Simulating Quantum Transport in Nanoscale Transistors: Real versus Mode-Space Approaches

R. Venugopal,* Z. Ren, S. Datta, and M. S. Lundstrom

School of Electrical and Computer Engineering,

Purdue University, 1285 Electrical Engineering Building,

West Lafayette, Indiana 47907-1285

D. Jovanovic

Computational Materials Group, Motorola Labs

(Dated: June 24, 2002)

Abstract

In this paper, we present a computationally efficient, two-dimensional quantum mechanical simulation scheme for modeling electron transport in thin body, fully depleted, n-channel, silicon-on-insulator transistors in the ballistic limit. The proposed simulation scheme, which solves the non-equilibrium Green's function equations self-consistently with Poisson's equation, is based on an expansion of the active device Hamiltonian in decoupled mode-space. Simulation results from this method are benchmarked against solutions from a rigorous two-dimensional discretization of the device Hamiltonian in real-space. While doing so, the inherent approximations, regime of validity and the computational efficiency of the mode-space solution are highlighted and discussed. Additionally, quantum boundary conditions are rigorously derived and the effects of strong off-equilibrium transport are examined. This paper shows that the decoupled mode-space solution is an efficient and accurate simulation method for modeling electron transport in nanoscale, silicon-on-insulator transistors.

*Electronic address: venugopr@ecn.purdue.edu

I. INTRODUCTION

As CMOS technology progresses, device dimensions have been scaled into the nanometer regime [1] [2]. Therefore, in the future, transistors may operate near their ballistic limit rendering it important to understand ballistic device physics. As transistor dimensions are scaled, quantum effects, which affect the threshold voltage (confinement), gate capacitance (charge centroid shift), off-current (source barrier tunneling) and gate leakage begin to manifest themselves, and semiclassical methods that disregard these effects are inadequate in capturing the physics of ballistic transport. This paper describes computationally efficient, quantum mechanical transport models for ballistic n-channel MOSFETs based on the non-equilibrium Green's function formalism (NEGF) [3] [4].

Quantum modeling approaches rely on a self-consistent solution of the Schrödinger and Poisson equations in order to obtain the charge distribution and current for a specific device geometry. A solution to the Schrödinger equation can be pursued at varying levels of complexity depending on the nature of the device under study and the desired degree of accuracy. In previous work, a simplified two-dimensional (2D) simulator for fully depleted, silicon-on-insulator (SOI) device geometries that coupled a 1D Schrödinger and a 1D Boltzmann solver to model 2D transport has been described [5] [6]. Although insightful, this solution has limited applicability as it can neither be easily extended to treat scattering within a quantum mechanical framework nor can it capture the effect of source-to-channel tunneling accurately. In order to treat these quantum effects, a general simulation scheme based on the NEGF formalism, which includes a 2D discretization of the Hamiltonian operator is necessary.

Two device geometries, namely, bulk and SOI are being studied currently from a scaling perspective. However, the SOI geometry with its good short channel immunity, is widely accepted as the device structure that may drive CMOS technology in the future [7]. For bulk devices, 2D solutions based on the Green's function formalism have been demonstrated by two groups [8] [9]. These solutions, which are based on a real-space discretization of the full 2D effective mass Hamiltonian, are computationally expensive. Such discretization is essential in order to treat quantum transport accurately in bulk MOSFETs because the confining effect of the gate is lost as carriers move from the source to the drain. However, this is not the case in thin body SOI MOSFETs where mobile charges are quantum confined

all the way from the source to the drain (due to the thin body and the two insulator geometry). For such geometries, the computational burden associated with the real-space solution can be greatly reduced (without compromising on accuracy) by expanding the 2D Hamiltonian in mode-space (characteristic modes of the Hamiltonian in the confinement direction) and by treating the first few occupied modes. To benchmark this simplified mode-space solution, we implement and apply, both, the 2D real-space solution and the simplified mode-space solution to simulate carrier transport in a thin body, fully depleted, DG n-MOSFET structure. The simulation results from the two approaches are compared and while doing so, the various approximations inherent in the mode-space solution, its realm of validity and the generality of the real-space solution are discussed. This paper aims to describe the numerical methods that one can use to simulate quantum transport in different transistor geometries with specific emphasis on the mode-space solution scheme. Quantum boundary conditions are also derived, and the quantum mechanical features of the simulation results are highlighted.

The paper is divided into the following sections: 1) Sec. II presents the real-space and mode-space solutions succinctly. The size of the problem associated with each method is highlighted. 2) Sec. III, compares simulation results obtained by applying both techniques to model ballistic transport in a thin body, fully depleted, DG MOSFET. 3) Sec. IV, critically examines the validity for the mode-space solution. 4) Sec. V summarizes key findings.

II. THEORY

The simulated device structure is shown in Fig.1a. A uniform rectangular grid with a grid spacing of a, along the x direction and b, along the z direction is used. The device is represented by a Hamiltonian matrix that is coupled to two infinite reservoirs, the source and drain. The source/drain (S/D) extension regions are terminated using open boundary conditions (no x dependence of the potential), and the Fermi level in these regions is specified by the applied voltage. The width of the device is assumed to be large, and the potential is assumed to be translationally invariant along the width (y dimension). A single band effective mass Hamiltonian is used to model carrier transport.

A. Real-Space solution

This section briefly explains the real-space simulation method with specific emphasis on the self-energy concept (which is used to quantum mechanically couple the active device to the infinite S/D contacts) and the size of the problem. In this simulation method, the effective mass Hamiltonian is discretized in 2D real-space (x, z), to solve for its retarded Green's function. We begin by expanding the 3D Hamiltonian for the device in terms of $\delta(x-x')\delta(z-z')$ and $e^{ik_j y/\sqrt{W}}$, where the plane wavefunction, $e^{ik_j y/\sqrt{W}}$, represents the device width (W). The quantum number, k_j, corresponds to the eigenenergy, $E_{k_j} = \hbar^2 k_j^2/2m_y^*$, where m_y^* is the electron effective mass in the y direction. The real-space delta functions, $\delta(x-x')$ and $\delta(z-z')$ with eigenvalues x' and z' respectively, combined with $e^{ik_j y/\sqrt{W}}$, form a complete and orthogonal expansion function set. On expansion, the Hamiltonian in block diagonal form is

$$H = \begin{bmatrix} h(x,z)+E_{k_1}I & 0 & \cdots & \cdots & \cdots \\ 0 & h(x,z)+E_{k_2}I & \cdots & \cdots & \cdots \\ 0 & \cdots & \ddots & 0 & \cdots \\ \cdots & \cdots & 0 & h(x,z)+E_{k_j}I & 0 \\ \cdots & \cdots & \cdots & 0 & \ddots \end{bmatrix} \tag{1}$$

where I is the identity matrix and

$$h(x,z) = \begin{bmatrix} \alpha_0 & \beta & 0 & \cdots & \cdots \\ \beta & \alpha_1 & \ddots & 0 & \cdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \cdots & 0 & \ddots & \alpha_{N_X} & \beta \\ \cdots & \cdots & 0 & \beta & \alpha_{N_{X+1}} \end{bmatrix} \tag{2}$$

