International Journal of Networking and Computing – www.ijnc.org
ISSN 2185-2839 (print) ISSN 2185-2847 (online)
Volume 2, Number 2, pages 206–215, July 2012

All about RIKEN Integrated Cluster of Clusters (RICC)

Maho Nakata
Advanced Center for Computer and Communications,
2-1, Hirosawa, Wako-City, Saitama
351-0198, Japan

Received: May 7, 2012
Revised: June 19, 2012
Accepted: June 29, 2012
Communicated by Koji Nakano

Abstract

This is an introduction to the RIKEN's supercomputer RIKEN Integrated Cluster of Clusters (RICC), that has been in operation since August 2009. The basic concept of the RICC is to "provide an environment with high power computational resources to facilitate research and development for RIKEN's researchers". Based on this concept, we have been operating the RICC system as a (i) data analysis environment for experimental researchers, (ii) development environment targeting the next-generation supercomputer; i.e., the "K" computer, and (iii) GPU (graphics processing unit) computers for exploring challenges in developing a future computer environment. The total performance of RICC is 97.94 TFlops, ranking it as the 125th on the Top500 list in Nov. 2011. We prepared four job class accounts, based on the researchers' proposals prior to evaluation by our Review Committee. We also provided backup services to RIKEN's researchers, such as conducting RICC training classes, software installation services, and speed up and visualization support. To encourage affirmative participation and proactive initiation, all the services were free of charge; however, access to RICC was limited to researchers and collaborators of RIKEN. As a result, RICC has been able to maintain a high activity ratio ($> 90\%$) since the beginning of its operation.

Keywords: RIKEN, Super computer, GPU, PC cluster

1 Introduction

The RIKEN Integrated Cluster of Clusters (RICC), initiated in August 2009, is a super computer system managed by the Advanced Center for Computing and Communication (ACCC) of RIKEN [1]. The main purpose of the cluster is to support research activity of RIKEN's researchers by providing an efficient and high performance computer environment, which might not be affordable by just one laboratory. Our three aims are: (i) technical support for massive experimental data-processing, (ii) providing an environment for application and development of the next-generation supercomputer or the "K" computer, and (iii) exploring and adopting novel challenging computing technologies such as graphics processing unit (GPU). The RICC is the successor model of the RIKEN Super Combined Cluster (RSCC), the former super computer of RIKEN. The RSCC is in fact the first combined system incorporating several different computer resources as a single system in Japan.

The RICC consists of the Massively Parallel PC Cluster (8384 cores), the Multi-purpose Parallel PC Cluster (800 cores), the MDGRAPE-3 cluster (256 cores), and the Large Memory Capacity

Server (36 cores). The Massively Parallel PC Cluster (peak performance: 98.2 TFlops) has been boosted with 24 nodes in October, 2010. The total LINPACK performance of RICC then indicated 97.94 TFlops [2].

The special features of each cluster are as follows: (i) the Multi-purpose Parallel PC Cluster is equipped with an NVIDIA C1060 Graphics Processing Unit (peak performance: 9.3 TFlops + 93.3 TFlops [in single precision]); (ii) the Large Memory Server has 512 GBytes of memory; and (iii) the MDGRAPE-3 server is equipped with an MDGRAPE-3 (a special computer developed by RIKEN for molecular dynamics) with a peak performance of 3.0 TFlops + 64 TFlops. These clusters are interconnected by a Double Data Rate (DDR) InfiniBand Host Channel Adapter (HCA) installed with a 550TB high-speed magnetic disk device and 2 PBytes tape library system for the storage system.

We also provided backup services and relevant technological supports by conducting RICC Training Classes, application installation service, speed up/parallelization support, high-grade programming support and visualization support for RIKEN researchers.

In this paper, we briefly describe our RICC system, services, and management with the relevant sections organized as follows: the specification and configuration of the RICC system (Section 2); installed applications (Section 3); operation status (Section 4); details about our unique services (Section 5); and summary (Section 6).

2 The specification and configuration of the RICC system

The RICC is a successor of the former RIKEN super computer, the RSCC. The RSCC had replaced a vector computer type supercomputer in March 2004. Since then, we have been focusing on supporting all the RIKEN's researchers. The basic concept of the RICC is to "provide an environment with high power computational resources for research and development of RIKEN's researchers". Based on this concept, we have been providing (i) data analysis environment for experimental researchers; (ii) a development environment targeting the next-generation supercomputer, or the "K" computer; and (iii) GPU computers for exploring challenges in developing a future computer environment.

