Atomistic simulation of band-to-band tunneling in III-V nanowire field-effect transistors

Dipanjan Basu*[1], Leonard F. Register[1], Matthew J. Gilbert[2] and Sanjay K. Banerjee[1].
1. Microelectronics Research Center, The University of Texas at Austin
10100 Burnet Road, Bldg. 160, Austin, TX 78758, U.S.A. *Correspondence: dipanjan@mail.utexas.edu
2. 2256 Micro and Nanotechnology Lab, University of Illinois at Urbana-Champaign
208 N. Wright Street, Urbana, Illinois 61801

***Abstract*— Efficient atomistic simulators are required for full band treatments in strongly quantum confined systems, and for simulation of transport in emerging materials and devices such as graphene. Here we present an efficient transmission matrix based approach to ballistic quantum transport calculation for full three-dimensional, nearest-neighbor tight-binding based atomistic simulations. The method is then used to demonstrate how band-to-band tunneling increases the leakage current in OFF state in field-effect transistors with low band gap semiconductors such as InSb as channel material.**

Keywords-component; atomistic, quantum transport

I. MOTIVATION

Phase-coherent ballistic transport effects manifest in many semiconductors as the channel lengths scale to few tens of nm [1], necessitating full quantum transport simulation. Field-effect transistors (FETs) using nanowires as the semiconducting channel have high confinement in the channel region in not only the vertical, but also the lateral direction. Furthermore, the source (S) and drain (D) regions may be comparatively large compared to channel cross-section and defects such as surface roughness or charge impurities may be present within the channel. As a result, nominally discrete propagating modes can mix and a quasi one-dimensional (1D) model of quantum transport with discrete transverse modes within the channel cross-section can be insufficient. Fully three-dimensional (3D) quantum transport simulation becomes necessary.

In addition, full band structure treatments are becoming increasingly important. Full-band treatments are necessary to model hole and strong electron quantum confinement. Furthermore, several III-V materials such as InAs, InSb, InGaAs and group IV material like Ge are possible candidates for replacing Si channel [2]. InSb with its very high electron mobility is particularly suitable for high-speed logic [3-4]. The attendant low band-gap, however, makes it difficult to shut off the InSb based transistors, necessitating more complex architectures like quantum well FETs [5-6]. Simple effective-mass based quantum transport simulation also fail to address the band-to-band tunneling in InSb nanowire transistors under high OFF-state gate bias, as well as the proper dependence of the bandgap on the degree of confinement.

For such reasons, efficient full-band atomistic simulators are increasingly in demand. In this work, we adapt an efficient transmission matrix based approach to quantum transport calculation previously implemented within an effective mass approximation [7] for full 3D, atomistic, nearest-neighbor tight-binding (TB) based simulations. These transport calculations represent an alternative numerical implementation for such systems of the widely accepted non-equilibrium Green's function [NEGF] approach [8]. We will detail our methodology in Section II. In Section III, we report on band-to-band tunneling (BTBT) current in strongly-confined nanowire FETs with InSb channels. In other works, we have adapted this same approach to model transport in single and multi-layer graphene-based devices [9].

II. METHOD

A. Crystal structure and basis set

In this approach, each atom in the device is incorporated explicitly in the Hamiltonian. The atoms are represented by orthogonal (Löwdin) orbitals and the on-site energies and hopping potentials are obtained from empirical TB theory [10]. The atomistic Hamiltonian changes for different materials, orbital bases and crystal orientations; however, the general method remains the same. For an optimum balance of accuracy and computational cost in the simulations of this work, we have restricted ourselves to the sp^3s^* basis with nearest neighbor interactions, as the energy bands are well modeled close to the Γ point [10]. Also, we have used the original empirical TB parameters given in [10] instead of the parameters optimized to obtain best fit of just the low lying conduction and valence bands [11]. The focus of this work is to demonstrate the capability of our atomistic simulator in a qualitative rather than quantitative fashion. And we concentrate here on [100] transport in III-V semiconductors that exhibit a zinc-blende crystal structure.

