School of Natural Sciences and Mathematics

Theoretical Simulation of Negative Differential Transconductance in Lateral Quantum Well nMOS Devices

UT Dallas Author(s):

P. B. Vyas
Mark Lee
William G. Vandenberghe
Massimo V. Fischetti

Citation:

Vyas, P. B., C. Naquin, H. Edwards, M. Lee, et al. 2017. "Theoretical simulation of negative differential transconductance in lateral quantum well nMOS devices." Journal of Applied Physics 121(4), doi:10.1063/1.4974469

Theoretical simulation of negative differential transconductance in lateral quantum well nMOS devices

P. B. Vyas,[1,a)] C. Naquin,[2] H. Edwards,[2] M. Lee,[3] W. G. Vandenberghe,[4]
and M. V. Fischetti[4,b)]

[1]*Department of Electrical Engineering, The University of Texas at Dallas, 800 W. Campbell Rd., Richardson, Texas 75080, USA*
[2]*Texas Instruments Inc., Richardson, Texas 75243, USA*
[3]*Department of Physics, The University of Texas at Dallas, 800 W. Campbell Rd., Richardson, Texas 75080, USA*
[4]*Department of Materials Science and Engineering, The University of Texas at Dallas, 800 W. Campbell Rd., Richardson, Texas 75080, USA*

(Received 21 October 2016; accepted 8 January 2017; published online 23 January 2017)

We present a theoretical study of the negative differential transconductance (NDT) recently observed in the lateral-quantum-well Si n-channel field-effect transistors [J. Appl. Phys. **118**, 124505 (2015)]. In these devices, p^+ doping extensions are introduced at the source-channel and drain-channel junctions, thus creating two potential barriers that define the quantum well across whose quasi-bound states resonant/sequential tunneling may occur. Our study, based on the quantum transmitting boundary method, predicts the presence of a sharp NDT in devices with a nominal gate length of 10-to-20 nm at low temperatures (~10 K). At higher temperatures, the NDT weakens and disappears altogether as a result of increasing thermionic emission over the p^+ potential barriers. In larger devices (with a gate length of 30 nm or longer), the NDT cannot be observed because of the low transmission probability and small energetic spacing (smaller than k_BT) of the quasi-bound states in the quantum well. We speculate that the inability of the model to predict the NDT observed in 40 nm gate-length devices may be due to an insufficiently accurate knowledge of the actual doping profiles. On the other hand, our study shows that NDT suitable for novel logic applications may be obtained at room temperature in devices of the current or near-future generation (sub-10 nm node), provided an optimal design can be found that minimizes the thermionic emission (requiring high p^+ potential-barriers) and punch-through (that meets the opposite requirement of potential-barriers low enough to favor the tunneling current). *Published by AIP Publishing.*
[http://dx.doi.org/10.1063/1.4974469]

I. INTRODUCTION

One little-explored route towards achieving the very aggressive high frequency and high sensitivity goals outlined in the International Technology Roadmap for Semiconductors (ITRS)[1] for the silicon semiconductor industry is to move beyond the conventional semi-classical device physics. This can be done by integrating explicitly the quantum mechanical transport into industrial Si complementary metal-oxide-semiconductor (CMOS) technology. This route would enable the Si CMOS to emulate the path of III–V devices that incorporate transport through electronic states that are localized and whose energy is discretized by quantum confinement. Esaki diodes[2,3] and resonant tunneling diodes (RTDs)[4–6] undoubtedly constitute the best-known examples of devices based on III-V compound semiconductors that have concretely realized this idea. These instances have yielded new capabilities in very high speed and very low-noise applications,[7–9] a level of performance that is beyond the reach of conventional semi-classical devices. However, Si CMOS quantum devices must be fabricated within the standardized progression of industrial process nodes (now approaching the 7 nm node at the high end of the performance scale, but still commonly relying on the processing technology of the 28 or 22 nm nodes) to guarantee economically scalable production. This restriction rules out the two primary methods of fabricating quantum structures in III–V devices, hetero-epitaxial layer growth, and electron-beam lithography, therefore rendering more difficult the possibility of entering the quantum-transport regime in Si CMOS devices. Recently, however, tantalizing hints of quantum transport have been observed in Si CMOS devices using lateral quantum wells (QWs) defined by ion implantation.[10,11] Here, we present a theoretical study of quantum electron transport in these devices, emphasizing the limitations of both their practical realization as well as of our understanding of the basic physical processes involved. At the same time, we also show that this type of quantum transport may be achieved in practice and indicate possible future promising paths that the technology may follow.

The paper is organized as follows: In Sec. II, we describe the experimental observations. In Sec. III, we present the physical and numerical methods we have used to study the devices. In Sec. IV, we present the results of our study for devices of gate length ranging from 40 to 10 nm and, finally, we draw our conclusions in Sec. V. Our main

[a)]Electronic mail: pbv130130@utdallas.edu
[b)]Electronic mail: max.fischetti@utdallas.edu

conclusions show that negative differential transconductance (NDT) may result only if several design criteria are met: (1) The p$^+$ potential barriers that define the lateral QW must be spatially close enough to result in an energetic spacing of the quasi-bound states larger than the thermal energy, k_BT. This requires the channel lengths of the order of 20 nm or shorter at cryogenic temperatures and even shorter for room-temperature operation. (2) The potential barriers must be high enough to prevent thermionic emission at high temperatures. (3) These barriers must be simultaneously small enough to favor the tunneling current across the quasi-bound states in the QW, rather than the punch-through current "around" the barriers.

II. DEVICE DESCRIPTION AND EXPERIMENTAL OBSERVATIONS

An explicit experimental demonstration of quantum transport in Si n-channel MOS (nMOS) transistors fabricated using an industrially standard 45 nm-node process technology has been recently reported.[10,11] These nMOS devices have a lateral quantum well (QW) built into the surface channel. This is obtained by reversing the ion-implantation dopant polarity of the shallow source/drain (S/D) extensions (pSDE) from the standard n-type (for an nMOS transistor) to p-type, as sketched in Fig. 1. The p-type extensions create an energy barrier for electrons between the n^{++} S/D and the surface channel beneath the gate. A two-dimensional (2D) electron quantum well (QW) is formed when the gate voltage V_{GS} is large enough to invert the channel between the p-type extensions. The depth of the QW can be controlled by the source/gate bias V_{GS}. Explicit evidence of quantum transport in these QW nMOS devices was shown in the form of a negative differential transconductance (NDT). This occurs when the drain-source current (I_{DS}) behaves non-monotonically, so that $g_m = \partial I_{DS}/\partial V_{GS} < 0$. This has been observed only in QW nMOS devices, but not in standard nMOS devices fabricated on the same chip as experimental controls, thus showing that indeed the quantized electronic states in the lateral QW play a fundamental role in controlling the electronic transport. Whereas NDT is an expected signature of direct or sequential tunneling through discrete QW bound states,[4,12] several quantitative aspects of the reported NDT indicate that the detailed physical mechanism causing the NDT is more complicated than a straightforward QW tunneling phenomenon. Among these apparently anomalous features is the observation of only a small number of NDT peaks (no more than 3) observed in any given device with a width and V_{GS}-separation much larger than expected, given the nominal QW lengths. Poorly understood is also the need to apply a positive body current or voltage bias that introduces the bipolar-like operational characteristics in order to observe the NDT.[11]

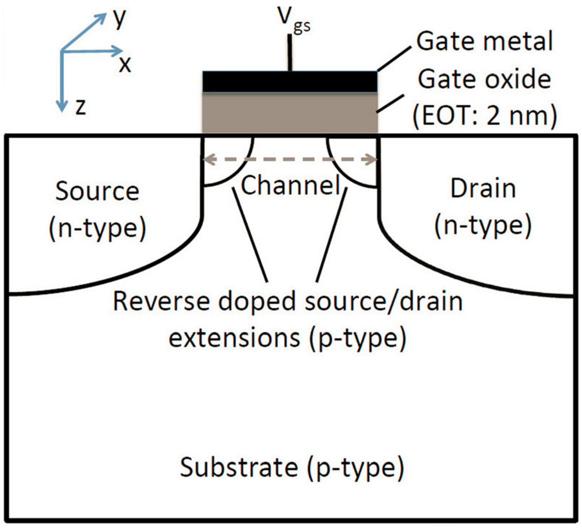

FIG. 1. Schematic cross-section of the lateral QW nMOSFET.

