Accelerated 3D Full Band Self-consistent Ensemble Monte Carlo Device Simulation Utilizing Intel MIC co-processors on TianHe II

Longxiang Yin[1], Minquan Fang[2], Lang Zeng[3 4], LiLun Zhang[2], Gang Du[1], Xiaoyan Liu[1*]
[1]Institute of Microelectronics, Peking University, Beijing, China
[2]School of Computer, National University of Defence Technology, Changsha, China
[3]School of Electrical and Information Engineering, Beihang University, Beijing, China
[4]Spintronics Interdisciplinary Center, Beihang University, Beijing, China

e-mail: xyliu@ime.pku.edu.cn

INTRODUCTION

3D Monte Carlo (MC) simulation method can properly simulate high-field non-local carrier transport that makes it suitable for nanoscale non-planar device simulations [1]. By appropriately accounting for quantum corrections, semi-classical MC simulation can give accurate results due to including scattering effect. However, both 3D simulation and quantum correction for MC demand huge computational resources [2].

In recent years, Intel has introduced its Intel® Xeon Phi™ Many Integrated Core (MIC) co-processor. It's famous for its powerful parallel processing capability and straightforward code portability [3]. We can easily acquire high parallel efficiency without rewriting codes as utilizing GPUs. In our previous work [4], we developed a 3D parallel Ensemble MC simulator with a max parallel speedup of around 30. In this work, we make use of Intel MIC co-processors on TianHe II to further improve parallel efficiency of our simulator. The results of the work indicate that MIC can make 3D MC device simulator more efficient for TCAD application.

SIMULATION METHOD

Our previous 3-D parallel full-band EMC simulator for Si devices [4,5], which takes into account the major scattering mechanisms including phonon, surface roughness (SR), impact ionization and ionized impurity scattering, can reproduce the bulk Si velocity-field characteristic and the inversion channel universal effective mobility curves [5]. The bulk Si full-band structure obtained by the local empirical pseudo-potential method is employed [5]. In addition, the effective potential (EP) method as the quantum correction method is used [4]. In 3D self-consistent Monte Carlo simulation, for each time step, 3D Poisson equation needs to be solved and millions of particles are simulated. In order to accelerate the 3D MC, by simulating a FinFET with different grid number, the simulation time for each part of the 3D MC is evaluated and shown in Fig.1. It can be seen that "Quantum Correction" part consumes about half of the CPU time and the time increases with the increasing grids number. Hence, following we will discuss the method to parallel the most time consuming part by MIC for clearness and simplification.

Fig.2 shows the flow chart of our simulator. Fig.3 shows a sample of compile statements needed when using MIC and it's easy and plain. Here, CPU is Intel Ivy Bridge Xeon processor and MIC is Intel Xeon Phi co-processor.

RESULTS AND DISCUSSIONS

SOI FinFETs with the structure shown in Fig.4 are simulated with different grids number to verify the simulator and evaulate the efficiency. The different grids is shown in Table I. The least grid number N is 44370.

A. Parallel Efficiency

For comparison, the initial condition keeps the same during simulation.

Firstly we compare one CPU (1CPU) with one CPU and multiple MICs (1CPU+xMIC) on one computing node as shown in Fig.5 and Fig.6.

Speedup ratio can be calculated as

$$\text{Speedup Ratio} = \frac{T_{CPU}}{T_{CPU+MIC}}$$

Fig.5 shows that for the least grid number N case, accelerated "Quantum Correction" can obtain 190-time speedup with 3-MIC acceleration and the simulation time is 11.9 mins,which is comparable to DD results by Sentaurus. Fig.6 shows that, by using one MIC, the simulation time of largest grid number is 19.6 hours and 3.2 times lower than without using MIC. Fig.7 shows that increasing node number can increase the speedup ratio. We can get 6.1 speedup when utilizing 5 computing node.

B. Verification

To verify the new developed MIC+CPU mode 3D MC simulator, a 14nm FinFET is simulated and the parameters is listed in Table II . The I-V curves and the corresponding distribution are compared with our original version of 3D MC simulator. Fig.8 shows the I-V curves and Fig.9 shows the potential distribution and electron energy distribution along channel direction. The results are of almost the same.Fig.10 shows some 3D electrical characteristic distribution results.