For [100] transport in a square III-V nanowire, we can visualize the nanowire as successive face-centered cubic (fcc) planes that are displaced from the previous plane along the body diagonal by (a/4, ±a/4, ±a/4) where a is the material lattice constant. The atomistic view of eight successive fcc planes along the transport direction is shown in Fig. 1. The atomic structure repeats itself every fourth layer along the transport direction, and therefore, these four layers constitute the primitive cell.

B. Band structure calculation for the leads

The time-independent Schrödinger equation for a tight-binding Hamiltonian can be written in a layer-to-layer coupled form as:

$$\mathbf{H}_{l,l-1}\boldsymbol{\psi}_{l-1} + \mathbf{H}_{l,l}\boldsymbol{\psi}_{l} + \mathbf{H}_{l,l+1}\boldsymbol{\psi}_{l+1} = E\boldsymbol{\psi}_{l}\,. \qquad (1)$$

Here $\boldsymbol{\psi}_l$ is the wavefunction (column matrix) where each row corresponds to a particular orbital of a particular atom in the l-th layer, $\mathbf{H}_{l,l\pm1}$ is the coupling (square matrix) from layer l to layer $l\pm1$, E is the eigenenergy and the $\mathbf{H}_{l,l}$ are populated with on-site energies including applied potential energy contributions, and transfer matrix elements between atoms in same layer, and $\mathbf{H}_{l,l\pm1}$ are populated with transfer matrix elements between atoms in neighboring layers. The entire device from source (S) to drain (D) can be broken up into N layers perpendicular to the transport direction. Fig. 2 shows these layers schematically. Within the leads the applied potential energy contributions are determined by the source and drain biases and the injected charge densities; within the simulation region, by self-consistent electrostatics.

The S/D contacts/leads can be assumed to be semi-infinite wires that are in thermal equilibrium with the applied bias, and are therefore equipotential. Denoting the four layers of the fcc lattice repeat unit from left to right as –2 to 1 (see Fig. 2) and assuming constant potential, which is adjusted consistent with source and drain biases and charge densities, (1) can be written explicitly for the left lead as:

$$\begin{aligned}
&\mathbf{H}_{-4,-3}\boldsymbol{\psi}_{-3}+\left(\mathbf{H}_{-2,-2}-E\right)\boldsymbol{\psi}_{-2}+\mathbf{H}_{-2,-1}\boldsymbol{\psi}_{-1}=0\\
&\mathbf{H}_{-3,-2}\boldsymbol{\psi}_{-2}+\left(\mathbf{H}_{-1,-1}-E\right)\boldsymbol{\psi}_{-1}+\mathbf{H}_{-1,0}\boldsymbol{\psi}_{0}=0\\
&\mathbf{H}_{-2,-1}\boldsymbol{\psi}_{-1}+\left(\mathbf{H}_{0,0}-E\right)\boldsymbol{\psi}_{0}+\mathbf{H}_{0,1}\boldsymbol{\psi}_{1}=0\\
&\mathbf{H}_{-1,0}\boldsymbol{\psi}_{0}+\left(\mathbf{H}_{1,1}-E\right)\boldsymbol{\psi}_{1}+\mathbf{H}_{1,2}\boldsymbol{\psi}_{2}=0
\end{aligned} \quad . \qquad (2)$$

where for Bloch states and also evanescent states within the leads we may write $\lambda\boldsymbol{\psi}_{-3}=\boldsymbol{\psi}_1$, $\boldsymbol{\psi}_2=\lambda\boldsymbol{\psi}_{-2}$, with Bloch factor $\lambda=\exp\left(\mp ik_x a\right)$ for the set of four layers each space by $a/4$ from the previous.

