AIP | Journal of Applied Physics

Methods for fast evaluation of self-energy matrices in tight-binding modeling of electron transport systems

Jun Z. Huang, Weng Cho Chew, Yumao Wu, and Li Jun Jiang

Citation: J. Appl. Phys. **112**, 013711 (2012); doi: 10.1063/1.4732089
View online: http://dx.doi.org/10.1063/1.4732089
View Table of Contents: http://jap.aip.org/resource/1/JAPIAU/v112/i1
Published by the American Institute of Physics.

Methods for fast evaluation of self-energy matrices in tight-binding modeling of electron transport systems

Jun Z. Huang,[1,a)] Weng Cho Chew,[1,2,b)] Yumao Wu,[1,c)] and Li Jun Jiang[1,d)]
[1]*Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong*
[2]*Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, Illinois 61801-2991, USA*

(Received 28 March 2012; accepted 30 May 2012; published online 9 July 2012)

Simulation of quantum carrier transport in nanodevices with non-equilibrium Green's function approach is computationally very challenging. One major part of the computational burden is the calculation of self-energy matrices. The calculation in tight-binding schemes usually requires dealing with matrices of the size of a unit cell in the leads. Since a unit cell always consists of several planes (for example, in silicon nanowire, four atomic planes for [100] crystal orientation and six for [111] and [112]), we show in this paper that a condensed Hamiltonian matrix can be constructed with reduced dimension ($\sim 1/4$ of the original size for [100] and $\sim 1/6$ for [111] and [112] in the nearest neighbor interaction) and thus greatly speeding up the calculation. Examples of silicon nanowires with $sp^3d^5s^*$ basis set and the nearest neighbor interaction are given to show the accuracy and efficiency of the proposed methods. © *2012 American Institute of Physics*. [http://dx.doi.org/10.1063/1.4732089]

I. INTRODUCTION

Non-equilibrium Green's function (NEGF) approach[1,2] has been widely adopted to simulate quantum transport in nanoscale devices. However, the large computational cost of this method limits its application to small problems. One major part of computational cost is the inversion of the large Hamiltonian matrix so as to obtain the Green's function of the device. Considerable effort has been made to reduce the complexity, such as recursive Green's function algorithm,[3] mode space approaches,[4,5] contact block reduction method,[6,7] and the recent R-matrix method.[8,9] Another major source of the cost is the open boundary treatment, which is expressed explicitly through the self-energy matrices. In the effective mass approximation,[4,8] the self-energy matrices can be obtained for the whole energy band once the eigenmodes of the leads are solved.[10] Beyond the effective mass approximation, such as the *ab initio* methods[11] and the empirical tight-binding approaches,[9] however, the self-energy matrices must be evaluated for each energy point individually, further increasing the computational burden. The tight binding models will be the focus of this work, as they are well-suited for nanodevice modeling due to limited-range interactions and reasonably sized basis sets.[12]

Traditionally, there are roughly two kinds of approaches for self-energy evaluation,[13] one is through iterative evaluation of the surface Green's function,[14] the other is by solving the Bloch modes of the leads.[15–17] The underlining assumption of both approaches is that the leads are characterized by a periodic potential and thus a principle layer[18,19] (usually a unit cell in tight-binding schemes) can be defined with translational invariance along the leads. The former approach usually requires many inversions of a Hamiltonian matrix of the size of the unit cell. The latter one, instead, requires solving a generalized eigenvalue problem (GEVP) for a matrix of the size twice that of the unit cell.

Several improvements that speed up the calculation have been developed over the past years. The widely used decimation algorithm[18] greatly improves the convergence of the iterative method by reducing the iteration steps from N to $\log(N)$. The shift-and-invert method transforms the GEVP to a normal eigenvalue problem (NEVP).[20] The Krylov subspace method reduces the cost of the GEVP approach by computing only a portion of the eigenmodes that have contribution to the transmission.[21] However, the calculation is still very slow when the size of the unit cell matrix becomes very large. We also notice that by imposing absorbing boundary conditions into the leads, the open system is transformed to a closed system and the surface Green's function (and then the self energy) can be constructed for any energy by spectral representation.[22] But this should be designed very carefully in order to eliminate possible reflections (less reflection with more absorbing layers, but with more computational cost).

However, if we take a closer look at the structure of the unit cell, it is easy to find that there are some redundancies when these traditional methods are applied to tight-binding schemes. Take silicon (or germanium), for example, the [100] crystal direction nanowire consists of four atomic planes in the unit cell and the [111] (or [112]) direction consists of six planes, as shown in Fig. 1. Moreover, take the nearest neighbor tight binding scheme,[23] for example, we do not need the surface Green's function of the size of the unit cell, but actually the size of an atomic plane is needed. Despite the method in Ref. 24, which transforms the GEVP to a NEVP of reduced size, it calculates the whole surface Green's function of the size of the unit cell and at the same

[a)]Electronic mail: huangjun@eee.hku.hk.
[b)]Electronic mail: w-chew@uiuc.edu.
[c)]Electronic mail: ymwu@eee.hku.hk.
[d)]Electronic mail: jianglj@hku.hk.

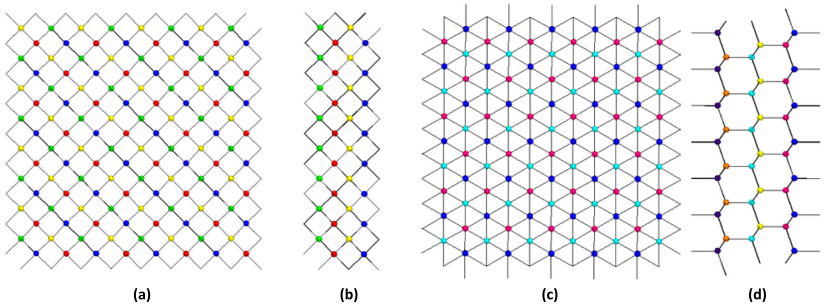

FIG. 1. Cross section and profile of a unit cell for silicon nanowires along the [100] direction (a and b) and the [111] direction (c and d). The unit cell consists of four and six atomic planes, respectively. Different planes are denoted with different colors.

time involves inverting a matrix of the size of the unit cell that incurs additional cost. In fact, due to the short-range interactions, it is possible to compress the Hamiltonian matrix of a unit cell to that of an atomic plane. Then, after some slight modifications, the decimation method and the eigenvalue methods can be employed to calculate the surface Green's function (and then the self energy). The gain is obvious, as we are now dealing with a much smaller matrix (approximately by a factor of 1/4 for [100] and 1/6 for [111] and [112]).

