Tunneling-lifetime model for metal-oxide-semiconductor structures

M. Ali Pourghaderi,* Wim Magnus,† Bart Sorée, Marc Meuris, Kristin De Meyer,‡ and Marc Heyns‡
IMEC, Kapeldreef 75, B-3001 Leuven, Belgium

(Received 15 May 2009; revised manuscript received 27 June 2009; published 21 August 2009)

In this paper we investigate the basic physics of charge carriers (electrons) leaking out of the inversion layer of a metal-oxide-semiconductor capacitor with a biased gate. In particular, we treat the gate leakage current as resulting from two combined processes: (1) the time-dependent decay of electron wave packets representing the inversion-layer charge and (2) the local generation of "new" electrons replacing those that have leaked away. As a result, the gate current simply emerges as the ratio of the total charge in the inversion layer to the tunneling lifetime. The latter is extracted from the quantum dynamics of the decaying wave packets, while the generation rate is incorporated as a phenomenological source term in the continuity equation. Not only do the gate currents calculated with this model agree very well with experiment, the model also provides an onset to solve the paradox of the current-free bound states representing the resonances of the Schrödinger equation that governs the fully coupled metal-oxide-semiconductor system.

DOI: 10.1103/PhysRevB.80.085315 PACS number(s): 73.40.Gk, 72.20.Jv, 72.10.−d

I. INTRODUCTION

In order to enhance the performance and the integration of modern field-effect transistors their size needs to be further scaled down. As is well known, this scaling trend, which is generally expected to hold for most device geometries that contain ultrathin oxide layers, leads to an exponential growth of the gate leakage current due to tunneling.[1,2] The latter, in turn, will increase the power dissipation inside the downscaled devices to unacceptable levels. In this light, it is utterly important to acquire detailed knowledge about the origin of the leakage currents and the underlying physical mechanisms.

During the past decades, many attempts have been made to construct simple and yet sufficiently accurate models accounting for the basic features of the gate tunneling processes. Most of them fall into two categories. The first category is based on the Bardeen approach[3,4] in which the current matrix element arising from the overlap of two sets of nonorthogonal wave functions, respectively, describing the isolated gate-oxide and oxide-semiconductor regions plays a key role. Being inspired by Gamow's theory for alpha-particle emission from atomic nuclei[5] and Breit-Wigner scattering theory,[6,7] the models from the second category[8–11] rely on a continuum of eigenstates describing the stationary states of the coupled gate-oxide-semiconductor system. The continuum states include a discrete subset of resonant states representing the quasibound states that are populated by electrons with a significant tunneling probability amplitude. Denoting, respectively, by $\psi_{\rm res}(\mathbf{r})$ and m the wave function of an arbitrary resonant state and the effective mass in the leakage direction, the quantum-mechanical current density carried by $\psi_{\rm res}(\mathbf{r})$ reads

$$\mathbf{J}_{\rm res}(\mathbf{r}) = -\frac{e\hbar}{m}{\rm Im}[\psi^*_{\rm res}(\mathbf{r})\nabla\psi_{\rm res}(\mathbf{r})]. \tag{1}$$

If the resonant energies exceeded $U_{\rm S}$, the constant potential energy of the substrate's neutral region, the wave functions $\psi_{\rm res}(\mathbf{r})$ would asymptotically behave like plane waves propagating toward either the substrate or the gate region. Accordingly, the energy spectrum would be doubly degenerate while their complex wave functions would represent traveling states. However, when the device operates in inversion mode, all occupied energy levels are below $U_{\rm S}$ and correspond to nondegenerate evanescent modes vanishing exponentially in the substrate. In particular, the resonant states are populated by electrons that spend a lot of time in the inversion-layer potential well. As such, these states are quasibound states, whereas the associated wave functions can be taken real due to the lacking degeneracy. Consequently, $\psi_{\rm res}(\mathbf{r})$ can be taken real and hence the corresponding resonant state cannot carry a tunneling current as $\mathbf{J}_{\rm res}(\mathbf{r})$ is seen to vanish.[12] Because of this paradox, the real gate current cannot be written as a statistically weighed sum over all individual tunneling current contributions associated with stationary resonant states.

In this work, we have attempted to resolve the above paradox by abandoning the stationary character of the global leakage process, studying the time-dependent tunneling decay of a charge packet initially and locally injected into the inversion layer, and invoking local carrier generation as a proper refilling mechanism to replace the leaked-out charge. The time-dependent decay of the inversion-layer charge package, including the extraction of tunneling lifetime is treated for a metal-oxide semiconductor capacitor in Sec. II, whereas three practical methods to describe uncoupled or localized inversion-layer electrons are discussed in Sec. III. A simple but consistent model representing the generation-based refilling mechanism is proposed in Sec. IV. Finally, some typical results obtained from the gate current model are presented and discussed as well as compared with related formalisms in Sec. V and the paper is concluded in Sec. VI.

II. DECAY OF THE INVERSION-LAYER CHARGE

We consider a metal-oxide-semiconductor (MOS) capacitor in which an n channel (p-type semiconductor) is induced by virtue of a positive voltage applied at the metallic gate side. The z axis is taken to be perpendicular to the three capacitor layers such that the plane $z=0$ coincides with the

 ©2009 The American Physical Society

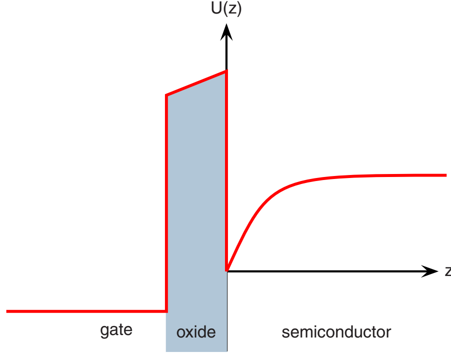

FIG. 1. (Color online) Potential-energy profile of a MOS capacitor biased with a positive gate voltage.

semiconductor/oxide interface while full translational invariance is assumed along the intervals $0 \leq x \leq L_x$ and $0 \leq y \leq L_y$ in the lateral x and y directions. From here on, $\mathbf{r} = (x, y)$ will denote the position vector in the (x, y) plane. The size of the gate and semiconductor regions is considered virtually infinite compared to t_{ox}, the thickness of the ultrathin oxide layer. The three-dimensional volume encompassing all layers of the coupled structure will be denoted by Ω whereas its semiconductor part hosting the inversion-layer potential well $z \geq 0$ will be labeled by Ω_{W}. The potential-energy diagram of the biased MOS capacitor is shown in Fig. 1. Anticipating the introduction of the generation mechanism locally compensating for the loss of inversion-layer electrons, we assume that at some initial time instant $t=0$ the inversion-layer charge is confined to the semiconductor potential well Ω_{W}. Consequently, the single-electron states populated by the corresponding electrons are not stationary energy eigenstates of the Hamiltonian governing the entire (MOS) capacitor, allowing for electron tunneling through the oxide layer. Indeed, due to the latter, electrons occupying the continuum of stationary eigenstates and, particularly, its subset of resonant states may have a significant nonzero amplitude of being found outside the inversion layer.

