Q,\omega}(z) \;=\; \begin{cases} a_{Q,\omega}\; e^{Qz} & (z < 0) \\ b_{Q,\omega}\; e^{-Qz} + c_{Q,\omega}\; e^{Qz} & (0 \leq z < t_{ox}) \\ d_{Q,\omega}\; e^{-Qz} & (z \geq t_{ox}) \end{cases} \;. \qquad (3)$$

The boundary conditions

$$\begin{cases} E_{\parallel,\omega}(\mathbf{R}, z=0^-, t) &= E_{\parallel,\omega}(\mathbf{R}, z=0^+, t) \\ E_{\parallel,\omega}(\mathbf{R}, z=t_{ox}^-, t) &= E_{\parallel,\omega}(\mathbf{R}, z=t_{ox}^+, t) \\ D_{z,\omega}(\mathbf{R}, z=0^-, t) &= D_{z,\omega}(\mathbf{R}, z=0^+, t) \\ D_{z,\omega}(\mathbf{R}, z=t_{ox}^-, t) &= D_{z,\omega}(\mathbf{R}, z=t_{ox}^+, t) \end{cases} , \quad (4)$$

(having replaced $\mathbf{E}_\parallel$ with its magnitude, again thanks to cylindrical symmetry) suitably rewritten in terms of the Fourier-transformed potential $\phi_{Q,\omega}$ determine a homogeneous linear problem in the four unknowns, $a_{Q,\omega}$, $b_{Q,\omega}$, $c_{Q,\omega}$, and $d_{Q,\omega}$. The vanishing of the determinant of the system (4) determines the dispersion $\omega(Q)$ of the interface modes. For a given solution $\omega(Q)$ of the secular equation, three of the unknowns (say, $b_{Q,\omega}$, $c_{Q,\omega}$, and $d_{Q,\omega}$, to fix the ideas) can be expressed in terms of $a_{Q,\omega}$. Fixing this last unknown amounts to normalizing the potential associated with that mode. This will be done as described in Appendix A of Ref. 1

In order to proceed we must be able to reformulate the boundary conditions (4) in terms of the electrostatic potential $\phi_{Q,\omega}$. For the first two of the Eqns. (4) this is trivially accomplished by Fourier transforming the in-plane component of the electric field, as in Eq. (1):

$$E_{\parallel,\omega}(\mathbf{R}, z, t) \;=\; \sum_{\mathbf{Q}} \; E_{\parallel,\mathbf{Q},\omega}(z) \; e^{i\mathbf{Q}\cdot\mathbf{R}} \; e^{i\omega t} \; , \quad (5)$$

with

$$E_{\parallel,\mathbf{Q},\omega}(z) = -i \; Q \; \phi_{\mathbf{Q},\omega}(z) \; . \quad (6)$$

The connection between the displacement field D_z and the potential ϕ is more complicated, since dielectric screening enters at this stage.

B. Gate-insulator boundary

Let us start by considering the gate/insulator boundary. As stated above, in the cases of interest the (poly-Si) gate will be depleted, with the electron density dropping from its charge-neutral value $n = N_{Dg}$ (the doping concentration in the gate) at the edge of the depletion layer, to a smaller value – which is function of the gate bias – at the gate-insulator interface. We idealize this inhomogeneous situation with a homogeneous and isotropic electron gas at a density given by the average electron density, n_g, in the depletion layer. Then, in terms of the Fourier transform $\epsilon_g(\mathbf{Q}, q_z; \omega)$ of the longitudinal dielectric function of the gate we have:

$$D_{z,\omega}(\mathbf{R}, z=0^-, t) \;=\; -\; e^{i\omega t} \; \sum_{\mathbf{Q}} e^{i\mathbf{Q}\cdot\mathbf{R}} \times \; a_{Q,\omega} \; \int_{-\infty}^{+\infty} \frac{dq_z}{\pi} \, \frac{\epsilon_g(\mathbf{Q}, q_z; \omega)}{1 + (q_z/Q)^2} \; . \quad (7)$$

For $z \to 0^+$, ignoring the wavelength-dependence of the dielectric function of the insulator, but retaining its frequency dependence (due to TO modes in the case of SiO_2), we can write:

$$D_{z,\omega}(\mathbf{R}, z=0^+, t) \;=\; e^{i\omega t} \; \sum_{\mathbf{Q}} e^{i\mathbf{Q}\cdot\mathbf{R}} \; \epsilon_i(\omega) \; Q \; [\; b_{Q,\omega} - c_{Q,\omega}] \; . \quad (8)$$

Thus, the boundary condition for the electric displacement field at $z = 0$ (the third of the Eqns.(4)) can be re-written as:

$$Q \; \epsilon_i(\omega) [\; b_{Q\omega} - c_{Q\omega}] \;=\; -\; a_{Q,\omega} \int_{-\infty}^{+\infty} \frac{dq_z}{\pi} \, \frac{\epsilon_g(\mathbf{Q}, q_z; \omega)}{1 + (q_z/Q)^2} \; . \quad (9)$$

As stated by Eq. (B20) of Appendix B of Ref. 1, we shall approximate the right-hand-side of Eq. (9) with $-a_{Q,\omega} \; \epsilon_g(2\sqrt{Q}; \omega)$, where $\epsilon_g(q; \omega)$ is the dielectric function of the gate, which, having been assumed isotropic, depends only on the magnitude q of the wave vector $\mathbf{q}$.

C. Substrate-insulator boundary

Dealing with the insulator-substrate boundary is slightly more complicated. As shown by Dahl and Sham[24], the response of a 2DEG is both nonlocal and anisotropic: nonlocal because a change of potential at z induces a change of wavefunctions at locations $z' \neq z$; anisotropic because when a field $\mathbf{E}_\parallel$ is applied on the plane of the interface, inter-subband transitions induce a redistribution of the inversion charge also along the z-direction, normal to the interface, resulting in a nonvanishing off-diagonal element $\epsilon_{z,\parallel}$ of the dielectric tensor. Here we retain the nonlocality, but since intersubband transitions will be ignored, we consider only the longitudinal (*i.e.*, diagonal) response. Thus, using the results and notation of Appendix A, the total potential resulting from an applied external potential $\phi^{(0)}_{Q,\omega}(z)$ can be written as:

$$\phi_{Q,\omega}(z) \;=\; \phi^{(0)}_{Q,\omega}(z) \;+\; \int_{t_{ox}}^{\infty} dz' \; G^{(0)}_Q(z, z') \times \sum_{\mu\mu'} \; \beta_{\mu\mu'}(Q, \omega) \; \phi_{Q,\omega;\mu\mu'} \; \zeta_\mu(z') \; \zeta_{\mu'}(z') \; . \quad (10)$$

Note that here we need the response of an 'isolated' 2DEG in absence of the interfaces, whose presence is already accounted for when solving of Eq. (2) with the boundary conditions (4). Thus, the Green's function of the Poisson equation appearing in Eq. (10) is the the 'free' Green's function, $G^{(0)}_Q(z, z') = -e^{-Q|z-z'|}/(2Q)$.

When considering only the longitudinal dielectric response

$$D_{z,Q,\omega}(z = t_{ox}^+) \;=\; -\epsilon_s^\infty \frac{d\phi_{Q\omega}(z = t_{ox}^+)}{dz} \; , \quad (11)$$

thus, inserting Eq. (10) into Eq. (11) we obtain:

$$D_{z,Q,\omega}(z = t_{ox}^+) \;=\; -\epsilon_s^\infty \frac{d\phi^{(0)}_{Q\omega}(z = t_{ox}^+)}{dz} - \epsilon_s^\infty \int_{t_{ox}}^{\infty} dz' \; \frac{dG^{(0)}_Q(z = t_{ox}^+, z')}{dz} \times \sum_{\mu\mu'} \; \beta_{\mu\mu'}(Q, \omega) \; \phi_{Q,\omega;\mu\mu'} \; \zeta_\mu(z') \; \zeta_{\mu'}(z') \; . \quad (12)$$

Now noticing that

$$D_{z,Q,\omega}(z = t_{ox}^-) \;=\; \epsilon_i(\omega) \; Q \; [\; b_{Q,\omega} e^{-Qt_{ox}} \;-\; c_{Q,\omega} e^{Qt_{ox}}] \; , \quad (13)$$

we can rewrite the last of Eqns. (4) in the following form:

$$\epsilon_i(\omega)\ Q\ [\ b_{Q,\omega}e^{-Qt_{ox}}\ -\ c_{Q,\omega}e^{Qt_{ox}}]\ =\epsilon_s^\infty d_{Q,\omega}\ Q \times\left\{e^{-Qt_{ox}}\ +\ e^{Qt_{ox}}\ \sum_{\mu\mu'}\ \frac{\beta_{\mu\mu'}(Q,\omega)}{2Q}\ \Phi_{Q,\omega;\mu\mu'}\ \Phi^{(0)}_{Q,\mu\mu'}\right\}\ , \tag{14}$$

where ϵ_s^∞ is the high-frequency dielectric constant of the substrate. In practice this is the background dielectric response of Si due to the valence electrons, $\epsilon_s^\infty = \epsilon_{Si}^\infty \approx 11.7\ \epsilon_0$, where ϵ_0 is the permittivity of the vacuum. (We shall similarly define the high-frequency limits of the permittivities of the gate semiconductor and of the insulator, ϵ_g^∞ and ϵ_i^∞, respectively. We take $\epsilon_g^\infty = \epsilon_{Si}^\infty$ and $\epsilon_i^\infty = \epsilon_{ox}^0 \approx 3.9\ \epsilon_0$ when ignoring the coupling to the SiO_2 transverse optical (TO) phonons, or $\epsilon_i^\infty = \epsilon_{ox}^\infty \approx 2.5\ \epsilon_0$ – see Eq. (31) below – otherwise.) The factor $1/(2Q)$ results from the evaluation of the term $dG_Q^{(0)}(t_{ox}^+, z')/dz$ under the assumption that the wavefunctions do not penetrate into the insulator (*i.e.*, $\zeta_\mu(z) = 0$ for $z \le t_{ox}$).

The unscreened 'normalized' form-factor $\Phi^{(0)}_{Q,\mu\mu'}$ is simply

$$\Phi^{(0)}_{Q,\mu\mu'}\ =\ \int_{t_{ox}}^\infty dz\ e^{-Qz}\zeta_\mu(z)\ \zeta_{\mu'}(z)\ , \tag{15}$$

and the screened form-factors $\Phi_{Q,\omega;\mu\mu'}$ are obtained from the unscreened form-factors after inversion of the dielectric matrix $\mathbf{1}-\mathbf{\Pi}^{(0)}$, where $\mathbf{\Pi}^{(0)}$ is the polarizability given in Appendix A, but with the Green's function $G_Q^{(0)}$ in place of $G_Q^{(s)}$, via Eq. (A11):

$$\Phi_{Q,\omega;\lambda\lambda'}\ =\ \sum_{\mu\mu'}\ [\mathbf{1}-\mathbf{\Pi}^{(0)}(Q,\omega)]^{-1}_{\lambda\lambda';\mu\mu'}\ \Phi^{(0)}_{Q;\mu\mu'}\ . \tag{16}$$

D. Secular equation

We can now summarize our results by rewriting the boundary conditions (4) as follows:

$$\begin{cases} a_{Q,\omega} & = & b_{Q,\omega} + c_{Q,\omega} \\ b_{Q,\omega}\ e^{-Qt_{ox}} + c_{Q,\omega}\ e^{Qt_{ox}} & = & d_{Q,\omega}\ e^{-Qt_{ox}} \\ \tilde\epsilon_g(Q,\omega)\ a_{Q,\omega} & = & \epsilon_i(\omega)\ (c_{Q,\omega} - b_{Q,\omega}) \\ \epsilon_i(\omega)\ [\ b_{Q,\omega}\ e^{-Qt_{ox}} - c_{Q,\omega}\ e^{Qt_{ox}}] & = & \tilde\epsilon_s^{(2D)}(Q,\omega)\ d_{Q,\omega}\ e^{-Qt_{ox}} \end{cases}\ , \tag{17}$$

with the 'effective' dielectric functions

$$\tilde\epsilon_g(Q,\omega)\ =\ \frac{1}{\pi}\ \int_{-\infty}^{+\infty}\ d\left(\frac{q_z}{Q}\right)\ \frac{\epsilon_g(\mathbf{Q},q_z;\omega)}{1+(q_z/Q)^2}\ , \tag{18}$$

and

$$\tilde\epsilon_s^{(2D)}(Q,\omega)\ = \epsilon_s^\infty\left\{1\ +\ e^{2Qt_{ox}}\ \sum_{\mu\mu'}\ \frac{\beta_{\mu\mu'}(Q,\omega)}{2Q}\ \Phi_{Q,\omega;\mu\mu'}\ \Phi^{(0)}_{Q,\mu\mu'}\right\}\ . \tag{19}$$

The dispersion of the interface modes is given by the solution of the secular equation (*i.e.*, from the vanishing of the determinant) associated with the linear homogeneous system of equations (17):

$$e^{Qt_{ox}}[\tilde\epsilon_g(Q,\omega)+\epsilon_i(\omega)][\tilde\epsilon_s^{(2D)}(Q,\omega)+\epsilon_i(\omega)] - e^{-Qt_{ox}}[\tilde\epsilon_g(Q,\omega)-\epsilon_i(\omega)][\tilde\epsilon_s^{(2D)}(Q,\omega)-\epsilon_i(\omega)]\ =\ 0\ . \tag{20}$$

Finally, for a given solution $\omega(Q) = \omega_Q$ of Eq. (20), the system (17) allows us to express the coefficients $b_{Q,\omega}$, $c_{Q,\omega}$, and $d_{Q,\omega}$ appearing in Eq. (3) in terms of $a_{Q,\omega}$:

$$b_{Q,\omega_Q}\ =\ \frac{\epsilon_i(\omega_Q)-\tilde\epsilon_g(Q,\omega_Q)}{2\epsilon_i(\omega_Q)}\ a_{Q,\omega_Q}\ , \tag{21}$$

$$c_{Q,\omega_Q}\ =\ \frac{\epsilon_i(\omega_Q)+\tilde\epsilon_g(Q,\omega_Q)}{2\epsilon_i(\omega_Q)}\ a_{Q,\omega_Q}\ , \tag{22}$$

$$d_{Q,\omega_Q}\ =\ \frac{\epsilon_i(\omega_Q)-\tilde\epsilon_g(Q,\omega_Q)}{\epsilon_i(\omega_Q)+\tilde\epsilon_s^{(2D)}(Q,\omega_Q)}\ a_{Q,\omega_Q}\ . \tag{23}$$

The coefficient a_{Q,ω_Q} can be determined by one of the methods described elsewhere[1]. For example, integrating the electrostatic energy associated with the potential (3), accounting also for the kinetic energy (equal in magnitude), taking the time average of the total energy so obtained, and equating it to the zero-point energy of the interface-plasmons, $(1/2)\hbar\omega_Q$, we get:

$$a_{Q,\omega_Q}\ =\ \left(\frac{\hbar\omega_Q}{2Q\mathcal{D}_Q}\right)^{1/2}\ , \tag{24}$$

where

$$\mathcal{D}_Q\ =\ \epsilon_g^\infty\ +\ \epsilon_i^\infty\left\{\frac{[\epsilon_i(\omega_Q)-\tilde\epsilon_g(Q,\omega_Q)]^2}{4\epsilon_i(\omega_Q)^2}\ (1-e^{-2Qt_{ox}}) + \frac{[\epsilon_i(\omega_Q)+\tilde\epsilon_g(Q,\omega_Q)]^2}{4\epsilon_i(\omega_Q)^2}\ (e^{2Qt_{ox}}-1)\right\} + \epsilon_s^\infty\left[\frac{\epsilon_i(\omega_Q)-\tilde\epsilon_g(Q,\omega_Q)}{\epsilon_i(\omega_Q)+\tilde\epsilon_s^{(2D)}(Q,\omega_Q)}\right]^2\ e^{-2Qt_{ox}}\ . \tag{25}$$

E. Random-phase approximation

In evaluating the effective electron mobility we shall make use of the long-wavelength ($q \to 0$) approximation for the bulk Si dielectric function $\epsilon_g(\mathbf{Q},q_z;\omega)$ and for the two-dimensional screening wave vector $\beta_{\mu\mu'}(Q,\omega)$. However, before taking this limit, it is interesting to estimate the range of validity of the long-wavelength approximation and to establish the range of Q-vectors in which interface plasmons exist. Indeed, in the literature the critical wave vector, q_c separating long-range from short-range excitations, as well as the wave vector, q_{LD}, beyond which excitations become Landau-damped, are often taken to be of the order of the Thomas-Fermi screening parameter[26,27] ($\beta_{s,3D}^2 = (e^2/\epsilon^\infty)(\partial n_g/\partial E_F)$ for bulk

