Hardware Technology of the SX-9 (2) - Internode Switch -

ANDO Noriyuki, KASUGA Yasuhiro, SUZUKI Masaki, YAMAMOTO Takahito

Abstract

The internode connection system of the SX-9 is a dedicated high-speed network with high scalability that accommodates up to 512 nodes. It not only provides high internode data transfer throughput but also implements short communication latency thanks to the functions provided by RCU and IXS. This paper is intended to introduce the configuration, features, functions, performance and architecture of the internode connection system that is built in to the SX-9 system.

Keywords

internode communication, transfer performance, shared memory, crossbar switch, RAS

1. Introduction

Recently, mainstream supercomputers are featuring cluster configuration systems to connect shared memory nodes via a high-speed network (multi-node system) and the advancements in node performances are tending to stress the importance of internode connection network performance. The Supercomputer SX-9 responds to these new needs by achieving a transfer performance 8 times higher than the previous IXS (Internode Crossbar Switch). In the following sections, we introduce the new architectures while focusing on this system.

2. Configuration of the Multi–node System

The multi-node system of the SX-9 series can connect up to 512 nodes by grouping shared memory type single nodes into clusters and connecting them to the IXS ultra-high-speed crossbar switch.

The multi-node system not only features a very wide bandwidth of internode data transfer but also implements a short communication latency thanks to the RCU (Remote Access Control Unit), which is the IXS connection unit at each node and to the functions provided by IXS.

Fig. 1 shows the configuration of the SX-9 multi-node system. Each node incorporates up to 16 RCUs, which are connected to the IXS via cables. Each RCU forms a single lane, which has two connection ports of 4G bytes/sec. × 2, offering a transfer performance of 8G bytes/sec. × 2. As a result, each node can incorporate a maximum of 16 lanes with 32 connection ports, and the total transfer performance reaches a maximum of 128G bytes/sec. × 2.

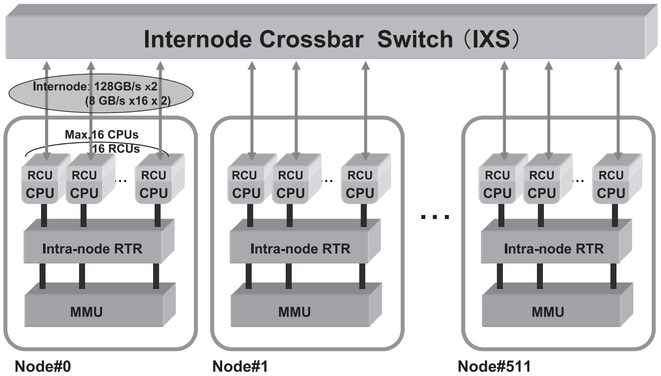

Fig. 1 SX-9 multi-node system configuration.

3. Configuration and Functions of the Remote Access Control Unit (RCU)

3.1 RCU Configuration

The RCU is packaged for each CPU LSI and is composed of the internode transfer controller, global address converter and data transmitter and receiver. The data transmitter and receiver are connected to the intra-node RTR (router), MMU (Main Memory Unit) and the IXS, which is a self-routing crossbar switch, via the cross-bus in the CPU. The internode transfer