On the other hand, being initially in the nonstationary states, the electrons will penetrate through the barrier into the gate region as time evolves. We investigate the dynamical behavior of these electrons under the assumption that no other interactions are affecting the tunneling processes. The corresponding quantum dynamics is conveniently treated in the Heisenberg picture tracing the time evolution of the electron field operator $\psi(\mathbf{r}, z, t)$ and convenient creation and annihilation operators. At an arbitrary time t, the electron field operator $\psi(\mathbf{r}, z, t)$ can be expanded in a complete set of continuum states $\phi_k(\mathbf{r}, z)$ representing the one-electron energy eigenstates for the fully coupled system Ω consisting of the gate, oxide, and well regions

$$\psi_\sigma(\mathbf{r}, z, t) = \int dk c_{k\sigma}(t) \phi_k(\mathbf{r}, z), \tag{2}$$

where $c_{k\sigma}^\dagger$ and $c_{k\sigma}$ are fermion creation and annihilation operators for the stationary continuum state $\phi_k(\mathbf{r}, z)$ with energy E_k and σ is a spin index. Clearly, the wave number k is a continuous variable characterizing $\phi_k(\mathbf{r}, z)$ as a standing wave in the gate region ($z < -t_{\mathrm{ox}}$). However, the integral over k should be understood to include also a summation over the two-dimensional wave vectors appearing in the plane waves

$$\frac{1}{\sqrt{L_x L_y}} \exp(i\mathbf{k} \cdot \mathbf{r}) \tag{3}$$

that represent the $\mathbf{r}$-dependent part of the wave functions. Presently, the envelope function approach is adopted and, consequently, it is assumed that the "parallel" wave vectors are being conserved during the barrier tunneling processes, whereas the energy eigenvalues E_k can be written as the sum of the kinetic energy related to the lateral motion and a term W_k describing the perpendicular motion that does not depend on the wave vector $\mathbf{k}$. This restriction may, however, be relaxed to generalize the formalism whenever appropriate.

The second-quantized Hamiltonian governing the time evolution of the coupled system Ω reads

$$\hat{H} = \sum_\sigma \int dk E_k c_{k\sigma}^\dagger c_{k\sigma}, \tag{4}$$

from which the trivial time dependence of the annihilation operators can be extracted

$$c_{k\sigma}(t) = \exp\left(-\frac{iE_k t}{\hbar}\right) c_{k\sigma}(0). \tag{5}$$

Similarly, we may construct a Hamiltonian for the uncoupled system Ω_{W} consisting merely of the inversion-layer well with an ensemble of strictly confined electrons

$$\hat{H}_0 = \sum_{n\sigma} E_{0n} c_{n\sigma}^\dagger c_{n\sigma}, \tag{6}$$

where the subband index n now labels the bound states with discrete energy eigenvalues E_{0n} and the corresponding "unperturbed" wave functions $\phi_n(\mathbf{r}, z)$ that are vanishingly small or tend to zero in the gate region. Again, it should be noted that n also runs over the set of two-dimensional wave vectors $\mathbf{k}$. Being complete only with respect to the subspace of localized well states, the set $\phi_n(\mathbf{r}, z)$ can only be used to expand the field operator at $t=0$,

$$\psi_\sigma(\mathbf{r}, z, 0) = \sum_n c_{n\sigma} \phi_n(\mathbf{r}, z). \tag{7}$$

Since the basis set in Eq. (2) can be considered orthonormal, the operators $c_{k\sigma}(0)$ behave as Fourier coefficients

$$c_{k\sigma}(0) = \int_\Omega d^3 r \phi_k^*(\mathbf{r}, z) \psi_\sigma(\mathbf{r}, z, 0). \tag{8}$$

Substituting Eq. (7) into Eq. (8), we may express $c_{k\sigma}(0)$ as the linear combination of the operators $c_{n\sigma}$

$$c_{k\sigma}(0) = \sum_n c_{n\sigma} \int_\Omega d^3 r \phi_k^*(\mathbf{r}, z) \phi_n(\mathbf{r}, z) \equiv \sum_n \Lambda_{kn} c_{n\sigma}, \tag{9}$$

with

$$\Lambda_{kn} = \int_{\Omega} d^3 r \phi_k^*(\mathbf{r},z)\phi_n(\mathbf{r},z). \tag{10}$$

Next, we introduce the operator $\hat{Q}$ counting the total electron charge in the inversion-layer well Ω_W

$$\hat{Q} = -e\sum_{\sigma}\int_{\Omega_W} d^3 r \psi_{\sigma}^{\dagger}(\mathbf{r},z)\psi_{\sigma}(\mathbf{r},z)$$
$$= -e\sum_{\sigma}\int dk' \int dk \Gamma_{k'k} c_{k'\sigma}^{\dagger} c_{k\sigma}, \tag{11}$$

where

$$\Gamma_{k'k} = \int_{\Omega_W} d^3 r \phi_{k'}^*(\mathbf{r},z)\phi_k(\mathbf{r},z) \tag{12}$$

is the overlap integral of the coupled-system eigenstates restricted to the well region. In the Heisenberg picture the time-dependent inversion-layer charge $Q(t)$ is obtained by taking the statistical average with the initial density matrix ϱ_0, while the time dependence is embedded in the Heisenberg operator

$$Q(t) = \langle \hat{Q}(t) \rangle \equiv \mathrm{Tr}[\varrho_0 \hat{Q}(t)]. \tag{13}$$

From Eq. (11) it follows that

$$Q(t) = -e\sum_{\sigma}\int dk' \int dk \Gamma_{k'k}\langle c_{k'\sigma}^{\dagger}(t) c_{k\sigma}(t)\rangle$$
$$= -e\sum_{\sigma}\int dk \int dk' \Gamma_{k'k}$$
$$\times \exp\left[\frac{i}{\hbar}(E_{k'} - E_k)t\right]\langle c_{k'\sigma}^{\dagger}(0) c_{k\sigma}(0)\rangle. \tag{14}$$

