A Computationally Efficient Method for Quantum Transport Simulation of Double-Gate MOSFETs

Yasser M. Sabry[1] Mohammed T. Abdel-Hafez[2] Tarek M. Abdolkader[3]
Wael Fikry Farouk[4]

[1]Department of Electronics and Communication, Faculty of Engineering, Ain Shams University, Egypt. ysabry@ieee.org

[2]Silicon Vision. mohamed.tawfik@si-vision.com

[3]Department of Basic Sciences, Higher Institute of Technology, Benha University, Egypt. tarik_mak@hotmail.com

[4]Department of Engineering Physics and Mathematics, Faculty of Engineering, Ain Shams University, Egypt. wael_fikry@ieee.org

Abstract

Quantum effects play an important role in determining the Double-Gate (DG) MOSFETs characteristics. The Non-Equilibrium Green's Function (NEGF) formalism provides a rigorous description of quantum transport in nanoscale devices. The traditional NEGF is heavy in computations and not suitable for 3D or even 2D device simulation. In this article, we propose a method that reduces the simulation time dramatically without loss of accuracy. The proposed method is used to simulate a 5 *nm* channel length DG MOSFET. The simulation time is reduced from 72 minutes to 11 minutes per bias point on a home PC: Intel® Pentium 4 CPU 2.4GHz, 768 MB RAM.

1. Introduction

Rapid device scaling pushes the dimensions of the field effect transistors to the nanometer regime [1]. The International Technology Roadmap for Semiconductors 2007 (ITRS) projection for the DG MOSFETs physical gate length is 4.5 *nm* for the year 2022. For these extremely scaled dimensions, the DG MOSFETS characteristics are greatly affected by the quantum effects. These effects can be accurately predicted only using quantum mechanical based device simulation [2].

The quantum transport in nanoscale devices can be rigorously described by the non-equilibrium Green's function (NEGF) formalism [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 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. The NEGF method has the advantage of being rigorous but the disadvantage of being heavy in computations [4].The 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.

In this work, we propose a method to accelerate the NEGF computations without loss in accuracy. The method is based on a clever manipulation of the matrices and making use of the sparse nature of some of them. The method can be applied to various device structures, but we used the DG MOSFET as an example to illustrate the method. In section 2, the model device geometry is described, the assumptions used in the transport model are listed, and a brief overview on the traditional NEGF method is given. In section 3, the proposed method is presented and its computational cost is compared to that of the traditional NEGF. In section 4, the results of the proposed method are compared to the NEGF's from both accuracy and simulation time points of view. Finally, this article is concluded in section 5.

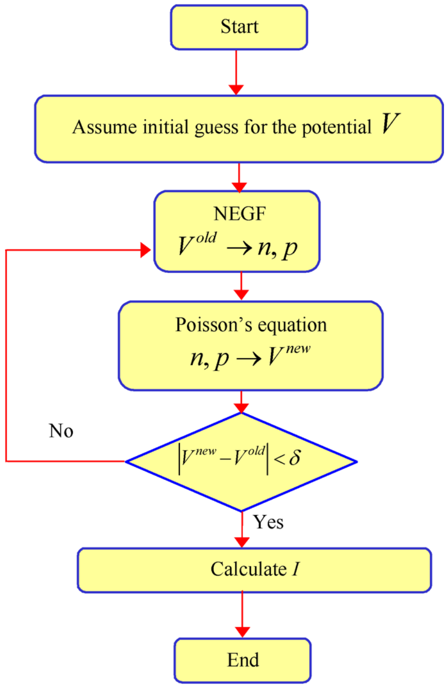

Start
Assume initial guess for the potential V
NEGF
$V^{old} \rightarrow n, p$
Poisson's equation
$n, p \rightarrow V^{new}$
No
$|V^{new} - V^{old}| < \delta$
Yes
Calculate I
End

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

2. Quantum Transport Using the Traditional NEGF

The DG MOSFET model device geometry is shown in figure 2. The following assumptions are usually encountered in the literature: 1) The channel length in *x*-direction is shorter than any characteristic scattering length such that the device is operating in the ballistic limit [3], 2) the width of the device in the *z*-direction is so large compared to other dimensions of the active device; such that the potential along that direction is rendered constant, 3) huge metal contacts; i.e. reservoirs, where thermal equilibrium is maintained and the Fermi level in these regions is determined by the applied voltage [3], 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, and 6) a single band effective mass Hamiltonian [5-6] is used to model the electron transport.

