ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Computational and Applied Mathematics 175 (2005) 87–99

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.com/locate/cam

A parallel adaptive finite volume method for nanoscale double-gate MOSFETs simulation

Yiming Li[a, b, *], Shao-Ming Yu[c]

[a]*Department of Computational Nanoelectronics, National Nano Device Laboratories, Hsinchu 300, Taiwan*
[b]*Microelectronics and Information Systems Research Center, National Chiao Tung University, Hsinchu 300, Taiwan*
[c]*Department of Computer and Information Science, National Chiao Tung University, Hsinchu 300, Taiwan*

Received 7 October 2003; received in revised form 20 March 2004

Abstract

We propose in this paper a quantum correction transport model for nanoscale double-gate metal-oxide-semiconductor field effect transistor (MOSFET) device simulation. Based on adaptive finite volume, parallel domain decomposition, monotone iterative, and a posteriori error estimation methods, the model is solved numerically on a PC-based Linux cluster with MPI libraries. Quantum mechanical effect plays an important role in semiconductor nanoscale device simulation. To model this effect, a physical-based quantum correction equation is derived and solved with the hydrodynamic transport model. Numerical calculation of the quantum correction transport model is implemented with the parallel adaptive finite volume method which has recently been proposed by us in deep-submicron semiconductor device simulation. A 20 nm double-gate MOSFET is simulated with the developed quantum transport model and computational technique. Compared with a classical transport model, it is found that this model can account for the quantum mechanical effects of the nanoscale double-gate MOSFET quantitatively. Various biasing conditions have been verified on the simulated device to demonstrate its accuracy. Furthermore, for the same tested problem, the parallel adaptive computation shows very good computational performance in terms of the mesh refinements, the parallel speedup, the load-balancing, and the efficiency.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Parallel algorithm; Domain decomposition; Adaptive computational method; Semiconductor device simulation; Quantum correction model; Nanoscale device; Double-gate MOSFETs

* Corresponding author. P.O. Box 25-178, Hsinchu 300, Taiwan. Tel.: + 886-930-330766; fax: +886-3-5726639.
E-mail address: ymli@mail.nctu.edu.tw (Y. Li).

0377-0427/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2004.06.001

1. Introduction

Development of nanoscale metal-oxide-semiconductor field effect transistors (MOSFETs) has recently been of great interest, in particular for the double-gate MOSFETs shown in Fig. 1 [3,4,15,16,18–20,24,25]. As these semiconductor devices are further scaled into the nanoscale regime (device channel length < 100 nm), it becomes extremely necessary to consider quantum mechanical effects when performing device modeling and simulation [1,3,9,15,16,18,19,22,24]. It is known that the growth of computational theory, method, and algorithms is an essential element in computational science and engineering. Parallel and adaptive computing methods for macroscopic semiconductor device models, such as Poisson, drift-diffusion (DD) and hydrodynamic (HD) equations are playing a crucial role in modern semiconductor device simulation [7,8,10,12,13]. Physically, the most accurate way of incorporation the quantum effect in the inversion layers is to solve the coupled Schrödinger–Poisson (SP) equations subject to an appropriated boundary condition at the interface of semiconductor and insulator [1,9,22], but it encounters numerical difficulties, such as convergence problem when coupling with DD or HD models, and is a time-consuming task in multi-dimensional nanodevice simulation [1,15,24]. There have been different approaches to modeling of quantum effects; one of them is adding quantum corrections to the classical transport models, such as Boltzmann, DD, and HD models [1,3,4,15,18,19,24].

Based on our recent works [9,14,15,24], in this paper, a new quantum correction model for nanodevice simulation is presented and solved with the parallel adaptive finite volume method on a PC-based Linux cluster with MPI libraries. Along the y direction, shown in Fig. 1, this model efficiently accounts for the quantum mechanical effects on the interface of gate oxide and silicon substrate (SiO_2/Si). Together with a two- or three-dimensional (2D or 3D) macroscopic transport model, it can be solved numerically by using the parallel adaptive finite volume method on triangular mesh. Numerical calculations are performed on a 20 nm double-gate n-type MOSFET with the developed quantum transport model and computational technique. Compared with a classical transport model, it is found that this model successfully characterizes the quantum effects of the nanoscale double-gate MOSFET quantitatively. Various biasing conditions have also been verified on the simulated device to demonstrate its accuracy. Furthermore, for the same tested