is block tridiagonal. Note that Eqs. 1 and 2 represent infinite matrices. This is because the device is coupled to infinite leads which have not yet been replaced by appropriate boundary conditions. If we think of the device as being composed of vertical slices adjoining each other, then the α's represent coupling along the z direction within each slice while the β's represent coupling between adjacent slices. The α's, with indices less than one, represent successive slices going into the source contact while those with indices greater than N_X represent successive slices into the drain. The α's and β's are themselves block

matrices and are,

$$\alpha[x] = \begin{bmatrix} 2t_x + 2t_z - qV_1(x) & -t_z & 0 & \cdots & \cdots \\ -t_z & 2t_x + 2t_z - qV_2(x) & \ddots & \cdots & \cdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \cdots & 0 & \ddots & \alpha_{N_X} & -t_z \\ \cdots & \cdots & 0 & -t_z & 2t_x + 2t_z - qV_{N_Z}(x) \end{bmatrix} \tag{3}$$

$$\beta = \begin{bmatrix} -t_x & 0 & \cdots \\ 0 & -t_x & \cdots \\ \cdots & 0 & -t_x \end{bmatrix} \tag{4}$$

The V's (Eq. 3), represent the potential along a vertical slice at site x and t_x and t_z, the coupling energies between adjacent grid points in x and z respectively. These site coupling energies are given by,

$$t_x = \frac{\hbar^2}{2m_x^* a^2} \quad \text{and} \quad t_z = \frac{\hbar^2}{2m_z^* b^2} \tag{5}$$

It is clear from Eq. 1 that blocks representing different plane wave states (along the device width) are decoupled, as there is no scattering within the device. Also, E_{k_j} ranges between 0 and $+\infty$, accounting for all possible plane wave states. There is no restriction on the solution domain and it can be easily extended to include the insulator regions provided changes in the electron effective mass is correctly accounted for within the insulator and at the silicon/insulator interface when discretizing the effective mass Hamiltonian.

For each plane wave eigenenergy E_{k_j}, we write the retarded Green's function relevant to 2D transport as,

$$G(E) = \left[EI - [h(x,z) + E_{k_j} I]\right]^{-1} = [E(k_x, k_z)I - h(x,z)]^{-1} \tag{6}$$

where the in-plane energy is defined as $E(k_x, k_z) \equiv E - E_{k_j}$ (note that that $E(k_x, k_z)$ is just a notation to identify the in-plane energy and that the Hamiltonian in the $X - Z$ plane is in a real-space basis). We then account for the infinite leads by introducing an appropriate self-energy function [4]. Details of the self-energy (Σ) calculation are presented in the Appendix. The self-energy represents the effects on the finite device Hamiltonian due to the outgoing wavefunctions from an impulse excitation within the device [10]. It allows us to eliminate the huge S/D reservoirs and work solely within the device subspace

whose dimensions are much smaller. The size of the self-energy matrix is $(N_X \times N_Z)^2$. On including the self-energies, the final form of the Green's function matrix is

$$G[E(k_x, k_z)] = [E(k_x, k_z)I - h(x, z) - \Sigma_S - \Sigma_D]^{-1} \tag{7}$$

Note that the Green's function in a real-space representation, has a size of $(N_X \times N_Z)^2$.

Once the retarded Green's function is evaluated, electron density and terminal current can be computed. We define a new quantity in terms of the lead self-energies [3] [11].

$$\Gamma \equiv i(\Sigma - \Sigma^\dagger) \tag{8}$$

Physically the function Γ, determines the electron exchange rates between the S/D reservoirs and the active device region [11]. But in general it can be viewed as the measure of interaction strength due to any perturbation source. It should be noted that, although the device itself may be in a non-equilibrium state, electrons are injected from the equilibrium S/D reservoirs (Fermi level is uniquely fixed for all carriers based on the applied voltage). The spectral density functions due to the S/D contacts can be obtained as [4]

$$A_S = G\Gamma_S G^\dagger \quad \text{and} \quad A_D = G\Gamma_D G^\dagger \tag{9}$$

where $\Gamma_S \equiv i(\Sigma_S - \Sigma_S^\dagger)$, and $\Gamma_D \equiv i(\Sigma_D - \Sigma_D^\dagger)$ (for clarity, we use Γ_S or Γ_D to denote matrices the same size as G, with nonzero diagonal blocks $\Sigma_S - \Sigma_S^\dagger$ or $(\Sigma_D - \Sigma_D^\dagger)$). The spectral functions are $(N_X \times N_Z)^2$ matrices. Although full in general, only the diagonal entries are of importance as they represent the state density at each lattice node. Therefore significant savings in computational cost are derived by solving for the diagonal blocks of the spectral functions as opposed to the entire matrix using a recursive algorithm [12]. The source-related spectral function is filled up according to the Fermi function in the source contact, while the drain-related spectral function is filled up according to the Fermi function in the drain contact. Both the source and drain spectral functions contribute to the 3D electron density, which, for each in-plane energy is [4],

$$\begin{aligned} n\left[E(k_x, k_z)\right] \;=\; & \frac{1}{2\pi ab} \\ & \times \int_0^{+\infty} D \cdot \left[f(\mu_S, E(k_x, k_z) + E_{k_j})A_S + f(\mu_D, E(k_x, k_z) + E_{k_j})A_D\right] dE_{k_j} \end{aligned} \tag{10}$$

where f is the Fermi-Dirac statistics function, and $D = (2/\pi\hbar)\sqrt{m_y^*/2E_{k_j}}$, represents the transverse mode state density (including spin degeneracy). Since the spectral functions (Eq. 9) depend on the in-plane energy alone, they can be moved out of the integral in Eq. 10 which reduces to [4],

$$\begin{aligned} n\left[E(k_x,k_z)\right] &= \frac{1}{ab}\sqrt{\frac{m_y^* k_B T}{2\pi^3\hbar^2}} \\ &\times \left[F_{-1/2}\left(\mu_S - E(k_x,k_z)\right)A_S + F_{-1/2}\left(\mu_D - E(k_x,k_z)\right)A_D\right] \end{aligned} \tag{11}$$

where the Fermi-Dirac integral, $F_{-1/2}$, accounts for all transverse mode contributions (for an analytical approximation to $F_{-1/2}$ see [13]). To obtain the total 3D electron density, we need to integrate Eq. 11 over $E(k_x,k_z)$ and sum contributions from every conduction band valley. The 3D electron density is fed back to the Poisson equation solver for self-consistent solutions.