The RICC system consists of four subsystems : (i) the Massively Parallel PC Cluster (1048 nodes, 2096 CPUs, and 8384 cores; the peak performance: 98.2 TFlops); (ii) the Multi-Purpose PC Cluster (100 nodes, 200 CPUs and 800 cores with one NVIDIA Tesla C1060 per node; the peak performance: 9.4+93.3 TFlops), (iii) the MDGRAPE-3 cluster (32 nodes, 64 CPUs, 256 cores with 32 MDGRAPE-3 boards for molecular dynamics simulations, the peak performance: 3.1+64 TFlops); and (iv) the Large Memory Capacity Server (512 GB of shared memory, the peak performance: 239 GFlops).

The LINPACK benchmark result of the RICC system is 97.94 TFlops by achieving 92.36% of efficiency of double-precision operation. Note that we measured the LINPACK benchmark in August 2009, and added 24 nodes to the Massively Parallel PC Cluster in August 2010. Therefore, the LINPACK performance has since improved (vs initiation of operation).

The Massively Parallel PC cluster, the Multi-Purpose PC Cluster, and the MDGRAPE-3 cluster are interconnected by an Infiniband DDR. Thus, we can combine and use these four systems as a single system.

The storage system of the cluster consists of 550 TBytes of magnetic storage (hard disks), and two PBytes of tape archiving (tape backup). The peak electricity consumption is approximately 850kVA, and the gross heating value is 710Mcal/h.

For better understanding of our concept, the specification details (Table 1) with configuration (Figure 1) and relevant pictorial illustration (Figure 2), of the RICC system are shown accordingly.

3 Installed Applications

In Table 2, we list some of major applications and their abstracts installed on the RICC. We also provide popular compilers like the Intel Compiler suite, the PGI Accelerator Compiler, and the Fujitsu Compiler (Fortran, C, C++). New application programs can be installed upon request,

Table 1: Specification of the RICC system.

System	Machine Type	Specification
The Massively parallel PC Cluster (98.3 TFlops)	Fujitsu PRIMERGY RX200S5 (1048 nodes) (24 nodes add in Oct. 2010)	CPU: Intel Xeon 5570 (2.93GHz) × 2 Memory:12 GBytes, HDD 400 or 800 GBytes (RAID0, SAS)
The Multi-Purpose PC Cluster (9.4+93.3 TFlops(Single precision))	NEC Express 5800/56Xg (100 nodes)	CPU: Intel Xeon 5570 (2.93GHz)× 2 Memory:24 GBytes, HDD 250 GBytes Accelerator: NVIDIA Tesla C1060
The MDGRAPE-3 Cluster (3.1+64 TFlops)	SGI Altix XE 250 (32 nodes)	CPU: Intel Xeon 5472 (3.0GHz)× 2 Memory:32 GBytes, HDD 750 GBytes Accelerator: MDGRAPE-3
The Large Memory Capacity Server (239 GFlops)	SGI Altix XE 450 (1 nodes)	CPU: Intel Itanium 9140M (1.66GHz)× 18 Memory:512 GBytes, HDD 12 TBytes I/O PCI-X (MDGRAPE-3 is available)
Magnetic storage (550 TBytes)	File System: QFX+SRFS File Server: SPARC Enterprise M9000 RAID: Etermas 2000 Model200 × 24	
Tape archiving system (2 PBytes)	HSM: High Performance Storage System Core Server: System p570 × 1 Mover: System p570 × 6 Cache: DS4800 × 6 (20 TBytes) Tape: TS1040 × 12 (LTO Ultrium4) Library: TS3500(L53+D53+S54)	
Networking system	InfiniBand: X4 DDR	Switch: Qlogic SilverStorm 9024 ×60 and 9120 ×2 Topology: Fat-tree (bisection, bandwidth: 240 GBytes/s)
	Ethernet: 10GbE, GbE	Switch Cisco 6509-E X2
The peak electricity consumption		850kVA
The gross heating value		710Mcal/h

