JAXA Supercomputer Systems with Fujitsu FX1 as Core Computer

● Takayuki Abe   ● Tomohide Inari   ● Ken Seki

(Manuscript received June 19, 2008)

In April 2009, the Japan Aerospace Exploration Agency (JAXA) will deploy an integrated supercomputer system called JAXA Supercomputer Systems (JSS) using the Fujitsu FX1 technical computing server as the core computer. JSS consists of a massively parallel supercomputing system, storage system, large-scale shared memory system, and remote access system. It features outstanding application-execution performance, large-capacity/high-speed storage, and remote access. This paper outlines JSS and describes its features.

1. Introduction

In line with its corporate message of "Reaching for the skies, exploring space", Japan Aerospace Exploration Agency (JAXA) is exploring the great possibilities that space and aviation hold through diverse research and development with the aim of contributing to the peace and happiness of mankind. A key example of these activities is the R&D of numerical simulation technology. Using the high-speed computational power of supercomputers, JAXA has been developing and expanding the use of computational fluid dynamics and other types of numerical simulation.[1]

JAXA's predecessors, the Institute of Space and Astronautical Science (ISAS), National Aerospace Laboratory of Japan (NAL), and National Space Development Agency of Japan (NASDA), operated as three independent organizations at Chofu Space Center, Kakuda Space Center, and Sagamihara Campus, respectively, each with its own supercomputer system. JAXA plans to use its integrated supercomputer system called JAXA Supercomputer Systems (JSS) to apply numerical simulation technology whole-heartedly to space development and its other work. Thus, the introduction of JSS symbolizes the merger of the above three aerospace organizations.[2]

This paper introduces the computing systems and infrastructure supporting these research activities; namely, the basic configuration and features of JSS, which uses the FX1, Fujitsu's new high-end technical computing server, as the core computer.

2. Outline of JSS

JSS consists of four main elements: a massively parallel supercomputing system as an ultrahigh-speed computing engine, large-capacity/high-speed storage system, large-scale shared memory system, and remote access system for improved usability from remote sites.

The massively parallel supercomputing system is a top-class scalar computing engine delivering a total computational performance of 135 teraflops (TFLOPS) and total memory space of 100 terabytes (TB). This system constitutes the core of JSS and consists of a main system that provides numerical simulation environments
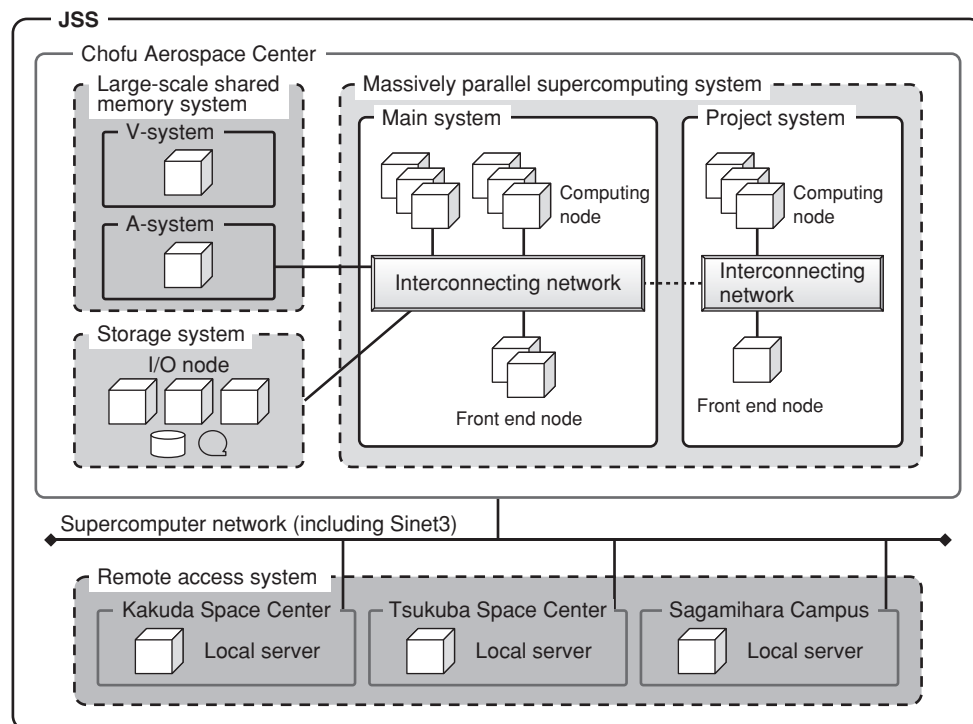
Figure 1
JSS configuration.

