White paper
FUJITSU Supercomputer PRIMEHPC FX100 Evolution to the Next Generation

Next Generation Technical Computing Unit
Fujitsu Limited

Contents

FUJITSU Supercomputer PRIMEHPC FX100 System Overview

Introduction

For over 30 years since developing Japan's first supercomputer in 1977, Fujitsu has successively developed leading-edge supercomputers. The FUJITSU Supercomputer PRIMEHPC FX100 is a new supercomputer from Fujitsu, containing a redesigned processor, memory, and interconnect with next-generation technologies to pave the path to exascale computing.

HPC-dedicated design for high performance

The PRIMEHPC FX100 is a massively parallel computer powered by the SPARC64™ XIfx processor and the Torus fusion (Tofu) interconnect 2, both of which have dedicated designs for HPC. The SPARC64™ XIfx processor features 32 compute cores and and instruction set extended for HPC workloads. Its main memory uses the state-of-the-art 3D-stacked Hybrid Memory Cube (HMC), achieving a high memory bandwidth of 480 GB/s. The Tofu interconnect 2 connects nodes with 12.5 GB/s high-speed links to construct a highly scalable 6D mesh/torus network.

Dedicated-core configuration for high parallel efficiency

The SPARC64™ XIfx processor has two assistant cores in addition to the 32 compute cores. To improve the efficiency of parallel processing, the compute cores are dedicated to parallel computation and the assistant cores process other tasks, including I/O. Thus, unlike the PRIMEHPC FX10, there is no I/O node in a PRIMEHPC FX100 system.

Direct water cooling for high reliability

The circulation of coolant through cold plates that cool the processor, memory, optics, and regulators keeps the chip temperature low. This reduces the failure rate of devices and improves reliability.

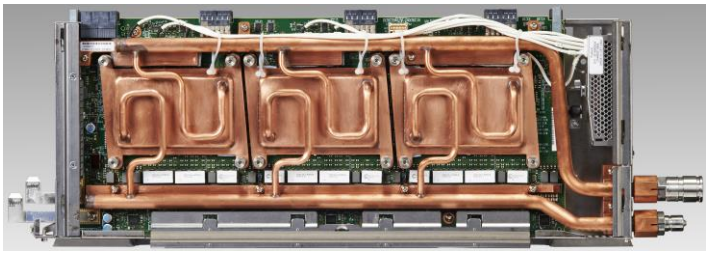

Figure 1 Direct water cooling on CPU memory board

Main unit and rack configuration

The main unit is small, with a size of 2U. Each main unit is populated with four CPU memory boards (CMBs), each of which incorporates three compute nodes, four power supply units, one boot disk, and one low-profile PCI Express expansion slot. The number of nodes in a main unit is 12.

The size of the supported rack is a standard 19 inches. Each rack houses a maximum of 18 main units. The maximum number of nodes per rack is 216.

Figure 2 PRIMEHPC FX100 main unit

System configuration

A PRIMEHPC FX100 five-rack system has a peak performance of over 1 Pflops. Since each PRIMEHPC FX100 system is scalable to over 500 racks, the maximum peak performance exceeds 100 Pflops.

Table 1 PRIMEHPC FX100 system specifications

	5 racks	512 racks
Number of main units	90	9,216
Number of nodes	1,080	110,562
Peak performance	> 1 Pflops	> 110 Pflops
Memory capacity	34 TiB	3.4 TiB
Memory bandwidth	518 TB/s	53 PB/s
Interconnect bandwidth	108 TB/s	11 PB/s
Number of expansion slots	90	9,216
Sample configuration	2x5x9x2x3x2	32x32x9x2x3x2

Figure 3 PRIMEHPC FX100 racks

Many-core Processor SPARC64™ XIfx

Dedicated design for HPC

The SPARC64™ XIfx is a new processor developed for the PRIMEHPC FX100 systems. Its design is based on the SPARC64™ X+. To improve the performance of HPC applications, the SPARC64™ XIfx has increased the number of cores and expanded the SIMD (Single Instruction Multiple Data) width without operating at a higher clock frequency so as to minimize the increase in power consumption.

