# Two-Dimensional Tunneling Effects on the Leakage Current of MOSFETs With Single Dielectric and High-$\kappa$ Gate Stacks

Mathieu Luisier and Andreas Schenk

***Abstract*—The gate leakage currents of single-gate silicon-on-insulator (SOI) n-type MOSFETs are investigated, assuming direct tunneling as the leakage mechanism and using either a 1-D Schrödinger–Poisson-based approach coupled to the conventional drift-diffusion transport model or a full quantum mechanical treatment. The first approach consists of calculating the transmission probability through the dielectric material along straight lines connecting the transistor channel to the gate. The second method is based on a 2-D Schrödinger–Poisson solver, where carriers are injected into the device from the source, drain, and gate contacts. The simulated structures have a physical gate length of 32 nm. The channel is isolated from the gate contact by a dielectric layer with an equivalent oxide thickness of 1.2 nm. This layer is composed of either pure $SiO_2$ or a high-$\kappa$ $SiO_2 - HfO_2$ stack. Irrespective of the dielectric material, the leakage currents calculated with the 1-D approach are about one order of magnitude smaller at low gate voltages and converge toward the same value as the channel potential barrier decreases. The difference is caused by the diffraction of the electron waves at both edges of the gate contact. This peculiar 2-D behavior of the gate leakage currents, as well as the limit of the 1-D model, is discussed in this paper for various dielectric configurations.**

***Index Terms*—Electron diffraction, gate leakage current, high-$\kappa$ gate stacks, silicon-on-insulator (SOI) MOSFETs, 2-D Schrödinger–Poisson solver.**

## I. Introduction

TO IMPROVE the performances of electronic devices, the size of their active components is scaled down according to the International Technology Roadmap for Semiconductors (ITRS) [1]. In this context, the current bulk complementary metal–oxide–semiconductor FETs are evolving toward nanoscale ultrathin body silicon-on-insulator (SOI) structures, which suffer less from short-channel effects and offer steeper subthreshold slopes [2]. This favorable behavior is attributed to better electrostatic control obtained by reducing the thickness of the silicon body on top of the buried oxide [3]. At the same time, the thickness of the traditional gate dielectric material $SiO_2$ has been scaled down below 2 nm [4]. At this size, the gate leakage currents can reach a value of 10 A/cm$^2$ [5] due to their exponential dependence on the oxide thickness. They may become the dominant leakage mechanism, deteriorate the device reliability, and cause most of the power consumption. To circumvent this problem, $SiO_2$ is replaced by "high-$\kappa$" dielectrics as $HfO_2$ or $ZrO_2$ [6] with an equivalent oxide thickness (EOT). These materials reduce the tunneling leakage due to the larger physical thickness but provide the same gate capacitance due to their higher permittivity. However, they also exhibit an increased interfacial trap concentration as compared to $SiO_2$ and cause a degradation of the channel mobility [7]. Hence, the high-$\kappa$ materials are often used in combination with a thin $SiO_2$ layer, forming a so-called gate stack.

In this paper, the gate leakage currents are studied from the device simulation perspective. Based on a n-doped single-gate SOI transistor designed to fulfill the requirements of the 32-nm technology node [1], we illustrate the following: 1) the benefit of $SiO_2 - HfO_2$ gate stacks over pure $SiO_2$; 2) the optimal choice of the spacer layers that isolate the gate from the flared out source and drain contacts; 3) the behavior of the gate current in the vicinity of the gate corners; and 4) the attribute of two different tunneling models.

In contemporary device simulators, the gate leakage currents are often calculated from the tunneling probability through the oxide barrier. This can be done in different ways. Very popular approaches include the Wentzel–Kramers–Brillouin (WKB) approximation [8] and the use of Airy or trigonometric functions in connection with Bardeen's perturbation theory [9], [10]. They all work well for a simple gate layout comprising one single dielectric layer only but fail or are not advantageous in the presence of gate stacks. Furthermore, their accuracy becomes questionable for ultrathin oxides. Therefore, more sophisticated models are required to simulate the gate leakage of next-generation devices. In this paper, two approaches are emphasized. They differ by their basic physics, their complexity, and their computational burden, but they qualitatively result in very similar characteristics. However, important features such as electron diffraction at the gate corners are only captured by the more evolved model.

This paper is organized as follows: In Section II, the two approaches previously mentioned are described in detail. The governing equations and their numerical implementation are reviewed. Then, the dimensions and characteristics of the

Manuscript received December 5, 2007. This work was supported in part by the Swiss National Science Foundation under Project NEQUATTRO SNF 200020-117613/1 and in part by the European Union under Project IST-4-026828-IP-PULLNANO. The review of this paper was arranged by Editor M. Reed.

M. Luisier was with Institut für Integrierte Systeme, ETH Zürich, 8092 Zürich, Switzerland. He is now with Network for Computational Nanotechnology, Purdue University, West Lafayette, IN 47907 USA (e-mail: mluisier@iis.ee.ethz.ch).

A. Schenk is with the Integrated Systems Laboratory, ETH Zürich, 8092 Zürich, Switzerland.

Digital Object Identifier 10.1109/TED.2008.922493

selected 32-nm SOI transistor are outlined in Section III. The material parameters used in the simulation, e.g., the effective masses, dielectric constants, and conduction band offsets, are also summarized. Simulation results are presented in Section IV. The two theoretical approaches are compared, and their advantages and disadvantages are discussed. Finally, this paper is concluded in Section V.

## II. Theoretical Method

In the first method, 1-D Schrödinger equations are solved along straight lines connecting the channel to the gate contact [11], [12]. This calculation is done on each line of a special-purpose grid, the generation of which is described here. The results are then self-consistently incorporated into a 2-D/3-D drift-diffusion or energy-balance simulator [12].

The special-purpose grid needed for the solution of the 1-D Schrödinger–Poisson system consists of straight lines that are attached to a semiconductor vertex and connect this vertex to the closest grid point on the gate contact. Vertices up to a distance of 5 nm may be connected this way to the gate electrode. In addition, those points not directly situated under the gate can be connected to the gate corners by defining a maximum possible angle measured to the normal of the gate contact line. Two length parameters allow the inclusion of regions below and above the stack. Hence, the transmission probability can be computed not only for the oxide barrier but also for a potential barrier in the semiconductor that might exist along the line. In practice, due to the exponential dependence of the tunneling probability on the tunneling length, only small angles and only small segments outside the stack are relevant.