of various sizes and forms to researchers and a project system that provides a flexible numerical simulation environment for specific projects.

The storage system features 1 petabyte (PB) of memory on hard disk drives and 10 PB of memory on a tape library. It stores a huge quantity of numerical simulation data and provides high-speed access to that data.

The large-scale shared memory system consists of the A-system, a scalar shared-memory parallel-processing (SMP) computer, and the V-system, a vector computer, each having 1 TB of shared memory. The A-system executes independent software vendor (ISV) applications that perform numerical simulations and processing requiring large amounts of shared memory. The V-system inherits program assets previously executed on the vector computers operated in Kakuda Space Center and Sagamihara Campus and also executes software programs for the vector computers.

The remote access system enables users at the Kakuda Space Center, Tsukuba Space Center, and Sagamihara Campus to access Chofu Aerospace Center and use JSS.

The JSS program development and execution environment was achieved using Parallelnavi.[note 1] This environment provides Fortran, C, and C++ as program development languages and XPFortran and the message passing interface (MPI) as a parallel execution environment. Parallelnavi provides an interface that can customize the management of central processing units (CPUs) and other resources and the scheduling of jobs. It supports the JAXA-developed job control system and detailed node allocation system and provides a fair and

Table 1
Constituents of JSS system.

System name	Node name	No. of nodes	No. of CPUs	Application
Massively parallel supercomputing system	Main system computing node	3008	3008	Job execution
	Main system front end node	2	16	Login, etc.
	Project system computing node	384	384	Job execution
	Project system front end node	1	20	Login and I/O processing
Storage system	I/O node	3	96	I/O processing
Large-scale shared memory system	A-system	1	32	ISV, etc.
	V-system	3	48	Job execution, etc.
Remote access system	Local server	3	24	Login, etc.

efficient numerical simulation environment. The basic configuration of JSS is shown in **Figure 1**. The following sections focus on the massively parallel supercomputing system and the storage system.

3. Massively parallel supercomputing system

3.1 Configuration

The massively parallel supercomputing system features 3392 FX1 computing nodes, each consisting of one CPU with either 32 or 16 GB of memory. The system is divided into a main system having 3008 nodes (each having 32 GB of memory) and a project system having 384 nodes (each having 16 GB of memory). Each node consists of an FX1 technical computing server. The constituents of JSS system are listed in **Table 1**.

Nodes are linked by a full bisectional bandwidth[note 2] (FBB) fat-tree interconnecting network having a transfer performance of 2 GB/s in each direction, achieving high-speed data transfer between nodes.

3.2 Components and features

This subsection describes the FX1 for scien-

tific and technical computing, which is a major component of the massively parallel supercomputing system, the high-speed high-functionality interconnecting network, and the outstanding application-execution performance of JSS.

3.2.1 FX1

The FX1 uses the high-performance SPARC64 VII processor[3] developed by Fujitsu as its CPU. The SPARC64 VII processor is a quad-core CPU featuring a clock rate of 2.5 GHz, prefetching, out-of-order execution, simultaneous execution of four floating-point (FP) operations, and other functions.

