# Model Order Reduction for Quantum Transport Simulation of Band-To-Band Tunneling Devices

Jun Z. Huang, Lining Zhang, *Member, IEEE*, Weng Cho Chew, *Fellow, IEEE*, Chi-Yung Yam, Li Jun Jiang, *Senior Member, IEEE*, Guan-Hua Chen, and Mansun Chan, *Fellow, IEEE*

***Abstract*— Simulations of nanoelectronic devices with nonequilibrium Green's function are computationally very intensive, in particular, when combined with multiband approaches, such as the $\mathbf{k}\cdot\mathbf{p}$ methods. To reduce the cost and make the simulation of realistic devices tractable, we have developed a model order reduction method for the simulation of hole transport in silicon nanowires using three- and six-band $\mathbf{k}\cdot\mathbf{p}$ models. It is shown in this paper that, with a spurious band elimination process, the method can be readily extended to the eight-band case that enables us to simulate band-to-band tunneling devices. The method is demonstrated via constructing reduced models for indium arsenide (InAs) nanowires and simulation of *I–V* characteristics of InAs tunneling field-effect transistors (TFETs). The results indicate that significant model reduction can be achieved with good accuracy retained. The method is then applied to study InAs TFETs with different channel orientations and source-pocket TFETs with n-p-i-p doping profiles.**

***Index Terms*— Band-to-band tunneling (BTBT), eight-band $\mathbf{k}\cdot\mathbf{p}$ model, indium arsenide (InAs) nanowires, model order reduction (MOR), nonequilibrium Green's function (NEGF), source-pocket TFETs, tunneling field-effect transistors (TFETs).**

## I. Introduction

BAND-to-band tunneling (BTBT) is a very interesting quantum phenomenon in electronic device applications. It accounts for a portion of the leakage current in the subthreshold region of carbon nanotube field-effect transistors (FETs) [1]. It has also been used to build novel devices, such as tunneling diodes [2] and tunneling FETs (TFETs) [3]. TFETs are energy-efficient switches since their subthreshold swing (SS) can be less than 60 mV/decade at room temperature [4]. This is impossible for conventional FETs, which are based on thermal injection. Therefore, TFET has been selected by ITRS as a very attractive candidate device for future low-power applications [5].

The BTBT process can be accurately simulated by combining nonequilibrium Green's function (NEGF) approach [6] with tight binding [7] or eight-band $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian [8] that describes both the conduction and valence bands. Unfortunately, these multiband NEGF studies require huge computational resource. To improve their efficiency, equivalent but greatly reduced tight-binding models are constructed for silicon nanowires (SiNWs) [9], which greatly speed up the simulation of p-type SiNW FETs even in the presence of inelastic scattering. For multiband $\mathbf{k}\cdot\mathbf{p}$ models, a mode space approach is proposed to simulate p-type SiNW FETs and indium arsenide (InAs) TFETs [8]. However, it selects the modes only at the $\Gamma$ point, i.e., at $k = 0$, which is inefficient to expand the modes that are far away from $k = 0$.

For three- and six-band $\mathbf{k}\cdot\mathbf{p}$ models, as is shown in [10], by sampling the Bloch modes at multiple points in the $k$ space and (or) $E$ space, a significantly reduced Hamiltonian can be constructed that describes very well the valence band top, based on which p-type SiNW FETs are simulated with good accuracy and efficiency. However, direct extension of this method to eight-band $\mathbf{k}\cdot\mathbf{p}$ model fails. The problem is that the reduced model constructed by multipoint expansion generally leads to some spurious bands, in addition to the normal bands, a situation similar to constructing the equivalent tight binding models [9] and rendering the reduced model useless. Therefore, it is essential to have these spurious bands eliminated, meanwhile the accuracy of the normal bands is retained.

Section II begins with a review of the eight-band $\mathbf{k}\cdot\mathbf{p}$ approach. Then, the model order reduction (MOR) method for the discretized NEGF equation is outlined, together with some discussion of the discretization. The problem of spurious bands and the procedure to eliminate them are addressed in detail. The method is validated by checking the band structures as well as the *I–V* curves. The usefulness of this method is demonstrated in Section III by simulating TFETs with different crystalline orientations and with source pockets. The conclusion is given in Section IV.

Manuscript received September 14, 2013; revised November 20, 2013; accepted December 17, 2013. Date of current version January 20, 2014. This work was supported in part by the Research Grants Council of Hong Kong under Grant GRF 712612, Grant GRF 711511, and Grant GRF 713011, in part by the University Grants Council of Hong Kong under Contract AoE/P-04/08, and in part by the Chinese Grant NSF 61271158. The review of this paper was arranged by Editor A. Schenk.

J. Z. Huang was with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong. He is now with the Network for Computational Nanotechnology, School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: junhuang1021@gmail.com).

L. Zhang and M. Chan are with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong (e-mail: lnzhang@ust.hk; mchan@ust.hk).

W. C. Chew is with the Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, IL 61801 USA (e-mail: wchew@uiuc.edu).

C.-Y. Yam is with the Beijing Computational Science Research Center, Beijing 100084, China (e-mail: yamcy@yangtze.hku.hk).

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

G.-H. Chen is with the Department of Chemistry, The University of Hong Kong, Hong Kong (e-mail: ghc@yangtze.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.2013.2295983

0018-9383 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

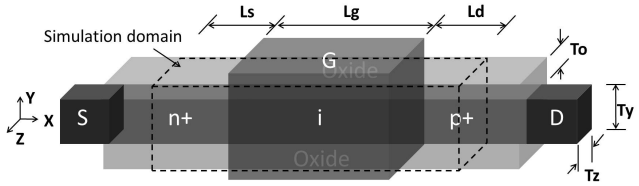



Fig. 1. GAA InAs TFET with n-i-p (or p-i-n) type doping profile. The transport direction is $x$. The source, channel, and drain lengths are $L_s$, $L_g$, and $L_d$, respectively. Nanowire thickness is $T_y$ and $T_z$. Oxide layer thickness is denoted by $T_o$ and dielectric constant is $\varepsilon_{\rm ox}$.

## II. Theory and Method

The gate-all-around (GAA) InAs nanowire TFET to be simulated is shown in Fig. 1. The InAs nanowire is n-type (p-type) doped in the source and p-type (n-type) doped in the drain, while it is intrinsic in the channel. The InAs nanowire is surrounded by the oxide layer, through which the gate controls the channel portion. The working principle of this device is based on BTBT, as described in [3] and [4]. InAs is chosen as the channel material because high ON-current is possible due to its small direct bandgap and light effective masses [7].

