A Reduced-Order Method for Coherent Transport Using Green's Functions

Ulrich Hetmaniuk, Dong Ji, Yunqi Zhao, and Manjeri P. Anantram

***Abstract*— A reduced-order method is presented to efficiently calculate Green's functions connecting contacts or leads to all the points in a nanostructure in the coherent transport limit. The proposed approach samples a small subset of spatial grid points on the lead and a small subset of energy grid points to build a reduced-order model. The efficacy of the algorithm is demonstrated by applying it to calculate both the electron density and transmission in a resonant tunneling structure, a MOSFET, and a bilayer graphene device. The match in features of both the electron density and transmission versus energy with conventional methods to model devices is excellent while a large reduction in computational time is demonstrated.**

***Index Terms*— Device modeling, graphene, quantum transport, reduced-order method (ROM).**

I. Introduction

QUANTUM mechanical effects play an important role in determining the characteristics of many nanoelectronic devices. Examples include devices, such as tunnel transistors, superlattice-based quantum cascade devices, and resonant tunneling diodes, where the wave nature of electrons are essential to device operation. In modeling these devices, the calculation of charge and current densities is computationally the most expensive step. Algorithms to speed up the computation are important and fall into two categories. The first category involves exact methods [1]–[4]. The second category involves approximate methods, where the dimensionality of the problem is greatly reduced from that of the exact methods. The contact block-reduction method [5] is one such method belonging to the second category, which expresses the current density (or transmission) between leads in terms of eigenstates of the corresponding closed system. A relatively small number of well-chosen eigenstates suffices to obtain an accurate approximation of the transmission function. The contact block-reduction method incorporates a model-order reduction as it solves a small linear system of equations. Eigenstates are also used in [6] to build a reduced-order model by projecting the $k \cdot p$ Hamiltonian into a smaller subspace consisting of sub-bands in the cross section of a nanowire. The low-rank approximation method [7] transforms the transport problem from the original basis representation to a reduced basis composed of eigenfunctions for the free particle Hamiltonian with the Neumann boundary conditions.

For many applications, computing eigenstates remains a challenging numerical task. An alternative approach to build a reduced-order model relies on matching moments and does not require any eigenstate but only a few solutions of linear systems of equations. André and Aïssi [8] used this technique to calculate the electric field in waveguides. Huang *et al.* [9] employ moment matching for quantum transport. They combine, in the context of the wave function approach, a Padé approximation and an efficient sampling of energy points. Moment matching has also been shown to be efficient without an explicit Padé approximation in circuit simulation [10]–[12].

In this paper, we propose a reduced-order method (ROM) based on moment matching to calculate the electron and current densities in nanodevices. Our method is based on finding an approximate solution to Green's functions at all energy points. The approximation is obtained from a combination of a small number of precisely computed Green's functions that are selected automatically by sampling a subset of energy and spatial points in every lead. In Section II, we discuss the equations for coherent quantum transport; and in Section III, we present our ROM. Section IV demonstrates the efficiency of our proposed approach on examples of current technological relevance.

II. Review of Equations for Coherent Transport

Green's function formalism solves for the retarded Green's function $G^r(E)$ at energy E

$$[EI - H - \Sigma_L(E) - \Sigma_R(E)]G^r(E) = I \tag{1}$$

where H is the system Hamiltonian and $\Sigma_\alpha(E)$ is the self-energy at energy E due to the coupling between the device and the lead ($\alpha = L, R$ stand for the left and right leads in Fig. 1). The charge density at grid point q, n_q is expressed in terms of Green's function elements G^r_{qL} and G^r_{qR}, which connect grid point q in the device to all grid points in the device that are connected to contacts L and R using the

Manuscript received October 8, 2014; revised December 10, 2014; accepted January 6, 2015. Date of publication February 5, 2015; date of current version February 20, 2015. This work was supported by the National Science Foundation through the Division of Electrical, Communications and Cyber Systems under Grant ECCS-1231927. The work of U. Hetmaniuk was supported by the Office of Naval Research under Grant N00014-11-1-0710. The review of this paper was arranged by Editor A. Schenk.

U. Hetmaniuk is with the Department of Applied Mathematics, University of Washington, Seattle, WA 98195 USA (e-mail: hetmaniu@uw.edu).

D. Ji was with the Visiting International Student Internship and Training Program, University of Washington, Seattle, WA 98195 USA. He is now with the Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287 USA (e-mail: dongji829@gmail.com).

Y. Zhao and M. P. Anantram are with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: anant@uw.edu).

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.2015.2395420

0018-9383 © 2015 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.

Authorized licensed use limited to: University of Washington Libraries. Downloaded on August 10,2020 at 21:15:01 UTC from IEEE Xplore. Restrictions apply.