CONCLUSION

We demonstrate that MIC is effective to accelerate the 3D MC simulation. MIC co-processors is suitable for parallel processing in 3DMC simulation with large grids number.

ACKNOWLEDGMENT

We would like to express our sincere thanks to the staff of TianHe II. This work is supported by 973 Program (2011CBA00600) and NSFC 61421005.

REFERENCES

[1] Mark Lundstrom, *Fundamentals of carrier transport*, second edition, Cambridge University Press, 2000, pp 247.

[2] H.Haug, A-P.Jauho *Quantum Kinetics in Transport and Optics of Semiconductors, Springer 1998.*

[3] *Intel® Xeon Phi™ Coprocessor (codename: Knights Corner) Software Developers Guide Ver.1.04.*

[4] W. Zhang, Gang Du, et al, IWCE, 2009

[5] X.Y.Liu, K.L.Wei, Gang Du, et al, ICSICT, 2012

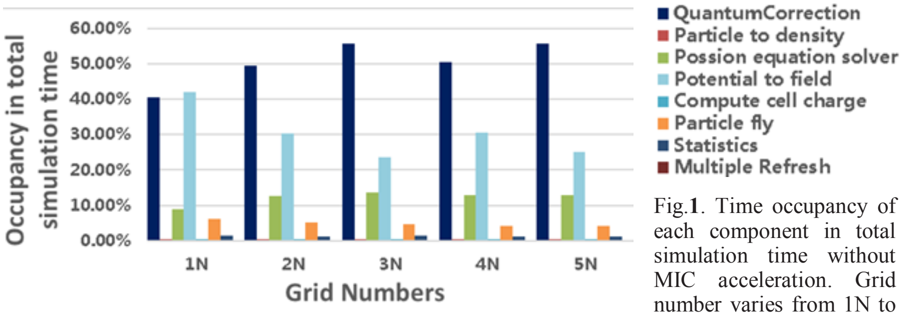
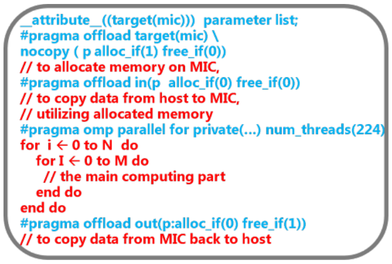


Fig.1. Time occupancy of each component in total simulation time without MIC acceleration. Grid number varies from 1N to 5N in Table I.



```
__attribute__((target(mic)))  parameter list;
#pragma offload target(mic) \
nocopy ( p alloc_if(1) free_if(0))
// to allocate memory on MIC,
#pragma offload in(p  alloc_if(0) free_if(0))
// to copy data from host to MIC,
// utilizing allocated memory
#pragma omp parallel for private(...) num_threads(224)
for  i ← 0 to N  do
   for I ← 0 to M do
      // the main computing part
   end do
end do
#pragma offload out(p:alloc_if(0) free_if(1))
// to copy data from MIC back to host
```

Fig.3. Compile statements used on MIC.

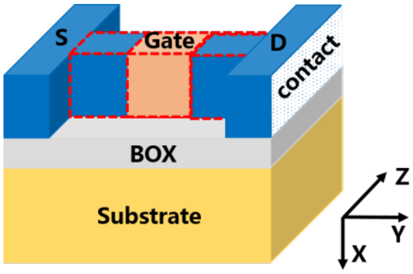


Fig.4. FinFET structure used in our simulation. QuantumCorrection region is indicated with dashed line.

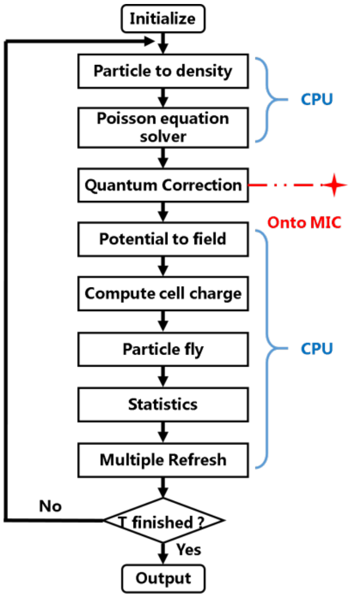


Fig.2. Flow chart of our 3D EMC simulator with "Quantum Correction" part accelerated by MIC.

