Overview of the K computer System

● Hiroyuki Miyazaki ● Yoshihiro Kusano ● Naoki Shinjou
● Fumiyoshi Shoji ● Mitsuo Yokokawa ● Tadashi Watanabe

RIKEN and Fujitsu have been working together to develop the K computer, with the aim of beginning shared use by the fall of 2012, as a part of the High-Performance Computing Infrastructure (HPCI) initiative led by Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT). Since the K computer involves over 80 000 compute nodes, building it with lower power consumption and high reliability was important from the availability point of view. This paper describes the K computer system and the measures taken for reducing power consumption and achieving high reliability and high availability. It also presents the results of implementing those measures.

1. Introduction

Fujitsu has been actively developing and providing advanced supercomputers for over 30 years since its development of the FACOM 230-75 APU—Japan's first supercomputer—in 1977 (**Figure 1**). As part of this effort, it has been developing its own hardware including original processors and software too and building up its technical expertise in supercomputers along the way.

The sum total of this technical expertise has been applied to developing a massively parallel computer system—the K computer[1), note)]—which has been ranked as the top performing supercomputer in the world.

The K computer was developed jointly by RIKEN and Fujitsu as part of the High Performance Computing Infrastructure (HPCI) initiative overseen by Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT). As the name "Kei" in Japanese implies, one objective of this project was to achieve a computing performance of 10^{16} floating-point operations per second (10 PFLOPS). The K computer, moreover, was developed not just to achieve peak performance in benchmark tests but also to ensure high effective performance in applications used in actual research. Furthermore, to enable the entire system to be installed and operated at one location, it was necessary to reduce power consumption and provide a level of reliability that could ensure the total operation of a large-scale system.

To this end, four development targets were established.

- A high-performance CPU for scientific computation
- New interconnect architecture for massively parallel computing
- Low power consumption
- High reliability and high availability

The CPU and interconnect architecture are defined in detail in other papers of this special issue.[2),3)] In this paper, we present an overview

note) "K computer" is the English name that RIKEN has been using for the supercomputer of this project since July 2010. "K" comes from the Japanese word "Kei," which means ten peta or 10 to the 16th power.

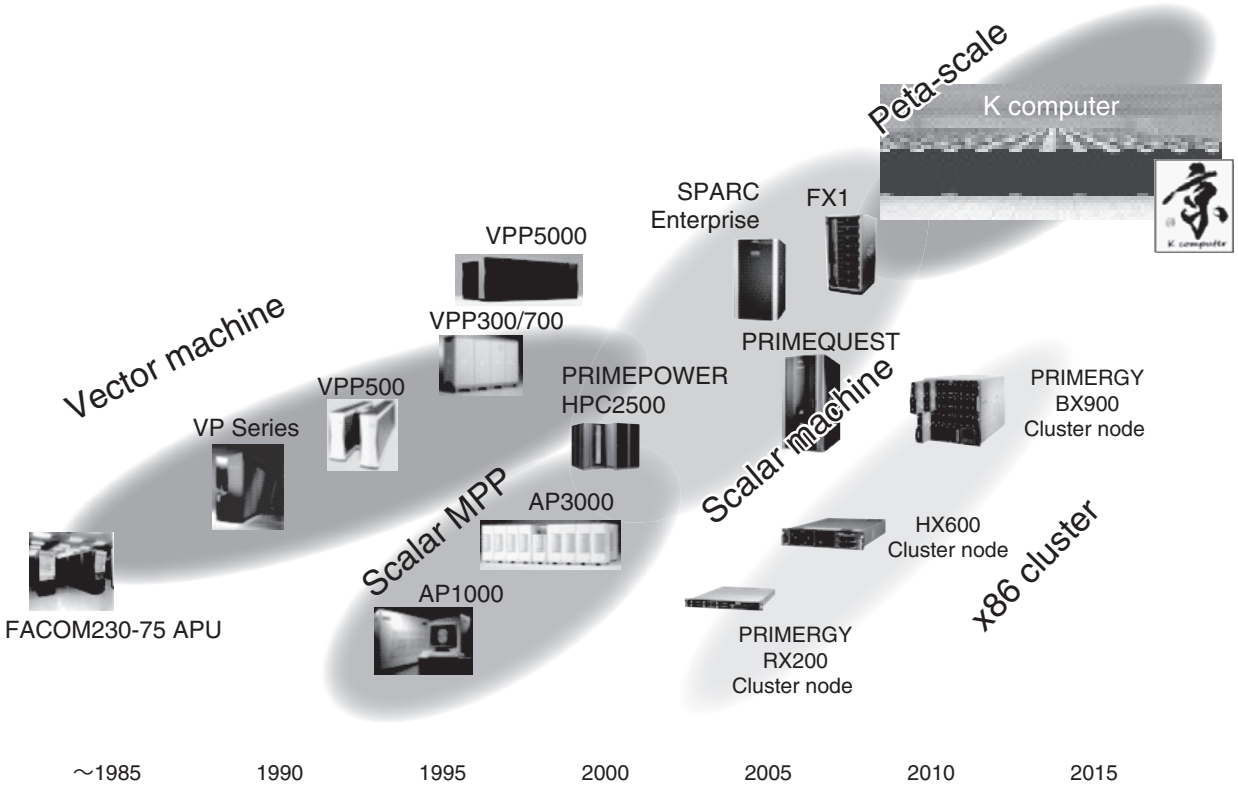

Figure 1
History of supercomputer development at Fujitsu.

of the K computer system, describe the measures taken for reducing power consumption and achieving high reliability and high availability at the system level of the K computer, and present the results of implementing those measures.

