INTERNATIONAL JOURNAL OF NUMERICAL MODELLING: ELECTRONIC NETWORKS, DEVICES AND FIELDS
Int. J. Numer. Model. 2011; **24**:322–334
Published online 2 July 2010 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/jnm.780

Simulation of quantum transport in double-gate MOSFETs using the non-equilibrium Green's function formalism in real-space: A comparison of four methods

Yasser M. Sabry[1,*,†], Tarek M. Abdolkader[2] and Wael F. Farouk[3]

[1]*Department of Electronics and Communication Engineering, Faculty of Engineering, Ain Shams Univeristy, Egypt*
[2]*Department of Basic Sciences, Benha Higher Institute of Technology, Egypt*
[3]*Department of Engineering Physics and Mathematics, Faculty of Engineering, Ain Shams University, Egypt*

SUMMARY

Quantum effects play an important role in determining the double-gate (DG) MOSFETs characteristics. The non-equilibrium Green's function formalism (NEGF) in real-space (RS) representation provides a rigorous description of quantum transport in nanoscale devices. Unfortunately, the traditional NEGF framework has the disadvantage of being heavy in computations. Methods that reduce the computations exist in the literature like the recursive Green's Function (RGF) algorithm, the contact block reduction (CBR) method, and Gauss elimination (GE) method. Comparison of the simulation time of the traditional NEGF, the RGF algorithm, the CBR method, and the GE method was always theoretical and based on approximate estimates. In this work, we carry out a real comparison between the four methods by implementing them inside the same simulator, using them to simulate the same device dimensions and parameters on the same machine. It is demonstrated that the RGF algorithm or the GE method introduce about one order of magnitude reduction in simulation time below that traditional NEGF, whereas the CBR method yields the smallest simulation time with about two orders of magnitude reduction. Copyright © 2010 John Wiley & Sons, Ltd.

Received 9 February 2010; Revised 20 April 2010; Accepted 8 May 2010

KEY WORDS: quantum transport; non-equilibrium Green's function; double-gate MOSFETs; recursive Green's function; contact block reduction

1. INTRODUCTION

Rapid device scaling pushes the dimensions of the field-effect transistors to the nanometer regime [1]. The International Technology Roadmap for Semiconductors projection for the double-gate (DG) MOSFETs physical gate length is 4.5 nm for the year 2022. For these extremely scaled dimensions, quantum effects play an important role in determining the DG MOSFETs characteristics. These effects can be accurately predicted only using quantum mechanical-based device simulation [2].

The non-equilibrium Green's function formalism (NEGF) provides a rigorous description of quantum transport in nanoscale devices [3]. Device simulation based on NEGF is carried out using the so-called self-consistent solution method shown in Figure 1. The method is composed of two main blocks, Poisson's equation solver and the quantum transport solver, which is based on the NEGF formalism. Poisson's equation gives the electrostatic potential distribution (V) in

*Correspondence to: Yasser M. Sabry, Department of Electronics and Communication Engineering, Faculty of Engineering, Ain Shams Univeristy, Egypt.
†E-mail: ysabry@ieee.org

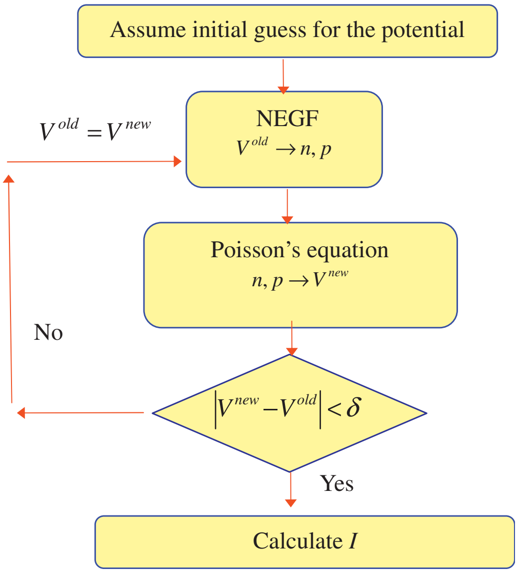
Assume initial guess for the potential
$V^{old} = V^{new}$
NEGF
$V^{old} \rightarrow n, p$
Poisson's equation
$n, p \rightarrow V^{new}$
No
$|V^{new} - V^{old}| < \delta$
Yes
Calculate I

Figure 1. Flow chart illustrating the self-consistent method used in device simulation using the NEGF.

the device for a given electron density (n) and hole density (p). The NEGF solver gives the n and p density and the electrical current (I) for a given potential V. The self-consistent method starts by assuming initial value for the potential, which is fed to the NEGF solver, to calculate the n and p densities. The calculated densities are fed to Poisson's solver to find the updated potential V_{new} in the device. We go forth and back between Poisson's solver and NEGF solver until the update in the potential drops below certain tolerance and then terminal currents are calculated.

Computational efficiency is needed to make the self-consistent method suitable for device design and characteristic prediction. Unfortunately, the NEGF method has the disadvantage of being heavy in computations [4]. Green's function is calculated by means of matrix inversion for the Hamiltonian matrix. This makes the NEGF formalism not suitable for 3D or even 2D devices. Therefore, several methods have been proposed to reduce the computational burden of the NEGF. Some of these methods sacrifice with the accuracy of the simulation by using the uncoupled-mode-space (UMS) representation [4, 5]. Other methods that reduce the computations in the real-space (RS) representation also exist like the recursive Green's Function (RGF) algorithm [6–8], the contact block reduction (CBR) method [9,10], and Gauss elimination (GE) method [11]. The RS representation has the advantage of being able to accurately: (1) predict the electrical characteristics of DG MOSFETs whether the Si body is ultra-thin or not [12], (2) simulate electronic devices with arbitrary-oriented wafer orientation [13], (3) account for non-coherent scattering, and (4) calculate the gate leakage current self-consistently [14].

Comparison of the simulation time of the traditional NEGF, the RGF algorithm, the CBR method, and the GE method was always theoretical and based on approximate estimates. In this work, we carry out a real comparison between the four methods by implementing them inside the same simulator, using them to simulate the same device dimensions and parameters on the same machine. The FETMOSS simulator has been chosen for this comparison [15]. It is a 2D simulator for nanoscale DG n-MOSFETs based on the UMS representation. In this work, the aforementioned RS methods will be implemented in FETMOSS and compared. The rest of this article is organized as follows. In Section 2, simulation of DG MOSFETs using the traditional NEGF is presented. Section 3 discuses the various computationally efficient methods used to reduce the traditional NEGF computational burden and their application to DG MOSFETs. These methods were implemented in the FETMOSS simulator and the results are given in Section 4. First, the new version of FETMOSS is benchmarked using a well-known simulator available online, then the simulation time of the four methods (the traditional NEGF, the RGF algorithm, the CBR method, and the GE method) is compared.

