Self-consistent modeling of longitudinal quantum effects in nanoscale double-gate metal oxide semiconductor field effect transistors

Frederik O. Heinz[a)]
Integrated Systems Laboratory, ETH Zurich, Gloriastrasse 35, CH-8092 Zürich, Switzerland

Andreas Schenk[b)]
Integrated Systems Laboratory, ETH Zurich, Gloriastrasse 35, CH-8092 Zürich, Switzerland, and Synopsys Switzerland LLC, Affolternstrasse 52, CH-8050 Zürich, Switzerland

(Received 28 October 2005; accepted 3 July 2006; published online 23 October 2006)

Ultrathin double-gate silicon-on-insulator transistors are studied in the quantum coherent limit. By treating electron-electron interaction on the level of a mean field approach, the density matrix of the device becomes diagonal when expressed in a basis that results from imposing scattering boundary conditions at the terminals. The self-consistent scattering wave functions are computed using a multisubband scattering matrix formalism. This allows us to retain the full dimensionality of the wave functions and eliminates the need for the adiabatic decomposition of the Schrödinger equation. Subband mixing is fully taken into account and a piecewise analytical representation of the wave functions can significantly reduce the number of sampling positions along transport direction. By self-consistent simulations the size of source-to-drain tunneling as a function of gate length is demonstrated for different body thicknesses. A strong forward bias is shown to increase the tunnel current due to the thinning of the source-drain potential barrier. The effect of channel orientation on the tunnel current is also discussed. © *2006 American Institute of Physics*.

[DOI: 10.1063/1.2355540]

I. INTRODUCTION

Ultrathin double-gate silicon-on-insulator (DGSOI) transistors are among the most promising devices for future very large scale integration (VLSI). Recently, DGSOI metal oxide semiconductor field effect transistors (MOSFETs) with silicon body thicknesses t_{Si} of 1 and 3 nm have been manufactured at Nippon Telephone and Telegraph Corporation (NTT).[1] These devices are still far from ideal because of strong film inhomogeneities. Nevertheless, it is interesting to study the ultimate behavior that may be obtained from such devices once the technological difficulties are overcome. Various quantum-mechanical (QM) theories have been investigated to face this complicated problem. To note the most important ones are as follows: the quantum transmitting boundary method (QTBM) from the International Business Machines (IBM) laboratory,[2] the nonequilibrium Green's functions (NEGFs) introduced by Kadanoff and Baym in the 1960s,[3] the Wigner transport equation,[4] the quasi-Pauli master equation,[5] and the Landauer-Büttiker formalism in absence of scattering.[6]

The most rigid application of the NEGF theory to a quantum transport problem is the software package NEMO developed at Texas Instruments at the end of the 1990s.[7] It allows the simulation of one-dimensional (1D) quantum-ballistic (QB) transport in III-V heterostructures with inclusion of scattering mechanisms via self-energies computed self-consistently with the NEGFs. The self-consistency with the Poisson equation demands high computer resources so that most of the self-energies considered in NEMO are diagonal. NEGFs also build the kernel of NANOMOS,[8,9] a software developed at Purdue University at the beginning of 2000. NANOMOS has two-dimensional capabilities although the transport direction is separated from the direction of confinement by virtue of the so-called mode-space approach[10,11] where the solution of the Schrödinger equation in the direction of confinement yields eigenenergies for the NEGFs in the transport direction. A mode-coupling version of the mode-space approach is presented in Ref. 12. A method that retains the full dimensionality of the system and is readily applicable to multiterminal systems can be found in Ref. 13. Scattering can be included by diagonal phenomenological decay times (Büttiker probes).

In contrast to the above references, that use finite difference or finite-element representations, we evaluate Green's functions in the framework of the so-called scattering matrix approach.[14,15] This allows for subband mixing while using piecewise analytical wave functions, which in some devices may considerably reduce the necessary number of grid planes in transport direction. In the present work the ultimate performance of DGSOI MOSFETs with silicon body thicknesses of 1 and 3 nm and gate lengths in the range from 3 to 30 nm has been assessed by QB transport simulations based on the self-consistent scattering matrix approach. The electrostatic effect of the injected charge is taken into account. To establish a connection with standard device simulation, the same transistors have also been studied by classical drift-diffusion (DD) simulations with constant mobility and DD simulations on top of a 1D Schrödinger-Poisson (1D-SP) equation system[16] that make use of a QM mobility model calibrated for long-channel MOSFETs.[17] In this model

[a)]Present address: Intel Corporation, Mailstop RA3-254, 2501 NW 229th Avenue, Hillsboro, OR 97124.

[b)]Electronic mail: schenk@iis.ee.ethz.ch the microscopic scattering rates (including intersubband scattering) are numerically computed based on the confined states from the Schrödinger solver, and the QM conductivity is subject to the self-consistent solution of the DD/1D-SP system. Recently, a DD/1D-SP variant with completely decoupled subbands[18] was proposed, where DD is applied to each subband separately with a subband density-averaged empirical mobility. Both methods can be expected to give similar results in cases of very strong confinement, i.e., if only the lowest subband contributes significantly.

The paper is organized as follows. The theory of QB transport is outlined in Sec. II. Section III summarizes the scattering matrix approach and Sec. IV provides the expressions for the current through quantum wires and quantum wells. Section V explains the difference between self-consistent and local-equilibrium charge densities within the scattering matrix method. The simulation results are presented in Sec. VI. Finally conclusions are drawn in Sec. VII.

II. QUANTUM-BALLISTIC TRANSPORT

For a many-particle system whose state is described by the statistical operator P, the expectation value $\langle A\rangle$ of a single-particle observable A is given by the trace of the product of the corresponding second-quantized operator,

$$A_{\text{many}} := \sum_{i,j} c_i^\dagger c_j \langle i|A|j\rangle, \tag{1}$$

where the $c_i^\dagger$ (c_j) are the creation (annihilation) operators for an arbitrary basis of single-particle states, with the statistical operator P,

$$\langle A\rangle = \text{tr}(A_{\text{many}}P). \tag{2}$$

In terms of the density matrix,

$$M_{i,j} \equiv \langle i|M|j\rangle := \langle c_j^\dagger c_i\rangle = \text{tr}(c_j^\dagger c_i P), \tag{3}$$

the expectation value of A may be recast as a trace over the space of single-particle states,

$$\langle A\rangle = \text{tr}(AM). \tag{4}$$

Now we consider an open quantum system with several terminals, that couple the system to particle reservoirs α (with electrochemical potentials $\epsilon_F^{(\alpha)}$ and a common temperature T) via ideal waveguides (cf. Fig. 1). For the analysis of such a system it is convenient to work in a single-particle basis in *scattering configuration*, i.e., each single-particle wave function is associated with an *injecting reservoir*. Only in the waveguide attached to the *injecting* reservoir the wave function contains a component that propagates *towards* the system; in the remaining waveguides the wave function only contains components propagating *away from* the system. Then the single-particle basis may relabeled by $|\alpha,i\rangle$, where α identifies the injecting reservoir (e.g., $\alpha\in\{\text{src},\text{drn}\}$ for a two-terminal system) and i is a collective label for the remaining quantum numbers. In this notation the statistical operator of the system takes the form