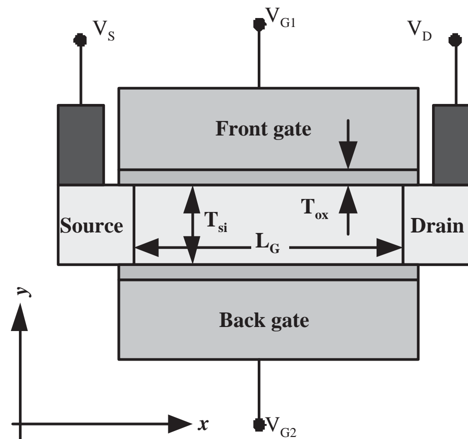

Fig. 1. A cross-section view of the simulated double-gate MOSFET device.

problem, the parallel adaptive computation shows very good computational performance in terms of the mesh refinements, the parallel speedup, the load-balancing, and the efficiency on our constructed 16-nodes PC cluster system. The mesh refinement mechanism precisely traces the errors of potential and carrier's density so that the quantum mechanical effects can be calculated accurately.

This article is organized as follows. Section 2 states the model to be solved. Section 3 describes the computing techniques. Section 4 shows the numerical results illustrating the preliminary model accuracy and efficiency of the method with different simulation cases. Section 5 draws the conclusions.

2. A quantum correction transport model

In classical transport models, there are more than three coupled partial differential equations (PDEs) have to be solved for a deep-submicron device simulation [2,8,12,13,21,23]. For example, the DD model includes: the Poisson equation, the electron current continuity equation, and hole current continuity equation. For the HD model, 5 PDEs have to be solved for the electrostatic potential, electron–hole densities, and electron–hole temperatures [21,23], respectively. Based on a phenomenological investigation from the SP solution [9,11,14,15,22,24] within the inversion regions of the SiO_2/Si interfaces, as shown in Fig. 1, the classical electron density is currently modeled to reflect the quantum confinement effect. Therefore, the DD and HD models can be directly applied to the nanoscale double-gate MOSFET device simulation without suffering numerical difficulties which is risen from the SP model. The proposed quantum correction equation for the quantum corrected inversion-layer charge densities n_{QM} is

$$\nabla \cdot (\varepsilon \nabla \phi) = -q(p - n_{CL} + N_D^+ - N_A^-) \tag{1}$$

and

$$n_{QM} = a_0 n_{CL} \left(1 - \exp\left(-a_1 \xi^2 \left(1 - \frac{1}{2}\left(\frac{\xi}{\xi_0}\right)^2\right) - a_2 \xi^3\right)\right). \tag{2}$$

All the symbols and physical quantities used here are followed [2,6–8,11–13,21,23], n_{CL} is the classical electron density solved from the Poisson equation (1). $\xi = y/\lambda_{th}$ and $\lambda_{th} = (\hbar^2/2m_0 k_B T)^{1/2}$ is the thermal wavelength, $\hbar$ is the reduced Planck constant, m_0 is the electron rest mass, k_B is the Boltzmann constant, T is the absolute temperature and $\xi_0 = T_{si}/2\lambda_{th}$ is for both the symmetric and asymmetric double-gate structures ("symmetric" means $V_{G1} = V_{G2}$ and "asymmetric" is $V_{G1} \neq V_{G2}$). The three model parameters a_0, a_1, and a_2 are functions of oxide thickness (T_{ox}), Si film thickness (T_{si}), and gate voltage (V_G) for both the SG and DG MOSFET structures. Together with the auxiliary equation (2), the conventional DD or HD model [2,6–8,10,12,13,21,23] forms a quantum correction transport model. The associated boundary condition for the model is the same with the DD or HD model depending on which model has been considered [2,6–8,10,12,13,21,23]. They not only can be solved numerically as usual but also can provide proper quantum correction to electron density quantitatively. This quantum correction approach for single-gate nanoscale MOSFET simulation has been used in our recent works [9,11,14,15,24].

