A Study on Computational Efficiency for Parallel Semiconductor Device Simulation

Jyun-Hwei Tsai[1], Shih-Ching Lo[1] and Yiming Li[2,3]
[1]National Center for High-Performance Computing, Hsinchu 300, Taiwan
[2]Department of Nano Device Technology, National Nano Device Laboratories, Hsinchu 300, Taiwan
[3]Microelectronics & Information Systems Research Center, National Chiao Tung Univ., Hsinchu 300, Taiwan

Abstract: - In this paper, the Drift-Diffusion (DD), Schrödinger –Poisson transport (SP) and Density Gradient transport models (DG) are computed by parallel direct method and three different meshes. According to the results, the DD and DG model cannot be good approximation of SP model with respect to electron density simulation. In addition, a dense mesh is necessary for simulation of quantum effect. Therefore, parallel computing is an important technique of semiconductor devices. Generally, simulation with two and four processors is about 1.6 ~ 1.8 and 2.8 ~ 3 times faster than that with one processor, respectively. In the case of efficiency, 0.8 ~ 0.9 and 0.7 ~ 0.75 are obtained for two and four processors, respectively.

Key-Words: - Quantum effects, Drift-Diffusion model, Schrödinger equation, Density Gradient model, Numerical simulation, Parallel computing.

1 Introduction

As the progress of semiconductor fabrication technology for the advanced metal oxide semiconductor field effect transistor (MOSFET) has been of great interests in recent years, computer-aided simulation for semiconductors, which provides a software driven approach to explore new physics and device also acquires a crucial role in the development of semiconductor. Semiconductor device models, such as Drift-Diffusion (DD) and Hydrodynamic (HD) models [1-8], is the classical device simulation. In order to understand the characteristics of nanoscale devices, it is important to take quantum mechanical effects into account with the classical models. In principle, the Schröedinger equation coupled with DD model (SP model) is the most accurate way to solve the carrier concentration, but it is not suitable for engineering applicationsespecially for the two- and three dimensional cases. This is not only because it is computationally expensive but also because it is difficult to model the multi-dimensional cases. Therefore, quantum correction models, which produce a similar results to quantum mechanically calculated one but requires only about the same computation cost as that of the classical calculation, are developed. Among the quantum correction models, such as Hänsch model [9], van Dort model [10], Effective-Potential (EP) approach [11-12], Density Gradient (DG) method [13-16] and modified local density approximation (MLDA) [17], DG model is considered a good approximation of the quantum effect. However, numerical results of DG and SP model are still different. The difference between them is more significant as the size of device is smaller. That is, solving SP model is still necessary.

Fortunately, the dilemma of time consuming or rough approximation can be overcame by advanced computing technique, such as parallel computing and adaptive computation. Parallel numerical simulation of semiconductor devices has been proven to be an indispensable tool for fast characterization and optimal design of semiconductor devices [8, 18 ~ 22]. In this paper, we employ a parallel direct solving method to simulate a 90 nm metal oxide semiconductor field effect transistor and compare the results of DD, DG and SP models. To show the difference between classical and quantum models, electron density distribution and drain current are discussed. The remaining content of this study is given as follows. Sec. 2 briefly explains the simulation models and the computational method. Sec. 3 shows the simulation results and discussion. Sec. 4 draws the conclusion.

2 Device Models and Algorithm

The DD, SP and DG models are employed to simulate and compare the difference of drain current and electron concentration by 1, 2 and 4 processors.

The models and computing algorithm are described as the following subsections.