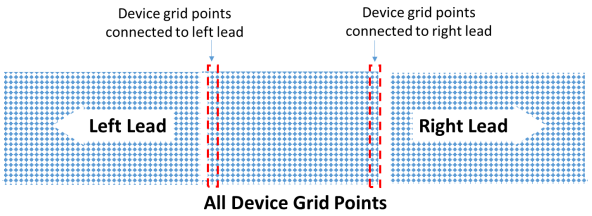


Fig. 1. Central region represents the grid points in the device. The two regions to the left and right represent the grid points in the leads. Only a small subset of Green's functions between the grid points connected to the leads (red boxes) and all grid points in the device are required to calculate the electron density. An even smaller subset of Green's functions only between the device grid points connected to the leads (the two red boxes) are required to calculate the transmission.

relationship [13]

$$n_q = 2\int \frac{dE}{2\pi} G^r_{qL}\,\Sigma^<_{LL}\left(G^r_{qL}\right)^\dagger + G^r_{qR}\,\Sigma^<_{RR}\left(G^r_{qR}\right)^\dagger \tag{2}$$

where $\Sigma^<_{LL}$ and $\Sigma^<_{RR}$ are the in-scattering self-energies and have nonzero elements between the device grid points connected to a lead, and have been defined in the following. The current from leads L to R is then expressed as

$$J = \frac{2e}{h}\int dE\; T_{LR}(E)\;[f_L(E) - f_R(E)] \tag{3}$$

where e is the electron charge and h is Planck's constant. The equilibrium distribution function within lead $\alpha = L, R$, which is the Fermi function, is denoted by $f_\alpha(E)$. The transmission function from leads L to R is given by

$$T_{LR}(E) = \mathrm{tr}\left(\Gamma_{LL}(E)G^r_{LR}(E)\Gamma_{RR}(E)\left(G^r_{LR}(E)\right)^\dagger\right) \tag{4}$$

where tr is the trace operator. $\Gamma_{\alpha\alpha}$ is given by

$$\Gamma_{\alpha\alpha}(E) = \imath(\Sigma_{\alpha\alpha}(E) - (\Sigma_{\alpha\alpha}(E))^\dagger) \tag{5}$$

and $\alpha = L, R$ and $\imath = \sqrt{-1}$. The self-energies are $\Sigma^<_{LL} = \imath\Gamma_{LL} f_L$ and $\Sigma^<_{RR} = \imath\Gamma_{RR} f_R$.

According to (2), the electron density can be evaluated from knowledge of only a small subset of all Green's function elements. The required subset of Green's function elements are those that connect all device grid points to those device grid points that are connected to contacts L and R (Fig. 1). The evaluation of the transmission requires an even smaller subset of Green's functions, which connects only those grid points that are connected to leads. The evaluation of Green's functions between the grid points connected to contact L and all other device grid points are obtained by solving a system of equations of the form

$$[EI - H - \Sigma_L(E) - \Sigma_R(E)]G^r_L(E) = I_L \tag{6}$$

where the right-hand side is a column vector with n rows and m_L columns. n is the total number of device grid points and m_L corresponds to the number of device grid points connected to the left lead. The nonzero elements of I_L have an entry of unity and their location is defined by rows and columns, which correspond to the same device grid point connected to lead L. If the grid points in Fig. 1 are labeled as vertical layers, the nonzero elements of I_L will form an identity matrix of size equal to $m_L \times m_L$. Similarly, the evaluation of Green's functions from the device grid points connected to contact R and all other device grid points are obtained by solving a matrix of the form

$$[EI - H - \Sigma_L(E) - \Sigma_R(E)]G^r_R(E) = I_R \tag{7}$$

where the right-hand side is a column vector with n rows and m_R columns. m_R corresponds to the number of device grid points connected to the right lead. The nonzero elements of I_R are defined in a manner identical to I_L except that they correspond to the device grid points connected to the right lead. Currently, (6) and (7) are solved without further approximations and this is referred to as the full-order method (FOM). The solution of the FOM uses efficient exact linear solvers that exploit the sparsity of the matrices.

Remark: For many ballistic simulations, the quantum transmitting boundary method (QTBM) [14] computes the transmission function, the currents between leads, and the electron density more efficiently than the outlined approach based on (6) and (7). Further investigation would be needed to combine QTBM with the reduced-order approach as described in the following.

III. Reduced-Order Method

We consider a device region that is connected to a number of leads represented by α (Fig. 1 corresponds to the case with two leads). The equation for the retarded Green's function between device grid points connected to lead α and all other device grid points are obtained by solving

$$\left[E\mathbf{I} - \mathbf{H} - \sum_\alpha \Sigma_\alpha(E)\right]\mathbf{G}^r_\alpha(E) = \mathbf{I}_\alpha. \tag{8}$$