3. Parallel adaptive computational method

To solve the 2D quantum correction DD and HD models for the nanoscale double-gate MOSFET, the parallel adaptive computing technique is applied. This simulation methodology has been successfully developed and applied to deep-submicron MOSFETs simulation in our recent works [7,8,10,12,13]. Based on the Gummel's decoupling algorithm and the finite volume discretization over the unstructured triangular mesh [6,21], the quantum corrected 2D DD (and HD) model is decoupled and approximated. Therefore, a corresponding system of nonlinear algebraic equations is obtained for each decoupled and discretized PDE. We solve the nonlinear system by means of the monotone iterative (MI) method [7] instead of the conventional Newton's iteration (NI) method [21] on our cluster system. The MI method is a constructive technique for the numerical solutions of PDEs. Compared with the NI method, application of the MI method to nanodevice simulation has some merits: (1) global convergence, (2) easier implementation, and (3) ready for parallelization [7,8,10,12,13]. By estimating the variations of the computed electrostatic potential and carrier's density, the error indicators and a global error estimator are calculated and tested for the convergence and mesh refinement [5,13,17,26]. The whole set of decoupled PDEs is solved self-consistently to obtain the convergent results. In parallelization, the dynamic domain decomposition approach is adopted and performed on our 16-nodes PC cluster system. The Linux cluster utilized for the simulation consists of 16 Pentium 1.7 GHz CPU with 512 MB memory and Intel 100 MBit fast Ethernet which are connected with 100 MBit 3Com fast Ethernet switch.

4. Results and discussion

By simulating a 20 nm double-gate MOSFETs, as shown in Fig. 1 , on our PC-based cluster system, several numerical testing examples are organized to demonstrate the model accuracy and achieved

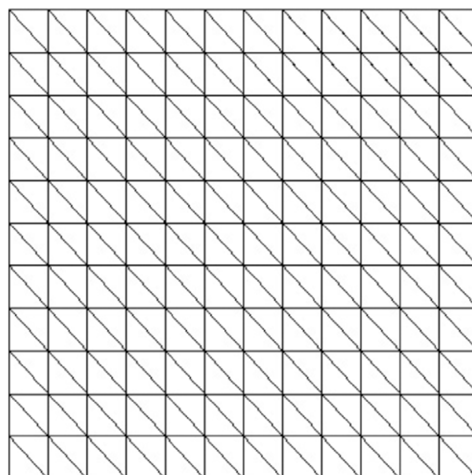

Fig. 2. Initial mesh. It contains 156 nodes.

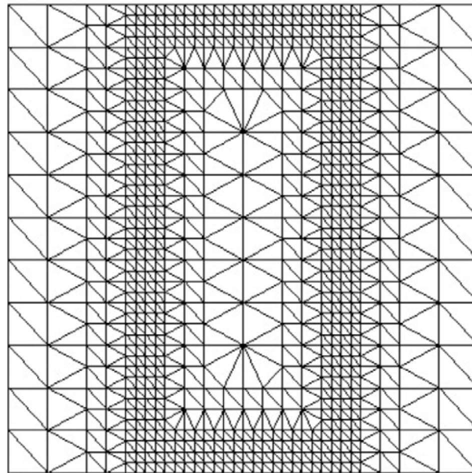

Fig. 3. The 3rd refined mesh. It has 794 nodes.

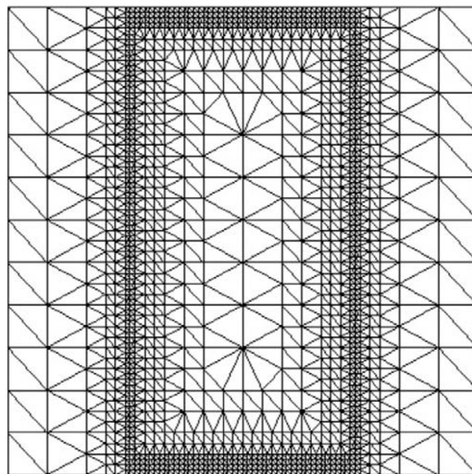

Fig. 4. There is 3104 nodes in the 5th refined mesh.

computational efficiency. Simulation considers the proposed 2D quantum correction HD model. The channel length of the device is 20 nm (total length is 40 nm), the thickness of Si film is 20 nm, and the oxide thickness on both sides is 2 nm. The doping profile on both the source and drain is 10^{20} cm^{-3} and the channel doping is with 10^{17} cm^{-3}.

First of all, the Figs. 2–5 shows the process of mesh refinements. The mechanism of mesh refinement is based on the estimation of solution error element by element. Starting from an initial triangular mesh, it is found that the mesh refinement adaptively focuses on both the top and bottom regions near the interfaces of SiO_2/Si. Therefore, the mesh is automatically concentrated on the inversion regions

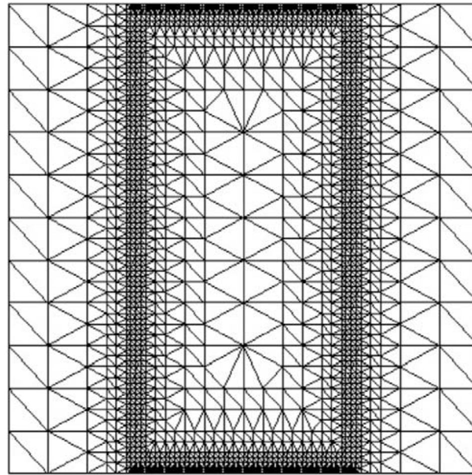

Fig. 5. The 7th refined mesh. It has 8437 nodes.