Once self-consistency is achieved, the terminal current can be expressed as a function of the transmission coefficient [11]. The transmission coefficient from the source contact to the drain contact is defined in terms of the Green's function as [4],

$$T_{SD} = \mathrm{Trace}[\Gamma_S G \Gamma_D G^\dagger] \tag{12}$$

The transmitted current at each in-plane energy (including spin degeneracy) is,

$$\begin{aligned} I\left[E(k_x,k_z)\right] &= \frac{q}{\hbar^2}\sqrt{\frac{m_y^* k_B T}{2\pi^3}} \\ &\times \left[F_{-1/2}\left(\mu_S - E(k_x,k_z)\right) - F_{-1/2}\left(\mu_D - E(k_x,k_z)\right)\right] T_{SD}\left[E(k_x,k_z)\right] \end{aligned} \tag{13}$$

The total current, like the 3D charge, is obtained by integrating over all $E(k_x,k_z)$ and summing over all conduction band valleys.

B. Mode-space solution

In this simulation scheme, the Green's function is solved for in a mode-space representation. These modes replace the $\delta(z-z')$ dependence of the basis, when compared to the real-space solution. This approach greatly reduces the size of the problem and provides

sufficient accuracy when compared to full 2D spatial discretization. We begin by solving a 1D, z directed, effective mass equation for each vertical device slice along x, to obtain a set of eigenenergies and eigenfunctions (modes) along the gate confinement direction. The equation that is solved is

$$-\frac{\hbar^2}{2m_z^*}\frac{\partial^2}{\partial^2 z}\Psi_i(x,z) - qV(x,z)\Psi_i(x,z) = E_i(x)\Psi_i(x,z) \tag{14}$$

where, m_z^* is the electron effective mass in the z direction, $\Psi_i(x,z)$, the wavefunction, and $E_i(x)$ the eigenenergy for subband i at slice x respectively. As with the real-space solution, the simulation domain in the confinement direction can be extended to include the insulator regions. Each vertical slice has a width, a, and within each slice, all quantities are assumed to be constant in the x direction.

The 3D Hamiltonian for the device is expanded in terms of $\delta(x-x')\Psi_i(x,z)$ and $e^{ik_j y/\sqrt{W}}$. The new basis functions, $\delta(x-x')\Psi_i(x,z)$ and $e^{ik_j y/\sqrt{W}}$ also constitute a complete and orthogonal expansion functions set. The Hamiltonian in this representation is

$$H = \begin{bmatrix} h[E_1(x)+E_{k_j}] & 0 & \cdots & \cdots & \cdots \\ 0 & h[E_2(x)+E_{k_j}] & 0 & \cdots & \cdots \\ 0 & \cdots & \ddots & \cdots & 0 \\ 0 & 0 & \cdots & h[E_i(x)+E_{k_j}] & 0 \\ 0 & 0 & 0 & 0 & \ddots \end{bmatrix} \tag{15}$$

where,

$$h[E_i(x)+E_{k_j}] = \begin{bmatrix} 2t_x+E_i(1)+E_{k_j} & -t_x & 0 & \cdots \\ -t_x & 2t_x+E_i(2)+E_{k_j} & \ddots & 0 \\ \cdots & 0 & \ddots & -t_x \\ \cdots & 0 & -t_x & 2t_x+E_i(N_X)+E_{k_j} \end{bmatrix} \tag{16}$$

is the Hamiltonian for subband i, with a planewave eigenenergy E_{k_j}. The subband index (i) in Eq. 16 runs over all subbands (replaces z in real-space calculations). Numbers 1 to N_X in parenthesis replace the position x because of the discretization.

The block diagonal nature of the Hamiltonian in Eq. 15 indicates that in the ballistic limit, subband coupling is neglected in the mode-space solution scheme (see Sec. IV). In the case of bulk MOSFETs, inversion layer electrons become unconfined as we move towards

the drain end of the device and strong electric fields near the drain couple different modes even in the ballistic limit. Therefore a mode-space solution, assuming decoupled modes is no longer applicable. Full real-space discretization provides the only accurate scheme to treat quantum ballistic transport in such devices.

Knowing the Hamiltonian for each subband, we write the retarded Green's function relevant to 1D transport as [4] [11],

$$G(E) = \left[EI - h\left[E_i(x), E_{k_j}\right] - \Sigma\right]^{-1} = \left[E_l I - h\left[E_i(x)\right] - \Sigma\right]^{-1} \tag{17}$$

where, the longitudinal (x) energy $E_l \equiv E - E_{k_j}$ (replaces $E[k_x, k_z]$) of the real-space solution). The self-energy for the leads (Σ) is a function of the longitudinal energy alone, and can derived based on the analysis in the Appendix as

$$\Sigma(E_l) = \begin{bmatrix} -t_x e^{ik_{l,1}a} & 0 & \cdots \\ 0 & \cdots & 0 \\ \cdots & 0 & -t_x e^{ik_{l,N_X}a} \end{bmatrix} \tag{18}$$

where, $E = E_{k_j} + E_i(n) + 2t_x(1 - \cos\mathrm{k_{l,n}a})$.The subband energy at the contact boundary (source or drain) is $E_i(n)$ and the subscript l represents the longitudinal dependence of k.

From a computational point of view, the size of the problem is measured by the size of the Hamiltonian. In a real-space representation the size of Hamiltonian is defined by the total nodal number in the 2D mesh, namely $(N_X \times N_Z)^2$; while in the mode-space representation, every subband is treated individually, and the size of the 1D Hamiltonian for each subband, is measured by the nodal number along the channel direction, namely $(N_X)^2$. In case of thin body, fully depleted SOI MOSFETs, strong body confinement causes the separation between modes to be large in energy. Therefore the Fermi level populates only a few modes even at high bias. Hence in practice, calculations, including the lowest few modes provide the desired accuracy. This, coupled with the reduced size of the mode-space Hamiltonian, implies that the mode-space approach provides enormous savings in computational burden without loss in accuracy as compared to a real-space solution which implicitly treats all modes (including mode coupling).

The mode-space spectral density functions due to the S/D contacts are analogous to those in Eq. 9. They differ from the real-space solution in that their diagonal entries represent the local density of states at site x, for mode i. The 2D electron density for mode i at a

longitudinal energy E_l is

$$n_i(E_l) = \frac{1}{a}\sqrt{\frac{m_y^* k_B T}{2\pi^3\hbar^2}}\left[F_{-1/2}(\mu_S - E_l)A_S + F_{-1/2}(\mu_D - E_l)A_D\right] \tag{19}$$

The net 2D electron density for each mode is obtained by integrating Eq. 19 over E_l. This 3D electron density at each lattice node of our 2D real-space grid is obtained by multiplying n_i with the corresponding distribution function $|\Psi_i(x,z)|^2/b$, and by summing over all subbands (index i) and conduction band valleys. Since the eigenvalue problem is solved exactly along the gate confinement direction, quantum effects associated with confinement and asymmetric gate design can be handled correctly within the mode-space modeling scheme. Once self-consistency is achieved, the terminal current is evaluated by summing contributions from each mode and conduction band valley.

III. RESULTS

The simulated device structure (Fig. 1a) is an ultra thin body, fully depleted, symmetric, n-MOSFET with S/D regions doped at $10^{20}\,cm^{-3}$ and an intrinsic channel. The gate length is 10 nm, and there is no gate-to-S/D overlap. The junctions are abrupt, and the oxide thickness for both top and bottom gates is 1.5 nm. A body thickness of 1.5 nm, and a power supply (V_{DD}) of 0.6 V has been used in this simulation study. The gate work function (4.25 V) has been adjusted to yield a threshold voltage (V_T) of 0.15 V. Gate oxides are treated as infinite potential barriers for electrons in all of the simulations.