The FX1 also adopts the Integrated Multicore Parallel Architecture,[4] a new technology that will lead the multicore era. This technology enables multiple cores (here, four) within a chip to be used as a single high-performance processing unit through cooperative operations and performance-enhancing techniques. These include an "automatic parallel compiler" that extracts the full performance of a multicore CPU, a "high-speed inter-core barrier mechanism" for achieving high-speed inter-core processing, and an "inter-core common L2 cache" that provides an effective means of avoiding false sharing. Furthermore, by linking with a high-performance compiler (Parallelnavi) and a dedicated chip set to achieve high memory bandwidth (Jupiter system controller [JSC]), this architecture speeds up the execution of vector-oriented code that runs

note 2) When an *N*-node network is divided into any two equal sections, the total theoretical communication performance between those two sections (bisectional bandwidth: bidirectional) is *N*/2 times the bandwidth of one node.

inefficiently on conventional scalar computers and extends the applicable range of loop parallelization. Within a node, moreover, multiple cores in one processor can be effectively used by thread parallelization.[5] The main specifications of the FX1 are listed in **Table 2**.

In short, the CPU features a hard barrier (high-speed thread synchronization) mechanism and allows the L2 cache to be shared among cores. The aim here is to improve thread parallelization performance by preventing synchronization overhead and cache false sharing (cache competition between multiple threads). In addition, improvements in latency[note 3] in the floating-point-operation unit and in parallel operability have resulted in an effective efficiency of 92% for the DGEMM[note 4] core routine of Linpack.[7] The LSI interconnections within a node are shown in **Figure 2**.

In the chip set, a theoretical peak throughput of 40 GB/s for memory can be achieved by operating the bus between the JSC and CPU at a high speed of 1.25 GHz and by using two JSC LSIs, each having four double data rate 2 (DDR2) memory interfaces. Actual measurements by the STREAM (Triad) benchmark[8] gave a memory throughput of more than 13.5 GB/s. Function values for SPARC64 processors ranging from the HPC2500 SPARC64 V to today's SPARC64 VII are compared in **Table 3**.

3.2.2 High-speed high-functionality interconnecting network

The interconnecting network uses an FBB fat-tree topology based on an InfiniBand DDR interface (peak theoretical performance of 2 GB/s full duplex) and interconnects 3008 nodes with a total computing performance of 120 TFLOPS using two-level leaf and spine (five LSI levels) InfiniBand switches. The objective here is to

note 3) Delay time from issuing a request to receiving results.
note 4) Calculates double-precision matrix products as commonly used in linear algebra.

Table 2
FX1 main specifications.

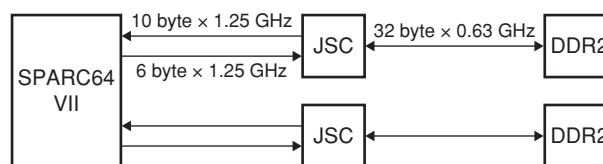| CPU	Processor	SPARC64 VII
	L2 cache	6 MB
Node	No. of CPUs	1
	Memory capacity	32 or 16 GB
	Memory bandwidth	40 GB/s
	Input/output	Hard disk drive (73 GB) × 1 InfiniBand[6] HCA (DDR) × 1 1000BASE-T × 1
Chassis	No. of nodes	4
	External dimensions	19-inch rack mount 5U

DDR: Double data rate
HCA: Host channel adapter

Figure 2
LSI interconnections within node.

Table 3
Function comparison among SPARC64 generations.

Function	SPARC64		
	V[note 1]	VI	VII
No. of CPU cores	1	2	4
Operating frequency (GHz)	1.3	2.4	2.5
FP product-sum operation latency	12 τ	7 τ	6 τ
Number of FP rename registers	32	48	48
Number of FP register write ports	2	4	4
Inter-core L2 cache sharing	No	Yes	Yes
Intra-CPU hard barrier	No	No	Yes
Number of TLB entries[note 2]	32	2048	
System bus width (B)	16	20 + 12	

note 1) Figures are for 130-nm version
note 2) For large-page operand
TLB: Translation lookaside buffer

minimize changes in communication time by using FBB connections to prevent degradation in effective performance caused by communication conflicts.