Hardware Technology of the SX-9 (2) - Internode Switch -

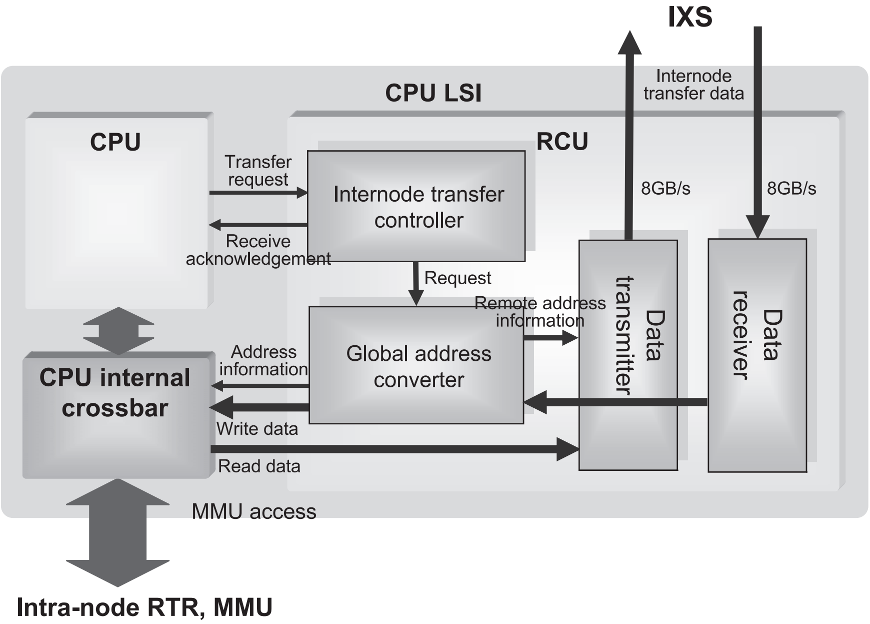

Fig. 2 RCU hardware configuration.

controller receives a transfer request from the CPU and issues an internode transfer request to start the internode data transfer. Each RCU is connected to the CPUs of other CPU LSIs via the intra-node RTR and can receive the transfer requests from the CPUs of other CPU LSIs at the internode transfer controller. The global address converter converts the logical address into a physical address before data transfer. The data transmitter transfers data from the MMU to the IXS with a performance of 8G bytes/sec. per RCU, or a maximum of 128G bytes/sec. per node. The data receiver transfers data from the IXS to the MMU, also with a performance of 8G bytes/sec. per RCU, or a maximum of 128G bytes/sec. per node. The data transmitter and data receiver act independently, thereby achieving a communication bandwidth of a maximum of 128G bytes/sec. × 2 per node (**Fig. 2**).

3.2 Remote Access Function

The CPUs of the SX-9 can transfer data between the memory of another node and that of local node by means of the IXS and RCU. This operation is called the remote memory access and is the fundamental operation of the multi-node system. The RCU has a data mover that functions independently from the CPU operation so that the CPU operations and the memory access can be executed completely concurrently in the internode data transfer.

The data mover inside the RCU supports two types of data transfer instructions (generically called the INA instructions). One of these is the async transfer instruction that prioritizes effective use of the CPU resources and the other is the sync transfer instruction that performs transfer synchronously with CPU operation. Both types of instruction can be executed by non-privileged users in order to minimize the overheads due to system calls during data transfer.

The SX-9 also supports a transfer function for providing short latency for async transfer instructions.

3.3 Asynchronous Ring Buffer

The SX-9 controls the async commands in the async ring buffer that is provided for each job on the MMU using the pointer in the RCU and is capable of queuing about 15,000 async instructions per job. This procedure minimizes the stoppage of subsequent processing by the CPU because of the impossibility of queuing of async instructions. The pointer of the async ring buffer is controlled by the hardware according to the number of configured RCUs, so the software can perform queuing without being concerned about the RCU configuration.

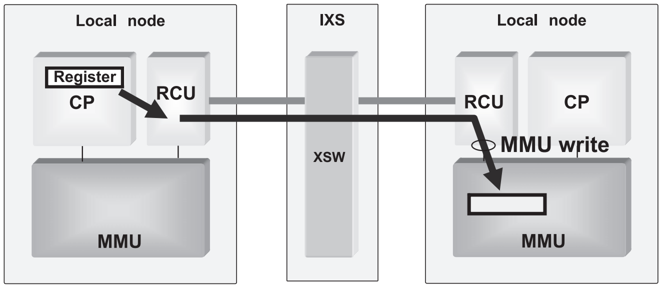

Fig. 3 P2P short message transfer.

3.4 Memory Protection Mechanism

In order to prevent mutual interference between programs and an unexpected destruction of memory, the RCU has the GSATB (Global Storage Address Translation Buffer) for memory protection.

