Tofu: Interconnect for the K computer

● Yuuichirou Ajima ● Tomohiro Inoue ● Shinya Hiramoto
● Toshiyuki Shimizu

Torus fusion (Tofu) is an interconnect for massive parallel computers, and it has been developed to build the K computer that interconnects more than 80 000 nodes. The Tofu interconnect achieves high scalability beyond 100 000 nodes, high performance, high reliability, and high availability. The network topology is a highly scalable six-dimensional mesh/torus. The link throughput is 5 GB/s in each direction. Each node can communicate in four directions simultaneously. The three-dimensional torus rank-mapping scheme improves the system availability and the Tofu barrier interface (TBI) processes collective communications with low latency. Network interfaces and a router of the Tofu interconnect are integrated into a newly developed chip called InterConnect Controller (ICC). This paper describes overviews and characteristics of the ICC chip, the six-dimensional mesh/torus network, high-performance and highly reliable communication functions and the TBI.

1. Introduction

Torus fusion (Tofu) is an interconnect developed for the purpose of achieving scalability of 100 000 nodes, which is two orders of magnitude larger than that of the existing parallel computers that use indirect networks. It has been developed to build the K computer[note)] that interconnects 88 128 nodes. Many technologies have been developed for the Tofu interconnect to achieve high performance, high reliability and high availability with an unprecedentedly large-scale and massively-parallel supercomputer.

This paper outlines and describes the characteristics of a dedicated chip, network and communication functions of the Tofu interconnect. First, an overview of the dedicated chip called InterConnect Controller (ICC) is described. Next, a six-dimensional mesh/torus network is presented. Then, high-performance, highly-reliable communication functions are introduced. Finally, a distinguishing technology called Tofu barrier interface (TBI) is explained.

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.

2. InterConnect Controller (ICC)

ICC is a chip which implements the Tofu interconnect and connects to a SPARC64 processor with a point-to-point connection. An ICC chip is composed of a Tofu network router (TNR), four Tofu network interfaces (TNIs), a TBI and PCI Express (**Figure 1**). A TNR transfers packets of the Tofu interconnect, TNIs are interfaces for the processor which transmit and receive packets to/from the network, and a TBI processes collective communication. A PCI Express connects extension I/O cards and is used with I/O nodes only. A TNR is provided with

10-port Tofu links and an ICC chip uses the Tofu links to interconnect with up to ten ICC chips integrated in other nodes. The specifications of ICC are shown in **Table 1**.

3. Six-dimensional mesh/torus network

3.1 Network configuration

A position in a six-dimensional mesh/torus network is given by six-dimensional coordinates X, Y, Z, A, B and C. The X- and Y-axes are coordinate axes that connect racks and the lengths of the X- and Y-axes correspond to the scale of the system. The Z- and B-axes connect system boards. The Z-axis has an I/O node at coordinate 0 and compute nodes at coordinates 1 and higher. The B-axis connects three system boards in a ring configuration to ensure redundancy. The A- and C-axes are coordinate axes with a length of 2 that connect processors on each system board.

The entire network topology is a structure with groups of ABC three-dimensional mesh/torus with size 2 × 3 × 2 connected by an XYZ three-dimensional mesh/torus. **Figure 2** shows a model representing this topology.

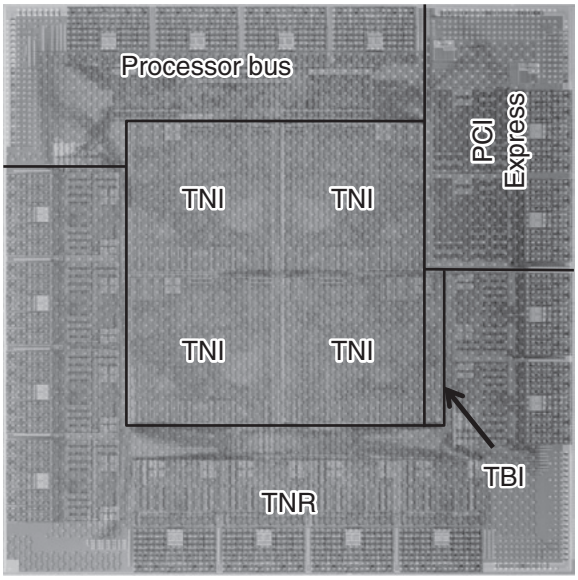

Figure 1
ICC chip.

