2012

Feasibility, Accuracy, and Performance of Contact Block Reduction Method for Multi-band Simulations of Ballistic Quantum Transport

Hoon Ryu
Korea Institute of Science and Technology Information

Hong-Hyun Park
Purdue University - Main Campus

Mincheol Shin
Korea Advanced Institute of Science and Technology

Dragica Vasileska
Arizona State University

Gerhard Klimeck
Purdue University - Main Campus, gekco@purdue.edu

Ryu, Hoon; Park, Hong-Hyun; Shin, Mincheol; Vasileska, Dragica; and Klimeck, Gerhard, "Feasibility, Accuracy, and Performance of Contact Block Reduction Method for Multi-band Simulations of Ballistic Quantum Transport" (2012). *Birck and NCN Publications.* Paper 866.
http://dx.doi.org/10.1063/1.3694740

AIP | Journal of Applied Physics

Feasibility, accuracy, and performance of contact block reduction method for multi-band simulations of ballistic quantum transport

Hoon Ryu, Hong-Hyun Park, Mincheol Shin, Dragica Vasileska, and Gerhard Klimeck

Citation: J. Appl. Phys. **111**, 063705 (2012); doi: 10.1063/1.3694740
View online: http://dx.doi.org/10.1063/1.3694740
View Table of Contents: http://jap.aip.org/resource/1/JAPIAU/v111/i6
Published by the American Institute of Physics.

Feasibility, accuracy, and performance of contact block reduction method for multi-band simulations of ballistic quantum transport

Hoon Ryu,[1,2,a)] Hong-Hyun Park,[2] Mincheol Shin,[3] Dragica Vasileska,[4] and Gerhard Klimeck[2]
[1]*Supercomputing Center, Korea Institute of Science and Technology Information, Daejeon 305-806, South Korea*
[2]*Network for Computational Nanotechnology, Purdue University, West Lafayette, Indiana 47907, USA*
[3]*Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, South Korea*
[4]*Department of Electrical Engineering, Arizona State University, Tempe, Arizona 85287, USA*

(Received 13 December 2011; accepted 17 February 2012; published online 21 March 2012)

Numerical utilities of the contact block reduction (CBR) method in evaluating the retarded Green's function are discussed for 3D multi-band open systems that are represented by the atomic tight-binding (TB) and continuum $k \cdot p$ (KP) band model. It is shown that the methodology to approximate solutions of open systems, which has been already reported for the single-band effective mass model, cannot be directly used for atomic TB systems, since the use of a set of zinc blende crystal grids makes the inter-coupling matrix non-invertible. We derive and test an alternative with which the CBR method can be still practical in solving TB systems. This *multi-band CBR* method is validated by a proof of principles on small systems and also shown to work excellent with the KP approach. Further detailed analysis on the accuracy, speed, and scalability on high performance computing clusters is performed with respect to the reference results obtained by the state-of-the-art recursive Green's function and wavefunction algorithm. This work shows that the CBR method could be particularly useful in calculating resonant tunneling features, but shows a limited practicality in simulating field effect transistors (FETs) when the system is described with the atomic TB model. Coupled to the KP model, however, the utility of the CBR method can be extended to simulations of nanowire FETs. © *2012 American Institute of Physics*. [http://dx.doi.org/10.1063/1.3694740]

I. INTRODUCTION

A. Needs for multi-band approaches

Semiconductor devices have been continuously down-scaled ever since the invention of the first transistor,[1] such that the size of the single building component of modern electronic devices has already reached a few nanometers (nm). In such a *nanoscale* regime, two conceptual changes are required in the device modeling methodology. One aspect is widely accepted, where carriers must be treated as quantum mechanical rather than classical objects. The second change is the need to embrace the multi-band models, which can describe atomic features of materials, reproducing experimentally verified bulk band structures. While the single-band effective mass approximation (EMA) predicts band structures reasonably well near the conduction band minimum (CBM), the subband quantization loses accuracy if devices are in a sub-nm regime.[2] The EMA also fails to predict indirect gaps, inter-band coupling, and non-parabolicity in bulk band structures.[3]

The nearest-neighbor empirical tight-binding (TB) and next nearest-neighbor $k \cdot p$ (KP) approach are most widely used band models of multiple bases.[3,4] The most sophisticated TB model uses a set of 10 localized orbital bases (s, s*, $3 \times$ p, and $5 \times$ d) on real atomic grids (20 with spin interactions), where the parameter set is fit to reproduce experimentally verified bandgaps, masses, non-parabolic dispersions, and hydrostatic and biaxial strain behaviors of bulk materials using a global minimization procedure based on a genetic algorithm and analytical insights.[3,5,6] This $sp^3d^5s^*$ TB approach can easily incorporate atomic effects, such as surface roughness and random alloy compositions, as the model is based on a set of atomic grids. These physical effects have been shown to be critical to the quantitative modeling of resonance tunneling diodes (RTDs), quantum dots, disordered SiGe/Si quantum wells, and a single impurity device in Si bulk.[7–10]

The KP approach typically uses four bases on a set of cubic grids with no spin interactions.[4] While it still fails to predict the indirect gap of bulk dispersions, since it assumes that all the subband minima are placed on the Γ point, the credibility is better than the EMA, since the KP model can still explain the inter-band physics of direct gap III-V devices and valence band physics of indirect gap materials, such as silicon (Si).[11,12]

B. Contact block reduction method

One of the important issues in modeling of nanoscale devices is to solve the quantum transport problem with a consideration of real 3D device geometries. Although the non- equilibrium Green's function (NEGF) and wave function (WF) formalism have been widely used to simulate the

a)Electronic mail: elec1020@gmail.com.

carrier transport,[2,11,13–15] the computational burden has been always a critical problem in solving 3D open systems, as the NEGF formalism needs to invert a system matrix of a degree-of-freedom (DOF) equal to the Hamiltonian matrix.[13] The recursive Green's function (RGF) method saves the computing load by selectively targeting elements needed for the matrix inversion.[16,17] However, the cost can be still huge depending on the area of the transport-orthogonal plane (cross-section) and the length along the transport direction of target devices.[18,19] The WF algorithm also saves the computing load if the transport is ballistic, as it does not have to invert the system matrix, and finding a few solutions of the linear system is enough to predict the transport behaviors. But the load still depends on the size of the system matrix and the number of solution vectors (modes) needed to describe the carrier-injection from external leads.[2,14] In fact, RGF and WF calculations for atomically resolved nanowire field effect transistors (FETs) have demonstrated the need to consume over 200 000 parallel cores on large supercomputing clusters.[20]

Developed by Mamaluy *et al.*,[19,21] the contact block reduction (CBR) method has received much attention, due to the utility to save computing expense required to evaluate the retarded Green's function of 3D open systems. The CBR method is thus expected to be a good candidate for transport simulations compared to the computing-intensive RGF method, since the method does not have to solve the linear system yet reducing the computing load needed for matrix inversion.[19] The method indeed has been extensively used such that it successfully modeled electron quantum transport in experimentally realized Si FinFETs[22] and predicted optimal design points and process variations in the design of 10-nm Si FinFETs.[23,24] However, all the successful applications for 3D systems so far have been demonstrated only for the systems represented by the EMA.

C. Goals of this work

While the use of multi-band approaches can increase the accuracy of simulation results, it requires more computing load, as a DOF of the Hamiltonian matrix is directly proportional to the number of bases required to represent a single atomic (or grid) spot in the device geometry. To suggest a solution to this *trade-off* issue, we examine the numerical utilities of the CBR method in multi-band ballistic quantum transport simulations, focusing on multi-band 3D systems represented by either the TB or KP band model.