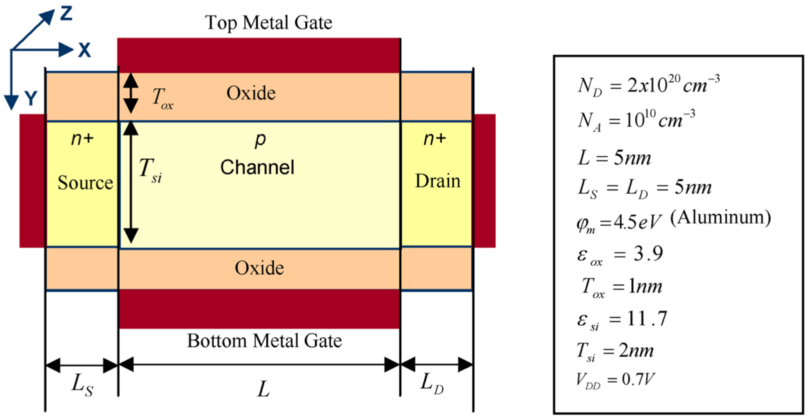


Figure 2. A model double-gate MOSFET used in this work

Generally, the 3D single band effective mass Schrödinger equation for a homogenous medium can be written as :

$$\left[-\frac{\hbar^2}{2}\left(\frac{1}{m^*_x}\frac{\partial^2}{\partial x^2}+\frac{1}{m^*_y}\frac{\partial^2}{\partial y^2}+\frac{1}{m^*_z}\frac{\partial^2}{\partial z^2}\right)+E_C\left(x,y,z\right)\right]\Psi\left(x,y,z\right)=E\Psi\left(x,y,z\right) \qquad (1)$$

where $\Psi(x,y,z)$ is the envelope wave function, m_x^*, m_y^*, and m_z^* are the electron effective mass in x- , y- and z-direction respectively, E_c is the conduction band edge and E is the total energy. The eigenstates in the transverse (z- direction) are plane waves as a consequence of the assumption of invariant potential in that direction; i.e. $E_c=E_c(x,y)$. The envelope wave function can be expanded in terms of the orthonormal basis $\psi(x,y)exp(jk_zz)/\sqrt{w}$ where the quantum number k_z corresponds to the transverse eigenenergy $E_{kz}=\hbar^2k_z^2/2m_z^*$ and w is the transistor channel width. The 2D wave function $\psi(x,y)$ is obtained from the solution of the 2D Schrödinger equation:

$$\left[-\frac{\hbar^2}{2}\left(\frac{1}{m^*_x}\frac{\partial^2}{\partial x^2}+\frac{1}{m^*_y}\frac{\partial^2}{\partial y^2}\right)+E_c\left(x,y\right)\right]\psi\left(x,y\right)=E_l\psi\left(x,y\right) \qquad (2)$$

where $E_l = E - E_{kz}$ is the longitudinal energy due to motion in x- and y- direction.

As long as we neglect elastic and inelastic scattering process that couple different transverse modes, we can think of the transverse modes as separate 2D devices connected in parallel [7]. Each transverse mode k_z has an extra transverse energy E_{kz} that should be added to the longitudinal energy whenever the total energy is needed. The Fermi Dirac function f needs the total energy as an argument and, therefore, it should be replaced by another function F that takes the transverse modes into account and is given by [8]:

$$F(E,E_f)=\sum_{k_z}f\left(E_l+E_{k_z},E_f\right)=\sqrt{\frac{2m_z^*k_BT}{\pi\hbar^2}}\Im_{-1/2}\left(\frac{E_f-E}{k_BT}\right) \qquad (3)$$

where $\Im_{-1/2}$ is the Fermi-Dirac integral of order –1/2. The summation was carried out by transforming it to integration over the transverse energy E_{kz} using the relation connecting it to the quantum number k_z. The above function accounts for all the transverse modes exactly as well as the sum over spins. Now the device can be treated as if it was purely 2D [8].