In Sec. II, the condensation of the Hamiltonian matrix (in the nearest neighbor tight-binding schemes) for the semi-infinite leads is derived in detail, followed by the applications of the decimation approach and the eigenvalue approach, respectively. Some numerical examples are provided in Sec. III to show the accuracy and the efficiency. In Sec. IV, we give a brief summary and also some possible extensions. In Appendix, we show that the methods in this paper can be generalized to second-near (and third-near) neighbor interaction schemes.

II. DESCRIPTION OF THE METHODS

A. Condensation of the Hamiltonian matrix

A typical two-probe system as illustrated in Fig. 2 is considered here, where the system Hamiltonian is divided into $\mathbf{H}_L$, $\mathbf{H}_D$, and $\mathbf{H}_R$. We focus on the self energy calculation for the right lead as the left lead can be done similarly. The Green's function matrix $\mathbf{g}_R$ for the right lead at energy point E is defined as

$$(E\mathbf{I} - \mathbf{H}_R)\mathbf{g}_R = \mathbf{I}, \tag{1}$$

where $\mathbf{I}$ is the identity matrix. We have assumed orthogonal basis; non-orthogonal basis case can be done by replacing $E\mathbf{I}$ with overlap matrix $E\mathbf{S}$.

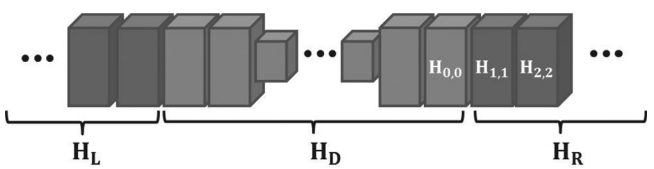

FIG. 2. Schematic representation of a two-probe system. The system consists of a device part with Hamiltonian $\mathbf{H}_D$ and two semi-infinite leads with Hamiltonian $\mathbf{H}_L$ and $\mathbf{H}_R$. The system is divided into many atomic planes and the right lead is described with atomic plane Hamiltonian $\mathbf{H}_{p,p}$ $(p = 1, 2, \ldots)$ as illustrated.

Take the nearest neighbor interaction scheme, for example (the generalization to second-near or third-near neighbor interaction schemes is discussed in Appendix), the matrix $\mathbf{H}_R$ can be written in a block tridiagonal form and $\mathbf{g}_R$ is usually a full matrix,

$$\mathbf{H}_R = \begin{pmatrix} \mathbf{H}_{1,1} & \mathbf{H}_{1,2} & \mathbf{0} & \cdots \\ \mathbf{H}_{1,2}^{\dagger} & \mathbf{H}_{2,2} & \mathbf{H}_{2,3} & \cdots \\ \mathbf{0} & \mathbf{H}_{2,3}^{\dagger} & \mathbf{H}_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

$$\mathbf{g}_R = \begin{pmatrix} \mathbf{g}_{1,1} & \mathbf{g}_{1,2} & \mathbf{g}_{1,3} & \cdots \\ \mathbf{g}_{2,1} & \mathbf{g}_{2,2} & \mathbf{g}_{2,3} & \cdots \\ \mathbf{g}_{3,1} & \mathbf{g}_{3,2} & \mathbf{g}_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{2}$$

where $\mathbf{H}_{p,q}$ with $p = q$ denotes the on-site Hamiltonian for atomic plane p and $\mathbf{H}_{p,q}$ with $p \neq q$ denotes the coupling Hamiltonian between atomic plane p and q, $\mathbf{H}_{p,q}^{\dagger}$ is the Hermitian conjugate of $\mathbf{H}_{p,q}$. We have made use of $\mathbf{H}_{1,0} = 0$ since the semi-infinite lead terminates at plane 1.

According to Eqs. (1) and (2), the Green's function $\mathbf{g}_{p,q}$ for $q = 1$ should satisfy the following equation,

$$\begin{pmatrix} E\mathbf{I}_{1,1} - \mathbf{H}_{1,1} & -\mathbf{H}_{1,2} & \mathbf{0} & \cdots \\ -\mathbf{H}_{1,2}^{\dagger} & E\mathbf{I}_{2,2} - \mathbf{H}_{2,2} & -\mathbf{H}_{2,3} & \cdots \\ \mathbf{0} & -\mathbf{H}_{2,3}^{\dagger} & E\mathbf{I}_{3,3} - \mathbf{H}_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \times \begin{pmatrix} \mathbf{g}_{1,1} \\ \mathbf{g}_{2,1} \\ \mathbf{g}_{3,1} \\ \vdots \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{1,1} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \end{pmatrix}. \tag{3}$$

Assuming that a unit cell in the lead consists of P atomic planes, the Hamiltonian then repeats every P atomic planes, i.e.,

$$\mathbf{H}_{nP+p,nP+q} = \mathbf{H}_{p,q}, \quad (p = 1,2,\cdots,P;\ q = p,\ p+1;\ n = 1,2,\cdots). \tag{4}$$