2. Compute-node configuration in the K computer

We first provide an overview of the compute nodes, which lie at the center of the K computer system. A compute node consists of a CPU, memory, and an interconnect.

1) CPU

We developed an 8-core CPU with a theoretical peak performance of 128 GFLOPS called the "SPARC64 VIIIfx" as the CPU for the K computer (**Figure 2**).[4] The SPARC64 VIIIfx features Fujitsu's advanced 45-nm semiconductor process technology and a world-class power performance of 2.2 GFLOPS/W achieved by implementing power-reduction measures from both process-technology and design points of view. This CPU applies High Performance Computing–Arithmetic Computational Extensions (HPC-ACE) applicable to scientific computing and analysis. It also uses a 6-MB/12-way sector cache as an L2 cache and uses a Virtual Single Processor by Integrated Multicore Architecture (VISIMPACT), whose effectiveness has previously been demonstrated on Fujitsu's FX1 high-end technical computing server. As a result of these features, the SPARC64 VIIIfx achieves high execution performance in the HPC field. In addition, the system-control and memory-access-control (MAC) functions, which were previously implemented on separate chips, are now integrated in the CPU, resulting in high memory throughput and low latency.

2) Memory

Commercially available DDR3-SDRAM-

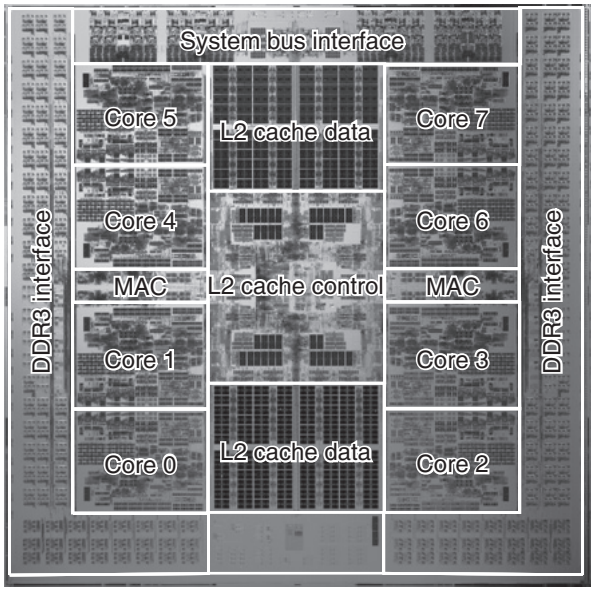

Figure 2
SPARC64 VIIIfx.

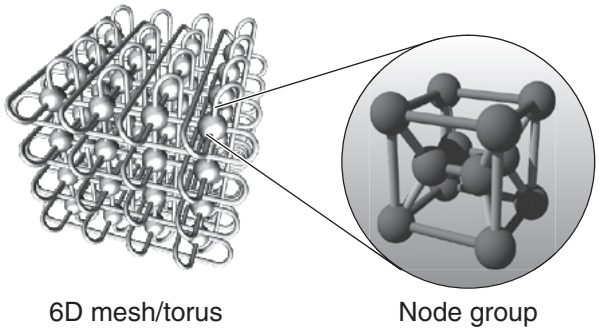

Figure 3
Tofu interconnect.

DIMM is used as main memory. The dual in-line memory module (DIMM) is a commodity module that is also used in servers and PC clusters, which provides a number of advantages. For example, it can be multi-sourced, a stable supply of about 700 000 modules can be obtained in a roughly one-year manufacturing period with a stable level of quality, and modules with superior power-consumption specifications can be selected. Besides this, an 8-channel memory interface per CPU provides a peak memory bandwidth of 64 GB/s, surpassing that of competitors' computers and the high memory throughput deemed necessary for scientific computing.

3) Interconnect

To provide a computing network (interconnect), We developed and implemented an interconnect architecture called "Tofu" for massively parallel computers in excess of 80 000 nodes.[5] The Tofu interconnect (**Figure 3**) constitutes a direct interconnection network that provides scalable connections with low latency and high throughput for a massively parallel group of CPUs and achieves high availability and operability by using a 6D mesh/torus configuration. It employs an extended-dimension-order routing algorithm that enables the provision of a CPU group joined in a 1–3 dimensional torus so that a faulty node in the system becomes non-existent as far as the user is concerned. The Tofu interconnect also makes it possible to allocate to the user a complete CPU group joined in a 1–3 dimensional torus even if part of the system has been extracted for use.

