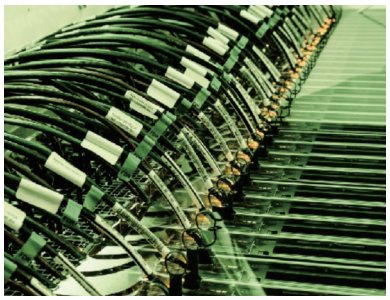
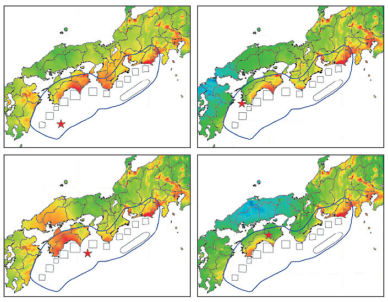
e-Science Journal
Global Scientific Information and Computing Center

Vol. 11

TSUBAME ESJ.





TSUBAME-KFC : the Greenest Supercomputer in the World With Liquid Submersion Cooling

Solving the Schrödinger Equations of Some Organic Molecules with Superparallel Computer TSUBAME

Application of GPGPU to Seismic Hazard Assessment

http://www.gsic.titech.ac.jp/TSUBAME_ESJ

TSUBAME-KFC : the Greenest Supercomputer in the World With Liquid Submersion Cooling

Toshio Endo Akira Nukada Satoshi Matsuoka
Global Scientific Information and Computing Center, Tokyo Institute of Technology

Modern supercomputer performance is principally limited by power. TSUBAME-KFC is a state-of-the-art prototype for our next-generation TSUBAME3.0 supercomputer and towards future exascale. In collaboration with Green Revolution Cooling and others, TSUBAME-KFC submerges compute nodes configured with extremely high density into oil coolant, and cooled using ambient / evaporative cooling tower, minimizing cooling power and leakage current. As a result, TSUBAME-KFC achieved world No.1 on the Green500 and the Green Graph500 simultaneously for the first time in Nov. 2013.

1 Introduction

The most predominant issue towards future exascale supercomputers is power consumption, and the importance of being "green'' has been, and still is a focus of many research in supercomputing. The DARPA exascale report [1], which established the goal of 20 Mega Watts for future exascale machines (50GFlops/Watt), analyzed the extreme challenges involved, and concluded that comprehensive low-power design would be required to even come close to the goal by 2018-20 timeframe.

Our research team at the Global Scientific Information and Computing (GSIC) Center, have been extremely cognizant of the low power requirements supercomputers. As an example, we have extensively investigating the use of extreme many-core GPUs in the HPC arena [2,3], which was resulted in TSUBAME2.0 supercomputer commissioned in Nov. 1st, 2010. Not only it became the Japan's first petascale supercomputer, it became #3 in the world in the Green 500 list, which is the power efficiency ranking (www.green500.org), with 958 MFlops/Watt, and in fact was recognized as the "greenest production supercomputer in the world".

Nonetheless we are still far away from the 50 GFlops/Watt goal. Thus, a new project, the Ultra Green Supercomputing was proposed and funded by MEXT (the Japanese Ministry of Education, Sports, Culture, Science and Technology). As a main deliverable of this project, *TSUBAME-KFC* was designed, constructed, and deployed in the fall of 2013.

TSUBAME-KFC facilitates extremely efficient cooling and durability via warm liquid (oil) submersion, in an unmanned container environment. Its cooling system is more power efficient than that of TSUBAME2 as discussed in Section 2 and 3.

Also TSUBAME-KFC serves as a prototype for TSUBAME3.0 to be commissioned in 2016, and features GPUs in extremely dense packaging, numerous sensors and control features of power and thermals. TSUBAME-KFC features over 600/200 TFlops performance (single precision and double precision namely) in a single rack, while maintaining the per-rack power consumption to the level of TSUBAME2 or approximately 35 KWatts per rack.

As a result, TSUBAME-KFC demonstrated world's top power efficiency on the Green 500 and the Green Graph 500 lists announced in November 2013, it became number one in the world in both.

2 Discussion on TSUBAME2 Power Consumption

One of the largest sources of power consumption of TSUBAME2 was identified to be cooling and semiconductor leakage power. TSUBAME2 cooling system utilizes low temperature water to cool a refrigerator-like semi-sealed rack by HP called the MCS rack, and inside the rack there is a forced circulation of cooled air, and the server inside is air-cooled. Here the inlet water temperature is approximately 8-9 degrees Celsius typical, while the outlet is 15-18 degrees. Except for winter, ambient temperature of Tokyo is well above such, and thus involves operating power-hungry compressors.

Although this solution is far more efficient than conventional air cooling in SC and IDC centers, due to the chilled water and rack/node fan requirements, with the observed PUE of 1.29 on the year average basis (refer to Section 4.2 for the definition of PUE). When we consider the fan, the PUE would rise, in that we are losing more than 30% of energy towards cooling.

Also, when GPUs are in full operation, their temperature rises to nearly 80-90 degrees even with our chilled water enclosed air cooling, raising up the semiconductor leakage power.

3 TSUBAME-KFC

3.1 Overview

In order to reduce overall total power usage, we have decided to build a liquid submersion cooled, highly dense cluster with extensive power and thermal monitoring, called TSUBAME-KFC (Kepler Fluid Cooling). TSUBAME-KFC was designed and built in collaboration with NEC, NVIDIA, and most importantly, Green Revolution Cooling (GRC) that provided the oil-based submersion cooling technology.

Fig 1 is the overview of KFC, and Fig 2 the external view of the container and the cooling tower. TSUBAME-KFC submerges the servers in warm oil of over 1000 liters. GRC's Carnot Jet oil submersion cooling (Fig 3) allows us to use standard servers with smaller degree of customization as is described later. Fig 4 shows how all the nodes are completely submerged in the oil coolant.

In order to cool the oil itself, there is a heat exchanger that transfers heat to the secondary water loop right next to the rack. The water in turn is cooled by evaporative cooling tower right outside the TSUBAME-KFC 20-feet container. The cooling tower is a standard drip cooler where the water is slowly flowed to the bottom, cooling the water with ambient air through radiation and evaporation in the process.

Fig. 2 Exterior view of TSUBAME-KFC: evaporative cooling tower right next to the 20-feet container for complete lights-out operation and low waterpump power

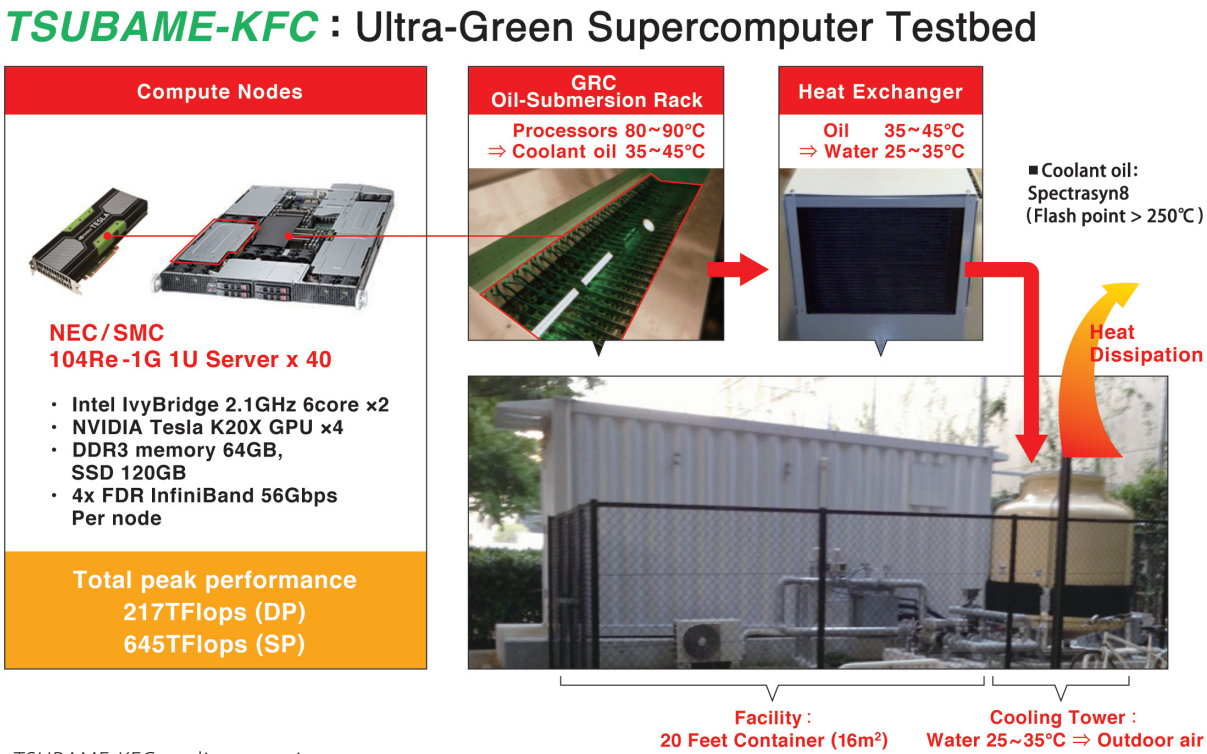


Fig. 1 TSUBAME-KFC cooling overview:
the heat emitted from the server is transferred o submerged oil than to water, than to ambient air.