We present here numerical simulations with explicitly quantum mechanical charge transport that we have performed in order to elucidate more clearly and quantitatively the device physics underlying the NDT phenomenon in such QW nMOS transistors. This also allows us to establish a route towards optimizing the quantum behavior of the devices.

III. THEORETICAL FORMULATION

The phenomenon that we wish to see and which is the motivation behind this whole project, is the presence of NDT in the current-voltage characteristics as a consequence of resonant or sequential tunneling in the channel of the QW nMOS device. The objective of the two-dimensional confinement created in the channel is to produce bound electronic states in that region. Electrons injected into the channel at energies equal to these bound states will undergo resonant tunneling that occurs with a very high transmission coefficient. The magnitude of the transmission coefficient depends on the relative height of the potential barriers. The gate bias acts as a control for the energy of the bound states. At certain gate biases, the energy of these bound states will coincide with the Fermi energy of the electrons in the source region. Therefore, at the appropriate gate bias, the source-to-drain current, I_{DS}, should exhibit sharp peaks, as the electrons that contribute most to the current have a high resonant transmission coefficient.

The simulation of the device characteristics follows an extensive work that has been performed on the ballistic simulation of RTDs (see, for example, Ref. 13) although RTD simulations including inelastic scattering, ignored here, have been performed using the non-equilibrium Green's function (NEGF) method with tight-binding models[14,15] or using the Wigner function formalism[16,17] and, more generally, on the use of the quantum transmitting boundary method (QTBM)[18]—employing the effective-mass approximation—also in two spatial dimensions.[19,20]

Before delving into technical details, it is convenient to summarize the main features of the calculation method employed. We proceed in two steps. First, we solve the two-dimensional (2D) Schrödinger equation under *closed* boundary condition self consistently with the Poisson equation[24,25] with a source-to-drain bias $V_{DS} = 0$ and for several values of the gate-to-source bias, V_{GS}. This gives quantitative information about the energetic positions of the confined states for various gate-bias conditions and provides the equilibrium electrostatic potential. As a second step, electron transport via tunneling through the confined states—the effect that gives rise to the NDT—is studied by solving the Schrödinger equation with *open* boundary conditions.[18,20,21] This allows us to calculate the current-voltage characteristics. We perform this calculation step by using the self-consistent potential obtained from the previous step. Indeed, self-consistent effects between the Schrödinger equation with open boundary conditions and the Poisson equation are not expected to alter the potential profile significantly at the low source-to-drain bias, V_{DS}, of interest (of the order of a few tens of mV). This choice of low V_{DS}, also employed in the experimental observations,[10,11] ensures that electrons remain in near-equilibrium conditions, so that the use of the self-consistent solution obtained for the closed system constitutes an excellent approximation also for the open system. Thus, having obtained the electrostatic potential, we calculate the current by applying a small shift V_{DS} of a few mV between the Fermi levels in the source and drain.

The model we employ approximates the conduction bands of silicon with six parabolic ellipsoidal valleys and the only ballistic transport of electrons is considered. We account for finite-temperature operation and the fact that the fabricated devices have the channel being oriented along the [110] direction. The device geometry attempts to reproduce as faithfully as possible the geometry of the devices used in Refs. 10 and 11. In particular, the doping profile we have used, shown in Fig. 2, has been obtained using the TCAD process simulation tool Sentaurus® by Synopsys, using the known fabrication process data such as the mask layout, implant types, energies, doses and angles, and post-implant annealing conditions.

Scaling the device from its nominal size (printed gate length of about 40 nm) to the smaller dimensions that we have considered (30, 20, and 10 nm gate length), has been done following the conventional scaling laws[22,23] as strictly as possible. Obvious exceptions had to be made regarding the scaling of the doping concentrations, since their high values in the 40 nm device cannot be realistically increased with scaling demands. Moreover, the SiO$_2$-equivalent thickness of the gate insulator (EOT) has been kept fixed at 2 nm. This is not a crucial parameter for the application of interest here. Crucial, instead, is the scaling of the doping profiles of the p-type substrate, S/D regions, and of the pSDEs: Preventing the short-channel effects (mainly punch-through) in short devices, requires increasing the p-type substrate doping. Since the peak pSDE doping is already quite large in the 40 nm device, the higher scaled substrate doping results in a reduced height of the pSDE potential barriers in shorter devices. This does result in the desired boost of the tunneling current across the QW and in a reduced punch-through current. On the other hand, it also results in the undesired occurrence of a large thermionic current (over the pSDE barriers) that hides the (resonant) tunneling current. The problem caused by this narrow "design window" will be discussed below.

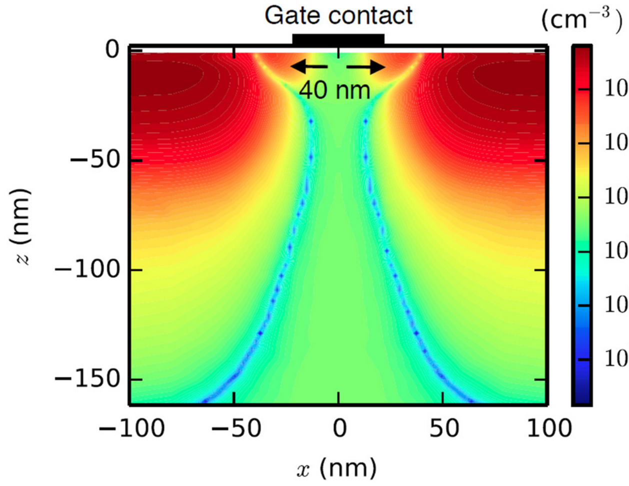

FIG. 2. Magnitude of the net doping profile of the device with a nominal gate length of 40 nm. The white region at the top represents the gate oxide.

We now present details regarding the physical models and their numerical implementation that are most relevant to our study. The simulation region (cross-section of the device) is 200 nm × 160 nm for the 40 nm device. Devices of smaller gate lengths, 10 nm, 20 nm, and 30 nm, are also simulated by shrinking the length and height of the 40 nm device doping profile (both the lateral S/D doping and the S/D-body junction-depth) by the relevant factors (of 0.25, 0.5, and 0.75, respectively).

A. 2D Schrödinger equation with closed boundary conditions

The finite-difference method is used to solve numerically the time-independent single electron "effective mass" Luttinger-Kohn (Schrödinger) equation[26] using the parabolic-band approximation to approximate the anisotropic ("ellipsoidal") electron dispersion close to the six minima of the conduction band. Only electrons in the first conduction band are considered. For the simplest case in which the channel is oriented along the [100] direction, the two-dimensional Schrödinger equation takes the form:[24,27]

$$-\frac{\hbar^2}{2}\left[\frac{1}{m_x}\frac{\partial^2\xi(x,z)}{\partial x^2}+\frac{1}{m_z}\frac{\partial^2\xi(x,z)}{\partial z^2}\right] \\ +V(x,z)\xi(x,z)=E_{xz}\xi(x,z)\,, \tag{1}$$

where the envelope wavefunction is $\psi(x,y,z)=e^{ik_y y}\xi(x,z)$ and m_x, m_y, and m_z are the transport mass (on the plane $(x,\,z)$-plane of the Si/gate-insulator interface and along the channel), the out-of-plane mass (along the y-direction perpendicular to the interface), and the quantization mass (on the plane of the interface, but along the z direction perpendicular to the transport direction). V is the potential distribution in the $(x,\,z)$ plane, E_{xz} is the energy of the two-dimensional wavefunction $\xi(x,z)$ and k_y represents the wavevector in the y direction. These quantities take different values for each of the six ellipsoidal valleys close to the X symmetry-point in the Si Brillouin zone. In order to render our notation more agile, we shall not explicitly introduce a "valley index" to reflect this six-fold degeneracy. However, all calculations presented in the following must be understood as repeated three times, once for each pair of inequivalent valley-orientations.

The electrostatic potential is assumed to be independent of the y coordinate. This amounts to assuming an infinitely wide channel. We denote by $E_{\mathrm{xz}} + \hbar^2 k_y^2/2m_{\mathrm{y}}$ the total energy of an electronic state described by the wavefunction $\psi(x, y, z)$.