Utilizing this fact, Eq. (3) can be rewritten in the following format with matrix partitioning,

$$\begin{pmatrix} E\mathbf{I}_{1,1}-\mathbf{H}_{1,1} & \mathbf{B} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{B}^{\dagger} & \mathbf{C} & \mathbf{D} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{D}^{\dagger} & E\mathbf{I}_{1,1}-\mathbf{H}_{1,1} & \mathbf{B} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{B}^{\dagger} & \mathbf{C} & \mathbf{D} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \bullet \begin{pmatrix} \mathbf{g}_{1,1} \\ \mathbf{g}_{2\sim P,1} \\ \mathbf{g}_{P+1,1} \\ \mathbf{g}_{(P+2)\sim 2P,1} \\ \mathbf{g}_{2P+1,1} \\ \vdots \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{1,1} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \end{pmatrix}, \tag{5}$$

where we have defined new blocks,

$$\mathbf{B} = (-\mathbf{H}_{1,2},\ \mathbf{0},\ \cdots,\ \mathbf{0}), \quad \mathbf{D} = \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ -\mathbf{H}_{P,P+1} \end{pmatrix}, \tag{6}$$

$$\mathbf{C} = \begin{pmatrix} E\mathbf{I}_{2,2}-\mathbf{H}_{2,2} & -\mathbf{H}_{2,3} & \cdots & \mathbf{0} \\ -\mathbf{H}_{2,3}^{\dagger} & E\mathbf{I}_{3,3}-\mathbf{H}_{3,3} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & E\mathbf{I}_{P,P}-\mathbf{H}_{P,P} \end{pmatrix}, \tag{7}$$

$$\mathbf{g}_{2\sim P,1} = \begin{pmatrix} \mathbf{g}_{2,1} \\ \mathbf{g}_{3,1} \\ \vdots \\ \mathbf{g}_{P,1} \end{pmatrix}, \quad \mathbf{g}_{(P+2)\sim 2P,1} = \begin{pmatrix} \mathbf{g}_{P+2,1} \\ \mathbf{g}_{P+3,1} \\ \vdots \\ \mathbf{g}_{2P,1} \end{pmatrix}. \tag{8}$$

Eliminating $\mathbf{g}_{2\sim P,1}, \mathbf{g}_{(P+2)\sim 2P,1}, \ldots$, in Eq. (5) results in,

$$\begin{pmatrix} E\mathbf{I}_{1,1}-\Xi_s & -\Pi & \mathbf{0} & \cdots \\ -\Pi^{\dagger} & E\mathbf{I}_{1,1}-\Xi & -\Pi & \cdots \\ \mathbf{0} & -\Pi^{\dagger} & E\mathbf{I}_{1,1}-\Xi & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \mathbf{g}_{1,1} \\ \mathbf{g}_{P+1,1} \\ \mathbf{g}_{2P+1,1} \\ \vdots \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{1,1} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \end{pmatrix}, \tag{9}$$

where,

$$\Xi_s = \mathbf{H}_{1,1} + \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^{\dagger}, \tag{10}$$

$$\Xi = \mathbf{H}_{1,1} + \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^{\dagger} + \mathbf{D}^{\dagger}\mathbf{C}^{-1}\mathbf{D}, \tag{11}$$

$$\Pi = \mathbf{B}\mathbf{C}^{-1}\mathbf{D}. \tag{12}$$

From Eq. (9), we can identify a condensed Hamiltonian that only consists of planes $p = nP+1$ $(n = 0, 1, \cdots)$, i.e.,

$$\mathbf{H}_{cnd} = \begin{pmatrix} \Xi_s & \Pi & \mathbf{0} & \cdots \\ \Pi^{\dagger} & \Xi & \Pi & \cdots \\ \mathbf{0} & \Pi^{\dagger} & \Xi & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{13}$$

where the blocks are of the size $\sim(N/P)\times(N/P)$ with N being the matrix dimension of a unit cell. Note that the condensed on-site Hamiltonian Ξ_s of plane 1 differs from condensed on-site Hamiltonian Ξ of plane $p = nP+1$ $(n = 1, 2, \ldots)$, as Ξ_s only includes the influences of right side planes (plane 2 to P) while Ξ includes the influences of both sides (plane $(n-1)P+2$ to nP and plane $nP+2$ to $(n+1)P$). The condensed coupling Hamiltonian Π connects plane $p = nP+1$ to plane $p = (n+1)P+1$ directly.

The problem now is to evaluate the expressions of Ξ_s, Ξ, and Π as shown in Eqs. (10)–(12). This requires inversion of matrix $\mathbf{C}$ of the size $\sim\left(\frac{P-1}{P}N\right)\times\left(\frac{P-1}{P}N\right)$, which can be done efficiently since it is highly sparse. Alternatively, we can avoid the full inversion by noticing that the matrix $\mathbf{B}$ or $\mathbf{D}$ consists of only one non-zero block and thus several blocks in $\mathbf{C}^{-1}$ are actually needed. In fact, by denoting $\mathbf{C}^{-1}$ as

$$\mathbf{C}^{-1} = \begin{pmatrix} \tilde{\mathbf{C}}_{2,2} & \tilde{\mathbf{C}}_{2,3} & \cdots & \tilde{\mathbf{C}}_{2,P} \\ \tilde{\mathbf{C}}_{3,2} & \tilde{\mathbf{C}}_{3,3} & \cdots & \tilde{\mathbf{C}}_{3,P} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{C}}_{P,2} & \tilde{\mathbf{C}}_{P,3} & \cdots & \tilde{\mathbf{C}}_{P,P} \end{pmatrix}, \tag{14}$$

due to Eq. (6), we find that only $\tilde{\mathbf{C}}_{2,2}$, $\tilde{\mathbf{C}}_{2,P}$, and $\tilde{\mathbf{C}}_{P,P}$ are relevant. Furthermore, these blocks can be calculated efficiently with forward and backward recursions since $\mathbf{C}$ is block tridiagonal. The details are as follows:

ALGORITHM 0 (Recursive Condensation of the Hamiltonian Matrix):