After some algebraic manipulation we can reduce the eigensystem to a basis set consisting of $\boldsymbol{\psi}_2$ and $\boldsymbol{\psi}_3$ as follows:

$$\begin{bmatrix} -\mathbf{H}_{0,1}^{-1}\mathbf{H}_{0,-1} & -\mathbf{H}_{0,1}^{-1}\left(\mathbf{H}_{0,0}-E\right) \\ -\mathbf{H}_{1,2}^{-1}\left(\mathbf{H}_{1,1}-E\right)\mathbf{H}_{0,1}^{-1}\mathbf{H}_{0,-1} & \mathbf{H}_{1,2}^{-1}\mathbf{H}_{1,0}-\mathbf{H}_{1,2}^{-1}\left(\mathbf{H}_{1,1}-E\right)\mathbf{H}_{0,1}^{-1}\left(\mathbf{H}_{0,0}-E\right) \end{bmatrix}\begin{pmatrix}\boldsymbol{\psi}_{-1}\\ \boldsymbol{\psi}_0\end{pmatrix} = \qquad (3)$$

$$\lambda\begin{bmatrix} \mathbf{H}_{-2,-3}^{-1}\left(\mathbf{H}_{-2,-2}-E\right)\mathbf{H}_{-1,-2}^{-1}\left(\mathbf{H}_{-1,-1}-E\right)-\mathbf{H}_{-2,-3}^{-1}\mathbf{H}_{-2,-1} & \mathbf{H}_{-2,-3}^{-1}\left(\mathbf{H}_{-2,-2}-E\right)\mathbf{H}_{-1,-2}^{-1}\mathbf{H}_{-1,0} \\ -\mathbf{H}_{-1,-2}^{-1}\left(\mathbf{H}_{-1,-1}-E\right) & \mathbf{H}_{-1,-2}^{-1}\mathbf{H}_{-1,0} \end{bmatrix}\begin{pmatrix}\boldsymbol{\psi}_{-1}\\ \boldsymbol{\psi}_0\end{pmatrix}$$

Equation (3) is a generalized eigenvalue system whose dimension depends on the size of the system modeled and the tight-binding basis set used. We solve (3) using standard commercially available math libraries such as IMSL [12].

The eigenfunctions of (3) can be identified as propagating modes if $|\lambda|=1$. A plot of the corresponding real values of k_x vs. the eigenenergy E provides the real band structure for the quantum wire leads. A similar procedure also allows calculation of the imaginary band structure for the evanescent states $|\lambda|\neq1$ which is relevant to understanding tunneling.

The probability current flow for any wavefunction $\boldsymbol{\psi}_l$ j is:

$$j=\frac{2}{\hbar}\mathrm{Im}\left[\boldsymbol{\psi}_l^{\dagger}\mathbf{H}_{l,l+1}\boldsymbol{\psi}_{l+1}\right]. \qquad (4)$$

The probability current carried per mode per unit energy by

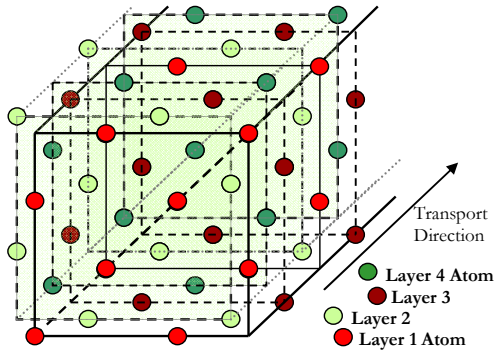

Figure 1. Eight fcc lattice planes, stacked along (100) direction, to form a square nanowire 2 atoms wide, showing the individual atomic locations in the plane. For zinc blende structures such as InSb, the layers are alternately anions and cations. For a diamond crystal structure such as Si or Ge, all the atoms are identical.

these occupied Bloch states should always be precisely equal to $2/h$ counting both spin states, where h is the Planck's constant. This relation is used to normalize the amplitude of the incident wavefunctions. The solutions to (3) with $|\lambda|=1$ and $j>0$ (although not necessarily $k_x>0$) for the left lead represent the right-going incident propagating waves.

C. *Transport calculation*

The eigenfunctions of the leads form a complete basis that is used for injecting probability density into and extracting probability density from the device region. For the two central layers in the left lead, and for each energy, the eigenfunctions can be arranged into a basis matrix $\mathbf{B}_{\mathrm{L}}$, as [7]:

$$\mathbf{B}_{\mathrm{L}}=\begin{bmatrix}\boldsymbol{\psi}_{-1}(\rightarrow) & \boldsymbol{\psi}_{-1}(\leftarrow)\\ \boldsymbol{\psi}_{0}(\rightarrow) & \boldsymbol{\psi}_{0}(\leftarrow)\end{bmatrix}, \qquad (5)$$

Here $\boldsymbol{\psi}_0$ is now a square matrix where each row represents a particular orbital of a particular atom as before, and each column represents a particular transverse mode in the left lead (source side). And $\rightarrow$ ($\leftarrow$) indicate the basis functions that either propagate or decay towards right (left). Similarly the basis function $\mathbf{B}_{\mathrm{R}}$ can be constructed for the two central layers, N+1 and N+2, in the right lead (drain side).

