Model Order Reduction for Multiband Quantum Transport Simulations and its Application to p-Type Junctionless Transistors

Jun Z. Huang, Weng Cho Chew, *Fellow, IEEE*, Jie Peng, Chi-Yung Yam, Li Jun Jiang, *Member, IEEE*, and Guan-Hua Chen

***Abstract*— An efficient method is developed for multiband simulation of quantum transport in nanowire electronic devices within nonequilibrium Green's function formalism. The efficiency relies on a model order reduction technique, which projects the $k \cdot p$ Hamiltonian into a much smaller subspace constructed by sampling the Bloch modes of each cross-section layer. Several sampling approaches are discussed to obtain a minimum and accurate basis with reduced computational overhead. The technique is verified by calculating the valence bands of silicon nanowires (SiNWs) and by solving I-V curves of p-type SiNW transistors. This enables us to study for the first time the performances of large cross-section p-type junctionless (JL) transistors in the quantum ballistic transport limit. The influences of doping density, transport direction, channel length, and cross-section size are examined. We find that larger doping densities may lead to worse sub-threshold slopes due to the enhanced source-to-drain tunneling. Compared with their counterparts, i.e., classical inversion-mode (IM) transistors, they have better sub-threshold behaviors, but they do not necessarily provide a better ON/OFF ratio except when the channel is short or thin. In addition, unlike IM transistors, [110] and [111] channel directions in JL transistors are very robust against channel thicknes scaling.**

***Index Terms*— Junctionless transistors, $k \cdot p$ approach, model order reduction (MOR), multiband simulation, nonequilibrium Green's function (NEGF), quantum transport, silicon nanowire transistors.**

I. Introduction

ONE-DIMENSIONAL nanowire structures, such as carbon nanotubes (CNTs), graphene nanoribbons (GNRs), and silicon (germanium or III–V material) nanowires (SiNWs), have attracted much attention during the past two decades. Due to their excellent physical properties, they are believed to have great potential in many applications, including the building blocks of future electronic devices.

To understand the electrical properties of nanodevices built upon these small structures, a quantum-mechanical method, nonequilibrium Green's function (NEGF) approach [1], has been widely used to simulate their carrier transport. As the computational cost of the real space (RS) NEGF approach is huge, mode space (MS) approaches have been successfully developed for simulating nanostructures with strong confinement in the lateral directions, such as the nanowires mentioned above [2]–[4]. These approaches expand the device Hamiltonian in the space spanned by the eigenmodes of the cross sections. Making use of the fact that usually a few modes participate in the transport process, the dimension of the Hamiltonian matrix in the MS can be greatly reduced and thus the Green's function in the MS can easily be solved. This is true for single-band effective mass approximation, since the eigenmodes for each cross section are wave vector k-independent. For more accurate multiband models, such as the tight-binding and $k \cdot p$ models, as pointed out in [4]–[6], the modes are generally k-dependent and thus the transformation from the RS to MS does not exist. It is recently shown that the k-dependent modes also make the contact block reduction method troublesome when it is combined with tight-binding model [7].

The exception is the tight-binding model of gate-all-around (GAA) CNT transistors, which has a rigorous MS approach [8]. However, to simulate general CNT transistors which do not possess GAA feature [9] and GNR transistors [10], some crude approximations have been made so that the MS approach can still be applied. As a result, the accuracy is compromised. To improve the accuracy, a criterion of mode selection for GNR transistors has been suggested [6] that works pretty well near the conduction band minima (and valence band maxima). Quite recently, low-dimensional equivalent transport models have been constructed for tight-binding Hamiltonians of SiNWs [11], thanks to a spurious mode elimination process. Another low rank approximation method has been tried [12], but it involves large eigenvalue problems requiring much computational cost.

A MS approach has also been proposed for multiband $k \cdot p$ models [5], which demonstrates great success to simulation of p-type SiNW transistors and InAs nanowire tunneling

Manuscript received December 22, 2012; revised February 21, 2013 and March 28, 2013; accepted April 22, 2013. Date of publication May 31, 2013; date of current version June 17, 2013. This work was supported in part by the Hong Kong Research Grant Council, under Grant HKU7009/09P, Grant 7008/08P, Grant 7011/06P, Grant 7013/07P, Grant 604709, Grant 711511E, Grant 713011E, Grant 711609E, Grant 712612, and Grant HKUST 9/CRF/08, the University Grant Council AoE/P-04/08, and National Science Foundation of China, under Contract NSFC 20828003 and Contract 61271158. The review of this paper was arranged by Editor A. Schenk.

J. Z. Huang and W. C. Chew are with the Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong, and also with the Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, IL 61801-2991 USA (e-mail: huangjun@eee.hku.hk; wchew@uiuc.edu).

J. Peng, C.-Y. Yam, and G.-H. Chen are with the Department of Chemistry, University of Hong Kong, Hong Kong (e-mail: pengjie@yangtze.hku.hk; yamcy@yangtze.hku.hk; ghc@yangtze.hku.hk).

L. J. Jiang is with the Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong (e-mail: ljiang@eee.hku.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2013.2260546

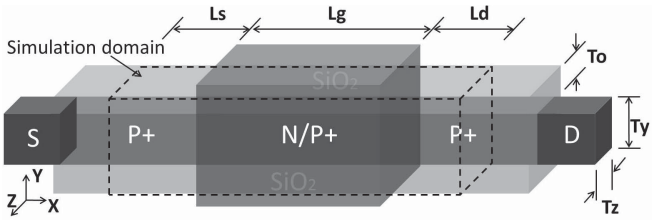


Fig. 1. P-type GAA SiNW transistor. If the channel is N-type doping, the device is a classical IM transistor. If the channel has the same doping type and doping density as the source (or drain) extension region, it is a JL transistor.

transistors. Unfortunately, the modes adopted can only accurately expand the wave function near the Γ point. Away from the Γ point, many modes are actually needed which limits its performance. Generally, to capture the feature of k-dependent modes in multiband simulations, as is commonly done in model order reduction (MOR) methods in electromagnetics [13] and asymptotic waveform evaluation [14], it is better to adopt multipoint expansion. This is adopted in this paper. We will demonstrate this approach by simulation of hole transport in p-type SiNW field effect transistors using the three-band and six-band $k \cdot p$ Hamiltonian.

Recently, JL transistors have been proposed [15] and experimentally demonstrated [16]; they show extraordinarily promising performance but with simpler fabrication. To characterize the performance and illustrate the physics, many simulations have been carried out, either semi-classically [15]–[22] or quantum-mechanically [23]–[26]. For quantum-mechanical study, as far as we know, only n-type ones with large cross sections have been carried out [24]–[26]. This is because the single-band effective mass model is enough for the description of conduction band and it can be done in the MS. Since the description of the valence band requires computationally more intensive multiband model, simulation of p-type ones has been limited to only 1.15-nm diameter [23] and performances of large cross-section ones remain unexamined. With our MOR technique, we are able to fill the gap. As a step toward more sophisticated full NEGF simulation, coherent transport will be assumed in this paper.

In Section II, we first describe the multiband model and outline the NEGF approach, then we present the MOR technique in detail and benchmark it with caution. In Section III, the method is applied to simulate p-type JL transistors with different channel materials and geometries. Various device figures-of-merit are extracted and compared with those of classical IM transistors. Conclusions are drawn in Section IV.