1. $\tilde{\mathbf{H}}_{P,P} = (E\mathbf{I}_{P,P} - \mathbf{H}_{P,P})^{-1}$
2. For $p = P-1, P-2, \cdots, 2$ (in this order), do {
3. $\tilde{\mathbf{H}}_{p,p} = (E\mathbf{I}_{p,p} - \mathbf{H}_{p,p} - \mathbf{H}_{p,p+1}\tilde{\mathbf{H}}_{p+1,p+1}\mathbf{H}_{p,p+1}^{\dagger})^{-1}$
4. $\tilde{\mathbf{H}}_{p,P} = \tilde{\mathbf{H}}_{p,p}\mathbf{H}_{p,p+1}\tilde{\mathbf{H}}_{p+1,P}$}
5. $\tilde{\mathbf{C}}_{2,2} = \tilde{\mathbf{H}}_{2,2}$, $\tilde{\mathbf{C}}_{2,P} = \tilde{\mathbf{H}}_{2,P}$
6. For $p = 3, \cdots, P$ (in this order), do {
7. $\tilde{\mathbf{C}}_{p,p} = \tilde{\mathbf{H}}_{p,p} + \tilde{\mathbf{H}}_{p,p}(\mathbf{H}_{p,p-1}\tilde{\mathbf{C}}_{p-1,p-1}\mathbf{H}_{p,p-1}^{\dagger})\tilde{\mathbf{H}}_{p,p}$}
8. Obtain $\Xi_s = \mathbf{H}_{1,1} + \mathbf{H}_{1,2}\tilde{\mathbf{C}}_{2,2}\mathbf{H}_{1,2}^{\dagger}$
9. Obtain $\Xi = \Xi_s + \mathbf{H}_{P,P+1}^{\dagger}\tilde{\mathbf{C}}_{P,P}\mathbf{H}_{P,P+1}$
10. Obtain $\Pi = \mathbf{H}_{1,2}\tilde{\mathbf{C}}_{2,P}\mathbf{H}_{P,P+1}$

With this condensed Hamiltonian (13) of reduced size, we are now ready to calculate the self energy either by the iterative approach or the eigenvalue approach as described separately in the following.

B. Iterative approach

As seen from matrix (13), the translational invariance is broken by the first block. Nevertheless, we can still apply the decimation method[18] to the chain in Eq. (9). The implementation is summarized into ALGORITHM I (for details, see supplementary material[25]). We want to emphasize that, although the decimation can be directly applied to the original chain in Eq. (3), our implementation is systematic and much simpler, as now all the layers (except the first one) in Eq. (13) are made identical no matter how many different atomic planes there are in a unit cell.

C. Eigenvalue approach

The eigenvalue approach,[16,17] however, cannot be directly applied to the chain in Eq. (13). Fortunately, we found that it can still be applied to the chain starting from layer 2, and the extra treatment of layer 1 can be done without too much effort.

First, we define a new semi-infinite chain that starts from layer 2 of Eq. (13). By using Bloch wave condition

$$\Psi_{p+P} = \lambda \Psi_p, \tag{15}$$

where $\lambda = e^{ikd}$ with k real (complex) for propagating (evanescent) modes, we have the following equation for Bloch waves,

$$-\lambda^{-1}\Pi^{\dagger}\Psi_p + (E\mathbf{I}_{1,1} - \Xi)\Psi_p - \lambda\Pi\Psi_p = \mathbf{0}. \tag{16}$$

This equation can be solved by transforming to a GEVP of size $2N_1$ (N_1 is the size of Ξ), i.e.,

$$\begin{pmatrix} \mathbf{0} & \mathbf{I}_{1,1} \\ -\mathbf{T}^{\dagger} & -\mathbf{D} \end{pmatrix} \begin{pmatrix} \Psi_p \\ \Psi_{p+P} \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{I}_{1,1} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{pmatrix} \begin{pmatrix} \Psi_p \\ \Psi_{p+P} \end{pmatrix}, \tag{17}$$

where the blocks are

$$\mathbf{D} = E\mathbf{I}_{1,1} - \Xi, \quad \mathbf{T} = -\Pi. \tag{18}$$

Second, we define a new Green's function $\mathbf{g}'$ for this new semi-infinite chain, the blocks $\mathbf{g}'_{p,q}$ for $q = 1$ should satisfy the following:

$$(E\mathbf{I}_{1,1} - \Xi)\mathbf{g}'_{1,1} = \mathbf{I}_{1,1} + \Pi\mathbf{g}'_{P+1,1}, \tag{19}$$

$$(E\mathbf{I}_{1,1} - \Xi)\mathbf{g}'_{P+1,1} = \Pi^{\dagger}\mathbf{g}'_{1,1} + \Pi\mathbf{g}'_{2P+1,1}, \tag{20}$$

$$\cdots$$

Then $\mathbf{g}'_{1,1}$ can be expanded through Bloch modes of the chain,

$$\mathbf{g}'_{1,1} = \mathbf{U}^{+}\mathbf{C}^{+}, \tag{21}$$

where matrix $\mathbf{U}^{+}$ (of size $N_1 \times M$) consists of M right-going normalized Bloch vectors constructed from the first N_1 elements of the solution of Eq. (17), and matrix $\mathbf{C}^{+}$ (of size $M \times N_1$) consists of N_1 vectors of corresponding expansion coefficients, i.e.,

$$\mathbf{U}^{+} = (\mathbf{u}_1^{+}, \mathbf{u}_2^{+}, \cdots, \mathbf{u}_M^{+}), \tag{22}$$

$$\mathbf{C}^{+} = (\mathbf{c}_1^{+}, \mathbf{c}_2^{+}, \cdots, \mathbf{c}_{N_1}^{+}). \tag{23}$$

