PRIMEHPC FX10 Specifications

	Node specifications	Theoretical peak performance	236.5 Gigaflops
		Processor	SPARC64 IXfx (1.848 GHz/16-core) x1
		Memory capacity	32GB, 64GB
		Memory bandwidth	85 GB/s
		Interconnect link bandwidth	5 GB/s x2(bi-directional)
	System specifications	Number of racks	4 to 1,024
		Number of compute nodes	384 to 98,304
		Theoretical peak performance	90.8 Teraflops to 23.2 Petaflops
		Total memory capacity	12TB to 6PB
		Interconnect	Tofu Interconnect
		Cooling method	Direct water cooling + air cooling (Option: Exhaust cooling unit)

■ The information in this catalog is subject to change without notice for improvement and other purposes.

First Edition, November 2011

Contact
FUJITSU LIMITED
Website: www.fujitsu.com

Fujitsu PRIMEHPC FX10
Supercomputer

shaping tomorrow with you

Petascale Computing Helps Create a Better World
Fujitsu's PRIMEHPC FX10

Computer simulation is an essential technology that is instrumental in solving many of today's most puzzling and complex problems. It enables organizations to address a large variety of topics from research and development to product design and optimization. It also enables scientists to better address "The Grand Challenges" of understanding our universe. But addressing these topics requires massive amounts of compute power.

Fujitsu's PRIMEHPC FX10 supercomputer provides the ability to address these high magnitude problems by delivering over 23 petaflops, a quantum leap in processing performance.

Ultra-high Speed and Ultra-large Scale Supercomputer
Problems previously constrained or impossible to solve due to performance limits are now able to be handled. This is due to the PRIMEHPC FX10's maximum peak performance of 23.2 Petaflops and memory that scales up to 6 PB with a 98,304 node configuration.

Green Credentials as well as High Performance
Mean Power Savings
In today's quest for a greener world the compromise between high performance and environmental footprint is a major issue. At the heart of PRIMEHPC FX10 are SPARC64 IXfx processors that deliver ultra high performance of 236.5 Gigaflops and superb power efficiency of over 2 Gigaflops per watt.

Application Performance and Simple Development
SPARC64 IXfx processor includes extensions for HPC applications known as HPC-ACE. This plus wide memory bandwidth, high performance Tofu interconnect, advanced compilers and libraries, enable applications to achieve the best performance ever. In addition, the time and effort to adapt to massively parallel processing is reduced through the use of VISIMPACT, which simplifies the implementation of hybrid parallel applications combining MPI and thread parallelism.

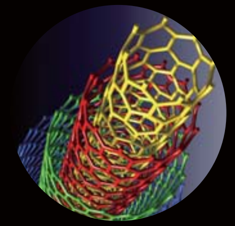
High Reliability and Operability in Large Systems
Incorporating RAS functions, proven on mainframe and high-end SPARC64 servers, SPARC64 IXfx processor delivers higher reliability and operability. The flexible 6D Mesh/Torus architecture of the Tofu interconnect also contributes to overall reliability. The result is outstanding operation: enhanced by the advanced set of system management, monitoring, and job management software, and the highly scalable distributed file system.