### A. Eight-Band $k \cdot p$ Approach

To describe the band structure involving both the conduction and valence bands of III–V compound semiconductor materials, a widely used approach is the eight-band $\mathbf{k} \cdot \mathbf{p}$ model. When the eight basis functions are chosen to be spinup and spin-down $s$ and $p$ atomic orbital-like states, the Hamiltonian can be written as [11], [12]

$$\mathbf{H}^8(\mathbf{k}) = \begin{bmatrix} \mathbf{G}(\mathbf{k}) & \Gamma \\ -\Gamma^* & \mathbf{G}(\mathbf{k})^* \end{bmatrix}. \tag{1}$$

The matrix $\mathbf{G}(\mathbf{k})$ is defined as

$$\mathbf{G}(\mathbf{k}) = \mathbf{G}_1(\mathbf{k}) + \mathbf{G}_2(\mathbf{k}) + \mathbf{G}_{\rm so} \tag{2}$$

where

$$\mathbf{G}_1(\mathbf{k}) = \begin{pmatrix} E_g & ik_xP & ik_yP & ik_zP \\ -ik_xP & -\Delta/3 & 0 & 0 \\ -ik_yP & 0 & -\Delta/3 & 0 \\ -ik_zP & 0 & 0 & -\Delta/3 \end{pmatrix} \tag{3}$$

$$\mathbf{G}_2(\mathbf{k}) = \begin{pmatrix} Ak^2 & Bk_yk_z & Bk_xk_z & Bk_xk_y \\ Bk_yk_z & Mk^2 + \widetilde{L}k_x^2 & Nk_xk_y & Nk_xk_z \\ Bk_zk_x & Nk_xk_y & Mk^2 + \widetilde{L}k_y^2 & Nk_yk_z \\ Bk_xk_y & Nk_xk_z & Nk_yk_z & Mk^2 + \widetilde{L}k_z^2 \end{pmatrix} \tag{4}$$

and

$$\mathbf{G}_{\rm so} = \frac{\Delta}{3}\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -i & 0 \\ 0 & i & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{5}$$

The matrix $\Gamma$ is

$$\Gamma = \frac{\Delta}{3}\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -i \\ 0 & -1 & i & 0 \end{pmatrix}. \tag{6}$$

Here, the parameter $E_g$ is the bandgap, $\Delta$ is the spin-orbit splitoff energy, and $P$ is proportional to the momentum matrix element and can be evaluated by its equivalent energy $E_p$. The parameter $A$ is determined from the conduction band effective mass and $B$ is set to be zero. $\widetilde{L} = L - M$ has been used to shorten the expression. The parameters $L$, $M$, and $N$ are related to the Luttinger parameters. For more discussion of the parameters, see [11] and [12].

For nanostructures, the periodicity is broken by the finite sizes and the external potentials. The wave function can be found by solving the following coupled differential equation for envelop function $F_m$ $(m = 1, 2, \ldots, 8)$:

$$\sum_{n=1}^{8}\left[\mathbf{H}_{mn}^8(-i\nabla) + V(\mathbf{r})\,\delta_{mn}\right]F_n(\mathbf{r}) = EF_m(\mathbf{r}) \tag{7}$$

where $V(\mathbf{r})$ is the slowly varying perturbed potential distribution and operator $\mathbf{H}_{mn}^8(-i\nabla)$ is the element of $\mathbf{H}^8(\mathbf{k})$ with $\mathbf{k}$ replaced by the differential operator $-i\nabla$.

The parameters of bulk InAs material [11] are used in this paper, except that the parameter $E_p$ is reduced to 18 eV according to [13]. Note that the adjustment of the bulk parameters may be needed to match other band structure models, such as in [14].

### B. Model Order Reduction

To solve (7) numerically with the NEGF approach, the operator needs to be discretized first. For reasons that will be stated later, finite-difference method (FDM) is adopted in the transport direction, whereas k-space discretization is employed in the transverse directions [8]. For simplicity, hard wall boundary condition is assumed at the interfaces between the oxide layer and the InAs channel. The resultant matrix equation for Green's function $\mathbf{G}(E)$ can be written as

$$[E\mathbf{I} - \mathbf{H}_0 - \Sigma(E)]\,\mathbf{G}(E) = \mathbf{I} \tag{8}$$

where $\mathbf{H}_0$ is the discretized $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian of the isolated device (including the potential term) and $\Sigma$ is the self-energy matrix due to the semi-infinite leads. For nanowire structures, such as Fig. 1, $\mathbf{H}_0$ can be written down layer by layer leading to block tridiagonal form (block size $N_t$).

As this equation can be large, to solve it efficiently for many different energy $E$, a reduced-order matrix equation can be constructed

$$\left[E\widetilde{\mathbf{I}} - \widetilde{\mathbf{H}}_0 - \widetilde{\Sigma}(E)\right]\widetilde{\mathbf{G}}(E) = \widetilde{\mathbf{I}} \tag{9}$$

and the reduced-order Green's function $\widetilde{\mathbf{G}}(E)$ is to be solved. Here, the reduced Hamiltonian, self-energy, and Green's function are

$$\begin{aligned} \widetilde{\mathbf{H}}_0 &= \mathbf{U}^\dagger\mathbf{H}_0\mathbf{U}, \quad \widetilde{\Sigma}(E) = \mathbf{U}^\dagger\Sigma(E)\,\mathbf{U} \\ \widetilde{\mathbf{G}}(E) &= \mathbf{U}^\dagger\mathbf{G}(E)\,\mathbf{U} \end{aligned} \tag{10}$$

where $\mathbf{U}$ is a block-diagonal unitary matrix containing the reduced basis $\mathbf{V}_i$ (with dimension $N_t \times N_m$, where $N_m$ is the number of reduced basis) for each layer $i$. Then, the problem is how to construct this transformation matrix $\mathbf{U}$ so that the reduced system is as small as possible, and yet it still accurately describes the original system.

To construct the reduced basis $\mathbf{V}_i$ for layer $i$, the Hamiltonian of layer $i$ is repeated to form an infinite periodic nanowire. The reduction comes from the fact that only the electrons near the conduction band bottom and valence band top are important in the transport process. To approximate the band structure over that small region, $\mathbf{V}_i$ then consists of the Bloch modes with energy lying in that region. Multiple-point construction based on $k$ space sampling and (or) $E$ space sampling can be employed [10]. Here, $k$ space sampling is adopted since $E$ space sampling is more costly and that the eight-band matrix is much larger than the six- or three-band case. Before going to the examples, there is a need to discuss the discretization because solving the Bloch modes for each layer is itself very costly when $N_t$ is large.

### C. Discretization

In [10], FDM is adopted and it results in extremely sparse matrices. Therefore, the Bloch modes can be obtained efficiently with sparse matrix solvers. With shift-and-invert strategy implemented, the Krylov subspace-based eigenvalue solver converges very quickly, as the interested eigenvalues (close to the valence band top) distribute in a very small area. However, it is found that the Krylov subspace method is less efficient in the eight-band case. The reason is that the interested eigenvalues distribute over a larger area, as both conduction and valence bands are of interest and between them there is a bandgap.