Using Eq. (9), we can rewrite the correlation function $\langle c_{k'\sigma}^{\dagger}(0) c_{k\sigma}(0)\rangle$ as

$$\langle c_{k'\sigma}^{\dagger}(0) c_{k\sigma}(0)\rangle = \sum_{nn'} \Lambda_{k'n'}^* \Lambda_{kn} \langle c_{n'\sigma}^{\dagger} c_{n\sigma}\rangle$$
$$= \sum_{n} \Lambda_{k'n}^* \Lambda_{kn} F(E_{0n} - \mu). \tag{15}$$

The latter equality results from the explicit assumption that the initially localized electrons are in a thermal equilibrium state, i.e.,

$$\langle c_{n'\sigma}^{\dagger} c_{n\sigma}\rangle = \delta_{n'n} F(E_{0n} - \mu), \tag{16}$$

where $F(E)$ is the Fermi-Dirac distribution function with chemical potential μ. Combining Eqs. (14) and (15), we arrive at the final expression for the time-dependent inversion-layer charge

$$Q(t) = -2e\int\int dk' dk M_{k'k} \exp\left[\frac{i}{\hbar}(E_{k'} - E_k)t\right], \tag{17}$$

with

$$M_{k'k} = \left[\sum_{n} \Lambda_{k'n}^* \Lambda_{kn} F(E_{0n} - \mu)\right]\Gamma_{kk'}. \tag{18}$$

In practice, the numerical algorithm extracting the time dependence of $Q(t)$ from Eq. (17) involves the following basic steps: (1) to solve self-consistently the Schrödinger and Poisson equations to get the stationary wave functions $\phi_k(\mathbf{r},z)$ of the coupled system, including the subset of resonant states, for a given gate voltage V_G; (2) to keep the potential profile of the coupled system and to solve the Schrödinger equation for an isolated inversion layer yielding the subband wave functions $\phi_n(\mathbf{r},z)$ and their energies E_{0n}; (3) to calculate the "memory matrix" $M_{k'k}$ by combining the results of the two previous steps and to extract $Q(t)$ from Eq. (17); and (4) to determine the tunneling lifetime by fitting $Q(t)$ against a purely exponential decay law.

Some remarks are in order. First, there are several ways to decouple the inversion-layer charge artificially from the other regions of the MOS capacitor. Three such approaches are discussed in the next section. Next, it should be noted that the computational effort required to obtain the subband structure is rather moderate because the resonant energies obtained from step (1) are sufficiently close to the subband energies of the decoupled electrons (except for very leaky structures).[8–10] Hence, they provide an excellent initial guess for the subband energy spectrum to be determined in step (2). Finally, the exponential decay which is conjectured in the last step to describe adequately the long-time tail of $Q(t)$ is not a general quantum-mechanical result, as pointed out clearly by Merzbacher[13] in a section on decay phenomena. In the present case, it can only be justified when the continuum resonances are so sharply peaked around the resonant energies E_{res} that their amplitudes can be accurately represented by distinguished bell-shaped curves (Lorentzians)[8–10] of the form $[(E-E_{res})^2+\Gamma_{res}^2]^{-1}$, where Γ_{res} is half the width of the resonance E_{res}. This condition is fulfilled—again—in case the real oxide layer is not extremely thin ($t_{ox} < 1$ nm).

As a final remark, we briefly discuss the boundary conditions obeyed by the continuum states. Since we restrict the present work to the case of MOS capacitors operating in inversion mode, we may fairly state that, within the continuous energy spectrum of the coupled system, no traveling states are occupied. However, as was mentioned in the introduction, the continuum states and the subset of resonant states are evanescent modes giving rise to the following boundary conditions for the corresponding wave functions in the deep substrate:

$$\phi_k(\mathbf{r},z) \to e^{-\alpha_k z} \quad \text{for } z \to \infty. \tag{19}$$

Within the envelope function approximation, the attenuation factor α_k reads

$$\alpha_k = \frac{\sqrt{2m(U_S - W_k)}}{\hbar}. \tag{20}$$

The gate region may be viewed as the interval $[-t_{ox}-L, -t_{ox}]$ where "closed" (Dirichlet) boundary conditions are imposed on the left side,

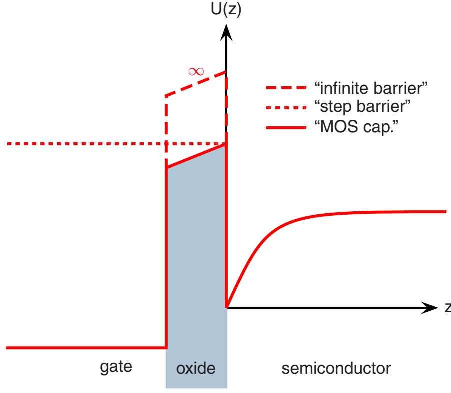

FIG. 2. (Color online) The potential profiles corresponding to three localization schemes. The first and second schemes, respectively, correspond to an infinite-barrier height and an infinite-barrier thickness while the third scheme exploits the potential profile of the whole MOS capacitor.

$$\phi_k(\mathbf{r},z=-t_{\mathrm{ox}}-L)=0. \tag{21}$$

In practice, having taken the limit $L\rightarrow\infty$, we end up with a continuous energy spectrum the wave functions of which are delta normalized according to

$$\langle\phi_{k'}|\phi_k\rangle=\int_\Omega d^3r\,\phi^*_{k'}(\mathbf{r},z)\phi_k(\mathbf{r},z)=\delta(k'-k). \tag{22}$$

As outlined in Refs. 8–10, the transfer matrix method provides a convenient numerical implementation of the continuum wave functions and, particularly, their boundary conditions, since its piecewise constant potential profile incorporates both the gate region and the deep substrate in a most natural way.

III. LOCALIZATION SCHEMES

Considering carrier localization as a gedanken experiment, we can propose several ways of preparing a packet of electrons being localized in the inversion layer at $t=0$, all giving rise to different decay profiles. On the other hand, the measurement of a stationary gate current involves a real experiment not being related to any gedanken experiment whatsoever. However, it goes without saying that carrier localization is a key ingredient for predicting gate currents in the theoretical model we propose herewith. Therefore, we need to ascertain that the calculated gate currents do not depend in essence on the details of the schemes adopted to localize the initial electron charge. To this end, we study three different localization schemes which are graphically summarized in Fig. 2.