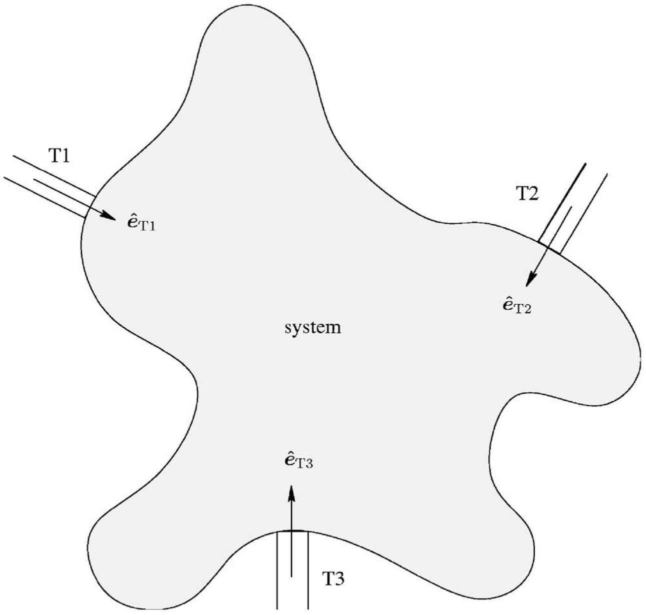

FIG. 1. The simulation domain (“system”) is coupled to the reservoirs by ideal waveguides.

$$P = \frac{1}{\Xi}\underbrace{\exp\left[-\frac{1}{k_BT}\left(H - \sum_{\alpha,i} c_{\alpha,i}^\dagger c_{\alpha,i}\epsilon_F^{(\alpha)}\right)\right]}_{=:\tilde{P}}, \tag{5}$$

with the normalization factor $\Xi := \text{tr}(\tilde{P})$.

In the interaction free case,

$$H = \sum_{\alpha,i} c_{\alpha,i}^\dagger c_{\alpha,i}\epsilon_{\alpha,i}, \tag{6}$$

P contains only pairs of creation and annihilation operators with the same quantum numbers. Consequently, the off-diagonal terms of the density matrix vanish,

$$M_{\alpha,i;\beta,j} = \bar{n}_{\alpha,i}\delta_{\alpha,\beta}\delta_{i,j}, \tag{7}$$

and the average occupation numbers $\bar{n}_{\alpha,i}$ are found to obey Fermi-Dirac statistics with the electrochemical potential of the injecting reservoir,

$$\bar{n}_{\alpha,i} = f\left[\frac{1}{k_BT}(\epsilon_{\alpha,i} - \epsilon_F^{(\alpha)})\right]. \tag{8}$$

In the context of finite-T density functional computations, Fermi-Dirac occupation factors are usually employed to populate the Kohn-Sham orbitals, despite the fact that these self-consistent orbitals are not interaction free. This approach works well for systems with spread-out wave functions, but is less accurate in the presence of localized states (e.g., inside quantum dots) and fails to reproduce effects such as Coulomb blockade, that are based on strong electron-electron interaction. Here we are mainly concerned with MOSFET devices, where wave functions are delocalized and the assumption of Fermi-Dirac occupation factors may be maintained. Then, the current through any surface A inside the device is given by

$$\langle I_A \rangle = \sum_{\alpha,i} \langle \alpha,i|I_A|\alpha,i\rangle f\left[\frac{1}{k_BT}(\epsilon_{\alpha,i}-\epsilon_F^{(\alpha)})\right]. \tag{9}$$

Here, I_A denotes the current operator for the surface A,

$$I_A = \frac{\hbar}{m}\int_A d^2\mathbf{r}|\mathbf{r}\rangle\langle\mathbf{r}|\Im(\hat{\mathbf{n}}(\mathbf{r})\cdot\boldsymbol{\nabla}), \tag{10}$$

where $\hat{\boldsymbol{n}}$ is the surface normal vector and $\Im$ stands for imaginary part.

In two-terminal devices with identical source and drain waveguides it is possible to identify pairs of degenerate single-particle wave functions which carry currents of equal magnitude but opposite signs. Then Eq. (9) simplifies to

$$\langle I_A \rangle = \sum_i \langle \mathrm{src},i|I_A|\mathrm{src},i\rangle\left[f\left(\frac{\epsilon_i-\epsilon_F^{(\mathrm{src})}}{k_BT}\right)-f\left(\frac{\epsilon_i-\epsilon_F^{(\mathrm{drn})}}{k_BT}\right)\right]. \tag{11}$$

This expression shows how current is driven by the voltage drop between source and drain. Numerical evaluation of the current matrix elements requires the computation of the Kohn-Sham basis of the device in scattering configuration, which is done in the next section.

III. COMPUTING THE KOHN-SHAM BASIS IN SCATTERING CONFIGURATION

For QB transport simulations through quantum wires/wells, the Schrödinger equation with scattering boundary conditions for injection of electrons from the source (or drain) contact is solved by means of a scattering matrix formalism, that makes allowance for subband mixing.[14,15] This approach uses piecewise analytical wave functions of the form

$$\Psi(\mathbf{r}) = \frac{1}{\sqrt{L}}\sum_n \chi_{\mathcal{I}_n}(x)\sum_i [a_i^{(n)}e^{ik_i^{(n)}(x-x_n)}+b_i^{(n)}e^{-ik_i^{(n)}(x-x_n)}] \times \begin{cases}\langle y,z|i\rangle^{(n)} & \text{“wire”}\\ e^{ik_\perp y}\langle z|i\rangle^{(n)} & \text{“well”}\end{cases} \tag{12}$$

where $\chi_{\mathcal{I}_n}$ is the characteristic function of the interval

$$\mathcal{I}_n := \left[\frac{x_{n-1}+x_n}{2},\frac{x_n+x_{n+1}}{2}\right]. \tag{13}$$