Equation (2) is discretized using finite difference technique. This results in a 2D grid of N_x and N_y points in the x and y directions separated by Δx and Δy respectively. This gives a set of N_xN_y linear equations terminated at the source and drain using closed boundary condition. The equations are cast in the matrix form:

$$\boldsymbol{H}_l\boldsymbol{\psi}=E_l\boldsymbol{I\psi} \qquad (4)$$

where

$$\boldsymbol{H}_l=\begin{bmatrix}\boldsymbol{\alpha}+E_{C1} & \boldsymbol{\beta} & \boldsymbol{0} & \cdots & \boldsymbol{0}\\ \boldsymbol{\beta} & \boldsymbol{\alpha}+E_{C2} & \boldsymbol{\beta} & \cdots & \boldsymbol{0}\\ \boldsymbol{0} & \cdots & \ddots & \cdots & \vdots\\ \boldsymbol{0} & \boldsymbol{0} & \cdots\ \boldsymbol{\beta} & & \boldsymbol{\alpha}+E_{C_{N_x.N_y}}\end{bmatrix}_{N_x.N_y\times N_x.N_y},\ \boldsymbol{\alpha}=\begin{bmatrix}2t_x+2t_y & -t_y & 0 & \cdots & 0\\ -t_y & 2t_x+2t_y & -t_y & \cdots & 0\\ 0 & \cdots & \ddots & \cdots & \vdots\\ 0 & 0 & \cdots & & -t_y\,2t_x+2t_y\end{bmatrix}_{N_y\times N_y}$$

$$\boldsymbol{\psi}=\begin{Bmatrix}\psi_1\\ \psi_2\\ \vdots\\ \psi_{N_x\times N_y}\end{Bmatrix}_{Nx.Ny\times 1},\ \boldsymbol{\beta}=\begin{bmatrix}-t_x & 0 & \cdots & 0\\ 0 & -t_x & \vdots & 0\\ \vdots & \cdots & \ddots & \vdots\\ 0 & \cdots & \cdots & -t_x\end{bmatrix}_{N_y\times N_y}$$

, $t_x=\hbar^2/2m^*_x(\Delta x)^2$, $t_y=\hbar^2/2m^*_y(\Delta y)^2$ and I is the identity matrix.

Within the NEGF framework, the correct boundary condition, accounting for the effect of the infinite contacts, is considered using the self energy concept [7]. This concept allows us to eliminate the source/drain (S/D) contacts and work only within the active device subspace. The S/D self energies are given by:

$$\Sigma_S = \begin{bmatrix} \boldsymbol{\beta g_S \beta} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \\ \boldsymbol{0} & \cdots & \cdots & \boldsymbol{0} \end{bmatrix}_{N_xN_y \times N_xN_y} \qquad \Sigma_D = \begin{bmatrix} \boldsymbol{0} & \cdots & \cdots & \boldsymbol{0} \\ \vdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{\beta g_D \beta} \end{bmatrix}_{N_xN_y \times N_xN_y} \tag{5}$$

where $\boldsymbol{g}_S$ and $\boldsymbol{g}_D$ are the surface Green's functions of the source and drain contacts respectively [8]. Once the self energies are calculated, the Green's function is given by:

$$\boldsymbol{G}(E_l) = \left[E_l \boldsymbol{I} - \boldsymbol{H}_l - \boldsymbol{\Sigma}_S - \boldsymbol{\Sigma}_D\right]^{-1} \tag{6}$$

Then, the electron density and the terminal current for a given longitudinal energy can be obtained by [7]:

$$n(E_l) = Diag[\ \boldsymbol{A}_S F(E_l, E_{f_S}) + \boldsymbol{A}_D F(E_l, E_{f_D})\ ]/\left(2\pi\Delta x\Delta y\right) \tag{7}$$

and:

$$I(E_l) = q[F(E_l, E_{f_S}) - F(E_l, E_{f_D})]T_{SD}(E_l)/\left(2\pi\hbar\right) \tag{8}$$