Fig. 3 The GRC Carnot Jet system installed inside the container

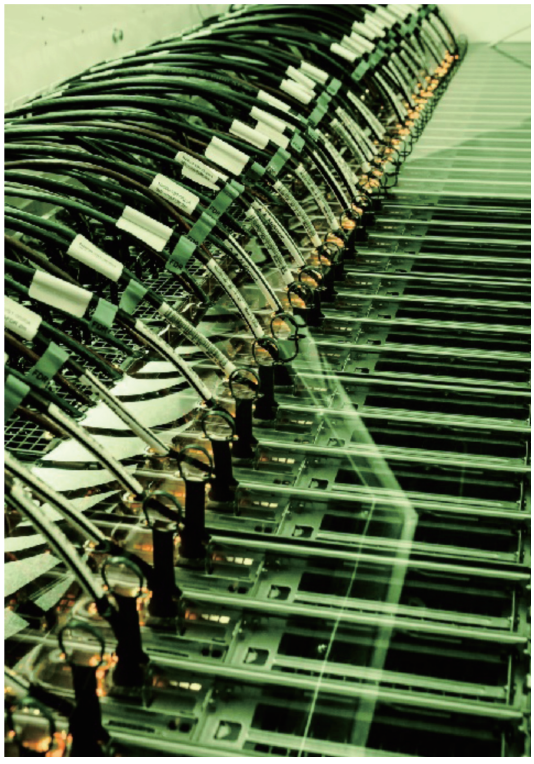


Fig. 4 The compute nodes submerged in the GRC Carnot Jet system.

3.2 Oil Submersion Cooling

TSUBAME-KFC's compute servers are submersed in a 42U rack, or more precisely speaking, a tub, as the rack is effectively placed sideways instead of vertically. The theoretical peak power consumption of the nodes are approximately 40 kilowatts, although the Carnot Jet system has the capability to handle up to 100 kilowatts per rack.

In order to conform to regulations in Japan, the cooling system was customized jointly with GRC and NEC as follows. The initial oil coolant proposed by GRC was non-toxic with low viscosity, but nonetheless had the flash point of 177 degrees Celsius. Due to this, it was considered as a flammable material under the Japanese fire laws subject to strict fire regulations, infeasible for large-scale operations. After extensive search, an alternative oil coolant Exxon Mobil Spectrasyn8 PAO (poly-alpha-olefin) was proposed, with flashpoint of 260 degrees, well above the threshold of 250 degrees avoiding such complications.

3.3 Compute Servers and their Customization

TSUBAME-KFC employs 40 nodes of a customized version of standard highly-dense compute server NEC/SMC 104Re-1G (Supermicro OEM) that embodies 2 CPUs and 4 GPUs in a dense 1U form factor. The node specifications are:

- Intel Xeon E5-2620 v2 (IvyBridge) 6 cores 2.1GHz x 2
- DDR3 memory 64GB
- NVIDIA Tesla K20X GPU x 4
- SATA3 SSD 120GB x 1+ 500GB x 2
- 4x FDR InfiniBand HCA x 1

The theoretical peak performance of each node is 16.1 TFlops in single precision and 5.4 TFlops in double precision floating point respectively. With 40 nodes comprising a single rack, the combined performance is approximately 217 TFlops in double precision and reaches over 645 TFlops in single precision, approaching the petaflop/rack metric for exascale.

Although standard servers were used as a baseline, the following customization were performed on the nodes jointly with GRC and NEC:

- Removal of moving components --- In order to submerse to high viscosity liquid, all moving parts such as server fans (12 units) were removed, as well as employing SSDs for storage. This has the additional benefit of lowering the node power requirements.
- Grease replacement --- Since silicone grease between the processor and the passive cooler will dissolve in oil, it was replaced with thin metallic sheets.

3.4 Power and Thermal Measurements

TSUBAME-KFC's embodies numerous power and thermal sensors, both on-line integral to the servers and other IT equipment. The following list shows the components measured by sensors:

- Power sensors
  - Each computer node
  - Network switch
  - Cooling tower
  - Two pumps for oil and water
- Thermal sensor
  - Each CPU and GPU
  - Oil and water
  - Outdoor/indoor air

The entire TSUBAME-KFC system was completed and began operation in October 2013, and will continue with various experimentations leading up to TSUBAME3.0, until Spring of 2016.

TSUBAME-KFC Evaluation 4

4.1 Effects of Oil Submersion Cooling to the Servers

We measured how the oil submersion cooling affects the power and thermals of individual TSUBAME-KFC node. For comparative purpose, we configured an air-cooled node with exactly the same CPU/GPU/memory hardware.

Fig 5 shows the comparison between air-cooled with inlet 26 degrees Celsius, versus 29 and 19 degrees inlet for the oil coolant. The servers are continuously running double precision matrix multiply using CUBLAS on all GPUs to incur the highest power and thermal load.

Cooling	Air-Cooled (26 °C)	Submersion (Oil 29 °C)	Submersion (Oil 19 °C)
Temp (°C)			
CPU1	46	42	33
CPU2	50	40	31
GPU1	52	47	42
GPU2	59	46	43
GPU3	57	40	33
GPU4	48	49	42
Node Power (W)	749	693	691

Fig. 5 Comparison of a TSUBAME-KFC submersed node and an air-cooled node

Comparing the air-cooled versus submersion, although the former has lower temperature input, the latter exhibits substantially lower temperature, especially in GPUs. Comparing the server power consumption, oil submersion is approximately 7.8% lower than air. This is largely the combined effect of fan removal and lower semiconductor temperature suppressing leakage current.

4.2 PUE Measurement

PUE (power usage effectiveness) is widely used in datacenters as a metric to measure the power efficiency of cooling. It is given by

$$PUE = (P_E + P_I) / P_I$$

where P_E stands for power consumption of IT equipment (computers, network switches, etc.) and P_I for power for infrastructure; in this article, we equate cooling power to infrastructure power. PUE=1 indicates the ideal case of no power required for cooling, while PUE=2, which is a fairly common value in classical datacenters, indicate that cooling is spending as much power as the IT equipment.

In order to measure TSUBAME-KFC PUE, we stressed the server with the load of GPU matrix multiply as in the previous subsection for all the nodes. Fig 6 shows the results. The power of TSUBAME-KFC is derived from the actual measurements from the real-time power sensors, while the air cooling was extrapolated from real power measurement of the server, with the assumption of PUE=1.29 as in TSUBAME2.

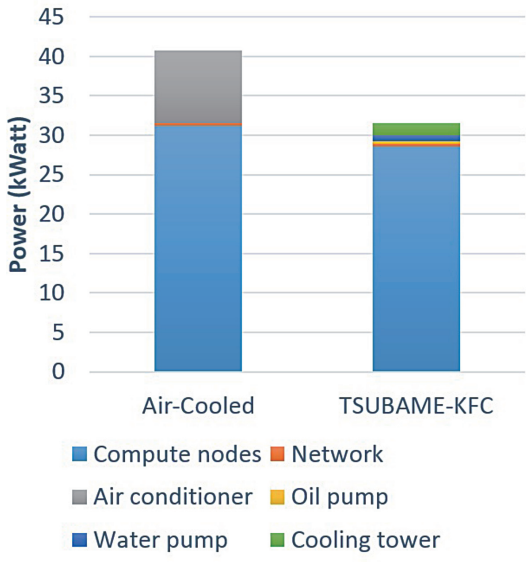


Fig. 6 PUE Evaluation of TSUBAME-KFC versus air-cooled system

According to the measurements, apparent PUE of TSUBAME-KFC is 1.09, which is larger better than 1.29. Furthermore, we should note that power consumption of computer nodes is also reduced as shown above; by making the denominator of the IT power to be that of the air-cooled server as a comparison, the adjusted PUE is 1.001(!)

4.3 Green500 Measurement

This section describes our challenge to the Green 500 lists. Green 500 measures the power efficiency of the system during the Linpack benchmark run; the systems are compared by (LinpackFlops / Power), where "Power" does not count cooling power. Although contributions from better cooling is not directly measured, we expect higher efficiency through improved node efficiency as described above, as well as improved tuning.

In order to obtain the maximum power efficiency, the following strategies in hardware and software are keys:

- At the architecture level we adopted the ratio of GPU to CPU of 2:1 rather than 1:1, thus decreasing the overhead of CPU.
- We employed a new, more efficient in-core Linpack kernel provided by NVIDIA.
- We tuned the HPL parameters by exhaustive search of the parameter space. This involved not only the standard tuning of HPL parameters such as the block size (NB), and process grid (P&Q), but also adjustment of the GPU clock and voltage.