Being adopted in the first scheme ("infinite barrier"), the simplest method to induce carrier localization inside the well relies on the assumption of an infinitely high oxide barrier prohibiting the penetration of electrons through the oxide layer. This amounts to complementing the Schrödinger equation

$$-\frac{\hbar^2}{2m}\nabla^2\phi_n(\mathbf{r},z)+U(z)\phi_n(\mathbf{r},z)=E_{0n}\phi_n(\mathbf{r},z) \tag{23}$$

with a hard wall boundary condition at the semiconductor/oxide interface plane $z=0$

$$\phi_n(\mathbf{r},0)=0. \tag{24}$$

Here, $\phi_n(z)$ denotes the envelope wave function of the nth subband whereas m denotes the electron effective mass in the confinement direction (z). In turn, the localized electron ensemble, constructed with the solutions of Eq. (23), gives rise to an electron concentration which equally vanishes at the interface.

In the second scheme ("step barrier") the oxide barrier height is taken to be a finite constant, while the oxide layer is assumed to be infinitely thick, acting as an abrupt potential step. In that case, the single-particle Schrödinger Eq. (23) needs to be extended to oxide region where it will give rise to exponentially damped wave functions. Furthermore, the wave functions and their probability currents must be continuous across the interface $z=0$, leading to

$$\phi_n(\mathbf{r},0^-)=\phi_n(\mathbf{r},0^+), \tag{25}$$

$$\frac{\kappa_n}{m_{\mathrm{ox}}}\phi_n(\mathbf{r},0^-)=\frac{1}{m}\left.\frac{\partial\phi_n(\mathbf{r},z)}{\partial z}\right|_{z=0^+}, \tag{26}$$

where m_{ox} is the electron effective mass for the oxide and κ_n determines the attenuation of the wave function inside the oxide, due to the presence of a constant, finite oxide barrier U_{ox}

$$\kappa_n=\sqrt{\frac{2m_{\mathrm{ox}}(U_{\mathrm{ox}}-E_n)}{\hbar^2}}. \tag{27}$$

As in the previous case, the step-barrier scheme leads to a set of the discrete subbands. However, in this situation the wave functions take a (small) nonzero value at the interface plane. For both cases the statistical average of $c^\dagger_{n'\sigma}c_{n\sigma}$ is given by Eq. (16), i.e.,

$$\langle c^\dagger_{n'\sigma}c_{n\sigma}\rangle=\delta_{n'n}F(E_{0n}-\mu), \tag{28}$$

where the chemical potential μ is determined such that the initial inversion-layer charge $Q_0=Q(0)$ is consistent with the potential profile of the coupled system.

The third method is referred to as "MOS capacitor" or shortly "MOS cap." because the localized charge distribution utilizes the eigenstates of the fully coupled MOS system. As such, the corresponding approach is somewhat different in the sense that the localized states are now constructed by taking appropriate linear combinations of the coupled system eigenstates. More specifically, the nth localized bound-state wave function is taken to be a superposition of stationary wave functions having energies in a narrow interval C_n around the nth resonant energy E_n

$$\phi_n(\mathbf{r},z) = \int_{C_n} dk \lambda_n(k) \phi_k(\mathbf{r},z). \tag{29}$$

Here, the functions $\lambda_n(k)$ denote the amplitudes of the continuum states $|\phi_k\rangle$ contributing to the superposition, whereas k_n defines the nth resonance, i.e., $E_{k_n}=E_n$. δk_n denotes the width of the nth resonance peak and the integration interval is given by $C_n=[k_n-M\delta k_n, k_n+M\delta k_n]$. M is an adjustable parameter that should chosen sufficiently large in order to encompass the relevant part of the resonant peak in the localized state $|\phi_k\rangle$ and sufficiently small in order to avoid overlap with neighboring resonances. This can only be achieved when (1) the widths of the resonant peaks are several orders of magnitude smaller than the resonant energies and (2) the same applies to the ratio of the peak values and the nonresonant amplitudes. Both conditions are obviously fulfilled in the relevant case where the MOS capacitor is not extremely leaky and M typically ranges between 100 and 200. In this light, the motivation for the superposition based approach is that, being part of the continuous spectrum, only the resonant states and their closest neighbors have an appreciable amplitude of hosting electrons that are localized inside the inversion layer, while the function $\lambda_n(k)$ is chosen so as to maximize the charge density of $\phi_n(\mathbf{r},z)$ in the inversion layer. Furthermore, as the states $|\phi_n\rangle$ are generally not eigenstates of an Hermitian operator, their orthogonality is not automatically warranted. On the other hand, the continuous eigenstates $|\phi_k\rangle$ have been chosen to satisfy delta normalization with respect to k, as mentioned in Eq. (22) from which the normalization of the states $|\phi_n\rangle$ can be determined straightforwardly

$$\begin{aligned}\langle\phi_{n'}|\phi_n\rangle &= \int_{C_{n'}} dk' \int_{C_n} dk \lambda^*_{n'}(k')\lambda_n(k)\langle\phi_{k'}|\phi_k\rangle \\ &= \int_{C_{n'}} dk' \int_{C_n} dk \lambda^*_{n'}(k')\lambda_n(k)\delta(k'-k) \\ &= \int_{C_{n'}\cap C_n} dk \lambda^*_{n'}(k)\lambda_n(k).\end{aligned} \tag{30}$$

Consequently, given two different resonances n and n' the cross-section $C_{n'}\cap C_n$ is empty provided M is chosen sufficiently small. Hence, we can safely consider $|\phi_n\rangle$ and $|\phi'_n\rangle$ orthogonal states for $n\neq n'$. On the other hand, imposing orthonormality, we infer from Eq. (30) the normalization requirement

$$\int_{C_n} dk |\lambda_n(k)|^2 = 1. \tag{31}$$

In summary, we need to maximize the charge inside the well region under the constraint of Eq. (31), which may be accomplished by calculating the extremum of the functional

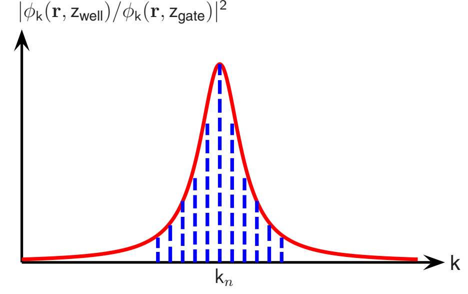

FIG. 3. (Color online) Typical ratio of the electron density in the well to the density in the gate as a function of k around the nth resonant peak, extracted from a continuum wave function. The dashed vertical lines indicate the N wave-function samples used to construct $\phi_n(\mathbf{r},z)$, running from $k_n-M\delta k_n$ to $k_n+M\delta k_n$.