2.1 Classical and Quantum Models

The three governing equations of DD model are listed as follows. The Poisson equation is

$$\nabla\varepsilon\cdot\nabla\psi=-q\left(p-n+N_D-N_A\right), \quad (1)$$

where ε is the electrical permittivity, q is the elementary electronic charge, n and p are the electron and hole densities, and N_D and N_A are the number of ionized donors and acceptors, respectively. The current densities are given by:

$$\mathbf{J}_n=-qn\mu_n\nabla\phi_n, \quad (2)$$

$$\mathbf{J}_p=-qp\mu_p\nabla\phi_p, \quad (3)$$

where $\mathbf{J}_n$ and $\mathbf{J}_p$ are the electron and hole current density satisfying the continous equations, μ_n and μ_p are the electron and hole mobility, and ϕ_n and ϕ_p are the electron and hole quasi-Fermi potentials, respectively. The mobility model used herein is Masetti's model. The continuity equations are as follows:

$$q\frac{\partial n}{\partial t}-\nabla\cdot\mathbf{J_n}=-qR, \quad (4)$$

$$q\frac{\partial p}{\partial t}+\nabla\cdot\mathbf{J_p}=-qR. \quad (5)$$

With continued scaling into the deep sub-micron regime, neither internal nor external characteristics of stateof-the-art semiconductor devices can be described properly using the conventional DD model.

To include quantization effects in a classical device simulation, a simple approach is to introduce an additional potential, such as quantity Λ, in the classical density formula, which reads:

$$n=N_C\exp\left(\frac{E_F-E_C-\Lambda}{k_BT}\right), \quad (6)$$

where N_C is the conduction band density of states, E_C is the conduction band energy, and E_F is the electron Fermi energy. It is not possible to describe all quantum mechanical effects in terms of a variable Λ. For the SP-DD model, we include the quantization effects in the classical DD model by considering a Schrödinger equation along the semiconductor substrate (z - direction)

$$\left(-\frac{\partial}{\partial z}\frac{\hbar^2}{2m_{z,v}(z)}\frac{\partial}{\partial z}+E_C(z)\right)\Psi_{j,v}(z)=E_{j,v}\Psi_{j,v}(z). \quad (7)$$

Together with the 2DEG formula [18] the Eq. (7) is introduced to the self-consistent DD model. $\hbar$ is the reduced Planck constant, E_C is the conduction band energy, v is the band valley, $m_{z,v}$, is the effective mass for valley in quantization direction, $\Psi_{j,v}$ is the j-th normalized eigenfunction in valley v; and , $E_{j,v}$ is the j-th eigenenergy. Solving the equations above, the device current can be directly computed. T denotes the carrier temperature, k denotes the Boltzmann constant, N_C is the conduction band density of states, E_C is the conduction band energy, and E_F is the electron Fermi energy. It is not possible to describe all quantum mechanical effects in terms of a variable Λ. Therefore, several quantum correction formalisms are suggested to approximate the effects. Among the models, the density gradient model gives a good approximation [13-16]. It gives a reasonable description of terminal characteristics and charge distribution inside the device. In addition, the adjustable parameter of density gradient approximation is nearly a constant for a wide range of applied gate voltage. Therefore, density gradient model is chosen as the quantum correction formalism in this study. For the density gradient model, Λ is given by

$$\Lambda=-\frac{\gamma\hbar^2}{12m}\frac{\nabla^2\sqrt{n}}{\sqrt{n}}, \quad (8)$$

where $\hbar$ is the reduced Planck constant, m is the density of states mass, and γ is a fit factor. The computing flowchart is illustrated as Fig. 1. Firstly, the stop criteria, mesh, output variables and simulation models are chosen. If density gradient model is chosen, the modified potential is added in Poisson equation. Then, Poisson equation is solved iteratively until the result converges. If Schrödinger equation is chosen, solving it until it converges is the next step. Otherwise, continuity equations are solved. After all equations converge, we'll check the whole system converges or not. If the whole system converges, then stop computing. Otherwise, the outer loop should be iterated again until the whole converges. This scheme makes sure the solution will self-consistent.