Using interpolation schemes, all data, as well as the refinement of the initial mesh, are transferred to the special-purpose grid. The 1-D Schrödinger equation is solved in the effective mass approximation (EMA) using the scattering matrix approach [13]. This can be done in either a one-band or a two-band model for the oxide layer. Coordinates on the lines of the special-purpose grid are denoted by $u$ or $r$, and have their origin at the metal contact ($0^-$ is infinitesimally smaller than the origin). The electron current density due to direct tunneling gate leakage (by conduction band electrons only) can be written as [10], [12]

$$j_n = -\frac{g_n A_0 T}{k_B}\int_{0^-}^{\infty} du\,\mathcal{T}_n\left[u, 0^-, E_c(u)\right]\left|\frac{dE_c}{du}(u)\right|\Theta\left[-\frac{dE_c}{du}(u)\right] \times \ln\left\{\frac{\exp\left[\frac{E_{\mathrm{F,n}}(u)-E_c(u)}{k_B T}\right]+1}{\exp\left[\frac{E_{\mathrm{F,n}}(0^-)-E_c(u)}{k_B T}\right]+1}\right\}. \quad (1)$$

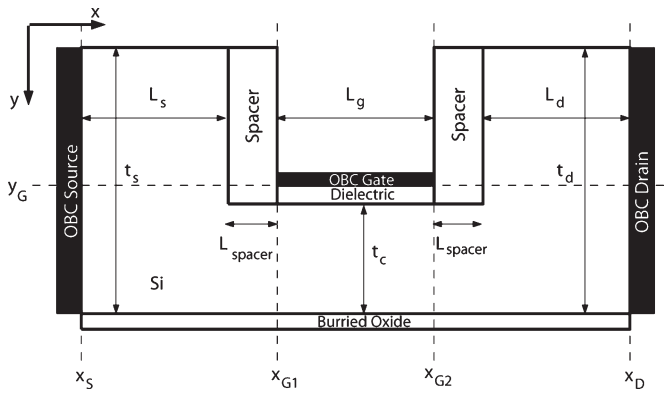


Fig. 1. Schematic view of a 32-nm SOI structure. The variables $t_s$, $t_c$, and $t_d$ refer to the thickness of the source, the channel, and the drain, $L_s$, $L_g$, $L_d$, and $L_{\mathrm{spacer}}$ refer to the length of the source, the gate, the drain, and the spacers, respectively. The three black layers labeled "OBC" represent regions where OBCs are defined. The source, drain, and gate are situated at $x = x_S$, $x = x_D$, and $y = y_G$, respectively. The latter starts at $x = x_{G1}$ and ends at $x = x_{G2}$.

Here, $A_0 = 4\pi m_0 k_B^2 q/h^3$ is the Richardson constant for free electrons, $T$ denotes the temperature (drift-diffusion model, no carrier heating), $k_B$ is the Boltzmann constant, $E_c(u)$ is the position-dependent conduction band edge, $E_{\mathrm{F,n}}(u)$ is the quasi-Fermi energy, and $\mathcal{T}_n$ is the tunneling probability. In the WKB approximation, the latter parameter would read (neglecting preexponential factors) as (2), shown at the bottom of the page.

Note that, in this paper, $\mathcal{T}_n$ was always computed by solving the Schrödinger equation with a one-band model.

Parameter $g_n$ can be used to change the effective density-of-state mass in the Richardson constant. For tunneling across a (100)-oriented interface, a reasonable choice is $g_n = 2m_t/m_0$ for the valley pair perpendicular to the interface and $g_n = 4\sqrt{m_t m_l}/m_0$ for the two valley pairs parallel to the interface. Separate simulations of the current have to be performed, because the effective mass of Si that enters the transmission probability $\mathcal{T}_n$ also changes.

The second approach presented in this paper treats the device and the gate contact as a single entity on a quantum mechanical level. A 2-D and real-space Schrödinger–Poisson solver is used to calculate the carrier and current densities, as well as the electrostatic potential of the device. It allows electrons (or holes) to enter and exit the simulation domain at the source, drain, and gate contacts. This is not possible with the mode-space approximation that separates the longitudinal ($x$-axis) and transverse ($y$-axis) directions [14]–[17].

A schematic view of a single-gate SOI MOSFET is given in Fig. 1. The simulation domain is discretized in the finite-difference method, so that $x(y)$ becomes a vector with $N_x(N_y)$ entries $x_i(y_j)$. The $z$-direction is assumed to be infinite and

$$\mathcal{T}_n\left[u, 0^-, E_c(u)\right] = \exp\left\{-2\int_{0^-}^{u} dr\sqrt{2m_c(r)\left|E_c(r)-E_c(u)\right|}/\hbar\Theta\left[E_c(r)-E_c(u)\right]\right\} \quad (2)$$

induces a $k_z$ dependence (not shown here) that modifies the injection probability of the electrons [17]. The EMA Schrödinger equation at each $(x_i, y_j)$ point can be written as

$$(E - H_{\mathrm{iijj}})\phi_{\mathrm{ij}} - H_{\mathrm{ii+1jj}}\phi_{i+1j} - H_{\mathrm{ii-1jj}}\phi_{i-1j} - H_{\mathrm{iijj+1}}\phi_{\mathrm{ij+1}} - H_{\mathrm{iijj-1}}\phi_{\mathrm{ij-1}} = 0 \quad (3)$$

where the Hamiltonian matrix element $H_{i_1 i_2 j_1 j_2}$ are defined as in [17], $\phi_{\mathrm{ij}}$ is the wave function $\phi(x_i, y_j)$, and $E$ is the injection energy. At the source, drain, and gate contacts, a single band scattering boundary ansatz [18], [19] is applied to model the open boundary conditions (OBCs)

$$(\mathbf{E} - \mathbf{H}_{\mathrm{nn}})\phi_{\mathbf{n}} - \mathbf{H}_{\mathrm{nn+1}}\phi_{n+1} - \mathbf{H}_{\mathrm{nn-1}}\phi_{n-1} = 0 \quad (4)$$

$$\phi_{n\pm1} = \phi_n e^{\pm i\mathbf{k}_n\Delta_n}. \quad (5)$$

Energy matrix $\mathbf{E}$ is diagonal, $\mathbf{H}_{\mathrm{nn}}$ is tri-diagonal and describes the on-site energies and connections within one grid line, $\mathbf{H}_{\mathrm{nn}\pm1}$ is diagonal and represents the connection of one grid line to the next ($+$ sign) or previous ($-$ sign) one, and $\phi_n$ is a vector containing the wave function along one grid line. Index $n$ is equal to $S$, $D$, or $G$ and is used to characterize the position of one grid line. For example, if the source contact is considered, $n$ denotes the vertical grid line situated at $x = x_S(\phi_n = \phi(x_S, y), \phi_{n\pm1} = \phi(x_S \pm \Delta_x, y))$. For the gate contact, $n$ refers to the horizontal line with the $y$ coordinates equal to $y_G(\phi_n = \phi(x, y_G), \phi_{n\pm1} = \phi(x, y_G \pm \Delta_y))$. In the scattering boundary theory, the contacts are just the extension of the device grid line they are connected to. Hence, the contact wave functions taken along one grid line orthogonal to the injection direction are identical up to a phase factor $e^{i\mathbf{k}_n\Delta_n}$. The variable $\Delta_n$ is the distance between two lines. Furthermore, $\mathbf{H}_{\mathrm{nn+1}} = \mathbf{H}^{\mathbf{T}}_{n+1n}$ holds, and it can be proven that

$$\mathbf{H_{nn+1}} = \mathbf{H_{nn-1}} = \mathbf{T_n} \quad (6)$$

is valid in the contacts since both $\mathbf{H_{nn+1}}$ and $\mathbf{H_{nn-1}}$ are diagonal matrices. Inserting (6) into (5) results in the eigenvalue problem

$$\underbrace{\mathbf{T_n^{-1}} \cdot (\mathbf{E} - \mathbf{H_{nn}})}_{\mathbf{M}} \phi_n = \underbrace{-2 \cdot \cos(\mathbf{k_n})}_{\lambda} \cdot \phi_n. \quad (7)$$

