Practical Use of FX10 Supercomputer System (Oakleaf-FX) of Information Technology Center, The University of Tokyo

● Yoshio Sakaguchi ● Takahiro Ogura

Information Technology Center, The University of Tokyo (ITC/UT) is a core organization of "Joint Usage/Research Center for Interdisciplinary Large-Scale Information Infrastructures" which consists of eight academic supercomputer centers in Japan. There, Oakleaf-FX, a system centered on Fujitsu's PRIMEHPC FX10 (76 800 CPU cores), which has been developed by further enhancing the technology applied to the K computer to realize both high performance and superior energy efficiency, started operations in April 2012 and achieved a performance of 1.04 PFLOPS (ranked 18th in the world) in June 2012. Supercomputing Division of ITC/UT provides services for operations of supercomputer systems and supports more than 1500 users from both inside and outside the university in various cutting-edge R&D and education activities. Fujitsu is working to develop products with high application execution performance, and create user-friendly systems that can work in various operational environments and that are energy efficient. This paper presents a system overview of Oakleaf-FX and the issues to be overcome to develop supercomputers, together with measures for solving them. The results of various studies making use of Oakleaf-FX have been reported and this paper also touches on four fields as successful examples of its application.

1. Introduction

Ever since its establishment as the Supercomputing Center (as it was known before 1999), a facility open to all scholars in Japan in 1965, Information Technology Center, The University of Tokyo (ITC/UT) has conducted operations promoting application to education and cutting-edge research in a variety of science and technology fields by using supercomputers. ITC/UT started operations of Oakleaf-FX, a large-scale massively parallel supercomputer, in April 2012 not only to meet the diverse and important needs for continuous support for existing supercomputer users, pioneering of new fields of application, improvement of the development environment for use of the K computer and human resources development, but also to offer a state-of-the-art supercomputer environment to more than 1500 users inside and outside the university including corporations. Oakleaf-FX is a system centered on Fujitsu's PRIMEHPC FX10, a supercomputer developed by further improving on Fujitsu's supercomputer technology employed in the K computer, which achieved the world's highest performance, to realize superior energy efficiency as well as high performance and reliability.

This paper presents an overview of Oakleaf-FX, issues to be overcome to develop supercomputers together with approaches to their solutions and successful examples of application of Oakleaf-FX.

2. Overview of Oakleaf-FX

Oakleaf-FX is composed of 50 racks of PRIMEHPC FX10 (4800 compute nodes) as the core of a large-scale massively parallel supercomputer system and, as sub systems, 74 FUJITSU Server PRIMERGY PC servers, 234 units of FUJITSU Storage ETERNUS and network equipment (**Figure 1**). The software includes HPC middleware Technical Computing Suite that supports petascale systems and its component Fujitsu Exabyte File System (FEFS), a high-capacity, high-performance and high-reliability parallel distributed file system.

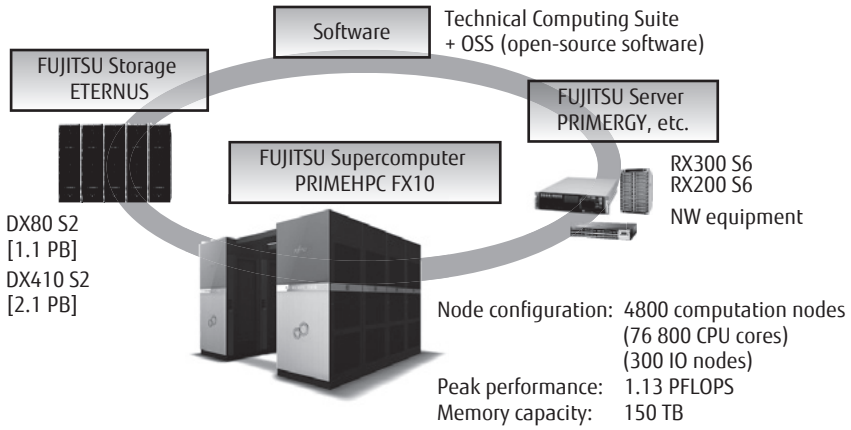


Figure 1
Oakleaf-FX overview.