where the spectral functions

$$\boldsymbol{A}_S = \boldsymbol{G\Gamma_S G}^+ \qquad \boldsymbol{A}_D = \boldsymbol{G\Gamma_D G}^+ \tag{9}$$

the transmission function

$$T_{SD} = Trace[\ \boldsymbol{\Gamma_S G \Gamma_D G}^+\ ] \tag{10}$$

and the broadening functions

$$\boldsymbol{\Gamma}_S = i\left(\boldsymbol{\Sigma}_S - \boldsymbol{\Sigma}_S^+\right) \qquad \boldsymbol{\Gamma}_D = i\left(\boldsymbol{\Sigma}_D - \boldsymbol{\Sigma}_D^+\right) \tag{11}$$

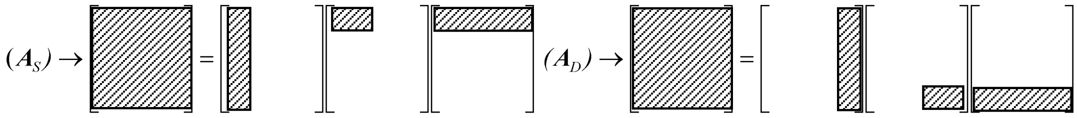
The total electron density and current are obtained by integrating over all E_l and summing over all the conduction band valleys [6].

3. The Proposed Method

The idea is based on the sparse nature of the broadening functions given by equation in the coherent transport case. This sparse nature can be deduced by substitution of equation (5) into (11) which gives:

$$\boldsymbol{\Gamma}_S = i\begin{bmatrix} \left[\boldsymbol{\beta(g_S - g_S^+)\beta}\right]_{N_y \times N_y} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \\ \boldsymbol{0} & \cdots & \cdots & \boldsymbol{0} \end{bmatrix}_{N_xN_y \times N_xN_y} \qquad \boldsymbol{\Gamma}_D = i\begin{bmatrix} \boldsymbol{0} & \cdots & \cdots & \boldsymbol{0} \\ \vdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \left[\boldsymbol{\beta(g_D - g_D^+)\beta}\right]_{N_y \times N_y} \end{bmatrix}_{N_xN_y \times N_xN_y} \tag{12}$$

Due to sparse broadening functions, the entire Green's function isn't needed to calculate the spectral functions given by equation (9). Schematically equation (9) can be seen as follows:



where the shaded area is the non-zero elements needed in (or resulted from) the multiplication process. As a result, we need only the two small parts of the Green's function denoted by $\boldsymbol{G}_S$ and $\boldsymbol{G}_D$ in equation (13):

$$G_S=\begin{bmatrix} G(1,1) & G(1,2) & \cdots & G(1,N_y) \\ G(2,1) & G(2,2) & \cdots & G(2,N_y) \\ \vdots & \vdots & \cdots & \vdots \\ G(N_xN_y,1) & G(N_xN_y,2) & \cdots & G(N_xN_y,N_y) \end{bmatrix}_{N_xN_y\times N_y} \qquad G_D=\begin{bmatrix} G(1(N_x-1)N_y+1) & G(1(N_x-1)N_y+2) & \cdots & G(1N_y) \\ G(2(N_x-1)N_y+1) & G(2(N_x-1)N_y+2) & \cdots & G(2N_y) \\ \vdots & \vdots & \cdots & \vdots \\ G(N_xN_y,(N_x-1)N_y+1) & G(N_xN_y,(N_x-1)N_y+2) & \cdots & G(N_xN_y,N_y) \end{bmatrix}_{N_xN_y\times N_y} \tag{13}$$

This can be understood from the physical meaning of the Green's function; it is the impulse response of the system [3]. More precisely, ***G(i,j)*** is the response at point *i* due to an impulse excitation at point *j*. In our case, excitation sources are the source and drain contacts. Thus we have excitations only at points 1, 2… N_y and at points $(N_x – 1)N_y + 1$, $(N_x – 1)N_y + 2…N_xN_y$ of the descretized grid where the points are counted vertically. The matrices in equation (13) can be obtained using Gauss elimination method as follows:

$$\boldsymbol{G}_S=\left[E_l\boldsymbol{I}-\boldsymbol{H}_l-\Sigma_S-\Sigma_D\right]\backslash\,\boldsymbol{I}_S \qquad \boldsymbol{G}_D=\left[E_l\boldsymbol{I}-\boldsymbol{H}_l-\Sigma_S-\Sigma_D\right]\backslash\,\boldsymbol{I}_D \tag{14}$$

