Single-Chip, 10-Gigabit Ethernet Switch LSI

● Takeshi Horie ● Takeshi Shimizu ● Akira Hattori

(Manuscript received October 1, 2005)

To develop flexible, highly reliable IT systems, there is an emerging need to provide compact, low-cost, and low-latency 10-Gigabit Ethernet switches to interconnect high-speed servers and large storage systems. To meet this need, Fujitsu has developed the world's first single-chip, 10-Gigabit Ethernet switch LSI. This LSI features twelve 10-Gigabit Ethernet interface ports that support layer-2 switching functions. It has a newly developed I/O circuit called the enhanced 10-Gigabit Attachment Unit Interface (eXAUI) that can transfer 10-Gigabit Ethernet signals over a 25 m copper cable, making it possible to reduce the size, cost, and power consumption of IT systems. The chip has been incorporated into 10-Gigabit Ethernet switches that are now deployed in data centers and high-performance computing applications. This paper describes the key technologies of this LSI, its functions and structure, the eXAUI circuit, and the integration of its circuits. It also includes an evaluation of the LSI's performance and a brief description of a reference board for the LSI.

1. Introduction

To realize flexible and highly reliable systems consisting of high-performance servers and large-capacity storages that are interconnected over networks, a high-performance and general-purpose interconnect technology is required. The 10-Gigabit Ethernet is considered promising as a standard solution to meet these requirements. The existing 10-Gigabit Ethernet switch equipment has been designed for large-scale applications and is intended for edge and wide-area networks. These switches have the disadvantages of a larger physical size, longer fall-through latency, and higher price. Therefore, there is a need for switches that are smaller, less expensive, and have a shorter fall-through latency to interconnect storages and servers, including blade servers, which integrate individual servers into a single enclosure with interconnects over a backplane.

We have developed the world's first single-chip 10-Gigabit Ethernet switch LSI, the MB87Q3140.[note)] The new chip features twelve 10-Ggigabit Ethernet interface ports, layer-2 switching functions, and a newly developed I/O circuit called the enhanced 10-Gigabit Attachment Unit Interface (eXAUI) that can transfer 10-Gigabit Ethernet signals over a 25 m copper cable. This LSI enables single-board, 10-Gigabit Ethernet switch systems to be built and electrical transfer through copper cables and backplane printed circuit board (PCB) connections without using expensive optical modules. As a result, it can reduce the size, cost, and power consumption of IT systems.

This paper describes the key technologies of this LSI, its functions and structure, the eXAUI circuit, and the integration of its circuits. It also includes an evaluation of the LSI's performance

note) The MB87Q3140 is a version of the MB87Q3070 with enhanced features and reduced power consumption.

and a brief description of a reference board for the LSI.

2. Key technologies

This product has four major technical features:

1) 12-port, 10-Gigabit Ethernet switch integrated in a single chip

Existing 10-Gigabit Ethernet switch equipment is physically large because it was designed as general telecommunications equipment; for example, it supports various interfaces, layer-3 and higher functions, and a large buffer memory for long-distance transmission. Therefore, it was difficult to build a single-chip switch LSI. However, by narrowing down the features to the functions needed for interconnects in IT systems and 10-Gigabit Ethernet interfaces, we realized a low-cost, low-power, high-performance single-chip switch LSI. Layer-2 switching is achieved through a new buffer memory sub-system, along with a control scheme for the crossbar switch that interconnects the shared buffer memory with the ports. The chip contains twelve 10-Gigabit Ethernet ports, along with high-speed buffer memories and high-speed I/O macros for switching.[1)]

2) Higher bandwidth of 240 Gb/s

We developed a new on-chip memory subsystem called the multi-port stream memory (MPSM) to effectively utilize multiple memory blocks in the chip and achieve a high-throughput, large-volume, and multi-port shared memory in the chip (**Figure 1**). This shared memory achieves a high bandwidth of 240 Gb/s, which allows the twelve 10-Gigabit Ethernet ports to simultaneously read and write without blocking.