The matrix $\mathbf{I}_\alpha$ is rectangular with only a few nonzero entries depending on lead α, as shown in (6) and (7). The dimension of matrix $\mathbf{G}^r_\alpha$ is $n \times m_\alpha$, where m_α is the number of grid points in the device connected to lead α. We seek to find a unitary matrix $\mathbf{V}_\alpha$ with dimension $n \times s$ such that

$$\mathbf{G}^r_\alpha(E) \approx \mathbf{V}_\alpha \tilde{\mathbf{g}}_\alpha(E) \tag{9}$$

where $\tilde{\mathbf{g}}_\alpha(E)$ is Green's function of reduced dimension $s \times s$ and $\mathbf{V}_\alpha\tilde{\mathbf{g}}_\alpha(E)$ is a good approximation of Green's function $\mathbf{G}^r_\alpha$ in (8). Substituting (9) in (8) and premultiplying by $\mathbf{V}^\dagger_\alpha$, we obtain the equation for $\tilde{\mathbf{g}}_\alpha$

$$[E\mathbf{I} - \tilde{\mathbf{H}} - \tilde{\Sigma}_L(E) - \tilde{\Sigma}_R(E)]\tilde{\mathbf{g}}_\alpha(E) = \mathbf{V}^\dagger_\alpha\mathbf{I}_\alpha \tag{10}$$

where $\tilde{\mathbf{H}} = \mathbf{V}^\dagger_\alpha\mathbf{H}\mathbf{V}_\alpha$ and $\tilde{\Sigma}_\alpha = \mathbf{V}^\dagger_\alpha\Sigma_\alpha\mathbf{V}_\alpha$ are the reduced Hamiltonian and self-energy matrices, which are of dimension $s \times s$. We now need to find a unitary matrix $\mathbf{V}_\alpha$ that does not depend on energy E. To find $\mathbf{V}_\alpha$, we calculate the precise Green's functions in (8) for a select subset of: 1) spatial grid points $\tilde{q}_i$ in lead α and 2) energy grid points $\tilde{E}_i$. The set of these precisely calculated Green's function is represented by

$$\mathbf{W}_\alpha = \left[\mathbf{G}^r_{\tilde{q}_1}(\tilde{E}_1), \ldots, \mathbf{G}^r_{\tilde{q}_s}(\tilde{E}_s)\right]. \tag{11}$$

$\mathbf{V}_\alpha$ is obtained from (11) by Gram–Schmidt orthogonalization (or a QR factorization).

Authorized licensed use limited to: University of Washington Libraries. Downloaded on August 10,2020 at 21:15:01 UTC from IEEE Xplore. Restrictions apply.

The overall quality of the approximation depends on the selection of samples $\tilde{q}_i$ and $\tilde{E}_i$, and the size of the reduced-order model s. The selection of samples $\tilde{q}_i$ and $\tilde{E}_i$ to construct $\mathbf{W}_\alpha$ (and subsequently $\mathbf{V}_\alpha$) and to determine the dimension s in (11) is discussed now. Starting from an empty matrix $\mathbf{W}_\alpha$, we insert new columns in the matrix $\mathbf{W}_\alpha$ until the residual norm

$$\|[E\mathbf{I} - \mathbf{H} - \Sigma_L(E) - \Sigma_R(E)]\mathbf{V}_\alpha \tilde{\mathbf{g}}_\alpha(E) - \mathbf{I}_\alpha\| \qquad (12)$$

is smaller than a chosen tolerance τ for any energy in the range $E_{\min} < E < E_{\max}$. Note that evaluating the residuals in (12) is much cheaper than solving (8) for one energy value E. The expression in (12) requires solving a system of s equations to compute $\tilde{\mathbf{g}}_\alpha(E)$ and doing a set of vector multiplications with sparse matrices.

The initial set of insertions into $\mathbf{W}_\alpha$ are column vectors at energy $E_{\min}$ and $\mathbf{G}^r_{\tilde{q}}(E_{\min})$, where the grid points $\tilde{q}$ lie in lead α and are chosen to enforce the largest decrease in the residual norm in (12). These insertions are continued until the residual norm becomes smaller than the tolerance τ. The second set of insertions into $\mathbf{W}_\alpha$ are at $E = E_{\max}$, and column vectors $\mathbf{G}^r_{\tilde{q}}(E_{\max})$ are added to the matrix $\mathbf{W}_\alpha$ until the residual norm becomes smaller than τ. At this point, the approximation is accurate for energies $E_{\min}$ and $E_{\max}$ but the residual norm may not be smaller than the tolerance τ at all energies in the interval $E_{\min} < E < E_{\max}$. In that case, one repeats the above procedure for the next energy point E_* where the residual norm is the largest. Vectors $\mathbf{G}^r_{\tilde{q}}(E_*)$ are added to the matrix $\mathbf{W}_\alpha$ following the previous procedure. The algorithm stops when it cannot find a value E_* where the residual norm is larger than the tolerance τ. In the examples presented below, the points $E_{\min} < E_* < E_{\max}$ are chosen to be Chebyshev points which traditionally have good interpolation properties [15]. Empirical knowledge and/or eigenvalue information may be used to adapt the energy sampling and to improve the quality of the resulting reduced-order model.

As a summary, we now have the unitary matrix $\mathbf{V}_\alpha$, which does not depend on energy E, and Green's function $\tilde{\mathbf{g}}_\alpha$ of reduced dimension [defined by (10)]. Using these two quantities in (9) gives us the approximate Green's functions connecting all grid points in the device to the grid points connected to lead α. This approximation is next inserted into (2) to efficiently calculate the electron density and into (3) to efficiently calculate the transmission.