$$\begin{aligned}G[\lambda_n(k)] &= \int_{\Omega_W} d^3r|\phi_n(\mathbf{r},z)|^2 - \alpha\int_{C_n} dk|\lambda_n(k)|^2 \\ &= \int_{C_n} dk' \int_{C_n} dk \lambda^*_n(k')\lambda_n(k)\Gamma_{k'k} - \alpha\int_{C_n} dk|\lambda_n(k)|^2,\end{aligned} \tag{32}$$

where α is a Lagrange multiplier. The stationary value of G corresponds to the zero of the variation δG which is obtained for all first-order variations $\delta\lambda_n(k)$ that are vanishing at the end points of C_n. This leads to the integral equation

$$\int_{C_n} dk' \Gamma_{k'k}\lambda_n(k') = \alpha\lambda_n(k), \tag{33}$$

the solution of which implicitly depends on α. From here on, the functions $\lambda_n(k)$ are taken real because the wave functions involved can, as well, be considered real. In turn, the Lagrange multiplier α is determined by fulfilling the normalization condition

$$\int_{C_n} dk |\lambda_n(k;\alpha)|^2 = 1. \tag{34}$$

In practice, we have discretized Eqs. (33) and (34) by introducing an equidistant mesh of N, k values specifying the interval C_n as shown in Fig. 3. Absorbing the mesh distance into the set λ_{ni}, $1\leq i\leq N$ which represents the restriction of $\lambda_n(k)$ to the mesh points, we are left with the remaining task of finding the normalized eigenvector $(\lambda_{n1},\ldots,\lambda_{nN})$ corresponding to the largest eigenvalue α of

$$\sum_{j=1}^{N} \Gamma_{nij}\lambda_{nj} = \alpha\lambda_{ni} \quad \text{for } 1\leq i\leq N, \tag{35}$$

where Γ_{nij} is the discretized version of $\Gamma_{k'k}$ for the interval C_n. Finally, we need to derive the correlation function $\langle c^\dagger_{n'\sigma}c_{n\sigma}\rangle$ to be used in the MOS cap. localization scheme. Starting from the inverse of Eq. (7), we obtain

$$\langle c^{\dagger}_{n'\sigma} c_{n\sigma}\rangle = \int\int dk' dk \Lambda_{k'n'} \Lambda^{*}_{kn} \langle c^{\dagger}_{k'\sigma}(0) c_{k\sigma}(0)\rangle. \quad (36)$$

Since this scheme adopts as well the Gibbs ensemble to determine the initial density matrix of the coupled system, we further obtain

$$\langle c^{\dagger}_{n'\sigma} c_{n\sigma}\rangle = \int\int dk' dk \Lambda_{k'n'} \Lambda^{*}_{kn} F(E_k - \mu)\delta(k'-k)$$
$$= \int dk \Lambda_{kn'} \Lambda^{*}_{kn} F(E_k - \mu), \quad (37)$$

where μ is now the chemical potential of the coupled system. Clearly, from the very definition of $\phi_n(\mathbf{r},z)$ and due to the delta normalization of the continuum eigenstates, it follows

$$\Lambda_{kn} = \int_{C_n} dk' \lambda_n(k')\delta(k'-k) = \begin{cases} \lambda_n(k) & \text{if } k \in C_n, \\ 0 & \text{else}. \end{cases} \quad (38)$$

Since, by construction, the intervals C_n do not overlap, the product $\Lambda_{kn'}\Lambda^{*}_{kn}$ can be nonzero only if $n=n'$, whence

$$\langle c^{\dagger}_{n'\sigma} c_{n\sigma}\rangle = \delta_{n'n} \int dk \lambda^2_n(k) F(E_k - \mu). \quad (39)$$

As an illustration, the steady-state configurations ($t \to \infty$) resulting from the three localization schemes are shown in Fig. 4. For the sake of comparison, the quasistatic density profile obtained from Ref. 9 by averaging the contributions from the continuum states within an *ad hoc* steady-state Gibbs ensemble, assigning different chemical potentials to left and right tunneling is also shown. It turns out that the MOS cap. scheme provides a very pronounced localization of the initial charge packet. Although it is composed by means of the basis of the continuum states penetrating in the oxide and gate regions, near the interface its distribution is very similar to that of the infinite-barrier case. This indicates that the solution of the coefficients in Eq. (29) is such that the sample wave functions $\phi_{in}(z)$ selected for the interval C_n cancel out each other near the interface while they seem to interfere constructively inside the well. Also from the lower curve of Fig. 4 it follows that the charge distribution produced by the step-barrier scheme which assumes an infinitely step oxide barrier, takes almost the same values near the interface as the one that corresponds to the fully coupled MOS capacitor extending to the gate region.

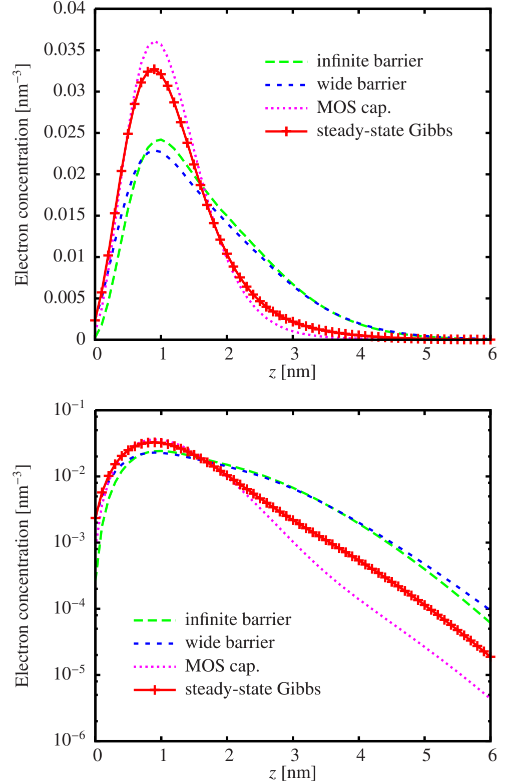

FIG. 4. (Color online) Electron concentration in a Si MOS capacitor with a 1.8 nm EOT and an acceptor concentration of 10^{18} cm^{-3}. The areal electron concentration Q_0/L_xL_y is taken to be 5×10^{12} cm^{-2} for all three localization schemes (dashed and dotted lines) as well as for the steady-state Gibbs average (solid lines). The corresponding gate voltage is V_G=0.48 V. The vertical axis of the upper (lower) plot has a linear (logarithmic) scale.