3) Substantially reduced fall-through latency

We developed a new scheduling control scheme for the shared memory to forward the incoming packets to the output ports with a shorter fall-through latency. This substantially reduces the conventional switching latency of several µs or longer to 450 ns.

4) Enabling 25 m copper cable transfer with high-speed I/O circuit

We developed a high-speed I/O circuit, the eXAUI, with equalization circuits in both the transmitter and receiver channels to compensate for frequency-dependent losses.[2)] This I/O circuit enables 25 m copper cable transmissions, which exceeds the 15 m specified in the 10GBASE-CX4 IEEE standard and also enables electrical transfer through copper cables between systems and through backplane PCBs within a system.

3. LSI features

Table 1 lists the principal specifications of the MB87Q3140. A single-chip solution was real-

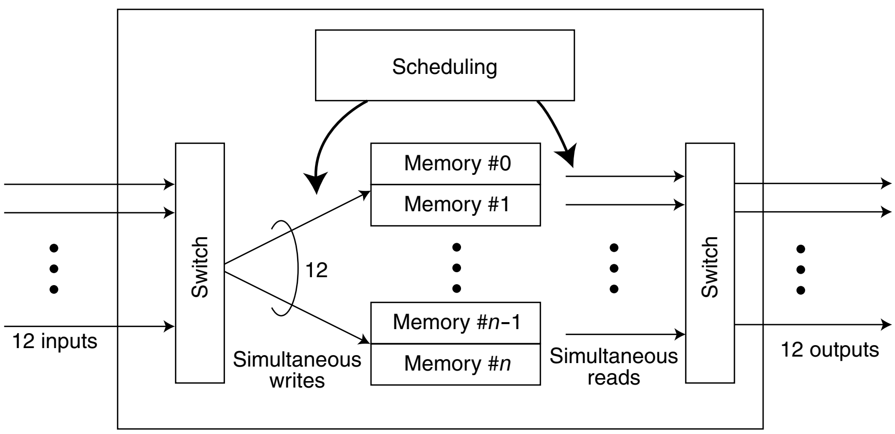

Figure 1
Multi-port stream memory.

Table 1
Principal specifications.

Parameter	Specifications
Number of ports	12
Interface	IEEE802.3ae XAUI and IEEE802.3ak CX4
Switch throughput	240 Gb/s
Fall-through latency	450 ns (under unloaded condition)
Switching system	Cut-through and store-and-forward
MAC address	8 K entry
VLAN	IEEE802.1Q
	Stacked VLAN
Link aggregation	IEEE802.3ad
Jumbo frame	Maximum 15 KB
Spanning tree	IEEE802.1D, IEEE802.1s (MSTP), IEEE802.1w (RSTP)
Layer-3 snooping	IGMP and MLD snooping
Quality of service	IEEE802.1Q/p 4 priority
	Diffserv for IPv4 and IPv6
	Shaper (CIR: committed information rate)
	Meter (PIR: peak information rate)
Port security and control	Source MAC address based filtering
	Broadcast storm control
Flow control	IEEE802.3ae full duplex
Management	RMON, SMON statistical information
	sFLow (RFC1276, wire-speed traffic monitoring)
	Port mirroring
CPU interface	MPC860 and MPC8260

ized by narrowing down the features to layer-2 switching and 10-Gigabit Ethernet interfaces.

For the layer-2 switching functions, the LSI has an 8 K-entry Media Access Control (MAC) address table with address learning and aging in hardware. It can also support a virtual LAN (VLAN) with up to 4 K addresses, which makes it possible to logically divide a network into subnets. The 4 K-address space of VLANs can be extended using stacked VLAN processing.

Support for Spanning Tree Protocol (SPT), which prevents infinite data circulation in a network with loops, enables redundant networks to be configured. A further enhancement of this protocol called Rapid Spanning Tree Protocol (RSTP) enables quick spanning-tree recovery, and Multiple Spanning Tree Protocol (MSTP) can be used to run only one spanning tree per VLAN instance.