DOI: 10.1002/jnm

2. DG MOSFETS SIMULATION USING THE TRADITIONAL NEGF

The DG MOSFET model device geometry is shown in Figure 2. The following assumptions are usually made in the nanoscale DG MOSFETs simulation:

(1) The channel length in x-direction is shorter than any characteristic scattering length such that the device is operating in the ballistic limit.
(2) The width of the device in the z-direction is so large compared with other dimensions of the active device such that the potential along that direction is rendered constant.
(3) The metal contacts are so large such that the thermal equilibrium is maintained, and the Fermi level in these regions is determined by the applied voltage.
(4) N-channel transistor where holes contribution, to both the transport and the electrostatic problems, can be neglected.
(5) No electron penetration in the insulator region.
(6) A single-band effective mass Hamiltonian [16] is used to model the electron transport.

The 2D wave function $\psi(x,y)$ is obtained from the solution of the 2D Schrödinger equation:

$$\left[-\frac{\hbar^2}{2}\left(\frac{1}{m_x^*}\frac{\partial^2}{\partial x^2}+\frac{1}{m_y^*}\frac{\partial^2}{\partial y^2}\right)+E_c(x,y)\right]\psi(x,y)=E_l\psi(x,y) \tag{1}$$

where m_x^* and m_y^* are electron effective mass in x- and y-direction, respectively, E_C is the conduction band edge and E_l is the longitudinal energy due to motion in x- and y- direction.

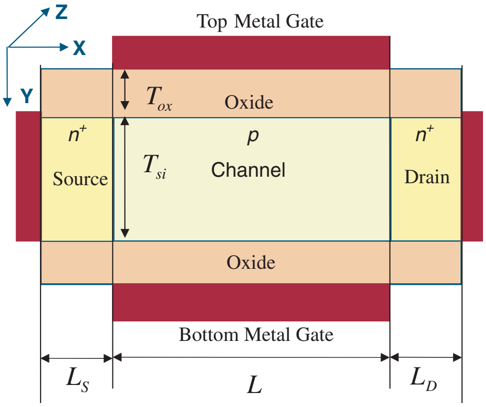

Figure 2. A model double-gate MOSFET used in this work.

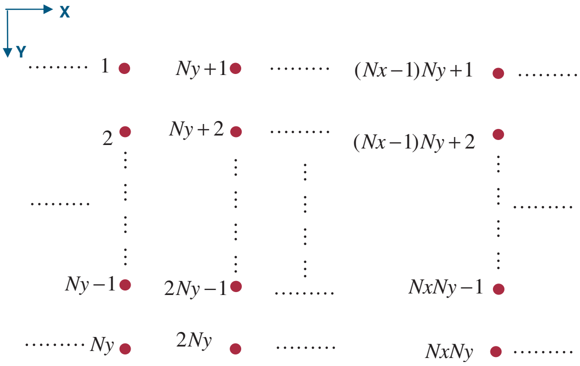

Figure 3. Two-dimensional simulation grid.

DOI: 10.1002/jnm

On discretization of Equation (1) using the grid shown in Figure 3, a set of linear equations is obtained and can be cast in the matrix form:

$$[\boldsymbol{H}_l]\{\boldsymbol{\psi}\}+[\boldsymbol{E}_C]\{\boldsymbol{\psi}\} = [E_l\boldsymbol{I}]\{\boldsymbol{\psi}\} \tag{2}$$

where

$$\boldsymbol{H}_l = \begin{bmatrix} \boldsymbol{\alpha} & \boldsymbol{\beta} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{\beta} & \boldsymbol{\alpha} & \boldsymbol{\beta} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \ddots & \cdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{\beta} & \boldsymbol{\alpha} \end{bmatrix}_{N_xN_y\times N_xN_y} \qquad \boldsymbol{\alpha} = \begin{bmatrix} 2t_x+2t_y & -t_y & 0 & \cdots & 0 \\ -t_y & 2t_x+2t_y & -t_y & \cdots & 0 \\ 0 & \cdots & \ddots & \cdots & \vdots \\ 0 & 0 & \cdots & -t_y & 2t_x+2t_y \end{bmatrix}_{N_y\times N_y}$$

$$\boldsymbol{\beta} = \begin{bmatrix} -t_x & 0 & \cdots & 0 \\ 0 & -t_x & \vdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & \cdots & \cdots & -t_x \end{bmatrix}_{N_y\times N_y} \qquad \boldsymbol{E}_C = \begin{bmatrix} Ec_1 & 0 & \cdots & \cdots & 0 \\ 0 & Ec_2 & 0 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & Ec_{N_xN_y} \end{bmatrix}_{N_xN_y\times N_xN_y} \qquad \boldsymbol{\psi} = \begin{Bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_{N_xN_y} \end{Bmatrix}_{N_xN_y\times 1}$$

$$t_x = \frac{\hbar^2}{2m_x^*(\Delta x)^2}, \quad t_y = \frac{\hbar^2}{2m_y^*(\Delta y)^2}$$

The retarded Green's function of the active device is given by:

$$\boldsymbol{G} = \left[E_l\boldsymbol{I} - \boldsymbol{H} - \sum_S - \sum_D\right]^{-1} \tag{3}$$

where $\boldsymbol{H} = \boldsymbol{H}_l+\boldsymbol{E}_C$; $\sum_{\boldsymbol{S}}$ and $\sum_{\boldsymbol{D}}$ are the source and drain contact self energy given by [12]:

$$\sum_{\boldsymbol{S}} = \begin{bmatrix} \boldsymbol{\beta g_S \beta} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \\ \boldsymbol{0} & \cdots & \cdots & \boldsymbol{0} \end{bmatrix}_{N_xN_y\times N_xN_y} \qquad \sum_{\boldsymbol{D}} = \begin{bmatrix} \boldsymbol{0} & \cdots & \cdots & \boldsymbol{0} \\ \vdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{\beta g_D \beta} \end{bmatrix}_{N_xN_y\times N_xN_y} \tag{4}$$

where $\boldsymbol{g}_S$ and $\boldsymbol{g}_D$ are the surface Green's functions of the source and drain contacts respectively [17].