IV. A SIMPLE REFILLING MECHANISM SOLVING THE CURRENT PARADOX

In order to resolve the paradox related to the gate leakage currents that are emerging from carriers leaving the inversion layer of a MOS capacitor, we start from the continuity equation governing the leakage current in the presence of generation-recombination processes

$$\nabla \cdot \mathbf{J}(\mathbf{r},z,t) + \frac{\partial \rho(\mathbf{r},z,t)}{\partial t} = -e[G(\mathbf{r},z,t) - R(\mathbf{r},z,t)], \quad (40)$$

where $\mathbf{J}(\mathbf{r},z,t)$ is the time-dependent leakage current density relating to the electron charge density $\rho(\mathbf{r},z,t)$. $G(\mathbf{r},z,t)$ and $R(\mathbf{r},z,t)$, respectively, describe the local generation and recombination rates which, in general, are functionals of $\rho(\mathbf{r},z,t)$ and the hole density. As such, the continuity equation does not merely represent a classical conservation law but may also be derived from rigorous quantum dynamics[14] and can thus be used in the present context. Since we explicitly assume that all generation-recombination events are exclusively localized in the semiconductor part of the structure, we integrate Eq. (40) over Ω_W and apply Gauss' theorem to the current-density term, yielding

$$\int_{\partial\Omega_W} \mathbf{dS} \cdot \mathbf{J}(\mathbf{r},z,t) + \frac{dQ(t)}{dt} = -e\gamma(t), \quad (41)$$

where $\partial\Omega_W$ is the boundary surface of Ω_W, $\mathbf{dS}$ denotes the outward pointing surface element and

$$\gamma(t) = \int_{\Omega_W} d^3r[G(\mathbf{r},z,t) - R(\mathbf{r},z,t)] \quad (42)$$

equals the total generation-recombination rate associated with the inversion layer. The contributions from the boundary planes perpendicular to the x and y axis are found to vanish while those from the remaining boundary planes yield the total current at the semiconductor/oxide interface,

$$- I(z=0,t) = - \int_0^{L_x} dx \int_0^{L_y} dy J_z(x,y,0), \quad (43)$$

and the vanishingly small total electron current in the substrate

$$\lim_{z\to\infty} I(z,t) \equiv \lim_{z\to\infty} \int_0^{L_x} dx \int_0^{L_y} dy J_z(x,y,z) = 0. \quad (44)$$

Combining Eqs. (41), (43), and (44), we obtain

$$I(0,t) = \frac{dQ(t)}{dt} + e\gamma(t). \quad (45)$$

In general, various mechanisms are contributing to the generation and recombination events in semiconducting materials.[4,15,16] For instance, electron-hole pairs may be generated by direct or phonon-assisted band-to-band tunneling events occurring where the local electric fields grow sufficiently high. Also transitions to intermediate defect levels are known to enhance generation and recombination currents. However, the aim of this paper is not to investigate the details of specific generation-recombination models but rather to propose a simple generation-recombination-based carrier refilling mechanism that may be integrated straightforwardly into the quantum-mechanical description of tunneling through gate oxide layers. Moreover, since extreme leakage currents are not considered, we explicitly assume that the tunneling processes are slow enough to ensure that the inversion-layer charge hardly changes in time, i.e., $|Q(t) - Q_0| \ll |Q_0|$. Therefore we propose a refilling model in which $\gamma(t)$ is linearly proportional to $Q(t) - Q_0$,

$$e\gamma(t) = \gamma_0(Q(t) - Q_0), \quad (46)$$

the proportionality constant γ_0 being a phenomenological parameter lumping together all relevant generation-recombination effects. It should be noted that the initial charge Q_0 packet is the thermal equilibrium ensemble which represents the localized inversion-layer electrons right before they are "unleashed" at $t=0$. Next, we need to integrate the refilling mechanism into the rate equation obeyed by $Q(t)$. The latter could, in principle, be provided by averaging the Heisenberg equation

$$\frac{dQ(t)}{dt} = - \frac{i}{\hbar}\langle[\hat{Q}(t),\hat{H}(t)]\rangle, \quad (47)$$

where the many-particle Hamiltonian would contain the potential profile induced by V_G and contributions from the electron and hole gases, as well as their interactions with phonons, photons, and defects, that would sustain the generation-recombination processes. Here we prefer a simpler approach based on the interplay between exponential decay due to oxide tunneling (see Sec. II) and the refilling model defined in Eq. (46). Hence, we adopt the following semiclassical rate equation for $Q(t)$:

$$\frac{dQ(t)}{dt} = - \frac{Q(t)}{\tau} - e\gamma(t) = - \frac{Q(t)}{\tau} + \gamma_0[Q_0 - Q(t)], \quad (48)$$

relating $Q(t)$ to the tunneling time τ and the generation rate γ_0. The solution of Eq. (48) is trivial

$$Q(t) = \frac{Q_0}{1+\gamma_0\tau}(\gamma_0\tau + e^{-(1+\gamma_0\tau)t/\tau}). \quad (49)$$

From Eqs. (45), (46), and (49) it follows that:

$$I(0,t) = - \frac{Q(t)}{\tau} = - \frac{\gamma_0 Q_0}{1+\gamma_0\tau}\left(1 + \frac{1}{\gamma_0\tau}e^{-(1+\gamma_0\tau)t/\tau}\right). \quad (50)$$

In particular, the stationary gate current ($t\to\infty$) reads

$$I_G = - \frac{\gamma_0 Q_0}{1+\gamma_0\tau}, \quad (51)$$

which clearly illustrates the competition between the tunneling time τ and the generation time $1/\gamma_0$. In particular, when the local generation is a substantially faster process than the leakage tunneling, we may exploit the inequality $\gamma_0\tau \gg 1$ to reduce the gate current formula to its simplest form

$$I_G = - \frac{Q_0}{\tau}. \quad (52)$$

Due to the simplicity of the above model, the major effort in calculating the gate current is in the four-step algorithm described in Sec. II, since the resulting decay characteristics immediately provide Q_0 and τ, the two basic quantities needed to evaluate I_G.

V. RESULTS AND DISCUSSION

Ignoring the refilling mechanism for now, we first compare the tunneling lifetime calculations obtained from the three localization schemes. Following the four-step method discussed in Sec. II, we have calculated the decay of an inversion-layer charge packet $Q(t)$ and plotted the resulting curves in Fig. 5. The Schrödinger-Poisson solver developed in Ref. 9 has been employed to determine the resonant states.