Within the simulation region, $\mathbf{B_0}$ can then be propagated to the right one layer at a time via transfer matrices $\mathbf{T}_l$

$$\mathbf{T}_l=\begin{bmatrix}\mathbf{0} & \mathbf{1}\\ -\mathbf{H}_{l,l+1}^{-1}\mathbf{H}_{l,l-1} & -\mathbf{H}_{l,l+1}^{-1}\left(E-\mathbf{H}_{l,l}\right)\end{bmatrix} \qquad (6)$$

such that

$$\begin{bmatrix}\boldsymbol{\psi}_l(\rightarrow) & \boldsymbol{\psi}_l(\leftarrow)\\ \boldsymbol{\psi}_{l+1}(\rightarrow) & \boldsymbol{\psi}_{l+1}(\leftarrow)\end{bmatrix}=\mathbf{T}_l\begin{bmatrix}\boldsymbol{\psi}_{l-1}(\rightarrow) & \boldsymbol{\psi}_{l-1}(\leftarrow)\\ \boldsymbol{\psi}_l(\rightarrow) & \boldsymbol{\psi}_l(\leftarrow)\end{bmatrix}. \qquad (7)$$

Here the Hamiltonians $\mathbf{H}$ are adjusted for changing coupling potentials, and onsite potentials including the self-consistent changes to the electrostatic potential.

In principle at least, by cascading the transfer matrices and imposing appropriate boundary conditions, the complex transmission and reflection matrices $\mathbf{t}$ and $\mathbf{r}$ coupling each incoming mode to each outgoing mode, — one column and

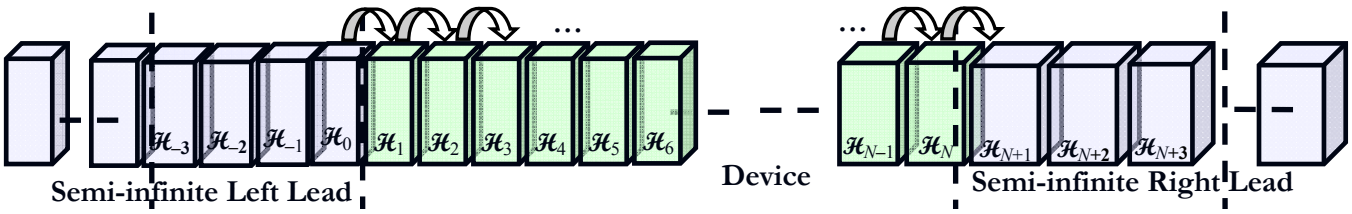

Figure 2. Schematic of the transport calculation showing the break-up of the simulation region into a device region that includes source, channel and drain regions (green shade), and semi-infinite source and drain contacts at the two ends. The layer to layer interaction matrix is shown for only one interaction in the schematic ($\mathbf{H}_{12}$). In practice, there are a large number of slices in the device region, depending on the length of the device simulated.

one row for each mode in the respective leads — for the entire device can then be calculated:

$$\begin{bmatrix}\mathbf{t}\\ \mathbf{0}\end{bmatrix} = \mathbf{B}_R^{-1}\mathbf{T}_{N+1}\mathbf{T}_N \ldots \mathbf{T}_2\mathbf{T}_1\mathbf{B}_L\begin{bmatrix}\mathbf{1}\\ \mathbf{r}\end{bmatrix}. \quad (8)$$

Propagation from the drain can be handled similarly.

In practice however (8) is extremely unstable. We have followed the stabilization method developed by Usuki *et al.* [7] to solve for the **t** and **r** matrices, but the wave-functions are obtained using a simpler, equivalent method given in [13].