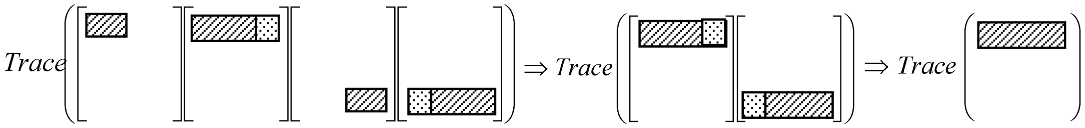
where the $A\backslash B$ denotes division of the B by A using Gauss elimination method, $\boldsymbol{I}_S$ and $\boldsymbol{I}_D$ are given by:

$$\boldsymbol{I}_S=\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ 0 & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix}_{N_xN_y\times N_y} \qquad \boldsymbol{I}_D=\begin{bmatrix} 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ 1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & \vdots \\ 0 & \cdots & \cdots & 1 \end{bmatrix}_{N_xN_y\times N_y} \tag{15}$$

It is important to note that, using equation (14), we calculate only $N_xN_y\times2N_y$ elements of the Green's function instead of calculating $N_xN_y\times N_xN_y$ elements. This is a great reduction in simulation time from the operations counts point of view. The spectral functions are calculated using the following equations:

$$\boldsymbol{A}_S=\boldsymbol{G}_S[\boldsymbol{\beta}(\boldsymbol{g}_S-\boldsymbol{g}_S^+)\boldsymbol{\beta}]\boldsymbol{G}_S^+ \qquad \boldsymbol{A}_D=\boldsymbol{G}_D[\boldsymbol{\beta}(\boldsymbol{g}_D-\boldsymbol{g}_D^+)\boldsymbol{\beta}]\boldsymbol{G}_D^+ \tag{16}$$

It is interesting to note that the transmission function can also be efficiently calculated without the full Green's function matrix. Schematically equation (10) can be seen as follows:



where the shaded area is the non-zero elements needed in (or resulted from) the multiplication process and the dotted area the elements that, after multiplication, results in the diagonal elements needed to perform the trace operation. Then, the transmission can be efficiently calculated using the following equation:

$$T_{SD}=Trace\left([\boldsymbol{\beta}(\boldsymbol{g}_S-\boldsymbol{g}_S^+)\boldsymbol{\beta}]\boldsymbol{G}_{DS}[\boldsymbol{\beta}(\boldsymbol{g}_D-\boldsymbol{g}_D^+)\boldsymbol{\beta}]\boldsymbol{G}_{DS}{}^+\right) \tag{17}$$

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

$$\boldsymbol{G}_{DS}=\begin{bmatrix} G(1,(N_x-1)N_y+1) & G(1,(N_x-1)N_y+2) & \cdots & G(1,N_y) \\ G(2,(N_x-1)N_y+1) & G(2,(N_x-1)N_y+2) & \cdots & G(2,N_y) \\ \vdots & \vdots & \cdots & \vdots \\ G(N_y,(N_x-1)N_y+1) & G(N_y,(N_x-1)N_y+2) & \cdots & G(N_y,N_y) \end{bmatrix}_{N_y\times N_y} \tag{18}$$

4. Results and Discussion

The NEGF in both its traditional and proposed form was implemented and integrated into FETMOSS [9]. It is a software tool that works under MATLAB environment and used for the 2D simulation of DG

MOSFETs with (100) oriented wafers. A sample of nanoscale DG MOSFETs has been simulated. The simulated device dimensions, doping concentration and material parameters were chosen to be compatible with the ITRS [1] and given in figure 2.

Figure 3 depicts the I_D-V_G characteristics of the device shown in figure 2. The simulation was carried out at $V_D=V_{DD}$ and $V_D=25\ mV$ while the $V_S=0.0\ V$ and V_G is swept from 0 to V_{DD} with a step of $0.1V$. There is an exact agreement between the traditional NEGF and the proposed one. In fact, the proposed method doesn't make any assumptions to reduce the computations and, therefore, there is no loss of accuracy at all.