Typically, the infinite-barrier and step-barrier localization schemes result in an exponential decay modulated by oscillations, while the decay predicted by the MOS cap. scheme is a very smooth and monotonic function of time. This may be understood by careful inspection of the wave functions of the initially bound states which, according to the very definition of the coefficients Λ_{kn} in Eq. (10), can be written as a Fourier integral

$$\phi_n(\mathbf{r},z) = \int dk \Lambda_{kn}\phi_k(\mathbf{r},z). \quad (53)$$

In principle all Fourier components are seen to contribute to the wave functions extracted from the infinite-barrier and

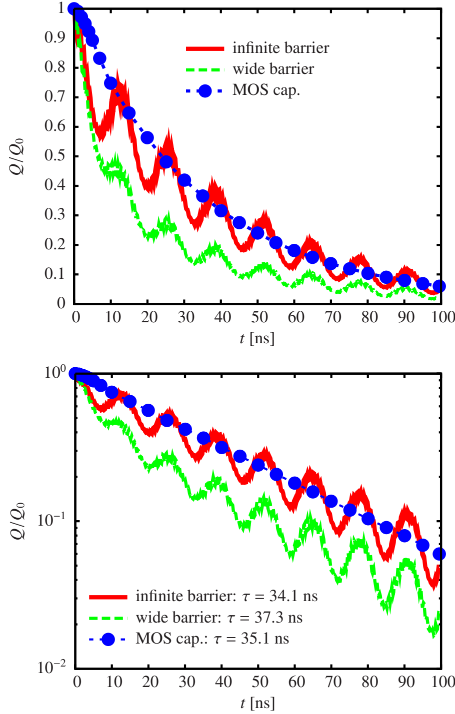

FIG. 5. (Color online) Charge evolution for a Si MOS capacitor with an 1.2 nm EOT and an acceptor concentration of 10^{18} cm^{-3}. The areal electron concentration Q_0/L_xL_y is set to be 10^{11} cm^{-2} for all three localization schemes (dashed and dotted lines). The corresponding gate voltage is V_G=0.48 V. The vertical axis of the upper (lower) plot has a linear (logarithmic) scale.

step-barrier schemes. In particular, the superpositions of resonant and off-resonant continuum states occurring in Eq. (53) are causing the oscillations observed in the function $Q(t)$. Indeed, the off-resonant states carry a significant probability of finding an electron in the gate region, whereas the resonant states are known to maximize the probability of finding an electron in the inversion-layer well. Electrons in a superposition electrons of the states $|\phi_n\rangle$ are therefore seen to propagate back and forth between the gate and the semiconductor until the charge packet has faded away, while the frequency of the oscillations is related to the energy difference of the resonant and off-resonant states. As such, the observed oscillatory behavior may be interpreted as a straightforward emanation of a quantum diffusion process governing the decay of strictly localized states into extended states.[17] On the other hand, the MOS cap. localization scheme is explicitly designed to select only a restricted set of Fourier components located in a narrow window around one particular resonance—provided M is chosen sufficiently small. The resulting smooth decay shown in Fig. 5 essentially reduces to the exponential decay that defines the tunneling lifetime, i.e.,

$$Q(t)|_{\text{tunneling}} = Q_0 \exp\left(-\frac{t}{\tau}\right), \tag{54}$$

while the absence of any oscillations can be predicted by noticing that only a single resonant state contributes to a given localized wave packet.

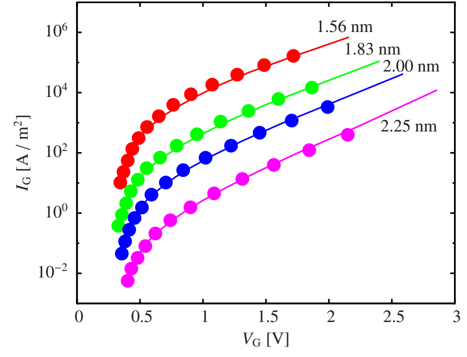

FIG. 6. (Color online) The solid lines are computed with the use of the tunneling lifetime model for four different EOTs, while circles are measured curves.

Finally, it can be observed from the lower part of Fig. 5 that the tunneling times extracted from the three initialization schemes are very close, which suggests that they share an almost identical long-time evolution. Stated otherwise, the tunneling lifetime depends mainly on the structural properties of the MOS capacitor and the applied gate voltage, rather than on the way the initial charge packet Q_0 is prepared.

Next, we return to the model that incorporates the generation events and use the results from the previous steps to calculate the gate current from Eq. (52). The resulting calculations reported in Fig. 6 involve an acceptor concentration of 10^{18} cm^{-3} while the effective oxide thickness (EOT) ranges from 1.56 to 2.25 nm. The simulated gate currents are compared with experimental data[18] in the same figure.

The calculated results show very good agreement with the measured gate currents for all reported EOT values, which justifies the use of the above developed tunneling lifetime model for a broad variety of leaky MOS devices. Furthermore, the dependence of the tunneling lifetime on the gate voltage, as plotted in Fig. 7 for different EOTs, strongly suggests the existence of an exponential relation. More explicitly, we are tempted to conjecture that

$$I_G = \frac{Q_0(V_G)}{\tau(V_G)} = \frac{Q_0(V_G)}{\tau_0}\exp\left(\frac{V_G}{V_0}\right), \tag{55}$$

where, for a given MOS structure, τ_0 and V_0 are positive constants. As suggested in Sec. IV, we implicitly assume that

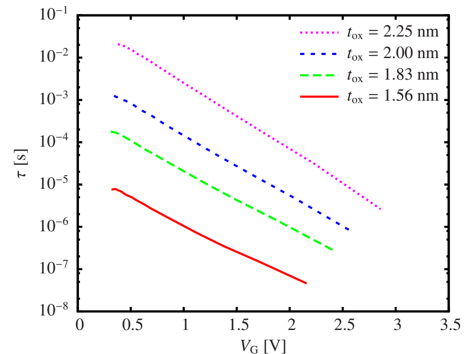

FIG. 7. (Color online) Tunneling lifetime versus gate voltage for different EOTs.

the generation process substantially exceeds the tunneling rate. As the latter relies on a self-consistent solution of the Poisson and Schrödinger equations, this assumption also implies that the generation of electron-hole pairs does not appreciably affect the potential profile throughout the MOS capacitor. This situation typically occurs when the MOS device suffers from poorly passivated electrically active states, while being left with lots of generation centers located near the interface, but is also characteristic for a genuine MOS field-effect transistor where the highly doped source contacts provide an almost instantaneous supply of channel electrons.

On the other hand, when the generation time is significantly larger than τ, the measured gate leakage current would be dominated by the generation events, as would be the case for properly surface-passivated devices. Correspondingly, the solution of Eq. (52) would reduce to

$$I_G = -\gamma_0 Q_0. \tag{56}$$

Though being a naive oversimplification, this result clearly indicates the predominance of the generation current. A more general description incorporating also this regime would require that the local generation-recombination rate be included explicitly on the right-hand side of Poisson's equation. This way, the resulting potential profile may account for the correct balance between the generation and tunneling rates.