Oakleaf-FX is a large-scale massively parallel supercomputer with a peak performance exceeding 1 PFLOPS and is utilized in wide-ranging research fields including earth science, cosmophysics, seismology, climatic modeling, material science, energy, biology, fluid dynamics and solid mechanics. It is also applied to high-performance computing (HPC) education for the development of human resources that will lead computational science.

3. Issues to be overcome to develop supercomputers

Until around 2004, improvement of supercomputer performance was achieved by increasing the clock frequency. Now, however, performance is improved by increasing the number of CPU cores, or by using many-core architectures. The supercomputer performance trends based on the top 10 systems in the TOP500 list[1] (ranking of the world's fastest supercomputers) show that the total theoretical computing performance has made rapid progress of 156% since 2010 but the CPU frequency has remained mostly unchanged at 99.8% since 2010 (**Figure 2**). Performance improvement by many-core architectures requires users to make changes to the programs or develop new programs according to the system for enhancing the program execution performance. From the perspective of providing a permanent-system-use environment, achieving high execution performance and reducing the development burden of massively parallel applications pose challenges.

With multiple methods of use for many different users and applications, efficient and flexible operations of systems are demanded of supercomputer center operations. In terms of use, Integrated Development Environment (IDE) that provides program development and improvement of open-source software (OSS) environments, which are increasingly in use globally as well as in Japan, are required.

The ultra-high-density packaging technology, which contributes to performance improvement of supercomputers, has aspects including no footprint increase in spite of the requirement for performance improvement and the demand for reduction of power consumption that has been revealed. One reason for this is that the tight energy supply and demand situation due to the impact of the Great East Japan Earthquake in March 2011 required energy efficiency from an ethical point of view as well. Accordingly, there are issues such as reduction of power consumption, visualization of energy for verifying the effect of such reduction, and monitoring and other operational issues.

4. Issue-solving approaches of Oakleaf-FX

4.1 Realization of high execution performance of applications

PRIMEHPC FX10 is equipped with a high-speed interconnect (Tofu interconnect) based on a six-dimensional mesh/torus architecture and Virtual Single Processor by Integrated Multicore Parallel Architecture

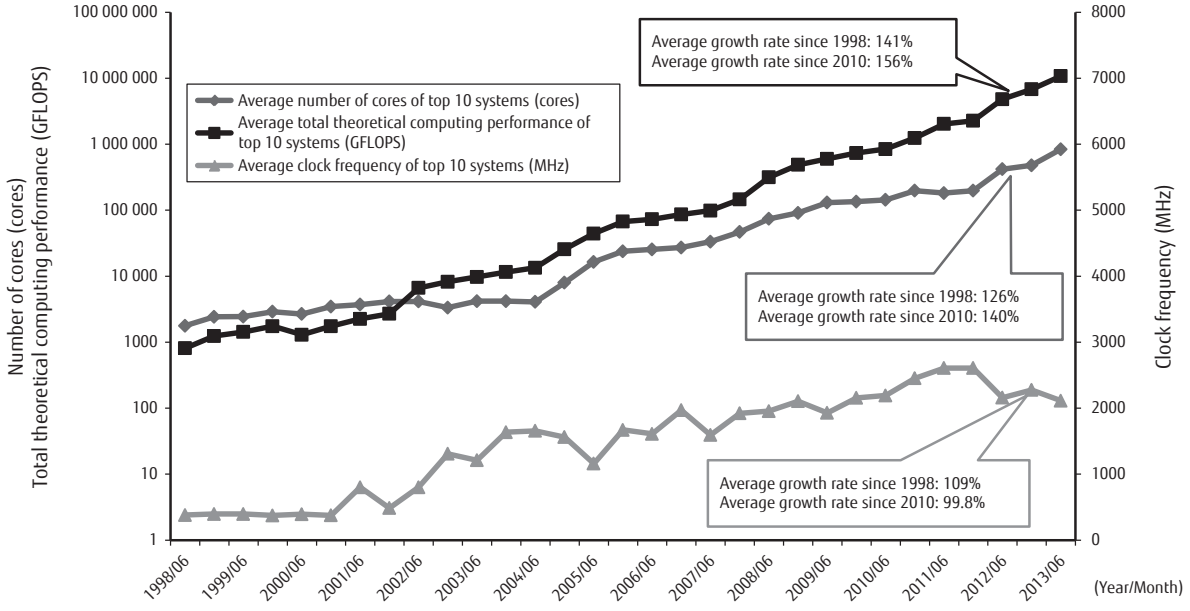


Figure 2
Trends of performance and scale of supercomputers.