In order to feel the reduction in simulation time, the time of each iteration within each bias point was recorded. We have eight bias points and for the first bias point; i.e. $V_G=0.0\ V$, 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, the initial guess for $V_G=0.1\ V$ was taken from the solution of $V_G=0.0\ V$. The simulation was carried out on a home PC: Intel® Pentium 4 CPU 2.4GHz, 768 MB RAM.

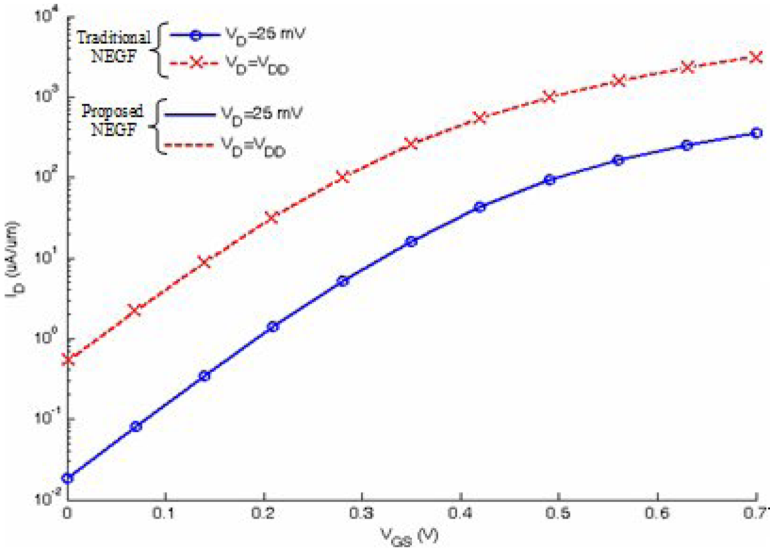


Figure 3. The I_D-V_G characteristics of the device shown in figure 2. The simulation was carried out at $V_D=V_{DD}$ and $V_D=25$ mV.

Figure 4 depicts the self-consistent error versus time for both the traditional NEGF and the proposed 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. It is important to observe that both the traditional NEGF and the proposed one have exactly the same values of error at various iterations. The simulation time, however, differs considerably between the two methods. The simulation time for the traditional NEGF is about 80 hours while it is about 12 hours for the proposed method. A summary of the total simulation time, average simulation time per bias point and the average simulation time per iteration is presented in table 1. The proposed method reduces the simulation time per iteration from 72 minutes to 11 minutes which represents an 85% reduction of the simulation time (61 minutes reduction out of 72 minutes). Thus the proposed method makes it practical to carry out quantum simulation on home PCs for device design and characteristics prediction.