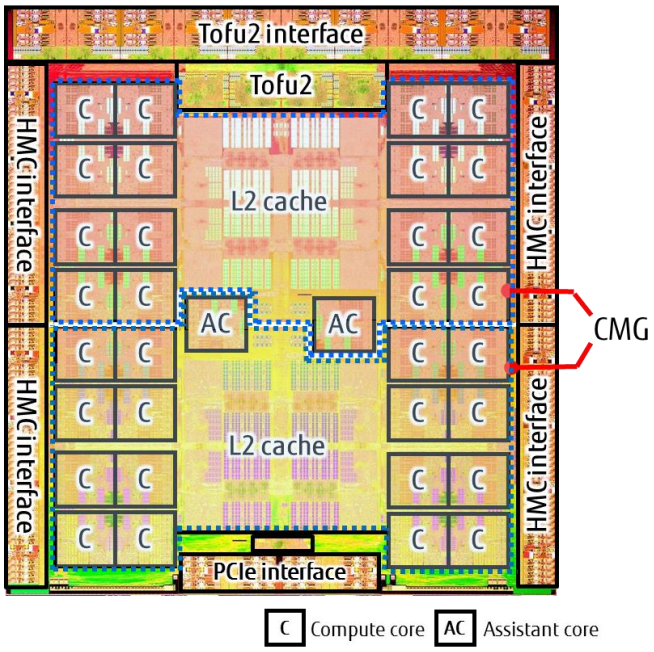

Figure 4 SPARC64™ XIfx

SPARC64™ XIfx overview

The SPARC64™ XIfx is composed of two core memory groups (CMGs), the Tofu2 controller, the PCI Express controller, and other components. Each CMG consists of 16 compute cores, 1 assistant core, 12 MiB of shared L2 cache, and memory controllers. Cache coherency is maintained between the two CMGs. The SPARC64™ XIfx processor uses state-of-the-art 20 nm process technology.

Each core consists of three units: IU (Instruction control Unit), EU (Execution Unit), and SU (Storage Unit). The IU controls the fetch, issue, and completion of instructions. The EU consists of two integer operation units, two address calculation and integer units, and eight floating-point multiply-add units (FMAs), and it performs integer and floating-point operations. A single FMA can execute two double-precision floating-point calculations (add and multiply). A single SIMD instruction can operate four FMAs. Each core executes two SIMD instructions simultaneously. Therefore, each core executes up to 16 double-precision floating-point calculations per clock cycle. The number of calculations doubles if the floating-point numbers are single-precision. The SU executes load/store instructions and contains 64 KiB of Level 1 instruction cache and 64 KiB of data cache per core.

Assistant cores

Process switching on a compute core to a process of the OS or system software degrades parallel processing efficiency. To address this issue, the SPARC64™ XIfx processor has two dedicated cores for the OS and system software.

HPC-ACE2, instruction set expansion for HPC

The SPARC64™ XIfx introduces an instruction set expansion called HPC-ACE2 (High Performance Computing - Arithmetic Computational Extensions 2) that doubles the computational throughput from that of its predecessor, HPC-ACE. HPC-ACE2 is described in detail on the next page.

Hybrid Memory Cube, 3D-stacked memory

To mitigate an increasing performance gap between processors and memory, the SPARC64™ XIfx uses the HMC, which is state-of-the-art 3D-stacked memory, to provide a maximum theoretical throughput of 480 GB/s. The HMC is described in detail on page 6.

Integration of the Tofu interconnect 2

The SPARC64™ XIfx integrates the interconnect controller that used to be a peripheral chip in the previous PRIMEHPC FX10 systems. The newly developed Tofu interconnect 2 uses 40 lanes of high-speed 25 Gbps SerDes and provides high throughput of 250 GB/s in total. The Tofu interconnect 2 is described in detail on pages 7 and 8.