In its discretized form, Eq. (1) can be recast as:

$$a_1\left[\frac{\xi(x_{i+1}, z_i)}{\Delta x_i^2} + \frac{\xi(x_{i-1}, z_i)}{\Delta x_i \Delta x_{i-1}}\right] + a_2\left[\frac{\xi(x_i, z_{i+1})}{\Delta z_i^2} + \frac{\xi(x_i, z_{i-1})}{\Delta z_i \Delta z_{i-1}}\right] - \left[\frac{a_1}{\Delta x_i^2} + \frac{a_1}{\Delta x_i \Delta x_{i-1}} + \frac{a_2}{\Delta z_i^2} + \frac{a_2}{\Delta z_i \Delta z_{i-1}} + V(x_i, z_i)\right]\xi(x_i, z_i) = E^i_{\mathrm{xz}}\xi(x_i, z_i)\,, \tag{2}$$

where $a_1 = -\hbar^2/2m_{\mathrm{x}}$, $a_2 = -\hbar^2/2m_{\mathrm{z}}$ and, i represents each of the N discretization points. We employ a two-dimensional mesh given by the tensor product of two one-dimensional meshes consisting of N_{x} and N_{z} points, so that $N = N_{\mathrm{x}} \times N_{\mathrm{z}}$. The $N \times N$ Hamiltonian matrix ($\mathbf{H}$) is created using the terms on the left-hand-side of the discretized equation. We assume the Dirichlet boundary conditions, assuming that the wavefunctions vanish outside the simulated region (the "device"). In particular, the wavefunctions are assumed to vanish on virtual mesh points just inside the "leads." We have also performed the simulation employing both Dirichlet and Neumann boundary conditions in order to verify that no numerical artifacts at the contacts affect the electron density of states, the electron density, and the transmission coefficient. Therefore, Eq. (2) takes the form of an eigenvalue problem of rank N

$$\mathbf{H} \cdot \boldsymbol{\xi}_\mu = E_\mu \boldsymbol{\xi}_\mu, \tag{3}$$

where $\boldsymbol{\xi}_\mu$ and E_μ are the μth eigenfunction and eigenvalue, respectively, and, $\mu = 1, 2, \ldots, N$. As we have already mentioned, these are calculated by solving the eigenvalue problem of Eq. (3) separately for each of the six equivalent valleys that are characterized by different effective masses. For channels oriented along the [110] direction, as in the fabricated devices of Refs. 10 and 11, the effective mass tensor is a full 3×3 matrix, instead of a diagonal matrix. In this case, the Schrödinger equation includes mixed second order derivatives, increasing the complexity of the problem,

$$-\frac{\hbar^2}{2}\left\{\frac{1}{2}\left[\frac{1}{m_{\mathrm{x}}} + \frac{1}{m_{\mathrm{y}}}\right]\left[\frac{\partial^2}{\partial x'^2} + \frac{\partial^2}{\partial y'^2}\right] + \frac{1}{m_{\mathrm{z}}}\frac{\partial^2}{\partial z^2} + \left[\frac{1}{m_{\mathrm{x}}} - \frac{1}{m_{\mathrm{y}}}\right]\frac{\partial^2}{\partial x' \partial y'} + V\left(x', y', z\right)\right\}\psi\left(x', y', z\right) = E_{\mathrm{xz}}\psi\left(x', y', z\right)\,, \tag{4}$$

where, $x' = \frac{1}{\sqrt{2}}\left(x + y\right)$, $y' = \frac{1}{\sqrt{2}}(x - y)$. Through rotation of the coordinate system and some mathematical manipulation, the mixed second derivatives can be removed and the modified Schrödinger equation for the [110] direction takes the form:

$$-\frac{\hbar^2}{2}\left[\frac{1}{m_{\mathrm{c}}}\frac{\partial^2 \xi(x,z)}{\partial x^2} + \frac{1}{m_{\mathrm{z}}}\frac{\partial^2 \xi(x,z)}{\partial z^2}\right] + V(x,z)\xi(x,z) = E_{\mathrm{xz}}\xi(x,z) - \frac{\hbar^2 k_y^2}{2}\left[\frac{1}{m_{\mathrm{c}}} - \frac{m_{\mathrm{c}}}{4m_{\mathrm{xy}}^2}\right]\xi(x,z)\,, \tag{5}$$

where the full (envelope) wavefunction $\psi'(x, y, z)$ is given by:

$$\psi'(x, y, z) = e^{\mathrm{i}k_y y}\, e^{-\mathrm{i}\frac{m_{\mathrm{c}}}{m_{\mathrm{xy}}}k_y x}\,\xi(x, z)\,, \tag{6}$$

with $1/m_{\mathrm{c}} = (m_{\mathrm{x}} + m_{\mathrm{y}})/(2m_{\mathrm{y}}m_{\mathrm{x}})$ and $1/m_{\mathrm{xy}} = 1/m_{\mathrm{x}} - 1/m_{\mathrm{y}}$. We calculate the electron wavefunctions $\xi_\mu(x, z)$ (labeled by the index μ) and the corresponding eigenvalues E^μ_{xz} by solving this eigenvalue problem. The second factor at the right-hand-side of Eq. (6) is incorporated into calculations at a later stage.

The more complicated form of the Schrödinger equation, Eq. (5), must be employed to treat only four of the six valleys, since the in-plane rotation from the [100] to the [110] direction does not affect the two ellipsoids with $m_{\mathrm{z}} = m_{\mathrm{L}}$. Here $m_{\mathrm{L}} = 0.91m_0$ (m_0 is the mass of an electron) is the longitudinal effective mass (we also assume $m_{\mathrm{T}} = 0.19m_0$ for the transverse effective mass). These valleys can be treated using the slightly simpler form given by Eq. (2).

The spatial mesh used to perform the simulation consists of 100 points in the x direction and 105 points in the z direction. The Hamiltonian $\mathbf{H}$ becomes a $10\,500 \times 10\,500$ matrix. The eigenvalues up to $5\,k_{\mathrm{B}}T$ above Fermi level are calculated, which typically amounts to around 2000 eigenvalues for the 40 nm device and around 400–1200 for the smaller devices.

B. Charge distribution and Poisson equation

The Poisson equation is solved on the two-dimensional cross-section of the device,

$$\nabla^2 V(x, z) = \frac{e^2}{\epsilon_{\mathrm{Si}}}\left[p(x, z) - n(x, z) + N_{\mathrm{A}}(x, z) - N_{\mathrm{D}}(x, z)\right], \tag{7}$$

where p, n, N_{A}, N_{D} are the hole, electron, donor, and acceptor doping distributions, respectively, and ϵ_{Si} is the permittivity of silicon. This is a linear system,

$$\mathbf{P} \cdot \mathbf{V} = \mathbf{D}\,, \tag{8}$$

where $\mathbf{P}$ is an $N \times N$ matrix containing the terms that represent the ∇^2 operator, $\mathbf{V}$ is a $N \times 1$ matrix containing the potential distribution at each of the N mesh points, and $\mathbf{D}$ is a $N \times 1$ matrix containing the charge terms.

The electron wavefunctions obtained from the solution of the closed-system are used to calculate the electron charge distribution in the device. This can be obtained by accounting for the density of states along the "homogeneous" (out-of-plane) direction y obtaining the well-known expression:

$$n(x, z) = \sum_\mu \frac{1}{\pi\hbar}\sqrt{\frac{m_{\mathrm{y}}k_{\mathrm{B}}T}{2}}F_{-\frac{1}{2}}\left(\frac{E_{\mathrm{F}} - E^\mu_{\mathrm{xz}}}{k_{\mathrm{B}}T}\right)|\xi_\mu(x, z)|^2\,, \tag{9}$$

where ξ_μ and E_μ are the wavefunction and energy of the quantum state μ, k_{B} is Boltzmann's constant, E_{F} is the Fermi energy measured from the bottom of the conduction band, and $F_{-\frac{1}{2}}$ is the Fermi-Dirac integral of order $-1/2$, defined by Eq. (11).

The hole charge distribution, p, is calculated using the classical three-dimensional density of states,

$$p(x,z) = \frac{1}{2\sqrt{\pi}}\left(\frac{2m_{\mathrm{h}}k_{\mathrm{B}}T}{\pi\hbar^2}\right)^{\frac{3}{2}} F_{\frac{1}{2}}[E_{\mathrm{F}} - V(x,z)]\,, \tag{10}$$

where $m_{\mathrm{h}} = 0.8m_0$ is the hole effective mass, the Fermi energy E_{F} is now measured from the top of the valence band, and, $F_{\frac{1}{2}}$ is the Fermi-Dirac integral of order 1/2, defined by Eq. (11).