IV. Numerical Modeling of Devices

In this section, we demonstrate the proposed method using the examples of realistic devices. The first device that we consider is a double-barrier resonant tunneling diode, which consists of sharp resonances through which transport occurs. The second example we consider is a MOSFET, where transport occurs mainly above the barrier through step-like transmission features at quantized energy levels in the 2-D channel. The third example considered is a bilayer graphene device where transport occurs by hopping of electrons between graphene layers, a device that has shown to exhibit negative differential conductance. While the Hamiltonian in the double-barrier and MOSFET device are based on the effective mass equation, the Hamiltonian in the bilayer graphene example is based on tight binding.

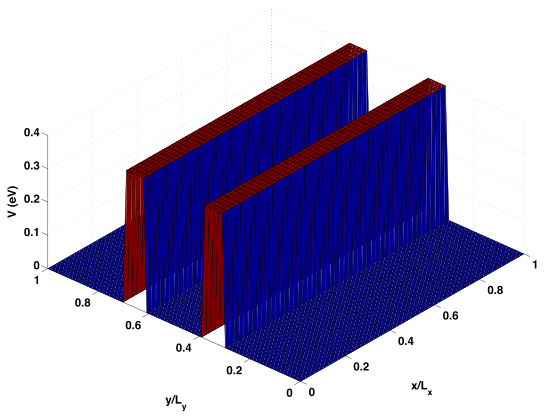


Fig. 2. Potential for the rectangular nanodevice.

We remark that all times reported are the total times, which include all aspects of the computation. In the case of FOM, these timings contain the evaluations of $\mathbf{G}^r_L$, of $\mathbf{G}^r_R$, and of the transmission functions at n_E energy values as well as the charge density n_q. The total time for the ROM includes the construction of $\mathbf{V}_L$ and $\mathbf{V}_R$, the evaluations of $\mathbf{V}_L\tilde{\mathbf{g}}_L(E)$, of $\mathbf{V}_R\tilde{\mathbf{g}}_R(E)$, and of the resulting approximations for the transmission function at n_E energy values as well as the charge density n_q. All the simulations are performed with MATLAB on a MacBook Air 1.7-GHz Intel Core i5 and 4-GB 1333-MHz DDR3.

A. Resonant Tunneling Device

We consider a 2-D double-barrier resonant tunneling device with $L_x = 10$ nm and $L_y = 15$ nm. There are hard walls at $x = 0$ and $x = L_x$, and semi-infinite boundaries with a potential energy of zero for $y < 0$ and $y > L_y$; these are reflection-less semi-infinite contacts. The effective mass Hamiltonian is discretized using the finite difference scheme with a single isotropic effective mass of $m^* = 0.25\ \mathrm{m}_o$. Fig. 2 shows the potential present over the rectangular device.

We first consider a uniform spatial grid that is composed of $N_y = 75$ layers in the y-direction, where each layer has $N_x = 50$ grid points. The conduction band offset of 400 meV at the barrier–well interface occurs over two adjacent grid points. The transmission is calculated using both the FOM and ROM proposed in the previous section with a tolerance of $\tau = 10^{-2}$. The same linear solver is used to solve the FOM and develop the ROM.

Fig. 3 compares the transmission obtained in the range of energies $0 < E < 0.4$ eV. We find that the resonances are accurately reproduced in magnitude, energy location, and width. The computational time using the FOM is 87 s, while our reduced-order model takes only 25 s. For each contact, the size of the reduced-order model is $s = 116 \ll 3750$. The number of energy values n_E is set at 401.

Fig. 4 shows the electron density obtained in the range of energies $0 < E < 0.4$ eV. The maximum relative error between the FOM computation and its ROM approximation is less than 0.03%.

Authorized licensed use limited to: University of Washington Libraries. Downloaded on August 10,2020 at 21:15:01 UTC from IEEE Xplore. Restrictions apply.

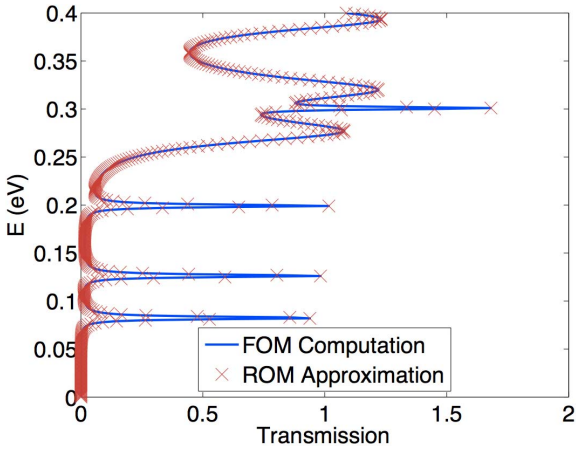


Fig. 3. Plot of the energy versus the transmission.