In order to highlight the quantum effects that one observes in nanoscale transistors and validate the simplified mode-space approach, we compare internal quantities from the real and mode-space solutions. It should be noted that the real-space solution, implicitly includes all modes and their coupling effects (if any). In Fig. 2a, we plot results from the solution to a 1D effective mass equation (modes), along slice $Z - Z$ (refer Fig. 1a) in the on-state ($V_{GS} = V_{DS} = 0.6V$). The local density of states (LDOS) spectrum *vs*. in-plane energy ($E[k_x, k_z]$), obtained from the real-space solution is also plotted alongside in Fig. 2b. Light areas in Fig. 2b represent regions of high state density while dark areas signify low state density. Spatially, the LDOS goes to zero at the silicon/oxide interface (infinite potential barriers at $z = 0$ and 1.5 nm) and exhibits single or multiple maxima points. If we compare this plot, to the result obtained from the first step of the mode-space solution (refer Fig. 1a

and Fig. 2a), we find that the spatial behavior of the LDOS along $Z-Z$, is clearly captured by the mode-space solution. Note that, on superposing the mode energies onto the LDOS spectrum (dotted lines in Fig. 2b), we find that the maximum density of states occurs at in-plane energies that are higher that the corresponding subband energy (non-classical behavior). This observation can be explained by examining Fig. 3, where we plot the classical and quantum 2D density of states (DOS) as a function of the in-plane energy for slice $Z-Z$ (note that the subband energy depends on the location of the slice in x).

In the classical case (inset), the 2D DOS ($x-z$ plane) is a convolution of discrete delta functions (subbands) with a 1D DOS, that has a $1/\sqrt{E(k_x)}$ dependence. This is because, in the classical case there is no information about the quantum mechanical coupling of the device to the S/D reservoirs. Since, x is treated as a free dimension, the classical 2D DOS exhibits singularities around each subband energy. On the other hand, the real-space solution includes coupling information through the self-energy terms associated with the source and drain. These self-energies are composed of real and imaginary parts. The effect of the real part is to shift the maxima of the DOS in energy, while that of the imaginary part is to broaden the singularity in the classical DOS around each subband energy (dotted lines in Fig. 3) leading to a tail in the quantum DOS below each subband as shown in Fig. 3. This tail in the DOS is the cause of source-to- channel tunneling. Note that, all of the quantum effects in the channel direction, are a result of coupling the active device Hamiltonian to the S/D reservoirs through the self-energy terms.

In order to capture quantum effects along the channel within the mode-space solution, we couple each subband to the S/D reservoirs through specific self-energy terms calculated using Eq. 18. On introducing this coupling, we obtain the final form of the mode-space solution, whose LDOS is illustrated in Fig. 4. We plot the LDOS spectrum along slice $X-X$ (Fig. 1a). The conduction band (solid line) and the first subband profile (dotted line) along the channel, is superposed on this plot. With the inclusion of coupling information, this LDOS spectrum is identical to that obtained from the real-space solution for the energy range considered. The presence of a forbidden energy region between the conduction and first subband and the broadening of the LDOS around the first subband energy is clearly visible. It should be noted that the source-to-channel barrier is with respect to the subband profile as opposed to the conduction band edge and that tunneling occurs at energies much greater than the classical conduction band energy. States injected from the drain end of the device are reflected off

a large barrier under high drain bias and interfere strongly. States injected from the source that have energies slightly above or less than the source barrier also interfere, resulting in the observed quantum oscillations in the LDOS. At energies much greater than the subband maxima, injected states are free and there is no visible quantum interference effect. The oscillations in the LDOS give rise to local oscillations in the 3D electron density, as charge density is a convolution of the S/D injected LDOS and the corresponding Fermi function (Eq. 10 and Eq. 19).

Figure 5, compares the I_{DS} *vs.* V_{DS} and I_{DS} *vs.* V_{GS} characteristics for our model device from real and mode-space solutions. It is clear from Fig. 5 that the two solution schemes are in close agreement with each other, thus indicating that the mode-quantum solution, which is computationally inexpensive (order of magnitude less in computational time), is an attractive simulation tool for modeling thin body, SOI n-MOSFETs in the ballistic limit. It should be noted that although the 3D charge exhibits local oscillations, the current *vs.* voltage characteristics from both real and mode-space solutions are smooth. This is because current is a function of the transmission coefficient (Eq. 13), which depends on the overall potential profile from the source to the drain. Since local charge density oscillations are washed out when solving Poisson's equation for the potential, the potential profile and hence current is a smooth function of the applied voltage.

The on-current spectrum *vs.* in-plane/longitudinal energy from real and mode space simulations is plotted in Fig. 6. The maximum of the first subband is also indicated (dotted line) to separate the thermionic and tunneling current components. The spectrum indicates that conduction in this thin body MOSFET is essentially through the first mode (only one peak). Also, exact agreement between the real and mode-space simulation results, highlights the validity of the approximations inherent in the decoupled mode-space model (Sec IV). Figure 6, indicates that tunneling carriers constitute $\sim 25\%$ of the simulated on-current. Therefore it is expected that a 1D Boltzmann treatment of a mode [5] [6], which does not include tunneling effects would under predict the on-current. To verify if this is the case, we compare the self-consistent current-voltage characteristics, obtained from a quantum and a Boltzmann (classical) treatment of modes, in Fig. 7.

Figure 7 indicates that the simulated on-currents from the Boltzmann solution are remarkably close to, but slightly higher than those predicted by the quantum solution even though the tunneling component is missing in the classical solution. This close agreement is

the result of self-consistency and can be understood by examining the subband profile and charge density along the channel from the two solution schemes (Fig. 8). Note that in the on-state, the 2D charge density at the subband maximum is prescribed by gate electrostatics irrespective of whether the charge is due to tunneling or thermionic emission. Therefore, the subband maximum in case of the Boltzmann solution is lowered to obtain roughly the same charge as in the quantum case. Also note that all of this charge is thermionic in nature in case of the Boltzmann model, whereas it has both tunneling and thermionic components in case of the quantum model. Since tunneling carriers have a lower velocity due to their lower longitudinal energy, the quantum model predicts a lower on-current compared to its classical analogue in the on-state.

In the off-state, all of the current is due to tunneling. Therefore, the quantum model predicts a degraded subthreshold swing and higher off-current as compared to its classical counterpart. From a ballistic simulation viewpoint, it seems that in order to model the on-current accurately, a Boltzmann treatment of the mode along the channel direction is adequate. However, the main advantage of the mode-quantum solution is that the self-energy concept used to model the S/D contacts can, and has been extended to treat scattering [14]. Also, source-to-channel tunneling imposes a scaling limit on the channel length (for lengths below 10 nm). The Boltzmann solution cannot capture this tunneling limit. For a detailed comparison of the classical and quantum models in mode-space, when applied to thicker body transistors, refer to [15].