Main Application Fields

Disaster prevention
and global environmental issues

Advanced product development

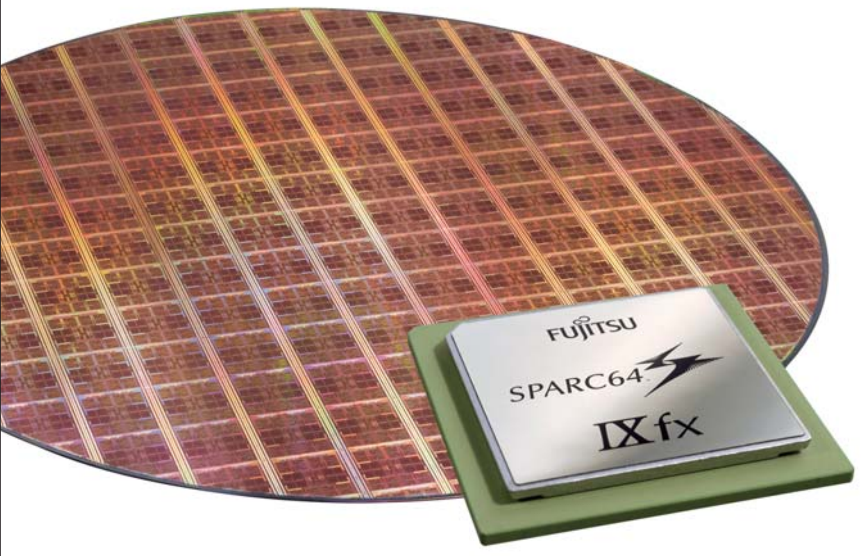
New energy development


Nanotechnology and
new material development

Scientific discovery and mystery
of the universe

New drug development and human science

Architecture for Petascale Computing



Green Supercomputer Credentials - Achieving High Performance with Power Savings

As supercomputer systems grow larger, not only high performance but also low power consumption is extremely important. The PRIMEHPC FX10 has green credentials that deliver high-levels of performance with a low power consumption.

■ SPARC64 IXfx Processors Achieve High Performance While Saving Power

SPARC64 IXfx integrates 16 cores on a chip. Coupled with an excellent power efficient architecture and water cooling system it achieves high performance of 236.5 Gigaflops and excellent power efficiency of over 2 Gigaflops per watt.

HPC-ACE*1, Extensions for HPC

SPARC64 IXfx has extended the SPARC-V9 instruction set architecture for HPC. Application execution performance is greatly improved by the following key enhancements: an increased number of integer and floating-point registers to exploit more instruction level parallelism; the use of sector cache to allow software level control of data to be cached; flexible SIMD instructions which can be applied for loops with conditional branches and non-continuous data processing.

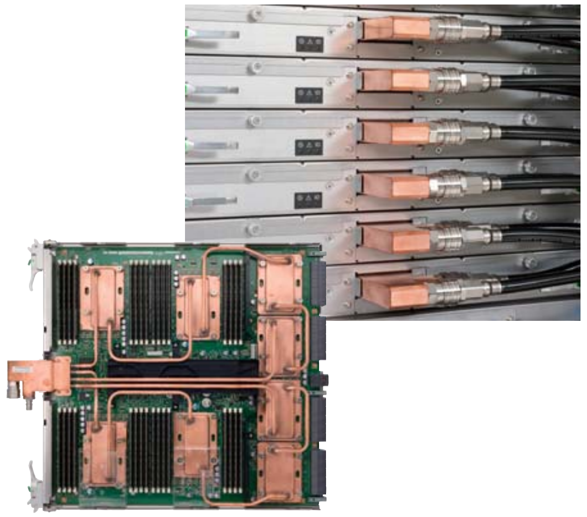
*1 HPC-ACE: High Performance Computing - Arithmetic Computational Extensions

High Memory Bandwidth

SPARC64 IXfx incorporates a memory controller with a wide memory bus, supporting 85 GB/s per processor. With high memory bandwidth and computational power, high application execution performance is achieved.

■Efficient Cooling Reduces Environmental Burdens

Direct water cooling of power consuming components such as processors, reduces exhaust heat, keeps semiconductor temperatures low, reduces current leakage, saves power, and improves reliability. An optional exhaust cooling unit uses water to completely cool any remaining heat generated from the rack. This results in overall minimization of cooling costs in the computer room including that for air-conditioning equipment.



Water Cooled System Board

Massively Parallel Technology Enabling Petascale Computing

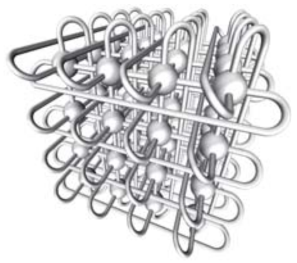
Petascale computing means large systems with huge numbers of computing nodes connected via high speed interconnect. Applications also need to be optimized for such massively parallel environments. PRIMEHPC FX10 provides the right high-speed interconnect to support large configurations, as well as functions for hybrid parallelization suited to massively parallel application processing.

■ Tofu*2 Interconnect Scales beyond 100,000 Nodes

PRIMEHPC FX10 uses the Tofu interconnect, a 6D Mesh/Torus architecture. This comprises up to 98,304 compute nodes with a maximum peak performance of 23.2 petaflops and 6,144 I/O nodes. The application view of the Tofu interconnect is a simple 3D torus.