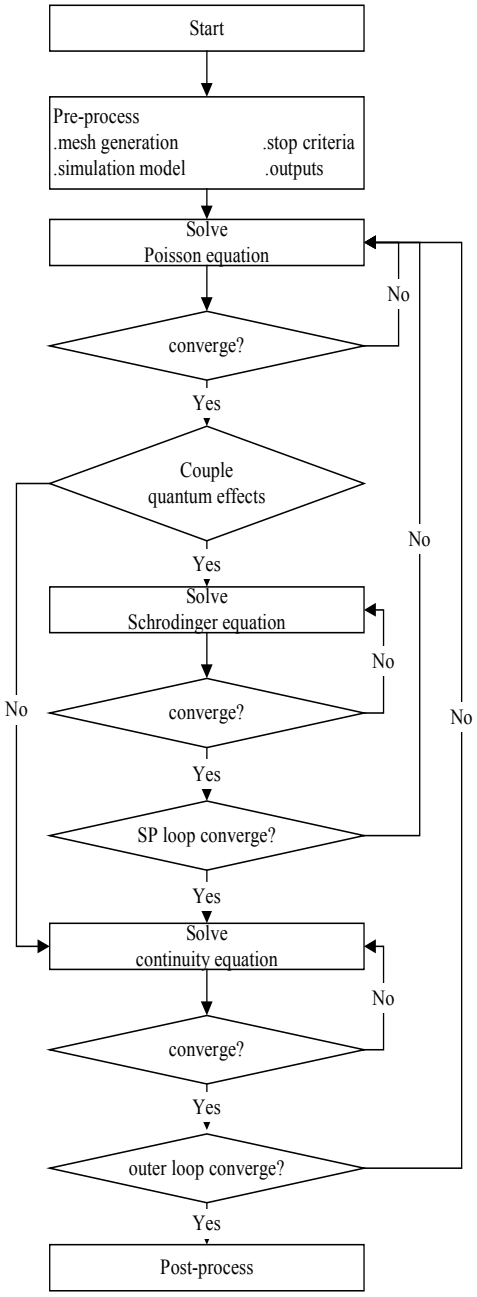

Fig. 1. Flowchart of a self-consistent simulation scheme

2.2 Parallel Algorithm

Parallel algorithm employed in this work is base on the parallel direct method of linear system. If a matrix **A** is factorized as **A** = **LU**. Direct solvers for sparse matrices involve much more complicated algorithms than for dense matrices. The main complication is due to the need for efficiently handling the fill-in in the factors **L** and **U**. The solving strategy of the fill-in reduction is integrated by multilevel recursive [19] or minimum-degree based approaches [20]. The numerical factorization algorithm utilizes the supernode structure of the numerical factors **L** and **U** to reduce the number of memory references. The result is a greatly increased sequential factorization performance. Furthermore, a left-right looking super node algorithm [21-22] for the parallel sparse numerical factorization on shared-memory multiprocessors is used.

3 Results and Discussion

In the numerical studies, a 90 nm MOSFETs is simulated. Oxide thickness is 2 nm. A super steep retrograde with S/D halo doping profile is given. Figure 2 illustrates the structure of the simulated device. DD, DG and SP models are simulated and discussed. Numerical results of MOSFETs are simulated by ISE-DESSIS ver. 8.0.3 [23] on HP 4000 workstation with 4 processors.

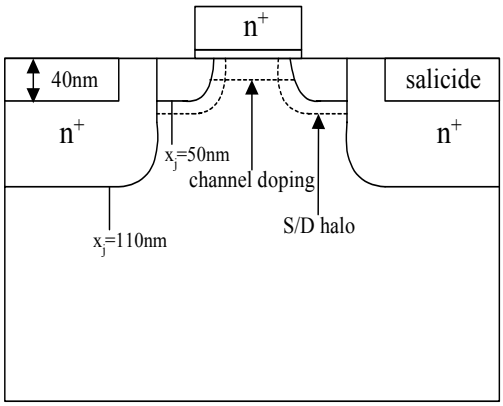

Fig. 2. A 90 nm NMOSFET with super steep retrograde channel doping and S/D halo

Figures 3 shows that drain current (I_{DS}) and electron density distribution simulated by classical model are different to that simulated by quantum model. Classical model overestimates a 30% drain current than quantum model. The simulated electron density distribution of both model are in different shape.