IV. DISCUSSION

The mode-space discretization of the Hamiltonian greatly reduces the size of the problem as compared to a full 2D spatial discretization (N_X^2 as opposed to $(N_X \times N_Z)^2$). It is important to look at the conditions under which this approach provides good simulation accuracy. In this section, we analytically expand the Schrödinger equation invoking the mode-space representation, and assess the approximations made in simplifying the Hamiltonian. This analysis explains why the decoupled mode-space approach provides the high degree of simulation accuracy in case of thin body DG and SG SOI MOSFETs.

We start with the 2D Schrödinger equation in the $x - z$ domain (the y dimension is

decoupled from the $x-z$ domain, and can therefore be treated separately)

$$-\frac{\hbar^2}{2m_x^*}\frac{\partial^2}{\partial^2 x}\Phi(x,z)-\frac{\hbar^2}{2m_z^*}\frac{\partial^2}{\partial^2 z}\Phi(x,z)-qV(x,z)\Phi(x,z)=[E-E_{k_j}]\Phi(x,z) \tag{20}$$

Left multiplying the mode-space eigenvectors and performing the integration in real-space we obtain

$$\begin{aligned}
&\int[\delta^*(x-x')\Psi_i^*(x,z)]\cdot\left[-\frac{\hbar^2}{2m_x^*}\frac{\partial^2}{\partial^2 x}\Phi(x,z)\right]dxdz\\
&+\int[\delta^*(x-x')\Psi_i^*(x,z)]\cdot\left[-\frac{\hbar^2}{2m_z^*}\frac{\partial^2}{\partial^2 z}-qV(x,z)\right]\Phi(x,z)dxdz\\
&=[E-E_{k_j}]\int[\delta^*(x-x')\Psi_i^*(x,z)]\cdot\Phi(x,z)dxdz
\end{aligned} \tag{21}$$

The third term in Eq. 21 becomes,

$$[E-E_{k_j}]\int[\delta^*(x-x')\Psi_i^*(x,z)]\cdot\Phi(x,z)dxdz=[E-E_{k_j}]\tilde{\Phi}_i(x') \tag{22}$$

Note that $\tilde{\Phi}_i(x')$, is the expansion coefficient of $\Phi(x',z)$, with respect to the mode-space eigenvector $\Psi_i(x',z)$ as defined by

$$\Phi(x',z)=\sum_{i=1}^{\infty}\tilde{\Phi}_i(x')\Psi_i(x',z)\quad\text{and}\quad\int\Psi_i^*(x',z)\Psi_j(x',z)dz=\delta_{ij} \tag{23}$$

where, δ_{ij} is the usual Kronecker delta. We can rewrite the second term in Eq. 21 as,

$$\int\Psi_i^*(x',z)\cdot\left[-\frac{\hbar^2}{2m_z^*}\frac{\partial^2}{\partial^2 z}-qV(x',z)\right]\Phi(x',z)dz=E_i(x')\tilde{\Phi}_i(x') \tag{24}$$

Finally the first term in Eq. 21 can be expressed as

$$\begin{aligned}
&\int\delta(x-x')\Psi_i^*(x,z)\left[-\frac{\hbar^2}{2m_x^*}\frac{\partial^2}{\partial^2 x}\Phi(x,z)\right]dxdz\\
&=\int\delta(x-x')\left[-\frac{\hbar^2}{2m_x^*}\frac{\partial^2}{\partial^2 x}\left\{\Psi_i^*(x,z)\Phi(x,z)\right\}\right]dxdz\\
&-\int\delta(x-x')\Phi(x,z)\left[-\frac{\hbar^2}{2m_x^*}\frac{\partial^2}{\partial^2 x}\Psi_i^*(x,z)\right]dxdz\\
&-2\int\delta(x-x')\left[-\frac{\hbar^2}{2m_x^*}\frac{\partial\Psi_i^*(x,z)}{\partial x}\frac{\partial\Phi(x,z)}{\partial x}\right]dxdz
\end{aligned} \tag{25}$$

Note that the second term in Eq. 25 reduces to $-\hbar^2/2m_x^*\partial^2/\partial^2 x'\tilde{\Phi}_i(x')$ after integration. If we assume that for all x (the shape of a mode does not change along the channel)

$$\frac{\partial\Psi_i^*(x,z)}{\partial x}=0, \tag{26}$$

Equation (20) becomes,

$$-\frac{\hbar^2}{2m_x^*}\frac{\partial^2\tilde{\Phi}_i(x')}{\partial^2 x'} + E_i(x')\tilde{\Phi}_i(x') = [E - E_{k_j}]\tilde{\Phi}_i(x'). \tag{27}$$

Equation 27 is the decoupled mode-space transformation of the 2D Hamiltonian invoking the assumption represented by Eq. 26. Equation 27 is indeed a 1D differential equation and greatly reduces the size of the original 2D problem. Note that Eq. 27 has two implications. The first is that subbands with different energies do not couple and the second is that some coupling information of a subband with itself is also lost.

Although the potential profile varies from the source to drain, if $V(x, z)$ retains the same shape in the z direction, at different locations along the channel, the eigenfunctions are the same at each x location, even though the eigenvalues are different. As a result, Eq. 27 is satisfied. In the case of SOI MOSFETs with uniform thin bodies, there is little room for the potential to vary vertically. Therefore, Eq. 27 is a valid approximation and results obtained from the decoupled mode-space solution are in close agreement with real-space simulation results. This observed agreement between the real and decoupled mode-space solutions holds true in the case of thicker bodies (upto 5 nm) as long as the device has a uniform SOI geometry.

If we perturb the uniformity of the potential profile by squeezing the channel region of our model device (Fig. 1b) and compare the real and mode-quantum solutions for a fixed potential profile, we see that the I_{DS} *vs.* V_{DS} characteristics of the real-space approach no longer agrees with that of the mode-space approach as indicated in Fig. 9. We know that in our model device transport is through the first subband. Therefore the mismatch between the real and mode-quantum solutions is because the mode-space solution does not completely capture the effect of a subband coupling with itself although a part of this information is built into the mode-space Hamiltonian as seen from its tridiagonal nature (Eq. 2). The differences in current at high V_{DS} could be as high as $\sim 15\%$. Note that the current from the real-space solution is lower. It has been pointed out that including the flared out portions of the S/D contacts would have a similar effect and this reduction in current can be thought of as arising due to a quantum spreading resistance that is not captured by the mode-space solution [16]. In the case of bulk transistors, channel depletion widths can vary considerably from the source to drain, resulting in significant changes in the vertical confinement potential profiles. Therefore the mode-space approach becomes inappropriate

for bulk device simulations or simulation of devices composed of heterostructures. A coupled mode-space representation, which includes all of the terms in Eq. 25, would be appropriate for such structures [17].