Furthermore, to raise the efficiency of parallel job execution and inhibit fluctuation in the performance of many-node jobs, the system

has a highly functional switch[note 5] for each of 768 nodes to execute inter-node high-speed barriers and reduction operations as well as adding synchronized interruptions for operating system (OS) scheduling. This synchronized interruption mechanism aims to inhibit performance degradation and fluctuation by equalizing the allocation periods for user time and OS time among the 768 nodes. The aim here is to prevent prolonged blocking or delaying of inter-node synchronized operations of user jobs by OS operations that differ for each node.

3.2.3 Outstanding application-execution performance of JSS

In this subsection, we present standard benchmark test results as indices of computing performance based on IMPACT. Measurements were taken using a pre-deployment FX1 system used for evaluation purposes.

Using the EuroBen Benchmark,[9] a benchmark test for measuring single CPU performance and MPI parallel performance of scientific and technical computers, we measured the pure performance of operations in the FX1. This benchmark test features many evaluation items and uses 31 types of calculation loops that extract basic operations to measure the performance of single-CPU operations. From among these, we introduce two examples of calculation loops that reflect the features of the SPARC64 VII processor.

1) Results of measuring single-core performance

The results of measuring single-core performance by using the EuroBen Benchmark are shown in **Figure 3**. Specifically, these are the
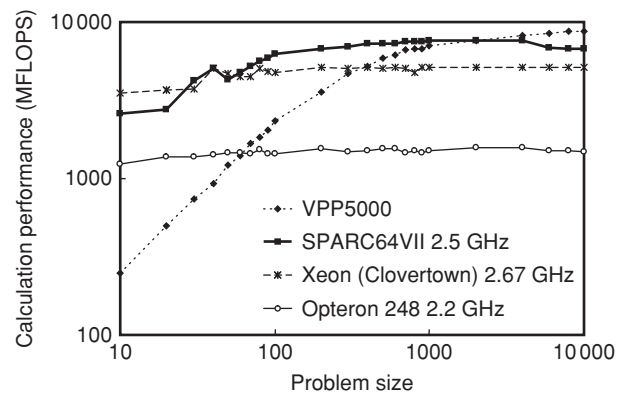
Calculation loop 14: 9th-order polynomial expression

Figure 3
Single-core performance measured by the EuroBen Benchmark.

results of measuring calculation performance for single-core execution while varying the problem size for calculation loop 14 (9th-order polynomial expression). The horizontal axis of the graph represents problem size and the vertical axis calculation performance. Compared with the performance of existing quad-core CPUs (Xeon and Opteron) and a vector machine (VPP5000), we can see that SPARC64 VII exhibits quick startup performance compared with that of the vector machine for small problems, and highly effective performance comparable with that of other CPUs or the vector machine for large problems.

2) Results of measuring automatic parallel performance

The results of measuring automatic parallel performance are shown in **Figure 4**. These results show calculation performance versus problem size for calculation loop 8 (vector multiplication by a constant and vector summation). Here, for other than the VPP5000, the system allocates four threads to four cores within a single CPU and performs parallel execution using automatic parallelization techniques. For relatively small problem sizes, the SPARC64 VII processor exhibits higher effective performance than the other systems, demonstrating its

---

note 5)   Applies some of the results obtained by the Petascale System Interconnect R&D project of the Fundamental Technologies for Next-generation Supercomputing R&D area of the R&D for Next-generation IT Infrastructure Building program of Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT).