In addition, the concept of the logical node is introduced to support memory protection. In order to enable processing even when the number of physical nodes executing distributed multi-node programs or the quantity of the physical nodes change, the GNATB (Global Node Address Translation Buffer) implements the mapping of logical and physical nodes at the hardware level.

3.5 High–speed Short Message Communication

With the remote access of traditional systems, transfer is performed via the MMU of the local node and that of the remote node, so the latency in internode transfer of short message length, for example the P2P (Peer to Peer) short message transfer has been large, making it impossible to offer an adequate performance.

When the SX-9 transfers such a short message, the RCU that controls the internode communication, writes the value of the scalar register in the CPU directly in the MMU of the remote node and simplifies handshaking between the CPU and the RCU to reduce latency and improve performance (**Fig. 3**).

4. Configuration and Functions of the IXS

The IXS of the SX-9 is composed of dedicated crossbar switches XSWs and enables 16 multi-lane network connections for up to 512 nodes. The transfer performance of 8G bytes/sec. × 2 gives the node a maximum total transfer performance of 128G bytes/sec. × 2 per node.

4.1 IXS Configuration

The XSWs in the IXS have 32 ports with a transfer performance of 4G bytes/sec. × 2 per port, and each port is connected to a port in the RCU with a 1-to-1 relationship. When the number of nodes is up to 32, one XSW is connected directly to the RCU of each node, forming a multi-node system with full-crossbar connections using a single XSW stage. Each RCU in a node is connected to a different lane and multiple lanes are used simultaneously to improve the throughput. When the number of nodes is small, a single XSW physically handles more than one lane but the connection between different lanes is logically isolated in XSW (**Fig. 4**).

When the number of nodes is 33 or more, the fat-tree configuration connecting two XSW stages is adopted. With this configuration, each of the XSWs in the first stage accommodates up to 16 nodes per XSW, and the XSW in the second stage bundles the first-stage XSWs. This arrangement makes it possible to build a multi-node system with up to 512 nodes. In a similar manner to the single-stage configuration usage with 32 or less nodes, the throughput can be improved by the simultaneous operation of multiple lanes, except that the lanes of the multi-stage configuration are physically independent, lane by lane (**Fig. 5**).

4.2 Data Transfer System

While the previous SX-8 adopted circuit switching, the IXS of the SX-9 adopts packet switching. The XSW which is a crossbar switch is equipped with an input buffer for saving packets from the nodes. The crossbar switch elements switch packet assignments to destination nodes, the output buffers for holding packets to be transferred to nodes, and the GCR (Global Communication Register) for use in efficient sync and exclusive control between nodes. The data packets from nodes are routed to the destination nodes according to the destination address information in the headers, and the throughput of this operation is improved by the input and output buffers. The input buffer arrangement is assigned the number of buffers that are sufficient for ensuring adequate throughput even when the length of the cable between the node and the IXS is 40 meters. This is done by considering the node cabinet placement with a maximum 512-node configuration. The crossbar switch elements are divided into two groups in order to reduce

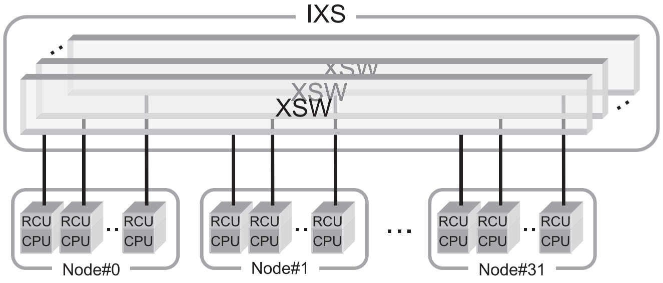

Fig. 4 32-node full-crossbar configuration.