All the eigenvalues $\boldsymbol{\lambda}$ of $\mathbf{M}$ are required. The computational burden associated with (7) increases in a cubic way as a function of the number of grid points taken along the open boundaries of the simulation domain. Therefore, a straightforward technique to symmetrize matrix $\mathbf{M}$ and to reduce the computational time is explained in Appendix A. Boundary wave functions $\phi_{\mathbf{S}}$, $\phi_{\mathbf{D}}$, and $\phi_{\mathbf{G}}$ and vectors $\mathbf{k_S}$, $\mathbf{k_D}$, and $\mathbf{k_G}$ resulting from (7) are used to calculate the source, drain, and gate boundary self-energies $\boldsymbol{\Sigma}_{\mathbf{S}}$, $\boldsymbol{\Sigma}_{\mathbf{D}}$, and $\boldsymbol{\Sigma}_{\mathbf{G}}$, respectively, as well as injection matrix $\mathbf{S_{inj}}$ [18], [19]. Finally, (3) is cast into a sparse linear problem with the following form:

$$\underbrace{(\mathbf{E} - \mathbf{H} - \boldsymbol{\Sigma}_{\mathbf{S}} - \boldsymbol{\Sigma}_{\mathbf{D}} - \boldsymbol{\Sigma}_{\mathbf{G}})}_{\mathbf{A}} \cdot \phi = \mathbf{S_{inj}}. \quad (8)$$

Matrices $\boldsymbol{\Sigma}_{\mathbf{S}}$ and $\boldsymbol{\Sigma}_{\mathbf{D}}$ vanish everywhere, except in the left and right corners of $\mathbf{A}$, $\boldsymbol{\Sigma}_{\mathbf{G}}$ occupies a large sparse block in the middle of $\mathbf{A}$ and destroys its block tridiagonal structure inherited from $\mathbf{H}$ [20]. The $N_S$, $N_D$, and $N_G$ states injected from the source, the drain, and the gate, respectively, are included in the $(N_S \cdot N_D \cdot N_G) \times (N_x \cdot N_y)$ matrix $\mathbf{S_{inj}}$.

The linear system in (8) is solved with a direct sparse linear solver like Umfpack 5.0.1 [21], Pardiso [22], or MUMPS 4.6.3 [23]. On a nonuniform finite-difference grid, matrix $\mathbf{A}$ is not symmetric, but it is possible to perform a basis transformation to obtain this highly desired property, as derived in Appendix A. Then, the factorization of matrix $\mathbf{A}$ is simplified, and the computational burden decreases. The advantage of working in the wave function formalism as in (8) over nonequilibrium Green's functions (NEGFs) is that all the elements of the boundary self-energy $\boldsymbol{\Sigma}_{\mathbf{G}}$ are easily taken into account. In the NEGF approach proposed in [20] or [24], only the first off-diagonal blocks of $\mathbf{A}$ are kept, but the higher order elements are neglected. To the best of our knowledge, the consequences of this omission have never been investigated. An NEGF alternative to this truncation scheme would be the contact block reduction method [25].

Equation (8) is solved for each injection energy $E$ and for the six degenerate conduction band valleys of Si. Once the wave functions $\phi(E)$ are known, carrier density $n(x_i, y_j)$ and ballistic current density $\mathbf{J}(x_i, y_j)$ are calculated according to

$$n(x_i, y_j) = \frac{1}{\Delta_x \Delta_y} \sum_n \int \frac{dE}{2\pi} \left|\phi_n(x_i, y_j; E)\right|^2 \quad (9)$$

$$\mathbf{J}(x_i, y_j) = -\frac{2e}{\Delta_x \Delta_y \hbar} \sum_n \int \frac{dE}{2\pi} \mathbf{Re} \times \begin{pmatrix} \phi_n^*(x_{i+1}, y_j; E) \cdot H_{i+1\mathrm{ijj}} \cdot \phi_n(x_i, y_j; E) \cdot \Delta_x \\ \phi_n^*(x_i, y_{j+1}; E) \cdot H_{\mathrm{iij+1}j} \cdot \phi_n(x_i, y_j; E) \cdot \Delta_y \end{pmatrix}. \quad (10)$$

Index $n$ ($S$, $D$, or $G$) refers to the origin of the wave function. It indicates from which port the state was injected. The Fermi levels of the contacts are already taken into account in the wave function. They determine the probability that a state injected at an energy $E$ is occupied [19]. The ultimate carrier density $n(x_i, y_j)$ is obtained after a self-consistent calculation of the 2-D electrostatic potential in the device. The drain and gate currents (in amperes per meter) then follow from

$$I_d(x_i) = \int dy J_x(x_i, y) \quad (11)$$

$$I_g(y_G) = \int_{x_{G1}}^{x_{G2}} dx J_y(x, y_G). \quad (12)$$

Current continuity implies that the difference in the drain current between the two gate corners exactly corresponds to what escapes from the gate, i.e., $I_d(x_{G2}) - I_d(x_{G1}) = I_g(y_G)$. This property was verified for all the results shown in Section IV.

## III. Device Description

The structure of the 32-nm SOI transistor simulated in this paper was originally designed in the framework of the European project PULLNANO [26]. It is schematized in Fig. 1. The silicon body has six degenerate conduction band valleys with a longitudinal and a transverse effective mass of $m_l^* = 0.92\ m_0$ and $m_t^* = 0.19\ m_0$, respectively; an affinity of $\chi_{\rm Si} = 4.05$ eV; and a relative dielectric constant of $\epsilon_{\rm Si} = 11.9$. The source and drain have a length $L_s = L_d$ of 30 nm, are elevated 10 nm above the channel, and have a thickness $t_s = t_d$ of 17 nm. They are doped with a donor concentration of $N_D = 5.8\text{e}19$ cm$^{-3}$. A process simulation [27] was carried out to find the exact doping profile. Incomplete ionization is not considered in the present simulations. The source is grounded ($V_s = 0$ V), whereas the drain is connected to a voltage source $V_d$.

The source and drain are separated from the gate contact by two 10-nm-long $SiO_2$ spacers. The total length $L_g$ of the gate electrode is 32 nm. It controls a 7-nm-thick p-doped channel ($N_A = 1.2\text{e}15$ cm$^{-3}$). The dielectric that isolates the channel from the gate contact is either pure $SiO_2$ or a $SiO_2 - HfO_2$ stack. It is well understood that the silicon oxide may not be $SiO_2$ but rather $SiO_x$. However, this goes beyond the scope of this paper. In the first configuration, a 1.2-nm-thick $SiO_2$ layer is selected since it is consistent with the ITRS specification for the 32-nm node. The gate stack is composed of a 0.8-nm $SiO_2$ layer on top of the Si channel and a 2.2-nm $HfO_2$ layer embedded between the silicon oxide and the gate contact. The EOT of the stack dielectric is the same as that of pure $SiO_2$. The band structure of the dielectric materials is assumed to be isotropic with an effective mass of $m^*_{\rm SiO_2} = 0.5\ m_0$ for $SiO_2$ and $m^*_{\rm HfO_2} = 0.08\ m_0$ for $HfO_2$ [28]. The dielectric constant of $SiO_2$ is $\epsilon_{\rm SiO_2} = 3.9$, and the band offset to the silicon conduction band is $\Delta_{E_C} = 3.07$ eV (corresponding to an affinity of $\chi_{\rm SiO_2} = 0.98$ eV). There is some scattering in the literature regarding the material parameters for $HfO_2$. Here, a dielectric constant of $\epsilon_{\rm HfO_2} = 23$ and a band offset of $\Delta_{E_C} = 1.95$ eV (corresponding to an affinity of $\chi_{\rm HfO_2} = 2.1$ eV) are assumed.