Figures 4 ~ 5 compares the numerical results of classical and quantum models with different meshes. Three meshes, which are 678, 2602 and 5409 vertices, are compared. The refinement of mesh is focused on 5 nm below Si/SiO_2 interface. Minimal lengths of mesh are 0.05, 0.02 and 0.01 nm, respectively. It is found that mesh size affects the result of quantum model, but does not affect the result of classical model. Since quantum effect can not be neglected, a dense mesh is needed when a deep sub-micron semiconductor device is simulated. However, the

computing time of SP model is much longer than that of DD and DG. Also, a simulation with dense mesh takes longer computing time than a sparse one and computing time increases rapidly when the model is complex. Figure 6 gives the computing time of different models with different meshes.

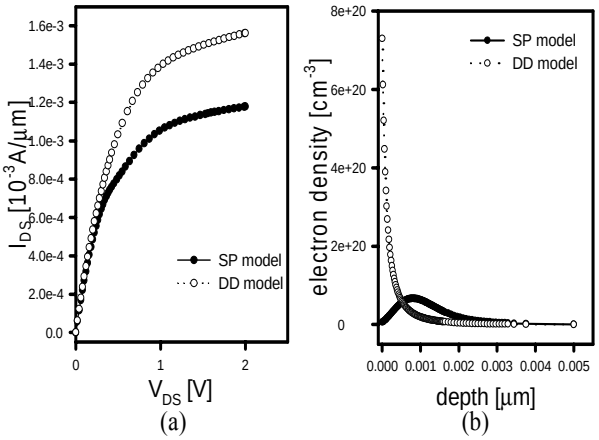

Fig. 3. Comparison of classical and quantum models by (a) drain current and (b) electron density distribution.

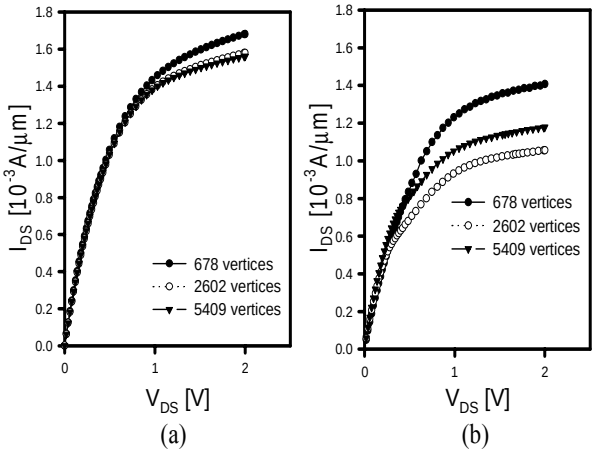

Fig. 4. Comparison of I_{DS}-V_{DS} curves by different meshes with (a) DD model and (b) SP model.

As mentioned above, simulation of deep sub-micron semiconductor device takes plenty of time. Parallel computing technique is employed to solve the difficulty. DD, DG and SP models are simulated with the three meshes. Two measures are used to discuss the performance. The first one is speedup, which is defined as

$$S(p)=\frac{T_1}{T(p)}, \tag{9}$$

where p is number of processors, T_1 is the runtime of the serial solution and $T(p)$ is the runtime of the parallel solution with p processors. The second one is efficiency, which is defined as

$$E(p)=\frac{S(p)}{p}=\frac{T_1}{pT(p)} \tag{10}$$

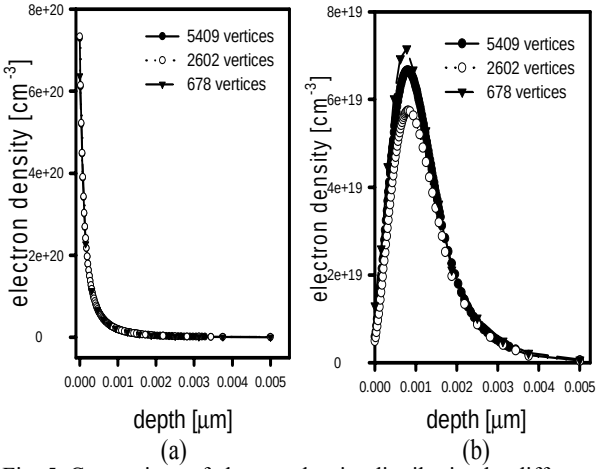

Fig. 5. Comparison of electron density distribution by different meshes with (a) DD model and (b) SP model.