(VISIMPACT), which realizes massive parallelization of applications.

Oakleaf-FX is a system composed of 76 800 CPU cores that allows large-scale massive parallelism and hybrid parallelism combining MPI parallelism, and thread parallelism is an effective parallelization technique that achieves high execution performance of applications.

VISIMPACT, a mechanism used to easily realize hybrid parallelization, reduces the application development burden on users and provides a suitable application development environment in view of petascale computing in the future. VISIMPACT is realized by combining the following functions.

- Hardware barrier mechanism between CPU cores
- 12 MB shared L2 cache
- Automatic parallelization function of compilers

High execution performance of a large-scale parallelism has been demonstrated by the result achieved with a High-Performance Linpack (HPL) benchmark program at 1.04 PFLOPS and execution efficiency of 91.86%. It ranked 18th in the world in the TOP500 list announced in June 2012 and, in comparison between the top 50 systems in terms of execution efficiency, came second in the world next to the K computer.

4.2 Provision of user-friendly systems and flexible operational environments

The supercomputer operations of ITC/UT have provided a variety of services including corporate use, educational use and use by young researchers in addition to the existing users from the perspectives of social contribution and human resources development. For Oakleaf-FX, ITC/UT and Fujitsu have worked on system design and tool development in order to create easy-to-understand systems that suit the purposes of various users, and user-friendly systems are offered to them.

1) Improved convenience by token system

For billing of Oakleaf-FX, a system is offered to users in which nodes can be used freely within the scope of the token assigned (number of nodes used × elapsed time), rather than the conventional system of limiting users to the maximum number of nodes that can be used.

2) Provision of user-friendly debugging/tuning environments

Consideration is given to ensure smooth flows of jobs for work such as program debugging/tuning, and interactive job and debugging environments are provided.

3) Realization of flexible operational environments

For educational use, a deadline scheduling system is developed and operated in which nodes are

efficiently assigned only during specified periods such as lecture hours. Furthermore, titled "large-scale HPC challenge" is also starting. Users of accepted proposal can occupy entire computing nodes (4800 nodes) of Oakleaf-FX for 24 hours to achieve research results.

4) Provision of rich open-source software

Oakleaf-FX widely provides an OSS, such as one containing frequently used applications, mathematical libraries and tools, to help facilitate users' research.

4.3 Green supercomputer combining high performance and energy efficiency

Oakleaf-FX achieved a system power consumption of 1176.8 kW combined with a performance of 1.043 PFLOPS in the Green 500 list[2] (ranking of the most energy-efficient supercomputers in the world) in June 2012 and ranked 43rd in the world with a power/performance ratio of 886.3 MFLOPS/W.

With the recent power circumstances taken into account, visualization of power is important in supercomputer operations. For that purpose, we have developed, introduced and operated a power monitoring system capable of real-time gathering of power consumption-related data for each subsystem and power usage prediction monitoring (demand monitoring). The power monitoring system can visualize power for each subsystem of Oakleaf-FX, unlike the conventional way of grasping the overall power usage based on the power equipment information (**Figure 3**). As a result, effect of fallback operation in response to a power-saving request, etc. can be more specifically understood and e-mail communication with regards to the tendency of approaching the permissible upper limit has been made possible by the demand monitoring function.