The arbitrary normalization length L is used in the computation of the one-dimensional density of states of the terminal waveguides and will cancel in the final expressions for current and charge density. The x_n denote the coordinates of the grid planes in transport direction, and $|i\rangle^{(n)}$ is the ith transverse bound state on interval $\mathcal{I}_n$. The potential is assumed piecewise constant along the x direction; hence, the transverse wave functions and subband energies remain the same throughout each interval $\mathcal{I}_n$, and each subband wave function evolves over the interval according to a superposition of forward and backward exponentials. The full Schrödinger equation including the lattice periodic potential would rather suggest an expansion of Ψ in terms of Bloch functions, but on the level of the effective mass approximation (EMA) exponentials are recovered. The forward component of the velocity associated with a function of wave vector $\mathbf{k}$, then, is given by $v_\parallel(\mathbf{k})=(1/\hbar)(\hat{\boldsymbol{e}}_x\cdot\boldsymbol{\nabla}_\mathbf{k})\tilde{\epsilon}(\mathbf{k})$.[19] In a wave function Ψ of total energy ϵ (including a possible kinetic energy contribution $\epsilon_\perp$ brought about by $k_\perp$), the wave number associated with the ith transverse mode on slice n is given by

$$k_i^{(n)} = \begin{cases}\sqrt{2m_x^*[\epsilon-\epsilon_\perp-\epsilon_i(x_n)]}/\hbar, & \epsilon-\epsilon_\perp-\epsilon_i(x_n)\geqslant 0\\ i\sqrt{2m_x^*[\epsilon_i(x_n)+\epsilon_\perp-\epsilon]}/\hbar & \text{otherwise.}\end{cases} \tag{14}$$

Here, $\epsilon_i(x_n)$ denotes the eigenenergy of the transverse mode $|i\rangle^{(n)}$. Hence, $\epsilon-\epsilon_i(x_n)$ is the kinetic energy available for transport along the x direction.

In devices that contain regions in which the potential does not vary strongly on the length scale of the electron wavelength, representing the wave function by piecewise analytic expressions allows using a discretization grid with much fewer grid planes[27] along the transport direction than are needed with a finite differences scheme in order to resolve short wavelengths.

Provided that the subband energies do not exhibit Dirac-δ singularities, Ψ must be continuous and possess a continuous first partial derivative along the x direction. By enforcing these conditions on the interval boundaries and invoking the orthogonality relation ${}^{(n)}\langle i|j\rangle^{(n)}=\delta_{i,j}$, the coefficients on interval $\mathcal{I}_{n+1}$ may be expressed in terms of the coefficients on interval $\mathcal{I}_n$ as

$$a_j^{(n+1)} = \frac{1}{2}e^{(1/2)ik_j^{(n+1)}\Delta x_n}\sum_i {}^{(n+1)}\langle j|i\rangle^{(n)} \times\left[\left(1+\frac{k_i^{(n)}}{k_j^{(n+1)}}\right)e^{(1/2)ik_i^{(n)}\Delta x_n}a_i^{(n)} + \left(1-\frac{k_i^{(n)}}{k_j^{(n+1)}}\right)e^{-(1/2)ik_i^{(n)}\Delta x_n}b_i^{(n)}\right] \tag{15}$$

and

$$b_j^{(n+1)} = \frac{1}{2}e^{-(1/2)ik_j^{(n+1)}\Delta x_n}\sum_i {}^{(n+1)}\langle j|i\rangle^{(n)} \times\left[\left(1-\frac{k_i^{(n)}}{k_j^{(n+1)}}\right)e^{(1/2)ik_i^{(n)}\Delta x_n}a_i^{(n)} + \left(1+\frac{k_i^{(n)}}{k_j^{(n+1)}}\right)e^{-(1/2)ik_i^{(n)}\Delta x_n}b_i^{(n)}\right]. \tag{16}$$

Here, it is assumed, that all the $k_j^{(n+1)}$ are nonzero. Mode mixing is afforded by the overlap matrix elements ${}^{(n+1)}\langle j|i\rangle^{(n)}$ between mode j on slice $(n+1)$ and mode i on slice n. Equations (15) and (16) may be combined into a vector equation,

$$\begin{bmatrix}\mathbf{a}^{(n+1)}\\ \mathbf{b}^{(n+1)}\end{bmatrix} = \mathbb{M}^{(n)}\begin{bmatrix}\mathbf{a}^{(n)}\\ \mathbf{b}^{(n)}\end{bmatrix}. \tag{17}$$

The matrix $\mathbb{M}^{(n)}$ is called the (nth partial) *transfer matrix*.

In this representation, scattering boundary conditions for injection into the ith subband of a two-terminal device are specified by

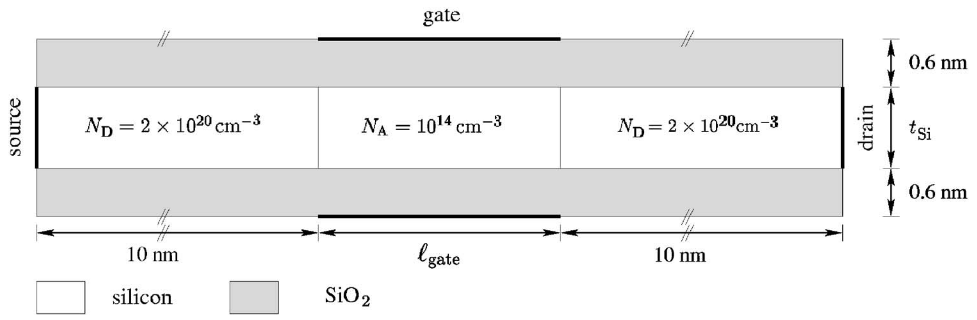
gate
source
drain
$N_D = 2\times10^{20}\,\mathrm{cm}^{-3}$
$N_A = 10^{14}\,\mathrm{cm}^{-3}$
$N_D = 2\times10^{20}\,\mathrm{cm}^{-3}$
0.6 nm
t_{Si}
0.6 nm
10 nm
ℓ_{gate}
10 nm
silicon
SiO_2

FIG. 2. Geometry of the DGSOI MOSFET structure.