Calculation loop 8: vector multiplication
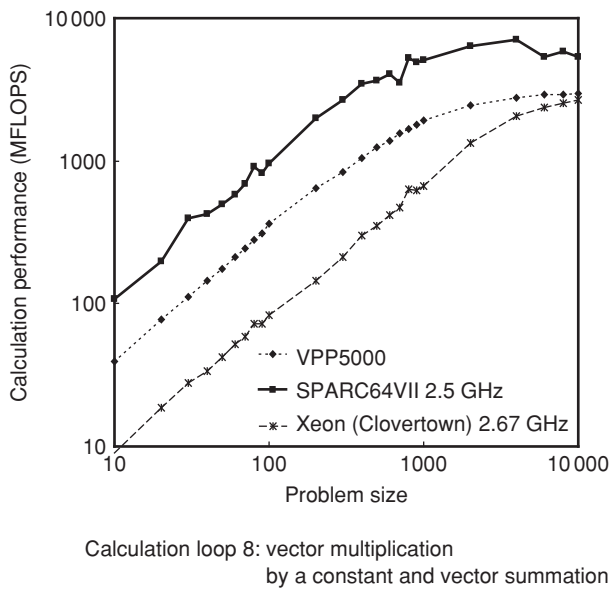by a constant and vector summation

Figure 4
Parallel performance measured by the EuroBen
Benchmark.

applicability to even fine-grained parallelization for which no effect can be obtained by conventional automatic parallelization. These results indicate that the FX1 can provide an effective high-speed mechanism for an even wider range of applications.

4. Large-capacity/high-speed storage system

The JSS large-capacity/high-speed storage system consists of a SPARC Enterprise M9000 server (3 units), 1-PB RAID5 disk array (ETERNUS2000: 90 units), and tape library with a total capacity of 10 PB (IBM TS3500 LTO drive: 48 units), resulting in high input/output (I/O) performance. The three SPARC Enterprise M9000 servers are connected to the disks and tape library by fiber channels via ETERNUS SN200 storage area network (SAN) switches.[note 6] This connection scheme aims to solve the problems of limited I/O performance due

to a single I/O node configuration and of system crashes caused by faulty I/O nodes by raising I/O performance through I/O node scale-out (horizontal scaling) and by improving availability through the use of a redundant format. The JSS storage system is outlined in **Figure 5**.

The storage system consists of two types of file systems. One is a local file system that controls the hard disk drives, tape drives, and storage devices connected to the storage system; the other is a shared file system for controlling the common use of a large-scale storage system by many computing nodes. The local file system uses SAM-QFS,[note 7] which has a high-speed disk access function and a hierarchical storage management function for tapes. The shared file system uses the Shared Rapid File System (SRFS)[note 8] to achieve high-speed access using the interconnecting network.

4.1 Transparent and high-speed storage

The SAM-QFS system features a disk striping function to provide a high-speed access environment and duplicates data onto tape to provide fault tolerance and a data backup function. The hierarchical storage management function of SAM-QFS enables file access without the user having to know whether a needed file exists on disk or tape. SAM-QFS also features an automatic data-staging mechanism that interfaces with the JAXA job control system and moves files if necessary from the tape library to disks during the job-execution wait period. This function provides users with high-speed disk access to files at all times.

We measured I/O performance to gauge the basic performance of the local file system. This

note 6)　Relay equipment required when building a SAN. These switches can interconnect multiple servers and storage products.

note 7)　Storage and Archive Manager—Quick File System (SAM-QFS): A file system capable of high-speed, large-capacity, and hierarchical storage management; a software product of Sun Microsystems.