Carrier density n is obtained from the probability density associated with the wave functions injected from source (left), $n(\rightarrow)$, as well as drain (right), $n(\leftarrow)$. $n(\rightarrow)$ is calculated for each atom α in the device by summing over the probability density in each orbital β in each incident propagating mode χ:

$$n_\alpha(\rightarrow) = \int dE \sum_\beta \sum_\chi \left|\psi_{\alpha,\beta,\chi}(E)\right|^2 f_s(E)\,. \quad (9)$$

Then n is fed into Poisson's equation to solve for the electrostatic potential which, in turn, is self-consistently fed back into (1).

Once a self-consistent solution is obtained, the total charge current due to injection from the source side (I) can be calculated from the normalized wavefunctions $\boldsymbol{\psi}_{l,\chi}(E)$ for arbitrary temperature and bias from

$$I(\rightarrow) = q\sum_\chi \int dE \frac{2}{\hbar}\mathrm{Im}\left[\boldsymbol{\psi}^\dagger_{l,\chi}(E)\mathbf{H}_{l,l+1}\boldsymbol{\psi}_{l+1,\chi}(E)\right] f_S(E) \quad (10)$$

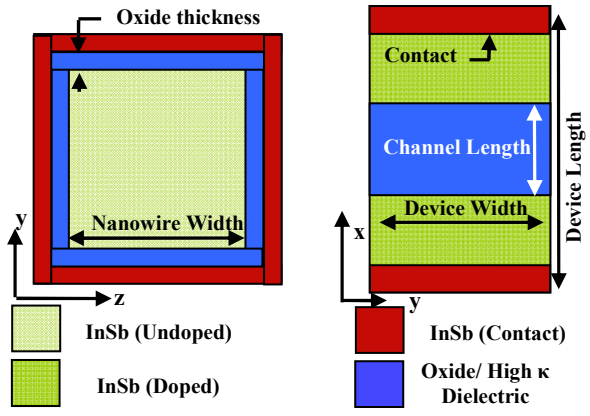

Figure 3. Schematic of the device simulated: (left) the cross-section of the channel region, perpendicular to the transport direction and (right) the view from the top showing the source and drain contact at the ends, and the gate oxide in the middle, surrounding the channel.

at any point within the simulation region, where q is the electronic charge, and χ labels the mode and l the layer. Here f_S is the Fermi function at the source contact, and the integration is performed over the range of applied bias plus or minus a few k_BT/q to account for non-zero temperature effects, where k_B is Boltzmann's constant and T is the temperature.

The method of calculating transmission probabilities that we follow [7] is similar to Ando's formalism [14], and Khomyakov *et al.* [15] have discussed the equivalence of this approach with that of the Green's function techniques [8, 16]. Gilbert and Ferry adapted this method to calculate 3D quantum transport in Si-on-insulator metal-oxide-semiconductor FET devices within an effective mass approach [17]. We note that the numerical approach used here is, beyond being an atomistic tight-binding approach, quite different than that used in [18].

III. Results

A. Band Structure of [100] square InSb nanowire

The schematic of a gate-all-around InSb nanowire FET is shown in Fig. 3. For computational ease, we have restricted the nanowire widths to 26.92, 32.40 and 38.88 Å, corresponding to 4, 5 and 6 atoms along the width, respectively. The device has an alumina (Al_2O_3) high-κ gate dielectric, with an effective oxide thickness of 6.48 Å, where the high-κ dielectric material is modeled pseudo-atomistically by taking appropriate empirical obtained TB parameters to produce a bandgap corresponding to Al_2O_3, with band offset to InSb as reported in [19] with a type I interface.

The band-structure of an ultrascaled square (100) InSb nanowire of width 26.92 Å is shown in Fig. 4. The conduction and valence band edges shift to -0.28 eV and 0.78 eV, respectively, from the bulk values being 0 and 0.23 eV for the set of TB parameters used [5] illustrating the considerable increase of band gap induced by very strong confinement.