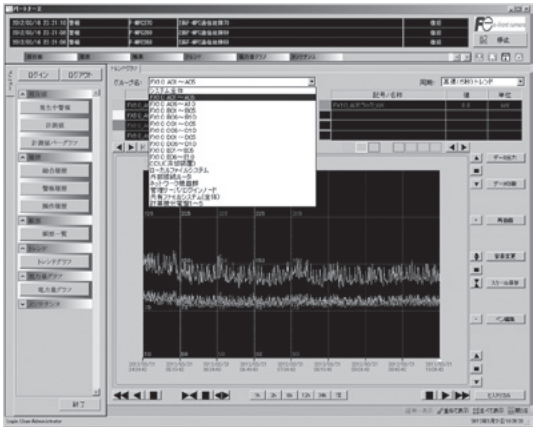


Figure 3
Power monitoring system screen.

5. Application Examples of Oakleaf-FX

There are 37 reported achievements of research (paper/poster presentations) using Oakleaf-FX made in the fields of physics, information science, applied physics (engineering fundamentals), earth and planetary science and mechanical engineering in the period between April and December 2012.

In the Graph 500 list[3] (ranking of supercomputers with the highest graph processing performance in the world) of June 2012, Associate Professor Toyotaro Suzumura at the Tokyo Institute of Technology achieved the world's fourth best performance by using Oakleaf-FX.

The ITC/UT has held a total of eight seminars between July 2012 and August 2013 for engineers of companies aiming to acquire skills in high-performance computing and parallel processing, in addition to users inside and outside the university who are considering using Oakleaf-FX, as part of its social contribution. In the seminars, trial accounts are given to allow people to experience parallel programming by actually using the supercomputer. In this way, Oakleaf-FX is not only used in a wide variety of research fields but also used for human resource development and social contribution.

The following presents four examples of achievement made by using HPC challenge operations and hybrid parallelism of Oakleaf-FX.

5.1 3D groundwater flow through heterogeneous porous media (Professor Kengo Nakajima, ITC/UT)[4)–6)]

Poisson equation is derived from finite-volume discretization of groundwater flow through heterogeneous porous media, and is solved by a conjugate gradient iterative solver with multigrid preconditioner (MGCG), which is scalable and suitable for large-scale problems (**Figure 4**). The 4096 nodes (65 536 CPU cores) of Oakleaf-FX are used for solving a large-scale linear equation with up to 17 179 869 184 unknowns in 10 seconds. It takes more than 100 seconds if using a conventional preconditioned CG method, such as ICCG (CG preconditioned by incomplete Cholesky factorization). Recent improvement of the parallel algorithm

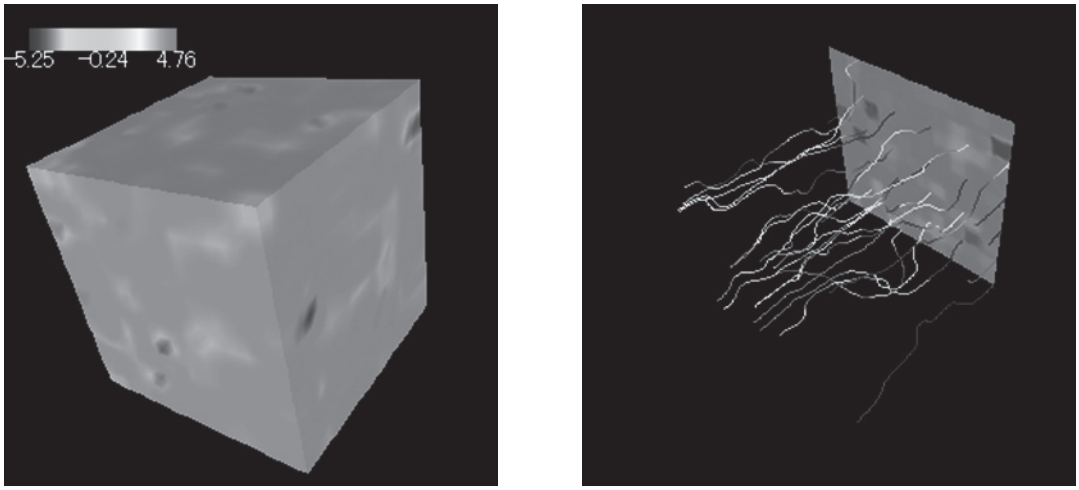


Figure 4
Result of simulation of 3D groundwater flow through heterogeneous porous media.