Table 2 SPARC64™ XIfx specifications	
Number of cores	32 + 2
Number of threads per core	1
L2 cache capacity	24 MiB
Peak performance	> 1 Tflops
Theoretical memory bandwidth	240 GB/s x 2 (in/out)
Theoretical interconnect bandwidth	125 GB/s x 2 (in/out)
Process technology	20 nm CMOS
Number of transistors	3.75 billion
Number of signal pins	1,001
HMC SerDes	128 lanes
Tofu2 SerDes	40 lanes
PCIe Gen3 SerDes	16 lanes

Instruction Set Expansion HPC-ACE2

HPC-ACE2 overview

HPC-ACE2 is the successor to HPC-ACE, an HPC-dedicated expansion to the SPARC-V9 instruction set architecture.

SIMD operation

SIMD is a technique for performing the same operation on multiple data items with a single instruction. A single SIMD instruction of HPC-ACE can execute two double- or single-precision floating-point multiply-add operations. HPC-ACE2 doubles the SIMD width to 256 bits, so a single SIMD instruction executes four double-precision or eight single-precision floating-point multiply-add operations.

Extended number of floating-point registers

The 32 floating-point registers in the SPARC-V9 is less than sufficient to maximize the performance of HPC applications. HPC-ACE extended the number of floating-point registers to 256 with a prefix instruction called SXAR (Set eXtended Arithmetic Register). The SPARC-V9 has a fixed 32-bit instruction size, and there is no space to extend the 5-bit register number fields. The SXAR instruction provides additional 3-bit fields to the subsequent one or two floating-point instructions. In HPC-ACE, 256 floating-point registers can be used as one-hundred twenty-eight 128-bit SIMD registers. In HPC-ACE2, the number of SIMD registers stays at 128 while the SIMD register width doubles to 256 bits. The capacity of SIMD registers doubles from that in HPC-ACE.

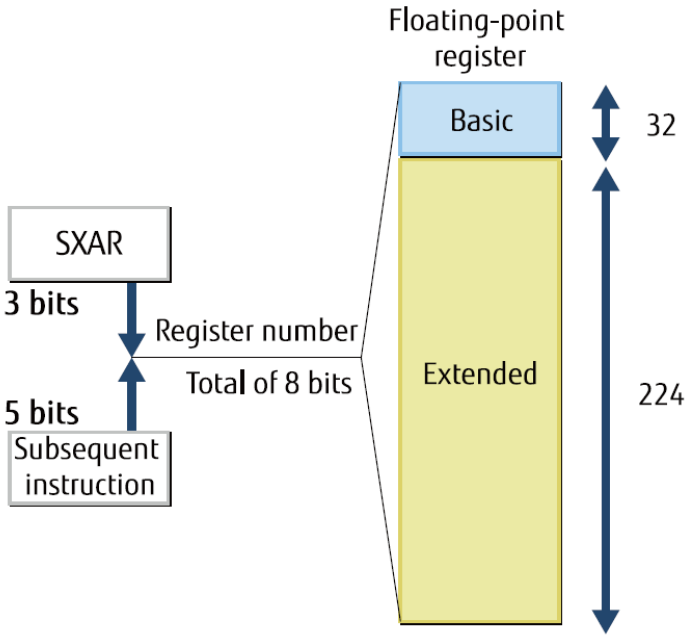

Figure 5 Increased number of floating-point registers

Software-controllable cache (sector cache)

A problem called "memory wall" relates to the growing disparity in speed between processors and the main memory supplying data to the processors. A cache and local memory are common means used to overcome the memory wall problem. The cache is controlled by hardware, with less-frequently reused data possibly pushing out frequently reused data from cache memory. This problem hampers performance improvement. The data access to the local memory is controlled by software, but this requires significant changes to the programs that use the local memory.

HPC-ACE introduced a new type of cache called "sector cache," which can be controlled by software. It combines the advantages of both a cache and local memory. Using the sector cache, software can categorize data into sectors and assign a cache memory capacity for each sector. In HPC-ACE, the L1 data cache and shared L2 cache each have two sectors. HPC-ACE2 doubles the number of sectors to four for both caches. This allows for more sophisticated control such as cache contention avoidance between compute and assistant cores.