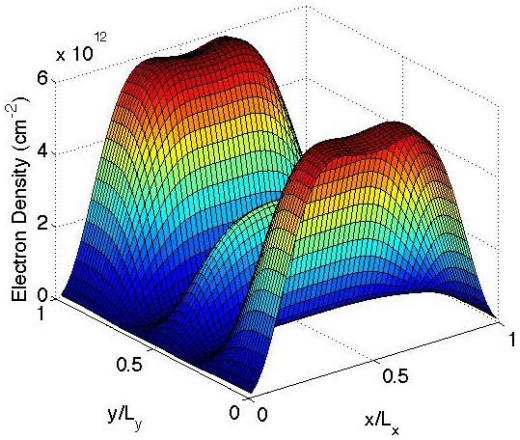


Fig. 4. Electron density for the nanodevice. The density shows a peak in the well due to the occupied resonance. The illustrated calculation uses the results from the ROM and matches the results from the FOM.

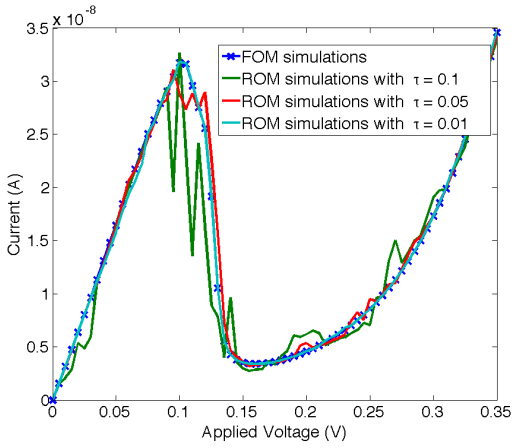


Fig. 5. Comparison of I–V characteristics obtained from FOM and ROM simulations with different tolerances.

Fig. 5 shows the I–V characteristics for this resonant tunneling device when a potential difference is applied between the two contacts and the potential varies linearly across the device. For a tolerance $\tau = 10^{-1}$, the I–V characteristics from the ROM simulations deviate from the reference computations. Decreasing the value of τ improves the quality of the I–V characteristics. When the tolerance is $\tau = 10^{-2}$, the I–V characteristics from the ROM simulations match the I–V characteristics computed with the FOM.

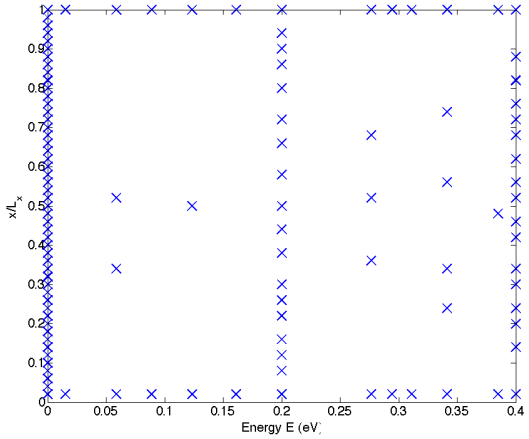


Fig. 6. Samples $(\tilde{q}_i, \tilde{E}_i)$ whose collection is used to build the ROM.

TABLE I
SIZES OF ROM AND TIMINGS AS A FUNCTION OF FOM

N_x	N_y	Size FOM	Size ROM	Time including setup	
				FOM	ROM
26	39	n=1,014	s=89	10 s	7 s (3 s)
50	75	n=3,750	s=116	87 s	25 s (17 s)
100	150	n=15,000	s=157	905 s	164 s (127 s)

Fig. 6 shows the x-coordinates of the grid points, $\tilde{q}_i$, located on the device layer adjoining the left lead (at $y = 0$) and the corresponding energy values $\tilde{E}_i$. Each sample $(\tilde{q}_i, \tilde{E}_i)$ yields a precisely calculated Green's function $\mathbf{G}^r_{\tilde{q}_i}(\tilde{E}_i)$, collection of which builds $\mathbf{W}_L$ and the set of orthonormal vectors in $\mathbf{V}_L$. Overall, the samples are sparsely distributed over the energy interval of interest and among the grid points on lead L. A similar distribution of samples occurs for the right lead.

We make a few remarks of how the calculations scale with dimension of both spatial and energy grid points. First, to evaluate Green's function $\mathbf{G}^r_L(E)$ at n_E energy values, the FOM computation solves $N_x n_E = 50 n_E$ linear systems (8), each with $n = 3750$ equations. On the other hand, the ROM approximation solves $s = 116$ linear systems (8) with $n = 3750$ equations to construct the matrix $\mathbf{V}_L$ followed by the solution of $N_x n_E = 50 n_E$ linear systems (10), each with $s = 116$ equations. The ROM approximation becomes more advantageous as the number of energy values n_E increases. The second remark is that the reduced-order model is also more effective as the number of spatial grid points becomes large. To demonstrate this, we keep the device dimensions L_x and L_y fixed but decrease the grid spacing by doubling both N_x and N_y. The conduction band offset at the barrier–well interface still occurs over two adjacent grid points. It is observed in Table I that the speedup increases as the number of grid points increases. Table I shows only a very modest increase in the size of the reduced-order models. Relative to FOM, the ROM performs increasingly better as the system size increases. The total time including the ROM setup is given in

Authorized licensed use limited to: University of Washington Libraries. Downloaded on August 10,2020 at 21:15:01 UTC from IEEE Xplore. Restrictions apply.