Therefore, the method used in [8] is adopted. In that method, the transport direction is still discretized by FDM, whereas the transverse directions are discretized by spectral method. Spectral method has high spectral accuracy (i.e., the error decreases exponentially with the increase of discretization points $N$) if the potential distribution is smooth [15]. This is true for devices that do not have any explicit impurities or surface roughness. Therefore, the Hamiltonian matrix size of a layer, i.e., $N_t$, can be kept very small (although it is less sparse or even dense), making direct solution of the eigenvalue problem possible. The discretized form valid for arbitrary nanowire orientation is provided in the Appendix.

### D. Spurious Band Elimination

As an example, Fig. 2(a) plots the $E$–$k$ dispersion for an ideal InAs nanowire orientated in the [100] direction. Fig. 2(b) shows the result using the reduced Hamiltonian $\widetilde{\mathbf{H}}_0$. The reduced basis $\mathbf{V}_i$ ($i$ is arbitrary here) is constructed by sampling the Bloch modes evenly in the Brillouin zone (at $k = 0, \pm\pi/4, \pm 2\pi/4$, and $\pm 3\pi/4$ [1/nm]), with the energy $E \in [-0.5\text{eV}, 1.7\text{ eV}]$, which results in $N_m = 262$ modes. Note that the modes at negative $k$ can be obtained by a transformation of those at positive $k$ [10]. Clearly, the reduced Hamiltonian reproduces quite well the dispersion bands in that energy window (except at the very bottom, which can be improved by sampling a slightly larger energy window or more $k$ points), demonstrating that the $k$ space sampling is effective. However, there are also some spurious bands appearing in the conduction and valence bands, and even in the bandgap, making the reduced model useless. Moreover, different sampling points or windows would change the number and position of the spurious bands. This situation is not encountered in the three- or six-band model involving only the valence bands, or in the one-band effective mass model involving the conduction band only. It should be caused by the coupling between the conduction and valence bands, which makes the eight-band model indefinite. The coupling is important for materials with narrow bandgaps.

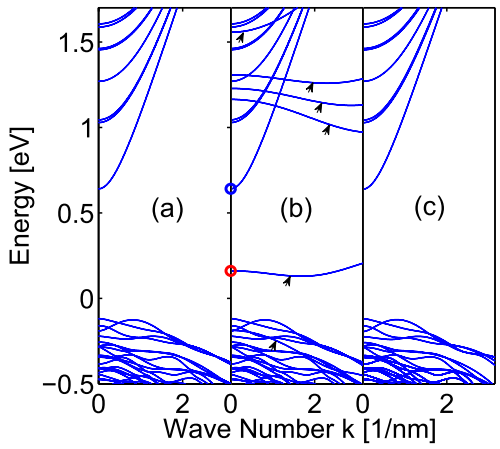
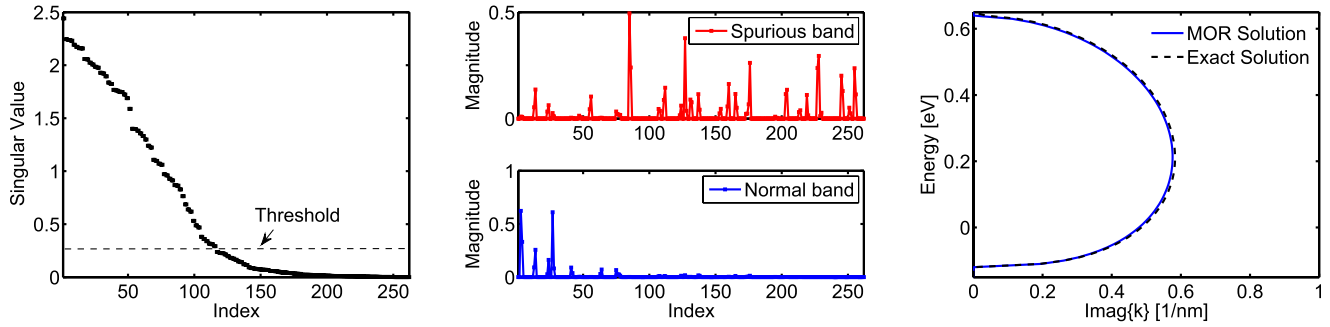



Fig. 2. $E$–$k$ diagram of a 5 nm × 5 nm InAs nanowire in [100] direction. The potential is assumed to be zero everywhere inside the nanowire. (a) Exact solution. (b) Reduced model solution with spurious bands marked. (c) Reduced model solution with spurious bands eliminated. Only $+k$ is shown as the band structure is symmetric with respect to $k = 0$.

The spurious bands must be suppressed. To this end, a singular value decomposition (SVD) is applied to the matrix $\mathbf{V}_i$. As plotted in Fig. 3(a), the singular values spread from a large value down to zero, suggesting there are some linearly dependent modes. These linearly dependent modes give rise to null space of the reduced model and therefore must be removed. It is further found that the normal bands are mainly contributed by singular vectors having large singular values, in contrast to the spurious bands where singular vectors with small singular values have large contribution. An example of this is shown in Fig. 3(b). By removing the vectors with small singular values, i.e., vectors with $\upsilon \leq \upsilon_{\text{th}}$, where $\upsilon_{\text{th}} = 0.25$ is the threshold, a new reduced basis $\widetilde{\mathbf{V}}_i$ is generated with $\widetilde{N}_m = 116$. Using this new reduced basis, a new reduced Hamiltonian is constructed with its $E$–$k$ diagram shown in Fig. 2(c). It is observed that all the spurious bands have been eliminated but at the cost of a slightly compromised accuracy.

For the BTBT process, the evanescent dispersion inside the bandgap is particularly important, and it is thus plotted in Fig. 3(c). Only the smallest imaginary $k$ is plotted, since evanescent waves decay exponentially and thus higher modes' contribution to the tunneling can be neglected. As can be observed, the MOR solution (after the spurious band elimination) agrees well with the exact solution.

The value of $\upsilon_{\text{th}}$ is found to be crucial. A small $\upsilon_{\text{th}}$ might be insufficient to remove all the spurious bands, whereas a large $\upsilon_{\text{th}}$ may degrade the accuracy severely. Moreover, adjustment of $\upsilon_{\text{th}}$ may be required when different sampling




(a) (b) (c)

Fig. 3. (a) Distribution of singular values of matrix $\mathbf{V}_i$. (b) Contribution of the singular vectors to the two modes at $k = 0$: one is the spurious mode with $E = 0.16$ eV [red circle in Fig. 2(b)] and the other is the normal mode with $E = 0.64$ eV [blue circle in Fig. 2(b)]. (c) Evanescent $E$–$k$ diagram in the bandgap of a 5 nm × 5 nm InAs nanowire in the [100] direction. Only the band with the smallest imag$\{k\}$ is shown.

points or sampling energy windows are used. To determine $\upsilon_{\text{th}}$ automatically, we propose a search process as follows.