The broadening functions, $\boldsymbol{\Gamma}_S$ and $\boldsymbol{\Gamma}_D$ are calculated using:

$$\boldsymbol{\Gamma}_S = \boldsymbol{i}\left[\sum_S - \sum_S^+\right] \quad \boldsymbol{\Gamma}_D = \boldsymbol{i}\left[\sum_D - \sum_D^+\right] \tag{5}$$

The retarded Green's function $\boldsymbol{G}$ is obtained using Equation (3) and the spectral functions filled by the source/drain contacts can be afterwards obtained as:

$$\boldsymbol{A}_S = \boldsymbol{G\Gamma}_S\boldsymbol{G}^+ \tag{6}$$

$$\boldsymbol{A}_D = \boldsymbol{G\Gamma}_D\boldsymbol{G}^+ \tag{7}$$

The correlation function is then calculated by:

$$\boldsymbol{G}^n(E_l) = \boldsymbol{A}_S(E)F(E_l, E_{f_S})+\boldsymbol{A}_D(E)F(E_l, E_{f_D}) \tag{8}$$

where E_{f_S} and E_{f_D} are the Fermi levels of the source and drain contacts, respectively, and the function F is given by:

$$F(E, E_f) = \sqrt{\frac{2m^*zk_BT}{\pi\hbar^2}}\Im_{-1/2}\left(\frac{E_f - E}{k_BT}\right) \tag{9}$$

and $\Im_{-1/2}$ is the Fermi-Dirac integral of order $-1/2$.

The longitudinal-energy-resolved electron density at a grid point i is obtained by:

$$\frac{n(i;E_l)=\boldsymbol{G}^{\boldsymbol{n}}(i,i;E_l)}{2\pi\Delta x\Delta y} \tag{10}$$

The longitudinal-energy-resolved electron density $n(i,E_l)$ is, further, summed over the Si six conduction band valleys and, finally, the total electron density $n(i)$ is obtained by integration over the longitudinal energy.

The transmission coefficient from the source contact to the drain contact is defined in terms of Green's function and the broadening function as:

$$\mathrm{T_{SD}} = \mathrm{Trace}\left[\boldsymbol{\Gamma}_S\boldsymbol{G}\boldsymbol{\Gamma}_D\boldsymbol{G}^+\right] \tag{11}$$

The longitudinal-energy-resolved terminal current in the ballistic limit is, afterwards, obtained as:

$$I(E_l)=\frac{q}{2\pi\hbar}T_{\mathrm{SD}}[F(E_l,E_{f_S})-F(E_l,E_{f_D})] \tag{12}$$

The terminal current is, further, summed over the six conduction band valleys and, finally, integrated over the longitudinal energy.

3. COMPUTATIONALLY EFFICIENT METHODS

The retarded Green's function is a central quantity in the NEGF. As seen from Equation (3), it is calculated by means of matrix inversion for the effective Hamiltonian matrix. The effective Hamiltonian matrix size is the same as the number of points in the grid which is $N_{\mathrm{grid}}=N_xN_y$ for 2D device. Numerical matrix inversion consumes a large number of operations that in the order of N_{grid}^3. Moreover, Green's function should be calculated for each energy point considered in the simulation. This makes total number of operations scales as $N_{\mathrm{op}}=N_{\mathrm{E}}N_{\mathrm{grid}}^3$ where N_{E} is the number of energy points. Therefore, efforts have been exerted to develop computationally efficient methods to reduce the computational burden. In this work, three methods are considered: the RGF algorithm [6–8], the CBR method [9,10], and GE method [11].

3.1. The RGF algorithm

The RGF algorithm builds up the Green's function recursively without full inversion of the Hamiltonian matrix [8]. It can be used only if the effective Hamiltonian matrix $[\boldsymbol{EI}-\boldsymbol{H}-\boldsymbol{\Sigma}]$ is block tri-diagonal. This means it allows only the nearest neighbor layers coupling in RS. Unfortunately, a lead couples all the layers connected to it [9] and, therefore, the RGF works when the device has no more than two contacts. The RGF algorithm is summarized in the following steps [8]:

Let $\boldsymbol{D}=[E_l\boldsymbol{I}-\boldsymbol{H}_d-\Sigma_S-\Sigma_D]$ and $\boldsymbol{D}_{n,m}$ denotes $D[(n-1)N_y:nN_y,(m-1)N_y:mN_y]$, then carry out the following steps for $\boldsymbol{G}$:

(1) $\boldsymbol{g}_{1,1}^{L1}=\boldsymbol{D}_{1,1}^{-1}$.
(2) For $q=1,2,\ldots,N_x-1$, compute $\boldsymbol{g}_{q+1,q+1}^{Lq+1}=(\boldsymbol{D}_{q+1,q+1}+\boldsymbol{D}_{q+1,q}\boldsymbol{g}_{q,q}^{Lq}\boldsymbol{D}_{q,q+1})^{-1}$.
(3) For $q=1,2,\ldots,N_x-1$, compute $\boldsymbol{g}_{q,q}^{+Lq}$.
(4) $\boldsymbol{G}_{N_x,N_x}=\boldsymbol{g}_{N_x,N_x}^{LN_x}$.
(5) For $q=N_x$-1, N_x-$2,\ldots,1$, compute $\boldsymbol{G}_{q,q+1}=-\boldsymbol{g}_{q,q}^{Lq}\boldsymbol{D}_{q,q+1}\boldsymbol{G}_{q+1,q+1}$, $\boldsymbol{G}_{q+1,q}=-\boldsymbol{G}_{q+1,q+1}\boldsymbol{D}_{q+1,q}\boldsymbol{g}_{q,q}^{Lq}$ and $\boldsymbol{G}_{q,q}=\boldsymbol{g}_{q,q}^{Lq}-\overset{Lq}{\boldsymbol{g}}_{q,q}\boldsymbol{D}_{q,q+1}\boldsymbol{G}_{q+1,q}$ in this order.
(6) For $q=1,2,\ldots,N_x-1$, compute $\boldsymbol{G}_{q,q+1}^+$ and $G_{q+1,q}^+$.

Let $\boldsymbol{\Sigma}^{in}(E_l)=\boldsymbol{\Gamma}_S(E_l)F(E_l,E_{f_S})+\boldsymbol{\Gamma}_D(E_l)F(E_l,E_{f_D})$, then carry out the following steps for $\boldsymbol{G}^n$:

(1) $g_{11}^{nL1}=g_{11}^{L1}\Sigma_{11}^{in}g_{11}^{L1^+}$.
(2) For $q=1,2,\ldots,N_x$-1, compute $\boldsymbol{g}_{q+1,q+1}^{nLq+1}=\boldsymbol{g}_{q+1,q+1}^{Lq+1}\left[\boldsymbol{\Sigma}_{q+1,q+1}^{in}+\boldsymbol{D}_{q+1,q}\boldsymbol{g}_{q,q}^{nLq}\boldsymbol{D}_{q,q+1}^+\right]\boldsymbol{g}_{q+1,q+1}^{Lq+1^+}$.
(3) $\boldsymbol{G}_{N_x,N_x}^n=\boldsymbol{g}_{N_x,N_x}^{nLN_x}$.