V. SUMMARY

We presented two approaches based on the NEGF formalism (real and decoupled mode-space), each with a different degree of complexity, that can be used to simulate 2D MOSFET structures under non-equilibrium conditions. These approaches were compared and contrasted by using them to simulate an ultra small DG n-MOSFET (with a body thickness of 1.5 nm). In doing so, quantum effects that are observed in nanoscale transistors, the treatment of open boundaries and the importance of self-consistency were highlighted. We showed that the real-space solution, which is the most general, is computationally expensive due to the 2D nature of the Hamiltonian; while the decoupled mode-space solution, which is specifically applicable to thin body, fully depleted, SOI device geometries is inexpensive (yet accurate) for two reasons: 1) A 1D Hamiltonian is used in the mode-space solution and 2) only few modes need to be considered as modes with high energies are not occupied by electrons and do not contribute to transport. For the model 1.5 nm body DG MOSFET, the simulation time per bias point on a one processor SUN workstation (300 MHz) was 40 secs in case of the mode-space solution as opposed to 1.5 hrs for the real-space solution. Finally, the validity of the mode-space solution and its regime of applicability were discussed by simulating a device with a squeezed channel region. These simulations indicated that including the flared out regions of the S/D contacts or simulating heterostructure devices at an effective mass level, would require a coupled mode-space simulation scheme.

VI. ACKNOWLEDEGEMENTS

The authors would like to thank the Semiconductor Research Corporation for funding this project. Useful discussions with M. Anantram from NASA and P. Damle from Purdue University are appreciated.

APPENDIX A: THE SELF-ENERGY CALCULATION FOR THE LEADS

To illustrate the self-energy calculation which accounts for the device leads, we consider the effect of coupling the active device Hamiltonian to the drain. The infinite Hamiltonian (Eq. 2) and its Green's function (Eq. 6) can be partitioned as follows,

$$\begin{bmatrix} G_{device} & G_{device,D} \\ G_{D,device} & G_D \end{bmatrix} = \begin{bmatrix} E(k_x,k_z)I - h(x,z)_{device} & -\beta \\ -\beta & E(k_x,k_z)I - h(x,z)_D \end{bmatrix} \tag{A1}$$

In Eq. A1, subscript "D" is used to indicate the infinite block of $h(x,z)$ and G, representing the drain. The matrix block we are interested in is G_{device} as we do not care about the Green's function within the drain. Using Eq. A1, G_{device} can be expressed in terms of known quantities as,

$$G_{device}\left[E(k_x,k_z)\right] = \left[E(k_x,k_z)I - h(x,z)_{device} - \Sigma_D\right]^{-1} \tag{A2}$$

where the drain self-energy matrix is,

$$\Sigma_D = \begin{bmatrix} 0 & 0 & \cdots \\ 0 & 0 & \cdots \\ -\beta & 0 & \cdots \end{bmatrix} \left[\begin{array}{c|cc} E(k_x,k_z)I - \alpha_{N_X+1} & -\beta & 0 \quad \cdots \\ \hline -\beta & E(k_x,k_z)I - \alpha_{N_X+2} & -\beta \quad 0 \\ 0 & -\beta & \ddots \quad \ddots \\ \cdots & 0 & \ddots \quad \ddots \end{array}\right]^{-1} \times \begin{bmatrix} \cdots & 0 & -\beta \\ \cdots & 0 & 0 \\ \cdots & \cdots & \cdots \end{bmatrix} \tag{A3}$$

Note that for evaluating the matrix product in Eq. A3, we only need the first block of the inverse of the infinite matrix associated with the drain. Also, note that the diagonal blocks of this infinite matrix are repeated due to translational invariance within the drain ($\alpha_{N_X} = \alpha_{N_X+1} = \cdots$). Using this property, and partitioning the matrix as shown in Eq. A3, a closed form expression for the first block of the inverse (denoted by g_D) of the infinite matrix, can be obtained as,

$$I = g_D\left[E(k_x,k_z)I - \alpha_{N_X+1} - \beta g_D \beta\right] \tag{A4}$$

Once g_D, has been solved for, we have,

$$\Sigma_D = \begin{bmatrix} 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ 0 & \cdots & \beta g_D \beta \end{bmatrix} \tag{A5}$$

Note that, only the last vertical slice of the device couples to the drain. Therefore the self-energy for the drain (Eq. A5) has a single non-zero block that perturbs the last diagonal block of the device Hamiltonian. To solve Eq. A4, a basis transformation has to be performed. The eigen vectors of $E(k_x, k_z)I - \alpha$, diagonalize g_D simultaneously. Therefore we change basis from 2D real-space to a basis that is composed of the eigenvectors of $E(k_x, k_z)I - \alpha$ (equivalent to a mode-space transformation at the boundary). This reduces Eq. A4 to a set of decoupled quadratic equations that can be solved for the diagonal entries g_D, in the transformed representation. It should be noted that each of these equations results in two roots. The root representing outgoing waves is selected as we are ultimately interested in obtaining the retarded Green's function for the device. An inverse basis transformation is then applied to evaluate g_D in 2D real-space. A similar procedure is invoked to solve for the self-energy part associated with the source. The final size of the self-energy matrix is $(N_X \times N_Z)^2$ for the real-space solution and $(N_X)^2$ for the decoupled mode-space solution.

REFERENCES

[1] Y. Taur and T. Ning, *Fundamentals of VLSI Devices* (Cambridge University Press Cambridge, UK, 1998).

[2] H. Wong, D. Frank, and P. Solomon, in *IEDM Tech. Digest* (1998), pp. 407–410.

[3] S. Datta, J. Phys. Condens. Matter **2**, 8023 (1990).

[4] S. Datta, Superlattices and Microstructures **28**, 253 (2000).

[5] Y. Naveh and K. Likharev, IEEE Electron Dev. Lett. **21**, 242 (2000).

[6] J. Rhew, Z. Ren, and M.S. Lundstrom, To appear in Solid State Electronics, *A Numerical Study of Ballistic Transport in a Nanoscale MOSFET* (2002).

[7] www.intel.co/labs (2001).

[8] D. Jovanovic and R. Venugopal, in *Presented at the 7th International Workshop on Computational Electronics* (University of Glasgow, UK, 2000).

[9] A. Svizhenko, M. Anantram, and T. Govindan, in *Presented at the 7th International Workshop on Computational Electronics* (University of Glasgow, UK, 2000).

[10] R. K. Lake and S. Datta, Phys. Rev. B **45**, 6670 (1992).

[11] S. Datta, *Electronic Transport in Mesoscopic Systems* (Cambridge University Press,UK, 1997).

[12] A. Svizhenko, M. Anantram, T. Govindan, B. Biegel, and R. Venugopal, J. Appl. Phys. **91**, 2343 (2002).

[13] P. V. Halen and D. Pulfrey, J. Appl. Phys. **57**, 5271 (1985).

[14] Z. Ren, R. Venugopal, S. Datta, and M. Lundstrom, in *IEDM Tech. Digest* (2001), pp. 107–110.

[15] Z. Ren, R. Venugopal, S. Datta, M. Lundstrom, D. Jovanovic, and J. Fossum, in *IEDM Tech. Digest* (2000), pp. 715–718.