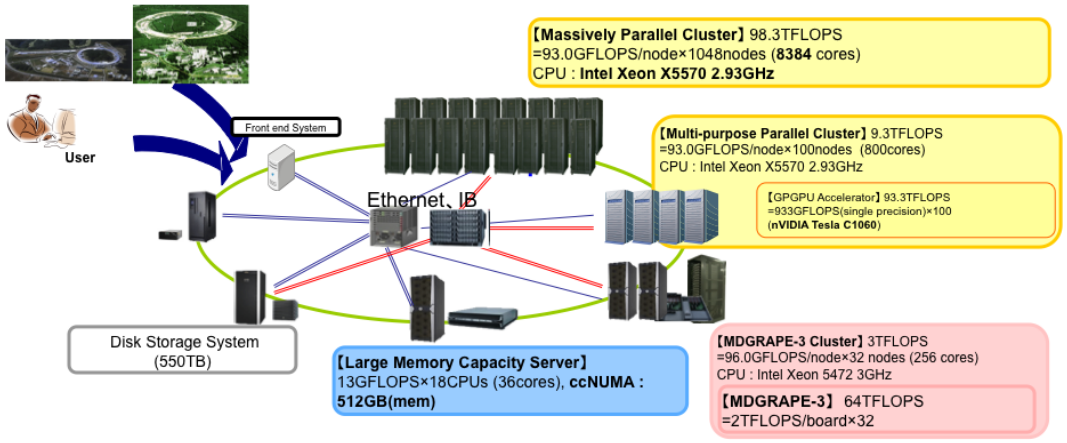


Figure 1: Configuration of the RICC system. In anticlockwise: 1) [Massively Parallel PC Cluster] 98.3 TFlops = 93.0 GFlops/node x 100 nodes (8384 cores) CPU: Intel Xenon X5570 2.93GHz, 2) [Multi-purpose Parallel PC Cluster] 9.3 = 93.0 GFlops/node x 100 nodes (800 cores), CPU: Intel Xenon X5570 2.93GHz, [GPU Accelerator] 93.3 GFlops = 93.3 GFlops (single precision) x 100 (NVIDIA Tesla C1060), 3) [MDGRAPE-3] 3 TFlops = 96.0 GFlops/node x 32 nodes (256 cores), CPU: Intel Xenon 5472 3GHz, [MDGRAPE-3] 64 TFlops = 2 TFlops/board x 32, 4) [Large Memory Capacity Server] 13 GFlops x 18 CPUs (36 cores), ccNUMA: 512 GB (memory), 5) Disk Storage System (550 TBytes), and 6) Front end System

The Massively Parallel PC Cluster

Air conditioner

The Multi-Purpose Parallel PC Cluster

The MDGRAPE-3 and the Large Memory Capacity Server

Figure 2: Pictorial illustrations of the RICC system.

Table 2: Installed applications and their abstracts on the RICC system.

Software	Abstract
ADF	A quantum chemistry software package based on Density Functional Theory (DFT)
AMBER	A set of molecular mechanical force fields for biomolecules simulations
ANSYS	Computer-aided engineering and engineering design analysis software
GAMESS	A computational chemistry software program (Hartree-Fock, DFT, GVB, and MCSCF)
Gaussian03, 09	A computational chemistry software program
GaussView	Graphical User Interface for Gaussian
GOLD	Protein-ligand docking
GROMACS	A molecular dynamics package designed for biomolecular systems
Hermes	A 3D visualizer with particular emphasis on functionality for the analysis of protein-ligand interactions. (for SuperStar, GOLD Suite, Relibase+)
Jmol	An open-source Java viewer for chemical structures in 3D with features for chemicals, crystals, materials and biomolecules
meep	A free finite-difference time-domain (FDTD) simulation software package
MIRA	A whole genome shotgun and EST sequence assembler for Sanger, 454, Solexa (Illumina) and PacBio data
molden	A visualization program for GAMESS, Gaussian, and Mopac/Ampac
NAMD	A parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems
ROOT	An object oriented framework for large scale data analysis
Smoldyn	A computer program for cell-scale biochemical simulations
VisIt	A free interactive parallel visualization and graphical analysis tool
VMD	A molecular visualization program for dislpaying, animating, and analyzing large biomolecular systems using 3-D graphics and built-in scripting

if repeated requests of a certain programs have been received from several different laboratories. Details of service are described in Section 5.1.

4 Operation status

In this section, we describe the operation status of the RICC system: viz, (i) our account classification by resources (four types of account classes); (ii) the Weekend Operation for Large Scale Parallel Jobs for the next generation parallel computers; and (iii) Statistics for RICC users related to the research fields, operation rate, numbers of running and waiting jobs. Note that we do not charge any of our services.