note 8)　A distributed file system operating on the FX1 and the SPARC Enterprise system; Fujitsu software.
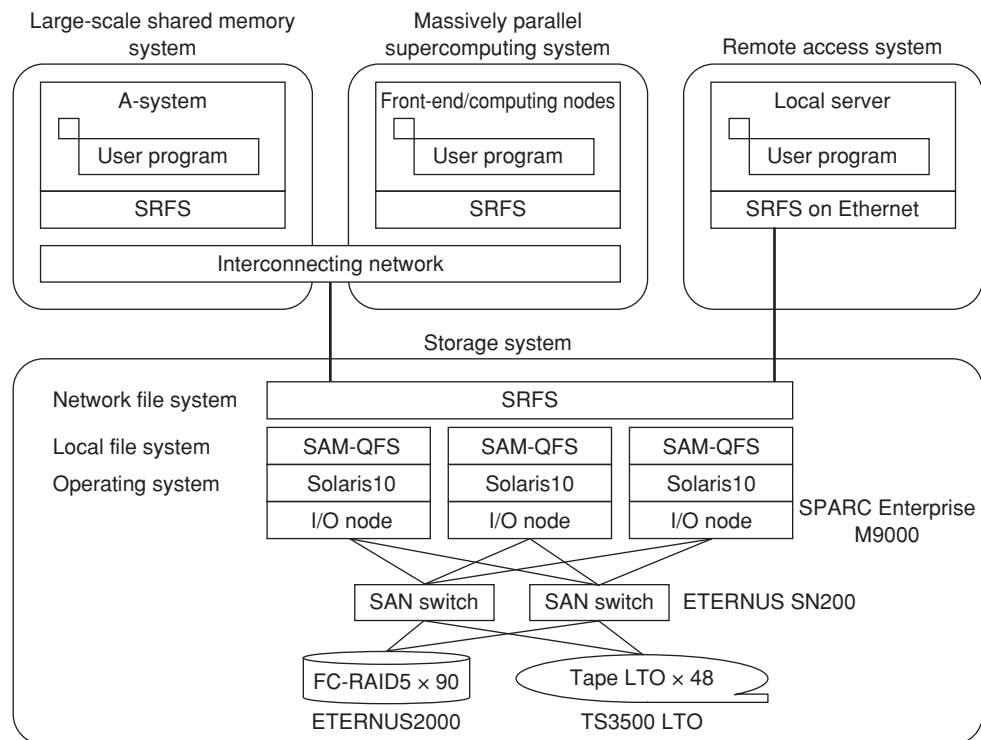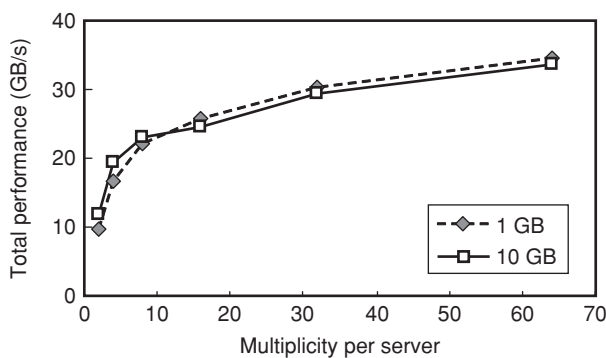
Figure 5
Outline of JSS storage system.

First read performance (15 stripes × 8 groups × 3 file systems)

Figure 6
I/O performance values.

thousand nodes and guarantees consistency of data targeted for updating by multiple nodes. It also has a server cache function to achieve high I/O throughput during massively parallel execution. For example, when the processing of many small files is concentrated in a short period as a result of requests made from SRFS clients to SRFS servers, the local file system may not be able to keep up because of insufficient processing power. However, the caching in SRFS server memory of I/O tasks concentrated in the local file system can help improve I/O turnaround.

4.2 High availability

In previous JAXA storage systems, an I/O server was connected directly to disks and tape equipment, so an I/O-server fault could lead to a system-wide fault that could bring down the entire storage system. In contrast, in JSS, the I/O servers are connected to the disk and tape equipment via SAN switches enabling three or more

system was found to read data at a maximum rate of 33 GB/s. The I/O performance values are shown in **Figure 6**.

The SRFS operates on systems (nodes) linked by the interconnecting network. It achieves file sharing and high-speed input/output of large-capacity data on a scale of several
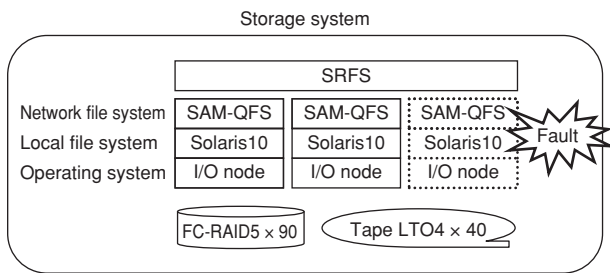
Storage system

Figure 7
Storage switching to maintain availability.