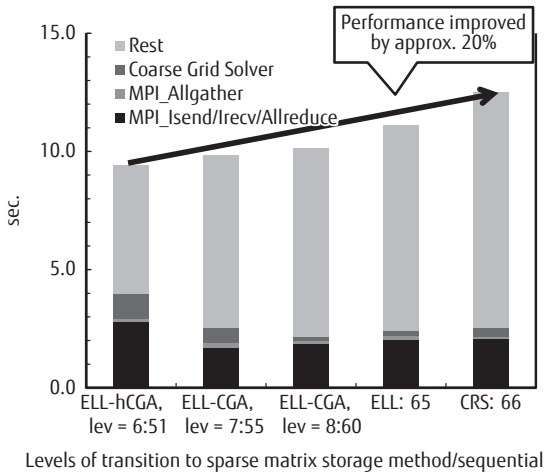


Figure 5
Parallel performance of MGCG solver on Oakleaf-FX.

for MGCG solver provided 20% or more improvement in the performance compared to that of the previous work. Performance of the application program based on the new algorithms of MGCG solver was improved by more than 20% (**Figure 5**).

5.2 Evaluation of effect of inflow from Atlantic on Arctic sea ice distribution (Professor Hiroyasu Hasumi, Atmosphere and Ocean Research Institute, The University of Tokyo)[7),8)]

Changes in oceanic sea ice fields from hour to hour are obtained by using a finite difference method to discretize and time-integrate a coupled system consisting of equations of motion of seawater represented by parameterizing phenomena of a few km or less horizontally, transport and diffusion equations of heat and salt and dynamic-thermodynamic equations of sea ice approximated as a continuum with viscoplastic rheology. While the computational domain is all the world's oceans, a grid system is used that intensively increases the resolution of the Atlantic Ocean side of the Arctic Ocean (**Figure 6**). The number of spatial grids is 1280 × 768 × 45 and use of Oakleaf-FX 128 nodes allows a one-year simulation data set to be obtained in about five hours. An oceanic sea ice field simulation data set for about 50 years is created to evaluate the behavior of high-temperature seawater flowing from the Atlantic Ocean into the Arctic Ocean and its effect on sea ice distribution (**Figure 7**).

5.3 Gyrokinetic simulation of multi-scale plasma turbulence (Dr. Shinya Maeyama, Japan Atomic Energy Agency)[9)]

Time evolution of turbulence in magnetically

confined plasma is numerically solved based on gyrokinetic equations (convection-diffusion equations of plasma distribution functions in a five-dimensional phase space). Examples of turbulent electrostatic potential are plotted in **Figure 8**. Spatio-temporal discretization is yielded by means of spectral/finite difference methods in conjunction with an explicit time integration method, which is parallelized by multi-dimensional domain decomposition using MPI and multiple threads using OpenMP. A typical problem size is 10^9 to 10^{10} grid points and 10^5 time steps. Use of Oakleaf-FX 512 nodes (8192 CPU cores) allows computation in about 48 hours per case. The newly developed optimized MPI-process mapping and computation-communication overlap techniques have improved single-node computing performance as well as parallelization efficiency for strong scaling, as shown in **Figure 9**.

5.4 Magnetohydrodynamics/radiation hydrodynamics simulation of black hole accretion flows (Professor Ryoji Matsumoto, Graduate School of Science, Chiba University)[10)]

The time evolution of a rotating disk (accretion disk) formed by matter accreting to a black hole

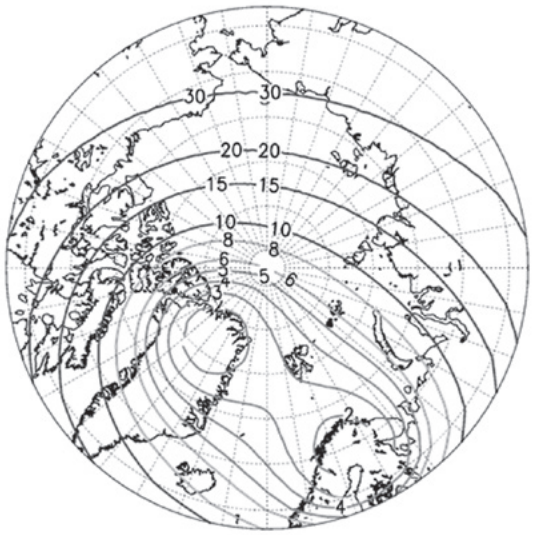


Figure 6
Result of simulation of oceanic sea ice fields.