Each node in the system has a 10-port Tofu link which directly connects it to the 10 adjacent nodes. This delivers low latency communication with high bandwidth of 5 GB/s x 2 (bi-directional) per link. The Tofu interconnect also provides hardware functions to support inter-node barrier and reduction operations and can process collective communications with low latency.

*2 Tofu: Torus Fusion



Tofu Interconnect Topology (Conceptual Diagram)

■ VISIMPACT*3 for Efficient Hybrid Parallel Programming

Hybrid parallel programming, combining MPI and multi threaded parallel programming, is an effective technique for implementing applications in massively parallel environment. PRIMEHPC FX10 includes a facility known as VISIMPACT*3 which helps to implement parallel programs for multicore environments. This facility makes use of the SPARC64 IXfx's inter-core hardware barrier function, 12 MB large-capacity shared L2 cache, and advanced automatic-parallelization functions of the compiler. Through the use of VISIMPACT, MPI based applications can be transformed easily to hybrid parallel application.

*3 VISIMPACT (Virtual Single Processor by Integrated Multicore Parallel Architecture)

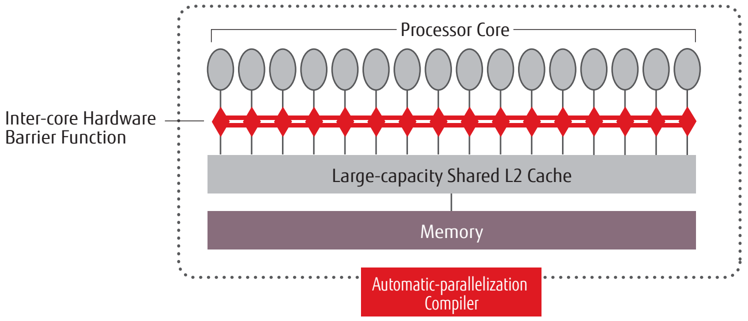

Processor Core
Inter-core Hardware Barrier Function
Large-capacity Shared L2 Cache
Memory
Automatic-parallelization Compiler

Image of VISIMPACT

High Operability and Reliability in Large Systems

Large scale systems contain large numbers of components. It is essential that each individual component is reliable, however the system also needs the capability to continue operation even if a component fails.

PRIMEHPC FX10 is designed to provide high reliability and fault tolerance, as well as high operability and maintainability.

■ High Reliability at Component Level

In PRIMEHPC FX10, high reliability is achieved through a wide range of detection and correction mechanisms covering both internally and externally induced errors. Example of such functions, inherited from mainframe architecture include, processor instruction retry and multi-bit memory error correction by ECC. In addition, direct water cooling reduces processor temperatures and extends component life. This significantly improves the overall reliability of very large systems.

■ Interconnect Supports Fault Tolerance and High Operability

The Tofu interconnect uses a 6D Mesh/Torus architecture to provide high operability and fault tolerance. The Tofu interconnect has the ability to maintain a 3D Torus configuration even when the interconnect is split arbitrarily or where the system has a failed node. This cannot be provided by a the conventional 3D Torus interconnect.

■ High Maintainability during Operation

High maintainability during operation is achieved through the automatic isolation of a failed node and an automated notification system that sends alerts to the Fujitsu call center in the event of a failure.

Highly Efficient I/O System

PRIMEHPC FX10 consists of two types of nodes: compute nodes and I/O nodes. A compute node accesses the external network and storage via the I/O node. PRIMEHPC FX10 supports a multi-tier storage solution including local and global file systems by cooperating with software functions. Since the local file system is reserved for specific jobs during calculation, interference with other user jobs is avoided.