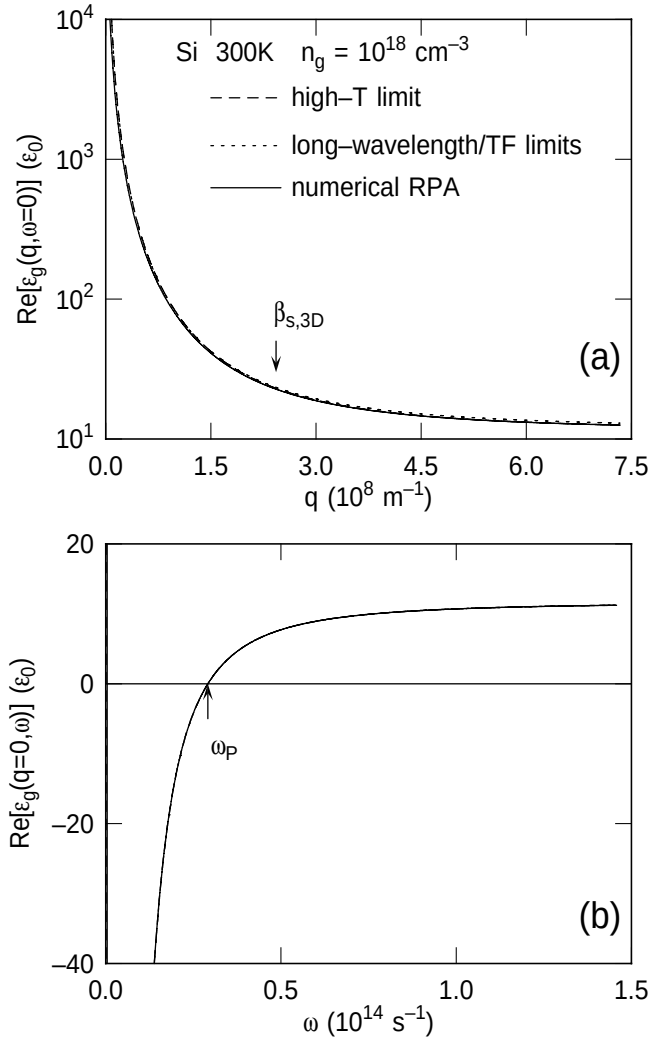

FIG. 1. Real part of the bulk dielectric function $\epsilon_g(q,\omega)$ for Si at 300 K with an electron concentration $n_g = 10^{18}$ cm^{-3} evaluated in (a) as a function of q at $\omega = 0$ and in (b) as a function of ω at $q = 0$ using three different models: The solid lines, labeled 'numerical RPA' show results obtained using the numerical evaluation of the RPA expression Eqns. (26)-(27); The dashed lines, labeled 'high-T limit', refer to the nondegenerate, high-T limit expressed by Fetter and Walecka[28] in terms of the plasma dispersion function. Finally, the dotted lines labeled 'long-wavelength/TF limits', have been obtained using the Thomas-Fermi limit, Eq. (26) – with $\beta_{s,3D}$ in place of the full $\beta_{3D}(q,\omega)$ – in (a), the long-wavelength limit, Eq. (33), in (b). Note the excellent accuracy of the numerical evaluation of the RPA expression, which coincides with the other results, as we should expect in the non-degenerate limit.

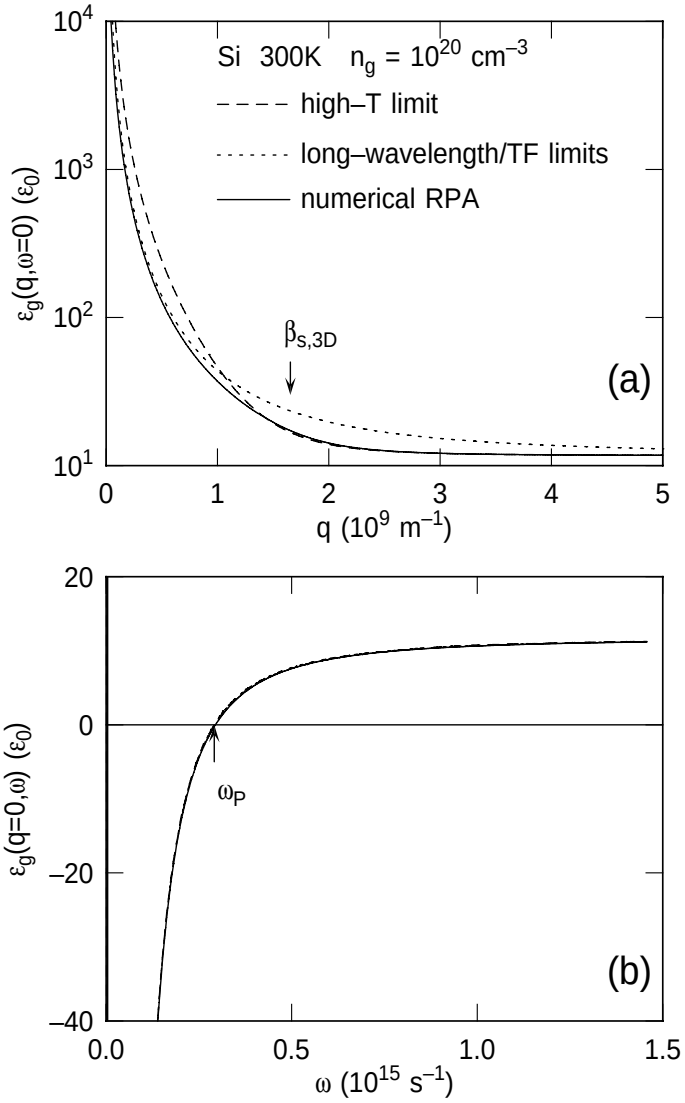

FIG. 2. As in Fig. 1, but for degenerate Si ($n_g = 10^{20}$ cm^{-3}). Note in (a) that the 'numerical RPA' results correctly match the degenerate TF-limit at low q and the q-dependence of the high-T limit at larger q. In (b), the RPA yields a slightly larger plasma frequency (*i.e.*, the frequency at which ϵ_g vanishes) than the other two approximations.

materials at density n_g, $\beta_{s,2D} = [e^2/(2\epsilon^\infty)](\partial n_s/\partial E_F)$ for the two-dimensional electron gas at a sheet density n_s), but more precise determinations of q_{LD} are lacking. Let us adopt the RPA expression

$$\epsilon_g(\mathbf{q};\omega) \;=\; \epsilon^\infty_{Si}\left[1 \;+\; \frac{\beta^2_{3D}(q,\omega)}{q^2}\right] \;, \qquad (26)$$

for the dielectric function of the (bulk) Si gate, where $\mathbf{q}$ is the three-dimensional wave vector, and the screening function $\beta_{3D}(q,\omega)$ is given by:

$$\beta^2_{3D}(q,\omega) \;=\; -\lim_{s\to 0^+}\frac{g_v e^2}{\epsilon^\infty_{Si}}\frac{1}{4\pi^2}\int_0^\infty dp\; p^2\int_0^\pi d\theta\sin\theta$$
$$\times\frac{f(p^2)-f(p^2+q^2+2pq\cos\theta)}{E(p^2)-E(p^2+q^2+2pq\cos\theta)+\hbar\omega+i\hbar s}\;, \qquad (27)$$

where g_v is the degeneracy of the conduction-band valleys (including spin) and $E(p^2)$ the dispersion in the first conducton band, here assumed to be parabolic. The imaginary part of this expression vanishes (*i.e.*, no Landau damping) as long as the denominator in the integrand does not vanish. Otherwise, the limit $s \to 0^+$ gives a δ-function which allows a quick evaluation of the inner (angular) integral: For a parabolic dispersion with effective mass m^*, and denoting by $\theta(x)$the step function, we have:

$$\lim_{s\to 0^+} \; mboxIm \int_0^\pi d\theta \sin\theta \; \frac{f(p^2) - f(p^2+q^2+2pq\cos\theta)}{E(p^2) - E(p^2+q^2+2pq\cos\theta) + \hbar\omega + i\hbar s}$$
$$= \frac{\pi m^*}{\hbar^2 qp} \; \frac{f[E(p^2)] - f[E(p^2)+\hbar\omega]}{[1+(m^*\omega)^2/(\hbar qp)^2 + (m^*\omega)/(\hbar p^2) + q^2/(4p^2)]^{1/2}}$$
$$\times\theta\left(1 - \left|\frac{m^*\omega}{\hbar qp} - \frac{q}{2p}\right|\right) . \qquad (28)$$

The integral over the momentum p can be easily performed numerically. A numerical evaluation of the real part of Eq. (27) requires more care, since taking the limit $s \to 0^+$ requires the evaluation of the principal part of the integral. Empirically, we have found that by giving ω a small imaginary part (that is, fixing $\hbar s \approx 1$ μeV) and by selecting sufficiently small integration intervals $\Delta\theta$ symmetrically selected around the 'pole' of the integrand, allows a sufficiently accurate evaluation of $\beta_{3D}(q,\omega)$. We have verified this by calculating $\epsilon_g(\mathbf{q};\omega)$ for Si in various cases and comparing it with known high-temperature, long-wavelength, and static limits. In Figs. 1(a) and 2(a), for example, we compare the wavelength-dependence of the static dielectric response of bulk Si, $\epsilon_g(q,\omega=0)$, computed from the numerical integration of Eq. (27) and using Eq. (26) (labeled 'numerical RPA' in the figures), with the nondegenerate high-temperature limit given by Fetter and Walecka[28] (labeled 'high-T limit'), or the simple Thomas-Fermi (Debye-Hückel in the non-degenerate case of Fig. 1) expression $\epsilon_{Si}^{\infty}(1+\beta_{s,3D}^2/q^2)$. In the bottom frames, the same three models are employed for the evaluation of the frequency dependence of the long-wavelength dielectric response, $\epsilon_g(q=0,\omega)$. As expected, in the nondegenerate limit (Fig. 1), the results obtained from the three models are virtually indistinguishable, while at a larger electron density ($n_g = 10^{20}$ cm^{-3} in Fig. 2), the static behavior obtained from the RPA approaches the correct limits at long and short wavelengths. Figures 3 and 4 illustrate a similar comparison bewteen results obtained from the RPA and the high-temperature limit for the case of the wavelength dependence at the plasma frequency $\epsilon_g(q,\omega=\omega_P)$, (top frames) and for the frequency dependence at a wavelength corresponding to the screening length, $\epsilon_g(q=\beta_{s,3D},\omega)$. The nonvanishing imaginary part of the dielectric function is also shown.

In dealing with the response of the 2DEG, $\beta_{\mu\mu'}(Q,\omega)$ is similarly approximated by its RPA expression

$$\beta_{\mu\mu'}(Q,\omega) \; = \; -\lim_{s\to 0^+} \frac{e^2}{\epsilon_{Si}^{\infty}} \frac{1}{4\pi^2} \int_0^\infty dP \; P \int_0^{2\pi} d\phi$$
$$\times \frac{g_\mu f_\mu(P^2) - g_{\mu'} f_{\mu'}(P^2+Q^2+2PQ\cos\phi)}{E_\mu(P^2) - E_{\mu'}(P^2+Q^2+2PQ\cos\phi) + \hbar\omega + i\hbar s} \; , \qquad (29)$$

where g_μ is the degeneracy of the μ-th subband, *including spin*, which can be evaluated in the same way, the only modification being Eq. (28) which, for the intra-subband processes we consider ($\mu = \mu'$), becomes:

$$\lim_{s\to 0^+} \; mboxIm \int_0^{2\pi} d\phi \frac{f_\mu(P^2) - f_\mu(P^2+Q^2+2PQ\cos\phi)}{E_\mu(P^2) - E_\mu(P^2+Q^2+2PQ\cos\phi) + \hbar\omega + i\hbar s}$$
$$= \frac{\pi m_\mu}{\hbar^2 PQ} \; \{f[E_\mu(P^2)] - f[E_\mu(P^2)+\hbar\omega]\}$$
$$\times\theta\left(1 - \left|\frac{m_\mu\omega}{\hbar QP} - \frac{Q}{2P}\right|\right) , \qquad (30)$$

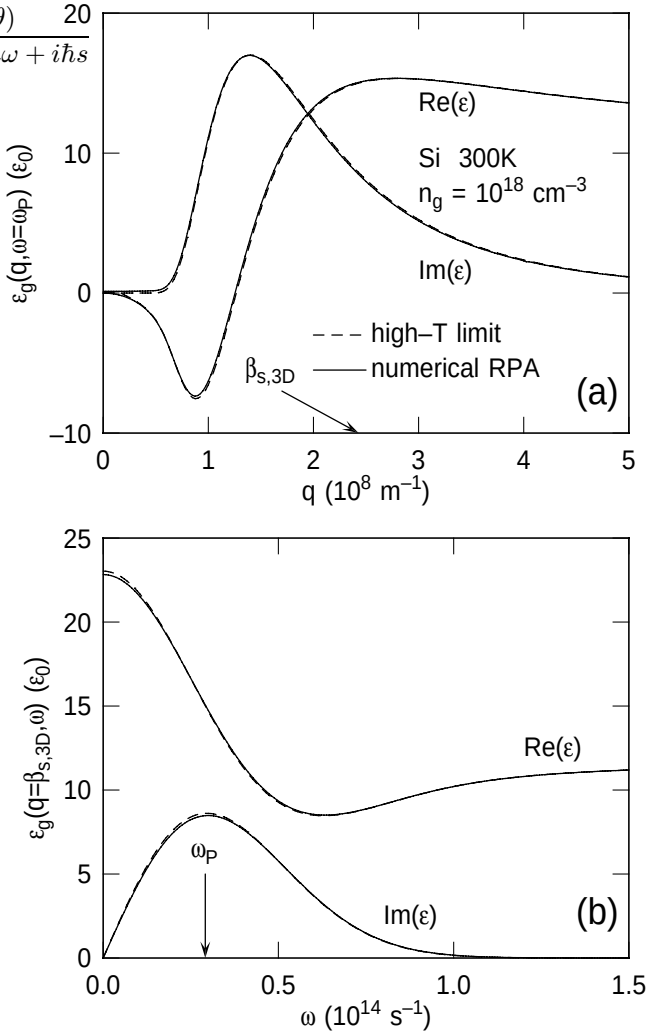

FIG. 3. As in Fig. 1, but comparing the real and imaginary parts of the bulk Si dielectric function at a non-degenerate electron concentration of 10^{18} cm^{-3} computed using either the numerical-RPA expression (solid lines) or the nondegenerate high-T expression (dashed lines) as a function of Q at the plasma frequency (a) or at the Thomas-Fermi screening wavelength $\beta_{s,3D}$ (b).

having assumed the dispersion in subband μ to be isotropic with effective mass m_μ.