The objective of this work is to provide detailed answers to the following questions through simulations of small two-contact ballistic systems focusing on a proof of principles: (1) Can the original CBR method be extended to simulate ballistic quantum transport of multi-band systems? (2) If the answer to the question (1) is "*yes*", what is the condition under which the multi-band CBR method becomes particularly useful? (3) How is the numerical practicality of the multi-band CBR method compared to the RGF and WF algorithms in terms of the accuracy, speed, and scalability on high performance computing (HPC) clusters?

II. METHODOLOGY

In real transport problems, a device needs to be coupled with external contacts that allow the carrier in-and-out flow. With the NEGF formalism, this can be done by creating an open system that is described with a non-Hermitian system matrix.[13] Representing this system matrix as a function of energy, we compute the transmission coefficient and density of states to predict the current flow and charge profile in non-equilibrium. This energy-dependent system matrix is called the retarded Green's function, G^R, for an open system (Eq. (1)).

$$G^R(E) = [(E + i\eta)I - H^o - \sum(E)]^{-1}, \quad \eta \to 0^+, \tag{1}$$

where H^o is the Hamiltonian representing the device and Σ is the self-energy term that couples the device to external leads. As already mentioned in the Sec. I, the evaluation of G^R is quite computationally expensive, since it involves intensive matrix inversions. The CBR method, however, reduces matrix inversions with the mathematical process based on the Dyson equation. We start the discussion revisiting the CBR method that has been so far utilized for EMA systems.

A. Revisit: CBR with EMA

The CBR method starts decomposing the device domain into two regions: (1) the boundary region c that couples with the contacts and (2) the inner region d that does not couple to the contacts. As the self-energy term Σ is non-zero only in the boundary region, H^o and Σ are decomposed as shown in Eq. (2), where subscripts (c and d) denote above-mentioned regions, respectively.

$$H = \begin{bmatrix} H^o_c & H^o_{cd} \\ H^o_{dc} & H^o_d \end{bmatrix}, \quad \sum = \begin{bmatrix} \sum_c & 0_{cd} \\ 0_{dc} & 0_d \end{bmatrix}. \tag{2}$$

Then, G^R can be evaluated with the Dyson equation defined in Eq. (3) and Eq. (4), where Σ^x and G^x are conditioned with a Hermitian matrix X to minimize matrix inversions by solving the eigenvalue problem (Eq. (5)).

$$A_c^{-1} = (I_c - G_c^x \sum_c^x)^{-1}, \tag{3}$$

$$\begin{aligned} G^R(E) &= (I - \sum^x G^x)^{-1} G^x \\ &= \begin{bmatrix} A_c^{-1} & 0_{cd} \\ G^x_{dc} \sum_c^x A_c^{-1} & I_d \end{bmatrix} \begin{bmatrix} G^x_c & G^x_{cd} \\ G^x_{dc} & G^x_d \end{bmatrix}, \end{aligned} \tag{4}$$

$$\begin{aligned} X &= \begin{bmatrix} x_c & 0_{cd} \\ 0_{dc} & 0_d \end{bmatrix}, \quad \sum^x = \sum - X, G^x = [EI - (H^o + X)]^{-1} \\ &= \begin{bmatrix} G^x_c & G^x_{cd} \\ G^x_{dc} & G^x_d \end{bmatrix} = \sum_\alpha \frac{|\Psi_\alpha\rangle\langle\Psi_\alpha|}{E - \epsilon_\alpha + i\eta}, \end{aligned} \tag{5}$$

where ϵ_α and Ψ_α are the αth eigenvalue and eigenvector of the modified Hamiltonian $(H^o + X)$. Here, we note that the matrix inversion is performed only to evaluate the boundary block, A_c (contact-block), for one time, while the RGF needs

to perform the block-inversion many times, depending on the device channel length. The computing load for matrix inversion is thus significantly reduced, and the method is also free from solving a linear system problem. Instead, the major numerical issue now becomes a normal eigenvalue problem for a Hermitian matrix $(H^o + X)$. For the numerical practicality, it is thus critical to reduce a number of required eigenvalues, and for EMA Hamiltonian matrices, a huge reduction in the number of required eigenvalues can be achieved via a smart choice of the *prescription matrix X*.

To find the matrix X and see if it can be extended to multi-band systems, we first need to understand how to couple external contacts to the device. Figure 1 illustrates the common approach, which treats the contact as a semi-infinite nanowire of a finite cross-section. Here, H_B is a block matrix that represents the unit-slab along the transport direction and W is another block matrix, which represents the inter-slab coupling. The eigenfunction of the plane wave at the mth mode in the nth slab, $\Psi_{(n,m)}$, should then obey the Schrödinger equation and the Bloch condition (Eq. (6)).

$$\begin{aligned}(EI - H_B)\Psi_{(n,m)} &= W^+\Psi_{(n-1,m)} + W\Psi_{(n+1,\mathrm{m})}, \Psi_{(n+1,\mathrm{m})} \\ &= \exp(ik_mL)\Psi_{(n,m)} (1 \le m \le M),\end{aligned} \tag{6}$$

where k_m is the plane-wave vector at the mth mode, L is the length of a slab along the transport direction, and M is the maximum number of plane-wave modes that can exist in a single slab and is equal to the DOF of H_B. Then, the surface Green's function, G_{surf}, and self-energy term, Σ, can be evaluated by converting Eq. (6) to the generalized eigenvalue problem for a complex and non-Hermitian matrix.[17] The solution for G_{surf} and Σ are provided in Eq. (7), where K and Λ are shown in Eq. (8),

$$\begin{aligned}G_{surf} &= K[K^{-1}(H_B - EI)K + K^{-1}W^+K\Lambda]^{-1}K^{-1}, \\ \sum &= W^+G_{surf}W,\end{aligned} \tag{7}$$

$$\begin{aligned}K &= [\Psi_{(0,1)}\Psi_{(0,2)}\ldots\Psi_{(0,M)}], \Lambda \\ &= diag[\exp(ik_1L)\exp(ik_2L)\ldots\exp(ik_ML)].\end{aligned} \tag{8}$$

In systems described by the nearest-neighbor EMA, each slab becomes a layer of common cubic grids, such that each grid on one layer is coupled to the same grid on the nearest layer. The inter-slab coupling matrix W thus becomes a *scaled identity matrix,* with which the general solution for G_{surf} and Σ in Eq. (7) can be simplified using a process described in Eq. (9) and Eq. (10). We note that previous literatures have shown only the simplified solution for G_{surf} and Σ.[19,21]

$$\begin{aligned}G_{surf} &= K[K^{-1}(H_B - EI)K + K^{-1}W^+K\Lambda]^{-1}K^{-1} \\ &= K[K^{-1}(H_B - EI)K + W^+\Lambda]^{-1}K^{-1} \\ &= K[-K^{-1}(W^+K\Lambda + WK\Lambda^{-1}) \\ &\quad + W^+\Lambda]^{-1}K^{-1}[\because (EI - H_B)K = W^+K\Lambda + WK\Lambda^{-1}] \\ &= -K[W\Lambda^{-1}]^{-1}K^{-1} = -K\Lambda W^{-1}K^{-1},\end{aligned} \tag{9}$$

$$\begin{aligned}\sum &= W^+G_{surf}W = W^+(-K\Lambda W^{-1}K^{-1})W \\ &= -W^+K\Lambda K^{-1}(\because W^+ = W) = -WK\Lambda K^{-1}.\end{aligned} \tag{10}$$

The original CBR method coupled to the EMA prescribes the Hermitian matrix X as $-W$ or its Hermitian component (if W is complex). The new self-energy term Σ_x in Eq. (5) then becomes (Eq. (11)),

$$\begin{aligned}\sum_x &= \sum -X = \sum +W = -WK\Lambda K^{-1} + W \\ &= -WK(\Lambda - I)K^{-1},\end{aligned} \tag{11}$$

where the matrix $(\Lambda - I)$ becomes zero at Γ point, on where EMA subband minima are always placed. The resulting new Hamiltonian $(H^o - W)$ becomes the Hamiltonian with the generalized *V on-Neumann* boundary condition at contact boundaries. The spectra of the matrix $(H^o - W)$ therefore become approximate solutions of the open boundary problem, and the retarded Green's function, $G^R(E)$, in Eq. (4) can be thus *approximated* with an incomplete set of energy spectra of the Hermitian matrix near subband minima.[19,25]

B. CBR with multi-band models

Regardless of the band model, the $G^R(E)$ in Eq. (4) can be accurately calculated with a complete set of spectra, since it then becomes the Dyson equation (Eq. (3)) itself. The important question here is then whether we can make the CBR method be still numerically practical for multi-band systems such that the transport can be simulated with as small a number of energy spectrums as possible. To study this issue, we focus on the inter-slab coupling matrix W of multi-band systems. A toy Si device that consists of two slabs along the [100] direction is used as an example for our discussion.