I/O servers to share that disk and tape equipment. Even if an I/O server of the storage system fails, a system-wide crash can be prevented by linking with the common file system of the local file system (SAM-QFS). The concept of storage switching to maintain availability is shown in **Figure 7**.

5. Remote access system

The remote access system consists of the SPARC Enterprise M5000 server and ETERNUS 2000 disk array. Users perform program development on a local server and make use of the computer system at the Chofu Aerospace Center for numerical simulation. In the past, this format required users to be conscious of the remote access system and the file storage locations of the storage system. Now, however, the system coordinates with the JAXA-developed job control system and achieves data transfer linked with job submittal from the remote access system. Thus, the system operates SRFS on an Ethernet network and provides file-system sharing by SRFS between the storage system and local servers. This enables users to use JSS from remote sites without having to worry about file storage locations.

6. Conclusion

This paper outlined JAXA Supercomputer Systems (JSS) and described its features with a focus on the FX1 technical computing server. JSS is the first supercomputing system procured by JAXA since its establishment through the merger of three independent aerospace organizations. In addition to promoting its use in the field of aviation as before, JAXA plans to apply JSS to rocket engine analysis, rocket plume acoustic analysis, and spacecraft design. In this way, JAXA is expected to make a substantial contribution to the use and development of supercomputers in the fields of space development, space science, and space exploration.

Looking to the future, we plan to work on improving the operation and usability of JSS with the FX1 as the core computer with the aim of promoting and expanding research activities in Japan's aerospace industry.

Acknowledgement

References
1) Fujitsu: JAXA Orders Fujitsu Supercomputer. Press release (Feb. 19, 2008).
http://www.fujitsu.com/global/news/pr/archives/month/2008/20080219-01.html
2) Y. Matsuo: Updating the Supercomputer, PLAIN Center News, No.173, 2008. (in Japanese).
http://www.isas.jaxa.jp/docs/PLAINnews/173_contents/173_contents.html
3) Fujitsu: FX1 High-end Technical Computing Server. (in Japanese).
http://jp.fujitsu.com/solutions/hpc/products/fx1.html
4) Fujitsu: FX1 key features and specifications.
http://www.fujitsu.com/downloads/PR/2008/20080219-01a.pdf
5) M. Tanaka et al.: Parallelnavi: A Development and Job Execution Environment for Parallel Programs on PRIMEPOWER (in Japanese), *FUJITSU*, Vol.52, No.1, pp.94-99 (2001).
http://img.jp.fujitsu.com/downloads/jp/jmag/vol52-1/paper20.pdf
6) InfiniBand.
http://www.infinibandta.org/home
7) Linpack Benchmark.
http://www.netlib.org/linpack/
8) STREAM Benchmark.
http://www.cs.virginia.edu/stream/
9) The EuroBen Benchmark.
http://www.euroben.nl/

FUJITSU Sci. Tech. J., Vol.44, No.4, (October 2008)

433

Takayuki Abe
Fujitsu Ltd.
Mr. Abe joined Fujitsu Ltd., Tokyo, Japan in 1988 and has been engaged in design and development of a large-scale system and support of JAXA Supercomputer Systems at the Technical Computing Solutions Unit.

Ken Seki
Fujitsu Ltd.
Mr. Seki joined Fujitsu Ltd., Tokyo, Japan in 1986 and has been engaged in development of hardware for high-performance computing servers at the Technical Computing Development Unit.

Tomohide Inari
Fujitsu Ltd.
Mr. Inari joined Fujitsu Ltd., Tokyo, Japan in 1992 and has been engaged in improving the application performance and support of JAXA Supercomputer Systems at the Technical Computing Solutions Unit.