A compute node of the K computer consists of the CPU/DIMM described above and an InterConnect Controller (ICC) chip for implementing the interconnect. Since the Tofu interconnect forms a direct interconnection network, the ICC includes a function for relaying packets between other nodes (**Figure 4**). If the node section of a compute node should fail, only the job using that compute node will be affected, but if the router section of that compute node should fail, all jobs using that compute node as a packet relay point will be affected. The failure rate of the router section is nearly proportional to the amount of circuitry that it contains, and it needs to be made significantly lower than the failure rate of the node section. ICC-chip faults are therefore categorized in accordance with their impact, which determines the action taken at the time of a fault occurrence. The end result is a mechanism for continuing system operation in the event of a fault in a node section.

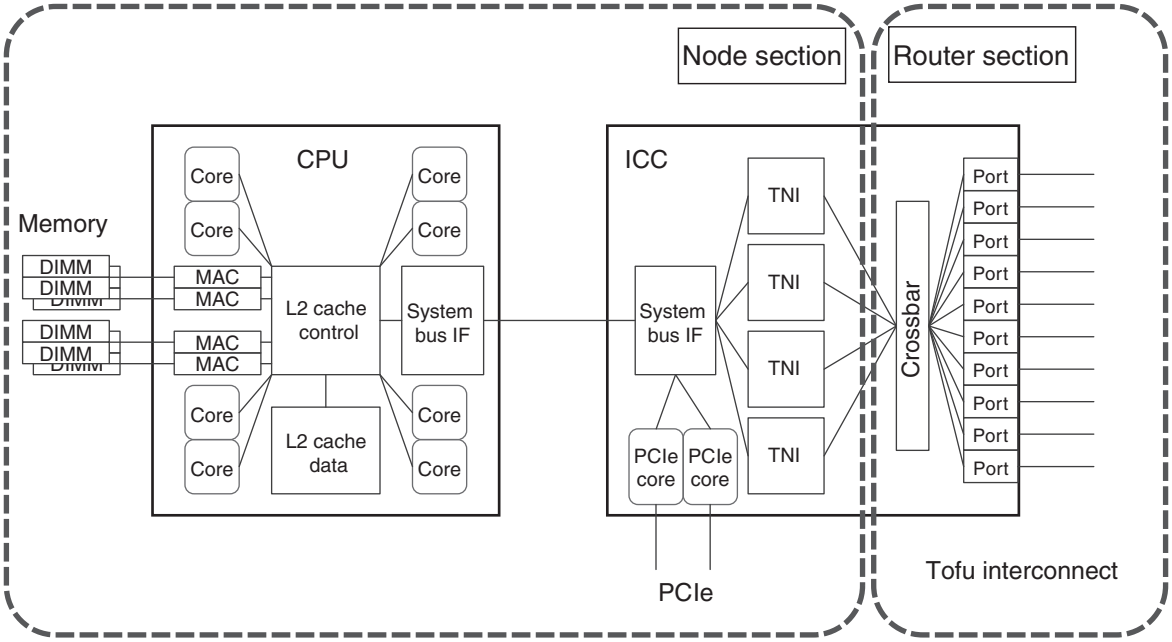

Figure 4
Conceptual diagram of node and router sections.

3. Rack configuration in the K computer

Compute nodes of the K computer are installed in specially developed racks. A compute rack can accommodate 24 system boards (SBs), each of which mount 4 compute nodes (**Figure 5**), and 6 I/O system boards (IOSBs), each of which mount an I/O node for system booting and file-system access. A total of 102 nodes can therefore be mounted in one rack.

Water cooling is applied to the CPU/ICCs used for the compute nodes and I/O nodes and to on-board power-supply devices. Specifically, chilled water is supplied to each SB and IOSB via water-cooling pipes and hoses mounted on the side of the rack. Water cooling is used to maintain system reliability, enable a high-density implementation, and lower power consumption (decrease leakage current). In contrast, the DIMMs mounted on the SBs and

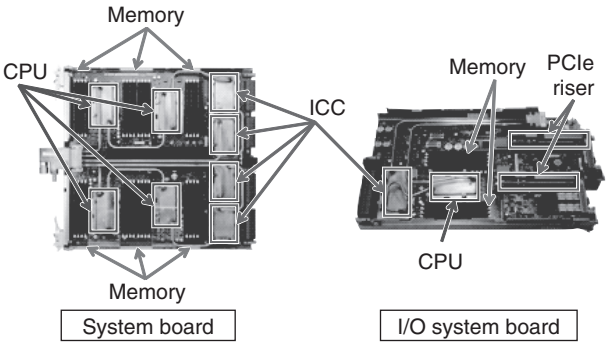

Figure 5
System board and I/O system board.

IOSBs are air cooled since they make use of commercially available components. Since these boards are mounted horizontally inside the rack, air must be blown in a horizontal direction. To enable the racks themselves to be arranged in a high-density configuration, the SBs are mounted diagonally within the rack to form a structure in which air is sucked in from the front, where

the boards are angled toward the opening, and passes through the SBs to cool the DIMMs before exiting from the rear (**Figure 6**).

The six I/O nodes mounted in each rack are classified by application:

- Boot-IO node (BIO): two nodes
- Local-IO node (LIO): three nodes
- Global-IO node (GIO): one node

The BIO nodes connect via a fiber-channel interface up to 8 Gbit/s (8G-FC) to the system-boot disk (ETERNUS DX80) installed in the rack. These nodes start up autonomously at power-on and function as a boot server via Tofu for the remaining 100 nodes in the rack.[6] One plays the role of active system, and one plays the role of standby system. If the one playing the role of active system fails, the standby one assumes that role.