Figure 2 shows the device geometry and corresponding Hamiltonian matrix built with the EMA, KP, and TB model, respectively. Here, we note that the simplifying process in Eq. (9) and Eq. (10) is not strictly correct if the inter-slab coupling matrix W is not an identity matrix, since, for any square matrix K and W, $K^{-1}WK$ cannot be simplified to W if W is neither an identity matrix nor a scaled identity matrix. When a system is represented with the KP model, a single slab is still a layer of common cubic grids, as the KP approach also uses a set of cubic grids. But, the non-zero coupling is extended up to next-nearest neighbors such that the inter-slab coupling matrix W is no more an identity matrix. The simplified solution for G_{surf} and Σ, however, can be still used to *approximate* the general solutions in Eq. (7),

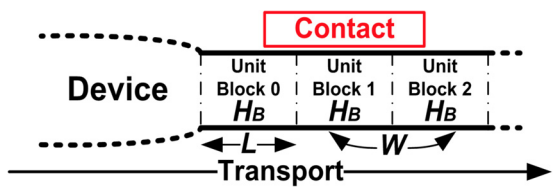

FIG. 1. Schematic of the semi-infinite contact to illustrate the treatment of the external contact that is normally assumed to be an infinite chain of the slab on the device boundary (the outmost slab in the device domain).

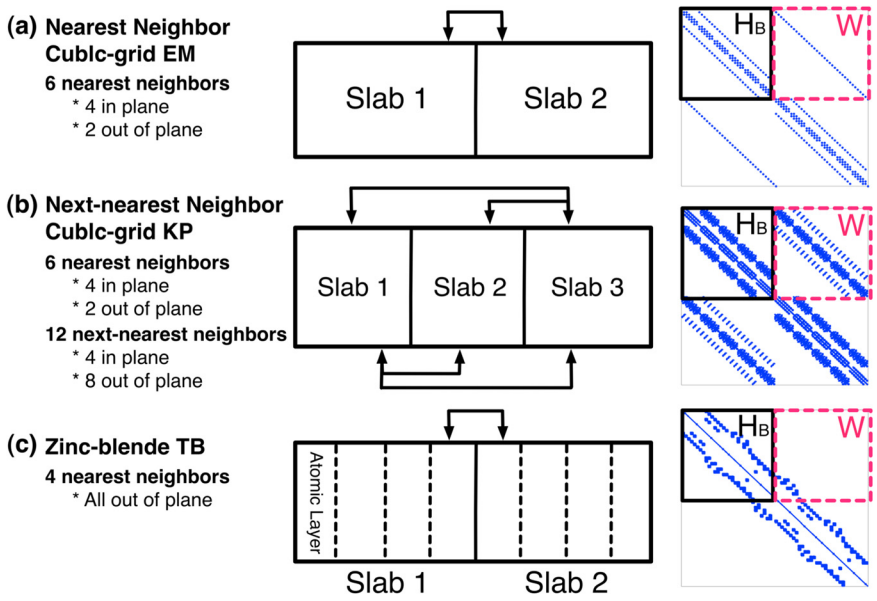

FIG. 2. Illustration of the geometry and the Hamiltonian matrix built for the (a) EMA, (b) KP, and (c) TB toy nanowire. Arrows represent the inter-slab coupling. The simplifying process in Eq. (9) and Eq. (10) are not strictly accurate for multi-band models, since the inter-slab coupling matrix is neither an identity matrix nor a scaled identity matrix. Especially, the coupling matrix becomes singular in TB model, which indicates that the simplified solution for G_{surf} and Σ are even mathematically invalid.

since the coupling matrix W is *diagonally dominant* and invertible. But the situation becomes tricky for TB systems that are represented on a set of real zinc blende (ZB) grids.

In the ZB crystal structure, a Si unit slab has a total of four unique atomic layers along the [100] direction. Because the TB approach assumes the nearest-neighbor coupling, only the last layer in one slab is coupled to the first layer in the nearest slab, while all the other coupling blocks among layers in different slabs become zero-matrices. As described in Fig. 2, this makes the inter-slab coupling matrix W *singular* such that matrix inversions become impossible. The simplified solution for G_{surf} and Σ in Eq. (9) and Eq. (10) are therefore mathematically invalid, and they cannot be even used to approximate the full solution (Eq. (7)). A new prescription for X is thus needed to make the CBR method still practical for ZB-TB systems, and we propose an alternative in Eq. (12),

$$X = \frac{\sum(\epsilon_{edge}) + \sum^{+}(\epsilon_{edge})}{2}, \tag{12}$$

where ϵ_{dge} is the energetic position of the CBM (valence band maximum (VBM)) of the band structure of the semi-infinite contact.

If only a few subbands near the CBM (or VBM) of the contact band structure are enough to describe the external contact, the prescription suggested in Eq. (12) works quite well, as X is the Hermitian part of the self-energy term, such that $(H^o + X)$ approximates the open system near the edge of the contact band structure. The approximation, however, becomes less accurate if more subbands in higher energy (in lower energy for valence band) are involved to the open boundaries. Away from the band edge, subband placement becomes denser and inter-subband coupling becomes stronger. The prescription X in Eq. (12) then would not be a good choice, as it only approximates the open boundary solution near band edges, and the CBR method thus needs more eigenspectrums to solve open boundary transport problems. So, for example, the multi-band CBR method would not be numerically practical to simulate FETs at a high source-drain bias, since a broad energy spectrum is then needed to get an accurate solution.

Before closing this section, we note that, if the inter-slab coupling matrix W is either an identity matrix or a scaled identity matrix, the prescription matrix X in Eq. (12) becomes *identical* to the one utilized to simulate 3D systems in the previous literatures,[22–24] where $(H^o + X)$ approximates the open system well near *every subband minima* if the system is represented by the EMA.[19,21,25] Once G_{surf} and Σ are determined from the prescription matrix X, evaluation of the transmission coefficient (TR) and the density of states (DOS) can be easily done.[13,19,21,25] Further detailed mathematics regarding derivation of TR and DOS will not be thus discussed here.

III. RESULTS AND DISCUSSIONS

The results are discussed in two subsections. First, we validate the CBR method for multi-band systems with the new prescription for X in Eq. (12). Focusing on a proof of principles, we compute the TR and DOS profiles for a toy TB and KP system, compare the result to the references obtained with the RGF algorithm, and suggest the device category where the multi-band CBR method could be particularly practical. Second, we examine the numerical practicality of the multi-band CBR method by computing TR and DOS profiles of a resonant tunneling device and a nanowire FET. The accuracy, the speed of calculations in a serial mode, and the scalability on HPC clusters are compared to those obtained with the RGF and WF algorithm. We assume a two-contact ballistic transport for all the numerical problems.