1) Sample enough Bloch modes and store them in matrix $\mathbf{B}$. Suppose $I$ points are sampled in the $k$ space, and $m_i$ modes with energy $E \in [E_1, E_2]$ are obtained at the $i$th point $k_i$ $(1 \le i \le I)$, then the size of matrix $\mathbf{B}$ is $N_t \times N_m$, where $N_m = \sum_{i=1}^{I} m_i$.
2) Do SVD of $\mathbf{B}$, i.e., $\mathbf{B} = \mathbf{U}\Sigma\mathbf{V}^{\dagger}$.
3) Set an initial value for $\upsilon_{\text{th}}$. Let us use $\upsilon_{\text{th}} = 0$ here.
4) Use $\upsilon_{\text{th}}$ to construct a reduced basis $\widetilde{\mathbf{U}}$ by removing the singular vectors with $\upsilon < \upsilon_{\text{th}}$ in $\mathbf{U}$. The size of $\widetilde{\mathbf{U}}$ will be $N_t \times \widetilde{N}_m$.
5) Use $\widetilde{\mathbf{U}}$ to build a reduced Hamiltonian $\widetilde{\mathbf{H}}$. For each layer of $\widetilde{\mathbf{H}}$, the size will be $\widetilde{N}_m \times \widetilde{N}_m$.
6) Solve the $E$–$k$ relation of $\widetilde{\mathbf{H}}$ for certain $k_i$, obtaining $\widetilde{m}_i$ modes with $E \in [E_1, E_2]$. It is found that $k_i = 0$ is a good choice.
7) If $\widetilde{m}_i > m_i$ (which means that there are still some spurious bands), increase $\upsilon_{\text{th}}$ appropriately and go back to step 4. Otherwise, stop.

The above search process is fast, since steps 5 and 6 are much cheaper than step 1 although they have to be repeated many times. The complexity of step 1 is $I \times O\left(N_t^3\right)$, step 2 is $O\left(N_t N_m^2\right)$, step 5 is $O\left(\widetilde{N}_m N_t^2\right)$, and step 6 is $O\left(\widetilde{N}_m^3\right)$. Note that $\widetilde{N}_m < N_m < N_t$.

The $\upsilon_{\text{th}} = 0.25$ used earlier is the result of the above search process. It has also been tested for energy windows $[-0.4$ eV, 1.5 eV] and $[-0.6$ eV, 1.9 eV], and for [110] and [111] directions, with good results obtained (not shown here). It should be mentioned that this process results in a smaller basis set, which is different from the method for tight binding models [9], where the basis is enlarged to eliminate the spurious modes. Further investigations are needed to check whether this method also applies to various tight binding models.

### E. Error Analysis

Now, this reduced model can be applied to simulate a TFET, as shown in Fig. 1. NEGF and Poisson equations are solved self-consistently. Phonon scattering has a very modest effect on the $I$–$V$ curve [16] and thus is excluded in this paper. The charge density involving both the electrons and the holes is calculated by the method in [17]. To improve the efficiency, the reduced basis is constructed for an ideal nanowire with its potential term set to zero (so the basis just needs to be solved in one layer and it remains unchanged during the self-consistent iterations). The potential term in real devices then merely causes transitions between these scattering states. This assumption has been adopted in [9] with good accuracy demonstrated. As will be shown soon, it is also a fairly good approximation for the GAA nanowire TFET here.

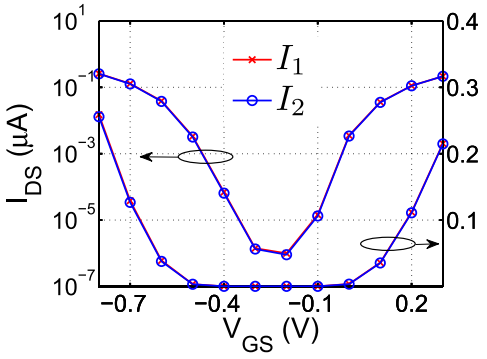
The $I_{\text{DS}}$–$V_{\text{GS}}$ transfer characteristics are obtained and plotted in Fig. 4(a), in both linear and logarithm scales. To check how large the sampling energy window is sufficient to produce the correct $I$–$V$ curve, $[-0.4$ eV, 1.5 eV] and $[-0.5$ eV, 1.7 eV] are tried, which result in $I_1$ and $I_2$. It is observed that $I_1$ is very close to $I_2$, showing that the results have converged. The relative errors of the two sets of currents are shown in Fig. 4(b). It is observed that the relative errors are very small for the region near ON-state, implying that the reduced basis is sufficient despite that, at ON-state, the longitudinal potential varies very rapidly around the junctions. This also validates the approximation made above. However, the errors are larger in the subthreshold region (up to 10%). The reason we believe is that the tunneling path is longer in this region and thus the tunneling current is very sensitive to the (evanescent) band structure errors induced by different sampling windows. In the following, energy window $[-0.5$ eV, 1.7 eV] will be used.

## III. Applications

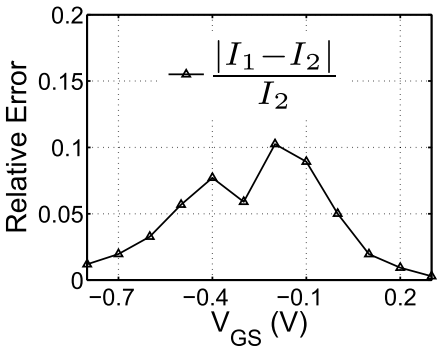
### A. Different Channel Orientations

Fig. 5(a) compares the $I_{\text{DS}}$–$V_{\text{GS}}$ characteristics of InAs nanowire TFETs oriented in the [100], [110], and [111] directions. It is observed that [100] has the best SS but the smallest ON-current; [111] has the worst SS but the largest ON-current. In addition, SS less than 60 mV/decade is observed in Fig. 4(a), but it is not observed here due to a shorter channel used. Note that the curves are not linear in the subthreshold region and several ways to define SS have been proposed [18].