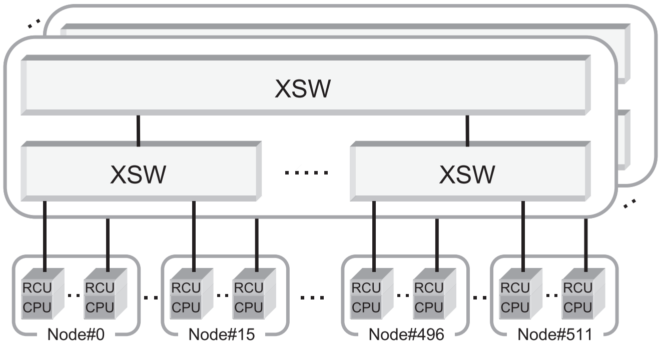

Fig. 5 512-node fat-tree configuration.

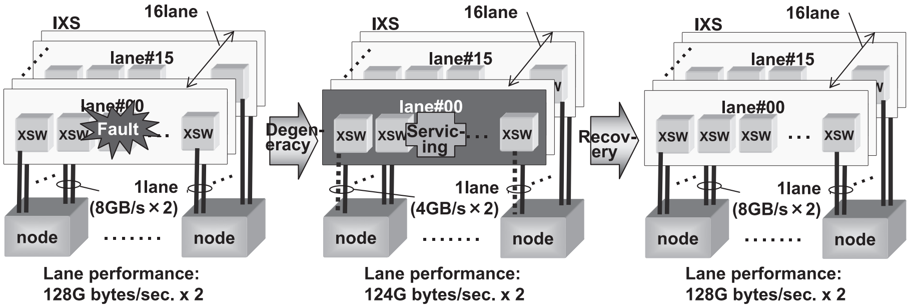

Fig. 6 Lane degeneracy of a multi-node system.

competition between packets and improve the throughput.

5. RAS Functions of Multi-node System

The SX-9 multi-node system is a large-scale system featuring excellent scalability based on the connection of multiple nodes via the IXS. In order to respond to the system performance requirement for high speed, it offers a maximum system performance of 0.8PFLOPS (1.6TFLOPS × 512 nodes).

Since the number of parts increases as the system scale increases, stocks of sufficient amounts of components are required to ensure high availability in preparation for potential failures.

The SX-9 inherits the excellent RAS functions from previous systems but also enhances the system availability by degenerating the performance per lane in case of a fault in the IXSs being introduced due to a new multi-lane connection.

With the multi-node system shown in **Fig. 6** , the internode transfer paths of 8G bytes/sec. × 2 transfer performance per lane are connected to the IXS via up to 16 lanes. Each lane is divided into two paths (4G bytes/sec. × 2 per path,) so that the parallel transfer using the 32 paths (2 paths × 16 lanes) achieves a high maximum internode transfer performance of 128G bytes/sec. × 2 per node.

When a fault occurs in one of the multiple IXS lanes composing the internode network, the system automatically isolates the faulty path in the faulty lane and starts the degenerated operation by switching the lane transfer performance from 8G bytes/sec. × 2 to 4G bytes/sec. × 2.

At such a time, degeneracy is not applied to the transfer paths of non-faulty lanes, so the drop in performance caused by the degenerated operation is minimized to 1/32 (approximately a 3% drop).

The availability of the system is further ensured by re-inputting multi-node jobs during the degeneracy of the lane transfer performance and by enabling re-incorporation of the lane after transfer recovery without stopping the operations at the node.

6. Conclusion

In the above, we introduced the IXS support for the high scalability and performance of the SX-9 multi-node system. In future, we intend to develop products with enhanced functions and performances in order to meet the predicted growth in the HPC usage.

Authors' Profiles

ANDO Noriyuki
Manager,
Computers Division,
1st Computers Operations Unit,
NEC Corporation

KASUGA Yasuhiro
Manager,
Computers Division,
1st Computers Operations Unit,
NEC Corporation

SUZUKI Masaki
Manager,
2nd Computer Technology Dept.,
NEC Computertechno, Ltd.

YAMAMOTO Takahito
Assistant Manager,
2nd Computer Technology Dept.,
NEC Computertechno, Ltd.