The LIO nodes connect to local disks (ETERNUS DX80) mounted on disk racks installed near the compute racks.

The GIO node connects to external global storage systems via an InfiniBand quad data rate (QDR) interface. These storage systems use the Fujitsu Exabyte File System (FEFS)[7] to provide a large-scale, high-performance, and high-reliability file system. If the LIO nodes or GIO node in a rack should fail, file-system access can be continued via a path to LIO or GIO nodes mounted in a neighboring rack.

Each rack also includes power supply units (PSUs), air conditioning fans, and service processors (SPs), all in a redundant configuration so that a single failure does not bring down the

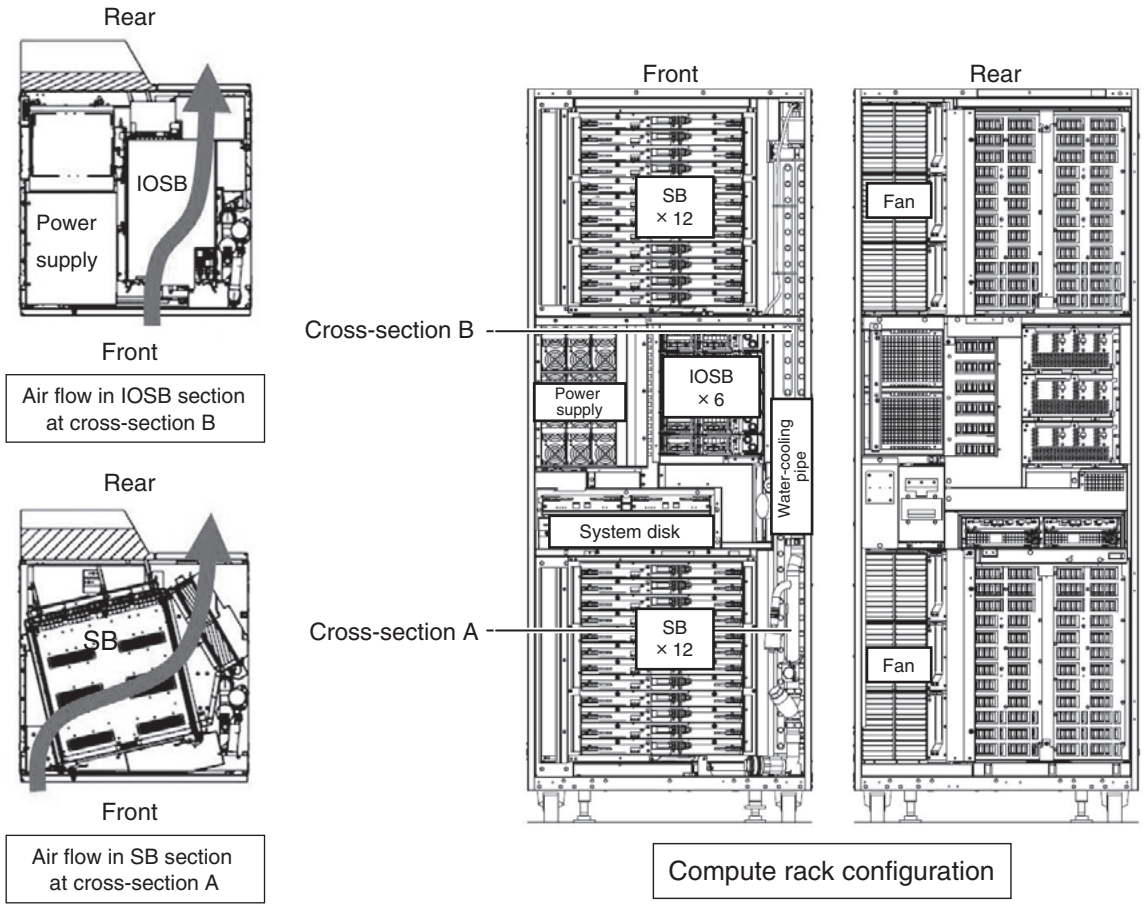

Figure 6
Compute rack configuration and direction of air flow.

entire rack.

4. Overall system configuration

The K computer system consists of 864 compute racks in total. As shown in **Figure 7**, two adjacent compute racks are connected by a Z-axis cable (connecting 17 nodes including I/O nodes). There is one disk rack for every four compute racks, and 45 racks are arranged along the Y-axis (36 compute racks + 9 disk racks) and 24 racks along the X-axis. The K computer therefore has a rectangular floor configuration of 24×45 racks. Each disk rack mounts 12 local disks and each local disk connects to 2 adjacent LIO nodes in the X-direction.

There's more to the K computer than simply compute racks, disk racks, and a parallel file system—a control mechanism consisting of peripheral server hardware and operation/management software for system and job management is an absolute necessity.[8] Maintenance operations based on a remote alarm system and maintenance server are also needed to replace hardware after a failure. An overview of the entire K computer system installed at RIKEN is given in **Figure 8**.