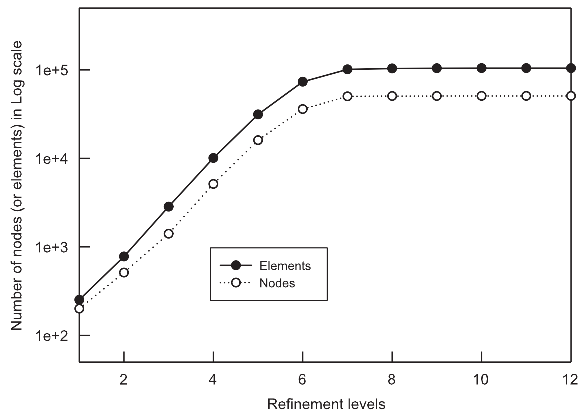

Fig. 6. The number of nodes (and elements) in Log scale versus the refinement levels.

to locate vary shape variation of carrier density. Fig. 6 reports the relationship of the number of nodes (and elements) versus the refinement levels. The increasing rate of the number of nodes (and elements) gradually becomes slow when the refinements increase. Therefore, it confirms the computational effectiveness of the adaptive computing method in the numerical simulation of the quantum correction transport model.

Figs. 7–14 show the computed potential and electron density for the device with different bias conditions, respectively. Figs. 7 and 8 are the computed potential of the device under symmetric bias conditions $V_{G1} = V_{G2} = 1.0$ V and different biases $V_{DS} = 0$ and 0.5 V, respectively. Ultra-thin regions on the both top and bottom sides are precisely calculated (red colors) with the quantum correction model. The adaptive refinement precisely locates the solution variation near the device surface efficiently. Similarly,

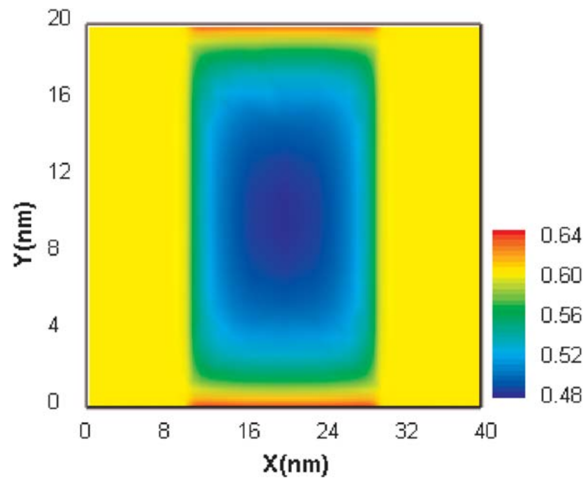

Fig. 7. Contour plots of the simulated potential of the 20 nm double-gate MOSFET with different biasing conditions: $V_{DS} = 0$ V and $V_{G1} = V_{G2} = 1.0$ V.

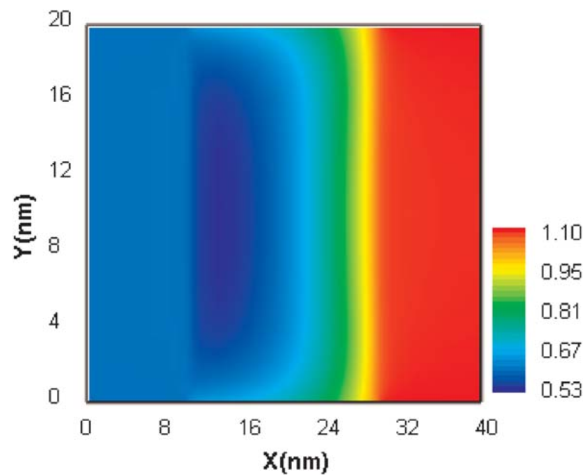

Fig. 8. Contour plots of the simulated potential of the 20 nm double-gate MOSFET with different biasing conditions: $V_{DS} = 0.5$ V and $V_{G1} = V_{G2} = 1.0$ V.