$$a_j^{(0)} = \delta_{i,j}, \quad b_j^{(n_{\max})} = 0, \tag{18}$$

for injection from the left reservoir and

$$a_j^{(0)} = 0, \quad b_j^{(n_{\max})} = \delta_{i,j}, \tag{19}$$

for injection from the right. Hence, the coefficient vectors $\mathbf{a}^{(n_{\max})}$ and $\mathbf{b}^{(0)}$ need to be computed from the coefficients $\mathbf{a}^{(0)}$ and $\mathbf{b}^{(n_{\max})}$. This may be written in matrix form as

$$\begin{bmatrix}\mathbf{a}^{(n_{\max})}\\ \mathbf{b}^{(0)}\end{bmatrix} = \mathbb{S}^{(0,n_{\max})}\begin{bmatrix}\mathbf{a}^{(0)}\\ \mathbf{b}^{(n_{\max})}\end{bmatrix}, \tag{20}$$

with the *scattering matrix*[28]

$$\mathbb{S}^{(0,n_{\max})} = \begin{bmatrix}\mathbb{S}_{0,0}^{(0,n_{\max})} & \mathbb{S}_{0,1}^{(0,n_{\max})}\\ \mathbb{S}_{1,0}^{(0,n_{\max})} & \mathbb{S}_{1,1}^{(0,n_{\max})}\end{bmatrix}. \tag{21}$$

Formally writing, the scattering matrix may be obtained from

$$\begin{bmatrix}\mathbf{a}^{(n_{\max})}\\ \mathbf{b}^{(n_{\max})}\end{bmatrix} = \mathbb{M}^{(n_{\max}-1)}\cdots\mathbb{M}^{(1)}\mathbb{M}^{(0)}\begin{bmatrix}\mathbf{a}^{(0)}\\ \mathbf{b}^{(0)}\end{bmatrix} =: \mathbb{M}^{\mathrm{tot}}\begin{bmatrix}\mathbf{a}^{(0)}\\ \mathbf{b}^{(0)}\end{bmatrix}, \tag{22}$$

where $\mathbb{M}^{\mathrm{tot}}$ is called the *total transfer matrix*. However, for reasons discussed in Ref. 20 $\mathbb{M}^{\mathrm{tot}}$ tends to have poor numerical quality. Therefore, we have used a numerical scheme,[14] that recursively constructs $\mathbb{S}^{(0,n+1)}$ from $\mathbb{S}^{(0,n)}$ and (the inverse of)[29] $\mathbb{M}^{(n)}$, starting with $\mathbb{S}^{(0,0)}=\mathbb{1}\oplus\mathbb{1}$.

IV. COMPUTATION OF THE CURRENT

In the notation of the preceding section, it is straightforward to evaluate the current matrix elements in Eqs. (9) and (11) for a cross-sectional plane A normal to $\hat{\boldsymbol{e}}_x$ as

$$\langle \mathrm{src},\epsilon,i,k_\perp|I_A|\mathrm{src},\epsilon,i,k_\perp\rangle = \frac{1}{L}v_{\|,i}^{(0)}\Bigg(1-\underbrace{\sum_j \frac{v_{\|,j}^{(0)}}{v_{\|,i}^{(0)}}|[\mathbb{S}_{1,0}^{(0,n_{\max})}]_{j,i}|^2}_{=:R_{\mathrm{src},i\to j}(\epsilon,k_\perp)}\Bigg) \tag{23}$$

and

$$\langle \mathrm{drn},\epsilon,i,k_\perp|I_A|\mathrm{drn},\epsilon,i,k_\perp\rangle = -\frac{1}{L}v_{\|,i}^{(n_{\max})}\underbrace{\sum_j \frac{v_{\|,j}^{(0)}}{v_{\|,i}^{(n_{\max})}}|[\mathbb{S}_{0,0}^{(0,n_{\max})}]_{j,i}|^2}_{=:T_{\mathrm{drn}\to\mathrm{src},i\to j}(\epsilon,k_\perp)}, \tag{24}$$

with the velocity

$$v_{\|,i}^{(n)} := \frac{1}{\hbar}(\hat{\boldsymbol{e}}_x\cdot\boldsymbol{\nabla}_{\mathbf{k}})\tilde{\epsilon}(k_i^{(n)},k_\perp). \tag{25}$$

When summing over all available states, the forward component of the density of states in the injecting waveguide cancels the velocity terms in Eqs. (23) and (24) and the total current (including a factor of 2 for spin degeneracy) becomes

$$I^{\mathrm{1D}} = \frac{2}{h}\sum_i\Bigg\{\int_{\epsilon_i^{(0)}} d\epsilon\Big[1-\sum_j R_{\mathrm{src},i\to j}(\epsilon)\Big]f\Bigg(\frac{\epsilon-\epsilon_F^{(\mathrm{src})}}{k_BT}\Bigg) - \int_{\epsilon_i^{(n_{\max})}} d\epsilon\sum_j T_{\mathrm{drn}\to\mathrm{src},i\to j}(\epsilon)f\Bigg(\frac{\epsilon-\epsilon_F^{(\mathrm{drn})}}{k_BT}\Bigg)\Bigg\}, \tag{26}$$

for a quasi-1D quantum wire. For transport through a quantum well, e.g., the channel of a double gate MOSFET as shown in Fig. 2, the current expression acquires an additional integration over $k_\perp$. In the case of a parabolic dependence of the band structure $\tilde{\epsilon}$ on $k_\perp$ this yields

$$I^{\mathrm{2D}} = \frac{2}{h}\sqrt{\pi}W\frac{\sqrt{2m_\perp^* k_BT}}{h}\sum_i\Bigg\{\int_{\epsilon_i^{(0)}} d\epsilon\Big[1-\sum_j R_{\mathrm{src},i\to j}(\epsilon)\Big] \times\mathfrak{F}_{-1/2}\Bigg(\frac{\epsilon_F^{(\mathrm{src})}-\epsilon}{k_BT}\Bigg) - \int_{\epsilon_i^{(n_{\max})}} d\epsilon\sum_j T_{\mathrm{drn}\to\mathrm{src},i\to j}(\epsilon) \times\mathfrak{F}_{-1/2}\Bigg(\frac{\epsilon_F^{(\mathrm{drn})}-\epsilon}{k_BT}\Bigg)\Bigg\}, \tag{27}$$

where the width W is assumed to be much larger than the wavelength of the charge carriers. The fraction to the right of W is the reciprocal of the thermal wavelength $\lambda_\perp$ normal to the plane of Fig. 2. Thus $W/\lambda_\perp$ may be regarded as an effective number of modes in normal direction. When using the EMA with indirect gap materials, charge density and current need to be summed up over all valleys.