5. Power-reduction measures at the system level

To reduce power consumption across the entire K computer, the first step was to implement drastic power-reduction measures

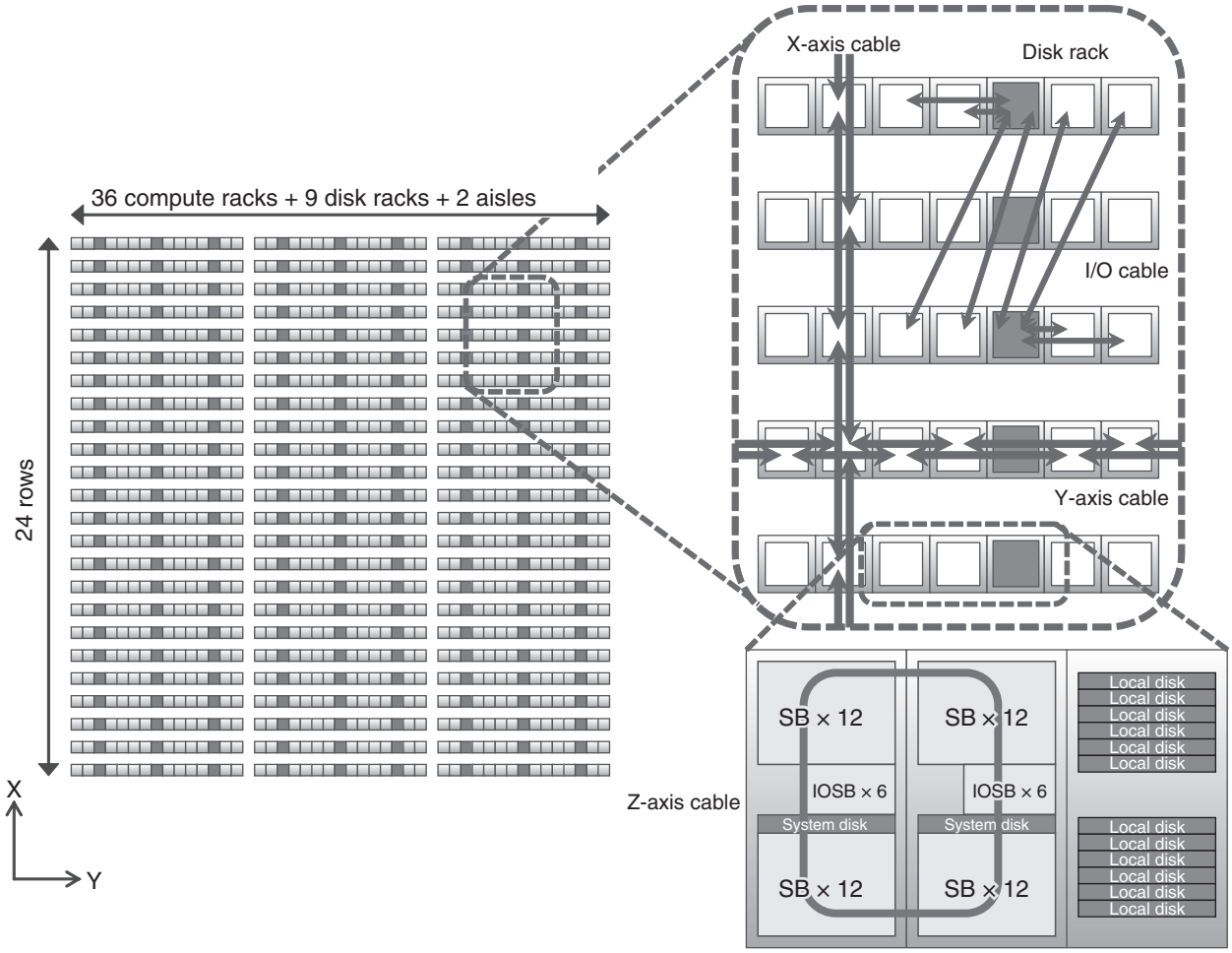

Figure 7
Configuration of compute racks.

at the CPU/ICC development stage. Dynamic current was reduced through clock gating, while operating frequency was lowered to 2.0 GHz and leakage current was significantly decreased by adopting high-Vth transistors.[9] It was also decided to employ water cooling since leakage current increases rapidly with the junction temperature. This had the effect of dropping junction temperature from the usual 85°C to 30°C, resulting in a significant decrease in leakage current.

The use of a clock-gating design to reduce LSI power consumption makes load fluctuation even more noticeable. This makes it necessary to improve transient response characteristics in the power supply. An intermediate bus converter system was adopted here to improve transient response characteristics and feed power efficiently to a large number of racks while enabling a high-density implementation. Specifically, transient response characteristics were improved by arranging non-isolated, point-of-load (POL) power supplies near loads, and power feeding to the entire rack was made more efficient by supplying 48 V of power to each SB from the rack power supply. The mounting of an insulated type of bus converter on each SB makes it possible to step-down the supplied 48 V and supply power to the POL power supplies. This approach reduces the number of isolated transformers and achieves a high-efficiency and high-density power-supply system.