Finally, a metallic gate made out of TiN is attached to the dielectric layer(s). The metal contact is characterized by its Fermi level $E_{\rm Fm}$, its work function $\phi_m = 4.6$ eV, its electron effective mass ($m^* = m_0$), and its conduction band edge $E_{\rm CB}$. Normally, the conduction band edge of a metal lies many electronvolts below its Fermi level. This wide energy range cannot be resolved in the 2-D Schrödinger–Poisson solver since it would require too much CPU time. Consequently, a virtual conduction band edge $E_{\rm CB}$ situated 2 eV below $E_{\rm Fm}$ is assumed, so that all the significant gate states can be injected into the device [29]. Voltage $V_g$ is applied to the gate contact.

## IV. Result

Fig. 2 shows the gate current characteristics $I_g$–$V_{\rm gs}$ at $V_{\rm ds} = 0$ V for the SOI transistor depicted in Fig. 1. The solid lines with symbols are calculated with the 2-D Schrödinger–Poisson (labeled SP) solver, whereas the dashed lines come from the drift-diffusion simulator (labeled DD) with 1-D gate tunneling. Results are presented for the two gate configurations described in Section III: a single 1.2-nm $SiO_2$ layer (black curves) and a $SiO_2 - HfO_2$ stack with an EOT (gray curves). Apart from the substantial reduction of the gate current obtained with the high-$\kappa$ gate stack, the differences between the SP and DD solvers deserve a special treatment. For the single $SiO_2$ dielectric, the full quantum mechanical simulator exhibits a gate current that is 10× larger than that of the 1-D approach at $V_{\rm gs} = 0.1$ V and 23× larger that that of the gate stack. At high gate voltages, the $I_g$–$V_{\rm gs}$ characteristics calculated with both SP and DD tend to the same value.

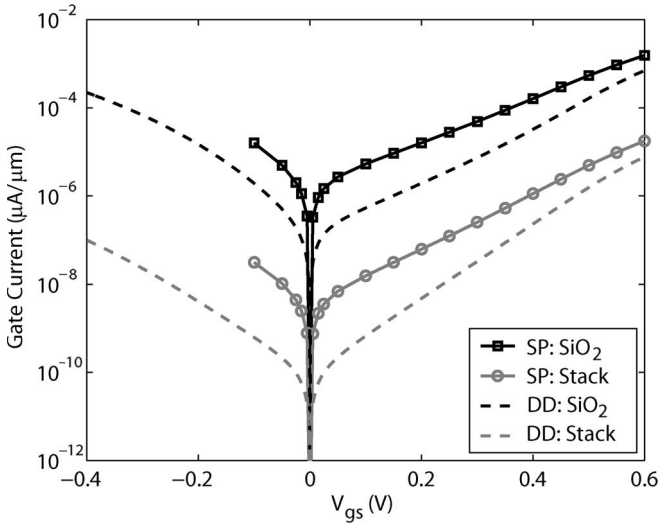


Fig. 2. Gate current characteristics $I_g - V_{\rm gs}$ at $V_{\rm ds} = 0$ V. The solid lines with symbols are calculated with the 2-D Schrödinger–Poisson solver labeled SP, and the dashed lines are calculated with the drift-diffusion solver and 1-D tunneling (DD). The black lines refer to the pure 1.2-nm-thick $SiO_2$ dielectric layer, and the gray lines refer to the $SiO_2 - HfO_2$ stack with EOT.

To physically explain the observed discrepancy at low gate bias, the electron flow issued from the 2-D Schrödinger–Poisson solver is drawn in Fig. 3(a) for the left extremity of the gate contact (around $x = x_{G1}$ and $y = y_G$). The pure $SiO_2$ dielectric configuration is chosen for that purpose. The direction and length of the gray arrows are directly proportional to the current vector $\mathbf{J}(x_i, y_j)$ given in (10). Two important features characterize the current behavior: 1) The electrons do not follow straight lines as imposed by the 1-D model but takes curved trajectories starting before ($x < 40$ nm) and below ($y < 7$ nm) the gate corner (trajectory labeled (a) in Fig. 3). 2) Some electrons do not bypass the $SiO_2$ spacer but tunnel through it to gate contact (b) or first reach the dielectric layer and then gate (c). This occurs in the region between $x < 40$ nm and $y > 7$ nm. As long as an electron takes a straight path, it can be modeled by the 1-D approach. This is the case for trajectory (b). Obviously, the behavior of path (c) cannot be captured by the 1-D model since it would require a straight line connecting the Si body to the dielectric layer and a second one between the dielectric and the gate.

To check whether tunneling paths (b) and (c) through the $SiO_2$ spacers are responsible for the differences between the SP and DD results, they are artificially suppressed. To do so, the $SiO_2$ spacers are replaced by a fictitious material with infinite band gap. Hence, the electrons are no more able to penetrate