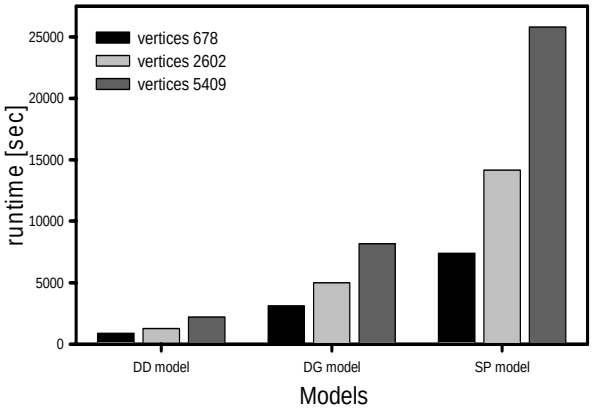

Fig. 6. Computing time of DD, DG and SP models with different meshes.

Speedup and efficiency are illustrated in Figs. 7 ~ 9. In the classical case, simulation with two and four processors are about 1.75 ~ 1.85 and 2.8 ~ 2.9 times faster than that with one processor, respectively. In DG case, simulation with two and four processors are about 1.6 ~ 1.7 and 2.5 ~ 2.7 times faster than that with one processor, respectively. In SP case, simulation with two and four processors are about 1.7 ~ 1.8 and 2.7 ~ 3 times faster than that with one processor, respectively. In the three cases, efficiency are not as good as expectation. Generally, 0.85 and 0.7 are obtained for two and four processors, respectively.

According to the numerical results mentioned above, we can make summary of the results.

Classical model takes lest runtime, but the simulated results, such as drain current and electron density distribution, are different to quantum models. Runtime of DG model is shorter than that of SP model. However, the results also have difference, especially the electron density distribution. In classical case, the accuracy of numerical result does not sensitive to the simulated mesh. Nevertheless, results of quantum models depend on the quality of mesh. To obtain a correct result, a dense mesh, which takes lots of runtime, is necessary. Therefore, parallel computing is considered as a solution to overcome the difficulty. The parallel algorithm employed in this study focus on solving a sparse matrix. Parallel algorithm speeds up the computation. Waiting time for data communication causes the lost of efficiency. It is the shortcoming of point parallelization, which solves a point paralleled. If a much denser mesh is used, the efficiency will become better. If a parallel algorithm is designed by line parallization, which solves a family curves paralleled with data communication, a better efficiency will be obtained.

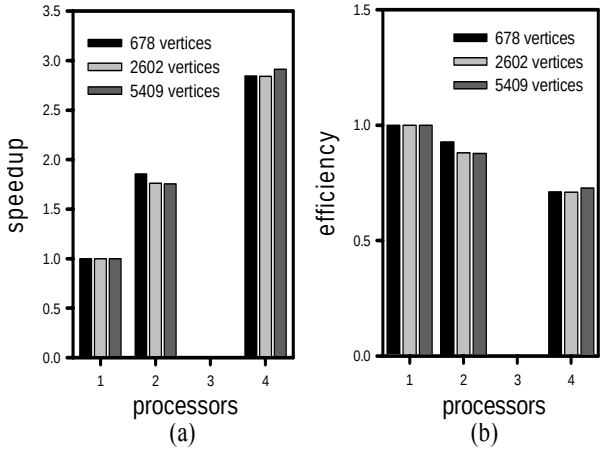

Fig. 7. (a) Speedup and (b) efficiency of simulating DD model by different number of processors and meshes .