Figs. 9 and 10 are the results for the asymmetric cases. As shown in Figs. 11–14, the corresponding electron densities are reported with the same biasing conditions of the Figs. 7–10. The peak location of electron density has about 1 nm shift from both the top and bottom sides [1,9,14,15,18,19,22,24].

To verify the parallel performance of the domain decomposition method for the nanoscale double-gate MOSFETs simulation, the same device under $V_{G1} = V_{G2} = 1.0$ V and $V_{DS} = 0$ V is considered for the following cases. Table 1 reports the details of the achieved sequential and parallel time, efficiency, and

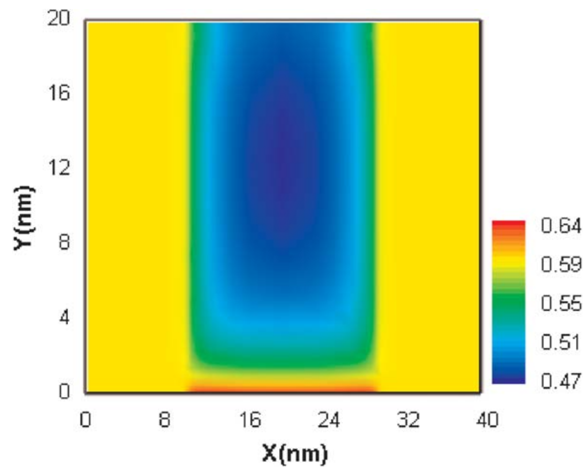

Fig. 9. Contour plots of the simulated potential of the 20 nm double-gate MOSFET with different biasing conditions: $V_{DS} = 0$ V, $V_{G1} = 0.5$ V, and $V_{G2} = 1.0$ V.

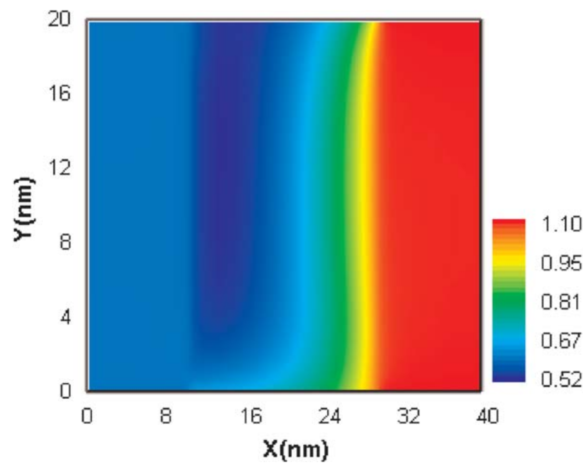

Fig. 10. Contour plots of the simulated potential of the 20 nm double-gate MOSFET with different biasing conditions: V_{DS}=0.5 V, $V_{G1} = 0.5$ V, and $V_{G2} = 1.0$ V.

speedup with respect to different number of nodes. It is performed on an 8-CPUs PC-based Linux cluster system. In our numerical experience, a 7.22 speedup factor is obtained on the tested 8-nodes system. Fig. 15 is the maximum difference versus the number of nodes. The maximum difference is defined as the maximum difference of the code execution time divided by the maximum execution time [13]. For the simulation with 2-,4-,8-, and 16-CPUs, the maximum difference decreases and tends to a stable value when the number of nodes increases. It shows a good dynamic load balancing for the domain decomposition. Fig. 16 is the achieved speedup and efficiency, where the speedup is the ratio of the code

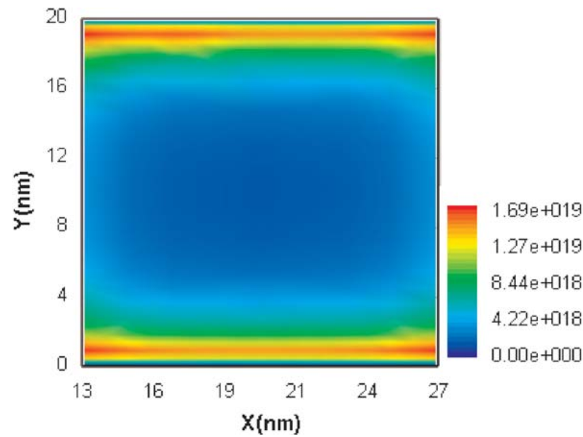

Fig. 11. Contour plots of the simulated electron density of the 20 nm double-gate MOSFET under both gates with different biasing conditions: $V_{DS} = 0$ V and $V_{G1} = V_{G2} = 1.0$ V.