By using link aggregation, multiple links can be used to increase the link speed beyond the limits of any single link. This has a two-fold advantage: it provides a higher link throughput and also offers redundant links for reliable and fail-safe communications.

Internet Group Multicast Protocol (IGMP) and Multicast Listener Discovery (MLD) are layer-3 protocols for establishing membership in a multicast group. With IGMP/MLD snooping, the multicast traffic of a group is only forwarded to ports that have members of that group; therefore, IGMP/MLS snooping significantly reduces multicast traffic without generating additional network traffic.

To differentiate the type and level of services for network traffic, IEEE802.1Q/p 4 priority tags and Differentiated Service (Diffserv) Code Point for IPv4 and IPv6 can be used for classification by mapping into four priorities. In addition to packet priority control, to improve the QoS, Meter is used for per port ingress rate limiting and Shaper is used for traffic shaping at each egress port.

The security of IT systems has recently become more important, and the LSI has port security features that prevent unauthorized access to the ports using secure source MAC addresses. The interface forwards only packets having source MAC addresses that match these MAC secure addresses.

The LSI has an interface for an MPC860 or MPC8260 CPU and can be initialized from EEPROM, which makes it possible to build an unmanaged switch system without switch management.

4. LSI configuration

Figure 2 shows the block diagram of the MB87Q3140.

Each port block consists of an eXAUI I/O circuit, 10-Gigabit Ethernet MAC, frame-filtering block, and input buffer. The input buffer is for store-and-forward and speed matching. The MPSM for the 12-port shared memory consists of a memory buffer, crossbar switch, and shared memory control. The LSI also contains a routing table, statistics counter, and CPU interface.

The MPSM has multiple memory banks connected with multi-stage networks. Its control scheme enables continuous access to variable-length data and uses credit-based flow control to realize cut-through switching, which forwards a packet before it has been completely received. The MPSM can achieve high-throughput and efficient multicast communication by using a shared buffer memory.

Note that the MPSM's architecture is

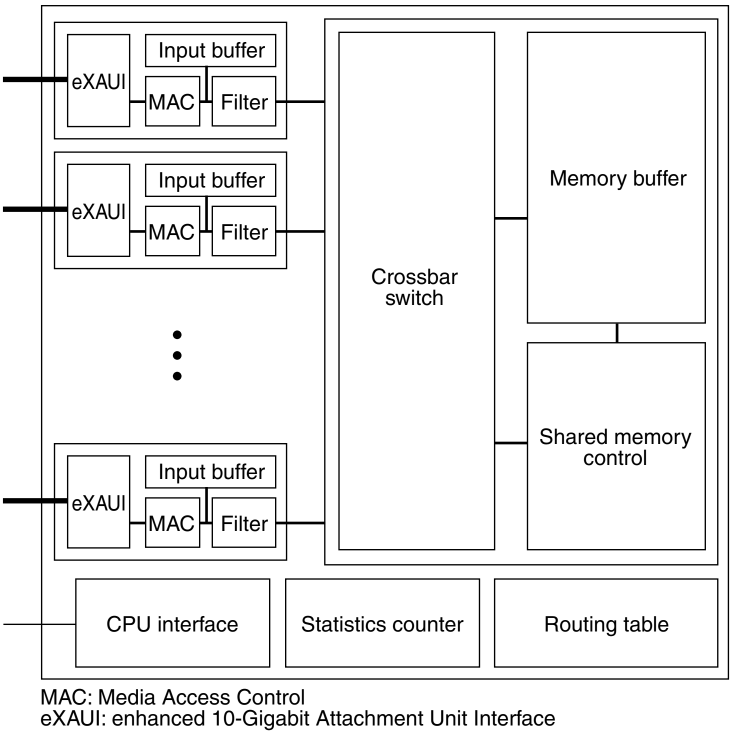

Figure 2
MB87Q3140 block diagram.

protocol-independent and does not need a special memory; therefore, it can easily be applied to LSIs that support protocols other than Ethernet.