The Fermi-Dirac integral of order σ is written as:

$$F_\sigma(\eta) = \int_0^\infty \frac{\varepsilon^\sigma \mathrm{d}\varepsilon}{1+\exp(\varepsilon-\eta)}\,. \tag{11}$$

Eq. (11) is computed using the Gauss-Legendre quadrature method.[28] The Poisson equation is solved over the same region as the Schrödinger equation and, therefore, the Poisson matrix **P** is also a 10 500 × 10 500 matrix.

C. Self-consistent scheme

The electrostatic potential in the device is obtained by solving the Schrödinger and Poisson equations self-consistently employing a conventional Newton iteration scheme.

The Fermi level of the device is first fixed at a value that results in charge neutrality deep in the substrate of the device. An initial "guess" for the electrostatic potential is made so as to satisfy the condition of charge neutrality at each point in the device, using the three-dimensional density-of-states for electrons and holes. The Schrödinger equation is then solved numerically using this initial guess for the potential and the electron wavefunctions $\xi_\mu(x,z)$ and the corresponding energies, E^μ_{xz}, are calculated. The electron charge distribution is then calculated from these wavefunctions and the hole charge distribution using the classical expression Eq. (10). Thereafter, the Poisson equation is solved numerically, using the electron, hole, and doping charge densities, to generate a new potential. The root-mean-square ("infinity-norm") error between the "new" and "old" potential is then determined. If the error is greater than a predefined minimum value, typically of the order of 10^{-7} eV, then the procedure is repeated using the "new" potential. Otherwise, the iterative procedure ends and the "new" potential is the equilibrium electrostatic potential in the device.

In order to accelerate the convergence of this iterative procedure, Newton's method[29] is used. In this method, the Poisson equation is not solved directly to generate the "new" potential. Rather, a Jacobian matrix **J** is constructed using first order derivatives (with respect to the potential at each mesh point) of the classical expressions of electron and hole charge distributions and the new potential is obtained from:

$$\mathbf{V}_{\mathrm{new}} = \mathbf{V}_{\mathrm{old}} - \mathbf{V}_{\mathrm{N}}, \tag{12}$$

where $\mathbf{J}\cdot\mathbf{V}_{\mathrm{N}} = \mathbf{P}\cdot\mathbf{V}_{\mathrm{old}} - \mathbf{D}$.

The self-consistent simulation typically achieves the desired degree of convergence in 25–30 iterations. In the IBM AIX7 P55 computer cluster we have used, this process requires about 48 central-processing-unit (CPU) hours for each bias point in the case of the 40 nm device. Devices with smaller channel lengths are less computationally intensive with the 10 nm device, for example, requiring only about 2 CPU hours to reach convergence.

D. 2D Schrödinger equation with open boundary conditions and electron current

In the closed boundaries system, the wavefunction is assumed to vanish outside the device, so there is no current flow through the device. However, we are interested in the open system in which electrons can flow into and out of the device. This describes the behavior of the system in the presence of an applied drain-source voltage V_{DS}. The method we follow is the QTBM proposed by Lent and Kirkner.[18] The source and drain contacts are imagined as infinite leads going into the device, as illustrated schematically in Fig. 3. A separate coordinate system (ω_r, κ_r) is defined for each lead r, as shown in the figure. The potential is assumed to be constant along the direction ω_r and, along κ_r, the potential profile is assumed to be the same as that along the lead-device interface. Another assumption made is that outside the device as well as outside the lead edges, the wavefunction vanishes. Using these conditions, the wavefunctions in the leads can be separated into two independent components—traveling waves along ω_r and wavefunctions with a discretized energy spectrum along κ_r due to the confinement in that direction. The latter part is determined by solving the one-dimensional Schrödinger equation along κ_r,

$$-\frac{\hbar^2}{2m_z}\frac{\partial^2\varphi^r_m(\kappa_r)}{\partial\kappa_r^2} + V_r(\kappa_r)\varphi^r_m(\kappa_r) = E^r_m\varphi^r_m(\kappa_r), \tag{13}$$

where φ^r_m and E^r_m represent the m_{th} eigen-state and eigenenergy, respectively in lead r, $V_r(\kappa_r)$ is the potential along κ_r in lead r and m_z is the quantization mass. These eigenstates are normalized along κ_r direction, and the total wavefunction in lead r is then given by:

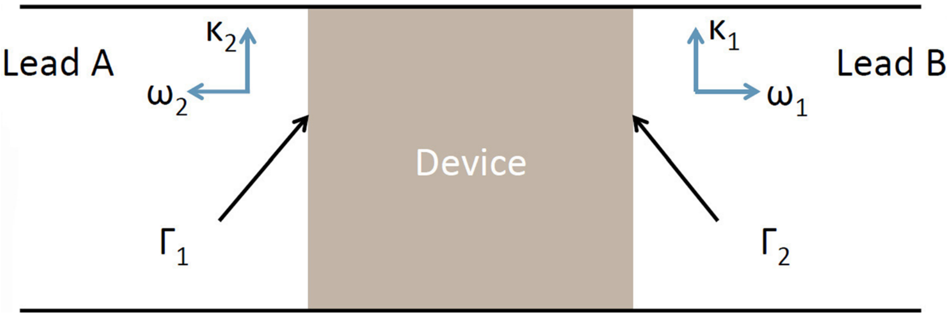

FIG. 3. Schematic illustration showing the implementation of the Quantum Transmitting Boundary Method used to simulate the QW nMOSFETs.

$$\Psi_{\beta,r}(\omega_r,\kappa_r)=\sum_{m=1}^{N_r}\left[a_m^r\varphi_m^r(\kappa_i)e^{-\mathrm{i}k^r_{m,\beta}\omega_r}+b^r_{m,\beta}\varphi_m^r(\kappa_i)e^{\mathrm{i}k^r_{m,\beta}\omega_r}\right]$$
$$+\sum_{m=N_r+1}^{\infty}b^r_{m,\beta}\varphi^r_m(\kappa_r)e^{-k^r_{m,\beta}\omega_r}\,. \tag{14}$$

Here, the terms inside the first summation on the right-hand-side represent the N_r traveling waves (traveling modes) with energy (along the x and z directions) $E_\beta > E^r_m$, going into and out of the device, respectively, through lead r. The index β denotes the different wavefunctions $\Psi_{\beta,r}$ and the corresponding energies E_β. The third term on the right-hand-side represents the evanescent modes with energy $E_\beta < E^r_m$. The coefficients a^r_m's are chosen as input for the different waves traveling into the device, while the coefficients $b^r_{m,\beta}$ need to be determined. The wavevectors k^r_m for the traveling modes are given by: $[2m_\mathrm{x}(E_\beta-E^r_m)]^{1/2}/\hbar$ and, for the evanescent modes, by: $[2m_\mathrm{x}(E^r_m-E_\beta)]^{1/2}/\hbar$. The energy E_β will be referred to as the "injection energy."

The boundary conditions at the interface Γ_r between the device and lead r involve the continuity of both the wavefunction at the interface, $\phi_{\beta,r}|_{\Gamma_r}=\Psi_{\beta,r}(\omega_r=0,\kappa_r)$ and the normal derivative, $\nabla\phi_{\beta,r}.\hat{\kappa_r}|_{\Gamma_r}=\nabla\Psi_{\beta,r}(\omega_r=0,\kappa_r).\hat{\kappa_r}$ for all $m\le N_r$. Here $\phi_{\beta,r}(x,z)$ represents the wavefunction inside the device. Using Eq. (14) and combining the two boundary conditions together:

$$\nabla\phi_{\beta,r}.\hat{\kappa_r}|_{\Gamma_r}=\sum_{m=1}^{N_r}\mathrm{i}k^r_m\varphi^r_m(\kappa_r)$$
$$\times\left(-2a^r_m+\int_0^{d_r}\mathrm{d}\kappa_\mathrm{r}\varphi^\mathrm{r}_\mathrm{m}(\kappa_\mathrm{r})\phi_{\beta,\mathrm{r}}(\omega_\mathrm{r}=0,\kappa_\mathrm{r})\right)$$
$$-\sum_{m=N_r+1}^{\infty}k^r_m\varphi^r_m(\kappa_r)\int_0^{d_r}\mathrm{d}\kappa_\mathrm{r}\varphi^\mathrm{r}_\mathrm{m}(\kappa_\mathrm{r})\phi_{\beta,\mathrm{r}}(\omega_\mathrm{r}=0,\kappa_\mathrm{r})\,. \tag{15}$$

Here, d_r is the vertical height of lead r, which in our case is same as the height of the device. In order to obtain this expression, we use the relation:

$$b^r_{m,\beta}=\int_0^{d_r}\mathrm{d}\kappa_\mathrm{r}\varphi^\mathrm{r}_\mathrm{m}(\kappa_\mathrm{r})\phi_{\mathrm{m},\beta,\mathrm{r}}(\omega_\mathrm{r}=0,\kappa_\mathrm{r})-\mathrm{a}^\mathrm{r}_\mathrm{m}\,. \tag{16}$$

Eq. (16) stems from the fact that the wavefunctions φ are eigenstates of a Hermitian operator (the one-dimensional Schrödinger Hamiltonian), and so are mutually orthogonal. Note that the wavefunctions ϕ in Eq. (16) depend on m as well, since they are calculated separately for each m.