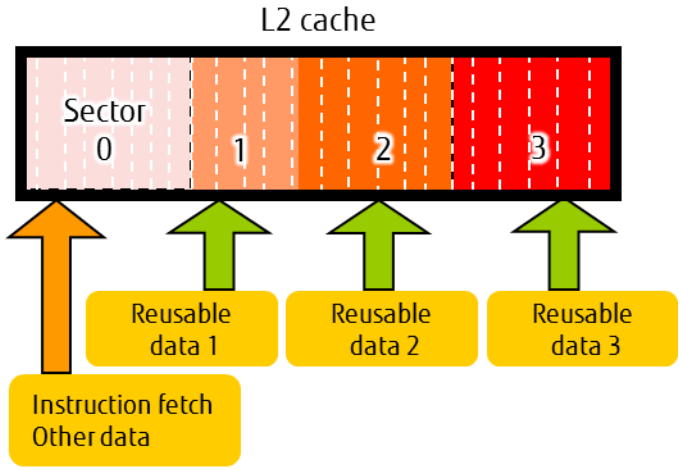

Figure 6 Example of usage of sectors

Arithmetic function auxiliary instructions

HPC applications use arithmetic functions frequently relative to applications in other fields. HPC-ACE introduced instructions to accelerate sin and cos trigonometric functions, division operations, and square root operations. HPC-ACE2 adds instructions for exponential calculation auxiliary and rounding operations.

Stride SIMD load/store instructions

HPC applications often perform parallel computation for multiple data items placed at a regular interval (stride). For a short stride, a single access to a cache line may contain multiple data items. However, load/store instructions accessing a single data item cannot utilize multiple data items on a single cache line. HPC-ACE2 expands stride SIMD load/store instructions to access multiple data items placed at a regular interval. The interval can be specified as a number of elements from two to seven.

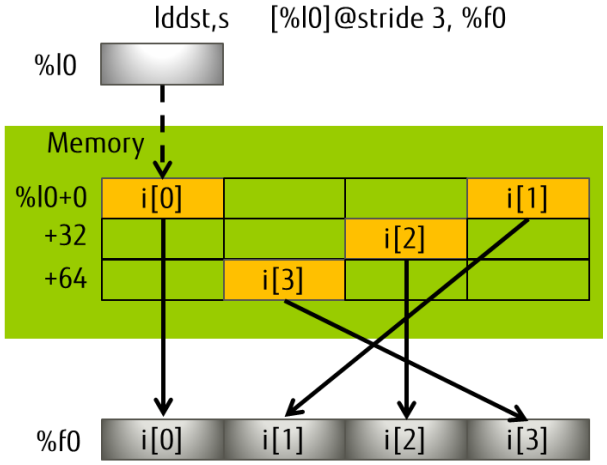

Figure 7 Example of stride 3 SIMD load

Indirect SIMD load/store instructions

HPC applications often access elements of an array indirectly by using an index stored in another array. HPC-ACE2 introduces indirect SIMD load/store instructions to enable SIMD parallel processing of indirect index access.

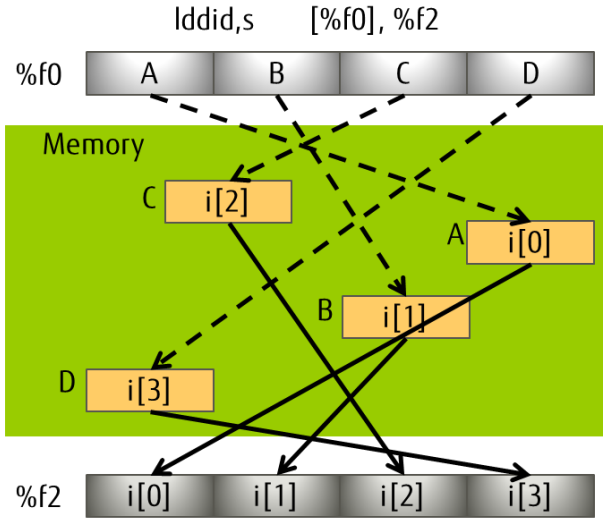

Figure 8 Example of indirect SIMD load

Hardware barrier mechanism for VISIMPACT

For a massively parallel computer like in the PRIMEHPC series, the overhead and memory consumption of communication on a large scale are issues. The hybrid parallel programming model combining process parallelism and thread parallelism is effective at addressing these issues because it reduces the number of processes.