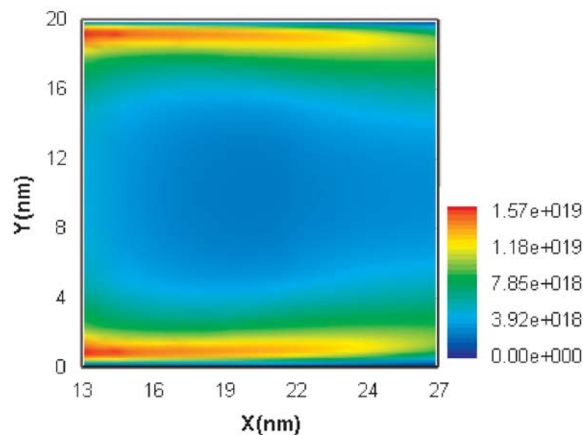

Fig. 12. Contour plots of the simulated electron density of the 20 nm double-gate MOSFET under both gates with different biasing conditions: $V_{DS} = 0.5$ V and $V_{G1} = V_{G2} = 1.0$ V.

execution time on a single processor to that on multiple processors. Efficiency is defined as the speedup divided by the number of processors [13]. The speedup is about 12.2 for the simulation running on a 16-CPUs system and 75% efficiency is maintained. In Fig. 17, we indicate the current–voltage $(I–V)$ characteristic difference by comparing the calculated current of the 20 nm double-gate MOSFET with and without the quantum correction transport model. This comparison shows that the classical mode overestimates the calculation of $I–V$ curves and demonstrate more than 30% difference in comparing with the result of the quantum correction model [3,19].

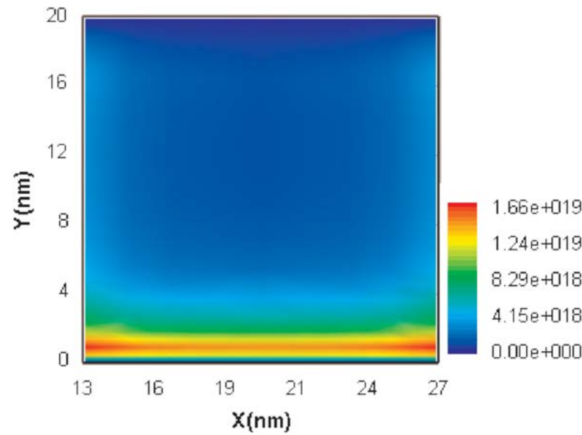

Fig. 13. Contour plots of the simulated electron density of the 20 nm double-gate MOSFET under both gates with different biasing conditions: $V_{DS} = 0$ V, $V_{G1} = 0.5$ V, and $V_{G2} = 1.0$ V.

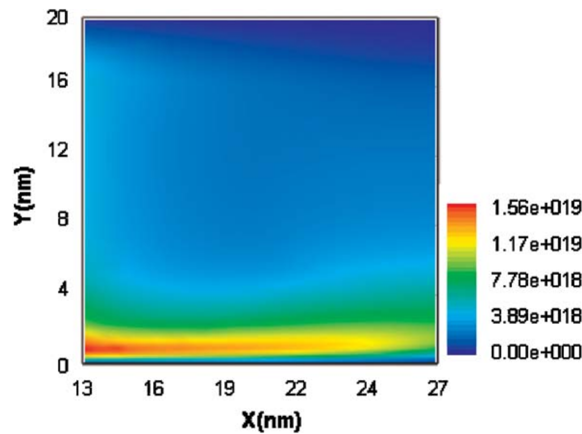

Fig. 14. Contour plots of the simulated electron density of the 20 nm double-gate MOSFET under both gates with different biasing conditions: $V_{DS} = 0.5$ V, $V_{G1} = 0.5$ V, and $V_{G2} = 1.0$ V.

5. Conclusions

We have demonstrated in this paper a quantum correction transport model for nanoscale double-gate MOSFET simulation. Based on an the adaptive finite volume, the parallel domain decomposition, the monotone iterative, and a posteriori error estimation methods, the model has been solved numerically on a PC-based Linux cluster with MPI libraries. Simulation of a 20 nm double-gate MOSFET has shown the accuracy of the developed model and computational efficiency successfully. Compared with a classical transport model, it is found that this model has accounted for the quantum effects of the nanoscale

Table 1
A list of the achieved sequential and parallel time, efficiency, and speedup with respect to different number of nodes

Nodes	Sequential time (S)	Parallel time (S) of the 8-processors system	Speedup	Efficiency (%)
1000	9.4	4.1	2.29	28.66
4000	87	26.1	3.30	41.67
8000	367	83.6	4.39	54.87
16 000	794	179.2	4.43	55.39
32 000	4041	735.6	5.49	68.67
64 000	9403	1993.6	5.30	66.27
90 000	27 775	4487	6.19	77.38
250 000	151 016	20 906	7.22	90.29

Performed on an 8-processors PC-based Linux cluster system.