Table 1
ICC specifications.

Item	Specifications
No. of simultaneous communication	4 transmission connections + 4 reception connections
Operating frequency	312.5 MHz
Switching capacity	100 GB/s
Link rate	5 GB/s × Bidirectional
No. of ports	10
Process technology	65-nm CMOS
Die size	18.2 mm × 18.1 mm
No. of logic gates	48 million gates
No. of SRAM cells	12 million bits
Differential I/O signals Tofu link Processor bus PCI Express	 6.25 Gb/s, 80 lanes 6.25 Gb/s, 32 lanes 5 Gb/s, 16 lanes

3.2 High scalability

With the Tofu interconnect, the number of nodes can be increased by simply connecting cables and the high system scalability beyond 100 000 nodes is achieved. A six-dimensional mesh/torus network is classified as a network called a “direct network” that does not use external switches and is characterized by the fairly high invariability in the average amount of hardware per node regardless of the scale of the system. The average amount of hardware per node required by the Tofu interconnect only includes an ICC chip and approximately 2.2 cables.

3.3 Extended dimension order routing

Packets are routed through coordinate axes in the order of B, C, A, X, Y, Z, A, C and B. The first ABC-axis routing may be addressed to twelve destinations, which can be specified for each transmission command. The communication library specifies the ABC-axis routing path to choose a path to avoid a failure. The system notifies the communication library of the location of a failed node at the start of a job.

3.4 Three-dimensional torus rank mapping

To make it easier to optimize communication patterns using nearest neighbor communication, the Tofu interconnect provides a one-/two-/three-dimensional torus space of a size specified by the user as the user view. A different rank number is provided for each process of an executed user program. The position of each process in the user-specified torus space is identified by a rank number. When a three-dimensional torus is specified, the system forms three spaces using a combination of one of the XYZ axes and one of the ABC axes. The system also assigns a rank number to ensure the adjacency of single-stroke drawing in each space. **Figure 3** shows an example of rank number assignment where a three-dimensional torus with size 8 × 12 × 6 is specified by the user.

3.5 High availability

Even when replacing a part (a failed system board) during maintenance work, the Tofu interconnect allows continued operation of other system boards, thereby ensuring the system has high availability. **Figure 4** shows an example of rank number assignment to avoid the coordinate undergoing maintenance or replacement.

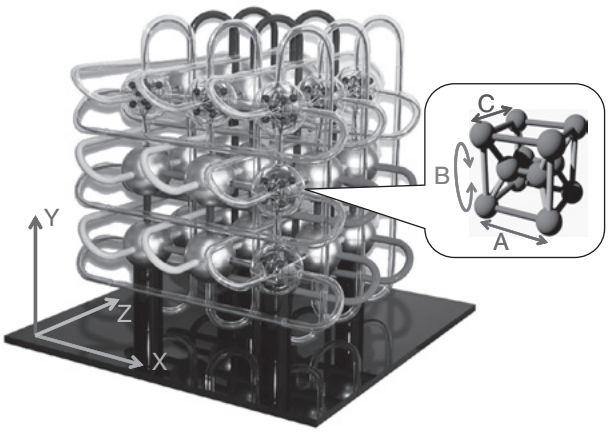

Figure 2
6D mesh/torus topology model.

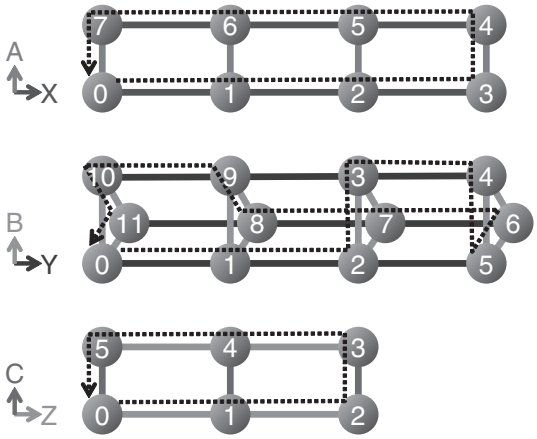

Figure 3
3D torus rank mapping example.