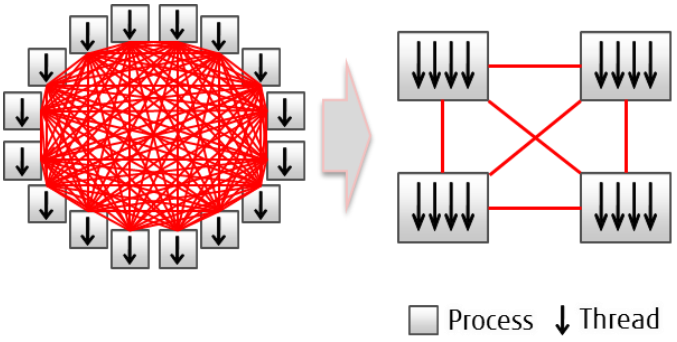

Figure 9 Reduction of number of parallel processes by hybrid parallel programming model

VISIMPACT (Virtual Single Processor by Integrated Multicore Parallel Architecture) is Fujitsu's automatic multi-thread parallelization technology introduced in the FX1. All PRIMEHPC series systems can execute a process parallel program in hybrid parallelism.

The core technology of VISIMPACT in the SPARC64™ processor is a hardware barrier mechanism that synchronizes cores with low latency. To improve the parallel efficiency, the compiler parallelizes a program to a fine-grain multi-thread program assuming low-latency synchronization.

The hardware barrier of the SPARC64™XIfx can synchronize eight groups with an arbitrary number of cores simultaneously to support various combinations of quantities of threads and processes.

3D-stacked Hybrid Memory Cube

The PRIMEHPC FX100 introduces state-of-the-art 3D-stacked memory, the HMC, providing high memory throughput of 480 GB/s. It also contributes to the high-density and water-cooled packaging.

Overview of the HMC

The HMC is a 3D-stacked memory module that stacks multiple DRAM layers and a logic layer using TSV (Through-Silicon Via) technology. The HMC significantly reduces the parts count by stacking multiple DRAM chips in a single module. Its logic layer provides advanced features that are difficult to implement in a DRAM chip, such as error correction, DRAM cell and TSV repair, and a high-speed serial interface. The HMC is an ideal memory solution for HPC requiring a high level of bandwidth, capacity, packaging density, and reliability.

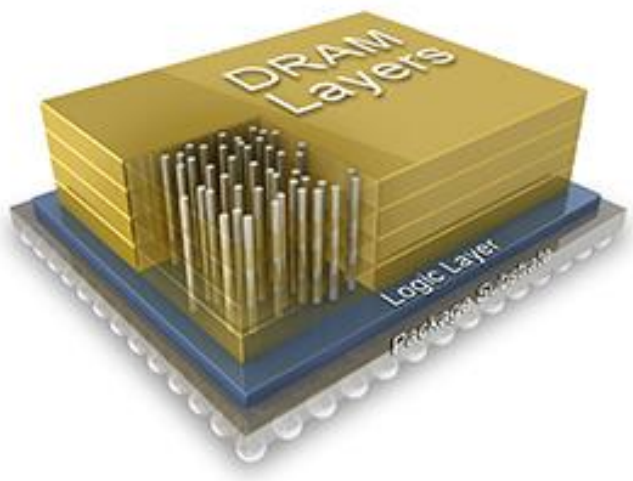

Figure 10 Hybrid Memory Cube structure

High-speed serial transmission

For HPC, not only the memory but also the interconnect requires a high bandwidth. Consequently, they compete for bandwidth, with a limited number of signal pins in a CPU package. The high-speed serial connection of the HMC is ideal for gaining a higher bandwidth with a limited number of pins.