Attempting to account for inter-subband excitations ($\mu \neq \mu'$) results in a large number of solutions of Eq. (17), actually too many to be correctly distinguished during the numerical search for the solutions of the secular equation. Thus, we have ignored the inter-subband modes, whose importance has been argued by Price[29] to be small, and considered only the intra-subband excitations. Finally, for the insulator we have used a model dielectric function appropriate to SiO_2:

$$\epsilon_i(\omega) \; = \; \epsilon_{ox}^{\infty} \; + \; (\epsilon_{ox}^{i} - \epsilon_{ox}^{\infty}) \; \frac{\omega_{TO2}^2}{\omega_{TO2}^2 - \omega^2}$$
$$+ \; (\epsilon_{ox}^{0} - \epsilon_{ox}^{i}) \; \frac{\omega_{TO1}^2}{\omega_{TO1}^2 - \omega^2} \; , \qquad (31)$$

where $\omega_{TO2} \approx 138.1$ meV and $\omega_{TO1} \approx 55.6$ meV are the frequencies of the TO-modes, and $\epsilon_{ox}^{\infty} \approx 2.5\epsilon_0$, $\epsilon_{ox}^{i} \approx 3.05\epsilon_0$, and $\epsilon_{ox}^{\infty} \approx 3.9\epsilon_0$ are the permittivities of SiO_2 at

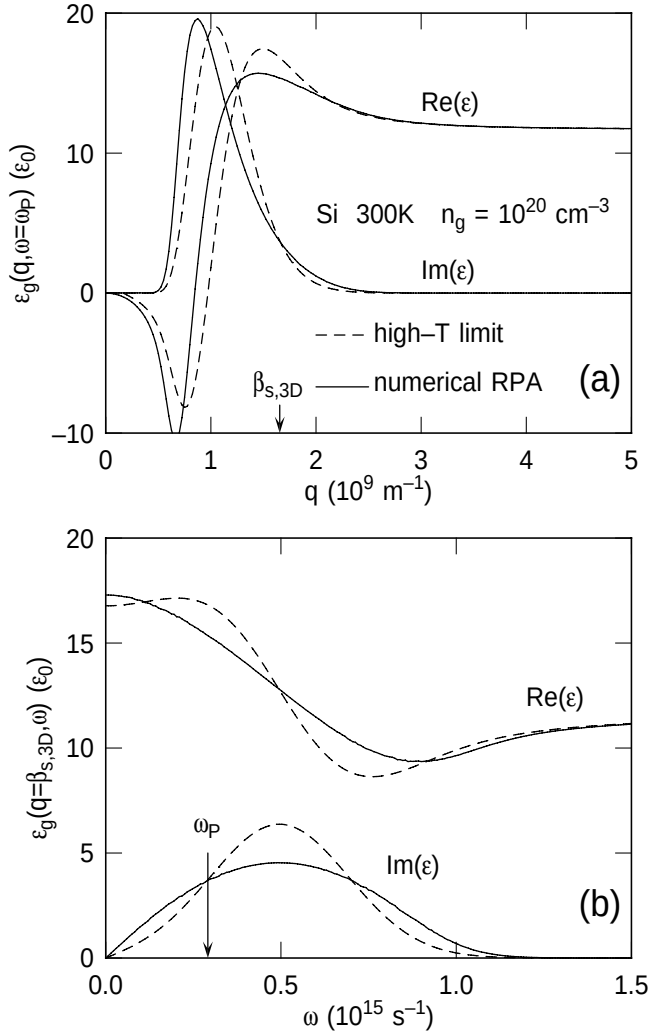

FIG. 4. As in Fig. 3, but in a degenerate case ($n_g = 10^{20}$ cm^{-3}).

small frequencies ($\omega << \omega_{TO1}$, at intermediate frequencies ($\omega_{TO1} << \omega << \omega_{TO2}$), and high frequencies ($\omega >> \omega_{TO2}$), respectively.

F. Long-wavelength limit

In the limit of long-wavelength perturbations ($q \to 0$), the imaginary part of the dielectric functions vanishes. Thus, setting $s = 0$, and expanding in powers of q, we can rewrite Eq. (27) as:

$$\beta_{3D}^2(q,\omega) \;=\; -\frac{g_v e^2}{\epsilon_{Si}^{\infty}} \int \frac{d\mathbf{p}}{(2\pi)^3} \; \frac{f(\mathbf{p}) - f(\mathbf{p}+\mathbf{q})}{E(\mathbf{p}) - E(\mathbf{p}+\mathbf{q}) + \hbar\omega} \;\to\; -\frac{e^2 g_v q^2}{\epsilon_{Si}^{\infty} m^* \omega^2} \int \frac{d\mathbf{p}}{(2\pi)^3}\, f(\mathbf{p}) \;=\; -\frac{\omega_P^2}{\omega^2} q^2 \;, \qquad (32)$$

where $\omega_P^2 = (e^2 n_g)/(m^* \epsilon_{Si}^{\infty})$ is the plasma frequency at a bulk electron density n_g. The bulk dielectric function of the semiconductor, Eq. (26), becomes the plasma expression

$$\epsilon_g(\mathbf{q} \to \mathbf{0};\omega) \;=\; \epsilon_{Si}^{\infty} \left(1 \;-\; \frac{\omega_P^2}{\omega^2} \right) \;. \qquad (33)$$

Finally, the 'effective dielectric function' $\tilde{\epsilon}_g(Q,\omega)$ in Eq. (18) simplifies to

$$\tilde{\epsilon}_g(Q,\omega) \;\to\; \epsilon_g(\mathbf{Q} \to \mathbf{0};\omega) \int_{-\infty}^{+\infty} \frac{dq_z}{q\pi} \; \frac{1}{1 + (q_z/Q)^2} \;=\; \epsilon_g(\mathbf{q} \to \mathbf{0};\omega) \;. \qquad (34)$$

Obtaining a similar simplification for $\tilde{\epsilon}_s^{(2D)}(Q,\omega)$ in Eq. (19) at long wavelengths requires some additional approximations. First, let us ignore the extension of the wavefunctions $\zeta_\mu(z)$ along the z-direction, assuming that the electron gas in the inversion layer is a two-dimensional sheet of charge. Thus, we assume $\zeta_\mu(z) \to \delta(z)$. In this limit, $\Phi^{(0)}_{Q,\mu\mu'} \to \delta_{\mu\mu'}\; e^{-Qt_{ox}}$. Treating screening at the lowest-order (Hartree approximation), $\Phi_{Q,\omega;\mu\mu'} \approx \Phi^{(0)}_{Q,\mu\mu'} \to \delta_{\mu\mu'}\; e^{-Qt_{ox}}$. Finally, expanding $\beta_{\mu\mu'}(Q,\omega)$ in powers of Q, as done in Eq. (32) above, from Eq. (19) we obtain:

$$\tilde{\epsilon}_s^{(2D)}(Q \to 0,\omega) \;=\; \epsilon_s^{\infty} \left\{ 1 \;-\; \sum_{\mu} \; \frac{e^2 n_\mu Q}{2\epsilon_{Si}^{\infty} m_\mu \omega^2} \right\} \;=\; \epsilon_{Si}^{\infty} \left[1 \;-\; \frac{\omega_{P,2D}^2(Q)}{\omega^2} \right] \;, \qquad (35)$$

where m_μ and n_μ are the effective mass and electron density in subband μ, respectively, and we have defined a two-dimensional plasma frequency

$$\omega_{P,2D}^2(Q) \;=\; \sum_{\mu} \; \frac{e^2 n_\mu Q}{2\epsilon_{Si}^{\infty} m_\mu}. \qquad (36)$$

G. Dispersion of the interface modes

In order to illustrate the basic properties of the interface modes, in this section we specify typical values of the average electron concentration in the depletion region of the poly-silicon gate, n_g, and for the electron sheet density, n_s, in the inverted region (channel) in the Si substrate. Furthermore, here and in the following we shall use the depletion approximation in the Si gate and the triangular well approximation for the channel. Thus, we use the following equations relating *i*) the surface potential in the gate and substrate, ΔV_g and ΔV_s, respectively, *ii*) the width of the depletion regions, t_g and t_s, respectively, *iii*) the electric fields in the oxide and at the surface in the substrate, F_i and F_s (see Ref. 30):

$$\Delta V_g \;=\; \frac{\epsilon_i^{\infty 2} F_i^2}{2e\epsilon_{Si}^{\infty} N_{Dg}} \;=\; \frac{\epsilon_{Si}^{\infty} F_s^2}{2eN_{Dg}} \;, \qquad (37)$$

where N_{Dg} is the donor concentration in the gate,

$$\Delta V_s \;=\; 2k_B T \ln\left(\frac{N_{As}}{n_i} \right) \;, \qquad (38)$$

which is the strong-inversion limit, n_i being the intrinsic carrier concentration and N_{As} the homogeneous acceptor concentration in the p-type substrate,

$$t_g \;=\; \left(\frac{2\Delta V_g \epsilon^{\infty}_{Si}}{e N_{Dg}}\right)^{1/2} \;, \tag{39}$$

$$t_s \;=\; \left(\frac{2\Delta V_s \epsilon^{\infty}_{Si}}{e N_{As}}\right)^{1/2} \;, \tag{40}$$

$$F_s \;=\; e\frac{n_{ds}+n_s}{\epsilon^{\infty}_{Si}} \;, \tag{41}$$

where $n_{ds} = N_{As} t_s$ is the areal density of charges in the depletion layer of the substrate, and

$$F_i \;=\; \frac{\epsilon^{\infty}_{Si}}{\epsilon^{\infty}_{i}}\, F_s \;. \tag{42}$$

When specifying n_g and n_s, we only require the additional specification of N_{As}, of the insulator thicknes t_{ox}, and Eq. (41). When specifying N_{Dg} instead of n_g, we make use of Gauss' law

$$n_s - n_{ds} \;=\; t_g(N_{Dg} - n_g) \tag{43}$$

and Eqns. (37), (39), and (41) to solve for n_g.

Having determined the surface field F_s, the energies E_μ and wavefunctions $\zeta_\mu(z)$ of the subbands in the inversion are obtained from the well-known expressions[31]:

$$E_\mu \;=\; \left(\frac{\hbar^2}{2m_{z\mu}}\right)^{1/3} \left[\frac{3\pi e F_s}{2}\left(\mu+\frac{3}{4}\right)\right]^{2/3} \;, \tag{44}$$

where $m_{z\mu}$ is the effective electron mass in subband μ along the direction perpendicular to the interface,

$$\zeta_\mu(z) \;=\; \mathcal{N}\ \mathrm{Ai}\left[\left(\frac{2m_z e F_s}{\hbar^2}\right)^{1/3}\left(z - t_{ox} - \frac{E_\mu}{eF_s}\right)\right] \;, \tag{45}$$

where $\mathrm{Ai}(x)$ is the Airy function, which we express in the following integral form suitable for its numerical evaluation:

$$\mathrm{Ai}(x) \;=\; \frac{1}{\pi}\int_0^\infty \cos\left(\frac{z^3}{3}+zx\right)\, dz \;, \tag{46}$$

and $\mathcal{N}$ is a normalization costant such that

$$\int_{t_{ox}}^{\infty} |\,\zeta_\mu(z)\,|^2 \;\; dz \;=\; 1 \;. \tag{47}$$

Finally, non-parabolic corrections, Fermi levels, and subband occupations are evaluated as in Ref. 32.

Let us start with the long-wavelength approximation (Sec. II F) assuming for now isotropic band structure both in the gate (using an effective mass $m^* = 0.32 m_0$, where m_0 is the free electron mass) and in the substrate (using an 'in-plane' – or density-of-states – effective mass $m_\mu = m_t = 0.19 m_0$ and a 'quantization mass $m_{z,\mu} = m_l = 0.91 m_0$ for the 'unprimed' μ-th subband and $m_\mu = (m_l m_t)^{1/2} = 0.42 m_0$ and $m_{z,\mu} = 0.19 m_0$ for the 'primed' μ-th subband). Let us take $n_g = 5\times 10^{19}$ cm^{-3} and $n_s = 10^{13}$ cm^{-2}. As shown by the dotted lines in Fig. 5(a), when fully isolated, the gate will exhibit the bulk plasma frequency $\omega_g = [e^2 n_g/(m^*\epsilon^{\infty}_{Si})]^{1/2}$ while the 2DEG in the substrate, assuming only the ground-state subband is populated, will oscillate at the frequency $\omega_s(Q) = \omega_{P,2D}(Q) = [e^2 n_s Q/(2\epsilon^{\infty}_{Si} m_t)]^{1/2}$ given by Eq. (36). Note that $\omega_s(Q)$ vanishes in the limit of long wavelengths for a simple physical reason: When ignoring inter-subband processes, the electron gas cannot respond with a redistribution of charge purely along the z-direction, since the electrons are 'frozen' in their sub-

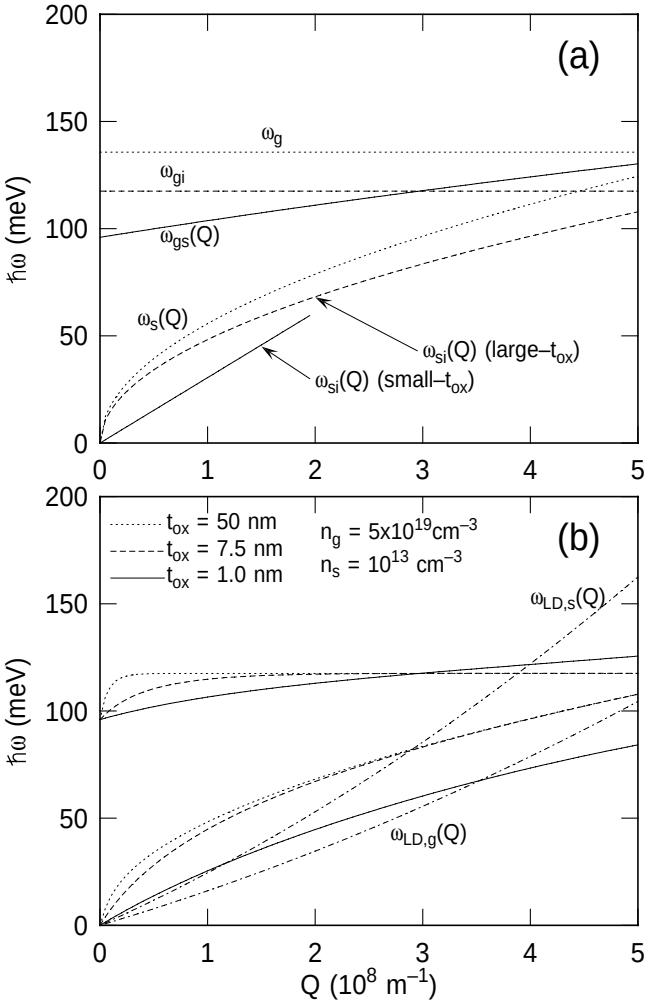

FIG. 5. **(a)** Theoretical limits for the gate and substrate plasma frequencies as functions of plasmon wave vector Q: ω_g is the long-wavelength bulk Si plasma frequency, ω_{gi} the plasma frequency at the Si/SiO_2 interface, $\omega_{gs}(Q)$ the frequency of the optical mode at the bulk-Si/inverted-Si interface (*i.e.*, the zero-oxide-thickness limit), $\omega_s(Q)$ is the frequency of the plasma oscillations of an isolated Si 2DEG, while two curves are shown for $\omega_{si}(Q)$, the frequency of the acoustic mode at the interface between inverted-Si and a thick or thin ($t_{ox} = 1.0$ nm) insulator. The latter curve is truncated since Eq. (51), used in this figure, is valid only in the limit of small Qt_{ox}. **(b)** Acoustic (lower energy) and optical (higher-energy) branches of the plasmons in a bulk-Si/SiO_2/inverted-Si system in the long-wavelength limit for three oxide thicknesses ($t_{ox} = 50$, 7.5, and 1.0 nm). The curves labeled $\omega_{LD,g}(Q)$ and $\omega_{LD,s}(Q)$ mark the onset of the zero-temperature Landau damping in the gate and substrate respectively.

bands. The 2DEG can repond only if the perturbing potential exhibits variations on the plane of the interface, that is, if there are nonvanishing perturbations at nonzero Q.

When placing the gate and substrate next to an insulating layer, but leaving the gate and substrate decoupled (that is, in the limit of $t_{ox} \to \infty$), the plasma frequency at the gate-insulator interface drops to the value

$$\omega_{gi} \;=\; \frac{\omega_g}{(1+\epsilon_i^\infty/\epsilon_g^\infty)^{1/2}} \;, \tag{48}$$

and similarly for the 2DEG:

$$\omega_{si}(Q) \;=\; \frac{\omega_s(Q)}{(1+\epsilon_i^\infty/\epsilon_s^\infty)^{1/2}} \;, \tag{49}$$

as shown by the dashed lines in Fig. 5(a). In the opposite limit of a very small (but non-zero) oxide thickness, the gate and substrate modes become fully coupled. The first of the two resulting modes represent a stiffer (optical) mode in which excess negative charges oscillate in phase on the two sides of the interface. The frequency of this coupled optical interface mode, present even for $t_{ox} = 0$, is given by

$$\omega_{gs}^2(Q) \;=\; \frac{\epsilon_s^\infty \omega_s(Q)^2 + \epsilon_g^\infty \omega_g^2}{\epsilon_s^\infty + \epsilon_s^\infty} \;. \tag{50}$$

This is shown by the higher-energy solid line in Fig. 5(a). The second mode is a lower-energy acoustic branch representing excess negative charges on the two sides of the interface oscillating out-of-phase. The frequency of this mode in the limit of small Qt_{ox} is given by:

$$\omega_{si}^2(Q) \;\approx\; \frac{e^2 n_s t_{ox} Q^2}{m_t \epsilon_i^\infty} \;, \tag{51}$$

shown for t_{ox} =1 nm by the lower-energy solid line in Fig. 5(a), arbitrarily truncated at $Q = 2\times 10^8$/cm.