Since the waves go outward from the δ source, we can express $\mathbf{g}'_{P+1,1}$ as,

$$\mathbf{g}'_{P+1,1} = \mathbf{U}^{+}\Lambda^{+}\mathbf{C}^{+}, \tag{24}$$

where the propagator Λ^{+} is a $M \times M$ diagonal matrix with elements

$$\Lambda^{+}_{mm} = \lambda^{+}_{m}. \tag{25}$$

By defining pseudo-inverse $\tilde{\mathbf{U}}^{+}$ of $\mathbf{U}^{+}$, i.e.,

$$\tilde{\mathbf{U}}^{+}\mathbf{U}^{+} = \mathbf{I}, \tag{26}$$

and using Eq. (24), $\mathbf{g}'_{P+1,1}$ can be related to $\mathbf{g}'_{1,1}$ through the following way

$$\mathbf{g}'_{P+1,1} = \mathbf{U}^{+}\Lambda^{+}\tilde{\mathbf{U}}^{+}\mathbf{U}^{+}\mathbf{C}^{+} = \mathbf{U}^{+}\Lambda^{+}\tilde{\mathbf{U}}^{+}\mathbf{g}'_{1,1} = \mathbf{F}\mathbf{g}'_{1,1}, \tag{27}$$

where we have defined a new propagator

$$\mathbf{F} = \mathbf{U}^{+}\Lambda^{+}\tilde{\mathbf{U}}^{+}. \tag{28}$$

Similarly, the following holds

$$\mathbf{g}'_{2P+1,1} = \mathbf{F}\mathbf{g}'_{P+1,1}. \tag{29}$$

Putting Eqs. (27) and (29) into Eqs. (19) and (20), we have

$$(E\mathbf{I}_{1,1} - \Xi - \Pi\mathbf{F})\mathbf{g}'_{1,1} = \mathbf{I}_{1,1}, \tag{30}$$

$$(E\mathbf{I}_{1,1} - \Xi - \Pi\mathbf{F})\mathbf{F}\mathbf{g}'_{1,1} = \Pi^{\dagger}\mathbf{g}'_{1,1}. \tag{31}$$

From above two we can solve for the surface Green's function $\mathbf{g}'_{1,1}$, which is

$$\mathbf{g}'_{1,1} = \mathbf{F}\Pi^{\dagger -1}. \tag{32}$$

In the case when $\Pi^{\dagger}$ is not invertable, we solve for self energy directly, i.e.,

$$\Sigma' = \Pi\mathbf{g}'_{1,1}\Pi^{\dagger} = \Pi\mathbf{F}. \tag{33}$$

Finally, the surface Green's function for the original chain (including layer 1 of Eq. (13)) is obtained as,

$$\mathbf{g}_{1,1} = (E\mathbf{I}_{1,1} - \Xi_s - \Sigma')^{-1}, \tag{34}$$

and the self energy is constructed using,

$$\Sigma = \mathbf{H}_{0,1}\mathbf{g}_{1,1}\mathbf{H}^{\dagger}_{0,1}. \tag{35}$$

We implemented the above approach (the approach is self-consistent since we can verify that Eqs. (21), (24), and (29) satisfy Eqs. (19) and (20) by direct substitution) in the following way,

ALGORITHM II (Eigenvalue method):

0. Do ALGORITHM 0.
1. Let $\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{I}_{1,1} \\ -\mathbf{T}^{\dagger} & -\mathbf{D} \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} \mathbf{I}_{1,1} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{pmatrix}$.
2. Instead of solving a generalized eigenvalue problem $\mathbf{A}\Psi = \lambda\mathbf{B}\Psi$, we resort to a normal eigenvalue problem by constructing $\tilde{\mathbf{A}} = (\mathbf{A} - \sigma\mathbf{B})^{-1}\mathbf{B}$, where σ is a shift. Note that the 2×2 block matrix $(\mathbf{A} - \sigma\mathbf{B})$ can be inverted efficiently by using the Schur complement block.[20]
3. Solve the normal eigenvalue problem $\tilde{\mathbf{A}}\Psi = \tilde{\lambda}\Psi$, obtain the eigenpairs $(\tilde{\lambda}, \Psi)$.
4. Obtain the eigenpairs of the original problem: $(\lambda = \tilde{\lambda}^{-1} + \sigma, \Psi)$.
5. Retrieve all the eigenpairs corresponding to the right-going propagating modes with $|\lambda| = 1$; Retrieve a part of the eigenpairs corresponding to the right-going evanescent modes with $\epsilon < |\lambda| < 1$, where ϵ can be truncated to

include only slowly decaying evanescent modes. Construct an $N_1 \times M$ matrix $\mathbf{U}^+$ and an $M \times M$ diagonal matrix Λ^+ from these eigenpairs.
6. Obtain pseudo-inverse $\tilde{\mathbf{U}}^+$ of $\mathbf{U}^+$ by factorizing $\mathbf{U}^+ = \mathbf{QR}$ and solving $\mathbf{R}\tilde{\mathbf{U}}^+ = \mathbf{Q}^\dagger$.
7. Construct $\mathbf{F}$ according to Eq. (28). Solve $(E\mathbf{I}_{1,1} - \Xi_s - \Pi\mathbf{F})\mathbf{Y} = \mathbf{H}_{0,1}^\dagger$ for $\mathbf{Y}$. Note that this is the only step where layer 1 (Ξ_s) comes in.
8. Obtain the self energy $\Sigma = \mathbf{H}_{0,1}\mathbf{Y}$.

D. Computational cost

To reduce the Hamiltonian to Eq. (13), as shown in ALGORITHM 0, we need $P-1$ inversions of the small matrices of the size $\sim(N/P)$. The cost is $(P-1)\times O((N/P)^3)$, which is very cheap.

Once will have Eq. (13), the computational cost of ALGORITHM I is $(M+1)\times O((N/P)^3)$ if the process converges in M steps (usually 20 to 50 steps, depending on the value of the introduced infinitesimal positive quantity η). This is a tremendous reduction compared with the original decimation method,[18] where the complexity is $(M+1)\times O(N^3)$ (here, we assume that the inversions are carried out for matrices of the size of a unit cell). The computational cost of ALGORITHM II is $O((2N/P)^3)+O((N/P)^3)$, where the first term is due to step 3, and the second term due to steps 2, 6, and 7. This is also a significant improvement over the original eigenvalue approach,[15–17] the cost of which is about $O((2N)^3)+O(N^3)$. Note that $P=4$ for [100] orientation and $P=6$ for [111] and [112].