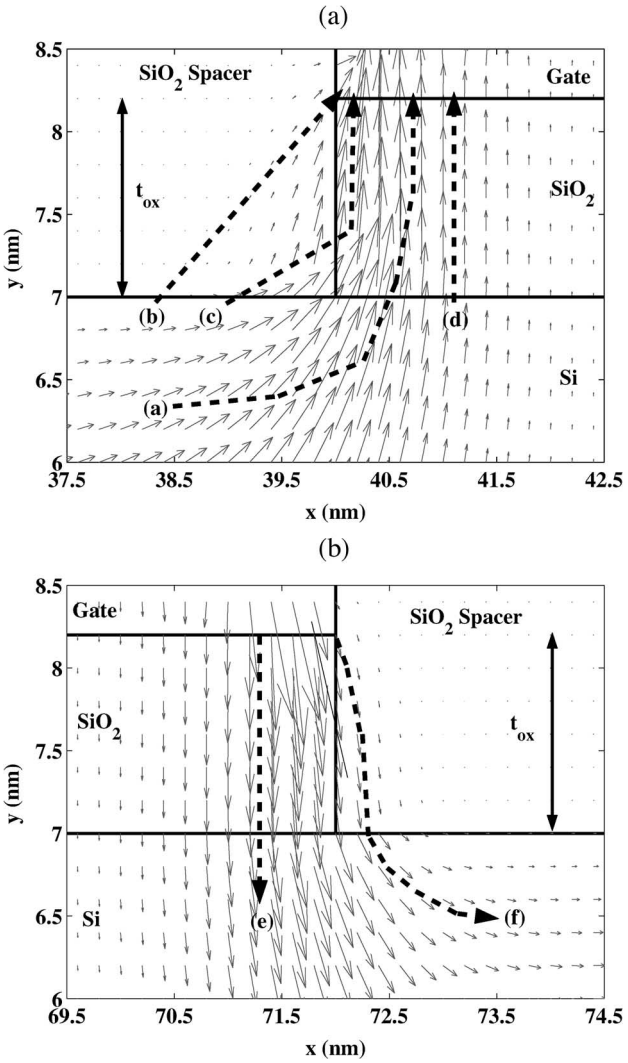


Fig. 3. Electron flow in the SOI FET in Fig. 1 with a 1.2-nm-thick $SiO_2$ oxide layer. (a) Out-tunneling of electrons at the left boundary of the gate contact, i.e., around $x_{G1}$ for $V_{\rm gs} = 0.1$ V and $V_{\rm ds} = 0$ V. (b) In-tunneling of electrons at the drain side (around $x_{G2}$) for $V_{\rm gs} = 0$ V and $V_{\rm ds} = 1$ V. Six types of trajectories labeled (a)–(f) are highlighted.

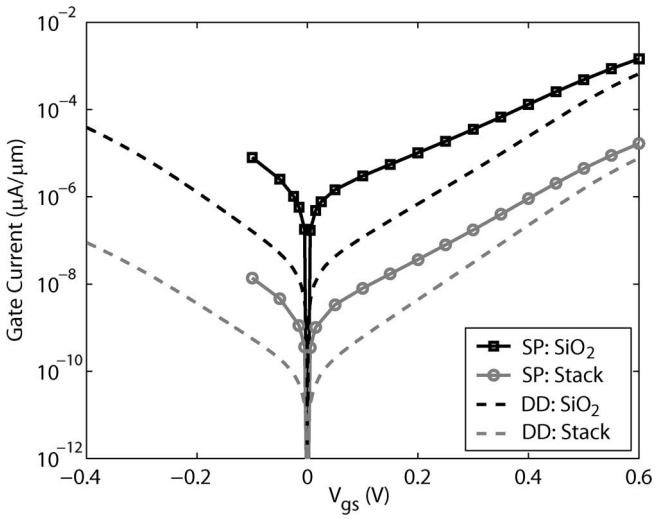


Fig. 4. Same as Fig. 2, except that the $SiO_2$ spacers are replaced by a fictitious material with an infinite band gap preventing the electrons from penetrating the spacers. Hence, the tunneling paths through the spacers vanish.

into the spacers and are forced to avoid these regions. The resulting $I_g$–$V_{\rm gs}$ at $V_{\rm ds} = 0$ V are plotted in Fig. 4. The gate currents coming from the two simulators still do not coincide, but they are smaller than those obtained with the $SiO_2$ spacers. Therefore, tunneling through the spacers does not differentiate the two simulation models. However, this leakage mechanism is important since about half of the gate current could be suppressed if the spacers were infinite potential barriers for the electrons, instead of $SiO_2$ ($I_g = 8$ pA/$\mu$m at $V_{\rm gs} = -0.1$ V, instead of $I_g = 16$ pA/$\mu$m, for the SP simulation in the case of pure $SiO_2$ dielectric). On the other hand, if the band gap offset of the spacers is reduced to $\Delta E_C = 2.15$ eV, as for nitride, the gate current increases by a factor of 1.7 as compared to the $SiO_2$ case ($I_g = 25$ pA/$\mu$m). Hence, the choice of the spacer material is crucial in order to control the level of the gate current.

The discrepancy between the SP and DD models does not originate from the trajectories labeled (b) and (c) in Fig. 3 but is caused by electrons moving on trajectories like path (a). In fact, the tunneling process starts not only when electrons enter the $SiO_2$ dielectric but also when carriers residing just before or after the gate contact penetrate into the Si potential barrier, where they change their direction of propagation by 90° before tunneling through the dielectric layer. The gate current paths are represented by curved lines, which describe the diffraction of the electron wave. Despite the fact that this tunneling path has lower probability than a straight path to the gate, like trajectory (d) in Fig. 3, it induces more leakage current since more carriers are available due to the higher doping level at the starting point. A 2-D quantum transport simulator fully accounts for such effects, whereas 1-D wave functions along straight lines are unsuited to obtain such trajectories. A treatment based on multiple line segments is also conceivable but would require an immense implementation effort. However, as the gate voltage increases, the source-to-drain potential barrier disappears, and (a)-like trajectories do not occur any more. The tunneling paths start in front of the dielectric and follow a straight line, so that the results from both simulation approaches converge toward the same gate current value, as shown in Figs. 2 and 4 at $V_{\rm gs} = 0.6$ V.

The total thickness of the $SiO_2 - HfO_2$ gate stack is 3 nm, so that the probability of finding (b)-like trajectories almost disappears. In effect, tunneling through the $SiO_2$ spacers over such a long distance is a very rare event. Consequently, the gate current calculated with the 1-D approach only counts contributions from (d)-like paths, if the dielectric is a gate stack. The enhancement due to (b)-like trajectories is only 12% at the most. With a single $SiO_2$ dielectric, however, (b)-like trajectories not only give significant contributions but also become the dominant paths at low $V_{\rm gs}$. One finds the maximum effect at $V_{\rm gs} \approx -0.15$ V, where $(b+d)/d \approx 12.5$. At zero gate voltage, this ratio is still $\approx 8$. This clearly demonstrates that simpler 1-D simulation models, which neglect path (b), strongly underestimate the gate current through single $SiO_2$ layers at small $V_{\rm gs}$. In the 2-D quantum mechanical model, the spacer effect can be defined by $(a+b+c+d)/(a+d)$,