In Fig. 5(b) we show the two solutions of the secular equation (20) ignoring TO-phonons in the insulator (*i.e.*, using $\epsilon_i(\omega) \approx \epsilon_{ox}^0$), and using the long-wavelength expressions (34) and (35). In this case Eq. (20) becomes a simple second-degree algebraic equation in ω^2 with solutions given by:

$$\omega_Q^2 \;=\; \frac{1}{2\alpha_Q}\left[\;\beta_Q \pm (\beta_Q^2 - 4\alpha_Q\gamma_Q)^{1/2}\right] \;, \tag{52}$$

where

$$\begin{aligned}\alpha_Q \;=\; & (\epsilon_g^\infty+\epsilon_i^\infty)(\epsilon_s^\infty+\epsilon_i^\infty)\; e^{Qt_{ox}} \\ & -(\epsilon_g^\infty-\epsilon_i^\infty)(\epsilon_s^\infty-\epsilon_i^\infty)\; e^{-Qt_{ox}} \;,\end{aligned} \tag{53}$$

$$\begin{aligned}\beta_Q \;=\; & -[(\epsilon_g^\infty+\epsilon_i^\infty)\epsilon_s^\infty\omega_s^2(Q) + (\epsilon_s^\infty+\epsilon_i^\infty)\epsilon_g^\infty\omega_g^2\;]\; e^{Qt_{ox}} \\ & +[(\epsilon_g^\infty-\epsilon_i^\infty)\epsilon_s^\infty\omega_s^2(Q) + (\epsilon_s^\infty-\epsilon_i^\infty)\epsilon_g^\infty\omega_g^2\;]\; e^{-Qt_{ox}} \;,\end{aligned} \tag{54}$$

$$\gamma_Q \;=\; \epsilon_g^\infty\epsilon_s^\infty\omega_g^2\;\omega_s^2(Q)\;(e^{Qt_{ox}} - e^{-Qt_{ox}}) \;, \tag{55}$$

Solutions relative to three insulator (SiO_2) thicknesses are shown: t_{ox} = 50, 7.5, and 1.0 nm. Notice how the optical branch for the thick-oxide case (t_{ox} = 50 nm) exhibits the expected thick-oxide limit $\omega_Q \approx \omega_{gi}$, Eq. (48), for all wavelengths shorter than the oxide thickness, approaching the limiting 'thin-oxide' fully-coupled limit, $\omega_{gs}(Q)$ only for values of $Q << 1/t_{ox}$. The acoustic branch is also practically indistinguishable from the value $\omega_{si}(Q)$ expected from Eq. (49). In the opposite case of a very thin oxide, t_{ox} = 1.0 nm, the optical branch approaches the frequency $\omega_{gs}(Q)$ of Eq. (50), while the acoustic branch approaches the small-Qt_{ox} limit $\omega_{si}(Q)$ of Eq. (51).

Two additional lines of great physical significance are also shown in Fig. 5(b): Labeled by $\omega_{LD,s}(Q)$ and $\omega_{LD,g}(Q)$ are the two frequencies defined by

$$E(k_F+Q) - E(k_F) \;=\; \hbar\omega_{LD,g}(Q) \;, \tag{56}$$

and

$$E_0(K_F+Q) - E_0(K_F) \;=\; \hbar\omega_{LD,s}(Q) \;, \tag{57}$$

where $k_F = (6\pi^2 n_g/g_v)^{1/3}$ and $K_F = (4\pi n_s/g_0)^{1/2}$ are the zero-temperature Fermi wave vectors in the gate and inversion layer in the extreme quantum limit (only the subband $\mu = 0$ being occupied), respectively. In Eqns. (56) and (57) the wave vectors $\mathbf{k}_F$ and $\mathbf{K}_F$ are taken to be parallel to $\mathbf{Q}$. For a given wave vector Q, collective modes oscillating at a frequency ω_Q smaller than $\omega_{LD,g}(Q)$ will decay into single-particle excitations in the gate, by exciting an electron at the Fermi surface and placing it at an energy $E(k_F+Q)$ given by Eq. (56). Similarly, collective modes oscillating at a frequency smaller than $\omega_{LD,s}(Q)$ will decay into single-particle excitations in the inversion layer. These processes, known as 'Landau damping', are not captured by the long-wavelength limit employed so far, since in Sec. II F we have explicitly ignored the imaginary part of dielectric functions. However, when we retain the imaginary parts of $\tilde{\epsilon}_g(Q,\omega)$ and $\tilde{\epsilon}_s^{(2D)}(Q,\omega)$, the secular equation (20) will, in general, admit complex solutions ω_Q. The imaginary part of ω_Q represents the damping of the oscillations as their energy is drained by exciting electrons at higher energies. Within the long wavelength approximation, we may assume that excitations associated with a field mainly localized at the gate/insulator interface will be damped by the Landau processes occurring mainly in the gate, and vice-versa for modes whose field is localized mainly at the insulator/substrate interface. Figure 6 shows the potential associated with the optical and acoustic modes for a system with t_{ox} = 2.5 nm, $n_g = 3\times 10^{19}$ cm^{-3} and $n_s = 10^{13}$ cm^{-2}. For a relatively short wavelength, $Q = 2/t_{ox}$ (dashed lines in the Fig. 6) the optical and acoustic potentials are indeed strongly confined around either interface. Only at a wavelength four times as large, $Q = 1/2t_{ox}$, do the two modes begin to act as strongly coupled oscillations, a significant fraction of their energy being present throughout the insulator.

Despite these considerations, some uncertainty remains: First, Eqns. (56) and (57) are zero-temperature expressions. The damped/undamped boundary marked by the curves $\omega_{LD,g}(Q)$ and $\omega_{LD,s}(Q)$ in Fig. 5(b) will become softer at finite temperatures. Secondly, interface

plasmons can still be considered sufficiently well-defined excitations whenever Im $\omega_Q \leq$ Re ω_Q (see note 33). Finally, an approximation often used in the literature is to assume that the Landau-damping wave vector q_{LD}, given by $E(k_F+q_{LD})-E(k_F) = \hbar\omega_P(q_{LD})$ for the bulk case, is of the order of the screening parameter $\beta_{s,3D}$ (See refs. 26 and 27). While this is qualitatively correct, numerical

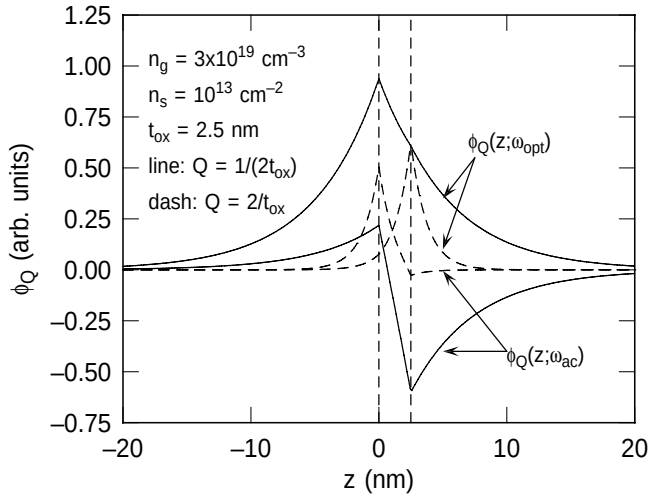

FIG. 6. Electrostatic potential energy associated with interface plasmons at the indicated electron densities. The dashed lines show the dependence of the potential $\phi_Q(z)$ associated with the optical and acoustic branches at a relatively small wavelength $\approx \pi t_{ox}$: The two branches appear largely decoupled, the optical branch being localized mainly at the SiO_2/inverted-Si interface, the acoustic branch at the bulk-Si-gate/SiO_2 interface. At a larger wavelength ($\approx 4\pi t_{ox}$) the two modes (shown by solid lines) appear more strongly coupled, exhibiting a potential energy distributed throughout the MOS structure.

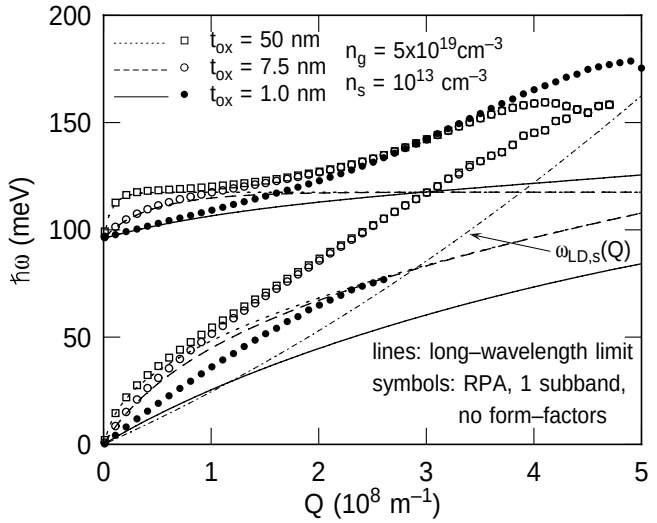

FIG. 7. Dispersion of the interface plasmons for three oxide thicknesses obtained employing the dielectric function evaluated in the long-wavelength approximation (lines) or using the RPA approximation using a single subband and having ignored the form-factors and the screening effects appearing in Eq. 19 (*i.e.*, having set $\Phi_{Q,\omega;\mu\mu'} = \Phi^{(0)}_{Q,\mu\mu'} = e^{-Qt_{ox}}$) (symbols).

differences of a factor of 2 or so can affect profoundly our mobility calculations, as we shall see below.

In order to clarify this issue, we should ideally employ the RPA expressions for the dielectric functions and solve Eq. (20) in the complex plane. In practice, numerical difficulties force us to simplify the problem: We shall indeed employ the RPA, but consider only the real part of the secular equation evaluated for real ω. The problem is now reduced to finding the real zeroes of a real function. However, solutions will be found only in the 'undamped' region: Oscillations which are Landau-damped appear spuriously within the long-wavelength limit, not within the RPA, since the imaginary part of $\tilde{\epsilon}_g(Q,\omega)$ and $\tilde{\epsilon}_s^{(2D)}(Q,\omega)$ are now explicitly accounted for. Even in this form, the problem is numerically challenging and we can obtain the RPA dispersions ω_Q only laboriously and in a limited number of cases. From these, however, we can estimate the errors we shall make in computing the electron mobility employing the long-wavelength limit below.

Figure 7 shows the t_{ox}-dependence of the dispersion obtained in the long-wavelength limit, as in Fig. 5(b), and using the full RPA expressions for the dielectric functions. In both cases, a single subband has been considered and the 'form factors' $\Phi^{(0)}_{Q;\mu\mu'}$ and $\Phi_{Q,\omega;\mu\mu'}$ have been approximated with their values for the zero-thickness limit of the inversion layer. This allows a comparison by isolating the effect of the full Q-dependence of $\tilde{\epsilon}_g(Q,\omega)$ and $\tilde{\epsilon}_s^{(2D)}(Q,\omega)$. Note that both the optical and the acoustic branches deviate from the long-wavelength limit for wavelengths comparable to the screening lengths in the gate (optical) and inversion layer (acoustic). Indeed, if in Eq. (32) we retain terms of order q^4, we find the leading correction to the bulk plasma dispersion

$$\omega_P \;\rightarrow\; \omega_P \left(1 + \frac{3}{2}\frac{q^2}{\beta^2_{s,3D}} \right) . \qquad (58)$$

Similarly, Eq. (36) would be modified as follows:

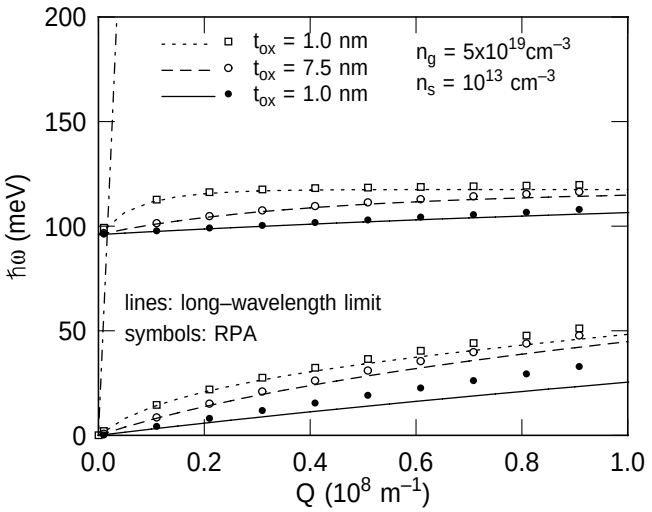

FIG. 8. As in Fig. 7, enlarging the long-wavelength region in order to show the range of wavelengths over which the long-wavelength limit is valid. The dashed-dotted line at left is the Si light-line, showing that retardation effects are important only for very large wavelengths.

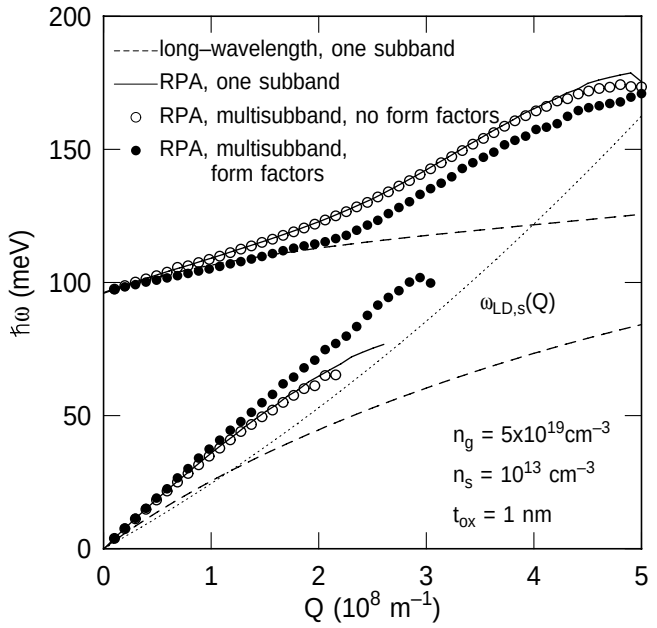

FIG. 9. Dispersion of the interface plasmons evaluated using the long-wavelength limits of the gate and substrate dielectric functions (dashed lines, as in Fig 5 for the case $t_{ox} = 1$ nm), using the RPA expression in the one-occupied-subband, no form-factors approximation (solid lines), as in Fig 7, also for the case $t_{ox} = 1$ nm), using the RPA expression, many subband occupied, no form-factors (open circles), or, finally the full RPA with the form-factor and screening corrections of Eq. (19) for $t_{ox} = 1$ nm.

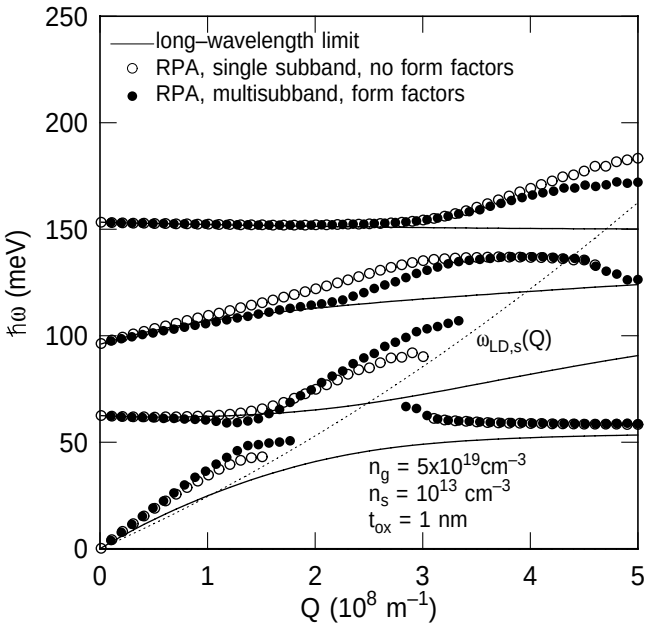

FIG. 10. Calculated dispersion of the coupled interface plasmons/TO-oxide phonon modes for three of the models described in Fig 9.

$$\omega_{P,2D}(Q) \;\rightarrow\; \omega_{P,2D}(Q)\left(\,1+\frac{3}{2}\frac{Q}{\beta_{s,2D}}\,\right)\,. \qquad (59)$$

All RPA results in Fig. 7 show this behavior. Figure 8 shows in more detail the low-Q region. The RPA differs from the long-wavelength approximation especially as far as the thin-oxide acoustic-branch is concerned. Note also the Si light-line (the almost vertical dot-dashed line) $\omega_{light}(Q) = cQ/(\epsilon^{\infty}_{Si}/\epsilon_0)^{1/2}$, where c is the speed of light and ϵ_0 the permittivity of the vacuum. Retardation effects, as noted above, are important only in the small region for $\omega_Q > \omega_{light}(Q)$. In addition, note in Fig. 7 that the acoustic branches are truncated by Landau damping in the substrate at approximately the expected wavelength.