4.1 Account classification for RICC system

Those who wish to use the RICC system must submit proposals for each project, and their proposals are evaluated by our Review Committee. As resources of the RICC are project-dependent the respective projects are divided into the following four categories: (i) General User, (ii) Special User (iii) Exclusive User and (iv) Quick User accounts.

1. General User

   Approved projects are allowed to use a computation time of 1% or more (i.e., 67,000 core × hours/month; equivalent to a cost of US $7,800/month).

2. Special User account.

Approved users are given priority to perform large-scale parallel computation using half of all of the computational resources on weekends.

3. Exclusive User account.

   Approved projects are allowed to use a portion of computational resources exclusively.

4. Quick User account.

   Approved users are allowed to use a computation time of $< 1\%$ (i.e., 67,000 core $\times$ hours/month; equivalent a cost of US $ 7,800/month).

Once approved, all account type users can use the system for at most one year. Use of the system may start immediately after the account has been approved, and users can use the system until the end of any fiscal year. An application can only be filed in once in a fiscal year. Application of research proposals are opened to all relevant parties of RIKEN.

Proposals in categories (1), (2) and (3) are evaluated by the Review Committee, which meets twice a year. For these relevant accounts, users must specify the required resources intended for use. Proposals for the Quick User account, or category (4), are reviewed by the staff of RICC, and are usually approved within one or two working days. Nonetheless, all the cores can be used with this account-type. Therefore, first-time users of the RICC system should apply for a Quick User account.

4.2 Weekend Operation for Large Scale Parallel Jobs

Since we started the operation of RICC, our policy has been to allow routinely the submission of large scale parallel jobs (up to a maximum of 8192 parallel jobs). This is facilitate development of the next-generation super computer "K", which has 640,000 cores. However, RICC has routinely been operated at a very high usage rate, and execution of large scale parallel jobs during its routine operation is therefore practically impossible. In order to perform tests on large-scale parallel jobs efficiently, we have reserved specific days for executing such jobs and are calling for proposals. Project leaders who wish to run large scale parallel jobs of 8192 parallel jobs are requested to read conditions for Weekend Operation for Large Scale Parallel Jobs and submit their requests by e-mail. If the total time of run requests exceeds 24 hours, the requests will be carried over to the next weekend operation. This limitation startegy is to keep the CPU usage ratio high without causing any impact on routine operation. Note that tuning know-how for "K" computer is still not yet accumulated as it is a new architecture. Tuning guide for "K" computer is really desirable.

4.3 Research fields of RICC users

As mentioned above, you have to be a member or research collaborators of RIKEN to use the RICC system. For researchers who wish to use the RICC in any given year, they must apply with their research projects to and then subjected to evaluation by the Review Committee for the supercomputer accordingly. The Review Committee meetings are usually held four times in a year, and 30 projects were evaluated in 2010. The Review Committee appropriates projects with due rationally, making sure that the projects fit in the available total CPU time. Currently, 174 research projects have been approved by the Review Committee, and 288 researchers are using the RICC system. According to number of users for each research field (Figure 3), 44% of researchers are involved in life sciences; the largest research field in RIKEN. Next is nuclear physics/high-energy physics (28%), following by chemistry (10%), engineering (8%), computer science/information science (5%), astronomy (3%), and brain science (3%) in that order.

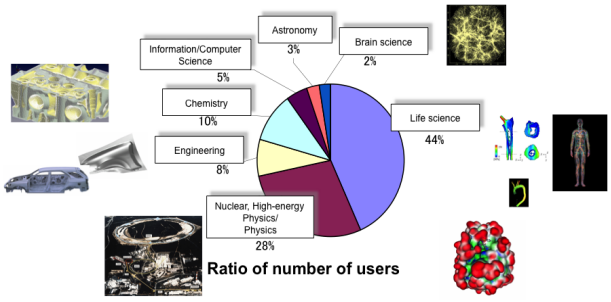


Figure 3: Research fields of RIKEN.

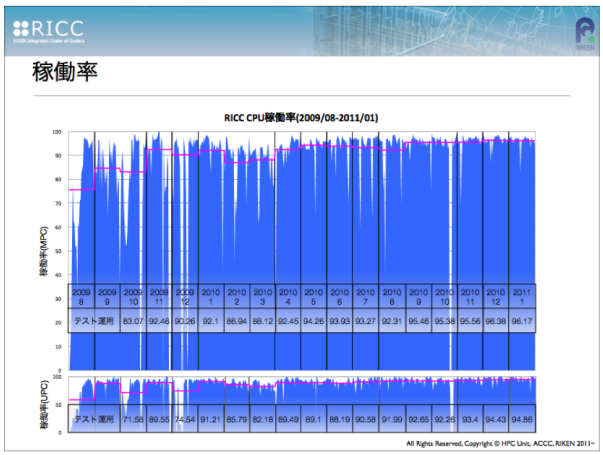


Figure 4: Operation rate of the RICC system.