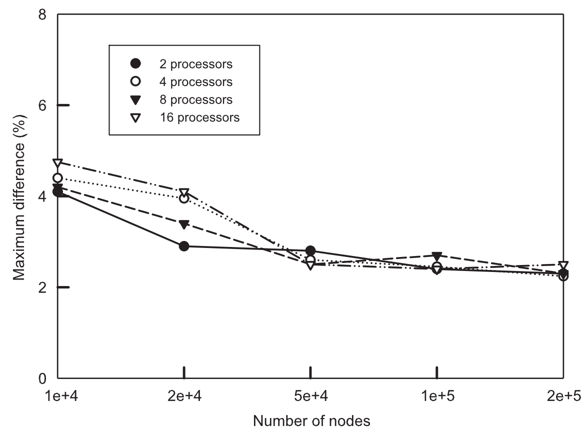

Fig. 15. The maximum difference versus the number of nodes.

double-gate MOSFET quantitatively. Various biasing conditions have been verified on the simulated device to demonstrate its accuracy. Furthermore, for the same tested problem, the parallel adaptive computation reported very good computational performance in terms of the mesh refinements, the parallel speedup, the load-balancing, and the efficiency.

Acknowledgements

This work is supported in part by the National Science Council (NSC) of TAIWAN under contract numbers: NSC-92-2112-M-429-001 and NSC-93-2752-E-009-002-PAE. It is supported in part by the grant of the Ministry of Economic Affairs, Taiwan under contract No. 92-EC-17-A-07-S1-0011 and the 2004 research grant of the Taiwan Semiconductor Manufacturing Company, Hsinchu, Taiwan.

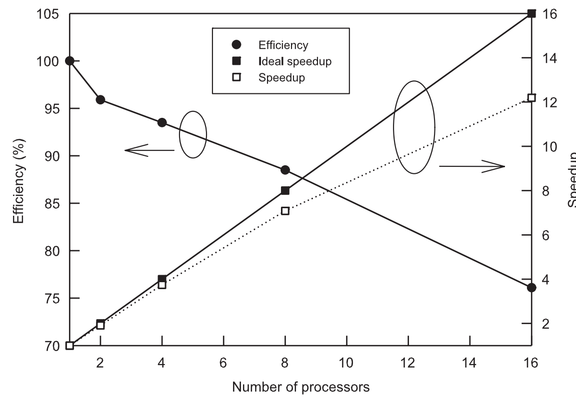

Fig. 16. The parallel speedup and efficiency versus the number of processors.

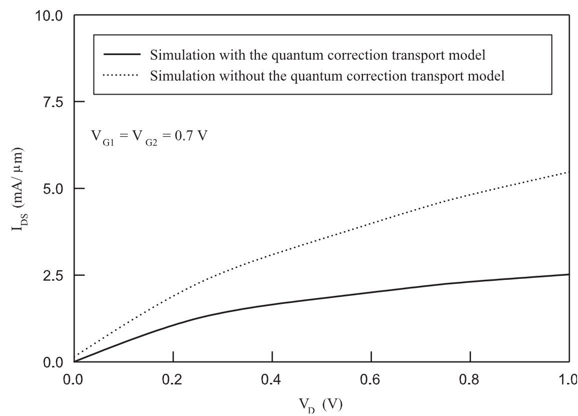

Fig. 17. A comparison of the calculated current of the 20 nm double-gate MOSFET with and without the quantum correction transport model.

References

[1] M.G. Ancona, H.F. Tiersten, Macroscopic physics of the silicon inversion layer, Phys. Rev. B 35 (1987) 7959–7965.

[2] K. Blotekjaer, Transport equations for electrons in two-valley semiconductors, IEEE Trans. Electron Dev. 17 (1970) 38–47.

[3] Y.-K. Choi, D. Ha, T.-J. King, J. Bokor, Investigation of gate-induced drain leakage (GIDL) current in thin body devices: single-gate ultra-thin body, symmetrical double-gate, and asymmetrical double-gate MOSFETs, Jpn. J. Appl. Phys. part I 42 (2003) 2073–2076.