The SPARC64™ XIfx connects 8 HMCs with a total of 128 lanes of 15 Gbps high-speed serial signals and provides a high memory bandwidth of 480 GB/s. Signal traces for memory are routed from two edges of the CPU package, and those for the interconnect are routed from another edge.

High-density packaging

In conventional computer systems, the parts count of DRAM is exceptionally greater than that of other major components. Therefore, a DIMM (Dual Inline Memory Module) is generally used since it has multiple DRAM chips mounted. With DIMMs installed in DIMM slots on a board, a small area of the board can be populated with many DRAM chips.

The HMC's smaller footprint and shorter signal traces than with DIMM slots contribute to the high-density packaging of the PRIMEHPC FX100 that implements 12 nodes per 2U. The packaging density is 1.5 times higher than that of the PRIMEHPC FX10, which has four nodes mounted on a 1U board.

Figure 11 High-density packaging of CMB

Water cooling

Computer systems for HPC and servers require a high memory capacity and implement a large number of DIMM slots on each board. An issue with DIMMs is that they make it difficult to introduce water cooling because of their complex 3D mechanical structure. In the PRIMEHPC FX10, DIMMs were the only major component cooled by air. All other major components were water-cooled.

In the PRIMEHPC FX100, water can cool all the major components including the HMC because they are all board-mounted and can be covered by cold plates easily.

Table 3 PRIMEHPC FX100 main memory specifications

Number of memory cubes per node	8
Memory capacity per node	32 GiB
Theoretical memory bandwidth	240 GB/s x 2 (in/out)
High-speed serial data rate	15 Gbps
Number of high-speed serial lanes	128
High-speed serial pin count	512

Tofu Interconnect 2

The Tofu interconnect 2 (Tofu2) is an interconnect integrated into the SPARC64™ XIfx processor. Tofu2 enhances the bandwidth and functions of the Tofu interconnect (Tofu1) of the previous PRIMEHPC FX10 systems.

6D mesh/torus network

Tofu2 interconnects nodes to construct a system with a 6D mesh/torus network, like with Tofu1. The sizes of the three axes X, Y, and Z vary depending on the system configuration. The sizes of the other three axes, A, B, and C, are fixed at 2, 3, and 2, respectively. Each node has 10 ports.

The network topology from the user's view is a virtual 1D/2D/3D torus. An arbitrary number of dimensions and size of each axis are specified by a user. The virtual torus space is mapped to the 6D mesh/torus network and reflected in the rank numbers. This virtual torus scheme improves the system fault tolerance and availability by enabling the region containing a failed node to be utilized as a torus.

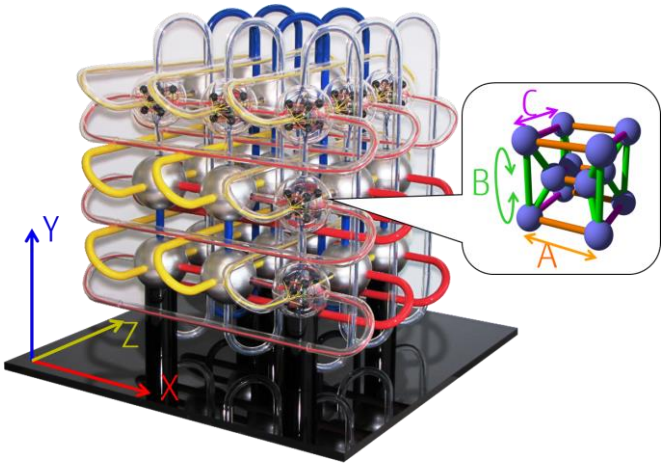

Figure 12 6D mesh/torus topology model

High-speed 25 Gbps serial transmission

Each link of Tofu2 consists of four lanes of signals with a data transfer speed of 25.78125 Gbps and provides peak throughput of 12.5 GB/s. The link bandwidth is 2.5 times higher than that of Tofu1, which uses 8 lanes of 6.25 Gbps signals and provides 5 GB/s of throughput.