A. Validation of multi-band CBR method

To validate the multi-band CBR method that has been discussed in Sec. II, we consider two multi-band toy Si systems represented by the 10-band $sp^3d^5s^*$ TB and 3-band KP

approach. Here, we intentionally choose extremely small systems to calculate a complete set of energy spectra of the Hamiltonian, with which the CBR method should produce results identical to the ones obtained by the RGF algorithm. For the TB system, the electron-transport is simulated while we calculate the hole-transport for the KP system due to a limitation of the KP approach in representing the Si material. (In principle, the 4-band KP model uses a total of four bases to model direct bandgap materials, such as GaAs and InAs, where one basis is used to model the conduction band and the remaining three bases are used to model the valence band. The valence band of indirect bandgap materials, such as Si, however, can be still modeled with three bases if VBM is at the Γ point (Ref. 11).)

1. TB system

Fig. 3 illustrates the dispersion, TR, and DOS profile that are calculated for the TB Si toy device, which consists of $(2 \times 2 \times 2)$ (100) unit-cells ($\sim$1.1(nm)). The device involves a complex Hermitian Hamiltonian matrix of 640 DOF, and electrons are assumed to transport along the [100] direction. The TR and DOS profiles are calculated using the CBR method for a total of four cases—with 6, 60, 190, and full (640) energy spectra that correspond to 1%, 10%, 30%, and 100% of the Hamiltonian DOF, respectively. The transport happens at the energy above 2.32(eV), which is the CBM of the contact band structure. We note that this energetic position is higher than the Si bulk CBM (1.13(eV)), due to the structural confinement stemming from the finite cross-section of the nanowire device. (Both the TB and KP model considered in this work place the VBM of Si bulk at 0 (eV).)

With the new prescription matrix X suggested in Eq. (12), the TR and DOS profile obtained by the CBR method become closer to the reference result, as more spectrums are used, and eventually reproduce the reference result with a full set of spectrums, as shown in the left column of Fig. 3(b). Here, the CBR result turns out to be quite accurate near the CBM, even with 1% of the total spectrums, indicating that the TB-CBR method could be a practical approach if most of the carriers are injected from the first one or two subbands of the contact band structure. This condition can be satisfied when (1) only the first one or two subbands in the contact band structure are occupied with electrons and (2) the energy difference between the source and drain contact Fermi level (the source-drain Fermi window) becomes extremely narrow. So the simulation of FETs at a high source-drain bias would not be an appropriate target of the TB-CBR simulations, since the source-drain Fermi window may include many subbands and many spectra may be thus needed for accurate solutions. (Assuming that the source contact is grounded, the Fermi window at $V_{DS} = \mathrm{V}$ becomes

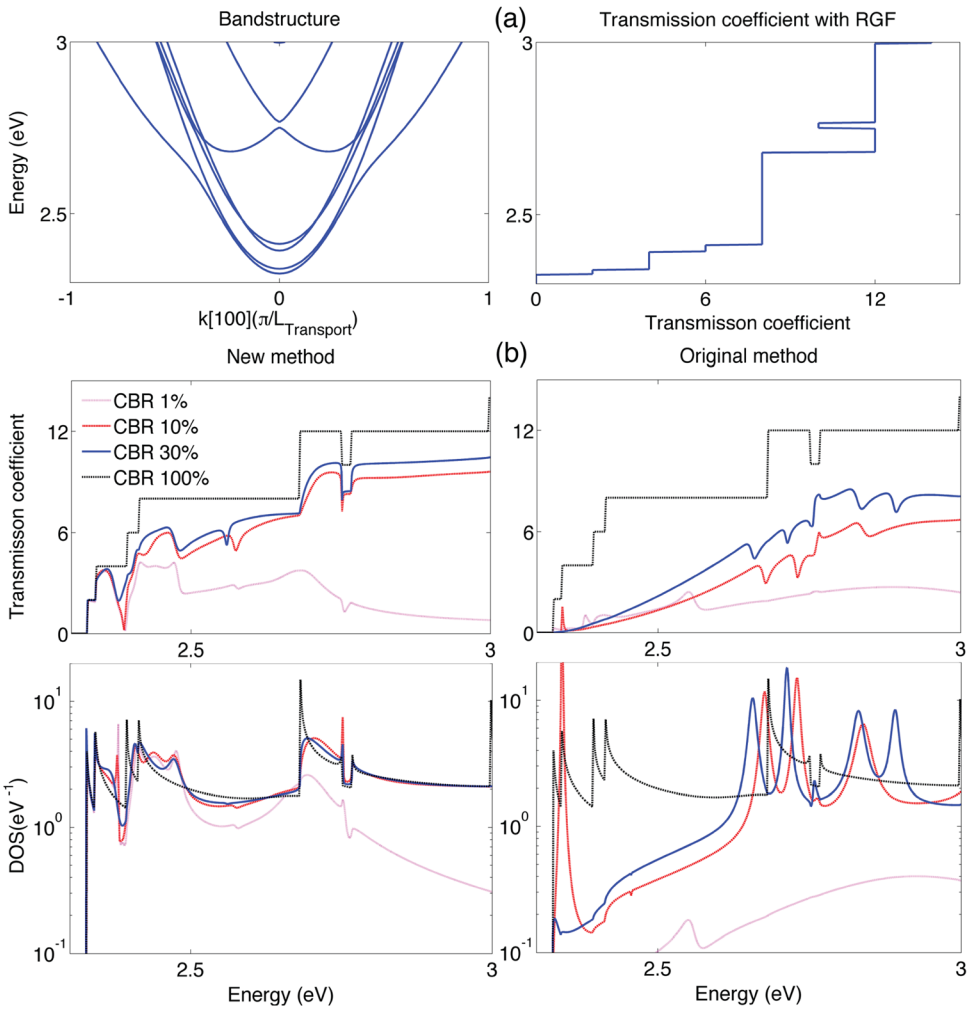

FIG. 3. Electron transport in a toy Si TB system. (a) Band structure along the transport direction and TR calculated with RGF. (b) TR and DOS profiles calculated by the CBR method: Results with a prescription suggested in Eq. (12) (new method) and an old prescription suggested for the EMA (original method). Note that, with the old prescription, using more energy spectra does not necessarily improve the accuracy of the CBR solution.