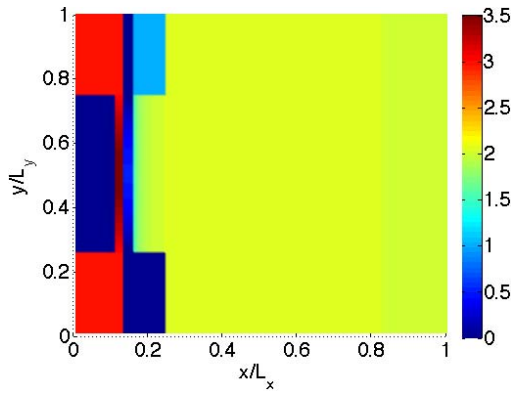


Fig. 7. Contour plot of potential profile. The source and drain regions are connected by the channel, all of which are in various shades of blue. The oxide is shown in red, embedded in which is the gate contact (shown in blue).

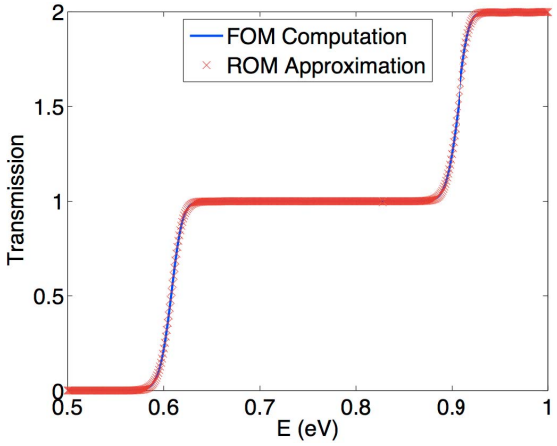


Fig. 8. Transmission between source and drain.

the last column. The portion of the total time required for ROM setup is given in the bracket of the last column. The timings include the calculation of both the electron density and transmission. The number of energy values n_E is set at 401.

B. MOSFET

We consider a silicon MOSFET with $L_x = 92$ nm and $L_y = 60$ nm (Fig. 7). The thickness and width of the oxide are 2 and 25 nm, respectively. The source, drain, and gate contacts are semiinfinite and periodic. Fig. 7 also shows a contour plot of the potential.

The Hamiltonian is discretized using the finite difference scheme with the number of grid points in the x- and y-directions equal to $N_x = 185$ and $N_y = 121$. The transmission between the source and drain contacts resulting from both the FOM and ROM methods are shown in Fig. 8. The ROM is constructed with the tolerance $\tau = 10^{-2}$. The ROM yields the correct transmission with very high fidelity. Both the location of the steps in transmission and the step heights are accurately reproduced.

The absolute error between the electron densities obtained from the FOM computation and the ROM approximation are presented in Fig. 9 (the electron density is four orders of magnitude larger).

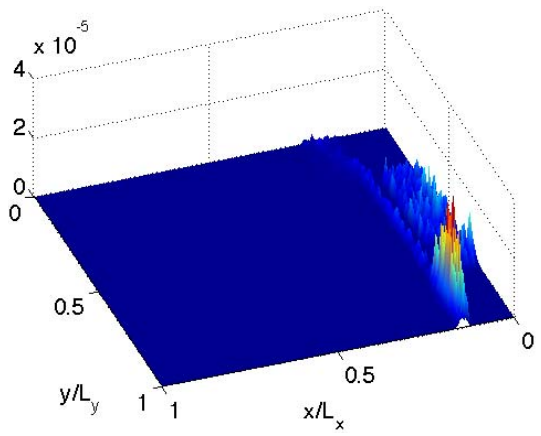


Fig. 9. Absolute error on electron density between the FOM computation and the ROM approximation. The absolute error is at least four orders of magnitude smaller than the electron density at each grid point. Note that we also calculate the electron density in the gate contact.

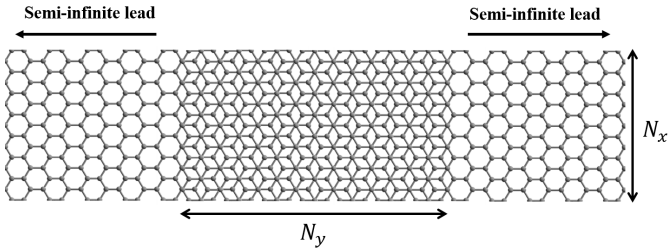


Fig. 10. Model bilayer graphene device. The separation between the two layers is 3.1 Å.

We construct reduced-order models for the three contacts: gate, source, and drain. For the tolerance $\tau = 10^{-2}$, the three separate unitary matrices, $\mathbf{V}_{\text{gate}}$, $\mathbf{V}_{\text{source}}$, and $\mathbf{V}_{\text{drain}}$, each have 22 385 rows and, 417, 307, and 197 columns, respectively. For this simulation, where $n_E = 500$ energy values were used, the total time for the FOM is 1274 s. The total time for the ROM including setup and calculation of the electron density and transmission is 365 s, yielding a speedup of 3.5. The construction of the ROM took 245 s and the evaluation of the electron density and transmission took 120 s. Note that if the number of energy points is larger within the window of energies in Fig. 8, the time taken for construction of the ROM does not change.