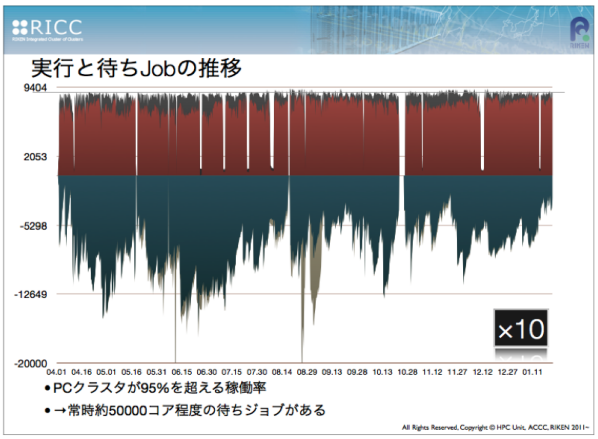
4.4 Operation rate and number of running and waiting jobs

Since initiation of the operations, the operation rate of RICC systems has registered higher than 90%. According to the operation rate of main nodes (the Massively Parallel PC Cluster or UPC; and the Multi-Purpose PC Cluster or MPC), the blue and red plots depict the daily operation rate, and month-averaged operation rate, respectively. Since April 2010, both these ratios have been gradually increasing, and the ratio registered 95% in January 2011. The operation of RICC was suspended in October for maintenance, thus accounting for a zero operation during this period.

User jobs are managed by our original job scheduler. RICC has a very high operation rate (see above), an application submitting may have to wait in line for application processing. Based on the number of running jobs on RICC, and waiting jobs in the queue, these numbers are actually counted as requested cores in a job since each job requires a different number of nodes and cores. For example, if a job for serial computing uses one core, it is counted as one; while job of a parallel type which uses 128 cores is counted as 128, and so on. Moreover, the horizontal axis is divided into positive (the number of running jobs) and negative (the number of waiting jobs divided by 10) parts (Figure 5). Accordingly, these are always ca. 50,000 waiting jobs in the queue: i.e., five times more than the RICC system can cope. Therefore, when the research project Review Committee controls and/or tunes parameters of the job scheduler by weighing the results versus usage for every user in August, the peak number of jobs in the queue reduces from September to October.

We also accept jobs to develop massively parallel applications using several thousands to 8,000 cores. In fiscal year 2010, special jobs consumed ca. 2,500,000 core CPU time, which is about 3%

Figure 5: Queue status of the RICC system.



of all available CPU resources of RICC.

5 Services and Supports for the RICC users

We have been providing the following services for the RICC users; (i) software installation, (ii) support for speed up/parallelization, high-grade programming and visualization, and (iii) RICC training classes. Again, all the services are free of charge. (details are given in the following subsections).

5.1 Software installation service

We accept proposals for application/software that researchers wish to use on the RICC at any time; installation costs for any application/software requested by two or more laboratories are usually borne by RIKEN within allocated budget. Even when free software is inappropriate or a commercial package is not available, or if application/software needs some improvement for research, the ACCC will take appropriate actions if requested.

5.2 Tuning and visualization support service

We have been providing the following three support services for RIKEN researchers: support services for speed up/parallelization, high-grade programming and visualization (see below for details of the support services).

- Speed up/parallelization support service

  We have been sponsoring training sessions on scalar tuning and parallel programming (MPI) several times a year to train researchers in speed up/parallelization. In addition to the training sessions, we tune programs to speed up or parallelize programs with MPI.

- High-grade programming support

  We innovate applications and programs used routinely by users to facilitate execution more easily with web or GRID computing technologies. As a result, it has become possible to run programs with a shorter turn-around time by using GRID computing technology that vitalizes specific computation resources.

Table 3: The tuning results.

Period	Abstract	Results
Apr. to May. 2010	Tuning of Commutation routine of in house molecular dynamics code. Speed up is observed by dividing the MPI communication parts into inter- and intra-nodes.	MPI_ALLTOALL : 0.33s to 0.042s MPI_ALLREDUCE : 7.66s to 1.84s MPI_REDUCE_SCATTER: 9.8s to 2.8s MPI_ALLGATHER: 0.7s to 0.1s

- Visualization support

  Through this service, we provide technical support for the development of visualization tools or the visualization of numerical data such as computational results and experimental results that are too large to be visualized in a personal computation environment.