During the tuning phase, we noticed that the best (absolute) performance does not equal best power efficiency. Fig 7 shows the best power efficiency is 24 % better than the case with the best speed performance.

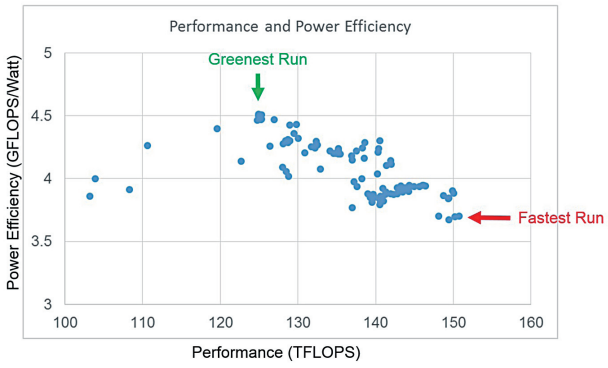


Fig. 7 Results of Linpack benchmark runs with various configurations. Each dot corresponds to a single Linpack run.

As a result, Flops/Watt reached 4.503GFlops/Watt. On November 18th, 2013, the Green 500 list was announced, in which TSUBAME-KFC was ranked #1 in the world, with 24% lead over Wilkes, the second ranked machine.

4.4 Green Graph 500 Measurement

The Green Graph 500 list (green.graph500.org) is a newly announced list, which is for the power efficiency in big data analysis. Instead of Linpack benchmark in Green 500, Green Graph 500 uses breadth first search (BSP) benchmark for large scale graph structures. The speed of this benchmark is expressed in TEPS (Traversed Edges per Second) value, and Green Graph 500 ranking is determined by TEPS/Watt value.

For the BSP benchmark, we have to be more careful for selection of computing resources. According to the results of preliminary comparison of the CPU implementation and the GPU implementation [4], both were well optimized, the former was selected for better power efficiency.

As a result, the power efficiency with 32 TSUBAME-KFC nodes reached 6.72 MTEPS/Watt (= 44.01GTEPS / 6.55kWatt). This result was ranked #1 in Green Graph500 list (the big data category) in November 2013.

4.5 Evaluation with the Phase-Field Simulation

Finally, we show the evaluation of power efficiency of a real application, namely the stencil based ``phase-field'' simulation, which was awarded the 2011 Gordon Bell prize[5]. This application simulates the micro-scale dendritic growth of metal materials during solidification phase. We show the results of this application since we have used it for evaluation of power efficiency for multiple years including TSUBAME1.0 commissioned in 2006.

Fig 8 shows the development of power efficiency by comparing several machines in different generations since 2006. Here we observe a gap between the 2006 (CPU only) and 2008 (with GPUs) numbers, which is the one-time performance leap with the many-cores transition. By extrapolating the results of multiple generations of GPU machines, we estimate that the expected performance circa 2016 being as 15.5 GFlops/Watt, which is about 1,210 times more efficient than the 2006 number. This result also supports the feasibility of achieving 50GFlops/Watt around 2020, though we still need intensive research towards better power efficiency.

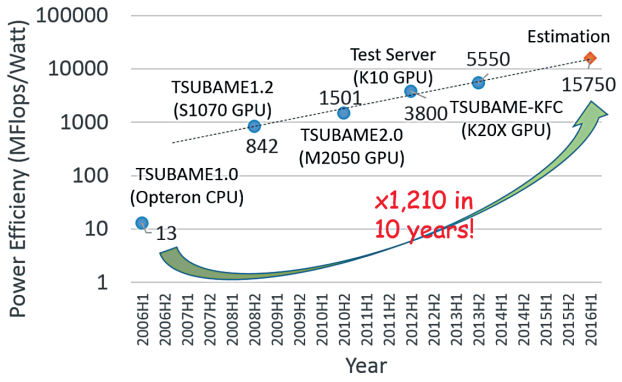


Fig. 8 Power efficiency of the phase-field simulation on machines in different generations.

Conclusion 5

We demonstrated a prototype supercomputer that combines most of the known the state-of-the-art architecture in both hardware and software, materialized as TSUBAME-KFC, achieving the world's top power efficiency in both Green 500 and Green Graph 500 rankings circa November 2013. TSUBAME-KFC is 24% better than the #2 machine. The most notable technology deployed by TSUBAME-KFC is the warm oil submersion cooling, which improves power efficiency over a similar production machine, namely TSUBAME2.5 with the same GPUs.

We will continue our experimentations of TSUBAME-KFC, some of the most up-to-date-result only obtainable during summer. As an experimental platform, we will conduct further updates either in software or hardware if affordable, to affect the design of TSUBAME3.0 as the bleeding-edge power efficient and big data supercomputer of the era.

References

[1] P. Kogge, K. Bergman et al. ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems, DARPA Technical report 2008-13, 2008.

[2] Akira Nukada and Satoshi Matsuoka. Auto-Tuning 3-D FFT Library for CUDA GPUs. In Proceedings of the ACM/IEEE conference on Supercomputing (SC09), 10 pages, ACM Press, Portland, November 2009.

[3] Takashi Shimokawabe, Takayuki Aoki, Chiashi Muroi, Junichi Ishida, Kohei Kawano, Toshio Endo, Akira Nukada, Naoya Maruyama, Satoshi Matsuoka. An 80-Fold Speedup, 15.0 TFlops, Full GPU Acceleration of Non-Hydrostatic Weather Model ASUCA Production Code. In Proceedings of the ACM/IEEE conference on Supercomputing (SC10), pp.1-11, New Orleans, November 2010.

[4] Koji Ueno and Toyotaro Suzumura, Parallel Distributed Breadth First Search on GPU, IEEE International Conference on High Performance Computing (HiPC 2013), India, Dec. 2013.

[5] Takashi Shimokawabe, Takayuki Aoki, Tomohiro Takaki, Akinori Yamanaka, Akira Nukada, Toshio Endo, Naoya Maruyama, Satoshi Matsuoka. Peta-scale Phase-Field Simulation for Dendritic Solidification on the TSUBAME 2.0 Supercomputer. In Proceedings of the ACM/IEEE conference on Supercomputing (SC11), pp. 1--11, Seattle, November 2011.

Solving the Schrödinger Equations of Some Organic Molecules with Superparallel Computer TSUBAME

Hiroshi Nakatsuji Hiroyuki Nakashima
Quantum Chemistry Research Institute

The purpose of "first-principle simulation" is to give a perfect prediction of phenomena. Since the Schrödinger equation is the governing principle of chemistry, one should be able to predict chemical phenomena based on the exact solutions of this equation. Since 2000, Nakatsuji and his colleagues have been successful to formulate a general theory of exactly solving this equation: the free complement (FC) theory combined with the local Schrödinger equation (LSE) method is applicable to any atomic and molecular systems. Recently, we performed the FC-LSE calculations of some simple organic molecules using the superparallel computer TSUBAME in the occasion of Grand Challenge and obtained accurate wave functions and absolute energies satisfying chemical accuracy (kcal/mol). In the calculations, the computational tasks due to the sampling can be equally distributed to each process (core). In fact, the parallel efficiency 111.5 % could be achieved with the test calculation of benzene using 4600 cores of the superparallel computer TSUBAME. The present accurate solutions of the Schrödinger equations of organic molecules are of significant importance in science that provides a basis for future development in theoretical and computational chemistry.

1 Introduction

For a perfect prediction of phenomena in simulation science, the basic theory must be exact and at the same time, can be solved in high speed. In the predictions of the motions of astronomical objects or spacecrafts, their governing principles are Newtonian mechanics plus a bit of relativity. Their simulation techniques to solve the principal equations have been clearly established. Many successful examples, therefore, are well known like Mars exploration and a remarkable accomplishment of "Hayabusa", etc. On the other hand, chemistry is governed by the Schrödinger and Dirac equations. If these equations were solved exactly in high speed, we would be able to predict and even simulate chemical phenomena. This was a dream of theoretical quantum scientists but was not realized for over 80 years after the Schrödinger equation was born. Then, quantum chemistry had been considered as a science just for understanding and explanation. Since 2000, however, Nakatsuji first discovered a general method for solving the Schrödinger equation and this theory was formulated as the free complement (FC) theory [1,2]. Combining with the local Schrödinger equation (LSE) method [3] which was introduced to overcome the integration difficulty, the FC method had become applicable to any atomic and molecular systems. Based on this theory, we are developing our original quantum chemistry package "Principia".

In the present study, we applied the FC-LSE method to solving the Schrödinger equations of some organic molecules, composed of the light atoms such as H, C, N, and O, with the superparallel computer TSUBAME in the grand challenge subject from Oct. 8th to 15th in 2013. Since any theory in quantum chemistry has not been able to give highly accurate solutions (in chemical accuracy) of their Schrödinger equations, the accurate results obtained in the present GC project may have a significant value in fundamental science. It was also expected that our method is suitable for massively parallel computing. So, we would like to demonstrate it, at the same time, with the use of the super computer TSUBAME.