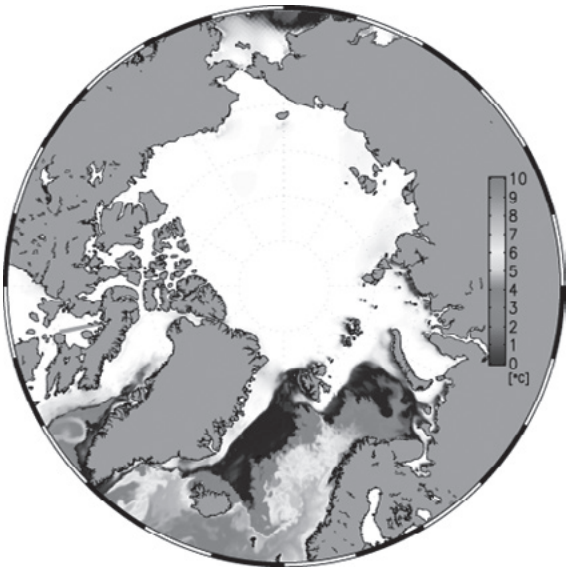
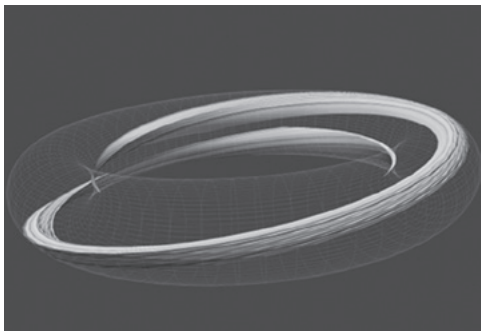


Figure 7
Result of evaluation of behavior of high-temperature seawater flowing from Atlantic Ocean into Arctic Ocean and its effect on sea ice distribution.



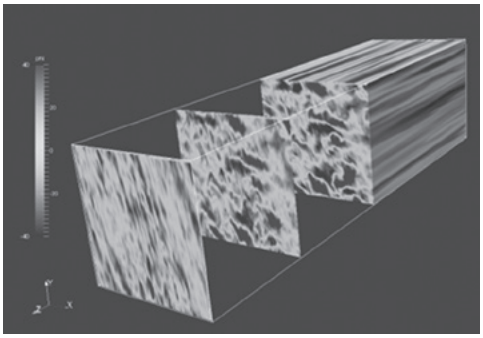


Figure 8
Examples of turbulent electrostatic potentials in magnetically confined plasma.

is studied by magnetohydrodynamic simulations. **Figure 10** shows a result of a numerical simulation of a gas cloud infalling toward the Galactic center supermassive black hole, whose mass is four million times the solar mass. We studied how the accretion rate increases when the gas cloud collides with an accretion disk surrounding the black hole. The numerical simulations were carried out by using a three-dimensional magnetohydrodynamic code in cylindrical coordinates based on the HLLD approximate Riemann solver. In order to reduce the numerical dissipation of currents, a spatially fifth-order MP5 scheme is implemented when computing the cell interface values. The number of mesh points is 256 × 64 × 256. Typical simulations carried out by using 64 nodes (1024 CPU cores) of Oakleaf-FX take 100 hours. **Figure 11** shows a result of numerically simulating the formation of a radiation pressure driven outflow when the accretion rate exceeds the threshold for the formation of a radiation pressure dominant disk. Axial symmetry is assumed in this simulation. For numerical computations, a relativistic, resistive radiation magnetohydrodynamic (R3MHD) code in cylindrical coordinates has been used. In this code, the 0-th and 1st moment equations of the radiation transfer equations coupled with the magnetohydrodynamic equations are solved. The code has been parallelized by using MPI/OpenMP. An effective efficiency of 10% has been attained by using 64 nodes (1024 CPU cores) of Oakleaf-FX. One future target is to carry out 3D relativistic radiation magnetohydrodynamic simulations before the competing groups in the world.