Figure 9 illustrates the significant effect of the finite extension of the inversion charge and the negligible effect of considering the occupation of many subbands. The results labeled 'RPA, one subband' (solid line) coincide with those labeled 'RPA, multisubband, no form factors' (circles), the latter differing from the former simply by permitting the occupation of many subbands. This is not a surprise at the large electron sheet density considered here, since about 96% of the electrons populate the gound-state subband anyway. On the contrary, when employing the 'exact' form factors $\Phi^{(0)}_{Q;\mu\mu'}$ and $\Phi_{Q,\omega;\mu\mu'}$, we see that the acoustic branch becomes slighty stiffer, while the energy of the optical branch is reduced. Again this should be expected: As we remove the inversion charge from the interface, as required by the 'correct' spatial distribution $\mid \zeta_\mu(z) \mid^2$, we lower the electrostatic repulsion of the in-phase optical oscillations while lowering also the electrostatic attraction of the out-of-phase acoustic branch. As a result, the 'gap' between the acoustic and the optical branch shrinks.[34]

Finally, Fig. 10 illustrates the effect of the coupling between the interface plasmons and the TO phonons in the insulator. Here we compare the results obtained in the long-wavelength limit (solid lines), those obtained using the RPA but using a single subband and ignoring the form factors (circles), and finally, those obtained by using the more correct RPA model and accounting for the screened form factors (dots). The RPA dispersion appears more 'interesting' since various branches repel each other as their dispersions in the uncoupled approximation (*i.e.*, no TO-phonon/plasmon coupling) would otherwise cross, as a comparison with Fig. 9. shows. Of interest is to note that while the acoustic and optical mainly-plasmon branches are Landau-damped at the expected wavelengths, the lowest-energy branch disappears at the Landau-damping wave vector in the substrate, as expected, but reappears at shorter wavelengths. This is caused by the fact that at sufficiently large Q the lowest-energy mode represents a low energy optical phonon mode with an ever decreasing plasmon content. Thus Landau damping will not affect this mode, once the plasmon content becomes sufficiently small.

We can draw a few conclusions from the results shown in Figs. 5 through 10: The long-wavelength limit yields an adequate dispersion only up to wave vectors of the order of 2×10^7 m^{-1} to 4×10^7 m^{-1} (see Fig. 8) for insulators as thin as 1 nm. The use of the RPA yields noticeable deviations for wave vectors of the order of the Fermi wave vector of the 2DEG at the sheet densities n_s of interest here. However, these corrections are larger for the acoustic branch, which we shall see plays a smaller role in fixing the electron mobility, than for the optical branch. Therefore, the use of the long-wavelength limit in the

following, while dictated purely by numerical necessities, should not result in major errors. In addition, the dispersions computed using the RPA push the Landau-damping limit to shorter wavelengths, as seen clearly in Fig. 7: The acoustic branch, which enters the Landau-damping region at $Q \approx 1.2 \times 10^8$ m^{-1} in the long wavelength limit, extends undamped all the way to $Q \approx 2.6 \times 10^8$ m^{-1} in the RPA. This is due to the higher plasma frequencies obtained within the RPA which pushes q_{LD} to larger values. The inclusion of the electronic form factors (that is, accounting for the thickness of the inversion layer) and their screening, while numerically hard, results only in a minor correction to the dispersion relations (see Fig. 9). Finally, the coupling between interface plasmons and SiO_2 TO phonons results in the expected 'repulsion' of the branches where they would cross (in the uncoupled model) and in a reduction of the phonon frequency from $\hbar\omega_{LO}$ to $\hbar\omega_{TO}$ at short wavelengths (see Fig. 10), as we had already found in the past[32].

III. EFFECTIVE MOBILITY OF THE 2D ELECTRON GAS

In this section we calculate the effective mobility of the 2DEG in the inversion layer of the Si substrate, accounting for scattering with phonons and interface plasmons. Scattering with the TO-phonon content of the coupled interface-plasmons/TO-phonon modes has been dealt with in the past[32] and has been shown to account for a small effect, of the order of 5% or so. Thus, we shall consider only scattering with the interface plasmons. The long-wavelength limit is used to determine their dispersion relation.

A. Effective mobility

The inversion layer is treated in the triangular-well approximation described above in Sec. II G (see Eqns. (44) and (45)).

From the linearization of the two-dimensional Boltzmann transport equation one can obtain the mobility tensor as:

$$\mu_{ij} \;=\; -\frac{e}{\hbar}\left\langle \tau_{p,i} v_i \left(\frac{\partial f}{\partial K_j}\right)\frac{1}{f}\right\rangle_{th} \;, \tag{60}$$

where the indices i and j run over the real-space coordinates x, y, $v_i(\mathbf{K})$ is the electron group velocity, $\tau_{p,i}(\mathbf{K})$ is the relaxation time for the i-th component of the crystal momentum $\hbar\mathbf{K}$, $f(\mathbf{K})$ is, as always, the equilibrium distribution function, and $< \mathcal{X} >_{th}$ denotes the thermal average of the dynamic quantity $\mathcal{X}(\mathbf{K})$,

$$< \mathcal{X} >_{th} = \sum_{\mu} \frac{g_\mu}{n_\mu} \int \frac{d\mathbf{K}}{(2\pi)^2}\; \mathcal{X}(\mathbf{K})\; f_\mu(\mathbf{K}) \;, \tag{61}$$

where n_μ is the electron sheet-density and f_μ the equilibrium distribution function in subband μ.

We consider the Si/SiO_2 interface to be parallel to the $< 100 >$ crystal plane and the x-axis along the (100) direction. Thus, we shall consider only the component μ_{xx} of the mobility defined as

$$\mu_{xx} \;=\; \frac{1}{n_s}\sum_{\mu}\; n_\mu\; \mu_{xx}^{(\mu)} \;, \tag{62}$$

where $\mu_{xx}^{(\mu)}$ the xx component of the mobility tensor in subband μ.

For the general case of elliptical subband structures (as is the case in the 'primed' subbands), we have:

$$\begin{aligned}\mu_{xx}^{(\mu)} \;=\; & \frac{e g_\mu}{4\pi^2 m_{\mu,x}^2 k_B T\; n_\mu}\int_0^{2\pi} d\beta \int_{E_\mu}^{\infty} dE\; K^2 \cos^2\beta \\ & \times \tau_{p,x}(K,\beta) f_\mu(E)[1-f_\mu(E)] \\ & \times \frac{1+2\alpha(E-E_\mu)+2\alpha < E_\mu - V >_\mu}{\cos^2\beta/m_{\mu,x} + \sin^2\beta/m_{\mu,y}} \;,\end{aligned} \tag{63}$$

where the (anisotropic) momentum relaxation time, $\tau_{p,x}(K,\beta)$ has been expressed as a function of energy and angle β between the $\mathbf{K}$-vector and the x-axis, $m_{\mu,x}$ and $m_{\mu,y}$ are the electron effective masses along the x and y axes, respectively. Note that we must consider three seperate 'kinds' of subbands: The 'unprimed' ladder is doubly-degenerate and isotropic with $m_{\mu,x} = m_{\mu,y} = m_t$, while the 'primed' ladder consists of doubly-degenerate high-mobility subbands for which $m_{\mu,x} = m_t$ and $m_{\mu,y} = m_l$, and of those of lower mobility for which $m_{\mu,x} = m_l$ and $m_{\mu,y} = m_t$, also doubly degenerate. We have also assumed a nonparabolic band structure characterized by a nonparabolicity parameter α= - 0.5 eV^{-1}, and have used first-order perturbation theory to express the nonparabolic subband structure[32]: If $E_\mu^{(0)}$ is the (unperturbed) bottom of the μ-th subband in the parabolic-band approximation, in the triangular well approximation the non-parabolic electron dispersion will be

$$\begin{aligned}E_\mu(\mathbf{K}) \;=\; & E_\mu^{(0)} \;+\; \frac{1}{5}\alpha {E_\mu^{(0)}}^2 \\ & + \gamma_\mu(\mathbf{K})\left[1 + \alpha\gamma_\mu(\mathbf{K}) \;+\; \frac{2}{3}\alpha E_\mu^{(0)}\right].\end{aligned} \tag{64}$$

having denoted with $\gamma_\mu(\mathbf{k})$ the parabolic dispersion. Moreover, we shall use the following expectation values

$$< E_\mu^{(0)} - V >_\mu = \; \frac{1}{3}E_\mu^{(0)} \;, \tag{65}$$

$$< (E_\mu^{(0)} - V)^2 >_\mu = \; \frac{1}{5}{E_\mu^{(0)}}^2 \;. \tag{66}$$

Here and in Eq. (63) the expectation value in subband μ is defined as:

$$< X >_\mu = \int_{t_{ox}}^{\infty} X(z)\; \zeta_\mu(z)^2 dz \;, \tag{67}$$

for wavefunctions normalized as in Eq. (47). Note that for spherical and parabolic subbands Eq. (63) reduces to the familiar two-dimensional Kubo-Greenwood expression

$$\mu^{(\mu)} = \frac{e}{m_\mu k_B T\, n_\mu} \int_{E_\mu}^{\infty} dE\ (E - E_\mu)\ \rho_\mu(E) \times \tau(E) f_\mu(E)[1 - f_\mu(E)]\ , \quad (68)$$

where, for $E > E_\mu$, $\rho_\mu(E)$ is the energy-independent two-dimensional density of states $g_\mu m_\mu/(2\pi\hbar^2)$.

B. Momentum relaxation time

We must now deal with the momentum relaxation time, $\tau_{p,x}(K,\phi)$, appearing in Eq. (63). Scattering with plasmons is an anisotropic and inelastic process, so that the derivation of a rigorous relaxation time is a troublesome issue. This has been discussed in detail elsewhere in the context of bulk-plasmons[35]. We start with the expression provided by Chattopadhyay and Queisser[36]. In the case of bulk semiconductors, we can rephrase their analysis as follows: We expand the nonequilibrium distribution function $f^{(ne)}(\mathbf{k})$ around its equilibrium value $f(\mathbf{k})$ to first order in the magnitude of the homogeneous external field F:

$$f^{(ne)}(\mathbf{k}) \approx f(\mathbf{k}) + \frac{e\hbar F}{m^*}\Phi(\mathbf{k})\frac{\partial f}{\partial E}. \quad (69)$$

Using this expansion and detailed balance,

$$S(\mathbf{k},\mathbf{k}')f(\mathbf{k}')[1 - f(\mathbf{k})] = S(\mathbf{k}',\mathbf{k})f(\mathbf{k})[1 - f(\mathbf{k}')]\ , \quad (70)$$

where $S(\mathbf{k}',\mathbf{k})$ is the transition rate from a state $\mathbf{k}$ to a state $\mathbf{k}'$, the collisional term of the Boltzmann equation,

$$\left(\frac{\partial f^{(ne)}}{\partial t}\right)_{coll} = \frac{1}{(2\pi)^3} \int d\mathbf{k}\{S(\mathbf{k},\mathbf{k}')f^{(ne)}(\mathbf{k}')[1 - f^{(ne)}(\mathbf{k})] - S(\mathbf{k}',\mathbf{k})f^{(ne)}(\mathbf{k})[1 - f^{(ne)}(\mathbf{k}')]\}\ , \quad (71)$$

can be rewritten as:

$$\left(\frac{\partial f^{(ne)}}{\partial t}\right)_{coll} \approx -\frac{f^{(ne)}(\mathbf{k}) - f(\mathbf{k})}{\tau_{p,x}(\mathbf{k})}\ , \quad (72)$$

where the relaxation time $\tau_{p,x}(\mathbf{k})$ is defined as:

$$\frac{1}{\tau_{p,x}(E,\theta)} = \frac{1}{4\pi^2}\int dk\ k^2 \int_0^\pi d\theta' \cos\theta' S(k',k;\theta,\theta') \times \frac{1 - f(E')}{1 - f(E)}\left\{1 - \frac{\Phi(\mathbf{k}')}{\Phi(\mathbf{k})}\right\}\ . \quad (73)$$

The problem stems from the fact that the function $\Phi(\mathbf{k})$ is proportional to $k\tau_{p,x}(\mathbf{k})\cos\theta$, where θ is the angle between the wave vector $\mathbf{k}$ and the field $\mathbf{F}$ (assumed to be along the x-axis) and thus it depends on the relaxation time itself, so that Eq. (73) is actually an integral equation for τ retaining the full complexity of the original Boltzmann equation. For isotropic band-structures $\Phi(\mathbf{k})$ is a function of $E(\mathbf{k})\cos\theta$ only. Thus, assuming that $S(\mathbf{k},\mathbf{k}')$ does not depend on the azimuthal angle and using the fact that $S(k',k;\theta,\theta') = S(k',k;-\theta,-\theta')$ in cubic crystals, the troublesome second term within the curly bracket in Eq. (73) becomes $[k'\tau(k')/(k\tau(k)]\cos\phi$ where ϕ is the angle between the intial and final wave vectors $\mathbf{k}$ and $\mathbf{k}'$. For isotropic scattering this term vanishes upon integration over the polar angle, while for elastic scattering $E = E'$ and this term simplifies to $(k'/k)\cos\phi$. In general, though, there is no simplification and one should solve the full integral equation Eq. (73). Fortunately, often sufficiently accurate results are obtained by setting $\Phi(\mathbf{k}')/\Phi(\mathbf{k}) \approx (k'/k)\cos\phi$, since anisotropic corrections due to scattering are typically small in diffusive, equilibrium situations in which isotropy (or a possible band-structure anisotropy) dominates.

In two-dimensional transport we encounter an identical situation and we shall approximate the relaxation time as

$$\frac{1}{\tau(E,\beta)} = \frac{1}{4\pi^2}\int dK K \int_0^\pi d\phi\ S(K',K;\beta,\beta') \times \frac{1 - f(E')}{1 - f(E)}\left\{1 - \frac{K'}{K}\cos\phi\right\}\ , \quad (74)$$

where β (β') is the angle between the wave vector $\mathbf{K}$ ($\mathbf{K}'$) and the field (aligned along the x-axis), and $\phi = \beta - \beta'$.