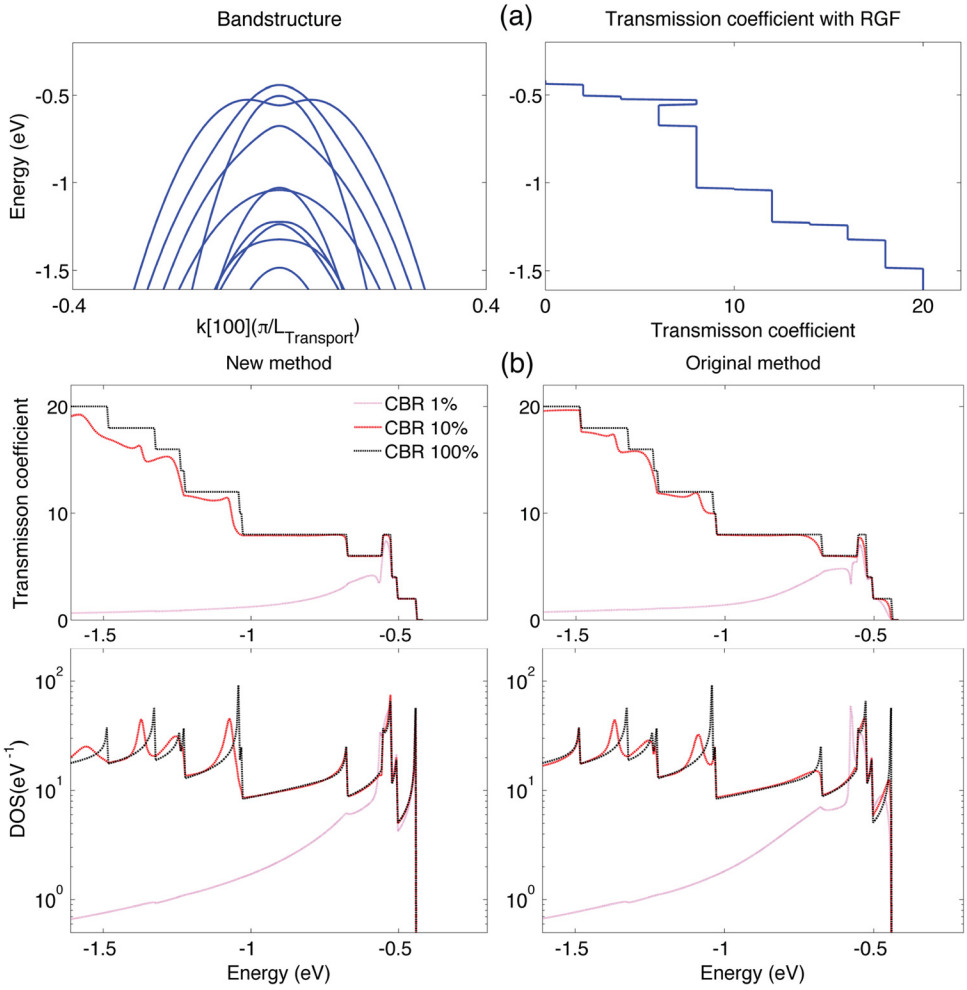

FIG. 4. Hole transport in a toy Si KP system. (a) Band structure along the transport direction and TR calculated with RGF. (b) TR and DOS profiles calculated by the CBR method: Results with a prescription suggested in Eq. (12) (new method) and an old prescription suggested for the EMA (original method). Note that the accuracy of the CBR solution is similar with both the new and old prescription.

$[E_{FD} - 3kT, E_{FS} + 3kT] = [E_F - qV - 3kT, E_F + 3kT]$, where q is the single electron charge and E_F is the Fermi level of the system in equilibrium. The maximum and minimum of the window are determined at the source and drain side, respectively.) Instead, we propose that RTDs could be one of the device categories for which the TB-CBR method is particularly practical, since the Fermi window for transport becomes extremely small in RTDs in some cases.[26]

The same calculation is performed again, but using the old prescription X suggested for the EMA, and corresponding TR and DOS profiles are shown in the right column of Fig. 3(b). The CBR method still reproduces the reference result with a full set of energy spectra, since the Dyson equation (Eq. (4)) should always work for any X's. The accuracy of the results near the CBM, however, turns out to be worse than the one with the new prescription. The results furthermore reveal that the accuracy with 10% of the total spectra does not necessarily become better than the one with 1%, indicating that the old prescription for X cannot even approximate the solution near the CBM of open TB systems.

2. KP system

The dispersion, TR, and DOS profile of the KP Si 2.0(nm) (100) cube are depicted in Fig. 4. The structure is discretized with a 0.2(nm) grid and involves a complex Hermitian Hamiltonian of 3000 DOF. Here, the DOF of the real-space KP Hamiltonian can be effectively reduced with the *mode-space* approach.[11] The effective DOF of the Hamiltonian therefore becomes 500, where we consider 50 modes per each slab along the transport direction. Again, we note that the VBM of the contact band structure is placed at –0.4(eV) and lower in energy than the VBM of Si bulk (0(eV)), due to the confinement created by the finite cross-section.

We claim that the CBR method works quite well for the KP system, since the TR and DOS profiles not only become closer to the reference results as more of the energy spectrums are used, but also exhibit excellent accuracy near the VBM of the contact band structure, as shown in Fig. 4(b). We, however, observe a remarkable feature that is not found in the CBR method coupled to TB systems: The KP-CBR method shows a good accuracy with both the old and new prescription matrix X, which supports that the simplified solution for G_{surf} and Σ (Eq. (9) and Eq. (10)) are still useful to approximate the full solution (Eq. (7)), as discussed in Sec. II. We also claim that the utility of the KP-CBR method could be extended to nanowire FETs, because the mode-space approach reduces the DOF of the Hamiltonian such that we save more computing cost needed to calculate energy spectra. In Subsection III B, we will come back to this issue again.

B. Practicality of multi-band CBR method

In this subsection, we provide a detailed analysis of the numerical utility of the multi-band CBR method in terms of the accuracy and speed. Based on discussions in Subsection III A, with a focus on a proof of principles on small systems, a RTD is considered as a simulation example of TB systems, while a nanowire FET is again used as an example of KP systems to discuss the numerical practicality of the method. The TR and DOS profiles obtained by the RGF and WF algorithm are used as reference results. We note that the WF case is added in this subsection to provide a complete and competitive analysis on the speed and scalability on HPC clusters.

1. TB system

A single phosphorous donor in host Si material (Si:P) creates a 3D structural confinement around itself. Such Si:P *quantum dots* have gained scientific interest, due to their potential utility for qubit-based logic applications.[27] Especially, the Stark effect in Si:P quantum dots is one of the important physical problems and was quantitatively explained by previous TB studies.[9,10] The electron-transport in such Si:P systems should be, therefore, another important problem that needs to be studied.

The geometry of the example Si:P device is illustrated in Fig. 5. Here, we consider a [100] Si nanowire that is 14.0(nm) long and has a 1.7(nm) rectangular cross-section. The first and last 3.0(nm) along the transport direction are considered as a densely N-type doped source- drain region, assuming a 0.25(eV) band-offset in equilibrium. (The band offset between the intrinsic channel and densely doped source-drain leads is taken from the work of Martinez *et al.*, where the equilibrium potential profile has been self-consistently obtained for a 14.0(nm)-long [100] Si nanowire that has a 2.0(nm) rectangular cross-section and 4.0(nm)-long source-drain regions (Ref. 28).) Then, a single phosphorous atom is placed at the channel center with a superposition of the impurity coulombic potential that has been calibrated for a single donor in Si bulk by Rahman *et al.*[9] The electronic structure has a total of 1872 atoms and involves a complex Hamiltonian matrix of 18 720 DOF.

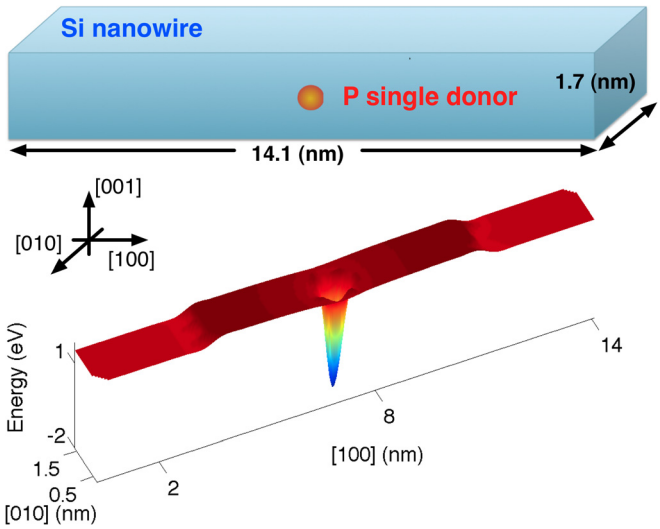

FIG. 5. Illustration of the geometry and potential profile of a Si:P RTD that is used as the example to examine the utility for the TB-CBR method. For the potential profile, 1.13(eV) is used as a reference value representing the Si bulk CBM. The single donor coulombic potential that has been calibrated by Rahman *et al.* (Ref. 9), with respect to the Si bulk, is superposed to the channel potential profile to consider the sharp structural confinement stemming from the single donor.