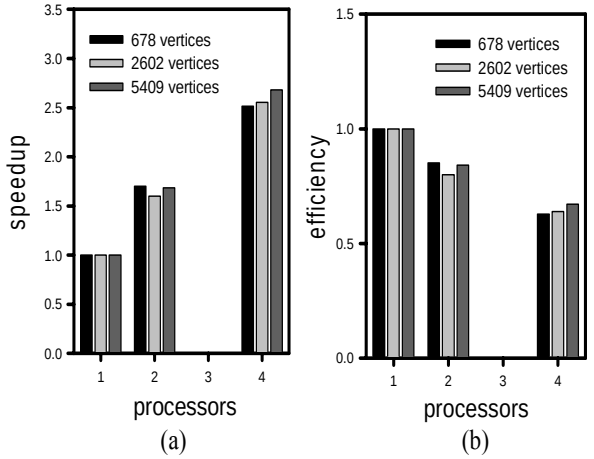

Fig. 8. (a) Speedup and (b) efficiency of simulating DG model by different number of processors and meshes .

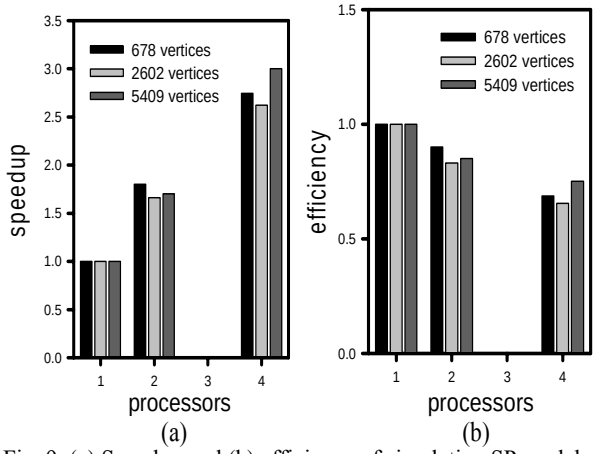

Fig. 9. (a) Speedup and (b) efficiency of simulating SP model by different number of processors and meshes .

4 Conclusions

In this paper, a 90 nm MOSFET is simulated by DD, DG and SP models with parallel algorithm under three different meshes. According to the numerical results, quantum effect should be simulated by a dense mesh, which takes lots of runtime. The parallel direct method is employed to speed up the simulation. Parallel computing accelerates the simulation. The efficiency is not as good as expectation because the mesh is not dense enough and the algorithm involves data communication. To obtain a better result, a parallel technique for solving family curves without data communication should be employed.

5 Acknowledgement

The work was partially supported by the National Science Council, Taiwan, R.O.C., under Contracts NSC-92-2215-E-429-010, NSC-92-2112-M-429-001, and NSC-92-2815-C-492-001-E. It was also partially supported by Ministry of Economic Affairs, R.O.C. under contract No. 91-EC-17-A-07-S1-0011.

References:

[1] H. K. Gummel, “A Self-Consistent Iterative Scheme for One-Dimensional Steady State Transistor Calculations,” *IEEE Transactions on Electron Devices*, Vol. 11, pp.455, 1964.

[2] R. Stratton, "Diffusion of hot and cold electrons in semiconductor barriers," *Physics Review*, Vol. 126, No. 6, pp. 2002–2014, 1962.

[3] K. Bløtekjær, “Transport Equations for Electrons in Two-Valley Semiconductors,” *IEEE Transactions on Electron Devices*, Vol. ED-17, No. 1, pp.38-47, 1970.

[4] T. W. Tang, "Extension of the Scharfetter-Gummel Algorithm to the Energy Balance Equation," *IEEE Transactions on Eletronic Devices*, Vol. 31, No. 12, pp.1912-1914, 1984.

[5] D. Chen, et al., "Dual energy transport model with coupled latticeand carrier temperatures," *SISDEP-5*, (Vienna), pp. 157–160, Sept. 1993.