[4] J.G. Fossum, L. Ge, M.H. Chiang, Speed superiority of scaled double-gate CMOS, IEEE Trans. Electron Dev. 49 (2002) 808–811.

[5] T. Gallouet, R. Herbin, M.H. Vignal, Error estimates on the approximate finite volume solution of convection diffusion equations with general boundary conditions, SIAM J. Numer. Anal. 37 (2000) 1935–1972.

[6] J.W. Jerome, Analysis of Charge Transport: A Mathematical Study of Semiconductor Devices, Springer, New York, 1996.

[7] Y. Li, A parallel monotone iterative method for the numerical solution of multidimensional semiconductor Poisson equation, Comput. Phys. Comm. 153 (2003) 359–372.
[8] Y. Li, C.-S. Wang, Numerical solution of hydrodynamic semiconductor device equations employing a stabilized adaptive computational technique, WSEAS Trans. Systems 1 (2002) 216–221.
[9] Y. Li, T.-S. Chao, S.M. Sze, A novel parallel approach for quantum effect simulation in semiconductor devices, Internat. J. Model. Simulation 23 (2003) 94–102.
[10] Y. Li, H.-M. Lu, T.-w. Tang, S.M. Sze, A novel parallel adaptive Monte Carlo method for nonlinear Poisson equation in semiconductor devices, Math. Comput. Simulation 62 (2003) 413–420.
[11] Y. Li, T.-w. Tang, S.-M. Yu, A quantum correction model for nanoscale double-gate MOS devices under inversion conditions, J. Comput. Electron. 2 (2003) 491–495.
[12] Y. Li, J.-L. Liu, T.-S. Chao, S.M. Sze, A new parallel adaptive finite volume method for the numerical simulation of semiconductor devices, Comput. Phys. Comm. 142 (2001) 285–289.
[13] Y. Li, S.M. Sze, T.-S. Chao, A practical implementation of parallel dynamic load balancing for adaptive computing in VLSI device simulation, Eng. Comput. 18 (2002) 124–137.
[14] Y. Li, S.-M. Yu, A unified quantum correction model for nanoscale single- and double-gate MOSFETs under inversion conditions, Nanotechnology 15 (2004) 1009–1016.
[15] Y. Li, T.-w. Tang, X. Wang, Modeling of quantum effects for ultrathin oxide MOS structures with an effective potential, IEEE Trans. Nanotechnol. 1 (2002) 238–242.
[16] G. Pei, J. Kedzierski, P. Oldiges, M. Ieong, E.C.-C. Kan, FinFET design considerations based on 3-D simulation and analytical modeling, IEEE Trans. Electron Dev. 49 (2002) 1411–1419.
[17] R. Ramakrishnan, Structured and unstructured grid adaptation schemes for numerical modeling of field problems, Appl. Numer. Math. 14 (1994) 285–310.
[18] S.M. Ramey, D.K. Ferry, Modeling of quantum effects in ultrasmall FD-SOI MOSFETs with effective potentials and three-dimensional Monte Carlo, Phys. B 314 (2002) 350–353.
[19] N. Sano, A. Hiroki, K. Matsuzawa, Device modeling and simulations toward sub-10 nm semiconductor devices, IEEE Trans. Nanotechnol. 1 (2002) 63–71.
[20] T. Schulz, W. Rosner, E. Landgraf, L. Risch, U. Langmann, Planar and vertical double gate concepts, Solid-State Electron. 46 (2002) 985–989.
[21] S. Selberherr, Analysis and Simulation of Semiconductor Devices, Springer, New York, 1984.
[22] F. Stern, W.E. Howard, Properties of semiconductor surface inversion layers in the electric quantum limit, Phys. Rev. 163 (1967) 816–835.
[23] S.M. Sze, Physics of Semiconductor Devices, 2nd Edition, Wiley, New York, 1981.
[24] T.-w. Tang, Yiming Li, A SPICE-Compatible model for nanoscale MOSFET capacitor simulation under the inversion condition, IEEE Trans. Nanotechnol. 1 (2002) 243–246.
[25] J. Walczak, B. Majkusiak, The remote roughness mobility resulting from the ultrathin SiO_2 thickness nonuniformity in the DG SOI and bulk MOS transistors, Microelectron. Eng. 59 (2001) 417–421.
[26] X.D. Zhang, J.-Y. Trepanier, R. Camarero, A posteriori error estimation for finite-volume solutions of hyperbolic conservation laws, Comput. Methods Appl. Mech. Eng. 185 (2000) 1–19.