(4) For $q = N_x\text{-}1,\ N_x\text{-}2,\ldots,1$, compute $\boldsymbol{G}^n_{q,q} = \boldsymbol{g}^{nLq}_{q,q} + \boldsymbol{g}^{Lq}_{q,q}(\boldsymbol{D}_{q,q+1}\boldsymbol{G}^n_{q+1,q+1}\boldsymbol{D}^+_{q+1,q})\boldsymbol{g}^{+Lq}_{q,q} - [\boldsymbol{g}^{nLq}_{q,q}\boldsymbol{D}^+_{q,q+1}\boldsymbol{G}^+_{q+1,q} + \boldsymbol{G}_{q,q+1}\boldsymbol{D}_{q+1,q}\boldsymbol{g}^{nLq}_{q,q}]$ and $\boldsymbol{G}^n_{q+1,q} = -\boldsymbol{G}_{q+1,q+1}\boldsymbol{D}_{q+1,q}\boldsymbol{g}^{nLq}_{q,q} - \boldsymbol{G}^n_{q+1,q+1}\boldsymbol{D}^+_{q+1,q}\boldsymbol{g}^{+Lq}_{q,q}$.

(5) Use $\boldsymbol{G}^n_{q,q+1} = \boldsymbol{G}^{+n}_{q+1,q}$.

Then longitudinal-energy-resolved electron density at a grid point i is obtained from the diagonal elements of the correlation function as given by Equation (9) and the longitudinal-energy-resolved terminal current is obtained as:

$$I = \frac{q}{2\pi\hbar} Trace\{[[\boldsymbol{G}^+_{1,1}(E) - \boldsymbol{G}_{1,1}(E)]F(E_l, E_{fs}) - i\boldsymbol{G}^n_{1,1}(E)][\boldsymbol{\beta}(\boldsymbol{g}_S - \boldsymbol{g}^+_S)\boldsymbol{\beta}]\} \tag{13}$$

The operation count of this algorithm scales approximately as $N_y^3 N_x$[8]. The dependence on N_y^3 arises because matrices of the sub Hamiltonian of the device vertical layers should be inverted, and the dependence on N_x corresponds to one such inversion for each of the layers. These operations are carried out for each energy step and, therefore, the total number of operations is estimated as $N_{\text{op}} = N_E N_y^3 N_x$.

3.2. The CBR method

There are three key points in the CBR method that makes it computationally efficient relative to the traditional NEGF [9]: (1) Dyson's equation is used together with a clever splitting of the simulation domain that makes the elements of Green's function can be calculated with inversion of a relatively small matrix, (2) the isolated device eigenstates are used as a basis for the transport problem, and the number of eigenstates needed to maintain acceptable accuracy is greatly reduced by applying von Neumann boundary condition for the isolated device Hamiltonian, and (3) the use of the leads propagating modes as a basis instead of the RS basis in single-band case where only propagating modes contribute to the current.

Greens' function of the isolated device $\boldsymbol{G}^0$ is given by its spectral representation [3]:

$$\boldsymbol{G}^0(i,j,E) = \sum_{\alpha=1}^{N_{eigen}} \frac{\boldsymbol{\psi}_\alpha(i)\boldsymbol{\psi}^*_\alpha(j)}{E - \varepsilon_\alpha + i\eta}, \eta \to 0 \tag{14}$$

where $\boldsymbol{\psi}_\alpha$ are the eigenfunctions of the isolated device, ε_α are the corresponding eigenenergies, i and j are the grid point's indices in RS. The simulation domain is spitted into two sub-domains, D: the interior part of the device and C: the boundary region that connects the interior parts to the contacts. Accordingly, the isolated device Hamiltonian matrix can be written as [9]:

$$\boldsymbol{H}^0 = \begin{bmatrix} \boldsymbol{H}^0_C & \boldsymbol{H}^0_{\text{CD}} \\ \boldsymbol{H}^0_{\text{DC}} & \boldsymbol{H}^0_D \end{bmatrix}_{N_{\text{grid}} \times N_{\text{grid}}} \tag{15}$$

where $\boldsymbol{H}^0_{CN_C \times N_C}$ is a relatively small matrix corresponds to N_C boundary points ($2N_y$ in our case) and $H^0_{DN_D \times N_D}$ is a huge matrix corresponds to the N_D interior points ($N_xN_y - 2N_y$ in our case.) Finally, Green's function of the coupled device is obtained:

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{X}_C^{-1}\boldsymbol{G}^0_C & \boldsymbol{X}_C^{-1}\boldsymbol{G}^0_{\text{CD}} \\ \boldsymbol{G}^0_{\text{DC}}\Sigma_C X_C^{-1}\boldsymbol{G}^0_C + \boldsymbol{G}^0_{\text{DC}} & -\boldsymbol{G}^0_{\text{DC}}\boldsymbol{\Sigma}_C\boldsymbol{X}_C^{-1}\boldsymbol{G}^0_{\text{CD}} + \boldsymbol{G}^0_D \end{bmatrix} = \begin{bmatrix} \boldsymbol{G}_C & \boldsymbol{G}_{\text{CD}} \\ \boldsymbol{G}_{\text{DC}} & \boldsymbol{G}_D \end{bmatrix} \tag{16}$$

where $\boldsymbol{X}_C = [I\boldsymbol{G}^0_C\boldsymbol{\Sigma}_C]_{N_C \times N_C}$ is a small matrix and $\boldsymbol{\Sigma}_C$ is the self-energy matrix in the contact region.

The transmission function given by [9]:

$$T = \text{Trace}[\boldsymbol{\Gamma}^S_C\boldsymbol{G}_C\boldsymbol{\Gamma}^D_C\boldsymbol{G}^+_C] \tag{17}$$

The spectral function filled by the source/drain contacts, $\boldsymbol{A}_{S,D}$ is given by [9]:

$$\boldsymbol{A}_{S,D}(i,j;E) = \sum_{\alpha,\beta} \psi_\alpha(i)\psi^*_\beta(j)\frac{\text{Trace}(\psi_\beta\psi^+_\alpha \boldsymbol{B}_C^{-1}\boldsymbol{\Gamma}^{S,D}_C(\boldsymbol{B}_C^{-1})^+)}{(E - \varepsilon_\alpha + i\eta)(E - \varepsilon_\beta + i\eta)}, \eta \to 0 \tag{18}$$

where $\boldsymbol{B}_C = I - \boldsymbol{\Sigma}_C\boldsymbol{G}^0_C$.