[16] M. Anantram , verbal communication (2001).

[17] P. Damle and S. Datta, unpublished (2001).

FIGURE CAPTIONS

Fig.1: (a) An ultra-thin body DG MOSFET structure with S/D doping of $10^{20}\ cm^{-3}$ and an intrinsic channel (channel thickness = 1.5 nm). (b) The squeezed DG MOSFET structure used to investigate the effect of mode coupling (channel thickness = 0.75 nm).

Fig.2: (a) The solution to a 1D effective mass equation (step 1 of the mode-space solution) along slice $Z - Z$ (refer Fig. 1a) in the on-state ($V_{GS} = V_{DS} = 0.6V$). (b) The local density of states from the real-space solution along slice $Z - Z$.

Fig.3: The classical (inset) and quantum 2D density of states (DOS) along slice $Z - Z$ (refer Fig. 1) in the on-state. The position of the slice along the channel determines the subband energies (dotted lines).

Fig.4: Computed local density of states (LDOS) and the longitudinal subband profile (dotted line) along slice $X - X$ in the on-state. Broadening in the LDOS is due to quantum mechanical coupling to the S/D and oscillations are due to quantum mechanical reflections. The conduction band (solid line) is far below the subband due to confinement.

Fig.5: I_{DS} *vs.* V_{GS} and I_{DS} *vs.* V_{DS} characteristics for the model device (refer Fig. 1a) from real (line) and mode-space (circles) solution schemes. Close agreement ($\sim 1\%$) between the two solutions validates the approximations inherent in the decoupled mode solution.

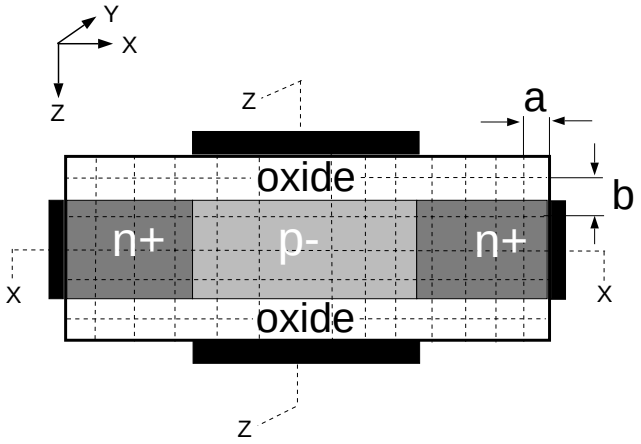
Fig.6: The energy distribution of the on-current from the real (solid line) and mode-space (circles) solution schemes. The top of the first subband (dotted line) is also indicated to separate the thermionic and tunneling current components.

Fig.7: I_{DS} *vs.* V_{GS} and I_{DS} *vs.* V_{DS} characteristics for the model device (refer Fig. 1a) from a quantum (line) and Boltzmann (circles) treatment of the modes. The off-current from the quantum solution is higher than the Boltzmann solution due to source-to-channel tunneling. The on-current is lower as a result of self-consistency.

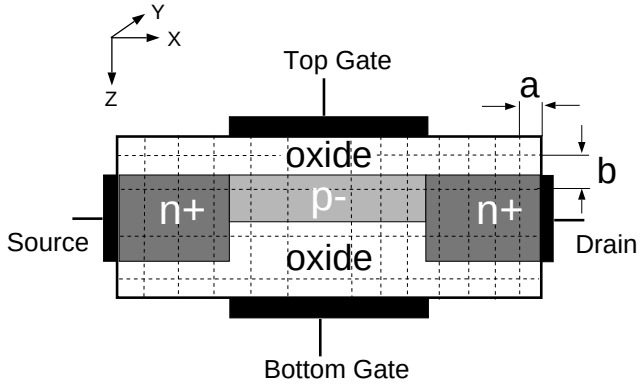
Fig.8: The 2D electron density and first subband profile, along the channel from a quantum (solid line) and classical (dashed line) treatment of the modes, in the on-state. Charge at the top of the source-barrier is primarily prescribed by gate electrostatics. Crossover between the quantum and classical charge profiles, is due to quantum reflections and tunneling.

Fig.9: I_{DS} *vs.* V_{GS} and I_{DS} *vs.* V_{DS} characteristics for the device with the squeezed channel (refer Fig. 1b) from mode-space (dashed line) and real-space (solid line) solution schemes. The mode-space solution does not account for mode coupling. Therefore, when the

vertical potential profile is perturbed strongly, the decoupled mode-space solution exhibits inaccuracies.



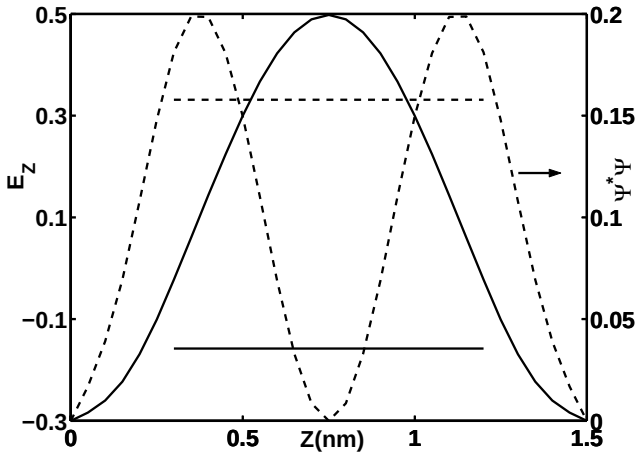
(a)



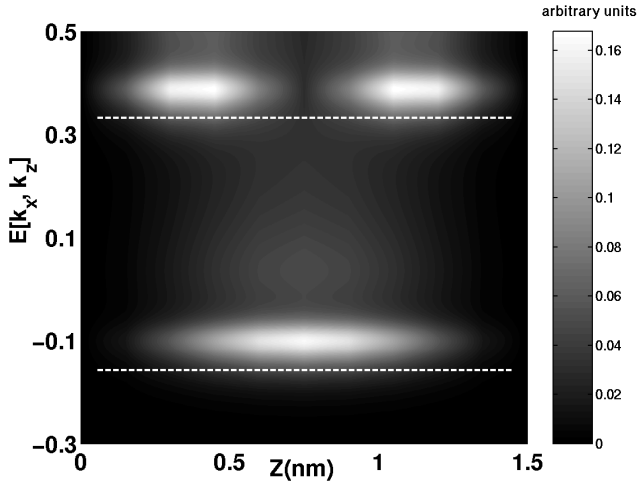
(b)

FIG. 1:

R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom and D. Jovanovic



(a)



(b)

FIG. 2:

R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom and D. Jovanovic

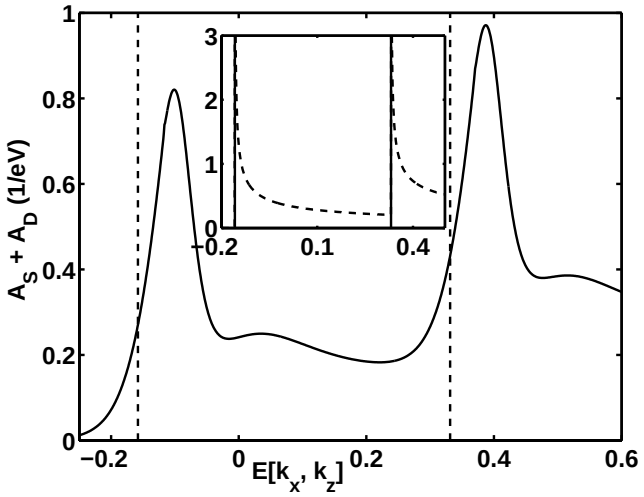


FIG. 3:

R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom and D. Jovanovic

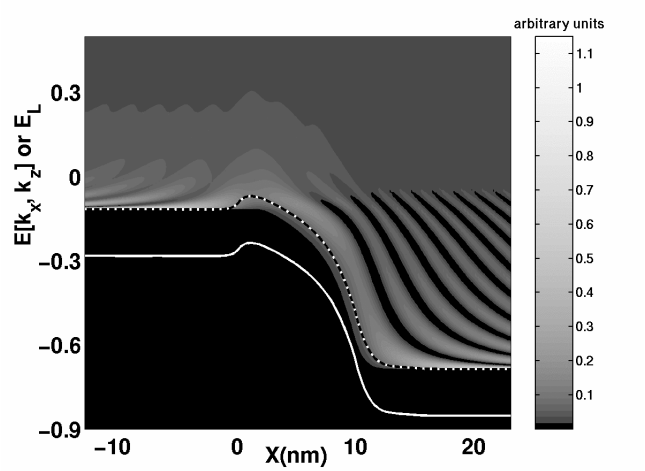


FIG. 4:

R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom and D. Jovanovic

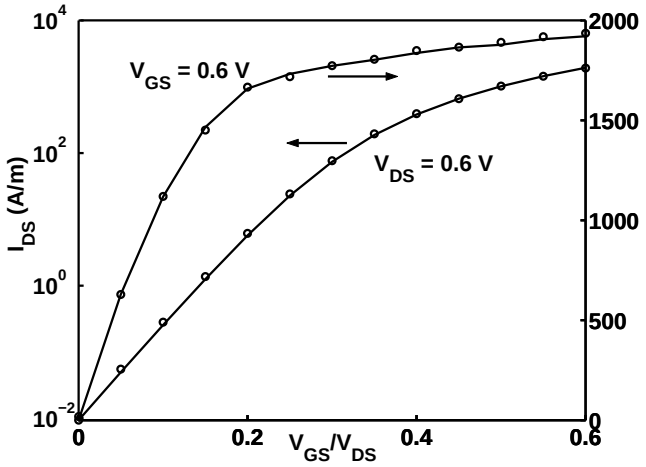


FIG. 5:

R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom and D. Jovanovic

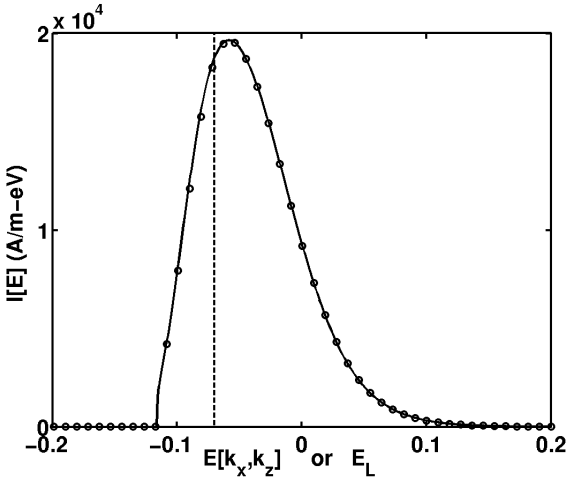


FIG. 6:

R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom and D. Jovanovic

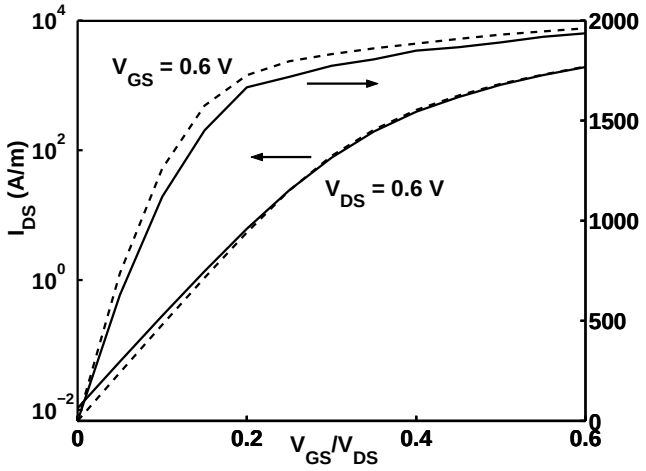


FIG. 7:

R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom and D. Jovanovic

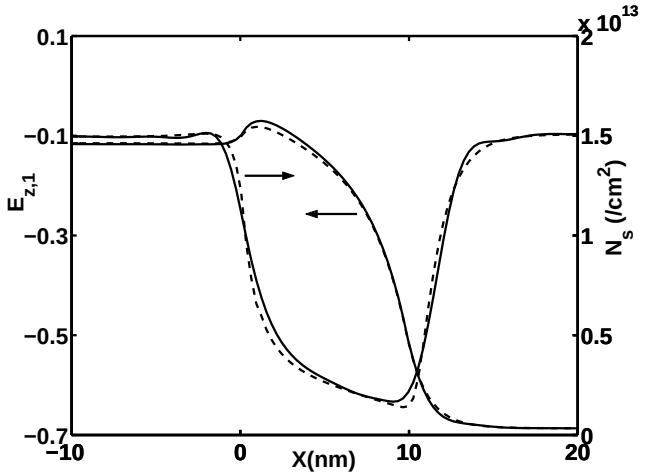


FIG. 8:

R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom and D. Jovanovic

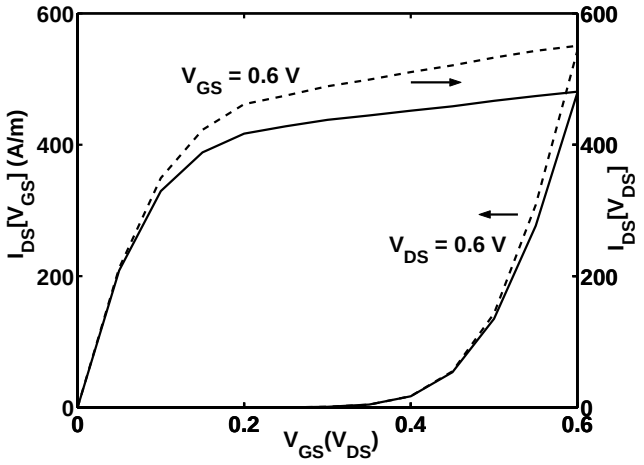


FIG. 9:

R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom and D. Jovanovic