Based on the results obtained in this GC project and further improvements of our method and computational techniques, we would like to do our best to open a way toward the exact-simulation theoretical chemistry in organic chemistry.

2 Free complement theory

First, we briefly explain our free complement (FC) method and introduce our review article [4] for an easy understanding. Figure 1 is a sketch of a scheme of the FC methodology.

2.1. FC method: Generation of the complement functions

According to the theory of the structure of the exact wave function, the FC method is based on the theory that the system's Hamiltonian creates its own complete space which spans the exact solutions of the system. By applying the Hamiltonian H several times to an appropriate initial function ψ_0, it is guaranteed that the wave function converges to the exact solution [1,2],

$$\psi_{n+1} = \left[1 + C_n g(H - E_n)\right]\psi_n \qquad (1)$$

where C_n and E_n denotes a variational coefficient and energy at iteration (or order) n. g is introduced to avoid the singularities that appear in the Coulomb potentials [2]. By expanding the right-hand-side of Eq. (1) into the independent functions, we obtain the complement functions (cf) $\{\phi_i\}$ that span the exact wave function as,

$$\psi = \sum_i^M c_i \phi_i \qquad (2)$$

where M denotes the number of the complement functions (dimension) and c_i is a unknown coefficient of each complement function ϕ_i. This $\{c_i\}$ is determined by the ordinary variational principle or by the LSE method explained below.

2.2. LSE method: Integral-free approach based on the sampling method

The variational method is applicable only when the Hamiltonian and overlap matrices are calculated by integration. In this case, very accurate results can be expected especially for the energy. This method, however, is only applicable to small atomic and molecular systems. On the other hand, the LSE method utilizes a necessary condition that the Schrödinger equation is satisfied at a local coordinates [3] $\mathbf{r}_\mu$ chosen as a sampling point

$$H\psi(\mathbf{r}_\mu) = E\psi(\mathbf{r}_\mu) \qquad (3)$$

This method does not require any analytical integration and is applicable to any atoms and molecules and to any functional forms.

The most time-consuming task, which is antisymmetrization of the wave function, is equally divided into each processor by distributing the sampling points. The LSE method is theoretically quite suitable for massively parallel computing.

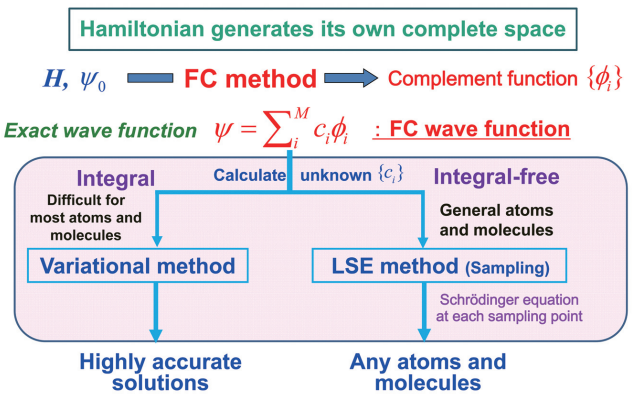


Fig. 1 Scheme of the FC methodology

2.3. Pauli principle and antisymmetrization

For solving the Schrödinger equation of the electronic system, the wave function must satisfy the Pauli principle, which demands the antisymmetrization of the wave function for the exchange of any pairs of electrons. However, it requires a huge computational cost for accurate calculations of many-electron systems.

We proposed a determinant-based antisymmetrization method [5]. It is referred to as Nk method. The computational order of this method is almost $O(N^3\text{-}N^5)$, where N is the number of electrons of the system. This is true, even when the wave function includes the explicitly correlated r_{ij} terms. This method is almost established and has been applied to various atoms and molecules. In the present study, we used this Nk method for the calculations of organic molecules.

Recently, an alternative antisymmetrization method called inter exchange (*i*Exg) theory [6] was discovered. It is theoretically proved that it makes the computational cost just in order-N. Thus, the iExg theory may bring a breakthrough for solving the Schrödinger equations of large molecules. The details will be published elsewhere in near future.

Superparallel algorithm of the FC-LSE method 3

The parallel algorithm of the FC-LSE method is summarized in Figure 2.

Step 1,2:
In this step, the complement functions are generated according to the system's Hamiltonian based on Eqs. (1) and (2). They are analytically formulated and this step requires only a small computational cost.

Step 3.1:
Based on the LSE method, Eq. (3) is applied to the FC wave function at given sampling points $\mathbf{r}_\mu$. Eq. (3) is replaced with a matrix eigenvalue equation, given by

$$\mathbf{AC} = \mathbf{BCE} \qquad (4)$$

where the elements of the matrices $\mathbf{A}$ and $\mathbf{B}$ are $\mathbf{A}_{\mu i} = H\phi_i(\mathbf{r}_\mu)$ and $\mathbf{B}_{\mu i} = \phi_i(\mathbf{r}_\mu)$, respectively, and $\mathbf{C}$ and $\mathbf{E}$ represents the coefficient vectors and energy (diagonal) matrices, respectively.

The evaluations of $\phi_i(\mathbf{r}_\mu)$ and $H\phi_i(\mathbf{r}_\mu)$ contains the antisymmetrization about the electrons and require large

computational cost. This is the most time-consuming step in the FC-LSE calculations. However, since the sampling points, which exactly have the same computational costs, are equally distributed to each process (or core), highly parallel efficiency can be expected. Moreover, there is no communication among the processors. Since 10^6-10^7 sampling points are usually employed, only 1-10 points are assigned to each process in a peta-flops superparallel computer. The computational order of this step is $M \cdot N_s \cdot o(N_e^{3-5})$, where Ns and Ne denote the numbers of the sampling points and electrons, respectively.

Step 3.2:

When the larger number of sampling points than the dimension M is employed, the matrices $\mathbf{A}$ and $\mathbf{B}$ of Eq. (4) become rectangular. In this case, applying $\mathbf{B}^\dagger$ from the left side of Eq. (4), one obtains

$$\mathbf{HC} = \mathbf{SCE} \qquad (5)$$

where $\mathbf{H} = \mathbf{B}^\dagger\mathbf{A}$ and $\mathbf{S} = \mathbf{B}^\dagger\mathbf{B}$ correspond to the Hamiltonian and overlap matrices, respectively. $\mathbf{S}$ is a positive definite matrix and $\mathbf{H}$ also approaches to a symmetry matrix. So, the eigenvalue equation of Eq. (5) can be stably solved (see step 4).

This step requires matrix products and one should use the BLAS3 library optimized in the machine environment.

The computational order of this step is $M \cdot N_s$ and there is no communication among processes.

Step 3.3:

This step collects all the $\mathbf{H}$ and $\mathbf{S}$ matrices from each process and MPI_Reduce is used with the summations. The amount of the data communication is M^2.

Step 4:

In this step, the eigenvalue equation of Eq. (5) is solved. A parallelized numerical library for the eigenvalue problem should be usually available and a moderate parallel efficiency is expected. To effectively use a superparallel resource, one should once stop the calculation after step 3 and this step should be performed with a small parallel machine.

Step 5:

Using the eigenvector (wave function) obtained in step 4, the physical observables including energy are evaluated in this step. The H-square error, which is an important indicator of the exactness of the wave function, is also evaluated. This step requires only a small computational cost and small memory and disc resources.