V. APPLICABILITY OF THE EMA

The silicon film thicknesses of the transistors studied in this work are very small ($\leqslant 3$ nm). In this regime the EMA cannot *a priori* be assumed to be valid. To assess the validity

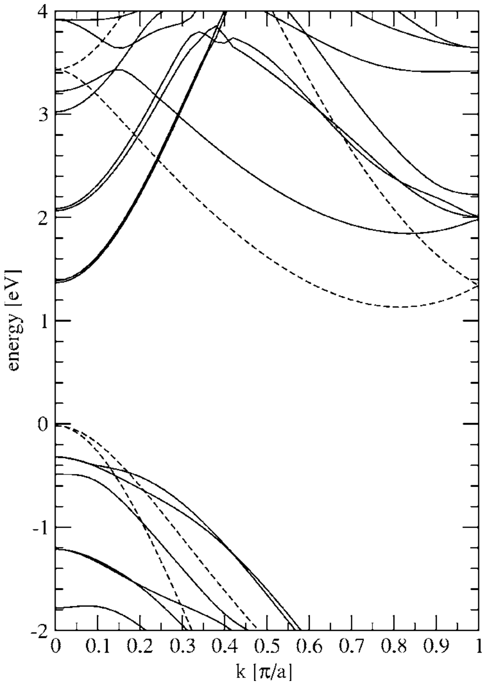

FIG. 3. Solid lines: tight-binding band structure of a silicon film of two cubic cell thickness with hydrogen passivation; dashed lines: bulk silicon. **k** along ⟨100⟩.

of the EMA in thin silicon films, we performed empirical tight-binding (TB) calculations of thin silicon film with and without hydrogen passivation. TB parameters were taken from Refs. 21 and 22. Figure 3 shows the band structure of a H-passivated silicon film of a thickness of two cubic cells (approximately 1.1 nm). The band energies are strongly shifted due to confinement, but the curvature near the conduction band extrema is only slightly changed relative to bulk silicon. We find transport masses of $0.21m_0$ and $0.78m_0$ for the valleys at $k_\parallel=0$ and $k_\parallel \approx 0.84\pi/a$, respectively—about a 10% change with respect to bulk silicon values. These values are very sensitive to details of the surface passivation—with dangling bonds instead of hydrogen passivation, the transport mass in the $k_\parallel \neq 0$ valley decreases from $0.78m_0$ to $0.58m_0$. This indicates that the impact of the silicon surface and its passivation are stronger than the finite size effect; the simplistic assumption of hydrogen passivation might not lead to accurate band curvatures. Hence, the results in the present work were obtained using bulk values for the effective masses. Band structure effects down to a film thickness of 1 nm are not expected to change these results in a qualitative way.

VI. SELF-CONSISTENT VERSUS LOCAL EQUILIBRIUM CHARGE DENSITY

The scattering matrix approach described in Sec. III may be used in a postprocessing step for the computation of the current, in which case the charge density for the nonlinear Poisson equation is constructed by populating the transverse wave functions $|i\rangle^{(n)}$ on each slice in a classical manner. Then, only the components of $\mathbb{S}^{(0,n_{\max})}$ that occur in the expressions for the transmission probabilities T and reflection probabilities R in Eqs. (23) and (24) need to be computed. Current-voltage (I-V) characteristics making use of this simplification are labeled "non-self-consistent" (nsc) in Figs. 5 and 6. Alternatively, the full scattering matrix formalism may be used for the computation of the carrier concentration $\rho(\mathbf{r}) = \mathrm{tr}(|\mathbf{r}\rangle\langle\mathbf{r}|M)$ inside the solution process for the nonlinear Poisson equation,

$$\rho(\mathbf{r}) = g_{\text{val}} \sum_{\alpha} \sum_{v} \sum_{i} \int_{\epsilon_{v,i,\alpha}}^{\infty} Z^{\text{inc}}_{v,i,\alpha}(\epsilon) f\left(\frac{\epsilon-\mu_\alpha}{k_B T}\right) \times |\psi_{v,i,\alpha,\epsilon}(\mathbf{r})|^2 d\epsilon. \tag{28}$$

Here, v denotes the valley index (a separate effective mass equation is solved in each valley) and $\epsilon_{v,i,\alpha}$ is the subband bottom energy of subband i in valley v at terminal α. The densities of state Z^{inc} are one-sided densities of state inclusive of spin degeneracy. The index ϵ in the wave function ψ denotes the injection energy. Evaluation of the injected wave function $\psi_{v,i,\alpha,\epsilon}$ at arbitrary x positions is equivalent to the computation of $\mathbf{a}^{(n)}$ and $\mathbf{b}^{(n)}$ for any n. To link $\mathbf{a}^{(n)}$ and $\mathbf{b}^{(n)}$ to the scattering boundary vectors $\mathbf{a}^{(0)}$ and $\mathbf{b}^{(n_{\max})}$, the two scattering matrix equations,

$$\begin{bmatrix} \mathbf{a}^{(n)} \\ \mathbf{b}^{(0)} \end{bmatrix} = \mathbb{S}(0,n) \begin{bmatrix} \mathbf{a}^{(0)} \\ \mathbf{b}^{(n)} \end{bmatrix},$$
$$\begin{bmatrix} \mathbf{a}^{(n_{\max})} \\ \mathbf{b}^{(n)} \end{bmatrix} = \mathbb{S}(n,n_{\max}) \begin{bmatrix} \mathbf{a}^{(n)} \\ \mathbf{b}^{(n_{\max})} \end{bmatrix}, \tag{29}$$

may be employed.

After eliminating the quantities $\mathbf{a}^{(n_{\max})}$ and $\mathbf{b}^{(0)}$, the resulting expressions for $\mathbf{a}^{(n)}$ and $\mathbf{b}^{(n)}$ are

$$\mathbf{a}^{(n)} = [\mathbb{1} - \mathbb{S}_{0,1}(0,n)\mathbb{S}_{1,0}(n,n_{\max})]^{-1} \times [\mathbb{S}_{0,0}(0,n)\mathbf{a}^{(0)} + \mathbb{S}_{0,1}(0,n)\mathbb{S}_{1,1}(n,n_{\max})\mathbf{b}^{(n_{\max})}], \tag{30}$$

$$\mathbf{b}^{(n)} = [\mathbb{1} - \mathbb{S}_{1,0}(n,n_{\max})\mathbb{S}_{0,1}(0,n)]^{-1} \times [\mathbb{S}_{1,0}(n,n_{\max})\mathbb{S}_{0,0}(0,n)\mathbf{a}^{(0)} + \mathbb{S}_{1,1}(n,n_{\max})\mathbf{b}^{(n_{\max})}]. \tag{31}$$

This approach is computationally considerably more costly, since it requires the computation of the full Kohn-Sham scattering basis for each evaluation of the charge density functional. Results making use of this extended scheme are labeled "self-consistent" (sc) in the aforementioned graphs. The sc approach maintains QM effects along the transport direction of the transistor. In the nsc approach only transverse quantization is considered; longitudinal quantum effects are disregarded. The resulting differences in charge density inside the device are depicted in Fig. 4.