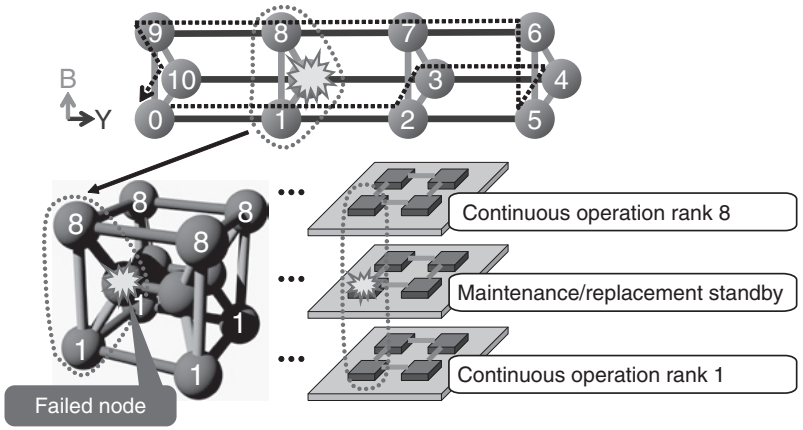

Figure 4
Rank number assignment when replacing a failed system board during maintenance work.

Specifically, the B-axis is exploited in the three-dimensional torus rank mapping. The B-axis is a ring of three nodes and a space containing the B-axis enables single-stroke drawing to avoid a single node.[1)]

4. High-performance, highly-reliable communication

4.1 RDMA communication

The Tofu interconnect provides remote direct memory access (RDMA) communication functions. RDMA is a communication which accesses the specified memory address on the destination node and imposes a low reception processing load on the destination node. The Tofu interconnect provides a virtual to physical address translation mechanism so that RDMA communication accesses a user process memory space. The address translation mechanism includes functions for memory protection, translation cache and address translation table search.

4.2 Low-latency, high-efficiency transfer

The Tofu interconnect has achieved both low-latency packet transfer of approximately 0.1 µs per hop by virtual cut-through switching, and high-efficiency data transfer with a theoretical bandwidth ratio of 90% or higher by a packet length of up to 2 KB. In virtual cut-through switching, the next hop transmission is started when the packet header has been received, which minimizes the latency per hop regardless of the packet length.

4.3 Four-way simultaneous communication

A TNI of the Tofu interconnect processes transmission and reception simultaneously. The four TNIs operate independently and each node is capable of parallel four-way transmission and four-way reception.

The communication library assigns different TNIs to multiple asynchronous transfers according to the destination. In processing a synchronous transfer, the library transfers data through different paths in parallel by splitting and assigning data to multiple TNIs. Collective communication uses multiple TNIs to achieve pipeline transfer in a virtual topology with branches such as trees.

4.4 Virtual channel

The Tofu interconnect employs four virtual channels: two for avoiding routing deadlocks and two for avoiding request and response deadlocks. Each receiver port equips a buffer of 8 KB per virtual channel, or 32 KB in total, and throughput degradation during congestion is mitigated by a newly developed virtual channel scheduling algorithm.[2),3)]

4.5 Link-level retransmission

A Tofu link uses link-level retransmission to correct bit errors in a high-speed transmission link at every hop. As compared with end-to-end retransmission of TCP/IP and InfiniBand, link-level retransmission significantly reduces the performance degradation caused by bit errors. Each transmitter port of a Tofu link has an 8 KB retransmission buffer for link-level retransmission.

4.6 High-reliability design

SRAM and all data path signals in the ICC are protected by error correction code and tolerate soft errors caused by secondary cosmic rays and such like. Signals for all controls except debugging are protected by parity bits and, for important control signals, fault detection has been enhanced by a monitoring state machine that detects abnormal state transitions. Each function module of the ICC is designed to block output signals when any fault is detected so that faults do not propagate beyond module boundaries.

5. Tofu barrier interface (TBI)

The TBI is a hardware module that executes

Barrier, Broadcast, Reduce and AllReduce collective communication processing in the respective nodes instead of software. It features flexibility that allows many algorithms to be implemented in addition to low latency. The TBI has eight barrier channels and is capable of parallel execution of barrier synchronization. One channel is reserved by the system for synchronization scheduling and the remaining seven channels are used by the communication library.

5.1 Low-latency processing

Figure 5 shows the difference between collective communication processing using software and hardware (TBI). When collective communication is processed by software, both the received and transmitted data pass through the main memory, which causes high latency. Communication processing by the TBI does not require main memory access and thus achieves low latency.