The two-dimensional Hamiltonian $\mathbf{H}$, determined in Sec. III A, is again used here to calculate the wavefunctions $\phi_{m,\beta,r}$ in the device. However, the matrix elements that correspond to the edges of the device are modified to account for the proper boundary conditions described above. The equilibrium electrostatic potential obtained from the self-consistent solution of the closed system is utilized here. The resulting linear system is solved to obtain the electron wavefunctions $\phi_{m,\beta,r}$ separately for different injection energies E_β and the different traveling modes m. These wavefunctions are then used to calculate the transmission coefficient, local density-of-states (LDOS) and, most importantly, the current. The transmission coefficient $T^\mathrm{SD}(E_\beta,m_1)$ for a traveling mode m_1 in lead $r=\mathrm{S}$ ("source") with injection energy E_β, going into lead $r=\mathrm{D}$ (drain), is defined as:

$$T^\mathrm{SD}(E_\beta,m_1)=\sum_{m_2}\frac{|b^\mathrm{D}_{m_2,\beta}|^2}{|a^\mathrm{S}_{m_1}|^2}\sqrt{\frac{E_\beta-E^\mathrm{B}_{m_2}}{E_\beta-E^\mathrm{S}_{m_1}}}\,, \tag{17}$$

where m_2 is the number of traveling modes in lead D. In order to compensate for discretization errors and maintain unitariety, the following discretized version of the transmission coefficient is used:

$$T^\mathrm{SD}(E_\beta,m_1)=\sum_{m_2}\frac{|b^\mathrm{D}_{m_2,\beta}|^2}{|a^\mathrm{S}_{m_1}|^2}\sqrt{\frac{E_\beta-E^\mathrm{B}_{m_2}}{E_\beta-E^\mathrm{S}_{m_1}}}$$
$$\times\sqrt{\frac{2\hbar^2-m_x\left(E_\beta-E^\mathrm{D}_{m_2}\right)\Delta x^2}{2\hbar^2-m_x(E_\beta-E^\mathrm{S}_{m_1})\Delta x^2}}\,. \tag{18}$$

The two-dimensional LDOS for lead r, $\mathcal{D}^r_\mathrm{loc}(E,x,z)$ can be calculated for each injection energy, E_β, from the expression:

$$\mathcal{D}^r_\mathrm{loc}(E_\beta,x,z)=\sum_{m=1}^{N_r}2\sqrt{\frac{2m_x}{\hbar^2}}\frac{1}{\sqrt{E_\beta-E^r_m}}|\phi_{m,\beta,r}(x,z)|^2\,. \tag{19}$$

Note that the wavefunction $\phi_{m,\beta,r}$ has been normalized assuming infinite-volume normalization along the x direction, but using a finite-volume normalization along the z direction. Therefore $|\phi_{m,\beta,r}|^2$ has dimensions of inverse length, and the local density of states given in Eq. (19) expresses the states per unit energy and area.

The current flowing through the device is obtained by solving the open boundary-condition's Schrödinger problem for both the left and right sides (drain and source) using the electrostatic potential calculated for $V_\mathrm{DS}=0$, and then subtracting the calculated current from both sides by shifting the Fermi level of the drain side by a magnitude V_DS above that of the source. This is a good approximation since the applied V_DS is very small, of the order of 1–10 mV. Using Eq. (18), the current I_DS (per unit height along y direction) is finally calculated as:

$$I_\mathrm{DS}=\sum_{r=\mathrm{S}}^{\mathrm{D}}\sum_\beta\sum_{m=1}^{N_r}\eta_r\frac{e}{2\sqrt{2}}\frac{1}{\hbar^3}(m_y)^{\frac{1}{2}}|a^r_m|^2T^{rr\prime}(E_{\beta,m})$$
$$\times\Delta E_\beta\sqrt{2\hbar^2-m_x(E_\beta-E^r_m)\Delta x^2}$$
$$\times\int_0^\infty\frac{\mathrm{d}E_y}{E_\mathrm{y}^{1/2}}f(E_\beta+E_\mathrm{y})\,. \tag{20}$$

Here f is the Fermi-Dirac distribution function, $r'=\mathrm{S}$ when $r=\mathrm{D}$ and vice versa, e is the electron charge, ΔE_β is the energy interval used to discretize the spectrum of the injection energy E_β, T is the source-to-drain transmission coefficient, E_y is the energy in the y direction, and $\eta_{r=\mathrm{S}}=1$ for the source-to-drain term, $\eta_{r=\mathrm{D}}=-1$ for the drain-to-source term. Note that different Fermi energies are used for $r=\mathrm{S}$ and $r=\mathrm{D}$, separated by a magnitude V_DS. All the calculations described in Sec. III D are repeated for the 6 equivalent Si energy valleys.

The mesh used to perform the simulation consists of 2000 points in the x direction and 105 points in the z direction. The Hamiltonian $\mathbf{H}$ becomes a $210\,000 \times 210\,000$ matrix. Note that the mesh is the same as used before in the closed boundaries simulation, only interpolated over a finer grid. A linear interpolation of the electrostatic potential generated from the closed boundaries system is done over the finer mesh. Typically 400–600 injection energies in an energy range that varies depending on the size of the device, roughly 0.2 meV around the Fermi level, are used to calculate the current I_{DS}. The simulation is parallelized to run over the multiple CPU cores, each core being assigned a specific set of injection energies. The CPU time required to calculate I_{DS} for each gate bias (parallelized over 100 cores) is approximately 35 min.

IV. SIMULATION RESULTS AND DISCUSSION

The simulation methods we have just described in detail are implemented for values of the gate bias V_{GS} varying from 0 to 3 V and with V_{DS} of 1–10 mV. As mentioned above, smaller devices are simulated by reducing the dimensions of the nominal 40 nm device following the well-known scaling laws,[22,23] with the exceptions and concerns that we have already discussed. The device characteristics are calculated at cryogenic temperatures of 10 K and 46 K, as well as at room temperature.

We present the current-voltage ($I_{DS} - V_{GS}$) characteristics of the 10 nm, 20 nm, and 40 nm devices in Figs. 4, 5, and 6, respectively. The 10 nm device exhibits NDT at $V_{GS} = 1.18$ V at a temperature of 10 K. The NDT is reduced to a small "kink" in the $I_{DS} - V_{GS}$ characteristics as the temperature is raised to 46 K, as shown in Fig. 4, and disappears altogether at room temperature. A similar behavior is seen in the 20 nm device: The NDT is seen at $V_{GS} = 1.93$ V at 10 K (Fig. 5), but not at room temperature. A "kink" is seen in the current voltage characteristics of the 30 nm device at 10 K (not shown), but no defined NDT is detected. Finally, the 40 nm device does not exhibit any NDT peaks at any temperature (Fig. 6). Note that in all devices as the temperature increases, the current I_{DS} increases exponentially, as a result of the thermionic emission over the pSDE potential barriers (discussed later).