If a finite bias is applied between the terminals of a ballistic device, the Fermi energy inside the device is no longer defined, and the charge density may differ significantly from an equilibrium charge density. As long as the tunneling electrons have negligible influence on the electrostatics, this nonequilibrium charge density may be modeled on the level of a classical treatment of the transport direction. Up to the barrier peak energy $\epsilon_{\max}$ both left and right propagating states are populated according to $\epsilon_F^{(\text{src})}$ for $x < x_{\max}$ ($\epsilon_F^{(\text{drn})}$ for $x \geq x_{\max}$) because all injected electrons are reflected back to their injection terminal. For energies greater than

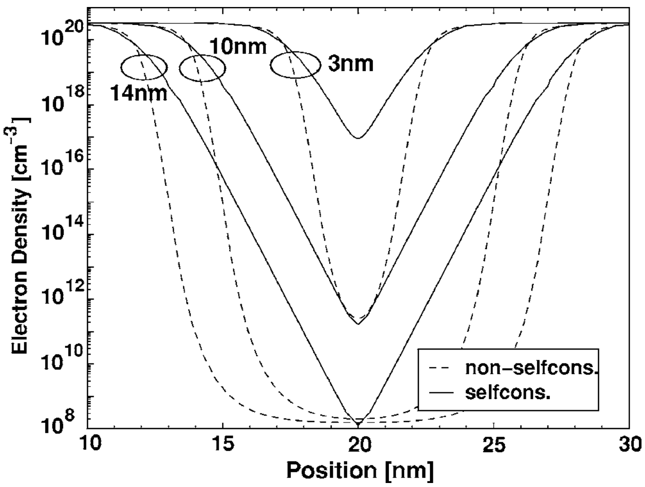

FIG. 4. Lateral density profiles in the middle of the channel for devices with $t_{\rm Si}=1$ nm and gate lengths of 3, 10, and 14 nm at $V_{\rm DS}=1$ μV and $V_{\rm GS}=0$ V. In very short devices, wave function penetration into the barrier may have a marked effect on the electrostatics.

$\epsilon_{\max}$ all electrons are transmitted over the barrier. Therefore, right propagating states are populated according to $\epsilon_F^{(\rm src)}$, whereas left propagating states are populated with $\epsilon_F^{(\rm drn)}$.[23,24] In the case of sizable tunneling currents, the charge density has to be computed from the full wave functions according to Eq. (28). This expression automatically treats different Fermi energies at the terminals in the correct way.

VII. RESULTS

A. QB transport at infinitesimal forward bias

DGSOI MOSFETs of the type shown in Fig. 2 with $\ell_{\rm gate}$ in the range of 3–20 nm were simulated using the two-dimensional (2D) QB transport model of Secs. II–VI. The resulting *I-V* characteristics are shown in Fig. 5 for devices with $t_{\rm Si}=1$ nm and in Fig. 6 for devices with $t_{\rm Si}=3$ nm, respectively. The solid curves were obtained by sc solving the nonlinear Poisson equation with the charge density from the full scattering matrix formalism; dashed lines result from simulations with local equilibrium population of the channel cross sections (nsc). Concerning source-drain tunneling, we find the same situation for both $t_{\rm Si}=1$ nm and $t_{\rm Si}=3$ nm: in transistors with "long" gates ($\ell_{\rm gate}\gtrsim 10$ nm) the entire drain current is thermionic current over the potential barrier between source and drain. This is apparent from direct comparison with ballistic transport results in which longitudinal tunneling was suppressed [$\hbar\rightarrow 0$ in the lower branch of Eq. (14)] as well as from the fact that the subthreshold slope of the (sc and nsc) QB *I-V* curves is constant and identical to the subthreshold slope obtained from (constant mobility) DD simulations. In devices of "intermediate" gate length ($\sim$7 nm) the subthreshold characteristics exhibit a steep slope similar to the thermionic slope for gate voltages not too far below the threshold voltage; far from the threshold voltage, the steepness of the curve diminishes, and a second shallower slope is established. This happens, when the thermionic contribution to the current becomes so small that source-drain tunneling dominates the charge transport from source to drain. In very short devices (3 and 5 nm) tunneling is dominant throughout the subthreshold range. This corroborates the results of $\ell_{\rm gate}$ comparisons based on the WKB approximation along the transport direction.[25]

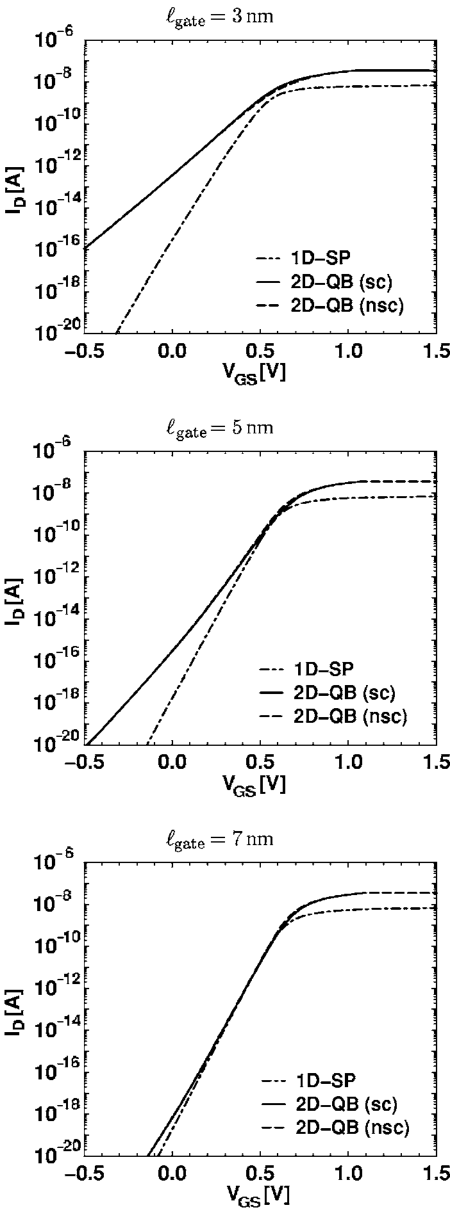

FIG. 5. Transfer characteristics with different transport models at $V_{\rm DS}=1$ μV for $t_{\rm Si}=1$ nm and gate length $\ell_{\rm gate}$.