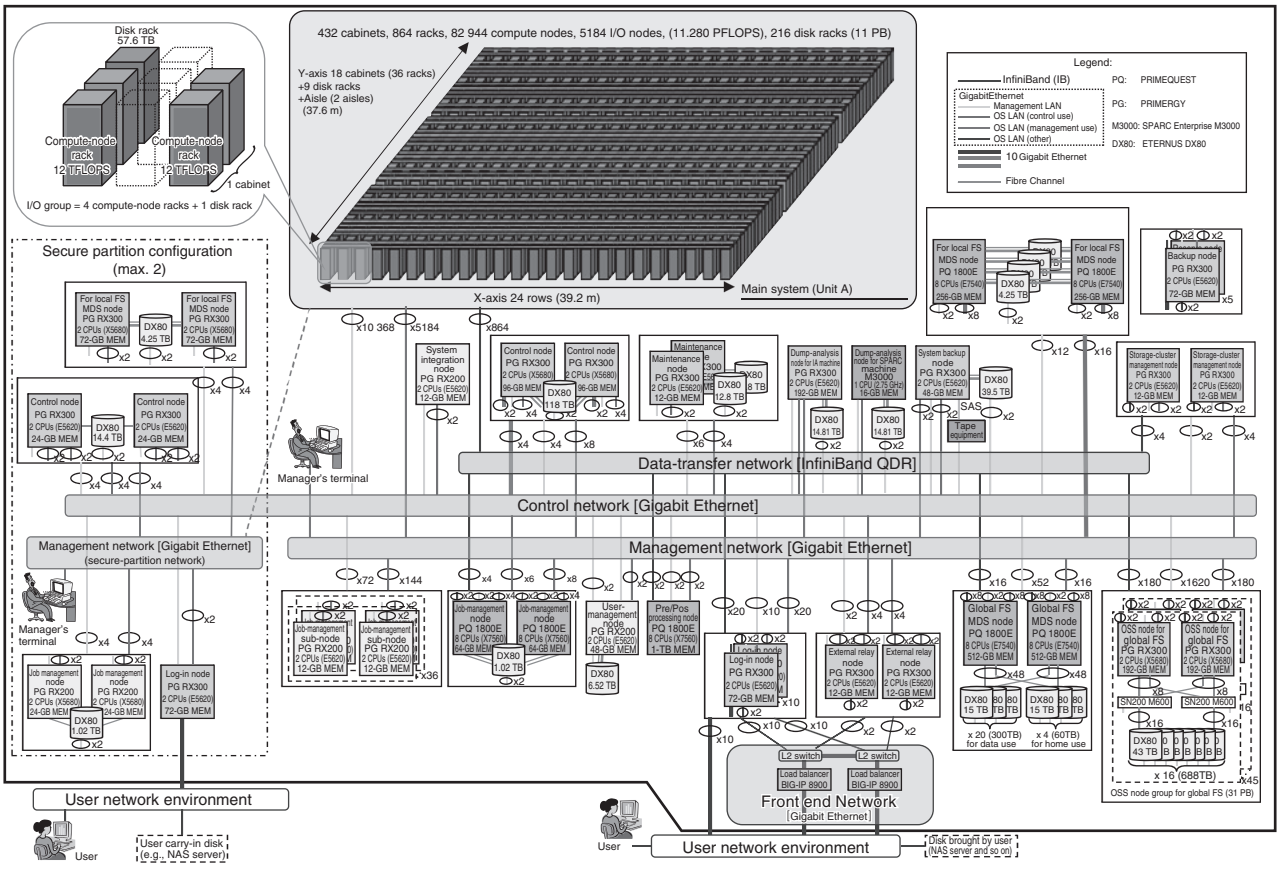

Figure 8
Overview of entire system.

6. High-reliability/high-availability measures at the system level

Adopting a system-on-a-chip (SOC) configuration means that the amount of circuitry on the chip increases while the number of components outside the chip decreases. Consequently, failures in large-scale chips account for a greater percentage of system failures, which means that reducing the failure rate of large-scale chips is essential for stable operation of the K computer.

6.1 High-reliability measures

In the K computer, reliability is improved by improving the reliability of individual components and adopting an appropriate method for using those components. In this regard, efforts to improve system reliability are performed within the frameworks of job continuity and system availability.

As shown in **Table 1**, giving major components a redundant configuration helps to ensure operational continuity and therefore job continuity, and replacing faulty components by "hot-swapping" without halting the entire system helps to ensure system availability.

In more detail, measures within the job-continuity framework include the application of redundant hardware configurations, a reduction in the number of components by using a 1-node/1-CPU configuration, error correction by using error correcting code (ECC) and retries, and reduction in the occurrence rate of semiconductor failure by lowering operating temperature through the water cooling of LSIs and POL power supplies. These measures lower the probability of a node or job crash due to a component failure. Another measure to help ensure job continuity is the use of electrical cables in the Tofu interconnect.

6.2 High-availability measures

From the viewpoint of ensuring system availability, a Tofu coordinate arrangement has been implemented within SBs so that the faulty-node-avoidance mechanism based on a 6D mesh/torus can be extended to SB-fault handling and hot maintenance. This enables detour routing despite the fact that each SB mounts only four nodes. Specifically, **Figure 9** shows how the four nodes of a SB are confined to the same Y/B coordinates and interconnected along the A-axis and C-axis on the SB so that a connection along the B-axis offers a way out of the SB.