Next, we compare the present model with related earlier work[5,6,19] exploiting as well the properties of the continuum states and their sharply peaked resonances. Among the first authors realizing the problem of the current paradox, Gamow,[5] Breit and Wigner[6] faced the essential inability of generating irreversible decay of alpha particles trapped in the potential well of an atomic nucleus and being described by a real eigenfunction. Alternatively, they proposed to assign to the energy eigenvalues an imaginary part that could be identified as an appropriate linewidth of the resonant state associated with the quasibound state of an alpha particle. Correspondingly, the inverse linewidth could be considered a measure of the "dwell time," i.e., the time an alpha particle spends inside the potential well. Moreover, the related imaginary part of the wave numbers describing the standing waves far away from the nucleus (equivalent to the gate region) artificially breaks the time-reversal symmetry of the standing waves, thereby allowing the newly formed complex wave functions to carry a nonzero current. This approach has been successfully extended to tackle the problem of gate currents in the recent past,[9,10,19] while the real and imaginary energy parts could uniquely be identified with the resonant energies and widths extracted from the energy-depending amplitudes solving the Schrödinger equation. Nevertheless, we considered the introduction of complex wave numbers and energies needed to generate current carrying states as well as the corresponding lack of an Hermitian Hamiltonian an essential drawback of the Gamow approach. In contrast, our model is merely using the (real) solutions of the Schrödinger equation to mediate the dynamics of localized charge packets. We conclude this section with a few remarks on the limitations of the present model and potential improvements or extensions. First, it's clear that the generation-limited regime $\gamma_0\tau \ll 1$ can only be treated quantitatively if the rate γ_0 can be assigned a meaningful numerical value to be extracted from a microscopic theory extending beyond the above phenomenological description of generation-recombination processes. The latter may be accounted for, in principle, by inserting into Heisenberg Eq. (47) an interaction Hamiltonian containing generation and recombination processes of the form $c^{\dagger}_{k\sigma}b^{\dagger}_{-k'\sigma}$, where $b^{\dagger}_{-k'\sigma}$ is a hole creation operator. Finally, it should be noticed that nonparabolic dispersion relations and the corresponding band-structure calculations[20] can be incorporated into the present formalism without any conceptual difficulty. In particular, the gate currents carried by holes tunneling out of a p-type channel can be treated in an analogous way. The computational burden, on the other hand, would be found to increase, as the longitudinal motion accounting for the tunneling could no longer be decoupled from the free-carrier motion in the lateral x and y directions and would make all subband-related quantities depend on the two-dimensional lateral wave vector.

VI. CONCLUSION

A model is presented to calculate the gate leakage currents in MOS devices operating in inversion mode. Relying on the dynamical evolution of a charge packet initially localized in the inversion layer, the model provides an algorithm to calculate the tunneling time which is the key quantity needed to unequivocally determine steady-state gate currents in the tunneling-limited regime, together with the initially localized charge. Calculated values of the gate currents agree very well with experimental data whereas, for a given applied gate voltage, the tunneling time is independent of the schemes adopted to construct the localized charge packets. More general, the model enables one to explore qualitatively the border between the generation-limited and tunneling-limited regimes. Quantitative predictions need not be restricted to the tunneling-limited regime either, provided that a microscopic model describing local generation-recombination events is included in the Heisenberg dynamics of the electron and hole operators. Moreover, the computational procedure leading to the tunneling lifetime, as outlined for n-type channels with parabolic conduction bands in the present work, can be generalized to treat as well holes in p-type inversion layers or electrons residing in nonparabolic subbands. Finally, solving the current paradox, the present model demonstrates that the occurrence of sustained tunneling leakage currents in MOS capacitors entails the local violation of the electron continuity equation due to the generation of electron-hole pairs. Correspondingly, the hole current which manifests itself as a measurable substrate current, violates as well the local conservation of holes, while the continuity equation for the total current density is found to hold all over the structure, as it should be.

ACKNOWLEDGMENTS

We are indebted to Geoffrey Pourtois for his technical support.

*pourgham@imec.be
[†]Also at Universiteit Antwerpen, Physics Department, Groenenborgerlaan 171, B-2020 Antwerpen, Belgium. FAX: +32-3-265 35 42.
[‡]Also at Katholieke Universiteit Leuven, Engineering Department, Kasteelpark, Arenberg 10, B-3001 Leuven, Belgium. FAX: +32-16-28 12 14.

[1] R. H. Fowler and L. W. Nordheim, Proc. R. Soc. London, Ser. A **119**, 173 (1928).
[2] J. Maserjian, J. Vac. Sci. Technol. **11**, 996 (1974).
[3] J. Bardeen, Phys. Rev. Lett. **6**, 57 (1961).
[4] A. Schenk, *Advanced Physical Models for Silicon Device Simulation* (Springer-Verlag, Wien, 1998), pp. 281–315.
[5] G. Gamow, Z. Phys. **51**, 204 (1928).
[6] G. Breit and E. P. Wigner, Phys. Rev. **49**, 519 (1936).
[7] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics (Non-Relativistic Theory)* (Pergamon, London, 1958), p. 441.
[8] A. K. Ghatak, K. Thyagarajan, and M. R. Shenoy, IEEE J. Quantum Electron. **24**, 1524 (1988).
[9] W. Magnus and W. Schoenmaker, J. Appl. Phys. **88**, 5833 (2000).
[10] W. Magnus and W. Schoenmaker, Microelectron. Reliab. **41**, 31 (2001).
[11] G. S. Lujan, B. Sorée, W. Magnus, and K. De Meyer, J. Appl. Phys. **100**, 033708 (2006).
[12] J. Suñé, P. Olivo, and B. Riccó, J. Appl. Phys. **70**, 337 (1991).
[13] E. Merzbacher, *Quantum Mechanics* (John Wiley & Sons, New York, 1970), pp. 481–487.
[14] W. Magnus, Phys. Status Solidi B **237**, 341 (2003).
[15] W. Shockley and W. T. Read, Phys. Rev. **87**, 835 (1952).
[16] J. G. Fossum, R. P. Mertens, D. S. Lee, and J. F. Nijs, Solid-State Electron. **26**, 569 (1983).
[17] B. J. Kim, H. Hong, and M. Y. Choi, Phys. Rev. B **68**, 014304 (2003).
[18] R. Clerc, G. Ghibaudo, and G. Pananakakis, Solid-State Electron. **46**, 1039 (2002).
[19] P. Price, Semicond. Sci. Technol. **19**, S241 (2004).
[20] M. Städele, B. R. Tuttle, and K. Hess, J. Appl. Phys. **89**, 348 (2001).