5.2 Communication algorithm

Each node is equipped with 64 barrier gates for reception, operation and transmission. By using the number X of barrier gates, a communication algorithm which requires X times of reception/transmission in each node can be realized. The communication library uses an appropriate communication algorithm according to the application. For example, the Recursive Doubling algorithm for N nodes offers low latency because the latency is in proportion to $\log_2 N$ but many barrier gates are consumed because $\log_2 N$ times of reception/transmission take place in every node. The Double Ring algorithm incurs high latency in proportion to N but the number of transmissions/receptions is only two. With the Tree algorithm, the latency is approximately double that of the Recursive Doubling and the number of transmissions/receptions is five, which strikes a good balance between low latency and saving on barrier gate consumption.

5.3 Types of reduce operation

Types of reduce operation supported by the TBI are AND, OR, XOR, MAX and SUM for 64-bit integer numbers and SUM for floating point numbers. In order to obtain identical results with low latency regardless of the order of operations, the floating-point SUM uses an original arithmetic method in which an intermediate result is represented by two 160-bit floating point numbers. The message length supported by the TBI is one element (scalar

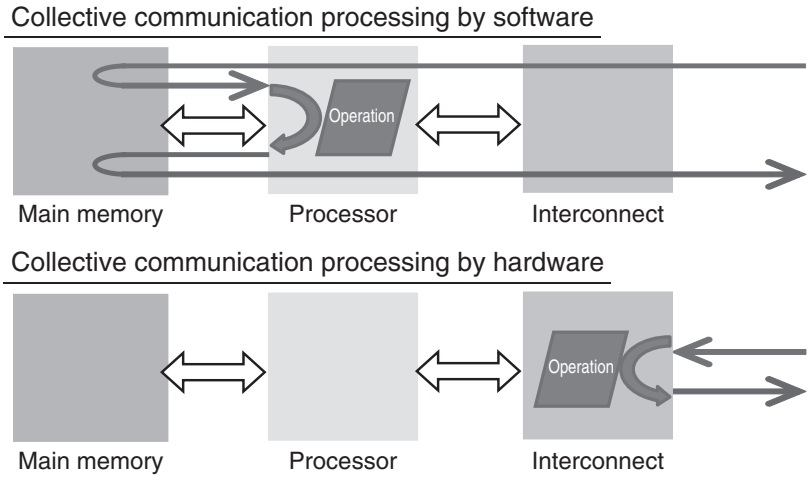

Figure 5
Difference between collective communication processing by software and that by hardware.

data).

5.4 OS jitter

Collective communication processing by software is affected by OS jitter. OS jitter is fluctuation in processing between computation processes by parallel computation, which is caused by interruptions to computation processing due to process switching to a daemon process and such like. A typical interruption time lasts several ten μs to a few ms but, in collective communication, many nodes wait for data and the delay in processing propagates to affect many nodes, which makes the drop in performance more serious. In contrast, collective communication processed by hardware has the advantage of being insusceptible to OS jitter.

6. Conclusion

This paper has presented an overview and description of the characteristics of the design, network and communication functions of the Tofu interconnect, which features scalability of 100 000 nodes. We intend to continue to further enhance and evolve the Tofu interconnect as an interconnect that combines high scalability, high performance, high reliability and high availability. These properties will become more important for future exascale supercomputers.

References

1) Y. Ajima et al.: Tofu: A 6D Mesh/Torus interconnect for Exascale Computers. *IEEE Computer*, Vol. 42, No. 11, pp. 36–40 (2009).
2) Y. Ajima et al.: The Tofu Interconnect. The 19th Annual Symposium on High-Performance Interconnects, pp. 87–94 (2011).
3) Y. Ajima et al.: The Tofu Interconnect. *IEEE Micro*, Vol. 32, Issue 1, pp. 21–31 (2012).

Yuuichirou Ajima
Fujitsu Ltd.
Mr. Ajima is currently engaged in research and development of supercomputers.

Shinya Hiramoto
Fujitsu Ltd.
Mr. Hiramoto is currently engaged in development of supercomputers.

Tomohiro Inoue
Fujitsu Ltd.
Mr. Inoue is currently engaged in development of supercomputers.

Toshiyuki Shimizu
Fujitsu Ltd.
Mr. Shimizu is currently engaged in development of supercomputers.