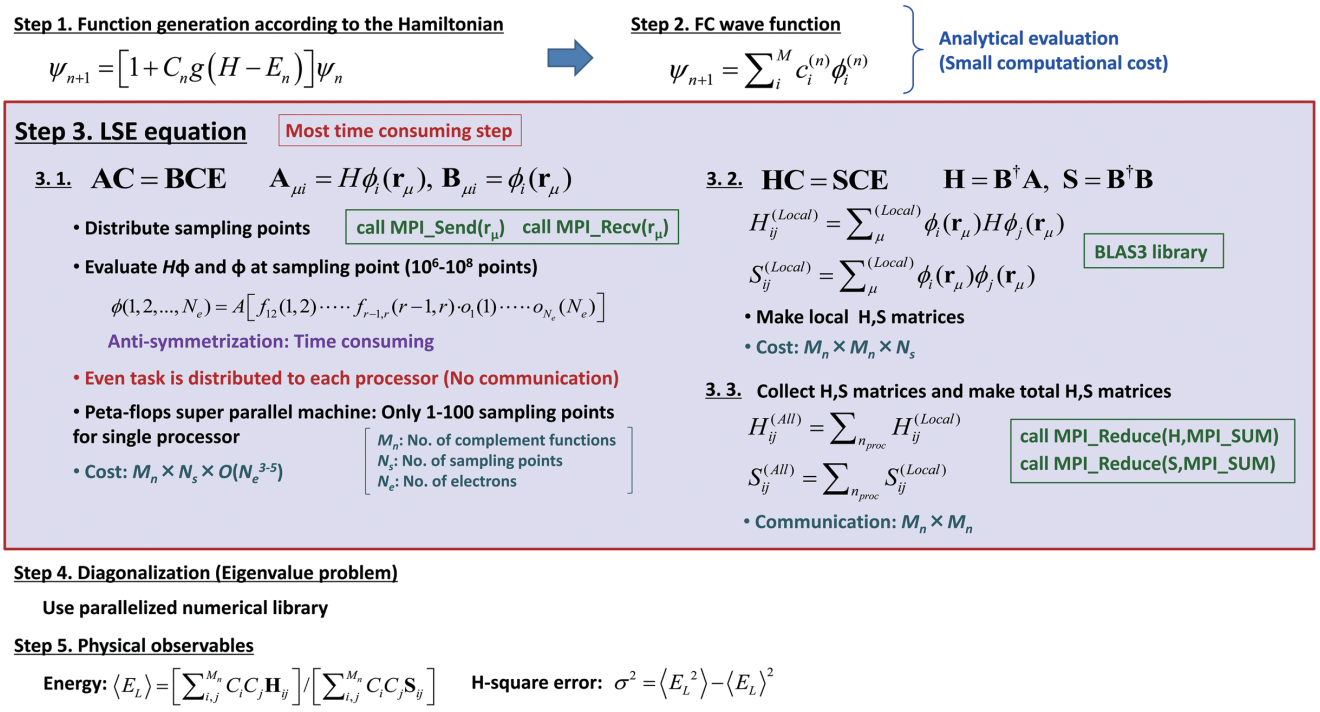


Fig. 2 Parallel algorithm of the FC-LSE method

Table. 1 Solving the Schrödinger equations of small organic molecules (Order of the FC method: n=2)

Molecule	No. of elec.	Dimension	Energy (a.u.)		$\Delta E=E_{FC\text{-}LSE}-E_{exact}$ (kcal/mol)
			FC-LSE	Estimated exact (experiment)	
Carbon hydride (CH)	7	1503	-38.480 41	-38.479 0	-0.88
Water (H_2O)	10	2075	-76.456 78	-76.457 8	0.67
Dicarbon (C_2)	12	1976	-75.923 69	-75.926 5	1.76
Nitrogen molecule (N_2)	14	1121	-109.542 07	-109.542 7	0.39
Acetylene (C_2H_2)	14	1709	-77.333 31	-77.335 7	1.49
Ethylene (C_2H_4)	16	2628	-78.577 95	-78.587 4	5.93
Formaldehyde (H_2CO)	16	4083	-114.505 35	-114.508 0	1.66

Table. 2 Test calculations of medium-size organic molecules (Order of the FC method: n=2)

Molecule	No. of elec.	Dimension	Energy (a.u.)		$\Delta E=E_{FC\text{-}LSE}-E_{exact}$ (a.u.)
			FC-LSE	Estimated exact (experiment)	
Furan (C_4H_4O)	36	161	-229.860 1	-230.027	0.167
Pyrrole (C_4H_5N)	36	174	-209.974 3	-210.173	0.199
Benzene (C_6H_6)	42	398 5092[a]	-232.409 3 -232.195 8[a]	-232.248	-0.161 -0.052[a]
Pyridine (C_5H_5N)	42	386	-247.704 1	(-248.290)	0.586

[a] Large FC order for the carbon atoms

Solving the Schrödinger equations of some organic molecules 4

We applied the FC-LSE method to some organic molecules given in Figure 3.

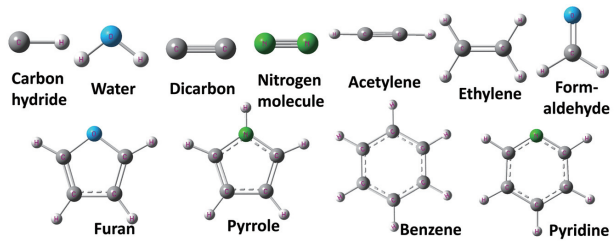


Fig. 3 Some organic molecules to which the FC-LSE method was applied

The analytical form of the complement functions is represented by

$$\phi_i(1,2,\ldots,N_e)=\prod_i^{N_e}\Big(\chi(\mathbf{r}_{iA_i})\cdot x_{iA_i}^{k_{x,iA_i}}y_{iA_i}^{k_{y,iA_i}}z_{iA_i}^{k_{z,iA_i}}\left|\vec{r}_{iA_i}\right|^{k_{r,iA_i}} \times\prod_{A(\neq A_i)}^{N_A}\left|\vec{r}_{iA}\right|^{k_{iA}}\times\prod_{j(>i)}^{N_e}\left|\vec{r}_{ij}\right|^{m_{ij}}\Big)\cdot\sigma_I \qquad (6)$$

where N_e and N_A is the numbers of electrons and nuclei, respectively. The indices: i and j denote electron and A denotes nucleus. A_i shows the nucleus that electron i belongs. $\chi(\mathbf{r}_{iA_i})$ is a atomic orbital of electron i and the Hartree-Fock orbitals of atoms were employed here [7]. σ_I is a spin eigenfunction and the chemical bonds were constructed according to the valence orbital theory. Note that, since the FC method is guaranteed to converge to the exact solution even with any starting function, one does not have to consider every possible configuration which appears in the valence bond theory.

Molecular sampling points are synthesized from the atomic sampling points by the local sampling method [8]. The H-square error stationary (HSES) algorithm based on the local variance contribution was also employed to optimize the sampling points and to reduce the sampling-points

dependency [8]. To simplify the calculations, we omit the complement functions including the explicitly correlated terms (Inter-r_{ij}) among different atoms. To investigate these effects in more details, we now perform the systematic examinations.

The calculations were performed on the TSUBAME grand challenge subject (category B). Table 1 summarizes the results from CH (7 electrons) to H_2CO (16 electrons) molecules with the FC order n=2. The energy differences ΔE between the calculated energies and exact values estimated from the experiments were almost in kcal/mol order or less for all the molecules: they satisfy chemical accuracy in absolute energy. Table 2 summarizes the test calculations of furan and pyrrole (36 electrons) and benzene and pyridine (42 electrons) with the FC order n=1. In this case, ΔE still show small differences but their absolute energies were accurate enough, when compared to the result of the conventional quantum chemistry methods. For instance, the absolute energy of benzene with the MP2-F12 method, which is one of the advanced electron correlation theories, was -231.835 6 a.u. [9] and our calculated energy was closer to the estimated exact value. The present results are still a testing phase and we will improve the results by increasing the order with more intensive calculations.

Thanks to the superparallel computer power of TSUBAME, all of the above calculations had finished within a week. In the period of the grand challenge, we could intensively improve, at the same time, the algorithms of the computations.

Parallel efficiency 5

To investigate the parallel efficiency of the FC-LSE method, we performed the test timing calculation of benzene. Figure 4 shows the acceleration by the parallelization, where we used 2300 and max 4600 cores and the timing was measured with reference to 460 cores. As a result, we obtained the parallel efficiency of 111.5 % with 4600 cores. The acceleration by the parallelization is expected for bigger systems because the rate of most time-consuming part, i.e. antisymmetrization step (step 3.1 in Fig. 2), increases. The reason why the efficiency was over 100 % is considered as a lack of tuning the program but it may also be due to the dispersion effect of the memory resource by increasing the nodes.

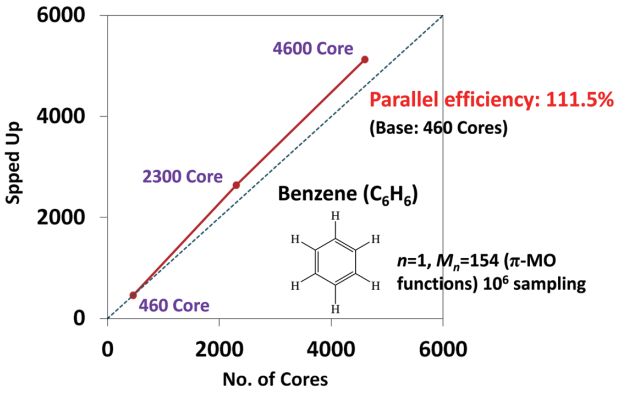


Fig. 4 Timing test with benzene for the parallel efficiency

Summary 6

Solving the Schrödinger equation was a dream in theoretical chemistry and we could realize it for the calculations of some simple organic molecules. Although we still need to examine the calculated results systematically, the present results were quite accurate, compared to the results from the conventional quantum chemistry methods. The TSUBAME grand challenge was also very much helpful to examine the theory and computational algorithms and program. Whereas most of the conventional quantum chemistry theories are not suitable for superparallel computations, the FC-LSE method is based on the theory that is very suitable for massively parallel computing. Furthermore, it is flexible for both the ordinary machine clusters and superparallel machines.