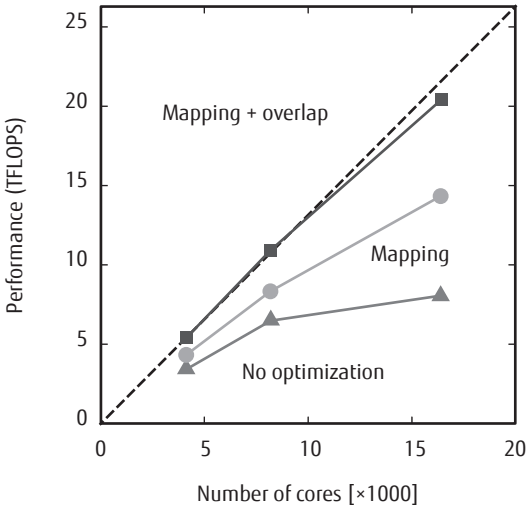


Figure 9
Result of strong scale parallel computing of gyrokinetic equations on Oakleaf-FX.

6. Future outlook

In March 2013, Center for Computational Sciences at the University of Tsukuba and Information Technology Center, The University of Tokyo established a new organization, called the Joint Center for Advanced High Performance Computer (JCAHPC), in order to design, operate and manage a next-generation supercomputer system provided by both universities.

JCAHPC promotes advanced computational sciences, and contributes to the fields of academia and science technology. In this effort, the supercomputer system of the two universities is installed at

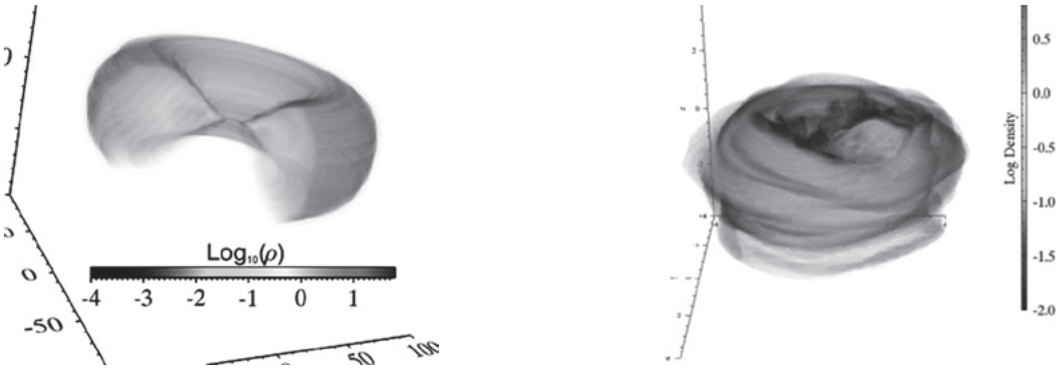


Figure 10
Result of simulation using 3D magnetohydrodynamics code in cylindrical coordinates based on HLLD scheme.