Figure 6 shows the TR and DOS profiles in five cases, where the first four cases are the CBR results with 10, 20, 40, and 80 spectra that correspond to 0.05%, 0.1%, 0.2%, and 0.4% of the Hamiltonian DOF, and the last one is used as a reference. Due to the donor coulombic potential, the channel forms a double-barrier system such that the electron transport should experience a resonance tunneling. As shown in Fig. 6, the CBR method produces a nice approximation of the reference result such that the first resonance is observed with just 10 energy spectra. It also turns out that 80 spectra are enough to capture all the resonances that show up in the range of energy of interest.

The accuracy of the solutions approximated by the CBR method is examined in a more quantitative manner by *integrating* the TR and DOS profile over energy. Figure 7 illustrates this *cumulative* TR (CTR) and DOS (CDOS) profile, which are *conceptually* equivalent to the current and charge profile, respectively. Here, again, the overall shape of the integrated results resembles the reference result more with an increasing number of energy spectrums. Some cases are, however, observed where the point-by-point absolute values are not quite improved, although more spectrums are considered in the simulation. For example, compared to the CTR profile with 0.1% spectrums, the CTR profile near the 2nd resonance ($\sim$1.55 eV) does not necessarily improve with 0.2% spectrums. This indicates that the channel mode near the 2nd resonance point is quite coupled to the other modes at the energy range that is not covered with 0.2% spectrums, such that more spectrums are needed for a better accuracy, as the result with 0.4% spectrums shows. In spite of a slight deviation in absolute values, however, the CTR profiles still confirm that the CBR method captures resonances quite precisely, such that the energetic positions where the TR sharply increases are almost on top of the reference result with just 0.4% spectrums. Compared to the CTR profile, the CDOS profile exhibits much better accuracy, such that the result with 80 spectra almost reproduces the reference result, even in terms of absolute values. We claim that the accuracy in the CDOS profile is particularly critical, since it is directly connected to charge profiles that are essential for charge-potential self-consistent simulations.

2. KP system

Si nanowire FETs obtained through top-down etching or bottom-up growth have attracted attention, due to their enhanced electrostatic control over the channel, and thus become an important target of various modeling works.[11,29] For KP systems, the CBR method could become a practical approach to solve transport behaviors of FET devices, since the computing load for solving eigenvalue problems can be reduced with the mode-space approach.

A [100] Si nanowire FET of a 15.0(nm) long channel and a 3.0(nm) rectangular cross- section is therefore considered as a simulation example to test the performance of the KP-CBR

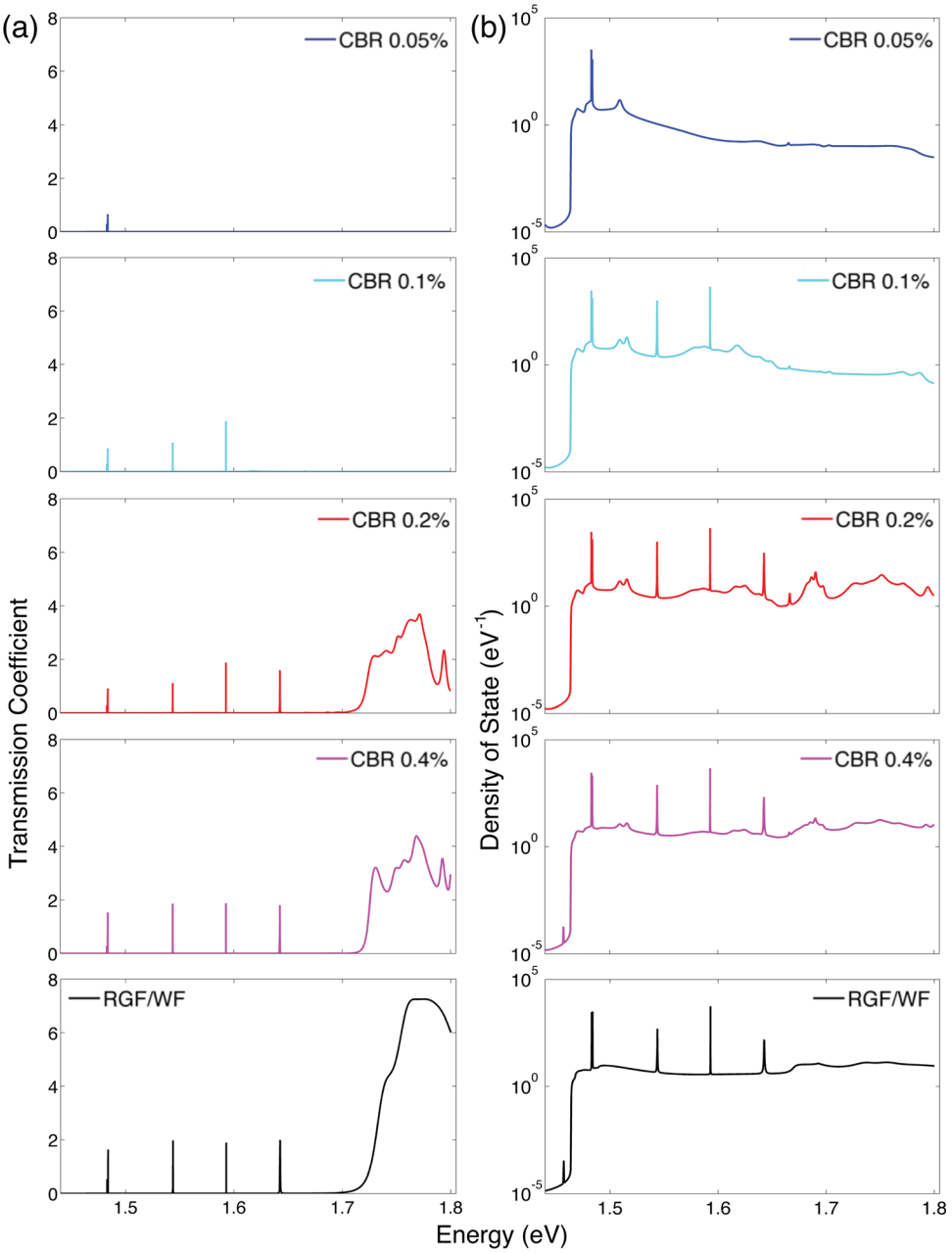

FIG. 6. Electron transport in a Si:P RTD. (a) TR and (b) DOS profiles calculated by the CBR method. Note that all the resonances in the range of energy are captured, even with just 80 spectra that corresponds to just 0.4% of the DOF of the TB Hamiltonian.

method. The hole-transport is simulated with the 3-band KP approach, where the simulation domain is discretized with a set of 0.2(nm) mesh cubic grids and involves a real-space Hamiltonian matrix of 50 625 DOF. As the device has a total of 75 slabs along the transport direction, the mode-space Hamiltonian has 9000 DOF with a consideration of 120 modes per slab. It has been reported that the wire band structure obtained with 120 modes per slab becomes quite close to the full solution for a cross-section smaller than $5.0 \times 5.0(\text{nm}^2)$.[11] The wire is assumed to be purely *homogeneous*, such that neither the doping nor band-offset are considered.