B. Comparison: 1D Schrödinger-Poisson with QM mobility versus standard DD

Besides the QB simulations, the DGSOI MOSFETS were also studied by DD simulations on top of a 1D-SP charge density using a QM mobility model, that correctly accounts for transverse quantization effects in long channel MOSFETs.[26] The on-current is strongly degraded by ionized impurity scattering in the source/drain regions (approximately by a factor of 3.5 in the case of $\ell_{\rm gate}=10$ nm). As this resistance effect depends on length, geometrical shape, and doping level of the contact regions, impurity scattering was excluded. For further comparison, DD simulations with constant mobility were performed. For each of these simula-

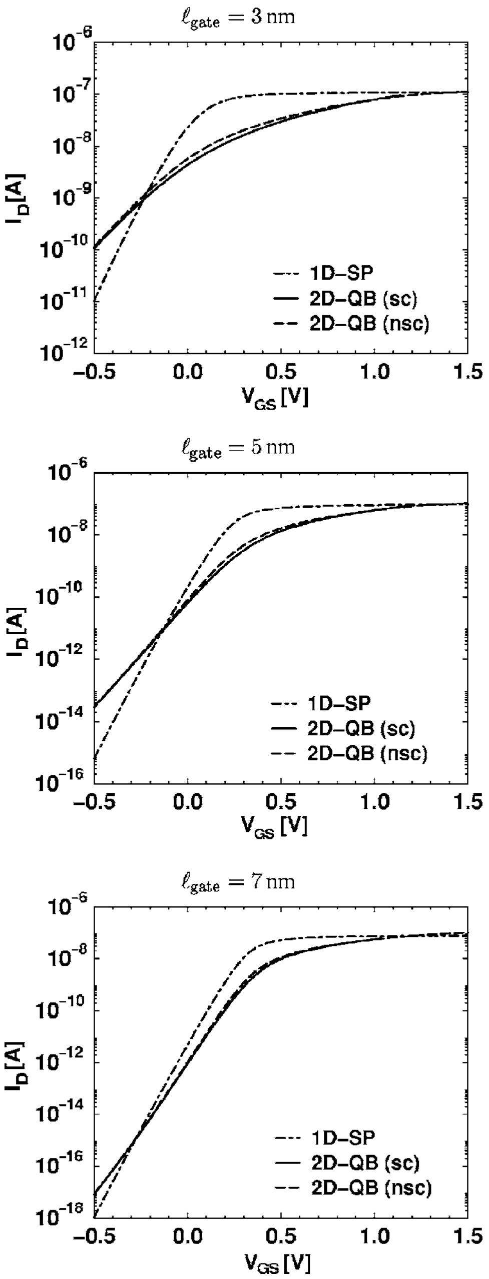

FIG. 6. Transfer characteristics with different transport models at $V_{\mathrm{DS}} = 1\ \mu\mathrm{V}$ for $t_{\mathrm{Si}} = 3$ nm and gate length ℓ_{gate}.

tions, the value of μ_{const} was chosen such that the on-current matched the corresponding DD/1D-SP simulation. Interestingly, apart from the QM threshold-voltage shift, only very minor differences were observed between the two DD-based approaches. Despite strong confinement orthogonal to the transport direction, the influence of transverse quantum effects on the *I-V* characteristics of the transistors with $t_{\mathrm{Si}} = 1$ nm turned out to be representable by a simple voltage shift of the whole curve. Devices of different gate lengths give rise to different effective mobilities, but the *I-V* curves of each transistor are governed by a single mobility value. Devices with a silicon body thickness of 3 nm, however, exhibited a considerable dependence of the mobility on the gate voltage even within a single *I-V* curve. This effect originates from the fact that in the $t_{\mathrm{Si}} = 1$ nm devices geometric confinement is so strong that the wave functions underneath the gate are not significantly deformed by the gate potential; in the 3 nm devices, on the other hand, the shape of the transverse wave functions and of the charge density distribution on the cross-sectional plane underneath the gate both may strongly be affected by a change in gate potential.

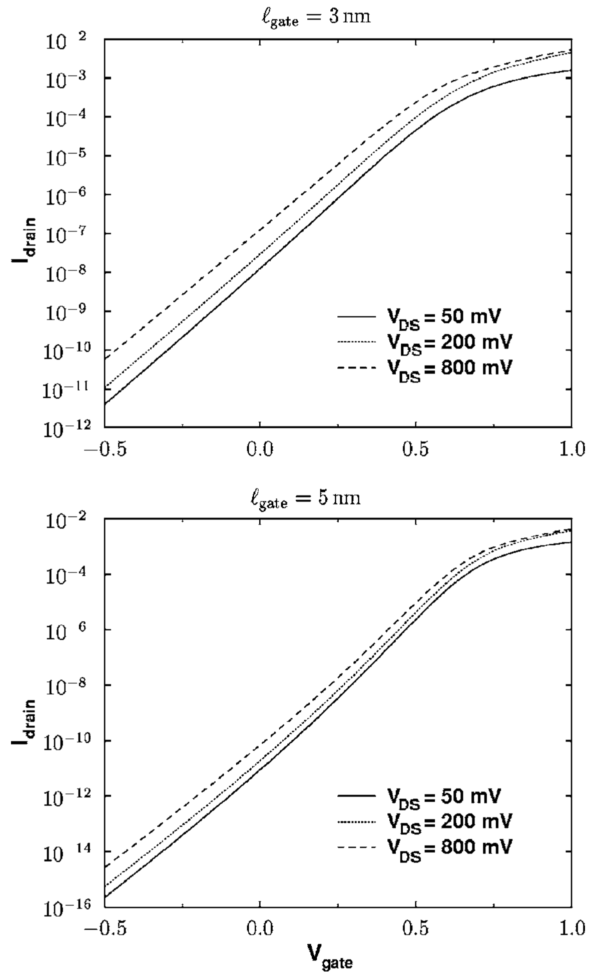

FIG. 7. I_{drain} (A/μm) vs V_{gate} (V) of DGSOI MOSFETs of 1nm body thickness and gate length ℓ_{gate} at various source-drain voltages.

C. QB transport: Finite forward bias

Figure 7 shows self-consistent 2D QB transfer characteristics for various values of the source-drain bias in transistors with $t_{\mathrm{Si}} = 1$ nm and gate lengths of 3 and 5 nm. For both channel lengths it can be seen that there is a considerable increase in drain current as the drain voltage increases from 50 to 200 mV. As drain voltage is increased further, the above-threshold current saturates faster in the longer transistors.[30] Regardless of the transistor length, current always was seen to increase with $V_{\mathrm{src\text{-}drn}}$ in the deep subthreshold regime.