C. Plasmon scattering

We can now calculate the scattering kernel $S(K',K;\beta,\beta')$ from the matrix element of the plasmon potential over the wavefunctions $\{\zeta_\mu\}$. From Eqns. (3), (20), and (24), the potential associated with an interface mode of frequency ω_Q in the Si substrate ($z \geq t_{ox}$) is

$$\phi_Q(z) = \frac{\epsilon_i(\omega_Q) - \tilde{\epsilon}_g(Q,\omega_Q)}{\epsilon_i(\omega_Q) + \tilde{\epsilon}_s^{(2D)}(Q,\omega_Q)}\left(\frac{\hbar\omega_Q}{2Q\mathcal{D}_Q}\right)^{1/2} e^{-Qz} \quad (75)$$

We must spend a word of caution about the use of this expression as a scattering potential for the electrons in the inversion layer. Equation (75) represents the potential energy associated with modes resulting from the coupling of modes oscillating both in the gate and in the inversion layer itself. It is not trivial to establish how much net loss of momentum of the 2DEG results from scattering with this last component. Indeed, as an electron of the 2DEG emits a plasmon localized at the insulator/substrate interface, the transfered momentum is simply redistributed among the electrons of the 2DEG. If things were this simple, processes of this kind would have no effect on the mobility. But, as also discussed in the past[32,35], plasmons can decay either via Landau damping or via collisions. If Landau damping dominates, indeed the 'lost' momentum is reabsorbed by 2DEG. But if collisional damping dominates, the momentum is indeed lost (to phonons, impurities, etc.). In this case, one would have a net contribution to the momentum relaxation time associated with these modes. On the contrary, momentum transfered from the 2DEG to plasmons localized at the gate-insulator interface does constitute a net loss of momentum. As shown in Fig. 6, the field of Eq. (75) is due to a mixture of oscillations localized at both interfaces. For this reason here we take a 'conservative' approach: We assume that modes

localized at the substrate/insulator interface *do not* contribute to the momentum relaxation processes. Thus, we shall evaluate an *upper bound* for the effective mobility. As a result of this discussion, we consider an 'effective' scattering potential: For optical modes we consider only the fraction of the potential which can be attributed to oscillations localized at the gate-oxide interface. From Eq. (3) this is

$$R_Q \;=\; \frac{\mid b_{Q,\omega_Q} \mid}{\mid d_{Q,\omega_Q} \mid} \; . \tag{76}$$

Indeed, for optical modes, whose potential is due to the sum of in-phase oscillations at both interfaces, the contributions to $\phi_Q(z)$ due to oscillations at both interface add constructively, so the $R_Q < 1$. For acoustic modes, we take the same fraction if the amplitude of the field at the oxide/substrate boundary is actually larger than the field at the gate/oxide. This is the case when the field due to larger oscillations at the oxide/substrate interface is mostly screened by gate charges, as shown in Fig. 6 by the longer wavelength acoustic mode (lower solid line). Thus, we effectively ignore the oscillation of the 2DEG and account only for the 'screening' oscillations induced at the gate/oxide interface. In the opposite case, as shown by the 'left-peaked' dashed line in Fig. 6, it is actually the 2DEG which oscillates while screening the potential induced at the gate/oxide interface. Therefore we leave the potential (75) unchanged, since in this case $R_Q > 1$, and rescaling Eq. (75) by the factor R_Q would amount to incorrectly ignoring the screening effects of the 2DEG. Thus, summarizing, in all cases the 'effective' scattering potential we consider is given by Eq. (75) multiplied by the correction factor $\min(1, R_Q)$.

Using the Fermi golden rule, the momentum relaxation time due to emission or absorption of an interface plasmon of either the acoustic or the optical branch causing electrons with wave vector $\mathbf{k}$ in subband μ to make an intra-valley, intra- or inter-subband transition to subband ν is

$$\begin{aligned}
\frac{1}{\tau^{(SP,\pm)}_{p,x;\mu\nu}(K,\beta)} \;&=\; \frac{e^2}{2\pi\hbar^3} \int_0^{2\pi} d\phi \\
\times \; & \frac{1 - 2\alpha(E' - E_\nu) - 2\alpha < E^{(0)}_\nu - V >_\nu}{\cos^2\beta'/m_{\nu,x} + \sin^2\beta'/m_{\nu,y}} \\
& \times \left\{ \begin{array}{c} n_Q \\ 1 + n_Q \end{array} \right\} \frac{1 - f_\nu(E')}{1 - f_\mu(E)} \mid \mathcal{A}_Q \mid^2 \mid \mathcal{F}_{\mu\nu}(Q) \mid^2 \\
& \times \left[1 - \frac{K'}{K} \cos\phi \right] \frac{1}{\mid g \mid} \; \theta(E' - E_\nu) \; ,
\end{aligned} \tag{77}$$

where the upper (lower) sign/symbols refer to absorption (emission), and $\theta(x)$ is the step function. Other symbols in the equation above are defined as follows:

$$\begin{aligned}
\mathcal{A}_Q \;&=\; \frac{\epsilon_i(\omega_Q) - \tilde{\epsilon}_g(Q,\omega_Q)}{\epsilon_i(\omega_Q) + \tilde{\epsilon}^{(2D)}_s(Q,\omega_Q)} \left(\frac{\hbar\omega_Q}{2Q\mathcal{D}_Q} \right)^{1/2} \\
& \qquad \times \, e^{-Qt_{ox}} \; \min(1, R_Q) \; ,
\end{aligned} \tag{78}$$

is the 'effective' amplitude of the plasmon field,

$$\begin{aligned}
\mathcal{F}_{\mu\nu}(Q) \;&=\; e^{-Qt_{ox}} \; \Phi^{(0)}_{\mu\nu}(Q) \\
&=\; \int_{t_{ox}}^{\infty} dz \; \zeta_\mu(z) \; e^{-Q(z - t_{ox})} \; \zeta_\nu(z) \; ,
\end{aligned} \tag{79}$$

is the electronic form factor,

$$E' = E \pm \hbar\omega_Q \tag{80}$$

is the final energy after emission $(-)$ or absorption $(+)$ of an interface plasmon of energy $\hbar\omega_Q$,

$$K' \;=\; \frac{\{2(E' - E_\nu)[1 - \alpha(E' - E_\nu) - 2\alpha < E^{(0)}_\nu - V >_\nu\}^{1/2}}{\hbar(\cos^2\beta'/m_{\nu,x} + \sin^2\beta'/m_{\nu,y})} \; , \tag{81}$$

is the final electron wave vector in the ν-th subband, E_ν being the nonparabolic energy of the bottom of the subband related to the parabolic energy $E^{(0)}_\nu$ via Eq. (64) for $\gamma_\mu(\mathbf{K}) = 0$, and

$$Q \;=\; (K^2 + K'^2 - 2KK'\cos\phi \;)^{1/2} \tag{82}$$

is the transferred wave vector. When using the long-wavelength limit for the plasmon dispersion, Landau damping must be accounted for 'by hand' by restricting the range of integration in Eq. (77) to values of Q (given by Eq. (82) above) smaller than a cut-off wave vector q_c. We shall discuss below the possible choices for q_c. Note that Eqns. (80)-(82) constitute a system of nonlinear equations which must be solved in order to obtain the final energy and wave vector satisfying their respective conservation laws. We have used an iteration process which has proven to be very efficient: An initial guess is made for the plasmon energy $\hbar\omega_Q$, the resulting energy E' is employed to compute the final wave vector K' and the wave vector transfer Q. The new plasmon energy $\hbar\omega_Q$ is computed and the iteration process is repeated until energy changes from one iteration step to the next are smaller than a preset error (typically, 1 μeV). Very few iterations (order of 5 to 10) are usually required. The term

$$\begin{aligned}
g \;&=\; 1 \pm \frac{d\omega_Q}{dQ} \left(1 - \frac{K'}{K} \cos\phi \right) \\
& \qquad \times \, \frac{1 - \alpha(E' - E_\nu) - 2\alpha < E^{(0)}_\nu - V >_\nu}{\hbar Q(\cos^2\beta'/m_{\nu,x} + \sin^2\beta'/m_{\nu,y})} \; ,
\end{aligned} \tag{83}$$

is the Jacobian resulting from the change of variable required to integrate the energy-conserving δ-function. This term is responsible for the 'spikes' observed in the relaxation rate *vs.* wave vector already observed by Hawrylak and co-workers[37] when computing the inverse lifetime of quasiparticles in layered electron gases. Finally, the plasmon frequency, computed in the long-wavelength limit, is given by Eq. (52).

A simplified expression can be obtained using a parabolic approximation (*i.e.*, setting α =0), in the limit in which only the ground-state subband is occupied, and using the variational expression for the wavefunction $\zeta_0(z)$ (see Ref. 31):

$$\zeta_0(z) \;\approx\; \left(\frac{b^3}{2}\right)^{1/2} (z - t_{ox})\; e^{-(b/2)(z-t_{ox})}\;, \qquad (84)$$

where $b = [12 m_t e^2 n^*/(\epsilon_s^\infty \hbar^2)]^{1/3}$, with $n^* = n_{ds} + (11/32) n_s$. In this case we obtain the isotropic expression by a much simpler evaluation (since in Eq. (63) the angular integral becomes trivial):

$$\frac{1}{\tau^{(SP,\pm)}_{p;00}(K)} \;=\; \frac{e^2 m_t}{2\pi\hbar^3} \int_0^{2\pi} d\phi\; \frac{b^6 e^{-2Qt_{ox}}}{(1+Q/b)^6} \left\{ \begin{array}{c} n_Q \\ 1+n_Q \end{array} \right\} \times \frac{1 - f_{\nu=0}(E')}{1 - f_{\mu=0}(E)} \mid \mathcal{A}_Q \mid^2 \left[1 - \frac{K'}{K}\cos\phi\right] \times \left| 1 \pm \frac{m_t}{\hbar Q}\frac{d\omega_Q}{dQ}\left(1 - \frac{K'}{K}\cos\phi\right)\right|^{-1}\;, \qquad (85)$$

E', K', and Q being defined by Eqns. (80)-(82) as functions of the scattering angle ϕ.

D. Electron-phonon scattering

Other additional scattering mechanisms which contribute to the total relaxation rate are intra- and intervalley phonon emission and absorption processes and scattering with roughness at Si/SiO_2 interface. We treat them as described elsewhere[31,32]. In particular, intravalley scattering with acoustic phonons is described by an anisotropic process. Using the elastic and equipartion approximation, the relaxation rate for a transition from an electron of wave vector $\mathbf{K}$ in subband μ to any state in subband ν assisted by an acoustic phonon of branch i = transverse (TA) or longitudinal (LA), is:

$$\frac{1}{\tau^{(i,\pm)}_{p,x;\mu\nu}(K,\beta)} \;=\; \frac{k_B T}{2\pi\rho\hbar^3 c_i^2}\; \theta(E - E_\nu) \times [1 - 2\alpha(E - E_\nu) - 2\alpha < E_\nu - V >_\nu] \times \int_{-\infty}^{\infty} \frac{dq_z}{2\pi} \mid F_{\mu\nu}(q_z) \mid^2 \int_0^{2\pi} d\phi\; \frac{\Xi_i^2(\eta_Q)\left[1 - \frac{K'}{K}\cos\phi\right]}{\cos^2\beta'/m_{x,\nu} + \sin^2\beta'/m_{y,\nu}} \qquad (86)$$

where c_i is the longitudinal (i=LA) or transverse (i=TA) sound velocity, ρ is the Si density, $\phi = \beta - \beta'$ is the scattering angle, β (β') the angle between the initial (final) wave vector $\mathbf{K}$ ($\mathbf{K'}$) and the x-axis, the magnitude of the final wave vector $\mathbf{K'}$ is given by Eq. (81) with $E = E'$, thanks to the elastic approximaton, and, finally, the 'deformation potential' Ξ_i is the anisotropic expression obtained by Herring and Vogt[38]

$$\Xi_i(\eta_Q) \;=\; \left\{ \begin{array}{ll} \Xi_d \;+\; \Xi_u \cos^2\eta_Q & (i = \mathrm{LA}) \\ \Xi_u \cos\eta_Q \sin\eta_Q & (i = \mathrm{TA}) \end{array} \right.\;, \qquad (87)$$

Ξ_u and Ξ_d being the uniaxial-shear and dilatation deformation potentials, respectively, and η_Q is the angle between the emitted/absorbed phonon and the longitudinal axis of the ellipsoidal equi-energy surface of the valley, that is:

$$\cos\eta_Q \;=\; \frac{K\cos\beta - K'\cos\beta'}{(Q^2 + q_z^2)^{1/2}}\;. \qquad (88)$$

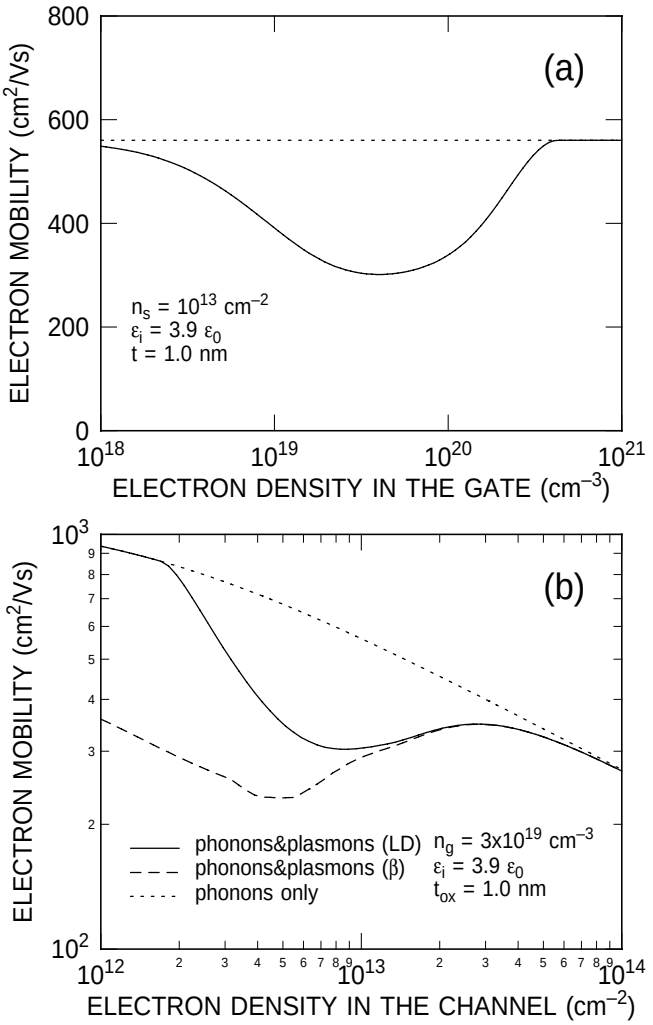

FIG. 11. Effective phonon- and plasmon-limited electron mobility at 300 K – using the single-subband and long-wavelength limit for the dielectric functions – as a function of electron density in the gate at a fixed electron density in the channel (a) and in the opposite case, as a function of electron density in the channel for a fixed density in the gate (b). The dotted lines represent the phonon-limited mobility in this model, the solid line has been calculated damping the interface phonons at the zero-T Landau-damping parameter q_{LD} in the gate, the dashed lines damping the plasmons at the Thomas-Fermi screening wave vector $\beta = \beta_{s,3D}$ in the gate.

Finally,

$$F_{\mu\nu}(q_z) \;=\; \int_{t_{ox}}^{\infty} dz\; \zeta_\mu(z)\; e^{iq_z z}\zeta_\nu(z) \qquad (89)$$

is the electronic form factor.

Intervalley phonon scattering is approximated as an isotropic process, so that the momentum relaxation rate coincides with the scattering rate, given by Eq. (31) of Ref. 32. Scattering with interface roughness is assumed to be an elastic process, so the associated momentum relaxation time will be similar to the scattering rate given by Eq. (49) of Ref. 32, with an extra factor $1 - (K'/K)\cos\phi$ appearing in the integrand. The parameters employed to treat these scattering processes will be discussed in the following section.

E. Numerical results

It is convenient to consider first the simplified model of Eq. (85), using also a simplified form of Eq. (86) with an isotropic deformation potential, $\Xi_i(\eta_q) = \Delta_{ac} = 10$ eV for both transverse and logitudinal acoustic phonons. In this case the momentum relaxation time due to electron-phonon scattering becomes simply $3m_t b\Delta_{ac}^2 k_B T/(64\hbar^3\rho c_s^2)$, where $c_s = 5.6\times10^5$ cm/s is an average sound velocity, and the mobility reduces to the form (68) with a single subband ($\mu = 0$), amenable to a fast numerical evaluation.

Figure 11(a) shows the electron mobility for a fixed carrier concentration in the channel (10^{13} cm^{-2}), at 300 K for an oxide thickness of 1 nm as a function of electron concentration in the Si gate, n_g. Two curves are shown in the figure, corresponding to two different choices for the Landau damping wave vector q_c. A first possible choice, corresponding to the solid lines in Figs. 11, 12, and 13, is the zero-temperature Landau-damping parameter q_{LD}, the solution of $\hbar\omega_Q = \hbar\omega_{LD,g}(Q)$, where $\hbar\omega_{LD,g}(Q)$ is given by Eq. (56). Indeed, since we are effectively suppressing the insulator/substrate modes via the factor R_Q given by Eq. (76), we have to worry only about gate/insulator modes decaying into single-particle excitations in the gate. A second choice, which we make only to conform to a practice commonly employed in the literature, is to approximate q_c with the screening wave vector $\beta_{s,3D}$ in the gate. This choice results in the dashed lines in the figures.