To see if the CBR method can be reasonably practical in simulating the hole-transport at a relatively large source-drain bias, we plan to cover the energy range at least larger than 0.4(eV) beyond the VBM of the wire band structure. For this purpose, we compute 50, 100, and 200 energy spectra that correspond to 0.5%, 1.1%, and 2.2% of the DOF of the mode-space Hamiltonian, respectively. Figure 8(a) shows the corresponding TR and DOS profiles. Here, the CBR solution not only becomes closer to the reference result with more spectra considered, but also demonstrates fairly excellent accuracy near the VBM of the wire band structure. The CTR and CDOS profiles provided in Fig. 8(b) further support the preciseness of the CBR solutions near the VBM. The cumulative profiles also support that the CBR solution covers a relatively wide range of energy, such that 50 energy spectra are already enough to cover ~0.4(eV) below the VBM quite well. We note that the solution obtained with 200 spectra almost replicates the reference result in the entire range of energy that is considered for the simulation (~0.8(eV) below the VBM).

C. Speed and scalability on HPCs

So far, we have discussed the practicality of the multi-band CBR method, focusing on the accuracy of the solutions for two-contact, ballistic-transport problems. Another

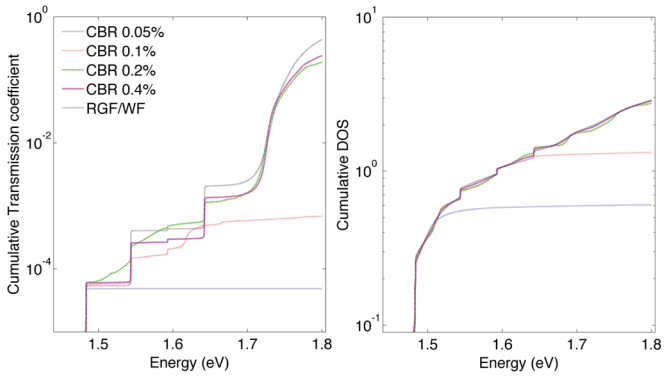

FIG. 7. Electron transport in a Si:P RTD. TR and DOS profiles integrated over energy. The cumulative profiles of TR and DOS effectively indicate the accuracy of the current and charge profiles. The cumulative DOS is especially important, as it is directly coupled to charge profiles that are needed for charge-potential self-consistent simulations.

important criterion to determine the numerical utility should be the speed of calculations. We therefore measure the time needed to evaluate the TR and DOS per single energy point for the TB Si:P RTD and the KP Si nanowire FET represented that are utilized as simulation examples. To examine the practicality of the multi-band CBR method on HPC clusters, we also benchmark the scalability of the simulation time on the *Coates* cluster under the support of the Rosen Center for Advanced Computing (RCAC) at Purdue University. The CBR, RGF, and WF methods are parallelized with Message Passing Interface (MPI)/C++, the MUltifrontal Massively Parallel sparse direct linear Solver (MUMPS),[30] and a self-developed eigensolver based on the shift-and-invert Arnoldi algorithm.[31] All the measurements are performed on a 64-bit, 8-core HP Proliant DL585 G5 system of 16-GB SDRAM and 10-gigabit ethernet local to each node.

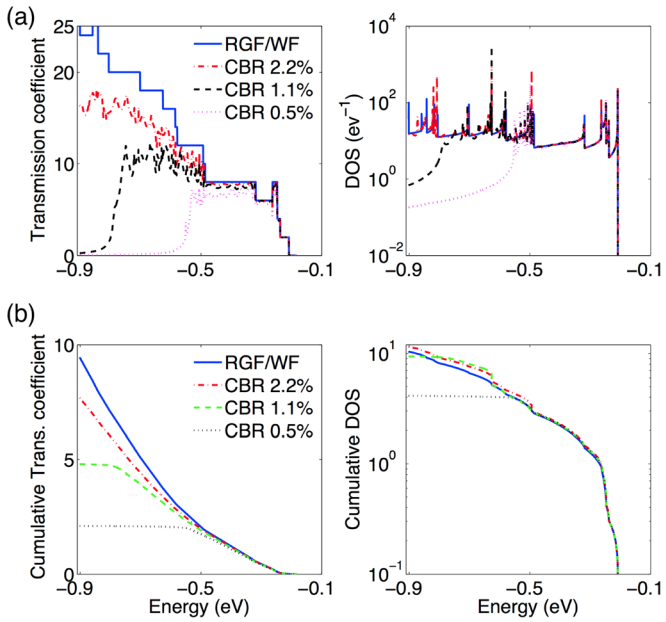

FIG. 8. Hole transport in a Si nanowire. (a) TR and DOS profiles and (b) corresponding cumulative profiles. The KP-CBR solutions exhibit excellent accuracy, such that 200 spectra, which corresponds to just 2.2%, turn out to be enough to almost reproduce the reference solutions in the entire range of energy of interest (0.8(eV) beyond the VBM of the wire band structure).

TABLE I. The time required to evaluate the TR and DOS per single energy point in a serial mode for the RTD and nanowire FET considered as simulation examples.

Approaches (TB)	time (s)	Approaches (KP)	time (s)
CBR 0.05(%)	11.5	CBR 0.5(%)	4.9
CBR 0.1(%)	11.8	CBR 1.1(%)	5.1
CBR 0.2(%)	12.0	CBR 2.2(%)	5.9
CBR 0.4(%)	12.7	RGF	5.0
RGF	19.0	WF	3.4
WF	6.5		

Table I summarizes the wall times measured for various methods in a serial mode. Generally, the simulation of the KP Si nanowire FET needs less computing loads, such that the wall times are reduced by a factor of two with respect to the computing time taken for the TB Si:P RTD. This is because the KP approach can represent the electronic structure with the mode-space approach such that the Hamiltonian matrix has a smaller DOF (9000) compared to the one used to describe the TB Si:P RTD (18 720).

Compared to the RGF algorithm in a serial mode, the CBR method demonstrates a comparable (KP) or better (TB) performance. Since a single slab of the KP Si nanowire is represented with a block matrix H_B (Fig. 1) of 120 DOF, the matrix inversion is not a critical problem anymore in the RGF algorithm, such that the CBR method does not necessarily show better performances than the RGF algorithm. The TB example device, however, needs a H_B of 720 DOF to represent a single slab (a total of 26 slabs), so the burden for matrix inversions become bigger compared to the KP example. As a result, the CBR method generally shows better performances. The CBR method, however, does not beat the WF method in both the TB and KP case, since, in a serial mode, the CBR method consumes time to allocate a huge

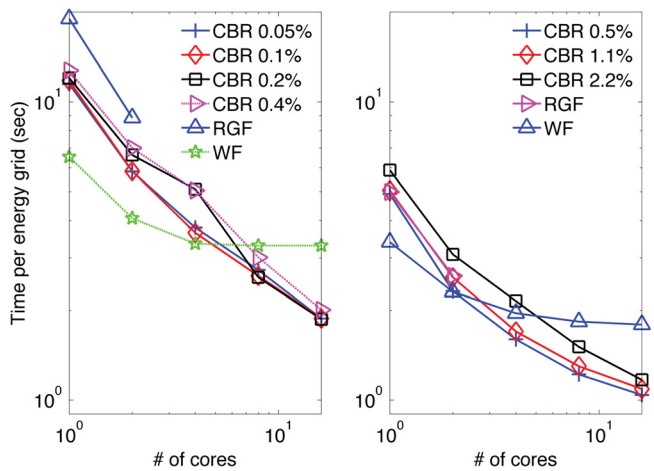

FIG. 9. Speed and scalability of the multi-band CBR method: For the example multi-band systems of TB Si:P RTD and KP Si nanowire FET, we measure the time required to calculate the TR and DOS per single energy point. Scalability of the calculation time is also measured to examine the numerical practicality of the method on HPC clusters.

memory space that is needed to store "full" complex matrices via vector-products (Eq. (5)).