In addition, I/O path redundancy is provided for the three types of I/O nodes (BIO, LIO, GIO) so that the I/O paths of compute nodes under a faulty I/O node can still be used. This is done by adopting a configuration that incorporates alternate I/O nodes having common connection

Table 1
High reliability by using redundant configurations and hot swapping.

Major components	Redundant configuration	Hot swapping
Rack power supply	Yes	Yes
Cooling fans	Yes	Yes
Service processors (SPs)	Yes (duplicate/switchable)	Yes
System boards (SBs/IOSBs)	No ⇒ Fault avoidance along B-axis	Yes
CPU/ICC	No ⇒ Reliability improved by using water cooling, retries, and error correcting code (ECC)	(SB hot swapping)
POL power supplies	No ⇒ Reliability secured by water cooling	(SB hot swapping)
Intermediate converters	Yes	(SB hot swapping)
DIMMs	No ⇒ Data rescue by using extended ECC	(SB hot swapping)
System disks	Yes ⇒ Controller and power supply: redundant HDD: RAID5 configuration + hot spare	Yes (module)

destinations, which enables a compute-node group to access alternate I/O nodes via the Tofu interconnect (**Figure 10**). There are two BIO nodes per compute rack in a redundant active/standby configuration, and there are three LIO nodes per compute rack that take on a redundant configuration with LIO nodes in another compute rack that share local disks mounted in a disk rack. Similarly, the single GIO node in a compute rack takes on a redundant configuration with the GIO node of the other compute rack.

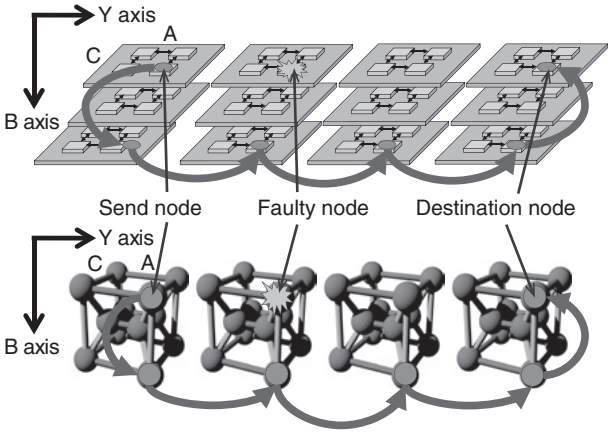

Figure 9
Conceptual diagram of Tofu detour.

7. Benchmark results

Though still in its configuration stage, the K computer developed according to the targets described at the beginning of this paper achieved a performance of 8.162 PFLOPS on the HPC-LINPACK benchmark test using only part of the system (672 racks). This achievement won