The aforementioned splitting of the simulation domain and the application Von Neumann boundary condition requires the modification of the Hamiltonian matrix given in

Equation (2) to:

$$\boldsymbol{H_l} = \begin{bmatrix} \boldsymbol{\alpha}+\boldsymbol{\beta} & \boldsymbol{0} & \boldsymbol{\beta} & \boldsymbol{0} & \cdots & \cdots & \cdots & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\alpha}+\boldsymbol{\beta} & \boldsymbol{0} & \cdots & \cdots & \cdots & \cdots & \cdots & \boldsymbol{\beta} \\ \boldsymbol{\beta} & \boldsymbol{0} & \boldsymbol{\alpha} & \boldsymbol{\beta} & \boldsymbol{0} & \cdots & \cdots & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\beta} & \boldsymbol{\alpha} & \boldsymbol{\beta} & \boldsymbol{0} & \cdots & \cdots & \boldsymbol{0} \\ \vdots & \boldsymbol{0} & \boldsymbol{0} & \ddots & \ddots & \ddots & \ddots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \ddots & \ddots & \ddots & \ddots & \boldsymbol{0} \\ \vdots & \cdots & \cdots & \cdots & \cdots & \boldsymbol{0} & \boldsymbol{\beta} & \boldsymbol{\alpha} & \boldsymbol{\beta} \\ \boldsymbol{0} & \boldsymbol{\beta} & \boldsymbol{0} & \cdots & \cdots & \cdots & \boldsymbol{0} & \boldsymbol{\beta} & \boldsymbol{\alpha} \end{bmatrix} \tag{19}$$

And the corresponding self-energy matrix is given by:

$$\sum = \begin{bmatrix} \begin{bmatrix} \sum_{\boldsymbol{S}} & \boldsymbol{0} \\ \boldsymbol{0} & \sum_{\boldsymbol{D}} \end{bmatrix}_{N_C \times N_C} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}_{N_{\text{grid}} \times N_{\text{grid}}} \tag{20}$$

where $\boldsymbol{\Sigma_S} = \boldsymbol{\beta g_S \beta} - \boldsymbol{\beta}$ and $\boldsymbol{\Sigma_D} = \boldsymbol{\beta g_D \beta} - \boldsymbol{\beta}$.

The double summation on the eigenstates in Equation (18) is composed of two terms, the first one is energy independent but position dependent and the second one is the opposite. Therefore, the number of operations can be estimated as $N_{\text{op}} = N_{\text{eigen}}^2 N_E + N_{\text{eigen}}^2 N_{\text{grid}}$[9] where N_{eigen} is the number of eigenstates to be used. The Numerical calculation effort of the transmission function and the spectral function can be further reduced in the single-band case by transforming Green's function and the self-energy into a basis of the leads mode-space [9]. The idea behind choosing these modes as a basis is that if the potential is constant inside a given lead, then its modes are truly uncoupled which results in diagonal self-energy matrices. Besides diagonal self-energy matrices, only few modes contribute to the transmission at a given energy [9]. Therefore, the number of operations can be estimated as $N_{\text{op}} = N_E N_{\text{eigen}} N_{\text{modes}} N_{\text{grid}}$ [10] in which higher orders of N_{grid} or N_{eigen} are absent.

3.3. The GE Method

The idea in this method is based on the sparse nature of the broadening function which can be seen from Equations (4) and (5). Consequently, the entire Green's function isn't needed to calculate the spectral functions in Equations (6) and (7). Instead, the spectral functions are calculated using the following equations [11]:

$$\boldsymbol{A_S} = \boldsymbol{G_S}[\boldsymbol{\beta}(\boldsymbol{g_s} - \boldsymbol{g_S^+})\boldsymbol{\beta}]\boldsymbol{G_S^+} \tag{21}$$

$$\boldsymbol{A_D} = \boldsymbol{G_D}[\boldsymbol{\beta}(\boldsymbol{g_D} - \boldsymbol{g_D^+})\boldsymbol{\beta}]\boldsymbol{G_D^+} \tag{22}$$

where $\boldsymbol{G_S}$ and $\boldsymbol{G_D}$ are submatrices of the retarded Green's function and are given by:

$$\boldsymbol{G_S} = \begin{bmatrix} \boldsymbol{G}(1,1) & \boldsymbol{G}(1,2) & \cdots & \boldsymbol{G}(1,N_y) \\ \boldsymbol{G}(2,1) & \boldsymbol{G}(2,2) & \cdots & \boldsymbol{G}(2,N_y) \\ \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{G}(N_xN_y,1) & \boldsymbol{G}(N_xN_y,2) & \cdots & \boldsymbol{G}(N_xN_y,N_y) \end{bmatrix}_{N_xN_y \times N_y} \tag{23}$$

$$\boldsymbol{G_D} = \begin{bmatrix} \boldsymbol{G}(1,(N_x-1)N_y+1) & \boldsymbol{G}(1,(N_x-1)N_y+2) & \cdots & \boldsymbol{G}(1,N_y) \\ \boldsymbol{G}(2,(N_x-1)N_y+1) & \boldsymbol{G}(2,(N_x-1)N_y+2) & \cdots & \boldsymbol{G}(2,N_y) \\ \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{G}(N_xN_y,(N_x-1)N_y+1) & \boldsymbol{G}(N_xN_y,(N_x-1)N_y+2) & \cdots & \boldsymbol{G}(N_xN_y,N_y) \end{bmatrix}_{N_xN_y \times N_y} \tag{24}$$

DOI: 10.1002/jnm

The matrices in Equations (23) and (24) can be obtained using GE method by the following equations:

$$\boldsymbol{G}_S = [E_l\boldsymbol{I} - \boldsymbol{H}_d - \boldsymbol{\Sigma}_S - \boldsymbol{\Sigma}_D]\backslash \boldsymbol{I}_S \tag{25}$$

$$\boldsymbol{G}_D = [E_l\boldsymbol{I} - \boldsymbol{H}_d - \boldsymbol{\Sigma}_S - \boldsymbol{\Sigma}_D]\backslash \boldsymbol{I}_D \tag{26}$$

where the $A\backslash B$ denotes division of the B by A using GE method, $\boldsymbol{I}_S$ and $\boldsymbol{I}_D$ are given by:

$$\boldsymbol{I}_S = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ 0 & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix}_{N_xN_y\times N_y} \tag{27}$$

$$\boldsymbol{I}_D = \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & \vdots \\ 0 & \cdots & \cdots & 1 \end{bmatrix}_{N_xN_y\times N_y} \tag{28}$$