II. Method Description

A. Multiband Effective Mass Equation

According to multiband effective mass theory [27], the wavefunction inside the nanostructures can be found by solving the following coupled differential equation for envelop function F_m $(m = 1, 2, \ldots, N)$

$$\sum_{n=1}^{N} \left[H_{mn}^{kp}(-i\nabla) + V(\mathbf{r})\,\delta_{mn} \right] F_n(\mathbf{r}) = E F_m(\mathbf{r}) \qquad (1)$$

where N is the number of bands considered, $V(\mathbf{r})$ is the slowly varying perturbed potential distribution, and operator $H_{mn}^{kp}(-i\nabla)$ is the element of the $k \cdot p$ Hamiltonian with k replaced by differential operator $-i\nabla$.

The six-band $k \cdot p$ Hamiltonian can be written as (if we arrange the basis in this order, three spin up p atomic orbital-like states and three spin down ones) [28]

$$\mathbf{H}^{kp} = \left(E_{VB,0} + \frac{\hbar^2 k^2}{2m_0} \right) \mathbf{I} + \begin{pmatrix} \mathbf{H}^{dkk} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}^{dkk} \end{pmatrix} + \mathbf{H}^{so} \qquad (2)$$

where $E_{VB,0}$ is the valence band edge, $\mathbf{I}$ is the identity matrix.

The Dresselhaus–Kip–Kittel (DKK) Hamiltonian $\mathbf{H}^{dkk}$ is

$$\mathbf{H}^{dkk} = \begin{pmatrix} \tilde{L}k_x^2 + Mk^2 & Nk_xk_y & Nk_xk_z \\ Nk_xk_y & \tilde{L}k_y^2 + Mk^2 & Nk_yk_z \\ Nk_xk_z & Nk_yk_z & \tilde{L}k_z^2 + Mk^2 \end{pmatrix} \qquad (3)$$

where $\tilde{L} = L - M$, and the parameters L, M, N are related to the effective masses, which can be found in [29]. The spin–orbit interaction $\mathbf{H}^{so}$ can be written as

$$\mathbf{H}^{so} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ -\mathbf{B}^* & \mathbf{A}^* \end{pmatrix} \qquad (4)$$

with

$$\mathbf{A} = \frac{\Delta}{3} \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{B} = \frac{\Delta}{3} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & -i \\ -1 & i & 0 \end{pmatrix} \qquad (5)$$

where Δ is the spin–orbit splitting parameter, which can be set to zero to reduce to three-band model.

To numerically solve (1), we need to discretize the differential operator, which can be done using finite difference method (FDM) provided in [30]. Note that for nanowire directions other than [100], coordinate transformation for (2) should be performed before the discretization.

B. NEGF Solution

For transport problems, it is required to solve (1) with open boundary conditions and then get the physical quantities of interest, such as charge density and current. This can be nicely formulated with NEGF approach [1].

In this formalism, we need to solve the retarded Green's function $\mathbf{G}$ of the device region (in RS) defined as

$$[E\mathbf{I} - \mathbf{H}_0 - \Sigma(E)]\,\mathbf{G}(E) = \mathbf{I} \qquad (6)$$

where $\mathbf{H}_0$ is the discretized $k \cdot p$ Hamiltonian of the isolated device (with potential term included), and Σ is the self-energy matrix due to the semi-infinite leads [31]. For nanowire structure like Fig. 1, $\mathbf{H}_0$ can take a block tridiagonal form, the diagonal block $\mathbf{H}_{i,i}$ (with size $N_t \times N_t$) is the on-site Hamiltonian for the ith layer; the off-diagonal block $\mathbf{H}_{i,i\pm1}$ (with size $N_t \times N_t$) is the coupling Hamiltonian between the ith and the $(i \pm 1)$th layer. Thus, $\mathbf{H}_0$ is of size $(N_t N_l) \times (N_t N_l)$ with N_l being the number of layers. In coherent transport limit, Σ just has two nonzero blocks, this first is $\Sigma_{1,1}$ and the last is Σ_{N_l,N_l}.

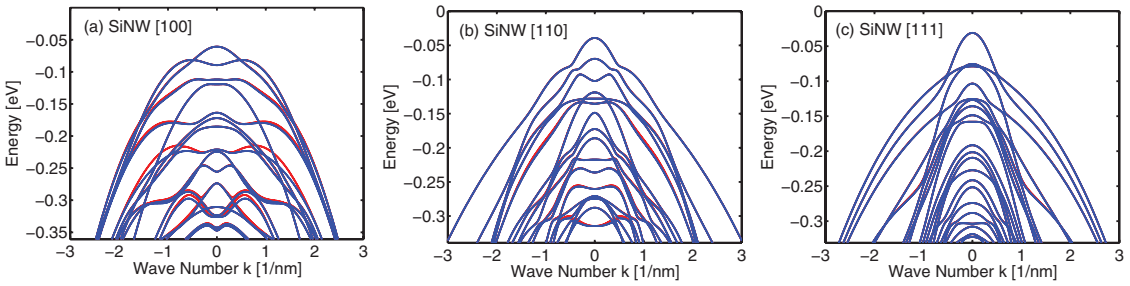


Fig. 2. E-k relations for 5 × 5 nm SiNWs in the [100], [110], and [111] directions. Red lines: exact solution. Blue lines: MOR solution.

The charge density $n(\mathbf{r})$ and the current $J_{i\to i+1}$ flowing between layer i and layer $i+1$ can both be expressed in terms of $\mathbf{G}(E)$ [1], [32]

$$n(\mathbf{r}) = 2\int \frac{dE}{2\pi} G^n(\mathbf{r},\mathbf{r},E) \tag{7}$$

$$J_{i\to i+1} = 2\frac{ie}{\hbar}\int \frac{dE}{2\pi}\mathrm{Trace}[\mathbf{H}_{i,i+1}\mathbf{G}^n_{i+1,i}(E) - \mathbf{H}_{i+1,i}\mathbf{G}^n_{i,i+1}(E)] \tag{8}$$

where correlation function $\mathbf{G}^n = \mathbf{G}\Sigma^{in}\mathbf{G}^\dagger$. Here, Σ^{in} takes the same format as Σ, but with the first and last diagonal blocks replaced by $-2\Im m\left(\Sigma_{1,1}\right) f(E-\mu_L)$ and $-2\Im m\left(\Sigma_{N_l,N_l}\right) f(E-\mu_R)$, where $f(E)$ is the Fermi–Dirac distribution function, μ_L and μ_R are the chemical potentials of the left and right leads.

The problem is that solving (6) for realistic systems is very difficult in terms of both CPU and memory requirements. Furthermore, we need to solve (6) for different E as required by (7) and (8), and we need to recalculate (7) once a new potential is generated by Poisson equation until self-consistency is achieved. In spite of the popular recursive Green's function (RGF) algorithm [32], its CPU and memory cost are $O(N_t^3 N_l)$ and $O(N_t^2 N_l)$, respectively, and therefore is only feasible for small cross-section size.