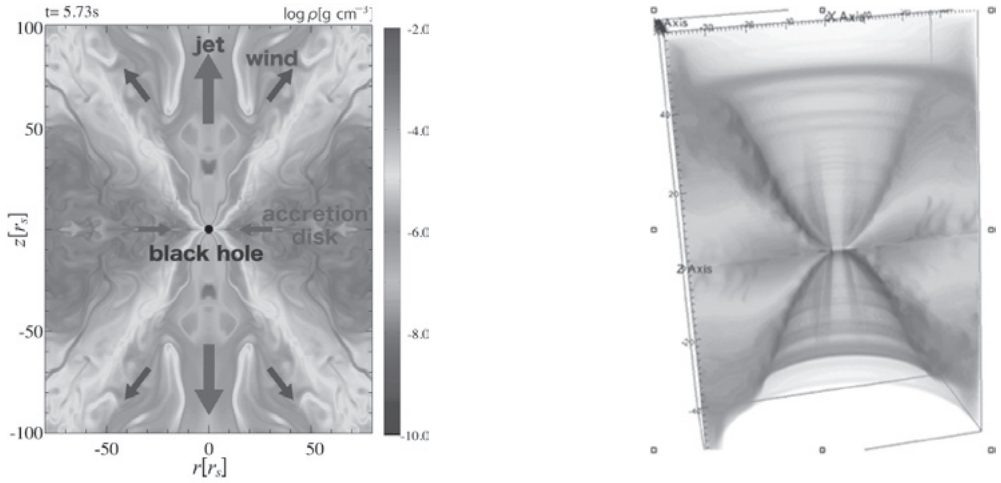


Figure 11
Result of simulation of formation process of jet accelerated by radiation pressure with high rate of accretion to black hole on assumption of axial symmetry.

the Advanced HPC Facility (Kashiwa Campus of The University of Tokyo) for installation and joint operation of the state-of-the-art, large-scale, high-performance computing infrastructure. It is an unprecedented challenge and is expected to promote leading-edge computational science and contribute to the promotion of scholarship and science and technology of Japan.

We intend to make use of the hardware technology that has achieved both the high performance and energy efficiency of Oakleaf-FX, software technology for the pursuit of stable operation and user-friendliness and operational know-how from the perspective of users to help maximize the future achievements of JCAHPC.

7. Conclusion

This paper has presented a system overview of Oakleaf-FX of ITC/UT, measures for solving issues and examples of application together with their results. In the future, we hope to make contributions to the promotion and advancement of science of Japan and the world by creating more versatile supercomputer systems.

We would like to take this opportunity to extend our deepest gratitude to the following persons for their provision of a large amount of application information:

Professor Kengo Nakajima, Information Technology Center, The University of Tokyo

Professor Hiroyasu Hasumi, Atmosphere and Ocean Research Institute, The University of Tokyo

Dr. Shinya Maeyama, Japan Atomic Energy Agency

Professor Ryoji Matsumoto, Faculty of Science, Chiba University

References

1) TOP500. *http://www.top500.org/*
2) Green500. *http://www.green500.org/lists/green201206*
3) Graph 500. *http://www.graph500.org/results_june_2012*
4) K. Nakajima: OpenMP/MPI Hybrid Parallel Multigrid Method on Fujitsu FX10 Supercomputer System. IEEE Proceedings of 2012 International Conference on Cluster Computing Workshops, 2012, pp. 199–206.
5) K. Nakajima: Large-scale Simulations of 3D Groundwater Flow using Parallel Geometric Multigrid Method. *Procedia Computer Science*, Vol.18, pp. 1265–1274 (2013).
6) K. Nakajima: Challenges in Sparse Linear Solvers for Post-Peta/Exascale Systems, Latest Advances in Scalable Algorithms for Large Scale Systems. ISC'13 (International Supercomputing Conference), Leipzig, Germany, 2013.
7) T. Kawasaki et al.: High-resolution modeling study on the Atlantic water inflow to the Arctic Ocean. IAHS-IAPSO-IASPEI Joint Assembly, Goteborg, Sweden, 2013.
8) T. Kawasaki et al.: High resolution modeling for long-term prediction of the Arctic sea ice. NIPR Symposium, Tokyo, Japan, 2012.
9) S. Maeyama et al.: Massively-parallelized spectral calculations of the gyrokinetic simulation code GKV on the

K computer. HPCS 2013, Tokyo, Japan, P1–1, 2013 (in Japanese).
10) R. Matsumoto: X-ray Flares Triggered by the Tidal Disruption of a Gas Cloud Approaching the Galactic Center Supermassive Black Hole. 12th International Workshop on the Interrelationship between Plasma Experiments in Laboratory and Space (IPELS), Hakuba, Japan, July 1–5, 2013.

Yoshio Sakaguchi
Fujitsu Ltd.
Mr. Sakaguchi is currently engaged in HPC business promotion in science and technology field and support for supercomputer center.

Takahiro Ogura
Fujitsu Ltd.
Mr. Ogura is currently engaged in operation support for Oakleaf-FX at ICT/UT.