The transmission function can also be efficiently calculated efficiently using the following equation [11]:

$$T_{\mathrm{SD}} = \mathrm{Trace}[(\boldsymbol{\beta}(\boldsymbol{g}_s - \boldsymbol{g}_S^{+})\boldsymbol{\beta}]\boldsymbol{G}_{\mathrm{DS}}[\boldsymbol{\beta}(\boldsymbol{g}_D - \boldsymbol{g}_D^{+})\boldsymbol{\beta}]\boldsymbol{G}_{\mathrm{DS}}^{+}) \tag{29}$$

where $\boldsymbol{G}_{DS}$ is a subset of $\boldsymbol{G}_D$ and given by:

$$\boldsymbol{G}_{DS} = \begin{bmatrix} \boldsymbol{G}(1,(N_x-1)N_y+1) & \boldsymbol{G}(1,(N_x-1)N_y+2) & \cdots & \boldsymbol{G}(1,N_y) \\ \boldsymbol{G}(2,(N_x-1)N_y+1) & \boldsymbol{G}(2,(N_x-1)N_y+2) & \cdots & \boldsymbol{G}(2,N_y) \\ \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{G}(N_y,(N_x-1)N_y+1) & \boldsymbol{G}(N_y,(N_x-1)N_y+2) & \cdots & \boldsymbol{G}(N_y,N_y) \end{bmatrix}_{N_y\times N_y} \tag{30}$$

Using Equations (24) and (25), we calculate only $N_xN_y\times 2N_y$ elements of Green's function instead of calculating $N_xN_y\times N_xN_y$ elements.

4. RESULTS AND DISCUSSION

The new version of FETMOSS using the RS methods discussed in Section 3 has been benchmarked using an online available simulator NanoMOS 3.0 [4] on nanohub: www.nanohub.org. NanoMOS3.0 is a 2D simulator for DG n-MOSFETs that uses the UMS representation for quantum ballistic transport simulation. The use of the UMS representation limits the simulator capability to thin bodies DG MOSFETs (less than 5 nm). It is expected to find well agreement between UMS NEGF in NanoMOS and RS NEGF in FETMOSS, for ultra-thin body thickness where the UMS is valid. A sample nanoscale DG MOSFET has been simulated using NanoMOS 3.0 and FETMOSS. The simulated device dimensions, doping concentrations, material parameters, simulator options, the finite difference grid spacing, and the supply voltage are given in Table I. The simulation results are shown in Figures 4 and 5 where good agreement between the two simulators can be observed.

Now, we have the three methods discussed in Section 3 implemented and integrated into FETMOSS. It is the time to compare their computational efficiency relative to the traditional

DOI: 10.1002/jnm

Table I. The simulated devices dimensions, doping concentration, material parameters, simulator options, the finite difference grid spacing, and the supply voltage.

Category	Parameter	Value
Dimensions	Channel length (L)	5 nm
	Source and drain length (L_S, L_D)	5 nm
	Oxide thickness (T_{ox})	1 nm
	Silicon (body) thickness (T_{Si})	2 nm
Doping	Channel doping	10^{10} cm^{-3}
	Source and drain doping	2×10^{20} cm^{-3}
	Junction doping profile	step
Material	Silicon relative permittivity (ε_{Si})	11.7 ε_0
	Oxide relative permittivity (ε_{ox})	3.9 ε_0
	Top and bottom gate work function (φ_m)	4.5 eV
	Longitudinal electron effective mass m_l^*	0.91 m_0
	Transverse electron effective mass m_t^*	0.19 m_0
	Self-consistence tolerance (δ)	10^{-3} V
	Poisson's tolerance	10^{-6} V
Grid	Vertical node spacing	0.1 nm
	Horizontal node spacing	0.2 nm
Supply voltage	V_{DD}	0.7 V

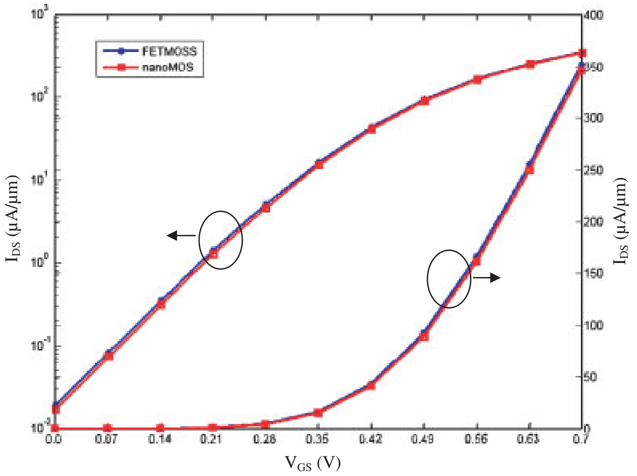

Figure 4. The I_{DS}-V_{GS} characteristics of the simulated device at $V_{DS} = 25$ mV. The left axis is the log scale while the right axis is the linear scale.

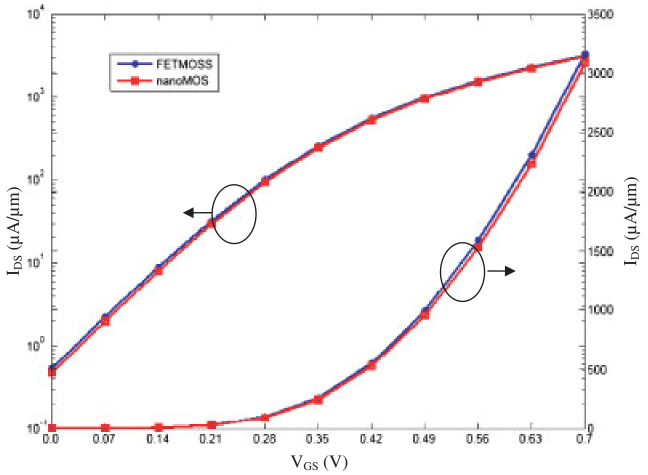

Figure 5. The I_{DS}-V_{GS} characteristics of the simulated device at $V_{DS} = 0.7$ V. The left axis is the log scale, whereas the right axis is the linear scale.