To explain it, Fig. 5(b) compares the lowest evanescent $E$–$k$ relations in the bandgap for the three cases. It is found that

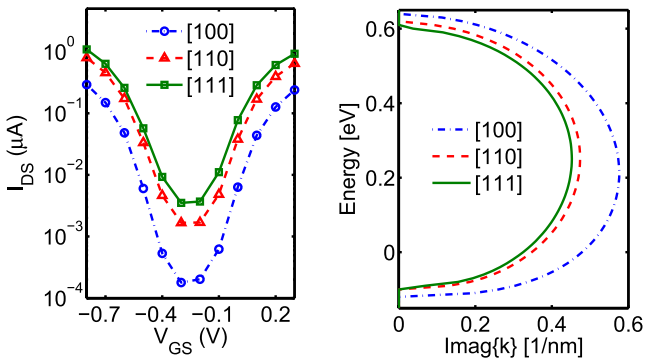



(a)




(b)

Fig. 4. (a) $I_{DS}$–$V_{GS}$ transfer characteristics of an n-i-p TFET, as shown in Fig. 1. The nanowire is oriented in the [100] direction. $T_{ox} = 1$ nm, $T_y = T_z = 5$ nm, $L_g = L_s = L_d = 15$ nm, and $\varepsilon_{ox} = 12.7$. The doping density is equal to $5 \times 10^{19}$ cm$^{-3}$ at both the source and the drain. The drain bias is fixed to $V_{DS} = -0.3$ V. $I_1$ and $I_2$ are obtained by sampling energy windows [−0.4 eV, 1.5 eV] and [−0.5 eV, 1.7 eV] (both at $k = 0$, $\pm\pi/4$, $\pm 2\pi/4$, and $\pm 3\pi/4$ [1/nm]) leading to $\widetilde{N}_m = 90$ and $\widetilde{N}_m = 116$. (b) Relative errors of the two sets of currents.




(a) (b)

Fig. 5. (a) $I_{DS}$–$V_{GS}$ curves of n-i-p InAs TFETs oriented in the [100], [110], and [111] directions. $L_g = 10$ nm, the other device parameters are the same as those in Fig. 4. (b) Evanescent $E$–$k$ relations in the bandgap, only the band with the smallest imag$\{k\}$ is shown.

[100] ([111]) has the largest (smallest) $\Im m\{k\}$ leading to the smallest (largest) tunneling probability. Denoting the tunneling length as $L_n$ ($L_f$) for the ON (OFF) state, the ON/OFF ratio can be estimated by Wentzel-Kramer-Brillouin method with uniform electric field approximation as $\exp\left[k_I\left(L_f - L_n\right)\right]$, where $k_I = \Im m\{k\}$. This means that: 1) large $k_I$ has large ON/OFF ratio when $L_f - L_n$ is fixed, which is the case for the [100] direction and 2) the ON/OFF ratio increases with $L_f - L_n$ and the increasing speed is larger for larger $k_I$, so in practice long channel (with long $L_f$) is employed to increase the ON/OFF ratio and [100] is expected to have much better ON/OFF ratio than the other directions when the channel is very long.

### *B. Source-Pocket TFETs*

Many TFETs suffer from low ON-current [4]. It is theoretically predicted and experimentally demonstrated that the source-pocket TFETs have significantly improved ON-current and steeper SS over the classical TFETs [19], [20]. In addition, the significantly degraded linear region $I_{DS}$–$V_{DS}$ characteristics of classical TFETs [21] can be improved by incorporating the source pocket [22]. The source-pocket TFET is formed by inserting a thin layer of p-type (n-type) doping between the n-type (p-type) source and the intrinsic channel, which results in an n-p-i-p (p-n-i-n) doping profile. Most of the simulations were based on semiclassical methods [19], [20]. Recently, 2-D quantum simulations were performed for all-Si and all-Ge double-gate structures [23], confirming the semiclassical simulation results. Here, we investigate InAs nanowire-based source-pocket TFETs for the first time, using 3-D quantum simulations.

Fig. 6(a) shows the $I_{DS}$–$V_{GS}$ of the n-p-i-p TFETs with three different pocket lengths, in comparison with the n-i-p one without source pocket. For the right part of the curve (due to the ambipolar nature of TFETs), as expected, these pockets have negligible influence on the turn-on property, since the conduction there is via tunneling through the drain junction. However, a better SS is observed due to the lower OFF-current at $V_{GS} = -0.2$ V. For the left part of the $I$–$V$ curve, which is of interest, the pockets merely shift the threshold voltage and less negative gate voltage can now turn the device on. However, the SS remains almost unchanged, in contrast to [19], [20], and [23].

Fig. 6(b) and (c) compare the band diagrams of the TFETs without and with the source pocket, at OFF- and ON-state, respectively. It is observed that the source pocket enhances the band bending at the source junction, which leads to longer source-to-drain direct tunneling path but shorter source-to-channel junction tunneling path. As the OFF (ON) state current is dominated by direct (junction) tunneling, the source pocket will decrease (increase) the OFF (ON) state current and thus improve the SS. However, the band bending of the n-i-p structure at the source-channel junction is already very abrupt; the insertion of source pockets does not improve it much. The reason is that, for the n-i-p structure here, the channel is fully controlled by the gate due to the small cross-sectional nanowire, high-$k$ gate oxide, and GAA geometry used, making the lateral electric field between the source and the channel very strong. While in [19], [20], and [23], thick silicon layer and single (or double) gate geometry were used, making the source pocket very effective in band bending improvement. To verify this, we plot in Fig. 7(a) the $I$–$V$ curves when the gate oxide is reduced to $\varepsilon_{ox} = 3.8$. Now the source pocket has

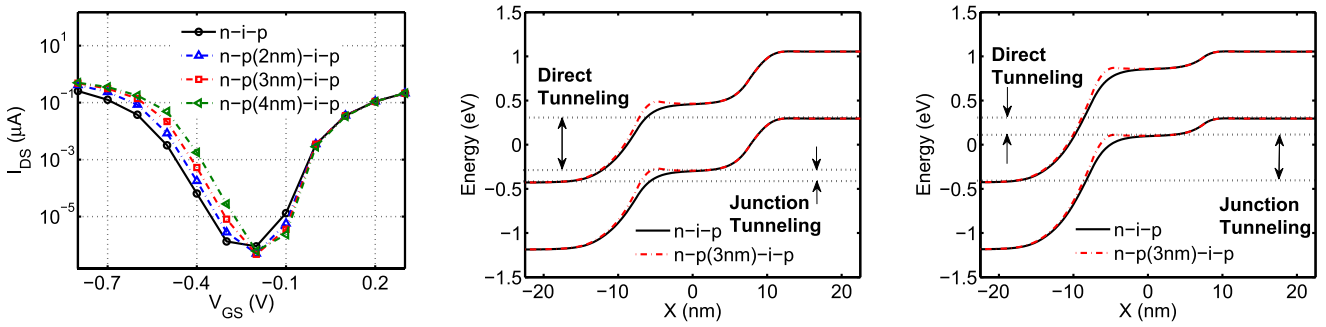



Fig. 6. Comparison of source-pocket n-p-i-p TFETs with different pocket lengths and n-i-p one without pocket. (a) $I_{DS}$–$V_{GS}$ curves. (b) Band diagrams at OFF-state ($V_{GS} = -0.2$ V). (c) Band diagrams at ON-state ($V_{GS} = -0.6$ V). The pocket doping density is $5 \times 10^{19}$ cm$^{-3}$, the other device settings are the same as those in Fig. 4. The source Fermi level is 0 eV, whereas the drain Fermi level is −0.3 eV.