The strength of the CBR method emerges in a *parallel* mode (on multiple CPUs), where the vector products are performed via MPI communication among distributed systems, and each node thus saves only a fraction of the full matrix. The scalability of the various methods is compared up to a total of 16 CPUs in Fig. 9. The common RGF calculation can be effectively parallelized only up to a factor of two, due to its recursive nature,[16] and the scalability of the WF method becomes worse in many CPUs, because it uses a direct-solver–based LU factorization to solve the linear system. As a result, the CBR method starts to show the best speed when more than 8 CPUs are used.

IV. CONCLUSION

In this work, we discuss numerical utilities of the CBR method in simulating ballistic transport of multi-band systems described by the atomic 10-band $sp^3d^5s^*$ TB and 3-band KP approach. Although the original CBR method developed for single-band EMA systems achieves an excellent numerical efficiency by approximating solutions of open systems, we show that the same approach cannot be used to approximate TB systems as the inter-slab coupling matrix becomes singular. We therefore develop an alternate method to approximate open system solutions. Focusing on a proof of principles on small systems, we validate the idea by comparing the TR and DOS profile to the reference result obtained by the RGF algorithm, where the alternative also works well with the KP approach.

Since the major numerical issue in the CBR method is to solve a normal eigenvalue problem, the numerical practicality of the method becomes better as the transport can be solved with a lower number of energy spectra. Generally, the practicality would be thus limited in multi-band systems, since multi-band approaches need a larger number of spectra to cover a certain range of energy than the single band EMA does. We, however, claim that the RTDs could be one category of TB devices, for which the multi-band CBR method becomes particularly practical in simulating transport, and the numerical utility can be even extended to FETs when the CBR method is coupled to the KP band model. To support this argument, we simulate the electron resonance tunneling in a 3D TB RTD, which is basically a Si nanowire, but has a single phosphorous donor in the channel center, and the hole-transport of a 3D KP Si nanowire FET. We examine numerical practicalities of the multi-band CBR method in terms of the accuracy and speed with respect to the reference results obtained by the RGF and WF algorithm and observe that the CBR method gives fairly accurate TR and DOS profiles near band edges of contact band structures.

In terms of the speed in a serial mode, the strength of the CBR method over the RGF algorithm depends on the size of the Hamiltonian such that the CBR shows a better performance than the RGF, as a larger block matrix is required to represent the unit-slab of devices. But the speed of the WF method is still better than the CBR method, as the CBR method consumes time to store a full complex matrix during the process of calculations. In a parallel mode, however, the CBR method starts to beat both the RGF and WF algorithm, since the full matrix can be stored into multiple clusters in a distributive manner, while the scalabilities of both the RGF and WF algorithm are limited, due to the nature of recursive and direct-solver–based calculations, respectively.

ACKNOWLEDGMENTS

H. Ryu, H.-H. Park, and G. Klimeck acknowledge the financial support from the National Science Foundation (NSF) under the contract No. 0701612 and the Semiconductor Research Corporation. M. Shin acknowledges the financial support from Basic Science Research Program through the National Research Foundation of Republic of Korea funded by the Ministry of Education, Science and Technology under the contract No. 2010-0012452. The authors acknowledge the extensive use of computing resources in the Rosen Center for Advanced Computing at Purdue University and NSF-supported computing resources on nanoHUB.org.

[1]G. E. Moore, Electronics **38**, 114 (1965).
[2]M. Luisier, A. Schenk, and W. Fichtner, Phys. Rev. B **74**, 205323 (2006).
[3]G. Klimeck, S. S. Ahmed, H. Bae, N. Kharche, R. Rahman, S. Clark, B. Haley, S. Lee, M. Naumov, H. Ryu, F. Saied, M. Prada, M. Korkusinski, and T. B. Boykin, IEEE Trans. Electron Devices **54**, 2079 (2007).
[4]Y. X. Liu, D. Z.-Y. Ting, and T. C. McGill, Phys. Rev. B **54**, 5675 (1996).
[5]G. Klimeck, R. Lake, R. C. Bowen, C. Fernando, and W. Frensley, VLSI Des. **8**, 79 (1998).
[6]T. B. Boykin, G. Klimeck, and F. Oyafuso, Phys. Rev. B **69**, 115201 (2004).
[7]R. C. Bowen, G. Klimeck, W. R. Frensley, and R. K. Lake, J. Appl. Phys. **81**, 3207 (1997).
[8]N. Kharche, M. Prada, T. B. Boykin, and G. Klimeck, Appl. Phys. Lett. **90**, 9 (2007).
[9]R. Rahman, C. J. Wellard, F. R. Bradbury, M. Prada, J. H. Cole, G. Klimeck, and L. C. L. Hollenberg, Phys. Rev. Lett. **99**, 036403 (2007).
[10]G. P. Lansbergen, R. Rahman, C. J. Wellard, I. Woo, J. Caro, N. Collaert, S. Biesemans, G. Klimeck, L. C. L. Hollenberg, and S. Rogge, Nature Phys. **4**, 656 (2008).
[11]M. Shin, J. Appl. Phys. **106**, 054505 (2009).
[12]C. Pryor, Phys. Rev. B **57**, 7190 (1998).
[13]S. Datta, Superlattices Microstruct. **28**, 253 (2000).
[14]M. Städele, B. R. Tuttle, and K. Hess, J. Appl. Phys. **89**, 348 (2001).
[15]N. Kharche, G. Klimeck, D.-H. Kim, J. A. del Alamo, and M. Luisier, *paper presented at the 2009 IEEE International Electron Devices Meeting, Baltimore, MD*, 7–9 December 2009.
[16]R. Lake, G. Klimeck, R. C. Bowen, and D. Jovanovic, J. Appl. Phys. **81**, 7845 (1996).
[17]C. Rivas and R. Lake, Phys. Status Solidi B **239**, 94 (2003).
[18]S. Cauley, J. Jain, C.-K. Koh, and V. Balakrishnan, J. Appl. Phys. **101**, 123715 (2007).
[19]D. Mamaluy, M. Sabathil, T. Zibold, P. Vogl, and D. Vasileska, Phys. Rev. B **71**, 245321 (2005).
[20]G. Klimeck and M. Luisier, Comput. Sci. Eng. **12**, 28 (2010).
[21]D. Mamaluy, D. Vasileska, M. Sabathil, and P. Vogl, Semicond. Sci. Technol. **19**, 118 (2004).
[22]H. R. Khan, D. Mamaluy, and D. Vasileska, IEEE Trans. Electron Devices **54**, 784 (2007).
[23]H. R. Khan, D. Mamaluy, and D. Vasileska, IEEE Trans. Electron Devices **55**, 743 (2008).
[24]H. R. Khan, D. Mamaluy, and D. Vasileska, IEEE Trans. Electron Devices **55**, 2134 (2008).
[25]D. Mamaluy, M. Sabathil, and P. Vogl, J. Appl. Phys. **93**, 4628 (2003).

[26]E. Lind, B. Gustafson, I. Pietzonka, and L.-E. Wernersson, Phys. Rev. B **68**, 033312 (2003).
[27]L. C. L. Hollenberg, A. S. Dzurak, C. J. Wellard, A. R. Hamilton, D. J. Reilly, G. J. Milburn, and R. G. Clark, Phys. Rev. B **69**, 113301 (2004).
[28]A. Martinez, N. Seoane, A. R. Brown, J. R. Barker, and A. Asenov, IEEE Trans. Electron Devices **8**, 603 (2009).
[29]N. Neophytou, A. Paul, M. S. Lundstrom, and G. Klimeck, IEEE Trans. Electron Devices **55**, 1286 (2008).
[30]See http://graal.ens-lyon.fr/MUMPS, for information about MUltifrontal Massively Parallel Sparse direct Solver (MUMPS).
[31]V. Mehrmann and D. Watkins, SIAM (Soc. Ind. Appl. Math.) J. Sci. Stat. Comput. **22**, 1905 (2001).