Recently, we proposed a theoretical synthesized method of the molecular wave function and a brand-new antisymmetrization theory and, thus, there was a big progress recently in both theory and algorithm [6]. Based on the results of the present study, we will further cultivate our theories and methodologies to develop the accurately predictive quantum chemistry.

Acknowledgments

The present study was performed on Oct.8-15, 2013 with the TSUBAME2 superparallel machine at Tokyo Institute of Technology as a subject of the TSUBAME grand challenge (category B). Parts of the calculations after the grand challenge

were performed at the computer center in the Institute of Molecular Science (IMS). We deeply acknowledge their supports. We also acknowledge Mr. Nobuo Kawakami for his support to the researches of QCRI.

References

[1] H. Nakatsuji: Structure of the Exact Wave Function, J. Chem. Phys., Vol. 113, pp.2949-2956 (2000). H. Nakatsuji and E. R. Davidson: Structure of the Exact Wave Function. II. Iterative Configuration Interaction Method, J. Chem. Phys., Vol. 115, pp.2000-2006 (2001)

[2] H. Nakatsuji: Scaled Schrödinger Equation and the Exact Wave Function, Phys. Rev. Lett., Vol.93, pp.030403-1-4 (2004). H. Nakatsuji: General Method of Solving the Schrödinger Equation of Atoms and Molecules, Phys. Rev. A, Vol. 72, pp.062110-1-12 (2005)

[3] H. Nakatsuji, H. Nakashima, Y. Kurokawa, and A. Ishikawa: Solving the Schrödinger Equation of Atoms and Molecules without Analytical Integration Based on the Free Iterative-Complement-Interaction Wave Function, Phys. Rev. Lett., Vol.99, pp.240402-1-4 (2007)

[4] H. Nakatsuji: Discovery of a General Method of Solving the Schrödinger and Dirac Equations That Opens a Way to Accurately Predictive Quantum Chemistry, Acc. Chem. Res., Vol. 45, pp.1480-1490 (2012)

[5] H. Nakashima and H. Nakatsuji: Efficient antisymmetrization algorithm for the partially correlated wave functions in the free complement - local Schrödinger equation method, J. Chem. Phys., Vol. 139, pp. 044112-1-16 (2013).

[6] H. Nakatsuji: ACS meeting, Indianapolis, U.S.A. (2013)

[7] E. Clementi and C. Roetti: Roothaan-Hartree-Fock Atomic Wavefunctions, Atomic Data and Nuclear Data Tables, Vol. 14, pp.177-478 (1974)

[8] H. Nakatsuji: Local sampling method and H-square error stationary algorithm, to be submitted.

[9] D. Yamaki, H. Koch, and S. Ten-no: Basis set limits of the second order Møller-Plesset perturbation energies of water, methane, acetylene, ethylene and benzene, J. Chem. Phys.,Vol. 127, pp.144104-1-5 (2007)

Application of GPGPU to Seismic Hazard Assessment

Shin Aoi* Takahiro Maeda* Takayuki Aoki**

* National Research Institute for Earth Science and Disaster Prevention

** Global Scientific Information and Computing Center, Tokyo Institute of Technology

A high-accurate and large-scale ground motion simulation is required for the seismic hazard assessment. We applied multi GPU technique to the GMS (Ground Motion Simulator) which is a practical tool for wave propagation simulation based on 3-D finite difference method using discontinuous grid. The performance test for the multi GPU calculation showed almost perfect linearity for the weak scaling test up to the simulation with 1024 GPUs; the model size for the 1024 GPUs case was about 22 billion grids. Lastly, we performed the long-period ground motion simulation for the Nankai Trough earthquake using the multi GPU code.

1 Introduction

Three-dimensional ground motion simulation is one of the key techniques for seismic hazard assessment using the hybrid method. In the method, high frequency component which is rather random process is simulated by the stochastic Green's function method. On the other hand, low frequency component is calculated by deterministic method such as FDM (finite difference method) and FEM (finite element method) assuming source and velocity structure models. The capacity (calculation and memory resources) of recent powerful computer is still not enough for detailed seismic hazard assessment because the practical model is rather large scale.

One of the prevailing approaches to overcome the problem of heavy computational expense is the utilization of the GPGPU (General Purpose computing on Graphics Processing Units)[1][2][3]. GPGPU is the technique of using a GPU to perform computation in applications traditionally handled by the CPU. In this study, we develop the FD code for the multi GPU simulation and examine the performance tests on TSUBAME which is one of the Japanese fastest supercomputer operated by Tokyo Institute of Technology.

2 3-D FD simulation of wave propagation using GPU

2.1 FD formulation using discontinuous grid

In this study, we employed the solver of the GMS (Ground Motion Simulator)[5] as the original code and developed the code for the multi GPU computation. GMS is a total and practical system for seismic wave propagation simulation based on 3-D FDM using discontinuous grids[4], which includes the solver as well as the preprocessor tools (parameter generation tool) and postprocessor tools (filter tool, visualization tool, and so on). The solver of the GMS employs the velocity-stress formulation using the FD operator having fourth-order of accuracy in space and second-order in time.

One of the problems in FD modeling using uniform grid is the extra computational requirements which are related with the oversampling of the models. This problem is getting obvious when the model includes the strong velocity contrast, because the grid spacing is determined by the shortest wavelength in the region to be calculated. In order to avoid this oversampling we use a discontinuous grid (Fig. 1) that is a kind of a non-uniform grid adapted to the velocity structure.

This grid system consists of two regions having the different grid spacing whose ratio is a factor of three, and the continuity of the wave field are kept by eliminating or inserting grids at the overlapping region of these two regions. The use of discontinuous grid significantly reduces the computational requirements, which is model dependent but typically one-fifth to one-twentieth, without a marked loss of accuracy.

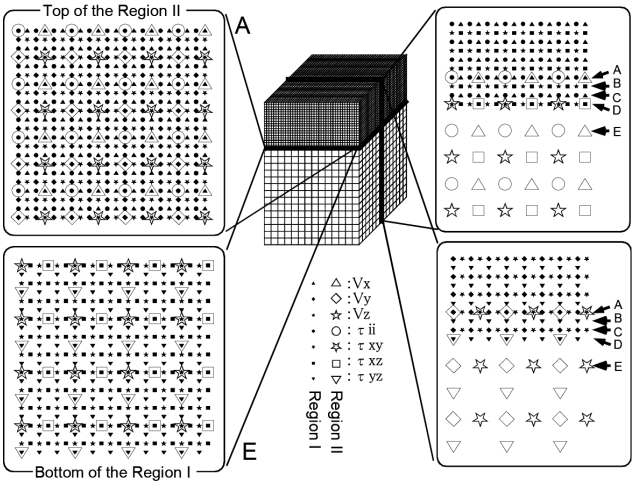


Fig. 1 (center) 3-D discontinuous grid system. (left) Two transections on the top and at the bottom of the overlapping region of Regions I and II, where the elimination or the insertion of grid points is necessary. (right) Two profiles of the discontinuous grid.

2.2 FD calculation on GPU

We developed the FD code using CUDA (Compute Unified Device Architecture) so that almost all FD calculations are executed on GPU. A device constituted by GPU and VRAM (global memory) is regarded as an accelerator in CUDA (Fig. 2 left); GPU executes calculation and VRAM stores processing data. In the FD calculation, blocks which are a group of thread (minimum unit of the execution on GPU) are distributed over the x-y plane. Each thread in a block advances a calculation for a grid point from the starting point to a terminal point in z-direction, and finally the whole three-dimensional domain will be calculated (Fig. 2 right). Because a memory bandwidth tends to be a bottleneck for the FDM which requires a large number of memory accesses, the number of access to a high-latency global memory is reduced by using the low-latency register and shared memory on GPU like a cash.

2.3 Parallel computing using multi GPU

For the parallel computing, the computational model is decomposed in two horizontal directions and each decomposed model is allocated to a different GPU. Because the values on the grid at the boundary of the neighbor decomposed models are necessary for the calculations, two grids from the boundary are overlapped each other and the values on these grids are exchanged by MPI (Message Passing Interface). Relative time required for the communication compared to the time for the calculation is longer for GPU than for CPU, because the calculation speed is much faster for GPU. Moreover, the overheads for the communication are larger for GPU because direct communications between GPUs are not available and values are transmitted to the target GPU via CPU using MPI. Therefore the time for the communication is not negligible and the concealment technique of the communication by overlapping the calculation and the communication is essentially important for achieve high performance parallel computation using GPU. Popular technique for concealing the communication is follows (Fig.3): Values on the overlapped grids are calculated first and then the communication of those values between neighbor decomposed models are performed during the calculation of rest grids. This technique is not efficient because it requires discontinuous memory accesses which are hard for GPU. Considering that our discontinuous grids have two regions having different size of grid spacing, exchanges of the values on the overlapped grid in one region are made during the calculation of another region(Fig.4). Our concealment technique makes it possible to avoid the discontinuous memory accesses.