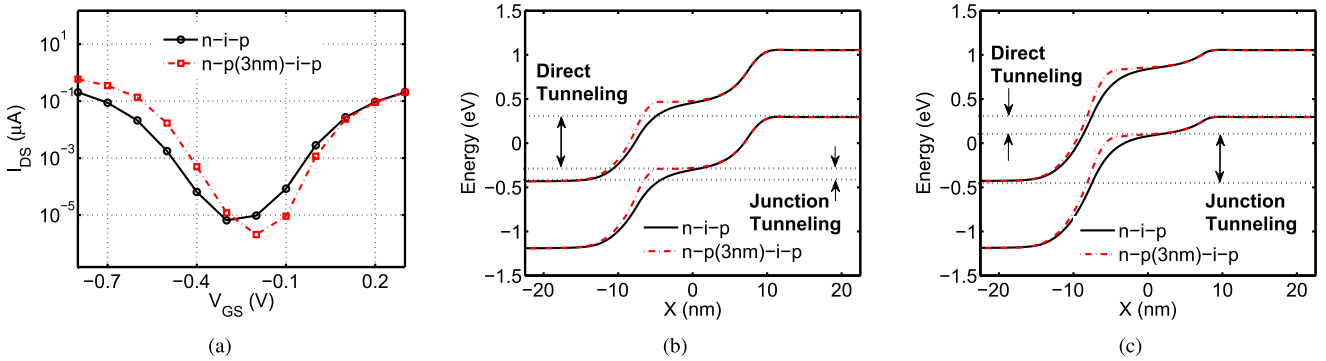



Fig. 7. Same plots as Fig. 6, except that $\varepsilon_{ox}$ is reduced from 12.7 to 3.8.

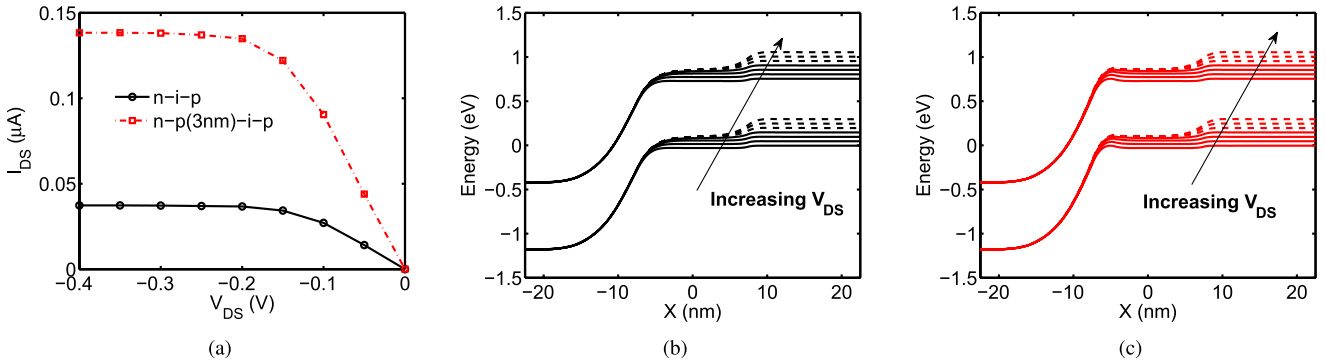



Fig. 8. (a) $I_{DS}$–$V_{DS}$ curve of source-pocket n-p-i-p TFET with 3-nm pocket length, comparing with n-i-p one without pocket. (b) Band diagrams of n-i-p TFET and (c) source-pocket n-p-i-p TFET with 3-nm pocket length for $V_{DS}$ varying from 0 V to −0.3 V with −0.05 V step. Solid (dashed) lines: the linear (saturation) region. The gate bias is fixed to $V_{GS} = -0.6$ V, the other device settings are the same as those in Fig. 6.

a larger impact on the curve and steeper SS is observed. This is supported by the band diagrams plotted in Fig. 7(b) and (c), which show that the band bending improvement is more significant than Fig. 6(b) and (c).

Fig. 8(a) shows the $I_{DS}$–$V_{DS}$ of the n-p-i-p TFET with 3-nm pocket length, in comparison with the n-i-p one. The quasi-linear dependence of $I_{DS}$ on small $V_{DS}$ is observed here for both cases, in agreement with [22] for the TFETs with source pockets, but in contrast to the exponential dependence observed for the p-i-n TFETs [21].

Fig. 8(b) and (c) compare the band diagrams by varying $V_{DS}$, for the TFETs without and with the source pocket, respectively. When $V_{DS}$ is small, as can be observed, the potential in the channel changes with $V_{DS}$. This should modulate the source-to-channel tunneling width leading to exponential change of the tunneling probability (and current), as the case in [21]. However, here, the tunneling width is almost unchanged, which is again due to the very abrupt source-channel band bending enabled by the good electrostatic integrity. The change of $V_{DS}$ largely affects the tunneling

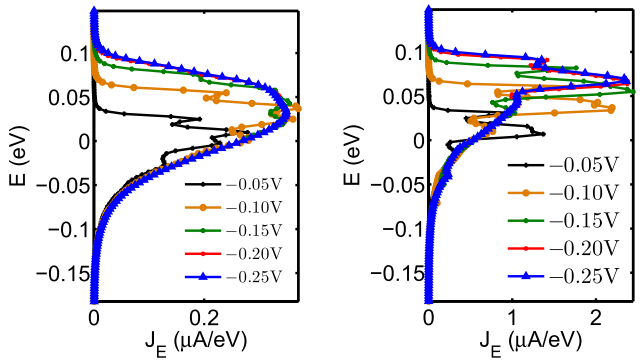



Fig. 9. (a) Current spectra of n-i-p TFET and (b) n-p-i-p TFET with 3-nm pocket length for $V_{DS}$ varying from −0.05 V to −0.25 V with −0.05 V step. Please refer to Fig. 8 for the device setting.

window leading to the quasi-linear behavior. When $V_{DS}$ is large, the potential in the channel (and thus the source-channel tunneling junction) no longer changes with $V_{DS}$ and the current saturates. The current spectra plotted in Fig. 9 confirm these observations.

## IV. Conclusion

An MOR method is developed for efficiently solving the NEGF equations employing the eight-band $\mathbf{k}\cdot\mathbf{p}$ model. By introducing a spurious band elimination process, reduced models can be constructed for reproducing the band structures in any energy window near the bandgap. The reduced models can also capture the $I$–$V$ characteristics of TFETs with acceptable accuracy. InAs TFETs with different channel orientations are compared, it is found that [100] direction has better SS but smaller ON-current than [110] and [111] directions, due to its larger imaginary wave vectors. Source-pocket TFETs with n-p-i-p doping profiles are also studied, it is observed that band bending at the source-to-channel junction is enhanced by the source pocket, which results in better ON/OFF ratio, SS, and output behaviors. However, such effects tend to be diminished when the electrostatic integrity of the devices is improved.