C. Model Order Reduction

Similar to the MS approach, the first step is to construct a unitary transformation matrix $\mathbf{U}$ of size $(N_t N_l)\times(N_m N_l)$ in the following format:

$$\mathbf{U} = \mathrm{diag}\left(\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_i, \ldots, \mathbf{V}_{N_l}\right) \tag{9}$$

where $\mathbf{V}_i$ $(i = 1, 2, \ldots, N_l)$ is a $N_t \times N_m$ $(N_m < N_t)$ submatrix that contains reduced basis for layer i.

Then, (6) can be transformed into this reduced basis

$$\left[E\mathbf{I} - \tilde{\mathbf{H}}_0 - \tilde{\Sigma}(E)\right]\tilde{\mathbf{G}}(E) = \mathbf{I} \tag{10}$$

where

$$\begin{aligned}\tilde{\mathbf{H}}_0 &= \mathbf{U}^\dagger\mathbf{H}_0\mathbf{U}\\ \tilde{\Sigma}(E) &= \mathbf{U}^\dagger\Sigma(E)\mathbf{U}\\ \tilde{\mathbf{G}}(E) &= \mathbf{U}^\dagger\mathbf{G}(E)\mathbf{U}.\end{aligned} \tag{11}$$

Note that $\tilde{\mathbf{H}}_0$ is still block tridiagonal and $\tilde{\Sigma}(E)$ can be directly calculated from lead Hamiltonian in the reduced space.

Solving (10) instead of (6) presents numerical advantages since the matrix involved is of reduced size $(N_m N_l)\times(N_m N_l)$, and this can be done efficiently by the standard RGF algorithm with CPU cost $O(N_m^3 N_l)$ and memory cost $O(N_m^2 N_l)$. With $\tilde{\mathbf{G}}(E)$, we can calculate the physical quantities in the reduced space with similar expressions as (7) and (8). After that the physical quantities in the RS can be recovered with inverse transformation. The accuracy and efficiency of this method are very much dependent on how we construct the reduced

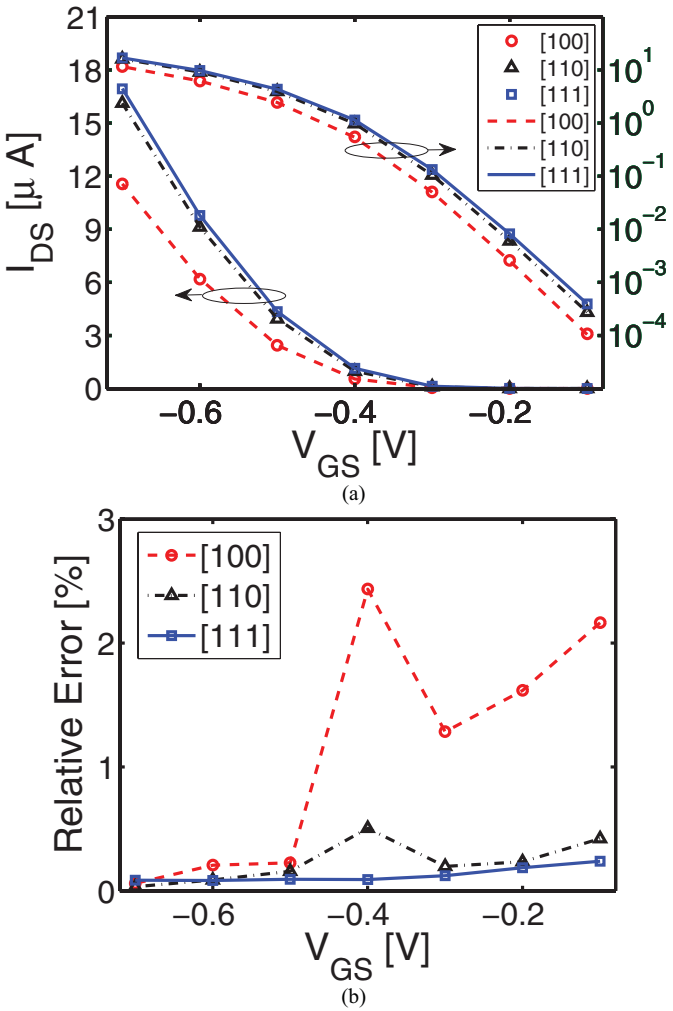


Fig. 3. (a) $I_{DS}-V_{GS}$ characteristics obtained by setting $E_m = E_t - 0.3$ eV (lines) and $E_m = E_t - 0.5$ eV (symbols). IM SiNW transistors in the [100], [110], and [111] directions are considered. Doping density N_d in the source and drain is 1×10^{20} cm^{-3}, while in the channel it is 1×10^{15} cm^{-3}. $T_y = T_z = 5$ nm, $L_g = 10$ nm, $T_o = 1$ nm. Drain bias is set to be $V_{DS} = -0.5$ V. (b) Relative error of the two sets of currents.

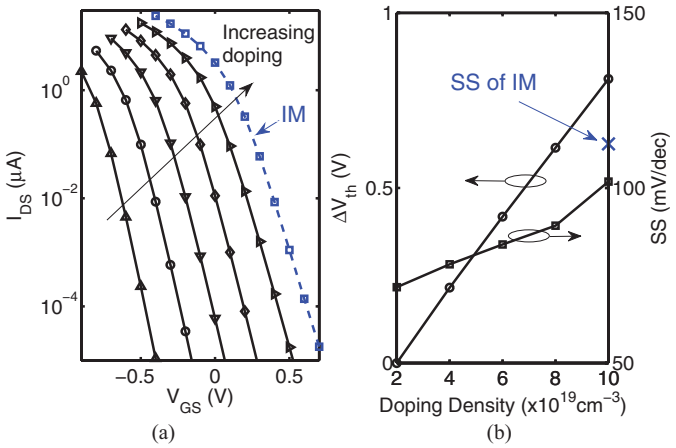


Fig. 4. (a) $I_{DS}-V_{GS}$ curves for JL transistors with different doping densities N_d. The curve for IM device is also plotted. The dimensions of the devices are all fixed to $T_y = T_z = L_g = 5$ nm. The N_d for JL devices is linearly varied from 2×10^{19} cm^{-3} to 1×10^{20} cm^{-3}. Transport direction [100] is considered here and $V_{DS} = -0.5$ V. (b) Extracted ΔV_{th} and SS with respect to doping density. Also shown is the SS for IM device.

basis set (9).

D. Construction of the Reduced Basis

We know that the $E-k$ relation of layer i repeating along the transport direction x can be obtained by solving the following eigenvalue problem (EVP):

$$\left(\mathbf{H}_{i,i} + \mathbf{H}_{i,i+1}e^{ik_x\Delta x} + \mathbf{H}_{i,i+1}^{\dagger}e^{-ik_x\Delta x}\right)\Psi_i = E\Psi_i \qquad (12)$$

where k_x is the wavenumber in the transport direction, Δx is the layer thickness, and Ψ_i is the eigenmode.