Grid Type	Grids (x*y*z)
1N	51*30*29
2N	51*60*29
3N	51*90*29
4N	51*120*29
5N	51*150*29

Table **I.** Different grids number in our simulation.

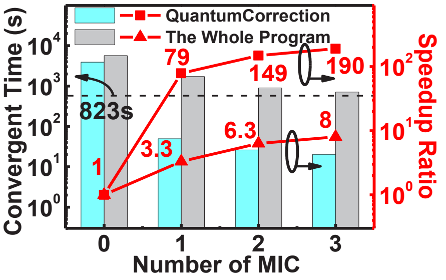


Fig.5. Convergent time and speedup ratio of the whole program and QuantumCorrection when utilizing 1,2,and 3 MIC co-processors on one computing node. Dashed line is the result of DD simulation by Sentaurus with one node. Grids is N.

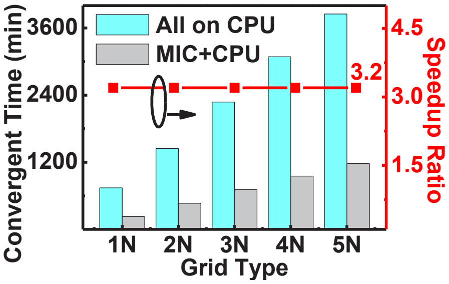


Fig.6. The relation of simulation time and speedup ratio along with different grids number when utilizing 1CPU and 1CPU+1MIC on one computing node. The speedup ratio keeps 3.2.

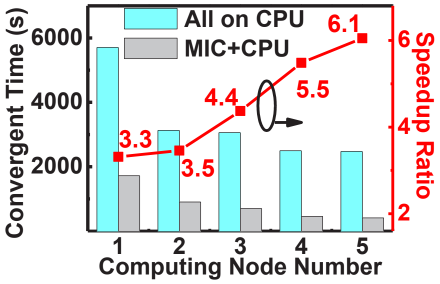


Fig.7. Convergent time and speedup ratio along with node number,1CPU and 1MIC on each computing node.Grids is N.

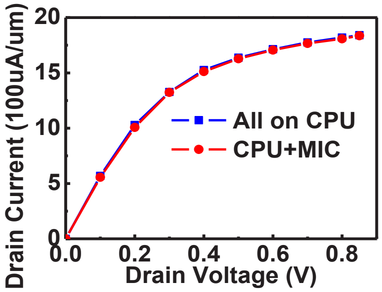


Fig.8. Comparison of I_D-V_D output characteristic curve between original simulator and MIC+CPU mode simulator.

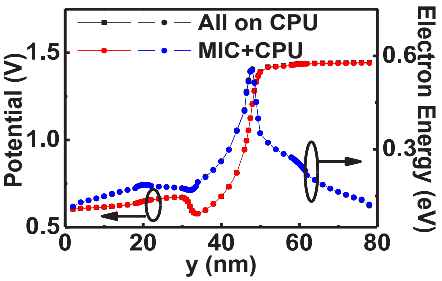
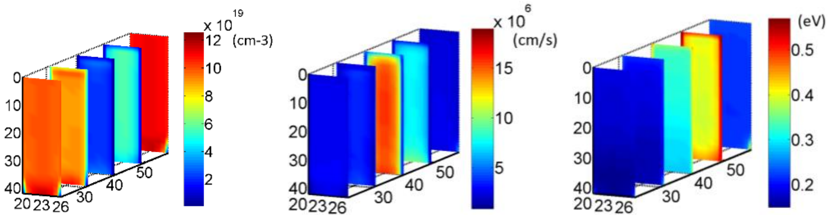


Fig.9. Comparison of electric characteristics along channel direction between original simulator and MIC+CPU mode simulator.



(a) electron density

(b) electron velocity

(c) electron energy

Fig.**10**. 3D electric characteristic distributions along y direction in QuantumCorrection region. Mesh grid number is 130000, particle number is about 2470000.

parameter	value
S/D doping	$1*10^{20}cm^{-3}$
channel doping	$1*10^{17}cm^{-3}$
channel length	16nm
fin height	42nm
fin width	6nm
BOX thickness	16nm
substrate thickness	100nm
substrate doping	$1*10^{18}cm^{-3}$

Table **II.** structure and operation parameters of 14nm FinFET used in our simulation.