The results of tuning in fiscal year 2010 are shown in Table 3.

5.3 RICC training service

We sponsor the following training courses sophisticated use of the RICC system. These classes are designed for elementary and intermediate level users. English versions are also available for the following classes; Introduction to RICC, OpenMP programming, XPFortran programming and MPI programming. Most of classes/courses are conducted only for RICC users, although some are opened to interested parties outside RIKEN. Moreover, texts for scalar tuning, MPI programming and GPGPU in Japanese are available from our website [3]. We conduct one or two classes for each course in a year (see below for the details of the classes).

- Introduction of the RICC.

  This course is designed for first-time users of the RICC. It covers how to use the RICC, access RICC from outside of RIKEN, and programming environment, etc. Both English and Japanese classes are available.

- Scalar tuning.

  Scalar tuning is a technique for getting better performance in terms of speed for a program. The performance of a program using scalar tuning may be 10 times – 100 times faster than without the technique. This course provides the basic techniques for scalar tuning; e.g., setting appropriate compiler options, measure performance, tuning for cache, I/O and, other tunings, how to use external libraries, etc. These techniques are almost language-independent, although in this course we explain in Fortran. This course is designed for beginner to intermediate level.

- MPI programming.

  The RICC and other contemporary super computer systems are usually of the massively parallel type; with distributed memories and are interconnected. Usually MPI (Message Passing Interface) programming is required for such systems for better performance, and this special programming model is not trivial. Mainly Fortran is used in this course. This course is designed for beginner to intermediate level, and only Japanese-medium classes are available.

- XPFortran programming.

XPFortran is a variant of Fortran parallel computers. It defines directives for partitioning instructions and data, and for communication among processors as well. This course is designed for beginner to intermediate level.

- CUDA programming.

  GPU (Graphics Processing Unit), is currently very popular among scientific computing for its spectacular performance and the RICC system is equipped with NVIDIA C1060 GPUs. However, CUDA programming is not trivial and is not easy for beginners, because the NVIDIA GPUs are suited for simple SIMD type programs used in merely repeating calculation like matrix-matrix multiplications, but not for very complicated calculations using conditionals and/or branches (e.g. using a lot of "if" statements). CUDA is very similar to C, and we require a basic C knowledge of the language. This course is designed for the intermediate level and only classed in Japanese are provided.

- OpenMP programming.

  OpenMP is an important parallel programming model for shared memory systems using multiple cores. To parallelize a certain program, OpenMP requires additional directives to C and/or Fortran. Note that the MPI parallel programming model is a distributed memory systems, therefore these two programming models are complementary. This course designed for the intermediate level and requires knowledge of C. Both English and Japanese classes are available.

6 Summary

In this short introduction, we presented the status of the RIKEN Integrated Cluster of Clusters (RICC) system, which consists of four cluster types. This system has been in operation since August 2009. The total LINPACK peak performance is 97.94 TFlops with 92.36% of efficiency. Our main objective with RICC is to facilitate, encourage, accelerate research activity of RIKEN by providing fast experimental data analysis with technical support and massively parallel clusters to establish the next-generation supercomputer "K", with a new type of computer environment: i.e., the GPU cluster. The operation ratio of RICC has scored very high ranking ($> 90\%$) since its operational launch. We also run very large-scale jobs using several thousand of cores at specific time periods. In fiscal year 2010 alone, we have processed 70 jobs corresponding 2,500,000 cores times or 3% of total core time resource of RICC. In addition, we provide tuning, visualization services, and training courses for the convenience of users. In this way, the RICC system has contributed and will continue to facilitate research activity and advancement of science endeavored by scientists at RIKEN.

Acknowledgment

Dedicated to the memory of our late Dr. SHIGETANI Takayuki, who passed away in 2010, at an early age. The current the RICC system development would not have been possible without his constructive contribution and devoted management of the RICC and other former super computer systems of RIKEN.

References

[1] Advanced Center for Computing and Communications, *Annual Report 2010 (in Japanese)*, 2010, RIKEN, Saitama, Japan.

[2] http://www.top500.org/list/2011/11/200.

[3] http://accc.riken.jp/HPC/training.html.