[6] Y. Li, "A Novel Approach to Carrier Temperature Calculation for Semiconductor Device Simulation using Monotone Iterative Method. Part I: Numerical Results," *Proceeding of 3rd WSEAS Symposium on Mathematical Methods Computing Technology Electrical Engineering* (MMACTEE 2001), Athens, Dec. 2001, pp. 5721-5726.

[7] Y. Li, "A Novel Approach to Carrier Temperature Calculation for Semiconductor Device Simulation using Monotone Iterative Method. Part I: Numerical Algorithm," *Proceeding of 3rd WSEAS Symposium on Mathematical Methods Computing Technology Electrical Engineering* (MMACTEE 2001), Athens, Dec. 2001, pp. 5671-5676.

[8] Y. Li, S. M. Sze and T. S., Chao, "A Practical Implementation of Parallel Dynamic Load Balanceing for Adaptive Computing in VLSI Device Simulation," *Engineering with Computers*, Vol. 18, pp.124-137, 2002.

[9] W. Hänsch, et al., "Carrier Transport Near the Si/SiO_2 Interface of a MOSFET," *Solid-State Electronics*, Vol. 32, No. 10, pp.839-849, 1989.

[10] M. J. van Dort, P. Woerlee and A. J. Walker, "A Simple Model for Quantization Effects in Heavy-doped Silicon MOSFETs at Inversion Conditions," *Solid-State Electronics*, Vol. 37, No. 3, pp.411-414, 1994.

[11] L. Shifren, R. Akis and D. K. Ferry, "Correspondence between Quantum and Classical Motion: Comparing Bohmian Mechanics with a Smoothed Effective Potential Approach," *Physics Letters A*, Vol. 274, pp.75-83, 2000.

[12] Y. Li, T. W. Tang and X. Wang, "Modeling of Quantum Effects for Ultrathin Oxide MOS Structures with an Effective Potential," *IEEE Trans. Nanotech.*, Vol. 1, No. 4, pp.238-242 2002.

[13] M. G. Ancona and H. F. Tierstein, "Macroscopic Physics of the Silicon Inversion Layer," *Physical Review B*, Vol. 35, No. 15, pp.7959-7965, 1987.

[14] M. G. Ancona., "Nonlinear Discretization Scheme for the Density-Gradient Equations," *SISPAD'00,* pp.196-199, 2000.

[15] T. W. Tang, X. Wang and Y. Li, "Discretization Scheme for the Density-Gradient Equations and Effect of Boundary Conditions," *J. Comp. Elec*., Vol.1, No. 3, pp. 389-393, 2002.

[16] X. Wang, Quantum Correction to the Charge Density Distribution in Inversion Layers, MS. Thesis, University of Massachusetts, U. S. A., 2001.

[17] G. Paasch and H. Ubensee, "A Modified Local Density Approximation," *Physics Stat Sol (b)*, Vol. 113, pp.165-178, 1982.

[18] Y. Li, et al., "A Novel Parallel Approach for Quantum Effect Simulation in Semiconductor Devices," *International Journal of Modelling and Simulation*, Vol. 23, No. 2, pp.1-8, 2003.

[19] G. Karypis and V. Kumar, "Analysis of Multilevel Graph Algorithms," Technical Report MN 95-037, Uni. Of Min., Dept. Comp. Sci. , Minneapolis, MN 55455, 1995.

[20] J. W. H. Liu, "Modification of the Minimum-Degree Algorithm by Multiple Elimination," *ACM Trans. Math. Software*, Vol. 11, No. 2, pp. 141-153, 1985.

[21] O. Schenk, K. Garter and W. Fichtner, "Efficient Sparse LU Factorizaytion with Left-right Looking Strategy on Shared Multiprocessors," *BIT*, Vol. 40, No. 1, pp.158-176, 2000.

[22] O. Schenk, K. Garter and W. Fichtner, "Efficient Sparse LU Factorizaytion with Left-right Looking Strategy on Shared Multiprocessors," *BIT*, Vol. 40, No. 1, pp.158-176, 2000.

[23] DESIS-ISE TCAD Release 8.0, ISE integrated Systems Engineering AG, Switzerland, 2002.