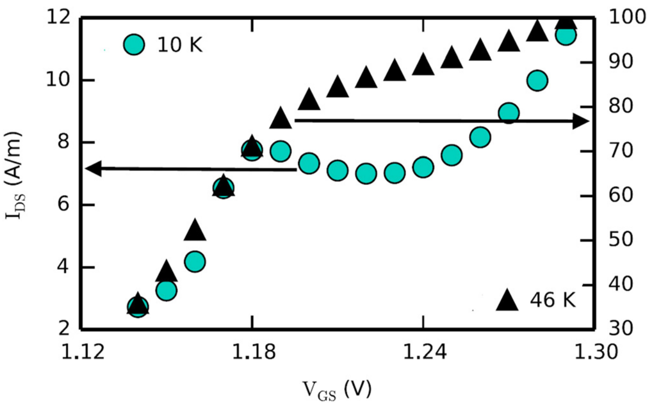

FIG. 4. Calculated $I_{DS} - V_{GS}$ characteristics at 46 K (black triangles) and 10 K (cyan circles) for the 10 nm device with $V_{DS} = 10$ mV.

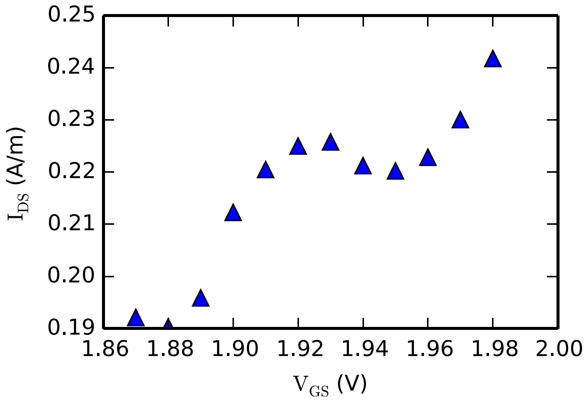

FIG. 5. Calculated $I_{DS} - V_{GS}$ characteristics for the 20 nm device at 10 K. $V_{DS} = 1$ mV.

To confirm that the NDT seen at low temperatures is indeed the result of resonant tunneling through the two-dimensional quantum well created by the pSDEs, we plot in Fig. 7, the average local density-of-states (LDOS) for all devices. The contour plots show the two dimensional LDOS computed along the length (x direction) of the device, averaged over a thin "vertical" region (z direction). The energy scale identifies different injection energies. The LDOS is shown only for injection from the source (left contact). In Fig. 8, we show the transmission coefficient as a function of injection energy for the traveling modes that exhibits the best resonant behavior (wherever applicable). Only the LDOS and transmission coefficient *vs.* total injection energy plots corresponding to the two Si ellipsoidal energy valleys having the longitudinal mass in the z direction ("unprimed subbands") are shown in Figs. 7 and 8, since these are the only valleys that exhibit NDT.

The dark "streaks" in the middle of the channel seen in Fig. 7 (first and second frames from the left, respectively) show the presence of the quasi-bound states in the 10 nm and 20 nm devices. The corresponding peaks seen in the transmission coefficient (first and second frames from the left in Fig. 8, respectively) at those energies confirm that resonant tunneling across the QW is indeed the origin of the NDT. More so, the NDT occurs at a gate bias for which the energy

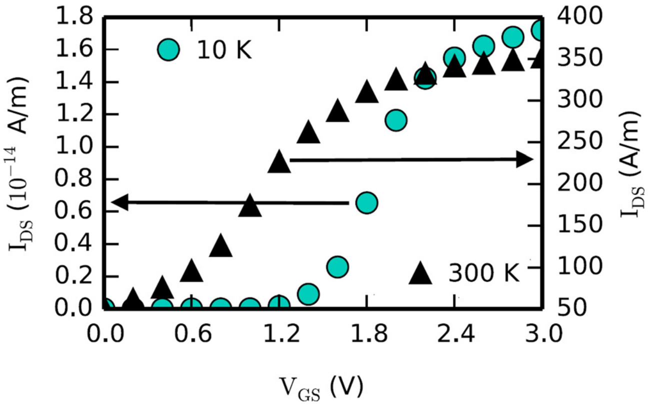

FIG. 6. Calculated $I_{DS} - V_{GS}$ characteristics at 10 K (cyan circle) and 300 K (black triangle) for the 40 nm device with $V_{DS} = 10$ mV. Negligible current is seen at 10 K. Thermionic emission is therefore the major cause of the vastly larger current observed at 300 K.

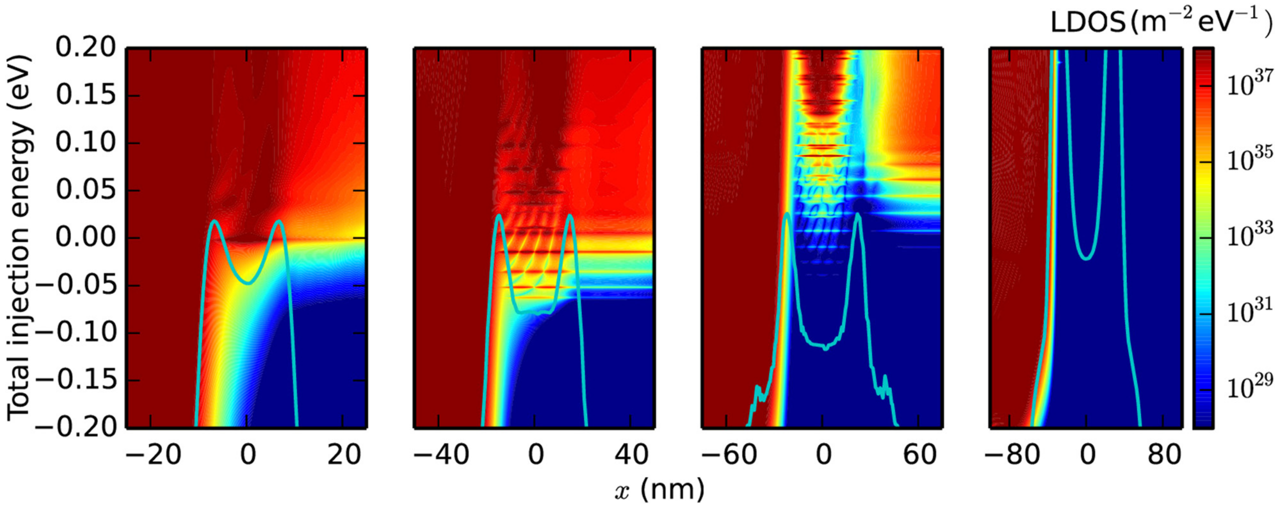

FIG. 7. Average LDOS in the channel for 10 nm, 20 nm, 30 nm, and 40 nm devices (from left to right) at 10 K. The cyan colored lines represent the potential-energy profile at the semiconductor/gate-insulator interface in each device. The energies are measured with respect to the Fermi energy in the source contact. The bias conditions are $V_{GS} = 1.18$ V, 1.93 V, 2.2 V, and 2.2 V, respectively (left to right).

of the first bound state (10 nm device) or fourth bound state (20 nm device) crosses the Fermi energy of the electron gas in the source region. The LDOS for the 30 nm device (third frame from the left in Fig. 7) shows closely spaced bound states, and no quasi-bound states can be seen in the LDOS of the 40 nm device (right-most frame in Fig. 7). The corresponding transmission coefficient (two rightmost frames in Fig. 8), show closely spaced peaks with very low transmission coefficients (30 nm) or no peaks (40 nm), explaining the absence of any NDT seen in the current-voltage characteristics. We should also emphasize that all devices are affected by a significant punch-through current. The severity of this problem is reduced in the shorter devices, thanks to the reduced height of the potential pSDE barriers, a result of device scaling, as we have mentioned above. However, this comes at the price of a larger thermionic current.

The absence of any NDT peak (or "kink") in the 40 nm device at any temperature can be explained from the behavior seen in the smaller devices. Going from the 10 nm to the 30 nm device, the transmission coefficient peaks (Fig. 8) become sharper and lower in magnitude. This can be expected, since scaling the device results in the pSDE potential barriers to become narrower and reduced in height. Narrower barriers cause a broadening of the peaks in the transmission coefficient, as the electron lifetimes in these bound states become shorter as a result of the higher probability of "leaking" out. The transmission coefficient scales inversely and exponentially with the height of the barriers. Already in the 30 nm device, the transmission peaks are very sharp, spanning an extremely small energy-width of the order of 10^{-7} eV, and have low transmission coefficients of the order $10^{-12} - 10^{-7}$. This implies that pSDE barriers in the 40 nm device are too opaque to allow any significant current through the channel of the device. Therefore, no NDT is seen theoretically and the current mostly consists of punch-through current. Moreover, the bound states are energetically grouped closely together. This results in a quasi-continuum of states, a situation that is far from ideal for the generation of NDT at any finite temperature.