Tofu2 connects 12 nodes in a PRIMEHPC FX100 main unit by electrical links, and inter-main unit links use optical transceiver modules because of a large transmission loss at 25 Gbps. Optical transceivers are even placed near the CPU to minimize the transmission loss. In contrast, Tofu1 does not use optical transceivers.

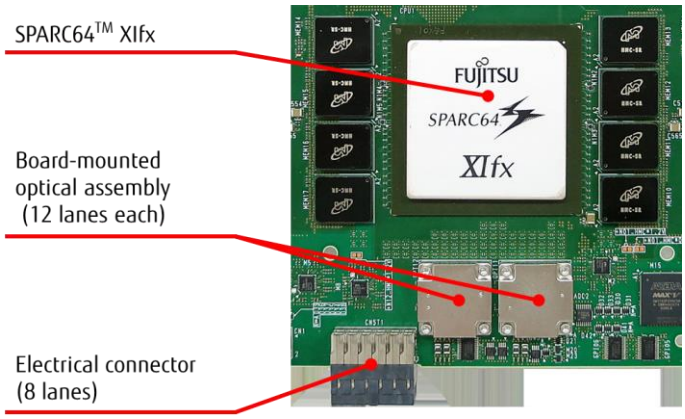

Figure 13 Close placement of optical modules and CPU

Optical-link dominant network

Twelve nodes in a main unit are connected in the configuration of (X, Y, Z, A, B, C) = (1, 1, 3, 2, 1, 2). The number of intra-main unit links is 20 (Figure 14). Therefore, 40 out of 120 ports are used for intra-main unit links, and the other 80 are used for inter-main unit links.

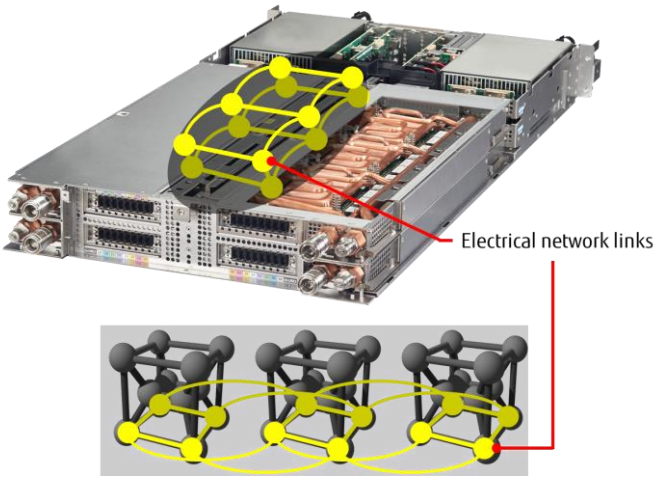

Figure 14 Connection topology in main unit

For conventional HPC interconnects using 10 Gbps generation transmission technology, the ratio of optical links in the total network was up to one-third (A, B, and C in Figure 15). These interconnects partially used optical transmission and only used it to extend the wire length. In contrast, the ratio of optical links in Tofu2 is far beyond that of electrical links. Tofu2 is recognized as a next-generation HPC interconnect that mainly uses optical transmission.