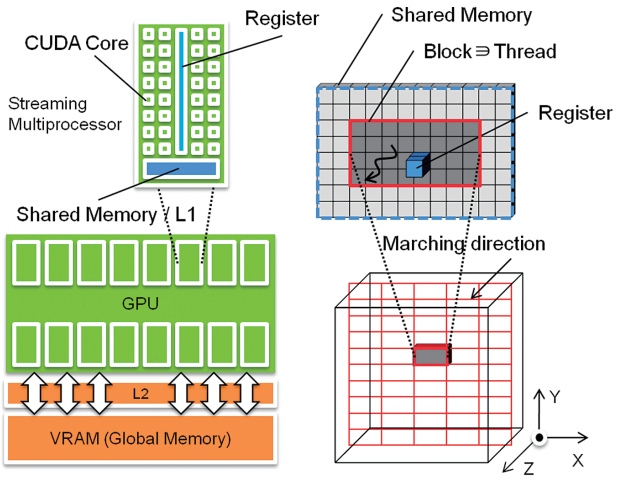


Fig. 2 (left) The architecture of GPU. (right) Calculation on GPU. Data on x-y plane and z direction are stored in sheared memory and register, respectively.

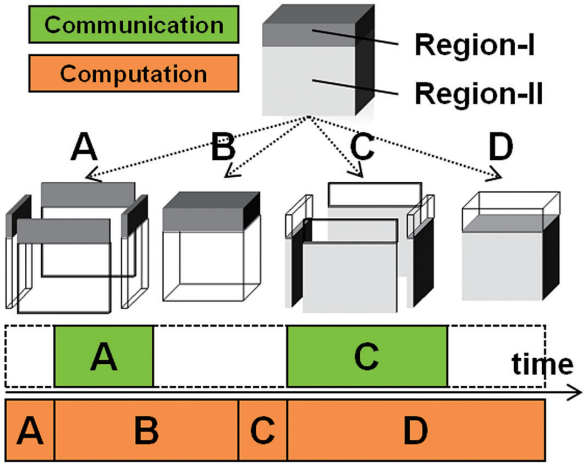


Fig. 3 Popular method for concealing the communication. The values on the overlapped grids are antecedently calculated to other grids.

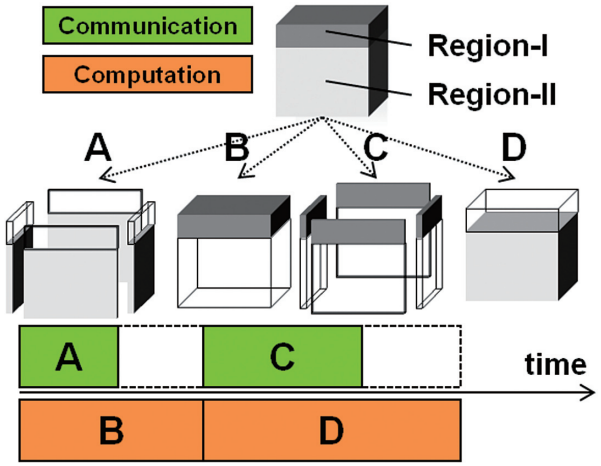


Fig. 4 Proposed method for concealing the communication. The values in one region are exchanged during the calculations of another region.

2.4 Performance test

We examine performance tests of the GMS on GPU using TSUBAME2.0. Each node has two six-core Intel Xeon X5670 2.93GHz CPUs and three NVIDIA Tesla M2050 GPUs and is connected with InfiniBand 4x QDR to each other. At first, we compared a performance for a single GPU computing with that for a single CPU core computing. We used a small model, unit420, which has about 22 million grids (420 × 420 grids for the two horizontal directions in Region I). Note that the floating point operation of the FD codes for GPU and CPU is a single-precision. The performance of the single GPU is faster than that of the single CPU core by a factor of 20.4 (Table 1). The maximum memory transfer rate between GPU and VRAM became 70 % or more of the peak transfer rate. It means that performance is bounded by the memory transfer rate and not arithmetic processing. Since the memory transfer in the main part of computing is high enough to theoretical performance, the room of the further speed-up is not so large.

Next we examined the two kinds of performance tests for parallel computing; weak and strong scaling tests. For the weak scaling test where the model sizes (number of grids) are increased in proportion to the degree of parallelism (number of GPUs), the result showed almost perfect linearity up to the simulation with 1024 GPUs (Fig.5). Here we used the unit420 as the unit model and the model size for the 1024 GPUs case is about 22 billion grids (unit13440). On the other hand, for the strong scaling test, we used a small model, unit420, and the model size is independent from the degree of parallelism. The speed-up is 3.2 and 7.3 for 4 and 16 GPUs cases, respectively (Fig.5). The reason of the decrease of the parallel performance is that the communication time increases so that it is no longer possible to conceal by calculation time. Moreover, the number of the threads for each GPU decreases because the model size allocated to each GPU becomes too small. Considering that the time steps for most model we use for simulation are up to hundreds thousand, the turn around times are several minutes to a few hours when the GPU resources appropriate to the size of the model are available. Thus, the performance of the GMS on GPU is practically satisfactory for most cases.

The calculation speed for unit13440 reaches to 79.7 TFLOPS which is about 34 times faster than the CPU calculation using the same number of cores (Fig.6).

Code	Execution time (sec)#	FLOPS (FP32)	Speed-up
GPU code	5.48E+01	7.95E+10	20.42
CPU code	1.12E+03	3.89E+09	1.0

Only FD calculation time is taken into account.

Table1 Comparison of performance of single GPU vs. CPU for Unit420, 1000 time steps

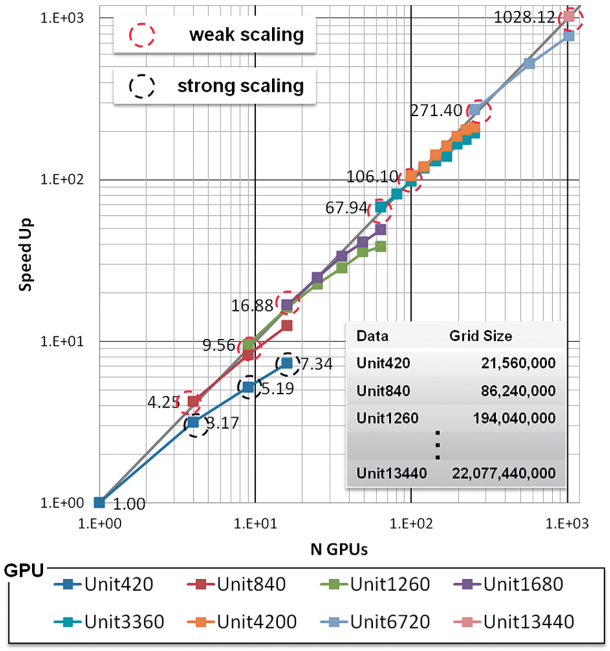


Fig. 5 Scalability of the parallel computing on GPU. Each color shows the result of performance test using different size of the model. Red and blue circles show the weak scaling and strong scaling, respectively.

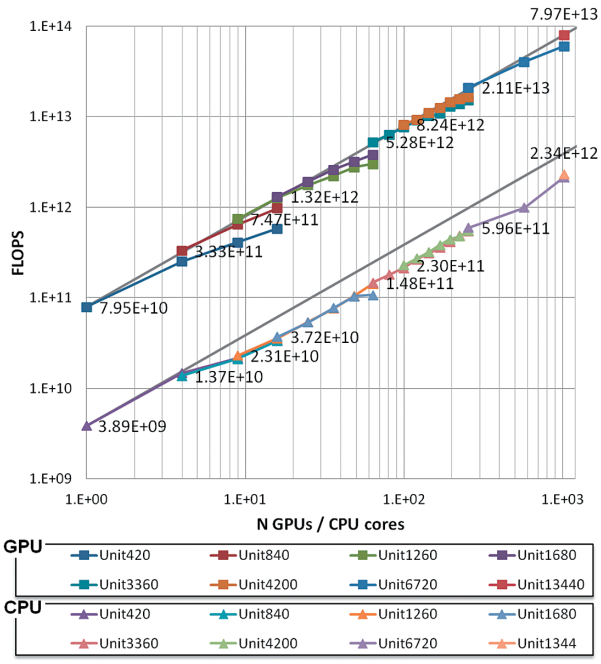


Fig. 6 Flops of the parallel computing on GPU (squares) and CPU (triangles). Each color shows the result of performance test using different size of the model.