The criteria for constructing $\mathbf{V}_i$ is that, while N_m is kept as small as possible, when $\mathbf{V}_i$ is applied to (12), the reduced EVP should produce the original $E-k$ relation as accurately as possible. The reduced EVP is

$$\left(\tilde{\mathbf{H}}_{i,i} + \tilde{\mathbf{H}}_{i,i+1}e^{ik_x\Delta x} + \tilde{\mathbf{H}}_{i,i+1}^{\dagger}e^{-ik_x\Delta x}\right)\tilde{\Psi}_i = \tilde{E}\tilde{\Psi}_i \qquad (13)$$

where

$$\begin{aligned}\tilde{\mathbf{H}}_{i,i} &= \mathbf{V}_i^{\dagger}\mathbf{H}_{i,i}\mathbf{V}_i \\ \tilde{\mathbf{H}}_{i,i+1} &= \mathbf{V}_i^{\dagger}\mathbf{H}_{i,i+1}\mathbf{V}_{i+1} \\ \Psi_i &= \mathbf{V}_i\tilde{\Psi}_i. \end{aligned} \qquad (14)$$

1) K Space Sampling: The reduced basis can be constructed by solving (12) for each layer i at n judiciously sampled k_x points [instead of solving (12) only at $k_x = 0$, as in [5]]. Suppose m_j eigenmodes are obtained at $k_x = k_j$ with eigenvalues inside the window $E_m \le E \le 0$, where E_m is the minimum energy of interest (usually several hundred meV below the top of the valence band), from which we construct a matrix $\mathbf{W}_i$

$$\mathbf{W}_i = \Big[\Psi_i^1(k_1), \ldots, \Psi_i^{m_1}(k_1), \Psi_i^1(k_2), \ldots, \Psi_i^{m_2}(k_2), \ldots, \Psi_i^1(k_n), \ldots, \Psi_i^{m_n}(k_n)\Big] \qquad (15)$$

which is then **QR** factorized. The unitary matrix **Q** then serves as the sub-matrix $\mathbf{V}_i$ in (9).

Solving (12) for several lowest eigenmodes can be done efficiently with iterative solvers since the matrix is highly sparse as a result of FDM. Moreover, we just need to solve it at positive k_x (or negative k_x) saving half the cost. Suppose we already have Ψ_i and E as the eigenpairs of (12) at k_x. When spin–orbit coupling is not considered, $\mathbf{H}_{i,i}$ and $\mathbf{H}_{i,i+1}$ are both real matrices. In this case, it is easy to prove that $\Psi_i' = (\Psi_i)^*$ and E are the eigenpairs of (12) at $-k_x$. When spin–orbit coupling is taken into account, instead, we do the following transformation to obtain those at $-k_x$ [which can be verified through (4)]

$$\Psi_i = \begin{pmatrix}\Psi_i \uparrow \\ \Psi_i \downarrow\end{pmatrix} \Rightarrow \Psi_i' = \begin{pmatrix}\Psi_i' \uparrow \\ \Psi_i' \downarrow\end{pmatrix} \qquad (16)$$

where $\Psi_i \uparrow$ and $\Psi_i \downarrow$ are the spin up and spin down components, respectively, $\Psi_i' \uparrow = (\Psi_i \downarrow)^*$, and $\Psi_i' \downarrow = -(\Psi_i \uparrow)^*$.

2) E Space Sampling: Alternatively, to obtain the eigenmode Ψ_i for each layer i, we can solve the following generalized eigenvalue problem (GEVP) [31] at n judiciously sampled E points:

$$\begin{pmatrix}\mathbf{0} & \mathbf{I} \\ \mathbf{H}_{i,i+1}^{\dagger} & \mathbf{H}_{i,i} - E\mathbf{I}\end{pmatrix}\begin{pmatrix}\Psi_i \\ \Psi_{i+1}\end{pmatrix} = \lambda\begin{pmatrix}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{H}_{i,i+1}\end{pmatrix}\begin{pmatrix}\Psi_i \\ \Psi_{i+1}\end{pmatrix} \qquad (17)$$

where $\lambda = e^{ik_x\Delta x}$. It is well known that the eigenpairs with $|\lambda| = 1$ correspond to the propagating modes; whereas the eigenpairs with $|\lambda| < 1$ ($|\lambda| > 1$) correspond to the decaying (growing) modes. Here, we only use the propagating modes to construct $\mathbf{W}_i$ as in (15), which is then orthonormalized to form $\mathbf{V}_i$.

To selectively solve the eigenpairs with $|\lambda| = 1$, one can adopt the Krylov subspace method with some shift-and-invert strategies [33]. As these eigenvalues distribute in a circle in the complex plane and for low energy range, they tend to cluster around 1 (since k_x is small), we choose the shift $\sigma = e^{i\theta}$ with $\theta = 0$. For the two cases mentioned before, we can further reduce the cost by choosing $\theta = \hat{\theta}$ (where $\hat{\theta}$ is a value slightly larger than 0) and solving the eigenmodes having eigenvalues in the upper half plane; then the corresponding transformation of these eigenmodes are those in the lower half plane.

3) Hybrid Sampling: Both of the above sampling schemes work well but each has its advantages and disadvantages. Sampling in the k space is fast since we only need to solve EVP (12) several times, the drawback is that it is not easy to determine the sampling points. While sampling in the E space has the advantage that we can easily establish the energy window to sample, it is slow as it is required to solve interior GEVPs (17) with matrix dimension twice the layer size. Therefore, we propose to hybridize the above two methods to construct $\mathbf{V}_i$.

It turns out that, as will be shown later, sampling at one particular k_x point ($k_x = 0$) and at one E point ($E = E_m$) can approximate very well the $E-k$ dispersion for energy range $E_m \le E \le 0$. The reason is that the modes at $k_x = 0$ can well produce the band structure near the Brillouin zone center, while the modes at $E = E_m$ are excellent for correcting the band structure far away from the center. In this simple scheme,

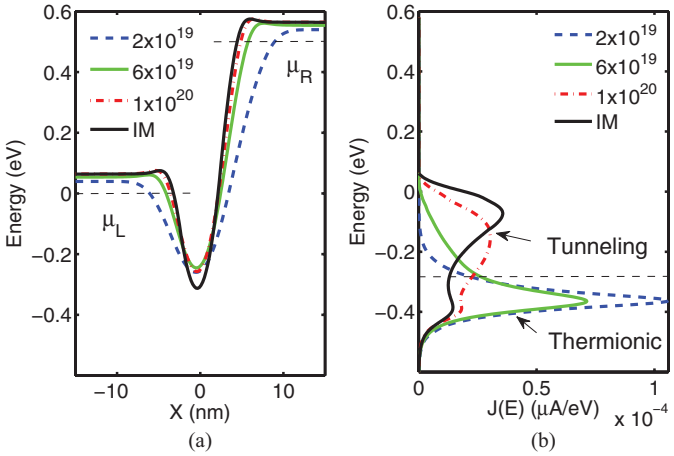


Fig. 5. (a) Potential distributions and (b) current spectra of JL transistors with different doping density N_d. The case of IM transistor is also shown. The potentials are sampled at the center of the silicon layer. All device parameters are the same as those in Fig. 4, with the gate voltages of all cases tuned to achieve the same current at $I_{DS} = 1 \times 10^{-5}\,\mu A$.