## Appendix
## Discretization of the k · p Hamiltonian in the Fourier Space

The $\mathbf{k}\cdot\mathbf{p}$ operator involves various differential operators, and thus the discretization usually results in many terms, particularly when the number of bands is large. To have a discretized form that is compact and valid for arbitrary nanowire orientation, we rewrite the eight-band $\mathbf{k}\cdot\mathbf{p}$ operator in (7) as [24]

$$\begin{aligned}\mathbf{H}\left(\frac{\partial}{\partial x},\frac{\partial}{\partial y},\frac{\partial}{\partial z}\right) = &-\mathbf{H}_{xx}\frac{\partial^2}{\partial x^2}-\mathbf{H}_{yy}\frac{\partial^2}{\partial y^2}-\mathbf{H}_{zz}\frac{\partial^2}{\partial z^2}\\ &-\mathbf{H}_{xy}\frac{\partial^2}{\partial x\partial y}-\mathbf{H}_{yz}\frac{\partial^2}{\partial y\partial z}-\mathbf{H}_{zx}\frac{\partial^2}{\partial z\partial x}\\ &+\mathbf{H}_x\frac{\partial}{\partial x}+\mathbf{H}_y\frac{\partial}{\partial y}+\mathbf{H}_z\frac{\partial}{\partial z}+\mathbf{H}_0+\mathbf{H}_{so}\end{aligned} \tag{A-1}$$

where the matrices $\mathbf{H}_0$, $\mathbf{H}_{so}$, $\mathbf{H}_x$, $\mathbf{H}_y$, $\mathbf{H}_z$, $\mathbf{H}_{xx}$, $\mathbf{H}_{yy}$, $\mathbf{H}_{zz}$, $\mathbf{H}_{xy}$, $\mathbf{H}_{yz}$, and $\mathbf{H}_{zx}$ are the coefficients containing contributions from Löwdin's renormalization and spin-orbit interactions.

For nanowire directions other than [100], coordinate transformations have to be performed. For example, to obtain the coordinate for [110] direction, we rotate the coordinate of [100] direction in $xy$ plane by $\phi = \pi/4$. Similarly, for [111] direction, we continue rotating the coordinate of [110] in $yz$ plane by $\theta$ with $\sin\theta = 1/\sqrt{3}$. Once we have $\mathbf{k}$ in terms of $\mathbf{k}'$, we plug them into the Hamiltonian expression (1) to obtain the new Hamiltonian (now in $\mathbf{k}'$). It turns out that the new Hamiltonian operator can still be written in the form of (A-1), the only difference is that now we have different coefficient matrices $\mathbf{H}_i$ and $\mathbf{H}_{ij}$ $(i, j = x, y, z)$.

To discretize the operator (A-1), the longitudinal component of the unknown envelope function is discretized with second-order central FDM, while the transversal components are expanded using Fourier series [8]

$$\phi_{p,q}(y_m, z_n) = \frac{2}{\sqrt{N_y N_z}}\sin\left(k_p y_m\right)\sin\left(k_q z_n\right) \tag{A-2}$$

where $N_y$ and $N_z$ are the number of grid points in the $y$ and $z$ directions respectively, $m$ and $n$ $\left(1 \le m \le N_y, 1 \le n \le N_z\right)$ are the coordinates of the $R$th grid point in real space, $p$ and $q$ $\left(1 \le p \le N_y, 1 \le q \le N_z\right)$ are the coordinates of the $S$th grid point in the Fourier space

$$k_p = \frac{p\pi}{L_y}, \quad k_q = \frac{q\pi}{L_z} \tag{A-3}$$

where $L_y$ $(L_z)$ is the nanowire length in the $y$ $(z)$ direction.

Operating (A-1) on (A-2), multiplying the result with (A-2) and performing integrations, we get the discretized form. It is block tridiagonal, with $\mathbf{D}$ being the on-site Hamiltonian for layer $l$ $(1 \le l \le N_x)$ and $\mathbf{T}$ being the coupling Hamiltonian between adjacent layers $l$ and $l+1$ $(1 \le l \le N_x - 1)$, where $N_x$ is the number of grids in the longitudinal direction $x$.

The $\left(S, S'\right)$ block of $\mathbf{D}$ can be written down using the very simple prescription

$$\begin{aligned}\mathbf{D}_{S,S'} = &\left(\mathbf{H}_y\frac{4k'_p}{\pi}\frac{p}{p^2-p'^2}\right)\delta_{p+p',\text{odd}}\delta_{q,q'}\\ &-\left(\mathbf{H}_{yz}\frac{4k'_p}{\pi}\frac{p}{p^2-p'^2}\frac{4k'_q}{\pi}\frac{q}{q^2-q'^2}\right)\delta_{p+p',\text{odd}}\delta_{q+q',\text{odd}}\\ &+\left(\mathbf{H}_0+\mathbf{H}_{xx}\frac{2}{(\Delta x)^2}+\mathbf{H}_{yy}k_p^2+\mathbf{H}_{zz}k_q^2+\mathbf{H}_{so}\right)\delta_{p,p'}\delta_{q,q'}\\ &+\left(\mathbf{H}_z\frac{4k'_q}{\pi}\frac{q}{q^2-q'^2}\right)\delta_{q+q',\text{odd}}\delta_{p,p'}\end{aligned} \tag{A-4}$$

where $\Delta x$ is the grid spacing in FDM, $(p, q)$ and $\left(p', q'\right)$ are the coordinates of the $S$th and $S'$th grid points respectively, and $\delta$ is Kronecker delta function. For instance, $\delta_{q+q',\text{odd}}$ is equal to 1 (0) if $q + q'$ is an odd (even) number.

Similarly, the $(S, S')$ block of $\mathbf{T}$ can be written as

$$\mathbf{T}_{S,S'} = \left(-\mathbf{H}_{xx}\frac{1}{(\Delta x)^2} + \mathbf{H}_x\frac{1}{2\Delta x}\right)\delta_{p,p'}\delta_{q,q'} - \left(\mathbf{H}_{xy}\frac{1}{2\Delta x}\frac{4k'_p}{\pi}\frac{p}{p^2 - p'^2}\right)\delta_{p+p',\text{odd}}\delta_{q,q'} - \left(\mathbf{H}_{xz}\frac{1}{2\Delta x}\frac{4k'_q}{\pi}\frac{q}{q^2 - q'^2}\right)\delta_{q+q',\text{odd}}\delta_{p,p'}. \quad \text{(A-5)}$$

In this paper, we use $\Delta x = 0.125$ nm and have limited $S$ to be $1 \leq S \leq 183$ by employing the index scheme in [8].