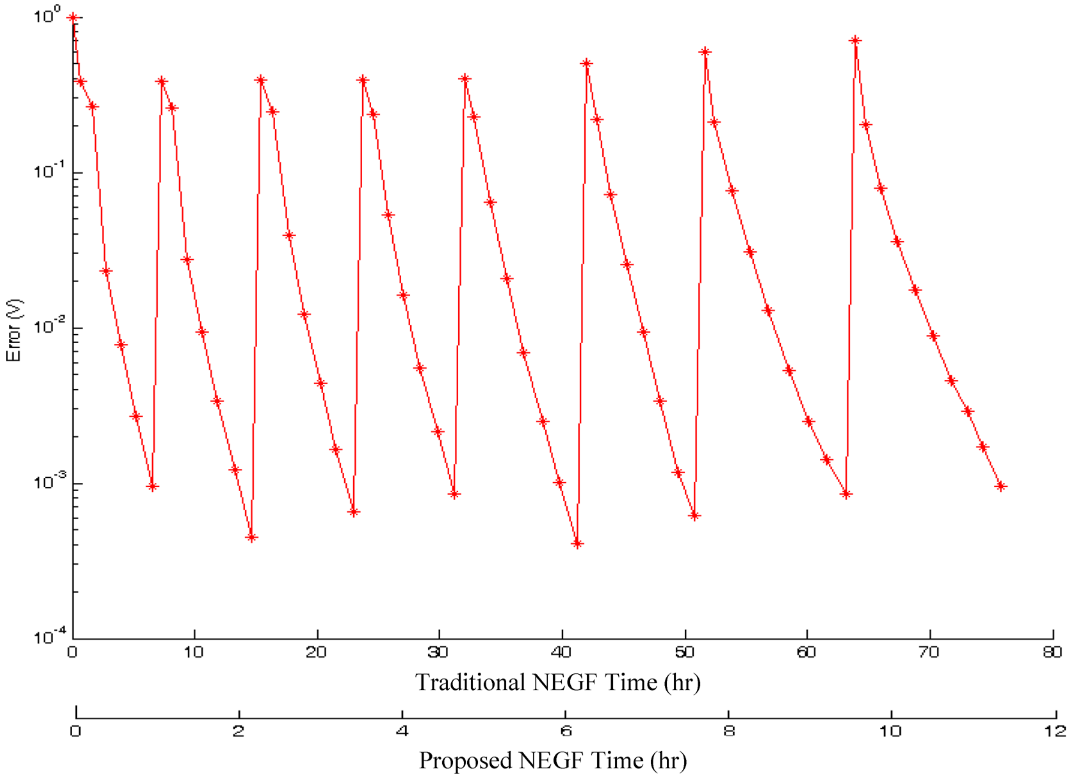


Figure 4. The self-consistent error versus time using both the traditional and the proposed NEGF methods, at V_D=V_{DD}

Table 1. The total simulation time, average simulation time per bias point and the average simulation time per iteration These simulations were carried out on a home PC: Intel® Pentium 4 CPU 2.4GHz, 768 MB RAM.

Method	t_{total} (hr)	t_{bias} (hr)	$t_{iteration}$ (hr)
Traditional NEGF	75.72	9.47	1.20
Proposed NEGF	11.65	1.46	0.18

5. Conclusion

The traditional NEGF for quantum transport of nanoscale devices was reviewed with emphasis on DG MOSFET. We proposed a method that is less computationally intensive than the traditional. Both the traditional and the proposed methods were implemented in FETMOSS and tested by simulation of a 5 *nm* DG MOSFET. The proposed method shows an exact match with the traditional method results. The simulation time, however, is reduced by about 85 %. The proposed method makes it practical to carry out quantum simulation on home PCs for device design and characteristics prediction. The method is generic and can be applied to simulate ballistic quantum transport in other nanoscale devices like FinFET and carbon nanotube FET.

References

1. International Technology Roadmap for Semiconductors 2007. Available at: http://www.itrs.net/Links/2007ITRS/2007_Chapters/2007_PIDS.pdf
2. V. Sverdlov, H. Kosina and S. Selberherr. "Modeling current transport in ultra-scaled field-effect transistors," *Journal of Microelectronics and Reliability – Elsevier*, Vol. 47, Issue no. 1, pp. 11–19, 2007.
3. Supriyo Datta, Electronic Transport in Mesoscopic Systems, Cambridge University Press, Cambridge, UK, 1995.

4. Z. Ren, R. Venugopal, S. Goasguen, S. Datta, and M. S. Lundstrom, "nanoMOS 2.5: A Two-dimensional simulator for quantum transport in Double-gate MOSFETs," *IEEE Trans. Electron Devices,* Vol. 50, Issue no. 9, pp. 1914–1925, Sep. 2003.
5. S. Datta, *Quantum Phenomena,* Addison-Wesley, 1989.
6. R. Pierret, *Advanced Semiconductor Fundamentals*, Prentice Hall, 2003.
7. S. Datta, "Nanoscale device modeling: the Green's function method," *Superlattices and Microstructures*, Vol. 28, Issue no. 4, pp. 253-278, 2000.
8. P. Damle "Nanoscale device modeling: From MOSFETs to molecules," Ph.D. Dissertation, Purdue University, West Lafayette, Indiana, May 2003.
9. T. Abdolkader, W. Farouk, O. Omar and M. Hassan. "FETMOSS: software tool for 2D simulation of double-gate MOSFET," *International Journal of Numerical Modeling*, Vol. 19, Issue no. 4, pp. 301-314, 2006.