The reason for this behavior is that increasing the source-drain voltage results only in a small decrease in barrier height—in the 5 nm transistor the drain induced barrier lowering at a forward bias of 800 mV was found to be 18 mV, but there is a significant reduction in the width of the barrier. Therefore, the tunneling contribution to the current increases much stronger with increasing source-drain bias than the thermionic current over the barrier. In very short devices, the tunneling contribution to the total drain current may be significant over a large part of the *I-V* curve. In longer devices, however, tunneling only shows up deep in the subthreshold regime; everywhere else, the current is almost entirely thermal. This discussion applies only to voltages $V_{\text{src-drn}} \gtrsim 100$ mV. For smaller voltages backward electron injection from the drain reduces the current, leading to an approximately linear $V_{\text{src-drn}}$ dependence of the current.

VIII. CONCLUSION

The simulations indicate that source-drain tunneling deteriorates the subthreshold behavior of ultrathin-film sub-10 nm MOSFETs. But they also show that such devices in principle are still operational. The simulations presented in the present work were obtained with a $\langle 100 \rangle$ orientation of the channel. Here, two of the three valley pairs of the silicon band structure have the small (transverse) component of the silicon effective mass tensor aligned with the transport direction. Hence, the tunneling current is at a maximum for this channel orientation. Other orientations of the channel may exhibit smaller tunneling contributions to the channel, but the summation over the various band minima calls for considerable care. For a $\langle 110 \rangle$ orientation of the channel, for example, the tunneling mass becomes $m_x' = \frac{1}{2}(m_x + m_y)$, a value of nearly three times m_t for the valleys with m_t along the quantization direction (z axis). This would suppress source-drain tunneling of electrons in these valleys down to gate lengths of about 3 nm. The lower energy valleys with the large (longitudinal) mass component in quantization direction, however, maintain their low tunneling mass for all channel orientations orthogonal to the (001) direction commonly found in silicon wafers. Hence, no qualitative change of the situation can be expected without changing the crystal orientation normal to the surface of the wafer. Should source-drain tunneling ever become a source of major concern, switching to a crystal orientation for which the quantization direction is not parallel to one of the principal axes of the effective mass tensor might help.

ACKNOWLEDGMENTS

This work was carried out in the context of the European Union research project IST-1999-10828 (NANOTCAD) with financial support by the Swiss Federal Office for Education and Science (BBW).

[1] Th. Ernst, S. Cristoloveanu, G. Ghibaudo, Th. Ouisse, S. Horiguchi, Y. Ono, Y. Takahashi, and K. Murase, IEEE Trans. Electron Devices **50**, 830 (2003).

[2] S. E. Laux, A. Kumar, and M. V. Fischetti, J. Appl. Phys. **95**, 5545 (2004).

[3] L. P. Kadanoff and G. Baym, *Quantum Statistical Mechanics* (Benjamin, New York, 1962).

[4] W. R. Frensley, Phys. Rev. B **36**, 1570 (1987).

[5] M. V. Fischetti, J. Appl. Phys. **83**, 270 (1998).

[6] R. Landauer, Z. Phys. B: Condens. Matter **68**, 217 (1987).

[7] R. Lake, G. Klimeck, R. C. Bowen, and D. Jovanovic, J. Appl. Phys. **81**, 7845 (1997).

[8] J. H. Rhew, Z. B. Ren, and M. S. Lundstrom, Solid-State Electron. **46**, 1899 (2002).

[9] S. Datta, Superlattices Microstruct. **28**, 253 (2000).

[10] Z. Ren, R. Venagupal, S. Goasguen, S. Datta, and M. S. Lundstrom, IEEE Trans. Electron Devices **50**, 1914 (2003).

[11] R. Venugopal, Z. Ren, and M. S. Lundstrom, IEEE Trans. Nanotechnol. **2**, 135 (2002).

[12] J. Wang, E. Polizzi, and M. S. Lundstrom, J. Appl. Phys. **96**, 2192 (2004).

[13] D. Mamaluy, A. Mannargudi, and D. Vasileska, J. Comput. Electron. **3**, 45 (2004).

[14] D. Y. K. Ko and J. C. Inkson, Phys. Rev. B **38**, 9945 (1988).

[15] A. C. Marsh and J. S. Inkson, Semicond. Sci. Technol. **1**, 285 (1986).

[16] A. Wettstein, A. Schenk, A. Scholze, G. Garretón, and W. Fichtner, Proceedings of the Spring Conference of the Electrochemical Society, Montréal, May 4–9, 1997, 191st Society Meeting, Vol. 97(1), p. 616.

[17] A. Schenk and A. Wettstein, Proceedings of the 2002 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD 2002), Kobe, Japan, 4–6 September 2002, p. 21.

[18] G. Curatola, G. Doornbos, J. Loo, Y. V. Ponomarev, and G. Iannaccone, IEEE Trans. Electron Devices **50**, 1851 (2005).

[19] N. W. Ashcroft and N. D. Mermin, *Solid State Physics* (Saunders College, Philadelphia/Harcourt Brace Jovanovitch, San Diego, 1976).

[20] G. Wachutka, Ph.D. thesis, Ludwig-Maximilians-Universität München, 1985.

[21] T. B. Boykin, G. Klimeck, and F. Oyafuso, Phys. Rev. B **69**, 115201 (2004).

[22] Y. Zheng, C. Rivas, R. Lake, K. Alam, T. B. Boykin, and G. Klimeck, IEEE Trans. Electron Devices **52**, 1097 (2005).

[23] G. Fiori and G. Iannaccone, Appl. Phys. Lett. **81**, 3672 (2002).

[24] G. Curatola, G. Fiori, and G. Iannacconne, Solid-State Electron. **48**, 581 (2004).

[25] D. Muntaneu and J. L. Autran, Solid-State Electron. **47**, 1219 (2003).

[26] A. Wettstein, *Quantum Effects in MOS Devices* (Hartung Gorre, Konstanz, 2000).

[27] In 1D, the required number of intervals may further be reduced by using a piecewise linear approximation of the potential and Airy functions instead of exponentials. This approach, however, does not readily generalize to higher dimensions: allowing a linear x dependence of the potential on an interval will in general cause the transverse wave functions to become nonconstant.

[28] Terminology seems to be nonuniform at this point. For example, Ref. 14 uses the definition of Eq. (20) for the scattering matrix; in Ref. 15 the name *scattering matrix* is used for the inverse of the transfer matrix. Other authors use yet other definitions.

[29] The *interface matrices* $\mathbf{I}(n+1)$ of Ref. 14 correspond to the *reverse transfer matrices* $\mathbb{M}^{(n)^{-1}}$ of the present work.

[30] For all voltages shown, there still is a small potential barrier left inside the channel.