C. Graphene Nanoribbon

In the third example, we demonstrate the reduced-order model in bilayer graphene nanoribbon. Graphene devices with insulators between them and bilayer graphene devices are being researched for applications such as resonant tunneling devices [16]–[19]. We first demonstrate the applicability of our ROM approach here for a device consisting of two graphene sheets. The graphene sheets have a width of $N_x = 32$ perpendicular to transport and have an overlap of $N_y = 32$ layers along the transport direction (Fig. 10).

The Hamiltonians are constructed using the nearest neighbor tight binding parameters proposed in [20], where each carbon atom has nonzero entries in the Hamiltonian for hopping to

Authorized licensed use limited to: University of Washington Libraries. Downloaded on August 10,2020 at 21:15:01 UTC from IEEE Xplore. Restrictions apply.

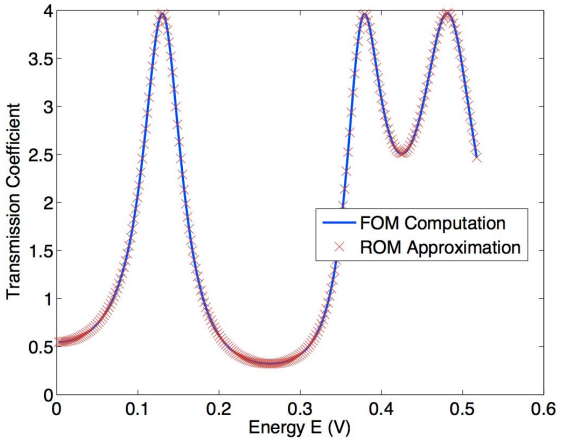


Fig. 11. Transmission versus energy for a bilayer graphene device.

TABLE II
TIMINGS FOR THE FOM AND THE ROM AS A FUNCTION OF THE NUMBER OF ENERGY VALUES, n_E. THE FOM SIZE IS 2176 WHILE THE ROM SIZE IS 39

n_E	Time FOM	Time ROM (including setup)
251	11 s	2.8 s
501	22 s	3.6 s
1001	45 s	5.0 s

the nearest neighbors both within a graphene sheet and to the neighboring graphene sheet. That is, there are three atoms within a graphene sheet to/from where electrons can hop to each carbon atom, and three additional atoms to which they can hop in the neighboring sheet. The procedure to construct the reduced-order model remains the same as in the previous two examples. The transmission between the left and right contacts resulting from both the FOM and ROM are shown in Fig. 11.

The ROM yields the correct transmission with very high fidelity. The ROM is constructed with the tolerance $\tau = 10^{-2}$ and the two separate unitary matrices, $\mathbf{V}_L$ and $\mathbf{V}_R$, have 2176 rows and 39 columns. For this simulation, where $n_E = 251$ energy values were used, the total time for the FOM is 11 s, and the total time for the ROM method is 2.8 s, yielding a speedup of four. Table II lists the total time for the FOM and ROM. The data indicate that the timings verify $t_{\text{FOM}} \approx 0.046 n_E$ and $t_{\text{ROM}} \approx 2.0 + 0.003 n_E$. The ROM approximation becomes more advantageous as the number of energy values n_E increases.

Finally, we comment on how the calculations scale with the width, N_x, perpendicular to transport and the number of overlapping layers, N_y, along the transport direction. The number of energy values n_E is set at 501.

The timings include the calculation of both the electron density and transmission. The total time including the ROM setup is given in the last column. The portion of the total time required for ROM setup is given in the bracket of the last column.

It is observed in Table III that the ROM remains faster than the FOM. However, the speedup decreases with increase in

TABLE III
SIZES OF ROM AND TIMINGS AS A FUNCTION OF N_x AND N_y (HERE n_E IS SET AT 501)

N_x	N_y	Size FOM	Size ROM	Time including setup	
				FOM	ROM
32	32	n=2,176	s = 48	22 s	3.6 s (2 s)
62	32	n=4,216	s = 74	95 s	13 s (9 s)
128	32	n=8,704	s = 205	509 s	197 s (162 s)
62	32	n=4,216	s = 74	95 s	13 s (9 s)
62	64	n=8,184	s = 99	195 s	54 s (44 s)
62	128	n=16,120	s = 112	409 s	146 s (129 s)

size of the device, in contrast to the double-barrier resonant tunneling structure. The underlying reason for this is that when N_x and N_y increase in the bilayer graphene device, the physical dimensions of the device change. As a result of this, Green's functions, and the corresponding transmission and electron density change significantly. In contrast, for the case of the double-barrier resonant tunneling structure, an increase in the number of grid points (matrix size) did not lead to an increase in the features in the underlying Green's functions because the underlying devices dimensions did not change.

V. CONCLUSION

We have proposed an ROM to speed up the modeling of nanoscale devices in the phase coherent limit. This approach is capable of calculating both the current and electron density. The proposed approach is based on constructing appropriate unitary matrices from the full-order model over a small subset of energies and contact grid points. This model has been implemented to calculate the phase coherent properties of nanodevices based on both the discretization of the continuum Green's function equations and the tight binding-based Green's function equations. Speedup of between three and seven times has been demonstrated for examples involving double-barrier resonant tunneling diodes, MOSFETs, and bilayer graphene device. We hope that this paper will spur interest in the further development of ROMs to model nanodevices both in the phase coherent limit and with decoherence.