## ACKNOWLEDGMENT

The authors would like to thank J. Sun, Y. Zhang, S. Koo for setting up the computer cluster, and J. Peng, Q. Chen, and S. Markov for helpful discussions.

## REFERENCES

[1] J. Deng and H.-S. Wong, "A compact SPICE model for carbon-nanotube field-effect transistors including nonidealities and its application—Part I: Model of the intrinsic channel region," *IEEE Trans. Electron Devices*, vol. 54, no. 12, pp. 3186–3194, Dec. 2007.

[2] D. A. Neamen, *Semiconductor Physics and Devices: Basic Principles*. New York, NY, USA: McGraw-Hill, 2003.

[3] J. Appenzeller, Y.-M. Lin, J. Knoch, Z. Chen, and P. Avouris, "Comparing carbon nanotube transistors—The ideal choice: A novel tunneling device design," *IEEE Trans. Electron Devices*, vol. 52, no. 12, pp. 2568–2576, Dec. 2005.

[4] A. M. Ionescu and H. Riel, "Tunnel field-effect transistors as energy-efficient electronic switches," *Nature*, vol. 479, no. 7373, pp. 329–337, Nov. 2011.

[5] (2013, Sep. 14). *ITRS* [Online]. Available: http://www.itrs.net/

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

[7] M. Luisier and G. Klimeck, "Atomistic full-band design study of InAs band-to-band tunneling field-effect transistors," *IEEE Electron Devices Lett.*, vol. 30, no. 6, pp. 602–604, Jun. 2009.

[8] 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, no. 5, pp. 054505-1–054505-10, Sep. 2009.

[9] G. Mil'nikov, N. Mori, and Y. Kamakura, "Equivalent transport models in atomistic quantum wires," *Phys. Rev. B*, vol. 85, no. 3, pp. 035317-1–035317-11, 2012.

[10] J. Z. Huang, W. C. Chew, J. Peng, C.-Y. Yam, L. J. Jiang, and G.-H. Chen, "Model order reduction for multiband quantum transport simulations and its application to p-type junctionless transistors," *IEEE Trans. Electron Devices*, vol. 60, no. 7, pp. 2111–2119, Jul. 2013.

[11] D. Gershoni, C. Henry, and G. 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.

[12] P. Enders and M. Woerner, "Exact 4×4 block diagonalization of the eight-band $k \cdot p$ Hamiltonian matrix for tetrahedral semiconductors and its application to strained quantum wells," *Semicond. Sci. Technol.*, vol. 11, no. 7, p. 983, 1996.

[13] R. G. Veprek, S. Steiger, and B. Witzigmann, "Ellipticity and the spurious solution problem of $k \cdot p$ envelope equations," *Phys. Rev. B*, vol. 76, no. 16, pp. 165320-1–165320-9, 2007.

[14] 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.

[15] A. Paussa, F. Conzatti, D. Breda, R. Vermiglio, D. Esseni, and P. Palestri, "Pseudospectral methods for the efficient simulation of quantization effects in nanoscale MOS transistors," *IEEE Trans. Electron Devices*, vol. 57, no. 12, pp. 3239–3249, Dec. 2010.

[16] F. Conzatti, M. G. Pala, D. Esseni, E. Bano, and L. Selmi, "Strain-induced performance improvements in InAs nanowire tunnel FETs," *IEEE Trans. Electron Devices*, vol. 59, no. 8, pp. 2085–2092, Aug. 2012.

[17] J. Guo, S. Datta, M. Lundstrom, and M. Anantam, "Toward multiscale modeling of carbon nanotube transistors," *Int. J. Multiscale Comput. Eng.*, vol. 2, no. 2, pp. 257–276, 2004.

[18] A. C. Seabaugh and Q. Zhang, "Low-voltage tunnel transistors for beyond CMOS logic," *Proc. IEEE*, vol. 98, no. 12, pp. 2095–2110, Dec. 2010.

[19] R. Jhaveri, V. Nagavarapu, and J. C. Woo, "Effect of pocket doping and annealing schemes on the source-pocket tunnel field-effect transistor," *IEEE Trans. Electron Devices*, vol. 58, no. 1, pp. 80–86, Jan. 2011.

[20] V. Nagavarapu, R. Jhaveri, and J. C. Woo, "The tunnel source (PNPN) n-MOSFET: A novel high performance transistor," *IEEE Trans. Electron Devices*, vol. 55, no. 4, pp. 1013–1019, Apr. 2008.

[21] Z. Chen, H. Yu, N. Singh, N. Shen, R. Sayanthan, G. Lo, *et al.*, "Demonstration of tunneling FETs based on highly scalable vertical silicon nanowires," *IEEE Electron Devices Lett.*, vol. 30, no. 7, pp. 754–756, Jul. 2009.

[22] H.-Y. Chang, B. Adams, P.-Y. Chien, J. Li, and J. C. Woo, "Improved subthreshold and output characteristics of source-pocket Si tunnel FET by the application of laser annealing," *IEEE Trans. Electron Devices*, vol. 60, no. 1, pp. 92–96, Jan. 2013.

[23] D. Verreck, A. Verhulst, K.-H. Kao, W. Vandenberghe, K. De Meyer, and G. Groeseneken, "Quantum mechanical performance predictions of p-n-i-n versus pocketed line tunnel field-effect transistors," *IEEE Trans. Electron Devices*, vol. 60, no. 7, pp. 2128–2134, Jul. 2013.

[24] 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, Nov. 2010.

**Jun Z. Huang** received the Ph.D. degree from The University of Hong Kong, Hong Kong, in November 2013. He joined Purdue University, West Lafayette, as a Post-Doctoral Researcher.

His current research interest is quantum transport simulation of nanoelectronic devices.

**Lining Zhang** (S'09–M'13) received the Ph.D. degree from the Hong Kong University of Science and Technology (HKUST), Hong Kong, in 2013.

He is currently a Research Associate with HKUST, and a Project Manager of i-MOS.

**Weng Cho Chew** (S'79–M'80–SM'86–F'93) received all his academic degrees from the Massachusetts Institute of Technology, Cambridge, MA, USA.

His research interests are in wave and field physics.

**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–SM'13) received the Ph.D. degree from the University of Illinois at Urbana-Champaign, Urbana, IL, USA, in 2004.

His current research interests include electromagnetics, computational electromagnetics, IC signal/power integrity, antennas, and multidisciplinary computational solutions.

**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.

**Mansun Chan** (S'92–M'95–SM'01–F'13) received the Ph.D. degree from the University of California at Berkeley, Berkeley, CA, USA, in 1995.

He has been with the Faculty of Electrical and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong, since 1996.