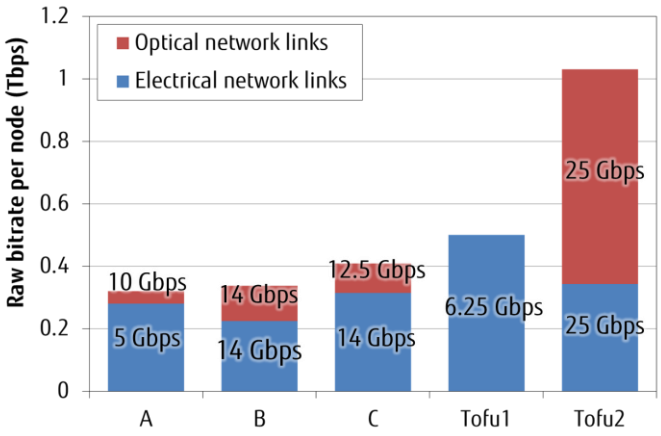

Figure 15 Aggregate raw bitrate comparison per node

RDMA communication functions

Tofu2 supports the new RDMA communication function "Atomic RMW (Read Modify Write)" in addition to Put and Get from Tofu1. Atomic RMW performs a 4- or 8-byte data operation on the destination node. The types of operation include compare-and-swap, swap, integer add, and bit operations.

Atomic RMW guarantees atomicity, which means that there is no other memory access of data while Atomic RMW executes read, modify, and write. Atomic RMW can perform fine-grained mutual exclusion effectively. Atomic RMW mutually guarantees the atomicity of the processor's atomic operations to reduce the overhead of sharing resources between process- and thread-parallel routines.

Communication interface

An RDMA engine receives communication commands and sends notification of results via an interface named CQ (Control Queue). The body of a CQ is placed on the main memory. The set of control registers of a CQ can be mapped independently to a different address space. Mapping the body and the set of control registers of a CQ to a user's address space allows the user process to bypass the OS kernel. The number of CQs for each node is 48, which is larger than the number of cores per node. Using a CQ does not require mutual exclusion because a user process can occupy at least one CQ.

Tofu1 has the direct descriptor function that directly transfers communication commands from CPU registers to the RDMA engine to reduce latency on the sender side. Tofu2 adds a new function called "cache injection" that writes received data directly to L2 cache to reduce latency on the receiver side.

Tofu2 introduces a new session mode for a CQ to allow other processes to control the execution of commands queued on the CQ for automatic execution of collective communications.

Table 4 shows the latencies estimated as the results from logic simulations. Put latency to memory is almost as the same as for Tofu1 (0.91 usec). Cache injection reduces latency by 0.16 usec. The overhead of an Atomic RMW operation is as low as 0.11 usec.

Table 4 Communication latency evaluation results

	Function	Latency
One-way	Put (to memory)	0.87 usec
	Put (to cache)	0.71 usec
Round-trip	Put ping-pong (CPU)	1.42 usec
	Put ping-pong (session)	1.41 usec
	Atomic Read Modify Write	1.53 usec

Tofu barrier

As with Tofu1, Tofu2 supports the Tofu barrier that is an interface for barrier or single element allreduce collective communications. The Tofu barrier implements dedicated logic for packet reception, operation, and transmission to execute various communication algorithms with lower latency than with the CPU. In addition, the collective communication offloaded to hardware is free from OS jitter. The Tofu barrier provides eight independent barrier channels per node.

Table 5 Tofu interconnect 2 specifications

Data transfer speed	25.78125 Gbps
Encoding scheme	64b/66b
Number of lanes per link	4
Theoretical link throughput	12.5 GB/s
Network degrees	10
Network topology	6D mesh/torus
Routing algorithm	Extended dimension order
Number of virtual channels	4
Maximum packet length	1,992 bytes
Packet transfer scheme	Virtual cut-through
Control flow scheme	Credit-based
Delivery guarantee scheme	Link layer retransmission
RDMA communication functions	Put/Get/Atomic RMW
Number of RDMA engines	4
Number of CQs per RDMA engine	12
Address translation scheme	Memory region + page table
Number of Tofu barrier channels	8
Communication protection scheme	Global process ID
Operating frequency	390.625 MHz

Reference
For more information about the PRIMEHPC FX100, contact our sales personnel or visit the following website:

http://www.fujitsu.com/global/products/computing/servers/supercompute/primehpc-fx100/

FUJITSU Supercomputer PRIMEHPC FX100 - Evolution to the Next Generation
Fujitsu Limited
November 17, 2014, First edition
2014-11-17-WW-EN