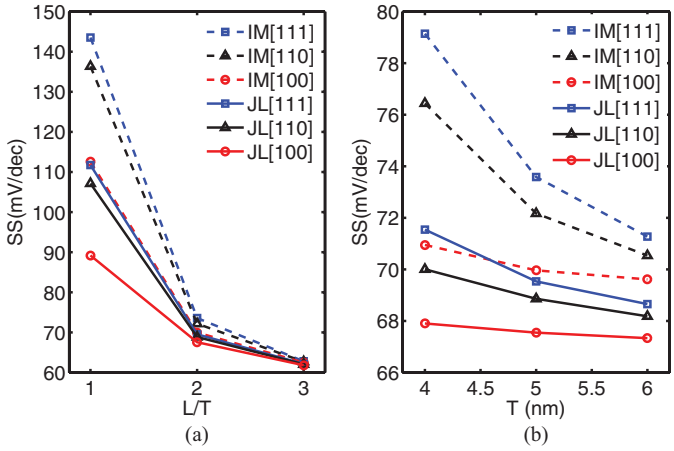


Fig. 6. (a) SS for different channel aspect ratio, the channel thickness is fixed to $T = 5$ nm. (b) SS for different channel thickness, the channel aspect ratio is fixed to $L/T = 2$.

for each layer, we only need to solve (12) once and solve (17) once. In addition, the only parameter we need to obtain is E_m. It is expected that a lower E_m may improve the accuracy since a wider energy range is approximated, but it also slows down the simulation as N_m becomes larger.

E. Validation of the Method

To validate our MOR method, which is essentially a new method of constructing the reduced basis, we follow two steps. First, we check if the reduced EVP can accurately capture the $E - k$ diagram. We examined SiNWs with cross-section size 5×5 nm in the [100], [110], and [111] directions. Six-band $k \cdot p$ model is used and the parameters are chosen to tune to tight-binding model provided in [34]. We choose $E_m = E_t - 0.3$ eV, where E_t is the top of the valence band, which results in N_m equal to 96, 102, and 126, respectively. Here, we use $\hat{\theta} = \pi/12$ as the eigenvalues fall within a very narrow region with $-\pi/6 < \theta < \pi/6$. As can be seen from Fig. 2, the $E - k$ diagrams obtained by our MOR method are very close to the exact solutions in all three cases. The reduction is tremendous, compared with the original matrix size $N_t = 9126$ as a result of 0.125-nm mesh size.

Then, one may wonder if the NEGF results are also correctly produced. A p-type IM SiNW transistor as shown in Fig. 1 is simulated, with NEGF and Poisson equations solved self-consistently. Again, we choose $E_m = E_t - 0.3$ eV and compare the drain currents to those obtained by setting $E_m = E_t - 0.5$ eV, which can be regarded as a more accurate solution. Three-band $k \cdot p$ model is used as spin–orbit coupling plays negligible role [34], which leads to N_m values that are roughly half of the values in six-band model. For simplicity, hard-wall boundary condition at the silicon-oxide interface is implemented, which is valid for large cross-section nanowires. The Poisson equation is solved with the same method and boundary conditions as in [14]. As can be seen in Fig. 3(a), the two solutions almost overlap with each other. The relative errors of the two solutions are further manifested in Fig. 3(b), which shows that errors of [110] and [111] directions for the whole bias range are within 0.5%, whereas the errors of [100] direction below threshold are larger, but still within 2.5%. Therefore, to be consistent with the $E - k$ calculation, $E_m = E_t - 0.3$ eV is justified. The relatively larger errors of [100] direction below threshold are due to the larger errors of the band structure approximation, as can be observed from Fig. 2(a). In particular, the subband edges near the middle of the energy window are not aligned, which may induce a small threshold voltage shift. To improve the accuracy, one can sample one more energy point at $(E_t + E_m)/2$. Note that for SiNWs with large cross sections ($\geq 4 \times 4$ nm), we have adopted 0.2-nm mesh size as this leads to negligible errors in the I-V curves (compared with 0.125-nm mesh size).

Although not shown here, we also examined the nanowires with different cross sections. It is found that for nanowires larger than 5×5 nm, choosing $E_m = E_t - 0.3$ eV is enough, while for cross sections smaller than 5×5nm, a slightly lower E_m is recommended to ensure that enough basis functions are included (for example, $E_m = E_t - 0.35$ eV for 4×4 nm nanowires). It should be mentioned that to achieve the same accuracy, N_m required in our MOR is expected to be less than that reported in [5] as our band structures are better approximated. In addition, thanks to the sparse solvers, construction of the reduced basis and the basis transformations do not occupy much CPU time, and most of the CPU time is spent on solving the reduced transport problem. As a result, simulation of an $I_{DS} - V_{GS}$ curve with ten bias points for a SiNW transistor with 5×5 nm channel cross-section size and 30 nm device length is within 10 h using single PC (Intel i5-2400 CPU at 3.10 GHz).

III. p-Type Junctionless Transistors

By applying the MOR method above, we are able to study hole transport in p-type JL SiNW FETs and compare them to similar IM ones. The impacts of various device structures and parameters on their performances will be examined. The gate oxide layer is assumed to be 1 nm, the channel length is varied

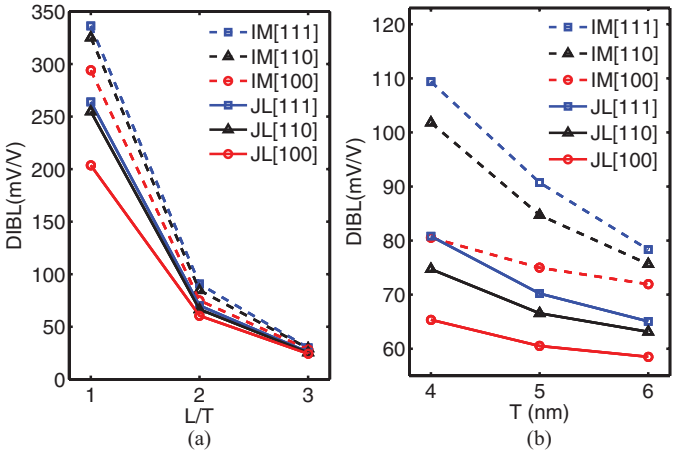


Fig. 7. (a) DIBL for different channel aspect ratio, the channel thickness is fixed to $T = 5$ nm. (b) DIBL for different channel thickness, the channel aspect ratio is fixed to $L/T = 2$.

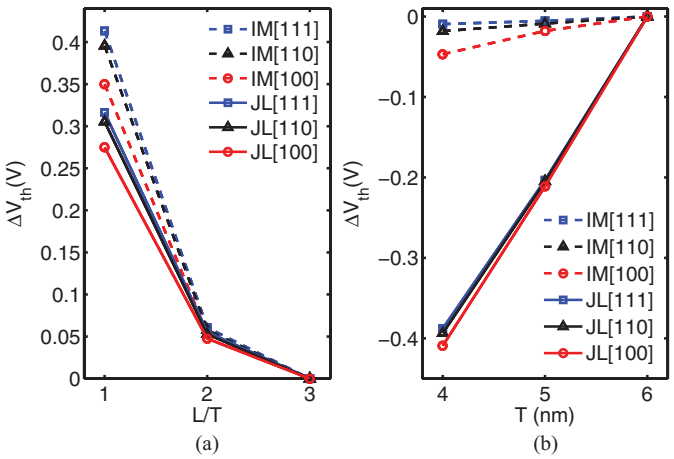


Fig. 8. (a) ΔV_{th} with respect to channel aspect ratio, the channel thickness is fixed to $T = 5$ nm, ΔV_{th} is obtained with reference to $L/T = 3$. (b) ΔV_{th} with respect to channel thickness, the channel aspect ratio is fixed to $L/T = 2$, ΔV_{th} is obtained with reference to $T = 6$ nm.

from 15 to 5 nm, and the channel cross-section size is varied from 6 $\times$6 nm to 4 $\times$4 nm. Channel orientations [100], [110], and [111] are considered. For IM FETs, the source and drain junctions are assumed to be abrupt, with N_d in the source/drain equal to 1×10^{20} cm^{-3} and in the channel it is 1×10^{15} cm^{-3}. The temperature is set to 300 K.