III. RESULTS AND DISCUSSION

The testing examples are rectangular silicon nanowires. The representation of the Hamiltonian matrix is through $sp^3d^5s^*$ tight binding scheme with nearest neighbor interaction (10 orbits per atom without spin-orbit coupling and 20 orbits per atom with spin-orbit coupling).[23] The dangling sp^3 hybridized bonds at the surfaces are passivated using hydrogen-like atoms.[26] This tight binding scheme has been widely employed to study nanowire transistors.

First, to validate our methods, we have calculated the transmission spectrum of an unbiased perfect silicon nanowire with Green's function approach.[1,2] The self energies involved were obtained by ALGORITHM I and II, respectively. The results are shown in Fig. 3, also shown are the $E-k$ dispersion and density of states (DOS) calculated for an infinite periodic nanowire. It is clearly seen that the transmission is an integer over the whole band and it steps up or down when a transmission channel is opened or closed. The transition points of the transmission match perfectly with the positions of the one dimensional DOS peaks (van Hove singularities), indicating that our transmission calculation is reliable, and in turn, validating our self energy calculations. Note that to explain the transmission in valance band, it is better to trace through the $E-k$ diagram since there are additional DOS peaks which do not correspond to the van Hove singularities and the number of transmission channel remains unchanged when one goes through these peaks. Similar phenomena can be observed for a [111] oriented silicon nanowire (see supplementary material[25]).

Next, to show the efficiency, we list the run times of our algorithms along with those of the existing methods in Table I. For the iterative methods (methods 1, 2, and 3), we choose $\eta = 10^{-9}$ eV so that the iterative processes converge in a certain number of steps. It is seen that ALGORITHM I can greatly speed up the simulation compared with the fastest iterative one, i.e., method 2. It should be mentioned that in this work, we implement method 2 by inverting the matrices of the unit cell. In particular, for [100] and [111] directions, we gain an acceleration factor of about 40 to 80. Note that for these two cases, we have implemented sparse matrix operations in method 2 as the matrices involved have many zero blocks. While among the eigenvalue approaches (methods 4, 5, and 6), ALGORITHM II is the best and it slightly outperforms the fastest existing one, i.e., method 5. Note that we have implemented sparse matrix operations in method 5 so that the matrix inversion involved is very efficient. To

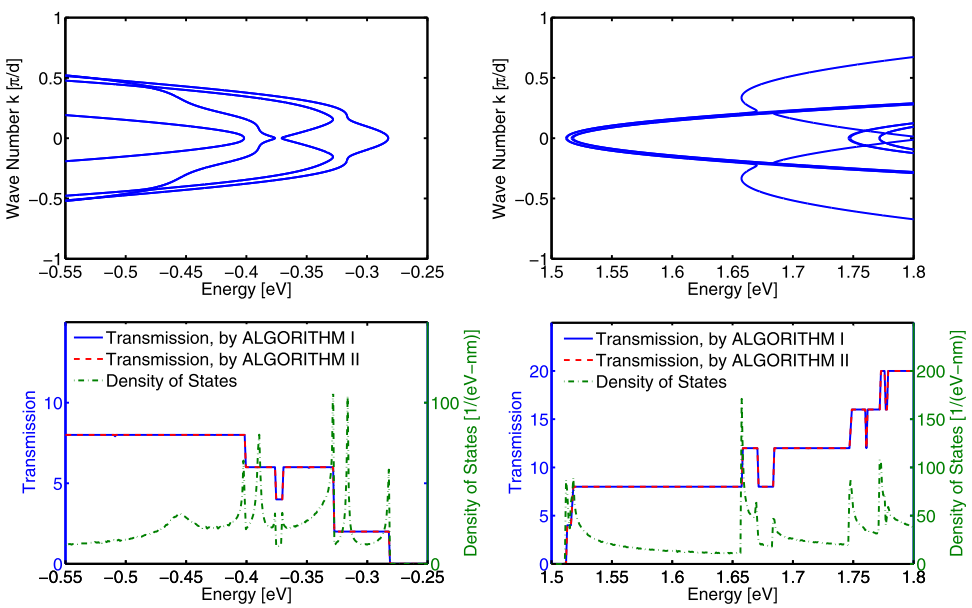

FIG. 3. Top: $E-k$ relation, bottom: transmission spectrum and DOS, for an ideal [100] oriented silicon nanowire with cross-section $\sim 2\,\text{nm} \times 2\,\text{nm}$. Left: for valance band, and spin-orbit coupling is included in the calculation, right: for conduction band, and spin-orbit coupling is not included in the calculation. No external bias is applied. The transmissions calculated by the two methods in this paper lie almost on top of each other.

TABLE I. List of run times (in seconds) for self energy evaluation in one energy point for silicon nanowires with cross section $\sim 2\,\text{nm} \times 2\,\text{nm}$. Calculations are carried out for three crystal directions ([110], [100], and [111]) and for two basis sets (without and with spin-orbital coupling). Six methods are implemented (in MATLAB). The quantities in the brackets are the speed degradation factors compared with the fastest method. The simulations are performed on an Intel Xeon processor (restricted to four cores, 2.66 GHz).

Orientation	[110]	[100]	[111]
Number of planes p.u.c	2	4	6
Number of atoms p.u.c	88	128	208
Matrix size p.u.c	880	1280	2080
1. Iterative method[a]	1816 (386×)	819.4 (394×)	239.9 (79.2×)
2. Decimation (Ref. 18)	34.3 (7.3×)	86.6 (41.6×)	169.2 (55.8×)
3. ALGORITHM I	4.71	2.08	3.03
4. NEVP method (Ref. 20)	5.04 (5.5×)	11.1 (21.8×)	38.9 (45.2×)
5. Advanced NEVP (Ref. 24)	1.52 (1.7×)	1.77 (3.5×)	3.43 (4.0×)
6. ALGORITHM II	0.92	0.51	0.86
Matrix size p.u.c	1760	2560	4160
1. Iterative method[b]	13475 (409×)	5468 (390×)	1590 (83.4×)
2. Decimation (Ref. 18)	262.6 (8.0×)	722.0 (51.5×)	1473 (77.2×)
3. ALGORITHM I	32.91	14.02	19.07
4. NEVP method (Ref. 20)	108.6 (7.1×)	314.2 (40.9×)	1302 (92.4×)
5. Advanced NEVP (Ref. 24)	22.36 (1.5×)	18.63 (2.4×)	36.47 (2.6×)
6. ALGORITHM II	15.26	7.69	14.09