DOI: 10.1002/jnm

NEGF. For this purpose, the four methods (the traditional NEGF, the RGF algorithm, the CBR method, and the GE method) were used to simulate the device given in Table I. The drain voltage was kept constant at 0.7 V and the gate voltage was swept from 0.0 to 0.7 V with a step of 0.1 V. Thus, we have eight bias points. For the first bias point, i.e. $V_{GS} = 0.0$ V, the initial guess was taken to be the zero potential at various grid points in the device. The initial guess for any other bias point was taken from the solution of the preceding bias point, for example initial guess for $V_{GS} = 0.1$ V was taken from the solution of $V_{GS} = 0.0$ V. It is important to mention that the traditional NEGF, the RGF algorithm, and the GE method are all giving exactly the same current for the same applied voltage. This is because neither the RGF algorithm nor the GE methods trades off the accuracy with the simulation speed, whereas the CBR method does. A key parameter in the CBR method is the number of eigenstates (N_{eigen}) used in the simulation. The lesser the eigenstates, the faster the simulation and the lesser accurate are the results. It has been demonstrated that the needed percentage of eigenstates for a given acceptable accuracy (less than 5% in the terminal current) is bias dependent, and can vary from 6% in the on-state to 40% in the off-state [18]. For this reason, N_{eigen} was decreased gradually from 40% at $V_{GS} = 0.0$ V to 6% at $V_{GS} = 0.7$ V.

Figure 6 depicts the self-consistent error versus time for the traditional NEGF, the RGF algorithm and the GE method, whereas Figure 7 depicts it for the CBR method. A solution is found when the error drops below 1 mV. Once this criterion is met, the terminal current is calculated and a new bias point is initiated. This causes the error to jump to a larger value, and the error starts decreasing again with iterations until the solution of the new bias point is found. The cycle was repeated until the eight bias points were completed. The simulation time differs considerably from one method to another. The traditional NEGF with full matrix inversion has the greatest simulation time, whereas the CBR method has the smallest one. A summary of the total simulation time (t_{total}), average simulation time per bias point (t_{bias}), and the average simulation time per iteration ($t_{\text{iteration}}$) is presented in Table II. These simulations were carried

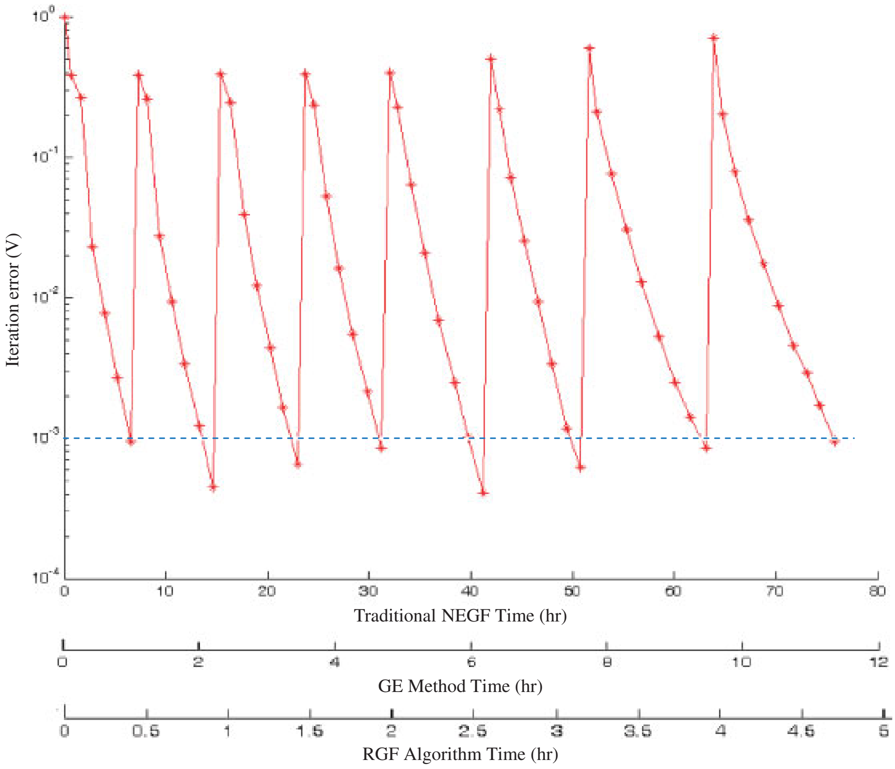

Figure 6. The self-consistent error versus time using the traditional NEGF, the RGF algorithm and the GE method.

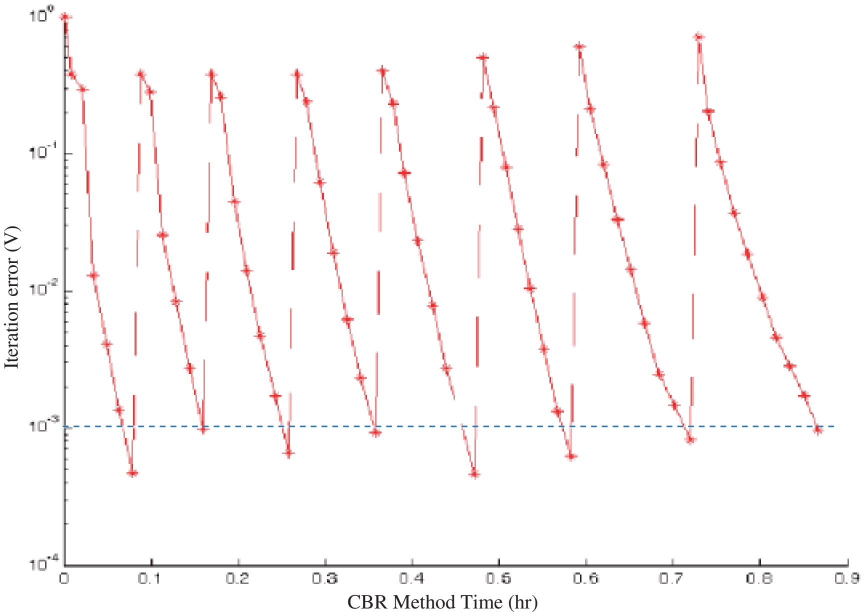

Figure 7. The self-consistent error versus time using the CBR method.

Table II. The simulation time comparison summary.

Method	t_{total} (h)	t_{bias} (h)	$t_{\mathrm{iteration}}$ (h)
Traditional NEGF	75.7218	9.4652	1.2022
The GE method	11.6525	1.4566	0.1850
The RGF algorithm	5.0872	0.6359	0.0807
The CBR method	0.8661	0.1108	0.0135

These simulations were carried out on a home PC: Intel® Pentium 4 CPU 2.4 GHz, 768 MB RAM.

out on a home PC: Intel® Pentium 4 CPU 2.4 GHz, 768 MB RAM. The RGF algorithm or the GE method introduces about one order of magnitude reduction in simulation time below that traditional NEGF. The CBR method yields the smallest simulation time with about two orders of magnitude reduction. By such a great reduction in the simulation time, the CBR method makes it practical to simulate and design DG MOSFETs using the RS simulations. The main disadvantage of the CBR method is the necessity to dynamically determine the number of eigenstates to achieve the desired accuracy [18, 19]. This is not necessary in either the RGF algorithm or the GE method.