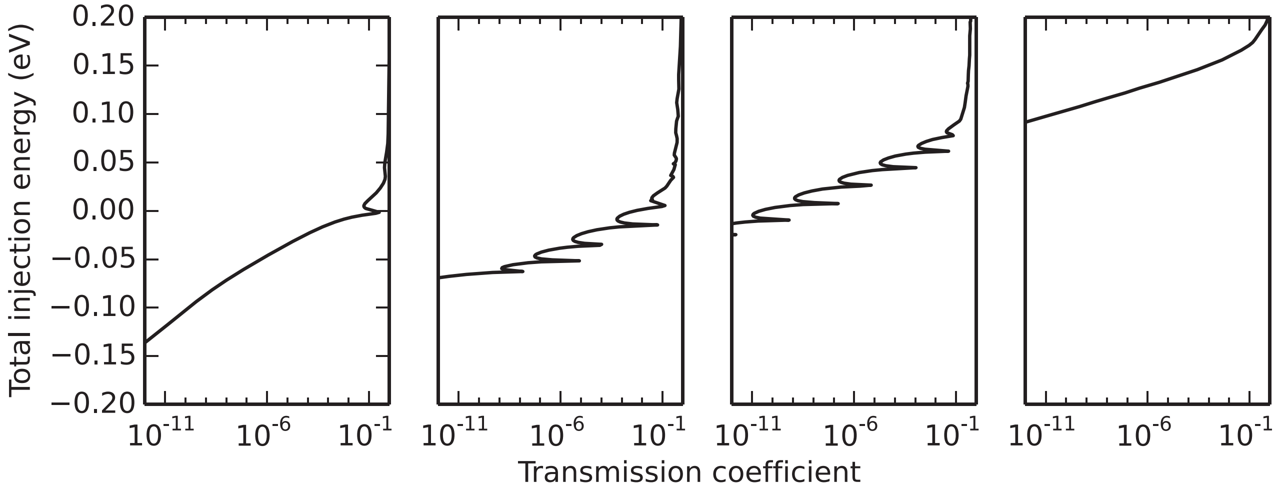

FIG. 8. Transmission coefficient *vs.* injection energy for a particular traveling mode is in the 10 nm, 20 nm, 30 nm, and 40 nm devices, respectively. The traveling mode energies E^r_m are, from left to right, −0.83 eV, −0.49 eV, −0.26 eV, and −0.04 eV. These energies are chosen since they exhibit the best resonant behavior in the respective devices (wherever applicable). $V_{GS} = 1.18$ V, 1.93 V, 2.2 V, and 2.2 V, respectively (left to right).

The discrepancy between the theoretical predictions and the experimentally observed device characteristics can only be explained by assuming that the doping profile used in the simulations is significantly different from the actual doping profile of the fabricated device. This is a common phenomenon not unlikely to occur in this particular case. As mentioned before, the doping profile used in our theoretical calculations is generated by Sentaurus® TCAD process-simulation tool. The actual doping profile of the fabricated device is generally somewhat different from the TCAD result because of inaccuracies in the modeling of implant dopant diffusion, even if the device is made using industrial CMOS processing standards. Moreover, in this case, the design of the pSDEs is a significant departure from the conventional CMOS design. These sharp, heavily doped, and highly localized p-type regions may easily be broadened by lateral diffusion enhanced by the defects (especially vacancies) caused by the ion implantation. This is an effect that is notoriously difficult to predict accurately.

In order to assess the sensitivity of the NDT on details of the doping profile of the pSDE, we have modified the original simulated doping profile of the 40 nm device by reducing the peak p-doping by a factor of 3 and by widening laterally the barriers by a factor of 1.8. This is done to roughly simulate the broadening of the pSDEs due to lateral diffusion. The modified doping profile and the corresponding LDOS distribution for a specific gate bias and applied V_{DS} are shown in Figs. 9 and 10, respectively. Interestingly, the modified doping profile results in the appearance of bound states, as indicated by the peaks of the transmission coefficient having a decent magnitude, and in the occurrence of resonant tunneling. This confirms that the device characteristics are very sensitive to small changes of the profile of the pSDEs. Therefore, we speculate that the devices may behave as intended, with the occurrence of resonant tunneling along the channel through the quasi-bound states in the lateral QW, provided the pSDE doping differs, and not too appreciably, from the original design specifications and from the Sentaurus® simulated profile.

More generally, in order to produce NDT, tunneling through the quasi-bound states in the lateral QW must occur with a probability large enough to overcome the undesired "leakage paths" of thermionic emission over the pSDE barriers and/or punch-through "around them." Peaks of the transmission coefficient in the range of 10^{-2} or larger are required to result in NDT. Clear examples are the first bound state in the 10 nm device and the fourth bound state in the 20 nm device, both giving rise to peaks in the NDT at 10 K. On the contrary, the 3 lowest-energy bound states (20 nm device) have much lower transmission coefficients and do not result in any significant tunneling current, even though the energy spacing between them might be sufficient enough to produce NDT at 10 K. On the contrary, in the 30 nm device, the transmission coefficient peaks are much weaker, of the order of $10^{-12} - 10^{-7}$, and no NDT is seen at low temperature, since the current flows in punch-through, or at high temperatures, because the current now is dominated by thermionic emission over the barriers.

The important question that we need to answer is: What is the maximum gate length for which NDT could be observed at room temperature? And what device design may be required to reach this goal? Performing simulations with different gate lengths, doping profiles, at different temperatures, or even considering different alternative device structures, is an almost impossible task as we scale devices down to the 5 or 7 nm gate length. However, the results we have presented so far allow us to formulate an "educated guess." We have already noticed that NDT can be observed when $k_B T \leq DE - dE$, where DE is the spacing of the quasi-bound state in the lateral QW (that is: the energetic spacing of the peaks in the transmission coefficient shown in Fig. 7) and dE is the full broadening of the confined states. We have also emphasized the fact that our device-scaling procedure results in a different width and height of the pSDE barriers at different channel lengths. Specifically, Fig. 8 shows that DE increases as the devices are scaled, from ≈20 meV for the

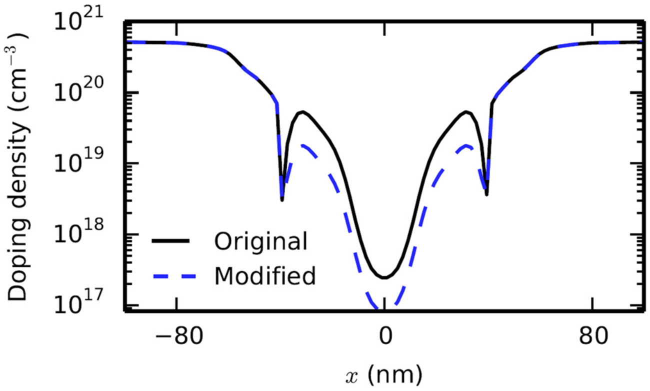

FIG. 9. Magnitude of the modified net doping profile along the Si/gate-insulator interface of the 40 nm device compared to the original doping profile. The pSDE's have been broadened and their doping concentration has been reduced, to increase the conduction via resonant tunneling.

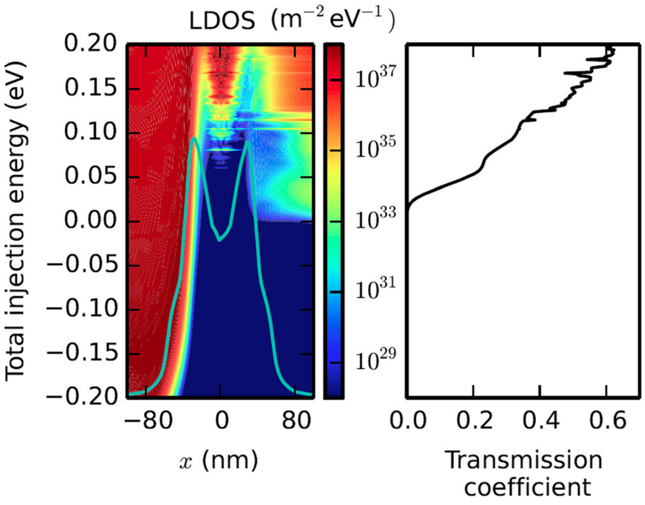

FIG. 10. Left: Average LDOS in the channel of the 40 nm device with the modified doping profile (shown in Fig. 9), at 10 K. The cyan colored line represents the potential-energy profile at the semiconductor/gate-insulator interface. Right: Transmission coefficient *vs.* total injection energy for the particular traveling mode that exhibits the highest resonant transmission. The energy E_m^r of the traveling mode is −0.16 eV and the gate bias $V_{GS} = 2.0$ V.