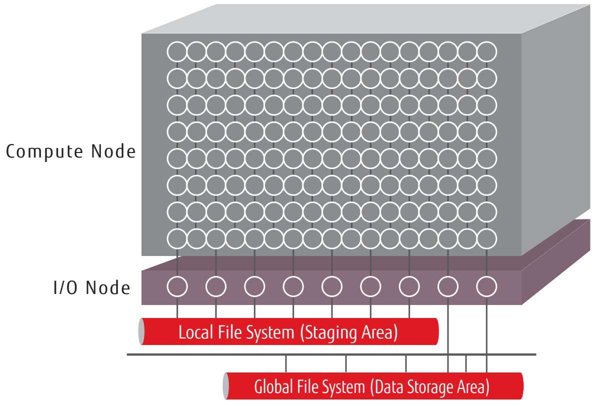

Compute Node
I/O Node
Local File System (Staging Area)
Global File System (Data Storage Area)

Image of PRIMEHPC FX10's I/O Configuration

Software for Petascale Systems

Industry Standard Operating System

The operating system is Linux, which is used as the standard in the HPC field. The portability of applications and tools is maintained, while the functions have been extended.
Extended functions include support for high-speed execution of applications through the use of HPC-ACE, synchronization scheduling and large page, as well as improvements in system operability achieved, for example, by isolating failed processors and memory.

Technical Computing Suite, HPC Middleware for Petascale Computing

System Management and Job operations Management Functions to Enable Efficient Operation in Large Systems

■ **System Management Functions**
The system is managed and operated using an efficient system monitoring and control facility capable of handling a large number of nodes. Management features include hierarchical management of compute nodes, synchronized start and stop across the whole system, simultaneous OS deployment of nodes, and execution of operational commands across specific nodes. The system can also be partitioned, depending on operational requirements.

■ **Job Operations Management Functions**
High system efficiency is maintained using a variety of job operations management functions. These include: resource assignment for jobs optimized to the 6D Mesh/Torus Tofu interconnect architecture; detailed assignment of resources according to resource selection priorities; a back fill function for effective use of unused resources; a deadline scheduling function that does not assign resources during pre-specified periods, such as in a maintenance period.

Large-capacity, High-performance, and Highly Reliable Distributed File System, FEFS

The FEFS distributed file system can be shared by over 100,000 nodes and provides high-capacity, high-performance, and highly reliable file systems. FEFS can support file systems of up to 8 exabytes with approximately 9,000 quadrillion files. It also has the ability to increase capacity and I/O throughput by scaling-out I/O servers.
FEFS achieves high performance through a caching and file striping function. Fault tolerance is supported through I/O server fail-over, file system journaling and other functions. In addition, fine control of the file system can be performed using functions such as the QoS control feature that guarantees I/O bandwidth for each user, and by the quota feature, available at the directory level.
FEFS includes the ability to use a global and local file system. The efficient file staging function between both file systems is provided in cooperation with the job operations management function. Files are arranged in the local file system for optimal access from job processes resulting in high-speed I/O and a reduction in variations in job execution times caused by inconsistent I/O processing times.

A Variety of Language Processing Systems Maximizing Hardware Performance

■ **Fortran/C/C++ Compilers**
International standards compliant Fortran 95, Fortran 2003, C99, and C++ are all part of the integrated development environment. These compilers maximize the performance of the SPARC64 IXfx using the extended register sets, sector cache, and SIMD instructions of HPC-ACE.
Moreover, automatic-parallelization and parallelization using OpenMP are supported, and highly efficient multi threaded processing can be achieved using VISIMPACT cooperating with the hardware functions.

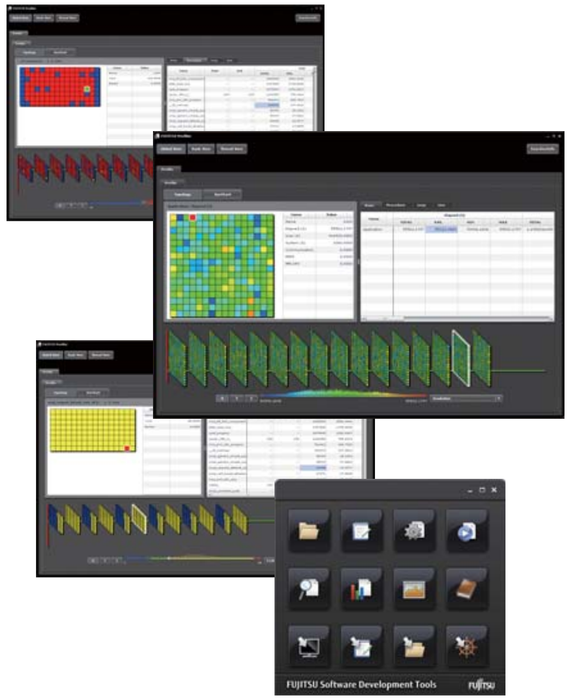
■ **Message Passing Library (MPI)**
Industry standard MPI 2.1 is supported. The MPI library is highly optimized for the Tofu interconnect with increased performance and a small memory footprint. In particular, special algorithms are used for functions that are frequently used in collective communication considering the Tofu interconnect topology. Implementation of a high-speed barrier and reduction function in the Tofu interconnect further reduces parallel application processing times.