B. I_D-V_G characteristics of a [100] square InSb nanowire

In Fig. 5, the self-consistent I_D-V_G characteristics are shown for ballistic transport in square (100) InSb nanowires of widths 26.92 and 32.40 Å. While these extremely narrow cross-sections (corresponding to 4 and 5 atoms wide) may not be realized in practice, the simulations serve to illustrate the band-to-band tunneling (BTBT) that ails any low band gap semiconductor.

For large negative V_G, the valence band is pulled up under the gate, allowing electrons injected from the conduction band of source to tunnel into the valence band in the channel, giving

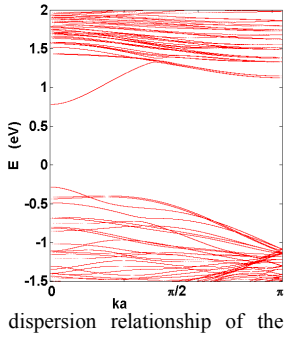

Figure 4. Energy dispersion relationship of the multiple quantum confined subbands of a square (100) InSb nanowire of width 26.92 Å.

rise to OFF state BTBT leakage current. This leakage current decreases with decreasing nanowire width, since wires with smaller cross-section have higher confinement-induced band gap, which reduces the overlap of bands, thereby reducing the current. The peak drain current, however also decreases slightly on reducing the cross-section (see the linear I_D-V_G in the inset of Fig. 5) likely due to an overall reduction in injected carrier velocities resulting from strong confinement combined with non-parabolicity.

In the OFF state, band-to-band leakage, and thus I_D is also be reduced by increasing the channel length (L_{ch}) due to an increase in the barrier thickness and an associated decrease in tunneling probability. This reduction can be seen in Fig. 6 for two devices with different L_{ch} (5.4 and 10.2 nm) but same nanowire width (W_{ch} = 3.24 nm) where the minimum subthreshold current decreases by ~ 3 orders of magnitude. In the ballistic limit, I_{ON} remains unaffected by the increase of L_{ch} (inset of Fig. 6) although gate capacitance and required drive current would increase. And with some diffusive transport, however, I_{ON} would also be expected to decrease with the increase of L_{ch}. As a consequence, L_{ch} would have to be optimized to achieve the best combination of I_{ON} and low I_{OFF}.

ACKNOWLEDGMENT

This work was supported in part by NRI SWAN Center and the GRC.

REFERENCES

[1] W. Chen, L. F. Register, and S. K. Banerjee, "Simulation of quantum effects along the channel of ultrascaled Si-based MOSFETs," *IEEE Trans. Electron. Dev.*, vol. 49, no. 4, pp. 652-657, Apr. 2002.

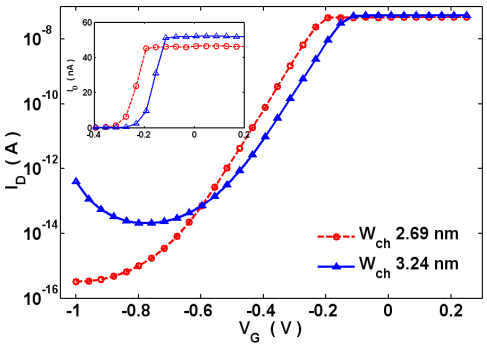

Figure 5. I_D-V_G for two square (100) InSb nanowire MOSFETs in a gate-all-around architecture, having identical channel length of L_{ch} = 5.2 nm. Inset: I_D-V_G in linear scale. Lower BTBT in narrow nanowires reduces I_{OFF}, but peak I_D is also slightly reduced for narrow nanowires.

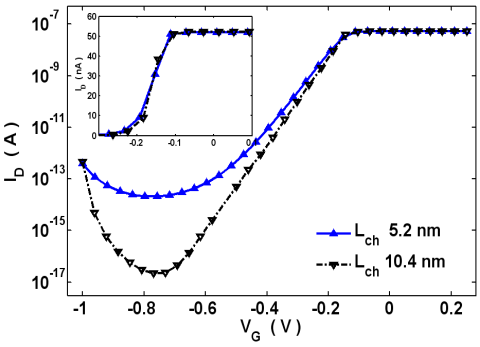

Figure 6. I_D-V_G for two square (100) InSb nanowire MOSFETs, having same W_{ch} = 3.24 nm, and different L_{ch}. Longer L_{ch} has lower minimum I_{OFF} due to reduced BTBT. In absence of scattering, peak I_D is unchanged with L_{ch}, as the inset (linear plot of I_D-V_G) confirms.