Long-period ground motion simulation of large subduction zone earthquake 3

Using the newly developed FD code, we simulated long-period ground motions of the hypothetical large earthquake in the Nankai Trough. Because a seismic source area of the Nankai Trough earthquake extends over wide area from the Suruga bay to the Hyuga-nada[6], we modeled huge area (950km in north-south, 1150km in east-west, and 100km in depth) for the FD simulation. To accommodate details of the 3-D subsurface structure model[7] into the FD model, we used a fine grid spacing of 200m in horizontal and of 100m in vertical for region shallower than 8km that contain low-velocity sedimentary layers, and a three times coarser grid spacing for deeper region. The total number of grid points is about 3.2 billion. A valid period of the FD simulation is more than two seconds. As for a seismic source model, we used 430 thousand point sources to model a source process that a fault rupture propagates on a fault plane radially from the hypocenter at an assumed rupture velocity and breaks asperities where seismic waves are strongly radiated. The FD simulation was performed using 81 GPUs (27 nodes) of TSUBAME2.5. Three-component (north-south, east-west, and up-down) velocity waveforms for 600 seconds (72000 steps) at about 80 thousand points are output to files (file size of 1 component is about 90GB).

Figure 7 shows distribution maps of peak velocity on the ground surface calculated from different 10 source models. These source models have the same source area and asperity configuration but have different hypocenters. From these figures, we can understand that the hypocenter location has strong impact on the ground motion distribution and that the large long-period ground motions can be observed even at long-distance basins like the Kanto plain in these cases. The calculation times of these models are about 2.5 hours, while the calculation time for the same model using 252 CPU cores (Intel Itanium 1.66GHz) is about 43 hours. The practical seismic hazard assessment based on a large number of FD simulation is required for large earthquakes in order to consider a variety of earthquake occurrence pattern. Such simulations have been difficult due to a limitation of computational resources. However, it is now implemented by using TSUBAME.

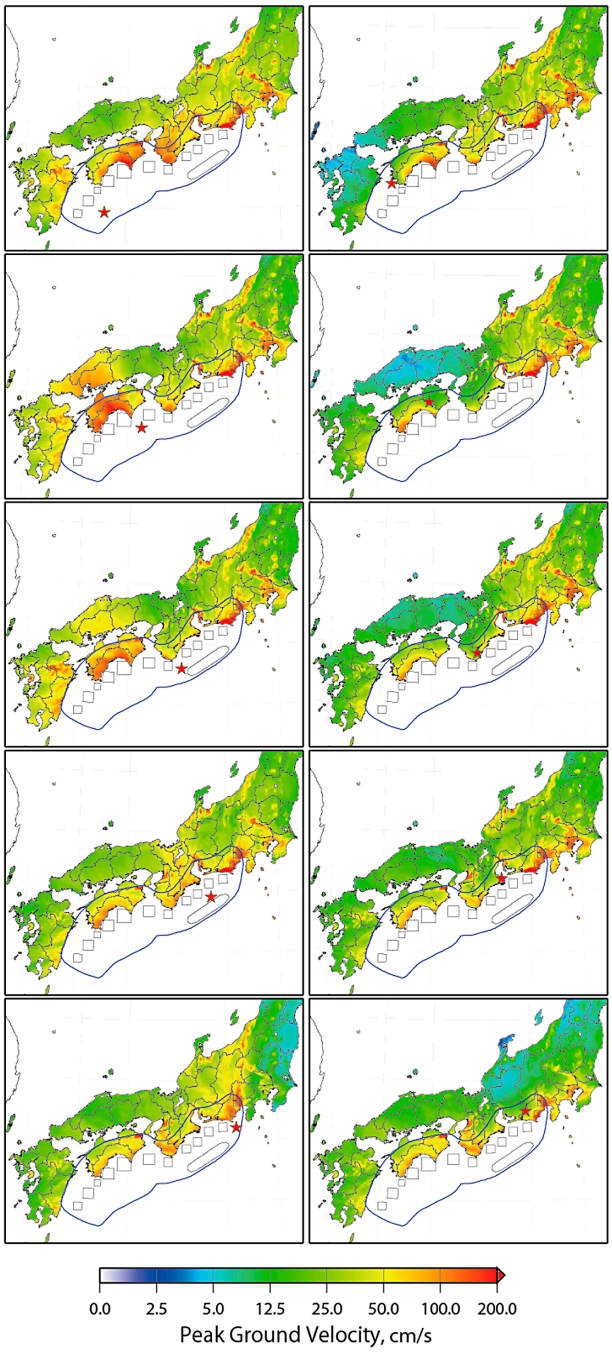


Fig. 7 Distribution maps of peak ground velocity value for the hypothetical Nankai Trough mega earthquakes using different 10 source models. The source models have the same source area (surrounded by blue lines) and asperities (black boxes), but have different hypocenter (red stars).

Conclusion 4

3-D FDM is one of the key techniques for a wave propagation simulation. We have developed the simulation code for GPGPU using CUDA based on the solver of GMS which is a total system for seismic wave propagation simulation. The computational model is decomposed in two horizontal directions and each decomposed model is allocated to a different GPU. For a high performance GPU calculation, we have proposed a new efficient concealing technique that successfully avoids the discontinuous memory accesses.

The performance test for the strong scaling using the model with about 22 million grids achieved 3.2 and 7.3 times of the speed-up by using 4 and 16 GPUs. The weak scaling test showed almost perfect linearity up to the simulation with 1024 GPUs. The calculation speed for the model with about 22 billion grids reaches to 79.7 TFLOPS which is about 34 times faster than the CPU calculation using the same number of cores.

Considering of the variety of occurrence pattern of large earthquakes has become an important issue in the seismic hazard assessment. Therefore, it is required high-accurate and large-scale seismic wave propagation simulations for many possible occurrence patterns which have been difficult due to the limitation of computational resources. We believe that the GPU adapted GMS will contribute to a progress in the seismic hazard assessment.

Acknowledgements

In this study, we used TSUBAME by supports from "Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures" and "High Performance Computing Infrastructure" in Japan. A part of the study was supported by the Support Program for Long-Period Ground Motion Hazard Maps of the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

References

[1] S. Aoi, N. Nishizawa, and T. Aoki, "3-D wave propagation simulation using GPGPU," Programme and Abstracts, Seismological Society of Japan 2009 Fall Meeting. A12-09, (2009) (in Japanese).

[2] D. Michéa, and D. Komatitsch, "Accelerating a three-

dimensional finite-difference wave propagation code using GPU graphics cards," Geophysical Journal International. vol. 182, pp. 389-402, (2010).

[3] T. Okamoto, H. Takenaka, T. Nakamura, and T. Aoki, "Accelerating large-scale simulation of seismic wave propagation by multi-GPUs and three-dimensional domain decomposition," Earth, Planets and Space. vol. 62, pp. 939-942, (2010).

[4] S. Aoi, and H. Fujiwara, "3-D finite difference method using discontinuous grids," Bulletin of the Seismological Society of America. vol. 89, pp. 918-930, (1999).

[5] S. Aoi, T. Hayakawa, and H. Fujiwara, "Ground motion simulator: GMS," Butsuri-Tansa. 57, pp.651-666, (2004) (in Japanese with English abstract).

[6] Earthquake Research Committee, "On the long-term evaluation of earthquakes in the Nankai Trough (2nd edition)," http://www.jishin.go.jp/main/chousa/13may_nankai/index.htm, (2013) (in Japanese).

[7] Earthquake Research Committee, "Long-period ground motion hazard maps for Japan," http://www.jishin.go.jp/main/chousa/12_choshuki/index.htm, (2012) (in Japanese).

● **TSUBAME e-Science Journal vol.11**

Published 6/20/2014 by GSIC, Tokyo Institute of Technology ©
ISSN 2185-6028
Design & Layout: Kick and Punch
Editor: TSUBAME e-Science Journal - Editorial room
Takayuki AOKI, Thirapong PIPATPONGSA,
Toshio WATANABE, Atsushi SASAKI, Eri Nakagawa
Address: 2-12-1-E2-6 O-okayama, Meguro-ku, Tokyo 152-8550
Tel: +81-3-5734-2085 Fax: +81-3-5734-3198
E-mail: tsubame_j@sim.gsic.titech.ac.jp
URL: http://www.gsic.titech.ac.jp/

TSUBAME

International Research Collaboration

The high performance of supercomputer TSUBAME has been extended to the international arena. We promote international research collaborations using TSUBAME between researchers of Tokyo Institute of Technology and overseas research institutions as well as research groups worldwide.

Recent research collaborations using TSUBAME

1. Simulation of Tsunamis Generated by Earthquakes using Parallel Computing Technique
2. Numerical Simulation of Energy Conversion with MHD Plasma-fluid Flow
3. GPU computing for Computational Fluid Dynamics

Application Guidance

Candidates to initiate research collaborations are expected to conclude MOU (Memorandum of Understanding) with the partner organizations/ departments. Committee reviews the “Agreement for Collaboration” for joint research to ensure that the proposed research meet academic qualifications and contributions to international society. Overseas users must observe rules and regulations on using TSUBAME. User fees are paid by Tokyo Tech’s researcher as part of research collaboration. The results of joint research are expected to be released for academic publication.

Inquiry

Please see the following website for more details.
http://www.gsic.titech.ac.jp/en/InternationalCollaboration

GSIC
Global Scientific Information and Computing Center