At small n_g the mobility approaches its phonon-limited value (dotted line), since the squared amplitude of the field associated with the gate/insulator interface plasmons (proportional to their frequency, in turn porportional to $n_g^{1/2}$) is small. As the concentration of the gate electrons increases, the coupling with the gate/insulator plasmons increases. If we chose $q_c = q_{LD}$, at sufficiently large n_g the frequency of the optical interface plasmons becomes too large: Electrons in the channel cannot absorb them, since their thermal occupation decreases fast with their energy, and cannot emit them once the energy of the optical modes exceeds the Fermi energy of the electrons in the channel. Acoustic modes, on the other hand, could couple to the 2DEG, but they are strongly Landau-damped. Therefore, the mobility rises again towards its phonon-limited value. However, the choice $q_c = \beta_{s,3D}$ allows undamped acoustic modes up to very short wavelengths. This results in a reduced mobility even at large values of n_g, since now the 2DEG can scatter with the undamped acoustic modes.

Figure 11(b) shows the complementary situation of the mobility calculated for a fixed n_g, chosen approximately at the minimum of the solid line in Fig. 11(a), but varying the sheet carrier density in the channel. Now the phonon-limited mobility decreases with n_s, because of the decreasing electronic form factor (89). At low n_s the Fermi energy of the 2DEG is so small that only acoustic plasmons can contribute to the mobility. When chosing $q_c = q_{LD}$ these are mostly Landau-damped and the mobility remains phonon-limited. When chosing $q_c = \beta_{s,3D}$, instead, their contribution is noticeable even at small n_s.

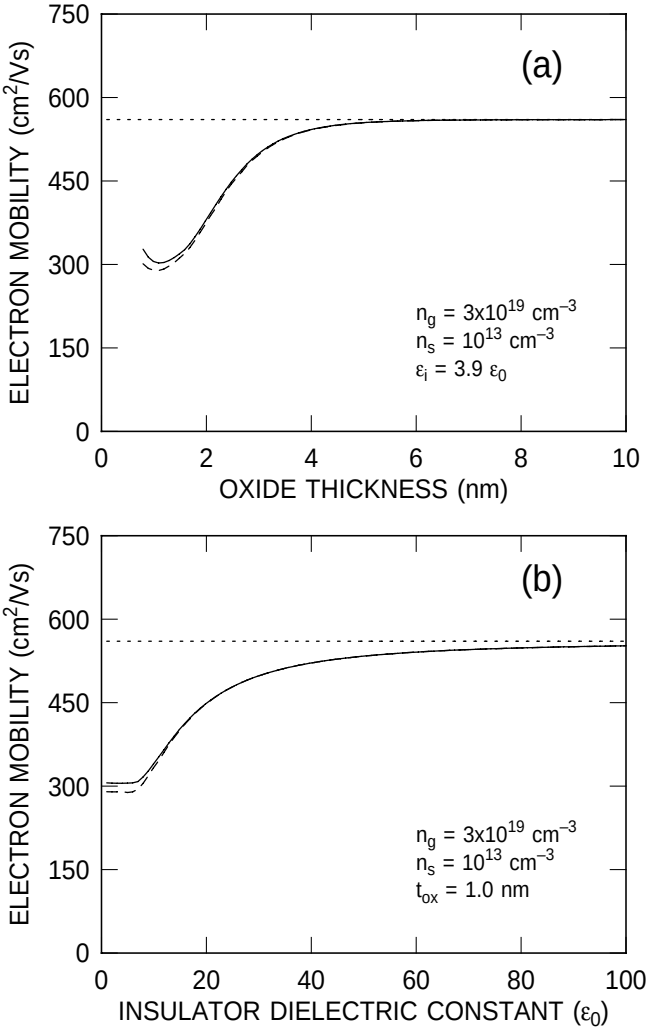

FIG. 12. The effective electron mobility as in Fig. 11, but computed at fixed electron densities in the gate and in the channel, but varying the oxide thickness (top) or the insulator dielectric constant (bottom) while maintaining the same 'oxide-equivalent' insulator thickness of 1 nm.

For large values of the electron sheet density in the channel the amplitude of the plasmon field in the channel, Eq. (78), decreases as the high-density 2DEG screens charge oscillations at the gate/insulator interface. Note that the 'dips' observed in the curves of Fig. 11 correspond to the peaks of the drag-rate calculated by Flensberg and Hu[16]. These are the signature of the plasmon-induced drag occurring whenever the mismatch between the Fermi velocities of the electron gases in the substrate and in the gate is minimized.

Figure 12 shows the dependence of the mobility at fixed n_g and n_s on the oxide thickness (a) and on the permittivity of the insulating layer for a fixed 'effective' insulator thickness $t_{ox,eq} = t_{ox}\epsilon_i^\infty/\epsilon_{SiO_2}^\infty$, (b). In Fig. 12(a) notice that the influence of interface plasmons disappears for oxides thicker than about 3.5 to 4 nm. Results for oxides thinner than about 1 nm show an increasing mobility, as acoustic plasmons become more severely damped for thin insulators. The calculations have not been pushed to the $t_{ox} \to 0$ limit since many assumptions we have made explicitly and implicitly break down in this limit. Most notably, as $t_{ox} \to 0$ and the electron

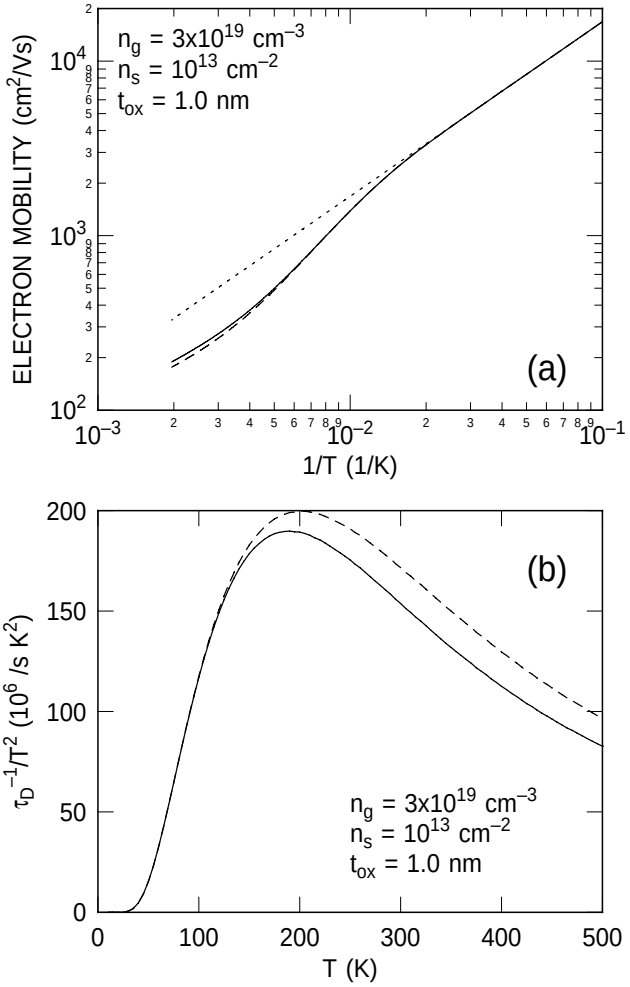

FIG. 13. **(a)**: The effective electron mobility as Fig. 11, computed as a function of temperature at fixed electron densities in the gate and in the channel. **(b)**: The Coulomb drag rate $1/\tau_D$, rescaled by T^2, as a function of temperature.

wavefunction penetrates into the oxide, an increasingly larger fraction of charge resides within the oxide layer. This will undoubltely alter the coupling between the gate and substrate electron gases and so the plasmon dispersion. Note in Fig. 12(b) how high-dielectric insulators show a beneficial effect.

Figure 13(a) shows the temperature dependence of the effective electron mobility in this simple one-subband model. As expected, the phonon-limited mobility increases linearly with $1/T$, while the effect of the interface plasmons disappears at low temperatures. This is illustrated in more detail in Fig. 13(b). Here we plot the 'drag rate' $1/\tau_D$ (defined as $1/\tau_D = e/(\mu_P m_t)$, where μ_P is the plasmon-limited mobility of Eq. (68)) rescaled by the squared temperature, as usually done in the literature[10,16]. The plot shows the 'usual' peak at temperatures slighlty lower than the Fermi temperatures T_F of the 2DEG in the substrate ($\approx$ 256 K at the density employed in the figure) or of th 3DEG in the gate ($T_F \approx$ 725 K). The T^2-dependence of the drag-rate, observed experimentally[8] in some systems at low temperature ($T < T_F$), is reproduced only qualitatively, in agreement with previous theoretical results accounting for long-range (plasmon) effects[16]. The behavior of

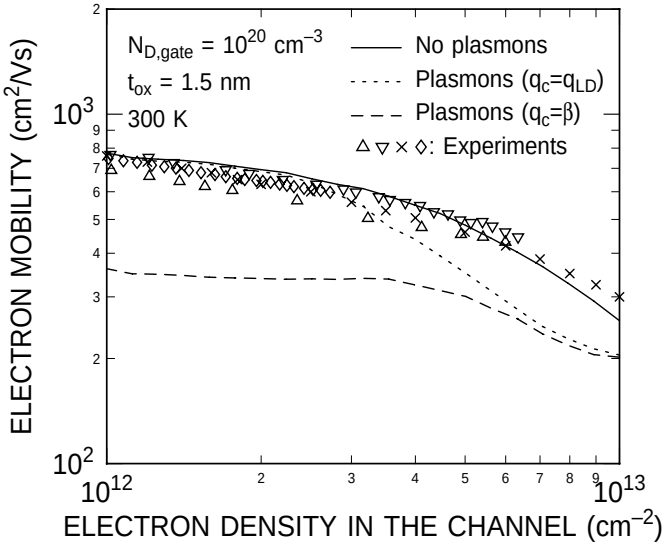

FIG. 14. Effective electron mobility in Si channels as a function of electron sheet density while also varying the average electron density in the gate depletion layer as required by Gauss' law. Scattering with intra- and inter-valley phonons, surface roughness, and interface plasmons has been included. The solid line shows the mobility in the thick-oxide limit (no interface plasmons), the dotted line accounting for interface plasmons damped at the wave vector q_{LD}, the dashed line damping them at the screening wave vector in the gate. Experimental data for lightly-doped substrates and thick oxides are as in Fig. 12 of Ref. 32.

$1/\tau_D$ at very low temperatures agrees with the results presented in Fig. 4 of Ref. 16 when using the plasmon-pole approximation. Indeed, we have not accounted for short-range Coulomb effects, being interested mainly in the behavior of the system at room temperature. But, like for all other Coulomb scattering processes, at sufficiently low T short-range electron-electron interactions cannot be ignored, since the origin of the the T^2-dependence is Pauli's exclusion principle acting on these processes.

Finally, Fig. 14 shows the electron mobility as a function of sheet carrier density n_s while also varying the electron density in the gate according to Eqns. (37)-(43). An oxide thickness of 1.5 nm has been assumed and the poly-Si gate has been taken to be n-type doped to a concentration of 10^{20} donors/cm^3. In this case, we have still used the long-wavelength limit, but have accounted for the occupation of many subbands, have employed the anisotropic expressions (63), (77) and (86), and have treated scattering with interface roughness with the model described in Ref. 32. One should keep in mind that, in Fig. 14, as we move along the n_s-axis we move also in the direction of decreasing n_g, as the poly-Si gate becomes more and more depleted. Roughly speaking, we walk through Fig. 11(b) right-to-left while walking along Fig. 11(a) in the opposite direction.

The first observation we must make is that the phonon-limited mobility itself (solid line) reproduces satisfactorily experimental data for thick oxides and lightly-doped substrates (symbols as in Fig. 12 of Ref. 32). This is the result of a choice of the deformation potentials, both intra- and inter-valley, different from Ref. 32. Previous

work on strained Si and on the calculation of electron and hole mobility in bulk Si has allowed us to obtain the values $\Xi_d = 1.1$ eV and $\Xi_u = 10.5$ eV for the dilatation and uniaxial deformation potentials.[39] Similarly, the high electron mobilities observed in strained Si have dictated the use of the intervalley deformation potential originally employed by Canali and co-workers[40]. It is no surprise that results regarding the electron mobility in strained Si should bear implications on the mobility in Si inversion layer: In both cases, the degeneracy of the six ellipsoidal valleys is removed, either by strain or by quantization, thus allowing a separate determination of the f- and g-scattering intervalley processes.

When scattering with interface plasmon is accounted for, the mobility is significantly depressed at large n_s. At smaller electron sheet densities the Fermi level of the 2DEG decreases, so that electrons in the channel can interact only with the acoustic branch of the interface plasmons. Similarly to what is found in Fig. 11(b), when strongly damping the acoustic branch (dotted line, $q_c = q_{LD}$), the mobility approaches its phonon-limited value. When leaving the acoustic branch largely undamped (dashed line, $q_c = \beta_{s,3D}$), the effective electron mobility remains low even at small n_s.

IV. DISCUSSION AND CONCLUSIONS

Figure 14, the main result of our work, shows that long-range Coulomb interactions between electrons in the Si channel and electrons in the depletion layer of the gate affect significantly the electron mobility for thin oxides. For sheet electron densities of technological importance (often exceeding 10^{13} cm^{-2}), this can be a 30% effect for 1.5 nm oxides, independent of the choice of q_c. At smaller densities the 'truth' will probably lie between the dashed and the dotted lines of Fig. 14. On the one hand, the RPA calculations show that the choice $q_c = q_{LD}$ may be too restrictive when using the long-wavelength limit. The choice $q_c = \beta_{s,3D}$, however, may err in the other direction. Recall also, as discussed above, that we have completely ignored the effect of the oscillations peaked at the channel/oxide interface, and this may result in an overestimation of the mobility. Sadly, despite the complexity of our work, we are not able to provide an unambigous result at small electron sheet densities.

A degradation of the electron mobility in thin-oxide MOS structures has been reported by several groups, but it has been attributed to alternative scattering processes: Chin and co-workers[41] have observed a reduction of the effective electron mobility by a factor of 2 even at low n_s going for $t_{ox} = 7$ nm to $t_{ox} = 2$ nm. Although the authors themselves warn us that a portion of this reduction may be only 'apparent', originating from an erroneous evaluation of the charge density n_s, they proposed remote scattering with impurities in the depletion region of the poly-Si gate as a possible mechanism. Calculations based on the proposal have been made by Krishnan and co-workers[42], using what they call the Remote Coulomb Scattering (RCS) model, prompting more experimental work showing an even more dramatic reduction of the mobility in the 1.4 to 1.0 nm range[43]. More recently, Yang and colleagues[44] have seen more moderate effects (a 15%-to-20% mobility reduction at large n_s going from a 3.2 nm to 1.5 nm-thick oxide). Their samples had a rather large substrate doping, so that low-n_s effects, if present at all, would be masked by Coulomb scattering with ionized impurieties. They went on to show that the RCS model cannot account for this variation, giving at most 5% effect, and blame enhanced scattering with surface roughness for the degradation of the mobility. Finally, Timp and co-workers[45], have observed a degradation of the saturated drain current in n-MOSFETs with oxides both thicker and thinner than an 'optimum' value of about 1.3 nm. While the poorer performance of devices with a thicker oxides can be easily understood, at least qualitatively, in terms the obvious reduction of the gate capacitance, they did not explain the loss of performance observed for oxides thinner than 1.3 nm.

Our (obviously biased) opinion is that the Coulomb channel-gate interaction can easily explain those findings. Indeed neither of the proposed mechanisms (RCS or enhanced scattering with surface roughness) seems too plausible. Scattering with interface roughness can easily be confused for the effect shown in Fig. 14, especially if we have to believe the $q_c = q_{LD}$-results (dotted line), and 'fitting' the data with any surface-roughness model does not necessarily prove that this is the 'right' mechanism: From a practical point of view, two empirical parameters (the rms step-height Δ and the correlation length λ) must be 'fitted' to the data in order to evaluate the roughness-limited mobility, thus 'explaining' almost any high-n_s mobility degradation by definition. In addition, theoretical estimates do not justify this claim: the dependence of the roughness-limited mobility on t_{ox}, while strong at small n_s, becomes negligible at the electron sheet densities of interest[46]. On the other hand, the RCS model, as considered either in Ref. 42 or in Ref. 44, is also likely to overestimate its effect on the mobility: Indeed, both versions of the model assume that the scattering centers in the depletion region of the gate are screened by the 2DEG, but not by the gate electrons. In typical cases, the gate 'depletion' region is not really 'depleted', strictly speaking: It simply exhibits an electron concentration smaller than what charge-neutrality would demand, and should be viewed as a quite thin region with a significant 'average' density of free carriers, n_g, so that its screening effects – as accounted for by the Green's function $\tilde{\mathcal{G}}_Q^{(g)}$ of Appendix B of Ref. 1 – should depress significantly the strength of the interaction.