20 nm to 37 meV for the 10 nm device. Also, in these devices $dE \approx 3$ meV (20 nm) and 5 meV (10 nm device). Therefore, at 300 K, 25 meV $= k_{BT} < DE - dE \approx 32$ meV for the 10 nm device. Thus, it is the thermionic "leakage path" that hides the expected NDT. Assuming $DE \sim 1/L^2$, where L is the gate length, DE would be as large as 75 or 150 meV in devices scaled to 7 and 5 nm, respectively. This would be more than sufficient to ensure the occurrence of NDT at 300 K. However, as we have already remarked, the possibility of having current flowing via the thermionic emission over low pSDE-barriers and via punch-through around high pSDE-barriers would have to be minimized. The latter requirement likely demands that we move from a "bulk" MOSFET design to ultra-thin-body (UTB) silicon-on-insulator (SOI) structures. This is required by scaling to the 5 nm gate-length, regardless of the particular goal we have in mind. This is already "conventional" Si CMOS VLSI technology and no major difficulties should be expected. Much more difficult is the constraint posed by the necessity of reducing the thermionic leakage path. This would require narrow and high pSDE barriers. This would have the welcome effect of reducing the energetic width of the quasi-bound states, dE, that scales as $dE \sim \exp(-\alpha W)$, where W is the width of the pSDE barriers, assumed to be proportional to L with proper scaling (α is a quantity that depends on the energy of the quasi-bound-state in the QW and on the electron effective mass in the gap, m_x). Clearly, a barrier-width W of the order of a few nm is hard to envision as achievable by ion implantation. However, the use of larger, slower-diffusion acceptor impurities, such as In in place of B, and use of doping techniques not relying on ion implantation and subsequent rapid-thermal annealing (RTA) steps, such as the low-temperature epitaxy employed for thin-base Si/SiGe epitaxial bipolar transistors,[31] could provide a solution if implemented in a "horizontal" epitaxy, such as the technology used to regrow S/D regions in uniaxially strained-Si p-channel MOSFETs.[32]

We should finally observe that our simulations have been performed assuming the ballistic quantum transport. Introducing scattering into the simulation, for example, via a Master equation[30] or a NEGF approach,[15] would lead to a larger broadening dE of the resonant states. Moreover, non-parabolic corrections to the electron dispersion, ignored here, would reduce the energetic separation DE of the transmission resonances. Therefore, our estimates should be considered moderately optimistic, but only "moderately" so.

V. CONCLUSIONS

We have shown with quantum-transport (QTBM) simulations that lateral-quantum-well Si nMOSFETs exhibit the desired NDT at low temperature and for gate lengths shorter than about 20 nm. Extrapolating from our results, devices with a gate length of 10 nm and lower should exhibit a sharp NDT signature even at room temperature, provided there is sufficient mitigation of the thermionic and punch-through currents. The former plays a crucial role in suppressing the NDT, the latter is favored over the tunneling current for high pSDE barriers. We have argued that alternative device designs (UTB SOI devices and/or epitaxial pSDE barriers) with sub-10 nm gate length are required to observe NDT at room temperature. Discussing processing and fabrication issues related to these scaled devices is a problem that transcends the scope of this work. However, the theoretical prediction of NDT in 10 nm devices leaves room for a moderate optimism. We have also shown the strong dependence of the resonant tunneling current on details of the doping profile of the pSDE pockets. This suggests that our inability to explain the NDT observed in 40 nm gate-length QW nMOS devices is likely due to the uncertainty of the actual doping profiles. This observation also bolsters our optimism.

ACKNOWLEDGMENTS

This work has been supported by the National Science Foundation under Grant No. ECCS-1403421.

[1] See http://www.itrs2.net/itrs-reports.html for International Technology Roadmap for Semiconductors 2.0, 2015 Edition.
[2] L. Esaki, Phys. Rev. B **109**, 603 (1958).
[3] L. Esaki, IEEE Trans. Electron. Devices **23**, 644 (1976).
[4] L. L. Chang, L. Esaki, and R. Tsu, Appl. Phys. Lett. **24**, 593 (1974).
[5] J. R. Söderström, D. H. Chow, and M. C. McGill, Appl. Phys. Lett. **55**, 1094 (1989).
[6] L. F. Luo, R. Beresford, and W. I. Wang, Appl. Phys. Lett. **55**, 2023 (1989).
[7] T. C. L. G. Sollner, W. D. Goodhue, P. E. Tannenwald, C. D. Parker, and D. D. Peck, Appl. Phys. Lett. **43**, 588 (1983).
[8] R. Kohler, A. Tredicucci, F. Beltram, H. E. Beere, E. H. Linfield, A. G. Davies, D. A. Ritchie, R. C. Iotti, and F. Rossi, Nature **417**, 156 (2002).
[9] O. Astafiev, S. Komiyama, T. Kutsuwa, V. Antonov, Y. Kawaguchi, and K. Hirakawa, Appl. Phys. Lett. **80**, 4250 (2002).
[10] C. Naquin, M. H. Edwards, G. Mathur, T. Chatterjee, and K. Maggio, Appl. Phys. Lett. **105**, 213507 (2014).
[11] C. Naquin, M. Lee, H. Edwards, G. Mathur, T. Chatterjee, and K. Maggio, J. Appl. Phys. **118**, 124505 (2015).
[12] F. Beltram, F. Capasso, S. Luryi, S. G. Chu, A. Y. Cho, and D. L. Sivco, Appl. Phys. Lett. **53**, 219 (1988).
[13] W. Pötz, J. Appl. Phys. **66**, 2458 (1989).
[14] R. Lake and S. Datta, Phys. Rev. B **45**, 6670 (1992).
[15] R. Lake, G. Klimeck, R. C. Bowen, and D. Jovanovic, J. Appl. Phys. **81**, 7845 (1997).
[16] N. C. Kluksdahl, A. M. Kriman, D. K. Ferry, and C. Ringhofer, Phys. Rev. B **39**, 7720 (1989).
[17] W. R. Frensley, Rev. Mod. Phys. **62**, 745 (1990).
[18] C. S. Lent and D. J. Kirkner, J. Appl. Phys. **67**, 6353 (1990).
[19] S. E. Laux, A. Kumar, and M. V. Fischetti, IEEE Trans. Nanotechnol. **1**, 255 (2002).
[20] S. E. Laux, A. Kumar, and M. V. Fischetti, J. Appl. Phys. **95**, 5545 (2004).
[21] E. Polizzi and N. Ben Abdallah, Phys. Rev. B **66**, 245301 (2002).
[22] R. H. Dennard, F. Gaensslen, H.-N. Yu, L. Rideout, E. Bassous, and A. LeBlanc, IEEE J. Solid-State Circuits **9**, 256 (1974).
[23] G. Baccarani, M. R. Wordeman, and R. H. Dennard, IEEE Trans. Electron. Devices **31**, 452 (1984).
[24] F. Stern, Phys. Rev. B **5**, 4891 (1972).
[25] J. Fang, W. G. Vandenberghe, B. Fu, and M. V. Fischetti, J. Appl. Phys. **119**, 035701 (2016).
[26] J. M. Luttinger and W. Kohn, Phys. Rev. **97**, 869 (1955).
[27] T. Ando, A. B. Fowler, and F. Stern, Rev. Mod. Phys. **54**, 437 (1982).
[28] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover Publications, Inc., New York, 1972).
[29] G. Dahlquest and A. Bjorck, *Numerical Methods* (Prentice-Hall, Englewood Cliffs, NJ, 1974).
[30] M. V. Fischetti, Phys. Rev. B **59**, 4901 (1999).
[31] D. L. Harame, J. H. Comfort, J. D. Cressler, E. F. Crabbé, J. Y.-C. Sun, B. S. Meyerson, and T. Tice, IEEE Trans. Electron. Devices **42**, 455 (1995).
[32] S. E. Thompson *et al.*, IEEE Trans. Electron. Devices **51**, 1790 (2004).