To characterize the device, we will extract from the $I - V$ curves the device metrics, such as sub-threshold slope (SS), doping density, or geometry induced threshold voltage change (ΔV_{th}), and drain-induced-barrier-lowering (DIBL). SS is expressed as millivolts of gate voltage needed for a decade change of drain current. Threshold voltage V_{th} is extracted using constant current method at 100 nA. DIBL is expressed as millivolts of ΔV_{th} induced by one volt change of drain voltage.

A. Doping Densities

For JL transistors, the first thing need to be ascertained is the doping density to use. Unlike the IM devices (which is lightly doped in the channel), the doping density actually affects the device performance greatly. In Fig. 4(a), we plot the $I - V$ curves of some short-channel JL devices, it is seen that SS increases as N_d increases. The SS values have been extracted and plotted in Fig. 4(b), which shows that for $N_d = 1 \times 10^{20}$ cm^{-3}, the SS can exceed 100 mV/dec. On the other hand, as seen from 4(a), the ON current for low N_d is very limited and it increases as N_d increases, as expected. Note that here we define the ON current as the current at flat band condition, which means no further increase of current is observed if the gate voltage is further increased. Therefore, one may suggest an N_d value based on a compromise between SS and ON current. It is also seen from 4(a) that as N_d increases, the $I - V$ curve shifts toward the positive direction, which means that more positive gate voltage is needed to turn the device off. In other words, V_{th} is shifted and ΔV_{th} is positive. As shown in 4(b), ΔV_{th} is almost linearly proportional to the change of N_d, say around 0.2V per 2×10^{19} cm^{-3} change of N_d, indicating

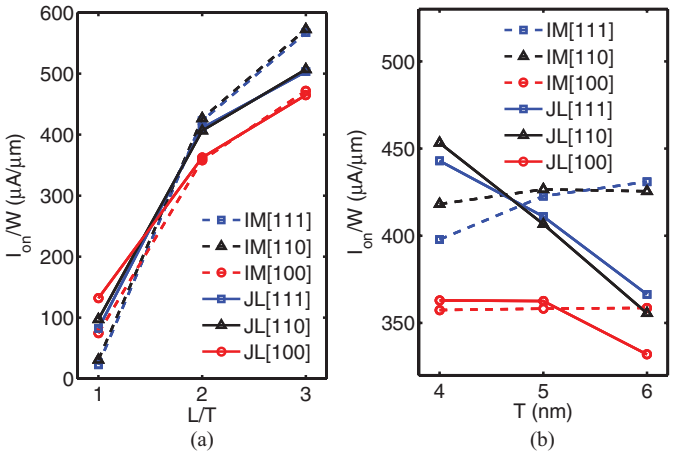


Fig. 9. (a) I_{on}/W for different channel aspect ratio, the channel thickness is fixed to $T = 5$ nm. (b) I_{on}/W for different channel thickness, the channel aspect ratio is fixed to $L/T = 2$.

that JL transistors are very sensitive to doping density variations.

To explain why SS degrades as doping density increases, we plot in Fig. 5 the potential distributions along the transport direction and their corresponding current spectra for three different doping densities. Those of IM transistor are also plotted in the same figure for comparison. We then roughly divide the current spectrum into two parts based on the peak of the potential barrier, the part below it can be attributed to thermionic current, while the part above it is largely due to tunneling contribution. It is clearly seen that as the doping density increases, the width of the potential barrier decreases, which results in larger source-to-drain tunneling current contribution that is known to degrade the SS. As this is due to electrostatics, similar behavior should be observed in n-type JL transistors as well. It is also obvious that the IM transistor has the thinnest potential barrier and the largest tunneling current, and thus the worst SS as shown in Fig. 4(b).

B. Channel Orientations and Scaling

In order to study the impacts of channel orientations and channel sizes, we fix N_d of JL transistors to be 8×10^{19} cm^{-3}. Note that different results may be obtained by choosing a different doping density, but the trends can be inferred based on the above analysis.

1) SS and DIBL: The SS as functions of channel aspect ratio and channel thickness are plotted in Fig. 6. From 6(a), it is observed that when channel length is long, the SS is very close to the 60 mV/dec limit. As the channel length is scaled down (with fixed channel thickness), the SS degrades as anticipated, for all nanowire directions and for both IM and JL cases. However, the JL devices degrade much slower than IM ones, which makes them very attractive for ultrascaled applications. The excellent SSs of JL devices are very much related to the effective gate length (EGL) concept that has been used to explain n-type JL transistors [26]. In fact, the EGL of JL device in the sub-threshold range is longer than that of IM device, which results in wider potential barrier that greatly helps the suppression of source-to-drain tunneling current. Our plots (for example, as shown in Fig. 5) confirm that this is also true for p-type devices. It is also observed that direction [100] has better SS although directions [110] and [111] do not differ too much, and this is more obvious when the channel becomes very short, indicating that [100] direction is more immune to short channel effect. The [100] direction's superior short channel performances can be explained by the fact that its larger effective mass of the first subband (as can be seen from Fig. 2) reduces the tunneling current contribution, as was first discovered in p-type double-gate transistors [35].

From Fig. 6(b), it is seen that, for IM devices, the SSs of [110] and [111] directions degrade when the channel is narrowed (with fixed aspect ratio), although SS of [100] direction does not degrade much. This is because the light hole effective mass of [100] direction grows much faster as the cross section becomes smaller, which suppresses the tunneling current and consequently compensates the otherwise faster increase of SS [34]. What we want to emphasize and is observed from 6(b) is that, unlike IM devices, SSs of JL ones change very little when the channel thickness is reduced, regardless of channel direction. Again, this is due to the longer EGL of JL devices, which makes the tunneling current less important.

The DIBL as functions of channel aspect ratio and channel thickness are plotted in Fig. 7. Similar trends as SS plotted in Fig. 6 are observed, as SS and DIBL are closely related quantities.

2) ΔV_{th}: The ΔV_{th} as functions of channel aspect ratio and channel thickness are plotted in Fig. 8. From 8(a), it is found that as the channel length is scaled down (with fixed channel thickness), more positive V_{th} is required to maintain the threshold current. The reason is twofold, one is that the negative drain voltage tends to raise the channel potential and the other is that tunneling current contribution becomes larger due to the narrower barrier. Direction [100] is again the most robust. In all cases, the JL devices outperform IM ones, particularly for ultrashort devices. These trends are similar to SSs in Fig. 6 and can be supported with similar arguments.