[a]This is done by repetitive use of relations, $\mathbf{g}_{p,p}^{(n)} = \left(E^{*}\mathbf{I}_{p,p} - \mathbf{H}_{p,p} - \mathbf{H}_{p,p+1}\mathbf{g}_{p+1,p+1}^{(n)}\mathbf{H}_{p,p+1}^{\dagger}\right)^{-1}$, for $p = P, P-1, \ldots, 1$, and $\mathbf{g}_{P+1,P+1}^{(n)} = \mathbf{g}_{1,1}^{(n-1)}$.
[b]As described in footnote (a) above.

include spin-orbit interaction, which is important for hole transport, the computational cost is significantly increased. The reason is two fold, one is that the number of orbits doubles, the other is the introduction of complex operations (in the eigenvalue approaches) as a result of complex Hamiltonian elements. Generally speaking, ALGORITHMS I and II are comparable in terms of speed when spin-orbit coupling is included; ALGORITHM II shows advantage when spin-orbit coupling is not included due to the real arithmetic, which is not the case in ALGORITHM I since a small imaginary part is introduced to ensure convergence.

IV. CONCLUSIONS

In order to efficiently simulate quantum transport in nanodevices within NEGF formalism, we have proposed two algorithms for the fast evaluation of self-energy matrices in tight binding schemes. The efficiency of the algorithms is based on constructing a condensed Hamiltonian with reduced size for the semi-infinite leads. The condensation successfully takes advantage of the crystal structures together with the short-range interactions of tight binding schemes. The reliability of our methods has been demonstrated by studying the transmission of an ideal silicon nanowire in the nearest neighbor interaction scheme. Extensive numerical examples and comparisons have shown that our methods can speed up the decimation approach by 7 to 80 times and can also outperform the advanced eigenvalue approach by several times.

Our methods are particularly useful when the unit cell in the leads is made very long due to the presence of doping atoms. This situation is very common in nano-electronics nowadays as the doping density (per nanometer) in the leads is usually very low as a result of the ultra-small cross sections. Furthermore, our methods can be applied to *ab initio* models as long as the interaction range is short compared with the unit cell length.

ACKNOWLEDGMENTS

This work was supported in part by the Research Grants Council of Hong Kong (GRF 711609, 711508, and 711511), in part by the University Grants Council of Hong Kong (Contract No. AoE/P-04/08). The authors would like to thank Dr. Hang Xie, Dr. Min Tang, and Dr. Chi Yung Yam for useful feedback and discussions.

APPENDIX: GENERALIZATION TO THE SECOND AND THIRD-NEAR NEIGHBOR (2NN AND 3NN) INTERACTION SCHEMES

The methods proposed in this paper are demonstrated through the nearest neighbor interaction scheme. In the next, we will show that they can be generalized to 2NN and 3NN interaction schemes. Take 2NN interaction, for example (3NN can be done in the same spirit), the Hamiltonian matrix in terms of atomic planes takes the form,

$$\mathbf{H}_R = \begin{pmatrix} \mathbf{H}_{1,1} & \mathbf{H}_{1,2} & \mathbf{H}_{1,3} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{H}_{1,2}^{\dagger} & \mathbf{H}_{2,2} & \mathbf{H}_{2,3} & \mathbf{H}_{2,4} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{H}_{1,3}^{\dagger} & \mathbf{H}_{2,3}^{\dagger} & \mathbf{H}_{3,3} & \mathbf{H}_{3,4} & \mathbf{H}_{3,5} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{H}_{2,4}^{\dagger} & \mathbf{H}_{3,4}^{\dagger} & \mathbf{H}_{4,4} & \mathbf{H}_{4,5} & \mathbf{H}_{4,6} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{H}_{3,5}^{\dagger} & \mathbf{H}_{4,5}^{\dagger} & \mathbf{H}_{5,5} & \mathbf{H}_{5,6} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{H}_{4,6}^{\dagger} & \mathbf{H}_{5,6}^{\dagger} & \mathbf{H}_{6,6} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \quad \text{(A1)}$$

which can be rewritten in a block tridiagonal form like that in Eq. (2),

$$\mathbf{H}_R = \begin{pmatrix} \bar{\mathbf{H}}_{1,1} & \bar{\mathbf{H}}_{1,2} & \mathbf{0} & \cdots \\ \bar{\mathbf{H}}_{1,2}^{\dagger} & \bar{\mathbf{H}}_{2,2} & \bar{\mathbf{H}}_{2,3} & \cdots \\ \mathbf{0} & \bar{\mathbf{H}}_{2,3}^{\dagger} & \bar{\mathbf{H}}_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \quad \text{(A2)}$$

where the blocks are

$$\bar{\mathbf{H}}_{1,1} = \begin{pmatrix} \mathbf{H}_{1,1} & \mathbf{H}_{1,2} \\ \mathbf{H}_{1,2}^{\dagger} & \mathbf{H}_{2,2} \end{pmatrix}, \quad \bar{\mathbf{H}}_{22} = \begin{pmatrix} \mathbf{H}_{3,3} & \mathbf{H}_{3,4} \\ \mathbf{H}_{3,4}^{\dagger} & \mathbf{H}_{4,4} \end{pmatrix},$$
$$\bar{\mathbf{H}}_{33} = \begin{pmatrix} \mathbf{H}_{5,5} & \mathbf{H}_{5,6} \\ \mathbf{H}_{5,6}^{\dagger} & \mathbf{H}_{6,6} \end{pmatrix}, \bar{\mathbf{H}}_{12} = \begin{pmatrix} \mathbf{H}_{1,3} & \mathbf{0} \\ \mathbf{H}_{2,3} & \mathbf{H}_{2,4} \end{pmatrix}, \quad \text{(A3)}$$
$$\bar{\mathbf{H}}_{23} = \begin{pmatrix} \mathbf{H}_{3,5} & \mathbf{0} \\ \mathbf{H}_{4,5} & \mathbf{H}_{4,6} \end{pmatrix}.$$

Now, the method in Sec. II A can be applied to Eq. (A2) to condense the Hamiltonian matrix into a small one which consists only the planes $p = nP + 1$ and $p = nP + 2$, where $n = 0, 1, 2, \ldots$. Thus, we gain a size reduction factor of

$\sim 1/2$ for [100] orientation and $\sim 1/3$ for [111] and [112]. With the condensed Hamiltonian matrix, the self energy matrix can be evaluated with the methods described in Secs. II B and II C.