In conclusion, there is no experimental evidence clearly supporting the oxide-thickness dependence of the electron mobility. There is 'circumstantial' evidence, but inconsistent in terms of the size of the effect (factors of 2 versus 15%) and the range of n_s over which the effect is most sensitive. Nevertheless, the results of Yang and co-workers[44] are consistent with our results: Considering the large amount of impurity-scattering present in the samples of Ref. 44 (which partially masks the effect we are interested in since it leads to an increasing impurity-limited mobility for decreasing t_{ox}), the observed 15-20% effect compares favorably with the 20-25%-effect predicted in Fig. 14 at large n_s for $t_{ox} = 1.5$ nm.

ACKNOWLEDGMENTS

A. B. Fowler and S. E. Laux have provided support and encouragement during the course of this work. Interesting and stimulating discussions with C. Hu and M. Krishnan regarding the RCS-model are also deeply appreciated.

APPENDIX A: DIELECTRIC FUNCTION OF A TWO-DIMENSIONAL ELECTRON GAS

In this appendix we derive the dielectric response of a two-dimensional electron gas (2DEG). Let us consider an external potential $\phi^{(ext)}_{Q,\omega}(z)e^{i\mathbf{Q}\cdot\mathbf{R}}e^{i\omega t}$ applied to the 2DEG. The total potential will be

$$\phi_{Q,\omega}(z) \;=\; \phi^{(ext)}_{Q,\omega}(z) \;+\; \int dz'\; G^{(s)}_Q(z,z')\;\frac{\delta\rho_{Q,\omega}(z')}{\epsilon^\infty_s}\;, \tag{A1}$$

where ϵ^∞_s is the valence-band permittivity of the semiconductor, $G^{(s)}_Q(z,z')$ the Green's function derived in Appendix B of Ref. 1, and $\rho_{Q,\omega}(z)$ is the polarization charge induced by the external potential. Assuming that all Fourier components of the potential can be considered independently, since their phases vary wildly so that 'cross terms' mixing different wavelengths cancel ('random phase approximation, RPA), and that the perturbation $\phi^{(ext)}_{Q,\omega}$ is 'weak' compared to the free electron Hamiltonian, the polarization charge can be evaluated by computing the change induced by the full (external *plus* polarization) potential on the 'unperturbed' wavefunctions. For a given subband μ, these are written as

$$\psi^{(0)}_{\mathbf{K},\mu}(\mathbf{R},z) = e^{i\mathbf{K}\cdot\mathbf{R}}\;\zeta_\mu(z)\;\upsilon_{\mathbf{K},k_z=k_{z,0}}(\mathbf{R},z)\;, \tag{A2}$$

where $\zeta_\mu(z)$ is the 'envelope' wave function along the quantization direction z, $\upsilon_{\mathbf{K},k_z}(\mathbf{R},z)$ are the (bulk) Bloch functions evaluated at the wave vector $(\mathbf{K}_0+\mathbf{K}, k_{z,0}+k_z)$, where $(\mathbf{K}_0,k_{z,0})$ is the minimum of the valley we are considering. Second-order perturbation theory gives for the perturbed (retarded) wavefunctions

$$\psi_{\mathbf{K},\mu}(\mathbf{R},z) = \psi^{(0)}_{\mathbf{K},\mu}(\mathbf{R},z) + \lim_{s\to 0^+}\sum_{\mathbf{K}'\mu'}\;\frac{e<\mathbf{K}'\mu'\mid\phi_\omega\mid\mathbf{K}\mu>}{E_\mu(\mathbf{K})-E_{\mu'}(\mathbf{K}')+\hbar\omega+i\hbar s}\;\psi^{(0)}_{\mathbf{K}',\mu'}(\mathbf{R},z)\;, \tag{A3}$$

where

$$<\mathbf{K}'\mu'\mid\phi_\omega\mid\mathbf{K}\mu> \;=\; \phi_{\mathbf{K}-\mathbf{K}',\omega;\mu\mu'} \;=\; \int dz\;\zeta_\mu(z)\;\phi_{\mathbf{K}-\mathbf{K}',\omega}(z)\;\zeta_{\mu'}(z)\;. \tag{A4}$$

Therefore the charge induced by the perturbation will be:

$$\delta\rho_\omega(\mathbf{R},z) \;=\; -\;e\;\sum_{\mathbf{K}\mu} g_\mu\;f_\mu(\mathbf{K})\;[\mid\psi_{\mathbf{K},\mu}(\mathbf{R},z)\mid^2\;-\;\mid\psi^{(0)}_{\mathbf{K},\mu}(\mathbf{R},z)\mid^2]\;, \tag{A5}$$

where $f_\mu(\mathbf{K})$ is the (Fermi-Dirac) equilibrium occupation of the subband μ at the wave vector $\mathbf{K}$, and g_μ is the degeneracy of the subband including spin. Assuming now that the external potential is also 'slowly varying', so that all wave vectors $\mathbf{K}$ of interest are much smaller than any wave vector $\mathbf{G}$ of the reciprocal lattice, using Eq. (A2) and (A3), we obtain:

$$\delta\rho_\omega(\mathbf{R},z) \;=\; -\;e^2\sum_{\mathbf{Q}} e^{i\mathbf{Q}\cdot\mathbf{R}}\lim_{s\to 0^+}\sum_{\mathbf{K}\mu\mu'}\;\frac{g_\mu f_\mu(\mathbf{K})-g_{\mu'}f_{\mu'}(\mathbf{K}+\mathbf{Q})}{E_\mu(\mathbf{K})-E_{\mu'}(\mathbf{K}+\mathbf{Q})+\hbar\omega+i\hbar s}\times\phi_{\mathbf{Q},\omega;\mu\mu'}\;\zeta_\mu(z)\;\zeta_{\mu'}(z), \tag{A6}$$

$E_\mu(\mathbf{K})$ being the energy in subband μ at wave vector $\mathbf{K}$. Thus, for the Fourier component $\delta\rho_{Q,\omega}(z)$ in Eq. (A1) we obtain:

$$\delta\rho_{Q,\omega}(z) \;=\; \epsilon^\infty_s\sum_{\mu\mu'}\;\beta_{\mu\mu'}(\mathbf{Q},\omega)\;\phi_{Q,\omega;\mu\mu'}\zeta_\mu(z)\zeta_{\mu'}(z)\;, \tag{A7}$$

having defined the 'screening wave vector' $\beta_{\mu\mu'}(\mathbf{Q},\omega)$ as:

$$\beta_{\mu\mu'}(\mathbf{Q},\omega) = -\lim_{s\to 0^+}\;\frac{e^2}{\epsilon^\infty_s}\;\sum_{\mathbf{K}}\;\frac{g_\mu f_\mu(\mathbf{K})-g_{\mu'}f_{\mu'}(\mathbf{K}+\mathbf{Q})}{E_\mu(\mathbf{K})-E_{\mu'}(\mathbf{K}+\mathbf{Q})+\hbar\omega+i\hbar s}\;. \tag{A8}$$

Equation (A1) should be regarded as an equation for the screened potential $\phi_{Q\omega}(z)$ once the unscreened potential $\phi^{(ext)}_{Q,\omega}(z)$ is known. In particular, we need only the matrix elements (*i.e.*, the 'screened form factors') $\phi_{Q,\omega;\mu\mu'}$, which can be obtained by taking the matrix elements of Eq. (A1) over the basis wavefunctions $\{\zeta_\lambda\}$, obtaining:

$$\phi_{Q,\omega;\lambda\lambda'} \;=\; \phi^{(ext)}_{Q,\omega;\lambda\lambda'} \;+\; \sum_{\mu\mu'}\;\mathcal{G}^{(s)}_{Q;\lambda\lambda';\mu\mu'}\beta_{\mu\mu'}(Q,\omega)\;\phi_{Q,\omega;\mu\mu'}\;, \tag{A9}$$

where

$$\mathcal{G}^{(s)}_{Q;\lambda\lambda';\mu\mu'} \;=\; \int dz\int dz'\zeta_\lambda(z)\zeta_{\lambda'}(z)\;G^{(s)}_Q(z,z')\;\zeta_\mu(z')\zeta_{\mu'}(z')\;. \tag{A10}$$

Equation (A9) is a matrix equation (in the double indices $(\mu\mu')$). Inversion of the dielectric matrix $\mathbf{1}-\mathbf{\Pi}$, where the polarizability $\mathbf{\Pi}$ is $\Pi_{\lambda\lambda';\mu\mu'}(Q,\omega) = \mathcal{G}^{(s)}_{Q;\lambda\lambda';\mu\mu'}\;\beta_{\mu\mu'}(Q,\omega)$, provides directly the desired screened form factors:

$$\phi_{Q,\omega;\lambda\lambda'} \;=\; \sum_{\mu\mu'}\;[\mathbf{1}-\mathbf{\Pi}(Q,\omega)]^{-1}_{\lambda\lambda';\mu\mu'}\;\phi^{(ext)}_{Q,\omega;\mu\mu'}\;. \tag{A11}$$

[1] M. V. Fischetti and S .E. Laux, *'Long-range Coulomb interactions in small Si devices. Part I: Performance and reliability'*, submitted to J. Appl. Phys. (2000).
[2] See, for example, H. S. Momose, M. Ono, T. Yoshitomi, T. Ohguro, S.-I. Nakamura, M. Saito, and H. Iwai, Solid-State Electron. **41**, 707 (1997).
[3] These results have been already pre-announced in M. V. Fischetti and S. E. Laux, Appl. Phys. Lett. **76**, 2277 (2000).
[4] M. B. Pogrebinskii, Fiz. Tekh. Poluprovodn. **11**, 637 (1977) [Sov. Phys. Semicond. **11**, 372 (1977)].
[5] P. J. Price, Physica B & C, **117**, 750 (1983).
[6] I. I. Boiko and Yu. N. Sirenko, Zh. Tekh. Fiz. **58**, 967 (1988) [Sov. Phys. – Tech. Phys. **33**, 586 (1988)].
[7] B. Leikhtman and P. M. Solomon, Phys. Rev. B **41**, 9921 (1990).
[8] T. J. Gramila, J. P. Eisenstein, A. H. MacDonald, L. N. Pfeiffer, and K. W. West, Phys. Rev. Lett. **66**, 1216 (1991).
[9] L. Zheng and A. H. MacDonald, Phys. Rev. B **48**, 8203 (1993).
[10] A.-P. Jauho and H. Smith, Phys. Rev. B **47**, 4420 (1993).
[11] K. Flensberg, B. Y.-K. Hu, A.-P. Jauho, and J. M. Kinaret, Phys. Rev. B **52**, 14761 (1995).
[12] See also the review article by A. G. Rojo, J. Phys.: Condens. Matter **11**, R31 (1999) for additional references.
[13] C. Jacoboni and P. J. Price, Solid-State Electron. **31**, 649 (1988).
[14] P. M. Solomon, P. J. Price, D. J. Frank, and D. C. L. Tulipe, Phys. Rev. Lett. **63**, 2508 (1989).
[15] B. Tanatar and A. K. Das, Phys. Rev. B **61**, 15959 (2000).
[16] K. Flensberg and B. Y.-K. Hu, Phys. Rev. B **52**, 14796 (1995).
[17] D. Bohn and D. Pines, Phys. Rev. **92**, 609 (1953).
[18] D. Pines, Rev. Mod. Phys. **28**, 184 (1956).
[19] This is indeed what we found in the semiclassical work presented in Ref. 1: At least at high drain-to-source bias, short-range electron-electron collisions do not alter the characteristics of the devices, even at the small oxide thicknesses.
[20] See Ref. 32 and references therein.
[21] See references cited in Ref. 31
[22] E. N. Economou, Phys. Rev. **182**, 539 (1969).
[23] K. L. Ngai and E. N. Economou, Phys. Rev. B **4**, 2132 (1971).
[24] D. A. Dahl and L. J. Sham, Phys. Rev. B **16**, 651 (1977).
[25] A. Eguiluz, T. K. Lee, J. J. Quinn and K. W. Chiu, Phys. Rev. B **11**, 4989 (1975).
[26] D. Pines, Phys. Rev. **92**, 626 (1953).
[27] P. Lugli and D. K. Ferry, Physica **129B**, 532 (1985).
[28] A. L. Fetter and J. D. Walecka, *Quantum Theory of Many-Particle Systems* (McGraw-Hill, New York, 1971), p. 305. See also Eqns. (73) and (74) of Ref. 32 below and Eqns. (A5) and (A6) in Appendix A of M. V. Fischetti, S. E. Laux, and E. Crabbé, J. Appl. Phys. **78**, 1058 (1995) for the explicit expressions we have used in the two- three-dimensional cases, respectively.
[29] P. J. Price, J. Vac. Sci. Technol. **19**, 599 (1981).
[30] See for example R. Rios, N. Arora, and C.-L. Huang, IEEE Electron Device Let. **15**, 129 (1995), from whose work one can derive Eq. (37). This equation is given explicitely by B. Yu, D.-H. Ju, W.-C. Lee, N. Kepler, T.-J. King, and C. Hu, IEEE Trans. Electron Dev. **45**, 1253 (1998), although with the implicit assumption $\epsilon_{Si}^{\infty} = 3\ \epsilon_{ox}^{\infty}$ and with the wrong exponent for the oxide field F_i.
[31] T. Ando, A. B. Fowler, and F. Stern, Rev. Mod. Phys. **54**, 437 (1982).
[32] M. V. Fischetti and S. E. Laux, Phys. Rev. B **48**, 2244 (1993).
[33] In this case the energy-conserving delta-function leading to the relaxation rate (77) should be replaced by a Lorentzian of width $\propto$ Im ω_Q. Equivalently stated, for Im $\omega_Q \neq 0$, Eq. (77) is the 'plasmon-pole approximation' of the exact result.
[34] For thicker oxides, as the coupling bewteen gate and substrate weakens, the effect of the finite thickness of the inversion layer on the optical mode disappears, while its effect on the acoustic mode has the opposite trend of reducing its frequency. In the limit of infinitely thick oxides this has been shown by D. E. Beck and P. Kumar, Phys. Rev. B **13**, 2859 (1976); Phys. Rev. B **14**, 5127(E) (1976). This work, as well as the work by A. K. Rajagopal, Phys. Rev. B **15**, 4264 (1977) and by M. Jonson, J. Phys. C **9**, 3055 (1977), also accounts for exchange-correlation effects, which we have ignored.
[35] M. V. Fischetti, Phys. Rev. B **44**, 5527 (1991).
[36] D. Chattopadhyay and H. J. Queisser, Rev. Mod. Phys. **53**, 745 (1981).
[37] P. Hawrylak, G. Eliasson, and J. J. Quinn, Phys. Rev. B **37**, 10187 (1988).
[38] C. Herring and E. Vogt, Phys. Rev. **101**, 944 (1956).
[39] M. V. Fischetti and S. E. Laux, J. Appl. Phys. **80**, 2234 (1996).
[40] C. Canali, C. Jacoboni, F. Nava, G. Ottaviani, and A. Alberigi-Quaranta, Phys. Rev. B **12**, 2265 (1975).
[41] A. Chin, W. J. Chen, T. Chang, R. H. Rao, B. C. Lin, C. Tsai, and J. C.-M. Huang, IEEE Electron Device Let. **18**, 417 (1997).
[42] M. Krishnan, Y. C. Yeo, Q. Lu, T.-J. King, J. Bokor, and C. Hu, Int. Electron Dev. Meet. Tech. Dig. 571 (1998).
[43] M. Krishnan, L. Chang, T.-J. King, J. Bokor, and C. Hu, Int. Electron Dev. Meet. Tech. Dig. 241 (1999).
[44] N. Yang, W. Kirklen Henson, J. R. Hauser, and J. J. Wortman, IEEE Trans. Electron Devices, **47**, 440 (2000)
[45] G. Timp, A. Agarwal, F. H.. Baumann, T. Boone, M. Buonanno, R. Cirelli, V. Donnelly, M. Foad, D. Grant, M. Green, H. Gossmann, S. Hillenius, J. Jackson, D. Jacobson, R. Kleiman, A. Kombilt, F. Klemens, J. T. C. Lee, W. Mansfield, S. Moccio, A. Murrell, M. O'Mally, J. Rosamilia, J. Sapjeta, P. Silverman, T. Sorch, W. W. Tai, D. Tennant, H. Vuong, and B. Weir, IEDM Tech. Dig. (1997), pp. 930-932.
[46] A. Gold, Appl. Phys. Lett. **48**, 439 (1985).