5. CONCLUSIONS

The traditional NEGF, the RGF algorithm, the CBR method, and the GE method were successfully implemented in the DG MOSFETs simulator FETMOSS. The new version of FETMOSS was benchmarked using the online available simulator NanoMOS 3.0. The existence of the mentioned methods inside the same simulator enables their simulation time comparison by using them to simulate the same DG MOSFET device on the same machine. The results showed that the CBR method is the most computationally efficient one with about two order of magnitude reduction in time with respect to the traditional NEGF. The RGF algorithm is a little bit faster than the GE method and both of them give about one order of magnitude only of simulation time reduction. From these results, one expects that the CBR method will make it practical to simulate and design true 2D and may be 3D devices on a home PC, in single-band case. For multi-band structure simulation, the CBR method may be comparable in speed to the other methods and the relative ranking of the methods is needed to be studied in the future.

DOI: 10.1002/jnm

REFERENCES

1. International Technology Roadmap for Semiconductors. Available from: http://www.itrs.net/about.html.
2. Sverdlov V, Kosina H, Selberherr S. Modeling current transport in ultra-scaled field-effect transistors. *Microelectronics Reliability* 2007; **47**:11–19.
3. Datta S. *Electronic Transport in Mesoscopic Systems*. Cambridge University Press: Cambridge, U.K., 1995.
4. Ren Z, Venugopal R, Goasguen S, Datta S, Lundstrom MS. nanoMOS 2.5: a two-dimensional simulator for quantum transport in double-gate MOSFETs. *IEEE Transactions on Electron Devices* 2003; **50**:1914–1925.
5. Sabry YM, Abdolkader TM, Farouk WF. Uncoupled mode-space simulation validity for double gate MOSFETs. *International Conference on Microelectronics* 2007; 351–354.
6. Svizhenko A, Anantram MP, Govindan TR, Biegel B, Venugopal R. Two-dimensional quantum mechanical modeling of nanotransistors. *Journal of Applied Physics* 2002; **91**(4):2343–2354.
7. Klimeck G, Oyafuso F, Boykin TB, Bowen RC, von Allmen P. Development of a nanoelectronic 3-D (NEMO 3-D) simulator for multimillion atom simulations and its application to alloyed quantum dots. *Computer Modeling in Engineering and Science* 2002; **3**(5):601–642.
8. Anantram M, Lundstrom M, Nikonov D. Modeling of nanoscale devices, condensed matter, mesoscopic systems and quantum hall effect, 2006 [online]. Available from: arXiv:cond-mat/0610247v2.
9. Mamaluy D, Vasileska D, Sabathil M, Zibold T, Vogl P. Contact block reduction method for ballistic transport and carrier densities of open nanostructures. *Physical Review* 2005; **B 71**:245321.
10. Khan H, Mamaluy D, Vasileska D. Self-consistent treatment of quantum transport in 10 nm FinFET using contact block reduction (CBR) method. *Journal of Computational Electronics* 2006; 77–80.
11. Sabry YM, Abdel-Hafez MT, Abdolkader TM, Farouk WF. A computationally efficient method for quantum transport simulation of double-gate MOSFETs. *National Radio Science Conference*, Cairo, Egypt 2009.
12. Venugopal R, Ren Z, Datta S, Lundstrom M. Simulating quantum transport in nanoscale MOSFETs: real versus mode space approaches. *Journal of Applied Physics* 2002; **92**:3730–3739.
13. Rahman A, Lundstrom MS, Ghosh AW. Generalized effective-mass approach for *n*-type metal-oxide-semiconductor field-effect transistors on arbitrarily oriented wafers. *Journal of Applied Physics* 2005; **97**:053702.
14. Khan HR, Mamaluy D, Vasileska D. Approaching optimal characteristics of 10 nm high performance devices: a quantum transport simulation study of Si FinFET. *IEEE Transactions on Electron Devices* 2008; **55**:743–753.
15. Adolkader T, Farouk W, Omar O, Hassan M. FETMOSS: software tool for 2D simulation of double-gate MOSFET. *International Journal of Numerical Modeling* 2006; **19**:214–301.
16. Datta S. *Quantum Phenomena*, *Modular Series on Solid-state Devices*, vol. VIII. Addison Wesley: New York, 1989.
17. Damle P. Nanoscale device modeling: from MOSFETs to molecules. *Ph.D. Dissertation*, Purdue University, West Lafayette, 2003.
18. Sabry YM, Attaby A, Abdolkader TM, Farouk WF. Inspection of the contact block reduction method for quantum transport simulation of FinFETs. *National Radio Science Conference*, Cairo, Egypt 2009.
19. Khan HR, Mamaluy D, Vasileska D. Quantum transport simulation of experimentally fabricated nano-FinFET. *IEEE Transactions on Electron Devices* 2007; **54**(4):784–796.

AUTHORS' BIOGRAPHIES

Yasser M. Sabry was born in Kuwait in 1982. He received the BS and MS degrees in Electronics and Communication Engineering from Ain Shams University, Cairo, Egypt in 2005 and 2009, respectively. He is currently pursuing the PhD degree in Electrical Engineering at the Microsystems Dept., ESIEE-Paris, Paris-Est University, France. From 2005 to 2008, he was with Mentor Graphics Corp., Egypt, where he was engaged in device modeling and simulation of IC circuits. Since 2009, he is with Si-Ware Systems, Egypt where he is involved in design, fabrication and characterization of MEMS devices. His research interests are modeling and simulation of nanoscale transistors, characterization of electronic and MEMS devices as well as design and fabrication of optical MEMS components.

Tarek M. Abdolkader was born in Cairo, Egypt in 1970. He received the BS degree in Electrical Engineering (electronics and communications) from the faculty of Engineering, Ain-Shams University, Cairo in 1992, another BS degree in Physics from the faculty of Science, Ain-Shams University, Cairo in 1996, and MS and PhD degrees in Engineering Physics from the faculty of Engineering, Ain-Shams University, Cairo in 2001 and 2005, respectively. He is currently a Post-Doctoral Research Associate in the School of Electrical and Computer Engineering, Purdue University, U.S.A. His research interests are modeling and simulation of Nano-electronic devices.

DOI: 10.1002/jnm

Wael F. Farouk was born in Cairo, Egypt in 1962. He received the BS degree in Electrical Engineering from Ain Shams University, Faculty of Engineering, Cairo, Egypt in 1984 and the MS and PhD degrees in Engineering Physics from the same university in 1989 and 1994, respectively. He is currently associate professor of solid state electronics in the Engineering Physics Department. His research interests include VLSI MOSFET and SOI devices characterization and modeling, solar cells, and silicon-electrochemical cells.

DOI: 10.1002/jnm