[1]S. Datta, *Electronic Transport in Mesoscopic Systems* (Cambridge University Press, 1995).
[2]S. Datta, *Quantum Transport: Atom to Transistor* (Cambridge University Press, 2005).
[3]A. Svizhenko, M. P. Anantram, T. R. Govindan, B. Biegel, and R. Venugopal, "Two-dimensional quantum mechanical modeling of nanotransistors," J. Appl. Phys. **91**(4), 2343–2354 (2002).
[4]J. Wang, E. Polizzi, and M. Lundstrom, "A three-dimensional quantum simulation of silicon nanowire transistors with the effective-mass approximation," J. Appl. Phys. **96**(4), 2192–2203 (2004).
[5]E. Polizzi and N. Ben Abdallah, "Subband decomposition approach for the simulation of quantum electron transport in nanostructures," J. Comput. Phys. **202**, 150–180 (2005).
[6]D. Mamaluy, M. Sabathil, and P. Vogl, "Efficient method for the calculation of ballistic quantum transport," J. Appl. Phys. **93**(8), 4628–4633 (2003).
[7]D. Mamaluy, D. Vasileska, M. Sabathil, T. Zibold, and P. Vogl, "Contact block reduction method for ballistic transport and carrier densities of open nanostructures," Phys. Rev. B **71**, 245321 (2005).
[8]G. Mil'nikov, N. Mori, Y. Kamakura, and T. Ezaki, "R-matrix theory of quantum transport and recursive propagation method for device simulations," J. Appl. Phys. **104**, 044506 (2008).
[9]G. Mil'nikov, N. Mori, and Y. Kamakura, "R-matrix method for quantum transport simulations in discrete systems," Phys. Rev. B **79**, 235337 (2009).
[10]J. Z. Huang, W. C. Chew, M. Tang, and L. Jiang, "Efficient simulation and analysis of quantum ballistic transport in nanodevices with AWE," IEEE Trans. Electron Devices **59**(2), 468–476 (2012).
[11]J. Taylor, H. Guo, and J. Wang, "*Ab initio* modeling of quantum transport properties of molecular electronic devices," Phys. Rev. B **63**, 245407 (2001).
[12]T. B. Boykin, "Recent developments in tight-binding approaches for nanowires," J. Comput. Electron. **8**, 142152 (2009).
[13]J. Velev and W. Butler, "On the equivalence of different techniques for evaluating the Green function for a semi-infinite system using a localized basis," J. Phys.: Condens. Matter **16**, R637–R657 (2004).
[14]L. M. Falicov and F. Yndurain, "Model calculation of the electronic structure of a (111) surface in a diamond-structure solid," J. Phys. C **8**, 147 (1975).
[15]C. Rivas and R. Lake, "Non–equilibrium Green function implementation of boundary conditions for full band simulations of substrate-nanowire structures," Phys. Status Solidi B **239**(1), 94–102 (2003).
[16]P. A. Khomyakov and G. Brocks, "Real-space finite-difference method for conductance calculations," Phys. Rev. B **70**, 195402 (2004).
[17]P. A. Khomyakov, G. Brocks, V. Karpan, M. Zwierzycki, and P. J. Kelly, "Conductance calculations for quantum wires and interfaces: mode matching and Green's functions," Phys. Rev. B **72**, 035450 (2005).
[18]M. P. L. Sancho, J. M. L. Sancho, and J. Rubio, "Highly convergent schemes for the calcualtion of bulk and surface Green functions," J. Phys. F: Met. Phys. **15**, 851–858 (1985).
[19]A. D. Carlo, M. Gheorghe, P. Lugli, M. Sternberg, G. Seifert, and T. Frauenheim, "Theoretical tools for transport in molecular nanostructures," Physica B **314**, 86–90 (2002).
[20]H. H. B. Sørensen, "Computational aspects of electronic transport in nanoscale devices," Ph.D dissertation (Technical University of Denmark, 2008).
[21]H. H. B. Sørensen, P. C. Hansen, D. E. Petersen, S. Skelboe, and K. Stokbro, "Krylov subspace method for evaluating the self-energy matrices in electron transport calculations," Phys. Rev. B **77**, 155301 (2008).
[22]J. A. Driscoll and K. Varga, "Calculation of self-energy matrices using complex absorbing potentials in electron transport calculations," Phys. Rev. B **78**, 245118 (2008).
[23]T. B. Boykin, G. Klimeck, and F. Oyafuso, "Valence band effective-mass expressions in the $sp^3d^5s^*$ empirical tight-binding model applied to a Si and Ge parametrization," Phys. Rev. B **69**, 115201 (2004).
[24]M. Luisier, A. Schenk, W. Fichtner, and G. Klimeck, "Atomistic simulation of nanowires in the $sp^3d^5s^*$ tight-binding formalism: From boundary conditions to strain calculations," Phys. Rev. B **74**, 205323 (2006).
[25]See supplementary material at http://dx.doi.org/10.1063/1.4732089 for ALGORITHM I (Iterative Method) and the calculation results for [111] orientation.
[26]S. Lee, F. Oyafuso, P. von Allmen, and G. Klimeck, "Boundary conditions for the electronic structure of finite-extent embedded semiconductor nanostructures," Phys. Rev. B **69**, 045316 (2004).