■ **Mathematical Libraries**
PRIMEHPC FX10 supports Fujitsu's optimized highly tuned mathematical libraries, SSL II, C-SSL II, and SSL II/MPI as well as industry standard libraries, such as BLAS, LAPACK, and ScaLAPACK. Since the majour routines are highly tuned to exploit performance of the SPARC64 IXfx, applications using these libraries can attain high levels of performance.

■ **Data Parallel Processing Compiler**
XPFortran facilitates the development of parallel applications based on Fortran. Since the XPFortran extensions are designed as directives, an XPFortran program can also be run as an ordinary Fortran program.

Advanced Application Development Environment

A GUI-based development environment provides a unified view through all phases of application development. The development process is accelerated with the use of highly functional development tools such as an interactive debugger usable with sequential and parallel applications written in Fortran, C, and C++ and profiler/tracer which helps with efficient application tuning.



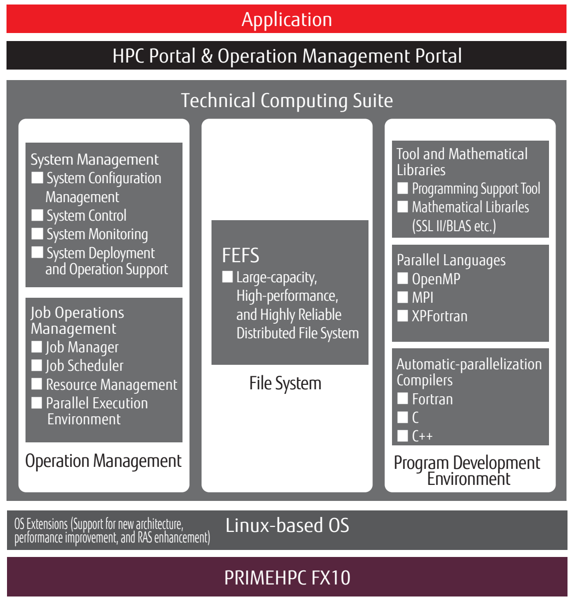
GUI for Apllication Development

User and Operation Management Support Solutions

Portal based solutions are provided to support the user and operational management of the system.

■ **HPC Portal**
HPC Portal is a Web portal that allows end users simple use of the PRIMEHPC FX10. Specifically, users can edit files, compile or submit/monitor/kill jobs, with the convenience of a standard Web browser.

■ **Operation Management Portal**
This Web portal allows administrators to monitor and operate the PRIMEHPC FX10. The portal enables a variety of views that present system operational status, display log information, review resource utilization, as well as manage jobs, and control the power supply via a Web browser. The result is a reduction in operation management work and lower costs for administrator training.


Application
HPC Portal & Operation Management Portal
Technical Computing Suite
System Management
□ System Configuration Management
□ System Control
□ System Monitoring
□ System Deployment and Operation Support
Job Operations Management
□ Job Manager
□ Job Scheduler
□ Resource Management
□ Parallel Execution Environment
Operation Management
FEFS
□ Large-capacity, High-performance, and Highly Reliable Distributed File System
File System
Tool and Mathematical Libraries
□ Programming Support Tool
□ Mathematical Librarles (SSL II/BLAS etc.)
Parallel Languages
□ OpenMP
□ MPI
□ XPFortran
Automatic-parallelization Compilers
□ Fortran
□ C
□ C++
Program Development Environment
OS Extensions (Support for new architecture, performance improvement, and RAS enhancement)
Linux-based OS
PRIMEHPC FX10

Software Structure