From Fig. 8(b), it is seen that while ΔV_{th} is very small for IM devices as the channel thickness is narrowed (with fixed aspect ratio), it is significantly larger for JL ones. This is consistent with the semi-classical studies [18], [19]. Regarding channel orientations, on the contrary, [100] is more sensitive to the channel thickness scaling and [111] turns out to be the most robust for both kinds of devices, although the distinction is less pronounced for JL ones. This has been attributed to the larger subband modulation in the [100] direction [36].

3) $I_{\rm ON}$: The $I_{\rm ON}/W$ (where $W = 4T$ is nanowire perimeter) as functions of channel aspect ratio and channel thickness are plotted in Fig. 9. The $I_{\rm ON}$ are obtained by setting $V_{GS} = V_{DS} = -0.5$ V after adjusting the gate work functions such that the $I_{\rm OFF}$ are all equal to 10 nA/μm [21]. From 9(a), it is found that JL devices have better $I_{\rm ON}/W$ only when the channel length is short, mainly due to their better short-channel SSs. However, when the channel length is long, the JL cases lose their advantages as the SSs become similar for both devices. In Fig. 9(b), it is seen that the JL devices have better $I_{\rm ON}/W$ only when the channel thickness is small, especially in the [110] and [111] directions, as a result of their better thin-channel SSs. Overall, [110] and [111] directions have similar $I_{\rm ON}/W$ and they are greater than those of [100] direction. The only exception is when $L/T = 1$ as shown in Fig. 9(a), where [100] direction provides the largest $I_{\rm ON}/W$. This is due to [100] direction's excellent short channel SS mentioned before. We have also lowered the $I_{\rm OFF}$ to examine their performances in low standby power applications, similar trends have been observed that JL devices have better $I_{\rm ON}/W$ only when the channel is short or thin, although they do have more margin to perform better.

C. Discussions

Our code is based on continuum $k \cdot p$ method and the dopants are modeled through doping concentration in the Poisson equation, which is valid when the devices are large. As the devices are aggressively scaled, the atomistic effects become crucial, calling for atomistic simulator. For example, just a few discrete dopants in the JL devices can result in a large doping density and there will be a doping density limit. Besides, the positions of these dopants matter, which may induce large performance variabilities, as reported recently by studying n-type JL transistors [25]. There is also an issue related to the dopant de-activation and dielectric screening at very small nanowires [37]. Such studies are outside of the scope of this paper and will be published elsewhere.

IV. Conclusion

In summary, an MOR technique was presented for efficient simulations of nanowire transistors based on the multiband $k \cdot p$ Hamiltonian and NEGF method. Numerical results showed that our method can correctly produce the band structures of SiNWs and I-V curves of p-type SiNW transistors, meanwhile significant reduction was achieved. With this

method, we then studied, for the first time, the influences of various device parameters on the performances of p-type JL transistors and compared them to IM devices.

Our method not only applies to GAA structures, but also applies to tri-gate structures, such as FinFETs. Alternative channels using Germanium or III–V materials could also be simulated in this framework. Moreover, strain effects can easily be incorporated into the $k \cdot p$ Hamiltonian. Further studies will be devoted to eight-band models, which enable us to simulate band-to-band tunneling devices.

ACKNOWLEDGMENT

The authors would like to thank their many AoE colleagues, in particular, J. Sun for setting up the computer cluster, and Q. Chen, S. Markov, and L. Zhang for helpful discussions.

REFERENCES

[1] S. Datta, *Quantum Transport: Atom to Transistor*. Cambridge, U.K.: Cambridge Univ., 2005.

[2] J. Wang, E. Polizzi, and M. Lundstrom, "A three-dimensional quantum simulation of silicon nanowire transistors with the effective-mass approximation," *J. Appl. Phys.*, vol. 96, no. 4, pp. 2192–2203, Aug. 2004.

[3] E. Polizzi and N. Ben Abdallah, "Subband decomposition approach for the simulation of quantum electron transport in nanostructures," *J. Comput. Phys.*, vol. 202, no. 1, pp. 150–180, 2005.

[4] M. Luisier, A. Schenk, and W. Fichtner, "Quantum transport in two- and three-dimensional nanoscale transistors: Coupled mode effects in the non-equilibrium Green's function formalism," *J. Appl. Phys.*, vol. 100, no. 4, pp. 043713-1–043713-12, 2006.

[5] M. Shin, "Full-quantum simulation of hole transport and band-to-band tunneling in nanowires using the $k \cdot p$ method," *J. Appl. Phys.*, vol. 106, p. 054505, Jan. 2009.

[6] R. Grassi, A. Gnudi, E. Gnani, S. Reggiani, and G. Baccarani, "Mode space approach for tight binding transport simulation in graphene nanoribbon FETs," *IEEE Trans. Nanotechnol.*, vol. 10, no. 3, pp. 371–378, May 2011.

[7] H. Ryu, H.-H. Park, M. Shin, D. Vasileska, and G. Klimeck, "Feasibility, accuracy, and performance of contact block reduction method for multi-band simulations of ballistic quantum transport," *J. Appl. Phys.*, vol. 111, no. 6, p. 063705, 2012.

[8] J. Guo, S. Datta, and M. Lundstrom, "Towards multi-scale modeling of carbon nanotube transistors," *Int. J. Multiscale Comput. Eng.*, vol. 2, pp. 257–276, Jan. 2004.

[9] G. Fiori, G. Iannaccone, and G. Klimeck, "Coupled mode space approach for the simulation of realistic carbon nanotube field-effect transistors," *IEEE Trans. Nanotechnol.*, vol. 6, no. 4, pp. 475–480, Jul. 2007.

[10] P. Zhao and J. Guo, "Modeling edge effects in graphene nanoribbon field-effect transistors with real and mode space methods," *J. Appl. Phys.*, vol. 105, p. 034503, Feb. 2009.

[11] G. Mil'nikov, N. Mori, and Y. Kamakura, "Equivalent transport models in atomistic quantum wires," *Phys. Rev. B*, vol. 85, no. 3, p. 035317, 2012.

[12] Y. He, L. Zeng, T. Kubis, M. Povolotskyi, and G. Klimeck, "Efficient solution algorithm of non-equilibrium Green's functions in atomistic tight binding representation," in *Proc. 15th Int. Workshop Comput. Electron.*, May 2012, pp. 1–3.

[13] C. Scheiber, A. Schultschik, O. Bíró, and R. D. Edlinger, "A model order reduction method for efficient band structure calculation of photonic crystals," *IEEE Trans. Magn.*, vol. 47, no. 5, pp. 1534–1537, May 2011.

[14] J. Z. Huang, W. C. Chew, M. Tang, and L. Jiang, "Efficient simulation and analysis of quantum ballistic transport in nanodevices with AWE," *IEEE Trans. Electron Devices*, vol. 59, no. 2, pp. 468–476, Feb. 2012.

[15] C. W. Lee, A. Afzalian, N. D. Akhavan, R. Yan, I. Ferain, and J. P. Colinge, "Junctionless multigate field-effect transistor," *Appl. Phys. Lett.*, vol. 94, pp. 053511-1–053511-2, Jan. 2009.

[16] J. P. Colinge, C. W. Lee, A. Afzalian, N. D. Akhavan, R. Yan, I. Ferain, P. Razavi, B. O'Neill, A. Blake, M. White, A. M. Kelleher, B. McCarthy, and R. Murphy, "Nanowire transistors without junctions," *Nature Nanotechnol.*, vol. 5, pp. 225–229, Mar. 2010.