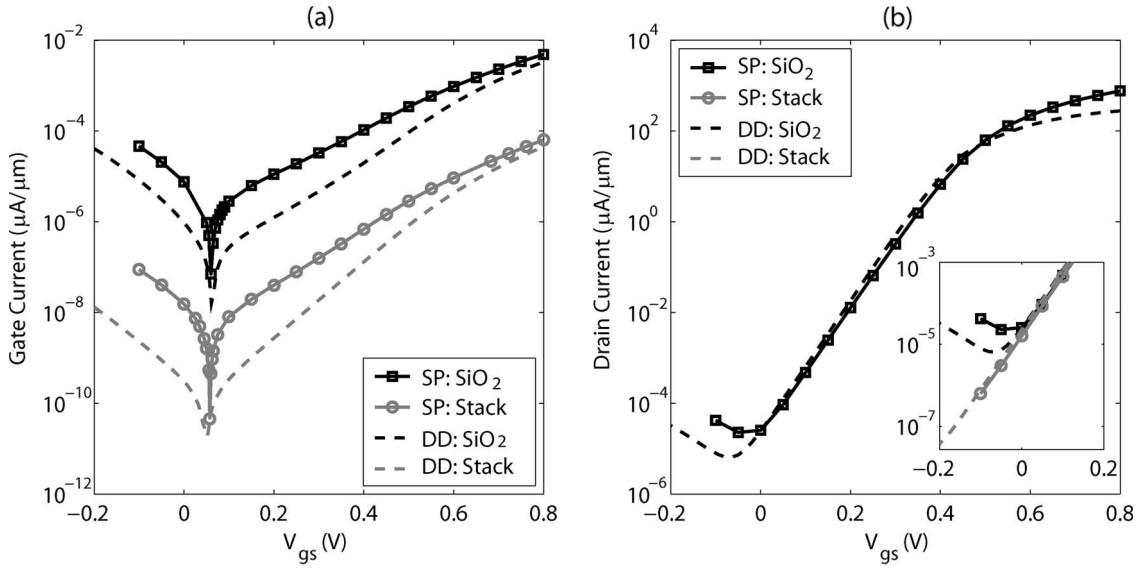


Fig. 5. $I-V$ characteristics of the SOI transistor with $SiO_2$ spacers at $V_{\rm ds} = 0.1$ V. (a) $I_g-V_{\rm gs}$ from the 2-D Schrödinger–Poisson (curves labeled SP) and from the drift-diffusion (DD) solver. The black lines are used for the pure $SiO_2$ dielectric configuration, and the gray lines are used for the gate stack. (b) $I_d-V_{\rm gs}$ from the 2-D Schrödinger–Poisson and the drift-diffusion solver. Results for the gate stack are only shown in the inset.

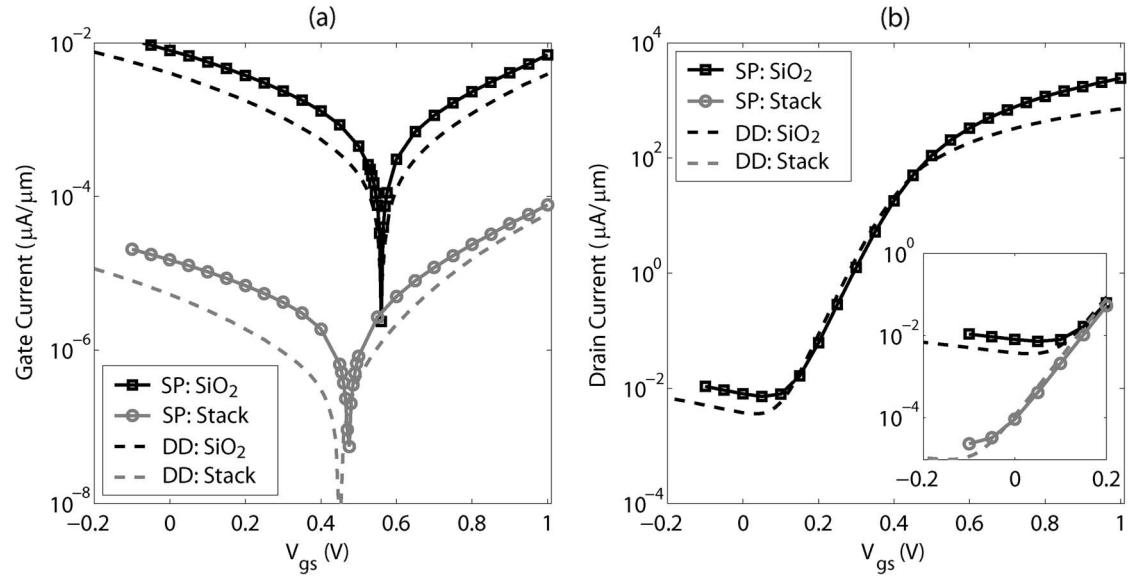


Fig. 6. Same as Fig. 5 but for $V_{\rm ds} = 1$ V.

and one finds a ratio of ≈2 around zero gate voltage for *both* dielectric configurations. Hence, the strong spacer effect in the 1-D model is caused by the absence of paths (a) and (c)! This also explains why the SP/DD ratio $(a + b + c + d/(b + d)$ is larger for the gate stack configuration than for the pure $SiO_2$ dielectric (23× versus 10× at $V_{\rm gs} = 0.1$ V). In general, the importance of 2-D effects increases with the physical thickness of the gate oxide. From the simulated data, it is furthermore possible to infer the relative contributions of (a) and (c). One finds c/a = 1.05 for pure $SiO_2$ and c/a = 0.91 for the gate stack at $V_{\rm gs} = 0.1$ V. Thus, paths that circumvent the spacer have approximately the same probability as paths that partially take course through the spacer.

In Figs. 5 and 6, the gate and drain current characteristics of the same SOI transistor as before are shown for $V_{\rm ds} =$ 0.1 V and $V_{\rm ds} = 1$ V, respectively. The results from the SP (solid lines) and DD (dashed lines) simulators are compared for the two gate dielectric configurations described in Section III and $SiO_2$ spacers. The 2-D Schrödinger–Poisson solver gives higher gate currents as previously discussed, but the shape and behavior of these currents are also reproduced by the simple 1-D model. The drain currents (calculated with the pure $SiO_2$ dielectric) do not only differ in the low-gate-voltage region but also at high gate bias. Since the 1-D model is incorporated into a drift-diffusion solver, scattering is automatically taken into account. This is not the case for the 2-D quantum-mechanical simulator, which computes the ballistic limit of the current.

At high drain voltages as in Fig. 6, the gate current is not equally distributed on both sides of the gate contact but is localized on the drain-side gate corner (i.e., around $x_{G2}$). It is

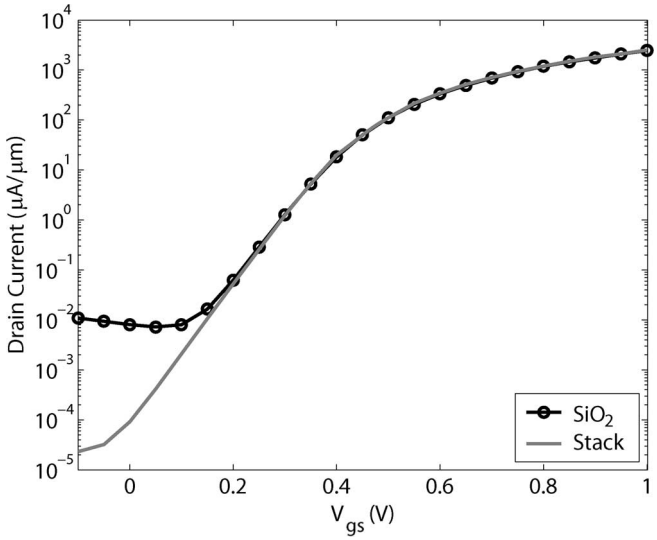


Fig. 7. Transfer characteristics $I_d$–$V_{\rm gs}$ at $V_{\rm ds} = 1$ V calculated with the 2-D Schrödinger–Poisson solver. The black line with circles represents the pure $SiO_2$ dielectric layer, and the gray line represents the gate stack configuration.

caused by in-tunneling of gate electrons in a narrow interval, as illustrated in Fig. 3(b). The straight path labeled (e) has higher tunneling probability than path (f), which can only be modeled by the 2-D Schrödinger–Poisson solver. However, since the electrostatic potential rapidly decreases on the drain side to compensate the applied bias of $V_{\rm ds} = 1$ V, the electrons following (f)-like trajectories have greater velocity than those propagating along (e)-like paths and contribute more to the gate current.