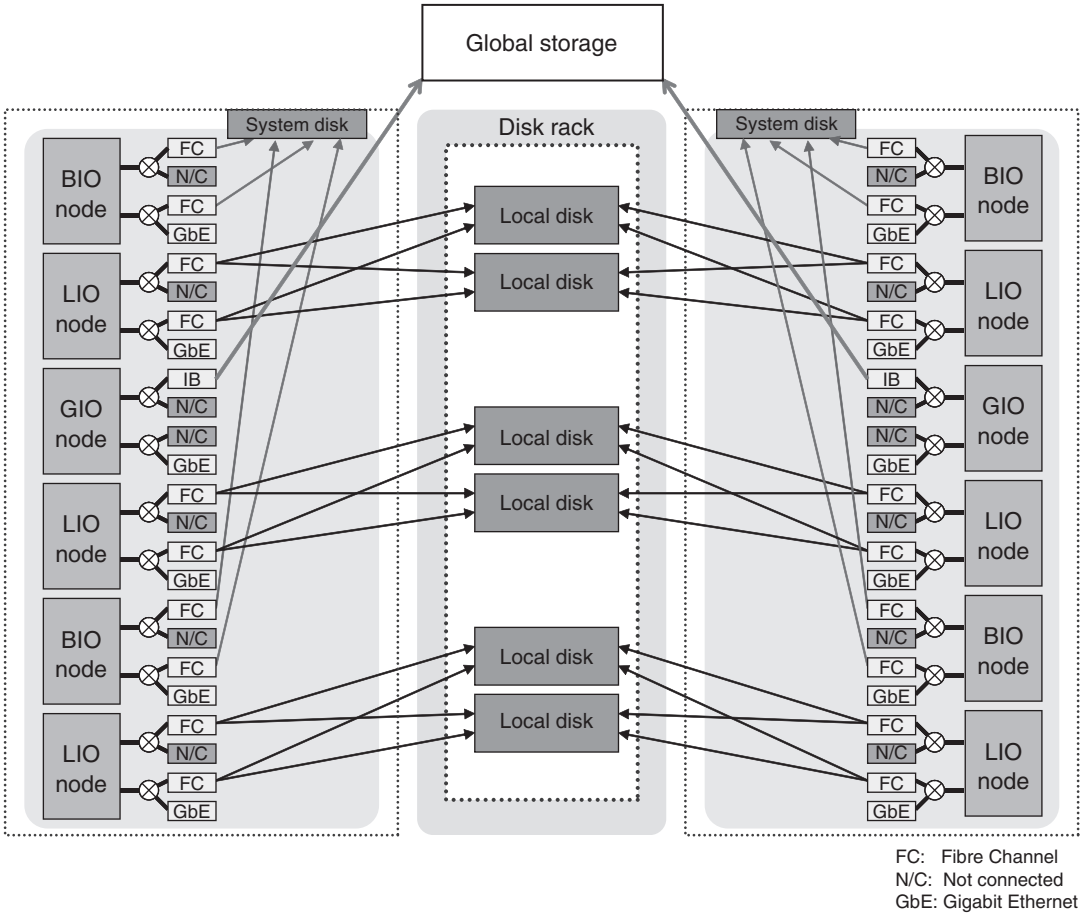

Figure 10
I/O path redundancy.

the K computer the No. 1 spot on the TOP500 list announced in June 2011.[10] Despite the fact that this LINPACK benchmark test ran for more than 28 hours, all of the 68 544 nodes involved operated continuously without a failure, demonstrating high system reliability and high job continuity. Details on these benchmark results and those announced in November 2011 are given in **Table 2**.

The K computer also participated in the HPC Challenge benchmark test that assesses the overall performance of a supercomputer. In this test, it received a No. 1 ranking in all four categories {Global HPL, Global RandomAccess, EP STREAM (Triad) per system, and Global FFT} of the HPC Challenge Award (Class 1) announced in November 2011. Additionally, the K computer was awarded the Gordon Bell Peak Performance prize announced in November 2011 in recognition of its achievements with actual applications. These results show that the K computer, far from being a machine specialized for the LINPACK benchmark, possesses general-purpose properties that can support a wide range of applications.

Table 2
TOP500/Green500 results.

	June 2011	November 2011	Unit
TOP500 ranking	1	1	Rank
No. of cores	548 352	705 024	Units
Performance (Rmax)	8162.00	10 510.00	TFLOPS
Power	9898.56	12 659.9	kW
Green500 ranking	6	32	Rank
Power performance	824.56	830.18	MFLOPS/W

8. Conclusion

The development objectives established for the K computer called for a large-scale high-performance computing system with importance placed on high reliability and high availability from the design stage. RIKEN and Fujitsu have been working to achieve target peak performance and application effective performance and to construct a high-reliability and high-availability system. The K computer was ranked No.1 on the TOP500 benchmark list of June of 2011, and kept to be ranked No.1 on that of next November of 2011. It also achieves 10 PFLOPS of LINPACK performance. These achievements demonstrate that objectives are being steadily met as planned. To enter a commencing full-service once the performance-tuning phase comes to an end, we will be evaluating the performance of actual applications and actual system operation.

References

1) M. Yokokawa et al.: The K computer: Japanese next-generation supercomputer development project. ISLPED, pp. 371–372 (2011).
2) T. Yoshida et al.: SPARC64 VIIIfx: CPU for the K computer. *Fujitsu Sci. Tech. J.*, Vol. 48, No.3, pp. 274–279 (2012).
3) Y. Ajima et al.: Tofu: Interconnect for the K computer. *Fujitsu Sci. Tech. J.*, Vol. 48, No.3, pp. 280–285 (2012).
4) T. Maruyama et al.: SPARC64 VIIIfx: A New-Generation Octocore Processor for Petascale Computing. *IEEE Micro*, Vol. 30, Issue 2, pp. 30–40 (2010).
5) Y. Ajima et al.: Tofu: A 6D Mesh/Torus Interconnect for Exascale Computers. *Computer*, Vol. 42, No. 11, pp. 36–40 (2009).
6) J. Moroo et al: Operating System for the K computer. *Fujitsu Sci. Tech. J.*, Vol. 48, No.3, pp. 295–301 (2012).
7) K. Sakai et al: High-Performance and Highly Reliable File System for the K computer. *Fujitsu Sci. Tech. J.*, Vol. 48, No.3, pp. 302–309 (2012).
8) K. Hirai et al: Operations Management Software for the K computer. *Fujitsu Sci. Tech. J.*, Vol. 48, No.3, pp. 310–316 (2012).
9) H. Okano et al.: Fine Grained Power Analysis and Low-Power Techniques of a 128GFLOPS/58W SPARC64™ VIIIfx Processor for Peta-scale Computing. Proc. VLSI Circuits Symposium, pp. 167–168 (2010).
10) TOP500 Supercomputing Sites. *http://www.top500.org*

Hiroyuki Miyazaki
Fujitsu Ltd.
Mr. Miyazaki is engaged in the drafting of specifications and the development of LSI circuits for next-generation HPC systems.

Yoshihiro Kusano
Fujitsu Ltd.
Mr. Kusano is engaged in the development work for next-generation HPC systems.

Naoki Shinjou
Fujitsu Ltd.
Mr. Shinjou is engaged in overall development work for next-generation HPC systems.

Fumiyoshi Shoji
RIKEN
Dr. Shoji is engaged in system development of system.

Mitsuo Yokokawa
RIKEN
Dr. Yokokawa is engaged in overall control of system development.

Tadashi Watanabe
RIKEN
Dr. Watanabe is engaged in overall control of development project.