[2] International Technology Roadmap for Semiconductors, 2007 Edition, Emerging Research Devices, Available at http://www.itrs.net.

[3] S. Datta, "III-V field-effect transistors for low power digital logic applications," *Microelec. Engg.*, vol. 84, no. 9-10, pp. 2133-37, Sep-Oct 2007.

[4] T. Ashley *et al.*, "Heterogeneous InSb quantum well transistors on silicon for ultra-high speed, low power logic applications," *Electron. Lett.*, vol. 43, no. 14, Jul 2007.

[5] T. Ashley *et al.*, "Novel InSb-based quantum well transistors for ultra-high speed, low power logic applications," *Proc. 7th International Conf. on Solid-State and IC Technology*, pp. 2253-2256, 2004.

[6] S. Datta *et al.*, "85 nm gate length enhancement and depletion mode InSb quantum well transistors for ultra high speed and very low power digital logic applications," *IEDM Tech. Dig.*, pp. 763-765, 2005.

[7] T. Usuki, M. Saito, M. Takatsu, R. A. Kiehl, and N. Yokoyama, "Numerical analysis of ballistic-electron transport in magnetic fields by using a quantum point contact and a quantum wire", *Phys. Rev. B*, vol. 52, no. 11, pp. 8244-55, Sep. 1995.

[8] S. Datta, Electronic Transport in Mesoscopic Systems, Cambridge University Press, 1995.

[9] D. Basu, M. J. Gilbert, L. F. Register, S. K. Banerjee, and A. H. MacDonald, "Effect of edge roughness on electronic transport in graphene nanoribbon channel metal-oxide-semiconductor field-effect transistors," *App. Phys. Lett.*, vol. 92, no. 042114, 2008.

[10] P. Vogl, H. P. Hjalmarson, J. D. Dow, "A semi-empirical tight-binding theory of the electronic structure of semiconductors," *Jour. of Phy. and Chem. of Solids*, vol 44, no. 5, pp. 365-379, 1983.

[11] G. Klimeck, R. C. Bowen, T. B. Boykin, T. A. Cwik, "sp^3s^* tight-binding parameters for transport simulations in compound semiconductors", *Superlatt and Microstruc.*, vol. 27, no. 5-6, pp. 519-524, 2000.

[12] Visual Numerics, Inc. http://www.vni.com/products/imsl/index.php.

[13] R. Akis, L. Shifren and D.K. Ferry, "Self-consistent modeling of open quantum devices," *Tech. Proc. of the International Conf. Computational Nanoscience and Nanotechnology*, 2002.

[14] T. Ando, "Quantum point contacts in magnetic fields," *Phys. Rev. B*, vol. 44, no. 15, pp. 8017-27, 1991.

[15] P. A. Khomyakov, G. Brocks, V. Karpan, M. Zwierzycki and P. J. Kelly, "Conductance calculations for quantum wires and interfaces: Mode matching and Green's functions," *Phys. Rev. B* vol. 72, no. 035450, 2005.

[16] C. Caroli, R. Combescot, P. Nozieres, D. Saint-James, "Direct calculation of the tunneling current," *J. Phys. C*, vol. 4, pp. 916-914, 1971.

[17] M. J. Gilbert and D. K. Ferry, "Efficient quantum three-dimensional modeling of fully depleted ballistic silicon-on-insulator metal-oxide-semiconductor field-effect-transistors", *J. App. Phys.*, vol. 95, pp. 7954-60, 2004.

[18] M. Luisier, A. Schenk, W. Fichtner and G. Klimeck, "Atomistic simulation of nanowires in the $sp^3d^5s^*$ tight-binding formalism: From boundary conditions to strain calculations," *Phys. Rev. B*, vol. 74, no. 205323, 2006.

[19] J. Robertson and B. Falabretti, "Band offsets of high K gate oxides on III-V semiconductors," *J. Appl. Phys.* vol. 100, no. 014111, 2006.