[17] C. W. Lee, I. Ferain, A. Afzalian, R. Yan, N. D. Akhavan, P. Razavi, and J. P. Colinge, "Performance estimation of junctionless multigate transistors," *Solid-State Electron.*, vol. 54, no. 2, pp. 97–103, Feb. 2010.

[18] S. J. Choi, D. I. Moon, S. Kim, J. P. Duarte, and Y. K. Choi, "Sensitivity of threshold voltage to nanowire width variation in junctionless transistors," *IEEE Electron Devices Lett.*, vol. 32, no. 2, pp. 125–127, Feb. 2011.

[19] J. P. Colinge, A. Kranti, R. Yan, C. W. Lee, I. Ferain, R. Yu, N. D. Akhavan, and P. Razavi, "Junctionless nanowire transistor (JNT): Properties and design guidelines," *Solid-State Electron.*, vols. 65–66, nos. 11–12, pp. 33–37, Nov.–Dec. 2011.

[20] R. T. Doria, M. A. Pavanello, R. D. Trevisoli, M. de Souza, C. W. Lee, I. Ferain, N. D. Akhavan, R. Yan, P. Razavi, R. Yu, A. Kranti, and J. P. Colinge, "Junctionless multi-gate transistors for analog application," *IEEE Trans. Electron Devices*, vol. 58, no. 8, pp. 2511–2519, Aug. 2011.

[21] G. Leung, and C. O. Chui, "Variability of inversion-mode and junctionless FinFETs due to line edge roughness," *IEEE Electron Devices Lett.*, vol. 32, no. 11, pp. 1489–1491, Nov. 2011.

[22] G. Leung, and C. O. Chui, "Variability impact of random dopant fluctuation on nanoscale junctionless FinFETs," *IEEE Electron Devices Lett.*, vol. 33, no. 6, pp. 767–769, Jun. 2012.

[23] L. Ansari, B. Feldman, G. Fagas, J. P. Colinge, and J. C. Greer, "Simulation of junctionless Si nanowire transistors with 3 nm gate length," *Appl. Phys. Lett.*, vol. 97, no. 6, pp. 062105-1–062105-3, 2010.

[24] N. D. Akhavan, I. Ferain, P. Razavi, R. Yu, and J. P. Colinge, "Improvement of carrier ballisticity in junctionless nanowire transistors," *Appl. Phys. Lett.*, vol. 98, no. 10, pp. 103510-1–103510-3, Mar. 2011.

[25] M. Aldegunde, A. Martinez, and J. R. Barker, "Study of discrete doping induced variability in junctionless nanowire MOSFETs using dissipative quantum transport simulations," *IEEE Electron Device Lett.*, vol. 33, no. 2, pp. 194–196, Feb. 2012.

[26] P. Razavi, G. Fagas, I. Ferain, R. Yu, S. Das, and J. P. Colinge, "Influence of channel material properties on performance of nanowire transistors," *J. Appl. Phys.*, vol. 111, no. 12, pp. 124509-1–124509-8, 2012.

[27] S. L. Chuang, *Physics of Optoelectronic Devices*. New York, NY, USA: Wiley, 1995.

[28] C. Galeriu, "$k \cdot p$ theory of semiconductor nanostructures," Ph.D. dissertation, Worcester Polytech. Inst., Leicester Univ., Leicester, U.K., 2005.

[29] D. Gershoni, C. H. Henry, and G. A. Baraff, "Calculating the optical properties of multidimensional heterostructures: Application to the modeling of quaternary quantum well lasers," *IEEE J. Quantum Electron.*, vol. 29, no. 9, pp. 2433–2450, Sep. 1993.

[30] M. A. Khayer and R. K. Lake, "Modeling and performance analysis of GaN nanowire field-effect transistors and band-to-band tunneling field-effect transistors," *J. Appl. Phys.*, vol. 108, no. 10, pp. 104503-1–104503-7, 2010.

[31] J. Z. Huang, W. C. Chew, Y. Wu, and L. J. Jiang, "Methods for fast evaluation of self-energy matrices in tight-binding modeling of electron transport systems," *J. Appl. Phys.*, vol. 112, no. 1, pp. 013711-1–013711-7, 2012.

[32] M. P. Anantram, M. S. Lundstrom, and D. E. Nikonov, "Modeling of nanoscale devices," *Proc. IEEE*, vol. 96, no. 9, pp. 1511–1550, Sep. 2008.

[33] H. H. B. Sørensen, P. C. Hansen, D. E. Petersen, S. Skelboe, and K. Stokbro, "Krylov subspace method for evaluating the self-energy matrices in electron transport calculations," *Phys. Rev. B*, vol. 77, no. 15, pp. 155301-1–155301-12, 2008.

[34] M. Shin, S. Lee, and G. Klimeck, "Computational study on the performance of Si Nanowire pMOSFETs based on the $k \cdot p$ method," *IEEE Trans. Electron Devices*, vol. 57, no. 9, pp. 2274–2283, Sep. 2010.

[35] N. Cavassilas, N. Pons, F. Michelini, and M. Bescond, "Multiband quantum transport simulations of ultimate p-type double-gate transistors: Influence of the channel orientation," *Appl. Phys. Lett.*, vol. 96, no. 10, pp. 102102-1–102102-3, Mar. 2010.

[36] M. Kobayashi and T. Hiramoto, "Experimental study on quantum confinement effects in silicon nanowire metal-oxide-semiconductor field-effect transistors and single-electron transistors," *J. Appl. Phys.*, vol. 103, no. 5, p. 053709, 2008.

[37] M. T. Björk, H. Schmid, J. Knoch, H. Riel, and W. Riess, "Donor deactivation in silicon nanostructures," *Nature Nanotechnol.*, vol. 4, pp. 103–107, Feb. 2009.

Jun Z. Huang received the M.S. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2010.

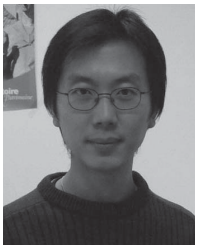
His current research interests include full-quantum and compact modeling of nanoscale devices.

Weng Cho Chew (S'79–M'80–SM'86–F'93) received the Academic degrees with MIT.

His current research interests include wave and field physics.

Jie Peng received the Ph.D. degree in physics from Sun Yat-Sen University, Guangdong, China, in 2007.

She joined the Department of Chemistry, University of Hong Kong, Hong Kong, in 2010, as a Post-Doctoral Fellow.



Chi-Yung Yam received the B.Sc. and Ph.D. degrees in chemistry from the University of Hong Kong, Hong Kong.

His current research interests include quantum transport and electronic device simulations.

Li Jun Jiang (S'01–M'04) received the Ph.D. degree from the University of Illinois at Urbana-Champaign, Urbana, IL, USA, in 2004.

His current research interests include electromagnetics, EMC/EMI, antennas, multidisciplinary EDA solutions, RF and microwave technologies, and high performance computing.

Guan-Hua Chen received the Ph.D. degree in physics from the California Institute of Technology, Pasadena, CA, USA, in 1992.

His current research interests include the development of first-principles method for open systems and O(N) first-principles methods for excited states.