REFERENCES

[1] R. Lake, G. Klimeck, R. C. Bowen, and D. Jovanovic, "Single and multiband modeling of quantum electron transport through layered semiconductor devices," *J. Appl. Phys.*, vol. 81, no. 12, pp. 7845–7869, 1997.

[2] A. Svizhenko, M. P. Anantram, T. R. Govindan, B. Biegel, and R. Venugopal, "Two-dimensional quantum mechanical modeling of nanotransistors," *J. Appl. Phys.*, vol. 91, no. 4, pp. 2343–2354, 2002.

[3] S. Li and E. Darve, "Extension and optimization of the FIND algorithm: Computing Green's and less-than Green's functions," *J. Comput. Phys.*, vol. 231, no. 4, pp. 1121–1139, 2011.

[4] U. Hetmaniuk, Y. Zhao, and M. P. Anantram, "A nested dissection approach to modeling transport in nanodevices: Algorithms and applications," *Int. J. Numer. Methods Eng.*, vol. 95, no. 7, pp. 587–607, 2013.

[5] H. R. Khan, D. Mamaluy, and D. Vasileska, "Quantum transport simulation of experimentally fabricated nano-FinFET," *IEEE Trans. Electron Devices*, vol. 54, no. 4, pp. 784–796, Apr. 2007.

[6] 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.

Authorized licensed use limited to: University of Washington Libraries. Downloaded on August 10,2020 at 21:15:01 UTC from IEEE Xplore. Restrictions apply.

[7] L. Zeng, Y. He, M. Povolotskyi, X. Liu, G. Klimeck, and T. Kubis, "Low rank approximation method for efficient Green's function calculation of dissipative quantum transport," *J. Appl. Phys.*, vol. 113, no. 21, p. 213707, 2013.
[8] F. André and A. Aïssi, "The equivalent matrices of a periodic structure," *IEEE Trans. Electron Devices*, vol. 57, no. 7, pp. 1687–1695, Jul. 2010.
[9] J. Z. Huang, W. C. Chew, M. Tang, and L. Jiang, "Efficient simulation and analysis of quantum ballistic transport in nanodevices with AWE," *IEEE Trans. Electron Devices*, vol. 59, no. 2, pp. 468–476, Feb. 2012.
[10] P. Feldmann and R. W. Freund, "Efficient linear circuit analysis by Pade approximation via the Lanczos process," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 14, no. 5, pp. 639–649, May 1995.
[11] D. S. Weile, E. Michielssen, and K. Gallivan, "Reduced-order modeling of multiscreen frequency-selective surfaces using Krylov-based rational interpolation," *IEEE Trans. Antennas Propag.*, vol. 49, no. 5, pp. 801–813, May 2001.
[12] A. C. Antoulas, *Approximation of Large-Scale Dynamical Systems*. Philadelphia, PA, USA: SIAM, 2005.
[13] M. P. Anantram, M. S. Lundstrom, and D. E. Nikonov, "Modeling of nanoscale devices," *Proc. IEEE*, vol. 96, no. 9, pp. 1511–1550, Sep. 2008.
[14] C. S. Lent and D. J. Kirkner, "The quantum transmitting boundary method," *J. Appl. Phys.*, vol. 67, no. 10, pp. 6353–6359, 1990.
[15] Z. Battles and L. N. Trefethen, "An extension of MATLAB to continuous functions and operators," *SIAM J. Sci. Comput.*, vol. 25, no. 5, pp. 1743–1770, 2004.
[16] G. Fiori, "Negative differential resistance in mono and bilayer graphene p-n junctions," *IEEE Electron Device Lett.*, vol. 32, no. 10, pp. 1334–1336, Oct. 2011.
[17] G. Fiori, D. Neumaier, B. N. Szafranek, and G. Iannaccone, "Bilayer graphene transistors for analog electronics," *IEEE Trans. Electron Devices*, vol. 61, no. 3, pp. 729–733, Mar. 2014.
[18] K. M. Masum Habib, F. Zahid, and R. K. Lake, "Negative differential resistance in bilayer graphene nanoribbons," *Appl. Phys. Lett.*, vol. 98, no. 19, p. 192112, 2011.
[19] L. Britnell *et al.*, "Resonant tunnelling and negative differential conductance in graphene transistors," *Nature Commun.*, vol. 4, Apr. 2013, Art. ID 1794.
[20] J. Sławińska, I. Zasada, and Z. Klusek, "Energy gap tuning in graphene on hexagonal boron nitride bilayer system," *Phys. Rev. B*, vol. 81, p. 155433, Apr. 2010.

Authors' photograph and biography not available at the time of publication.

Authorized licensed use limited to: University of Washington Libraries. Downloaded on August 10,2020 at 21:15:01 UTC from IEEE Xplore. Restrictions apply.