5. eXAUI

The high-speed eXAUI I/O complies with the XAUI specification of the 10-Gibabit Ethernet IEEE 802.3ae standard and the 10GBASE-CX4 IEEE802.3ak standard. Therefore, the eXAUI can connect not only with optical modules and 10-Gigabit Ethernet LSIs with an XAUI interface, but also with 10-Gigabit Ethernet adaptor cards with a CX4 interface.

In 10-Gigabit Ethernet data transmission through long cables or interconnects with PCB traces and connectors, frequency-dependent losses cause inter-symbol interference (ISI). To compensate for signal distortion due to ISI, the eXAUI is equipped with equalization circuits in both the transmitter and receiver channels (**Figure 3**). The eXAUI can transfer data over a backplane of more than 1 m or up to 25 m of copper cable; it therefore exceeds the distances specified in XAUI and 10GBASE-CX4.

The eXAUI macro has four channels, and each channel can transfer data at 3.125 Gb/s. The transmitter has a 5-tap Finite Impulse Response (FIR) filter circuit, while the receiver has a second-order-derivative filter. The transmitter and receiver equalizers can compensate for losses of up to about 30 dB. The equalizer parameters can be flexibly controlled; also, adaptive equalization can be performed to compensate for high-frequency losses on a transmission line by using the residual ISI monitor circuit contained in the eXAUI.

To reduce power consumption, the eXAUI has a port power-down mode that can be set individually for each port. It also has a receiver equalizer disable mode that can be used for less lossy transmission lines. These modes are useful in power-sensitive applications such as blade server switch systems.

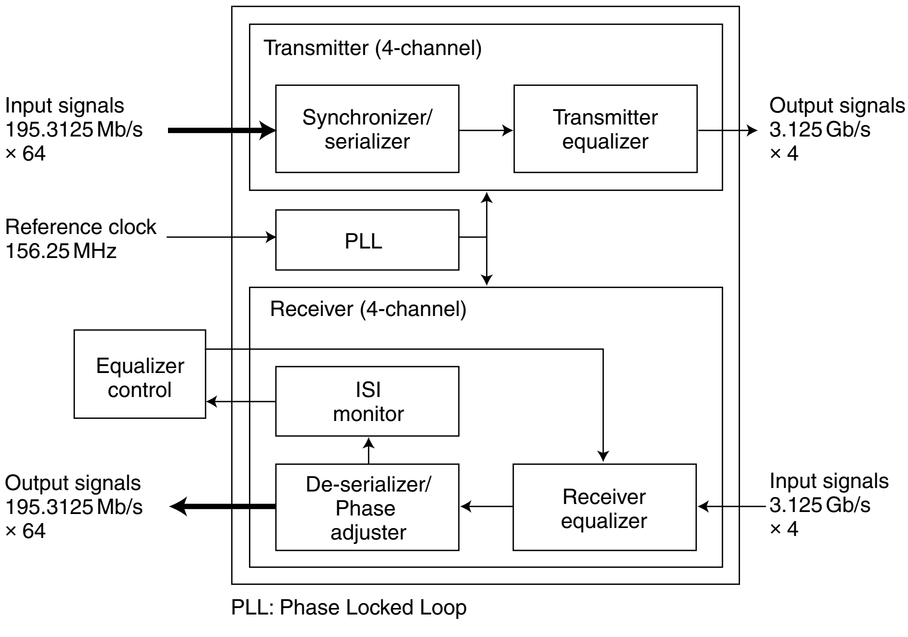

Figure 3
eXAUI configuration.

6. LSI implementation

Figure 4 shows a photograph of the MB87Q3140 chip.

The LSI uses 0.11 µm CMOS technology and operates at 312.5 MHz with a 1.2 V and 2.5 V power supply. The chip is 16 mm × 16 mm, the package is a 728-pin Flip Chip Ball Grid Array (FCBGA, 35 mm × 35 mm), and the high-speed signals on the package were routed to avoid crosstalk between the transmitter and receiver lines.

The LSI floorplan was designed symmetrically with eXAUI macros on each side and regular placement of internal blocks so timing closure between blocks could be achieved more quickly.