The benefit of high-$\kappa$ gate stacks over pure $SiO_2$ dielectrics is illustrated in Fig. 7. The transfer characteristics $I_d$–$V_{\rm gs}$($V_{\rm ds} = 1$ V) of the SOI FET in Fig. 1 is simulated with the 2-D Schrödinger–Poisson solver. The black line with symbols represents the pure $SiO_2$ gate dielectric, and the gray line represents the $SiO_2 - HfO_2$ gate stack. The OFF-current ($I_d$ at $V_{\rm gs} = 0$ V, $V_{\rm ds} = 1$ V) is reduced by about two orders of magnitude if the gate stack is used ($I_{\rm OFF} = 91$ pA/$\mu$m, instead of 8100 pA/$\mu$m). Since the EOT of both gate configurations are the same (1.2 nm), the electrostatic and ON-current ($I_d$ at $V_{\rm gs} = V_{\rm ds} = 1$ V) properties of the transistors do not vary. Note that the value of the OFF-current strongly depends on the choice of effective mass $m^*_{\rm HfO_2}$ and band offset $\Delta E_C$. Here, a rather small value for $m^*_{\rm HfO_2}$ is chosen (0.08 $m_0$), so that the tunnel probability through the high-$\kappa$ gate stack is facilitated. Comparisons with measured gate currents could clarify whether these theoretical results are too large.

## V. Conclusion

In this paper, the gate currents of SOI transistors with either a single dielectric or a high-$\kappa$ gate stack were investigated using a 1-D approach incorporated into a drift-diffusion simulator and a 2-D and real-space Schrödinger–Poisson solver. The main conclusion is that a 1-D treatment will always underestimate the gate current as it fails to include the effect of electron diffraction at the gate corners. At low gate voltages, this effect yields the dominant contribution to direct tunneling leakage. In the case of out-tunneling, the major part of the difference comes from those trajectories that start outside the gate region at points with higher carrier density. This electron diffraction around the spacers is completely absent in the 1-D approach. Due to this effect, the 2-D model gives ten times more current for a pure $SiO_2$ dielectric, and this difference increases with the physical thickness of the oxide to reach a factor of about 23 for the gate stack. Only at very large gate voltages both methods converge.

The actual OFF-state leakage is determined by in-tunneling electrons in a narrow interval ($< 2$ nm) at the drain-side gate corner. The 1-D approach underestimates the OFF-current for both kinds of gate dielectric. Again, curved trajectories are advantageous leakage paths as they end at points of lower potential energy and higher carrier velocity. These points exist because of the rapid voltage drop in the pinchoff region. For the 32-nm FET studied here, this more than doubles the gate current through the pure $SiO_2$ dielectric. As in the case of out-tunneling, the 2-D effects become stronger with increasing physical thickness of the dielectric. Thus, for the optimum design of high-$\kappa$ stack configurations, a 2-D and full quantum-mechanical treatment of gate leakage should be envisaged.

## Appendix
## Symmetric Matrices in the Finite-Difference Scheme

When the Schrödinger equation is discretized on a nonuniform finite-difference grid, the resulting Hamiltonian matrix $\mathbf{H}$ is not symmetric. However, working with symmetric matrices is highly recommended since it allows a reduction in the computational burden during the factorization process. The elements $H_{i_1 i_2 i_1 j_2}$ of $\mathbf{H}$ are proportional to [17]

$$H_{\rm ii\pm 1jj} \propto \frac{1}{x_{i+1} - x_{i-1}} \tag{13}$$

$$H_{\rm iijj\pm 1} \propto \frac{1}{y_{j+1} - y_{j-1}}. \tag{14}$$

Since $H_{\rm ii\pm 1jj} \neq H_{i\pm 1\rm ijj}$ and $H_{\rm iijj\pm 1} \neq H_{\rm iij\pm 1 j}$, Hamiltonian matrix $\mathbf{H}$ is not symmetric. One introduces now diagonal matrix $\mathbf{R}$ and its inverse $\mathbf{invR}$ with the entries

$$R^{\rm ij}_{\rm ii} = \frac{1}{\sqrt{x_{i+1} - x_{i-1}}} \cdot \frac{1}{\sqrt{y_{j+1} - y_{j-1}}} \tag{15}$$

$$invR^{\rm ij}_{\rm ii} = \sqrt{x_{i+1} - x_{i-1}} \cdot \sqrt{y_{j+1} - y_{j-1}}. \tag{16}$$

Index $i$ refers to the $i$th vertical grid line, which is described by a block whose $j$th diagonal element is $R^{\rm ij}_{\rm ii}$ or $invR^{\rm ij}_{\rm ii}$. Matrix $\mathbf{H}$ is symmetrized by applying the following basis transformation to the wave function $\phi$ in (8):

$$\phi = \mathbf{R} \cdot \tilde{\phi}_{\mathbf{SYM}} \tag{17}$$

$$\tilde{\mathbf{H}} = \mathbf{invR} \cdot \mathbf{H} \cdot \mathbf{R}. \tag{18}$$

After this transformation, the nondiagonal elements $H_{\rm nm}$ of $\mathbf{H}$ are replaced by the geometric mean value $\sqrt{H_{\rm nm} \cdot H_{\rm mn}}$, and $\tilde{\mathbf{H}}$ becomes symmetric as well as $\tilde{\mathbf{A}}$, leading to a faster solution

of (8). The same transformation can be performed on $\mathbf{H}_{\mathrm{nn}}$ and $\phi_n$ in (7). The resulting matrix $\mathbf{M}$ is still not symmetric, but since $\mathbf{T_n}$ is diagonal, a second basis transformation can be applied, so that

$$\phi_n = \sqrt{\mathbf{T_n}} \cdot \tilde{\phi}_{\mathbf{n},\mathbf{SYM}}. \tag{19}$$

This identity symmetrizes $\mathbf{M}$ and reduces (7) to the solution of a symmetric eigenvalue problem.

## ACKNOWLEDGMENT

The authors would like to thank Dr. A. Wettstein (Synopsys LLC., Switzerland) for the many valuable discussions.

## REFERENCES

[1] [Online]. Available: http://www.itrs.net/reports.html
[2] L. Risch, "Pushing CMOS beyond the roadmap," *Solid State Electron.*, vol. 50, no. 4, pp. 527–535, Apr. 2006.
[3] B. Doris *et al.*, "Device design considerations for ultra-thin SOI MOSFETs," in *IEDM Tech. Dig.*, 2003, pp. 27.3.1–27.3.4.
[4] T. Irisawa, T. Numata, T. Tezuka, N. Sugiyama, and S. Takagi, "Electron transport properties of ultrathin-body and tri-gate SOI nMOSFETs with biaxial and uniaxial strain," in *IEDM Tech. Dig.*, 2006, pp. 1–4.
[5] D. A. Buchanan, "Scaling the gate dielectric: Materials, integration, and reliability," *IBM J. Res. Develop.*, vol. 43, no. 3, pp. 245–264, May 1999.
[6] G. D. Wilk, R. M. Wallace, and J. M. Anthony, "High-$\kappa$ gate dielectrics: Current status and materials properties consideration," *J. Appl. Phys.*, vol. 89, no. 10, p. 5243, 2001.
[7] J. C. Lee *et al.*, "High-*k* dielectrics and MOSFET characteristics," in *IEDM Tech. Dig.*, 2003, pp. 4.4.1–4.4.4.
[8] N. G. Tarr, D. L. Pulfrey, and D. S. Camporese, "An analytic model for the MIS tunnel junction," *IEEE Trans. Electron Devices*, vol. ED-30, no. 12, pp. 1760–1770, Dec. 1983.
[9] J. Bardeen, "Tunneling from a many-particle point of view," *Phys. Rev. Lett.*, vol. 6, no. 2, pp. 57–62, 1961.
[10] A. Schenk and G. Heiser, "Modeling and simulation of tunneling through ultra-thin gate dielectrics," *J. Appl. Phys.*, vol. 81, no. 12, pp. 7900–7908, Jun. 1997.
[11] A. Wettstein, A. Schenk, and W. Fichtner, "Simulation of direct tunneling through stacked gate dielectrics by a fully integrated 1D-Schrödinger–Poisson solver," *IEICE Trans. Electron. Jpn.*, vol. E83-C, pp. 1189–1193, Aug. 2000.
[12] *Sentaurus Device User Guide*, Synopsys Inc., Mountain View, CA, 2007. version Z-2007.03.
[13] A. C. Marsh and J. C. Inkson, "Scattering matrix theory of transport in heterostructures," *Semicond. Sci. Technol.*, vol. 1, no. 4, pp. 285–290, Oct. 1986.
[14] Z. Ren, R. Venugopal, S. Goasguen, S. Datta, and M. S. Lundstrom, "NanoMOS 2.5: A two-dimensional simulator for quantum transport in double-gate MOSFETs," *IEEE Trans. Electron Devices*, vol. 50, no. 9, pp. 1914–1925, Sep. 2003.
[15] R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom, and D. Jovanovic, "Simulating quantum transport in nanoscale transistors: Real versus mode-space approaches," *J. Appl. Phys.*, vol. 92, no. 7, p. 3730, Oct. 2002.
[16] E. Polizzi and N. Ben Abdallah, "Subband decomposition approach for the simulation of quantum electron transport in nanostructures," *J. Comput. Phys.*, vol. 202, no. 1, pp. 150–180, Jan. 2005.
[17] M. Luisier, A. Schenk, and W. Fichtner, "Quantum transport in two- and three-dimensional nanoscale transistors: Coupled mode effects in the nonequilibrium Green's function formalism," *J. Appl. Phys.*, vol. 100, no. 4, pp. 043 713-1–043 713-12, Aug. 2006.
[18] W. R. Frensley, "Boundary conditions for open quantum systems driven far from equilibrium," *Rev. Mod. Phys.*, vol. 62, no. 3, pp. 745–792, Jul. 1990.
[19] M. Luisier, G. Klimeck, A. Schenk, and W. Fichtner, "Atomistic simulation of nanowires in the $sp^3d^5s^*$ tight-binding formalism: From boundary conditions to strain calculations," *Phys. Rev. B, Condens. Matter*, vol. 74, no. 20, p. 205 323, 2006.
[20] A. Svizhenko, M. P. Anantram, T. R. Govindan, R. Biegel, and R. Venugopal, "Two-dimensional quantum mechanical modeling of nanotransistors," *J. App. Phys.*, vol. 91, no. 4, p. 2343, Feb. 2002.
[21] T. A. Davis, "A column pre-ordering strategy for the unsymmetric-pattern multifrontal method," *ACM Trans. Math. Softw.*, vol. 30, no. 2, pp. 165–195, Jun. 2004.
[22] O. Schenk and K. Gärtner, "Solving unsymmetric sparse systems of linear equations with PARDISO," *J. Future Generation Comput. Syst.*, vol. 20, no. 3, pp. 475–487, Apr. 2004.
[23] P. R. Amestoy, I. S. Duff, and J.-Y. L'Excellent, "Multifrontal parallel distributed symmetric and unsymmetric solvers," *Comput. Methods Appl. Mech. Eng.*, vol. 184, no. 2, pp. 501–520, Apr. 2000.
[24] V. Nam Do and P. Dollfus, "Oscillation of gate leakage current in double-gate metal-oxide-semiconductor field-effect transistors," *J. Appl. Phys.*, vol. 101, no. 7, p. 073 709, Apr. 2007.
[25] D. Mamaluy, M. Sabathil, and P. Vogl, "Efficient method for the calculation of ballistic quantum transport," *J. Appl. Phys.*, vol. 93, no. 8, pp. 4628–4633, Apr. 2003.
[26] [Online]. Available: http://www.pullnano.eu
[27] *Sentaurus Process User Guide*, Synopsys Inc., Mountain View, CA, 2007. version Z-2007.03.
[28] A. Campera, G. Iannaccone, and F. Crupi, "Modeling of tunneling currents in Hf-based gate stacks as a function of temperature and extraction of material parameters," *IEEE Trans. Electron Devices*, vol. 54, no. 1, pp. 83–89, Jan. 2007.
[29] M. Luisier, A. Schenk, and W. Fichtner, "Three-dimensional modeling of gate leakage in Si nanowire transistors," in *IEDM Tech. Dig.*, 2007, pp. 733–736.

**Mathieu Luisier** received the Dipl.-Ing. degree (with honors) in electrical engineering and the Ph.D. degree from ETH Zürich, Zürich, Switzerland, in 2003 and 2007, respectively. His Ph.D. dissertation was about atomistic and full-band simulation of nanowire transitors.

In 2008, he joined the Network for Computational Nanotechnology, Purdue University, West Lafayette, IN, as a Research Assistant Professor. His current research interests are quantum transport in nanoscale devices, parallel numerical algorithms, and development of next-generation technology computer-aided design tools.

Mr. Luisier was the recipient of the ETH silver medal for his diploma thesis in 2003 and for his Ph.D. dissertation in 2007.

**Andreas Schenk** was born in Berlin, Germany, in 1957. He received the Dipl. degree in physics and the Ph.D. degree from Humboldt University Berlin (HUB), Berlin, in 1981 and 1987, respectively. In 1987, he became a Research Assistant with the Department of Semiconductor Theory, HUB. In 1988, he joined the R&D Division of WF Berlin. From 1987 to 1991, he was working on various aspects of the physics and simulation of optoelectronic devices. In 1991, he joined the Integrated Systems Laboratory (ISL), ETH Zürich, Zürich, Switzerland, as a Senior Research/Teaching Assistant, where he qualified to give lectures at university in 1997 for "Physics and Modeling of Microelectronic Devices." In 2004, he became a Titular Professor at the ISL. He is also heading the Device Physics Group at ISL. He authored and coauthored two books and 125 papers. His main activities are focused on physics-based modeling of micro- and nanoelectronic devices: quantum effects in ultrashort transistors, physics and simulation of single-electron devices, many-body effects, contacts, heterojunctions, degradation effects, transport processes like resonant and sequential tunneling, generation–recombination, and noise.