One of the challenges in designing a single-LSI implementation that has many high-speed I/Os like the MB87Q3140 is to prevent interference from the digital circuits from affecting noise-sensitive analog circuits. This was achieved for the analog circuits of the eXAUI by completely isolating the analog power and ground planes from the digital power and ground planes and by using on-chip bypass capacitors. Furthermore, the eXAUI circuit was designed to be highly noise tolerant, for example, by using a triple-well CMOS structure.

To reduce power consumption, clock gating for the logic and SRAM has been aggressively used. When the eXAUI's receiver equalizers are disabled, the LSI consumes about 16 W under typical 100% traffic conditions.

7. LSI performance evaluation

In this section, we present evaluation results for the switching performance of the MB87Q3140 when sending and receiving ports of the LSI are paired with each other. Under maximum traffic conditions, the LSI achieved almost a 100% 10-Gigabit wire-speed at each port, even for small Ethernet frames, which cause especially severe conditions (**Figure 5**).

Figure 6 shows the results of switching la-

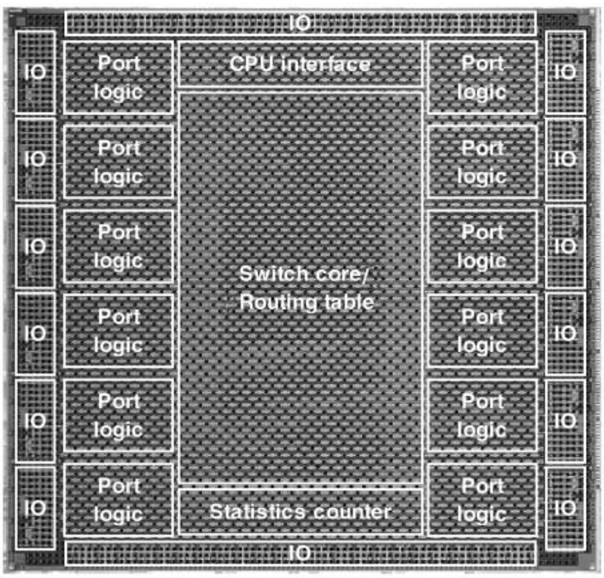

Figure 4
MB87Q3140 chip photograph.

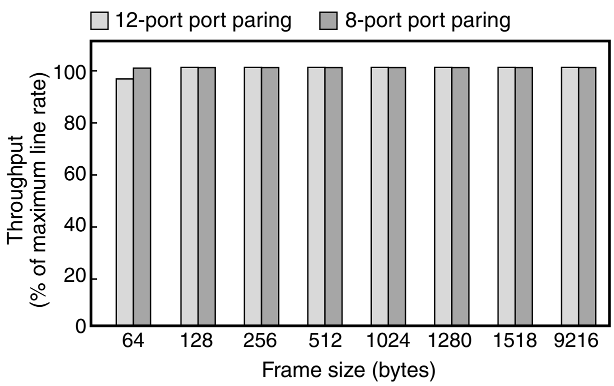

Figure 5
MB87Q3140 throughput evaluation.

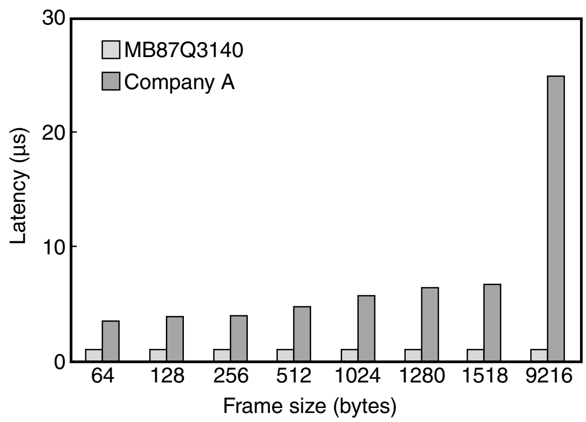

Figure 6
MB87Q3140 latency evaluation.

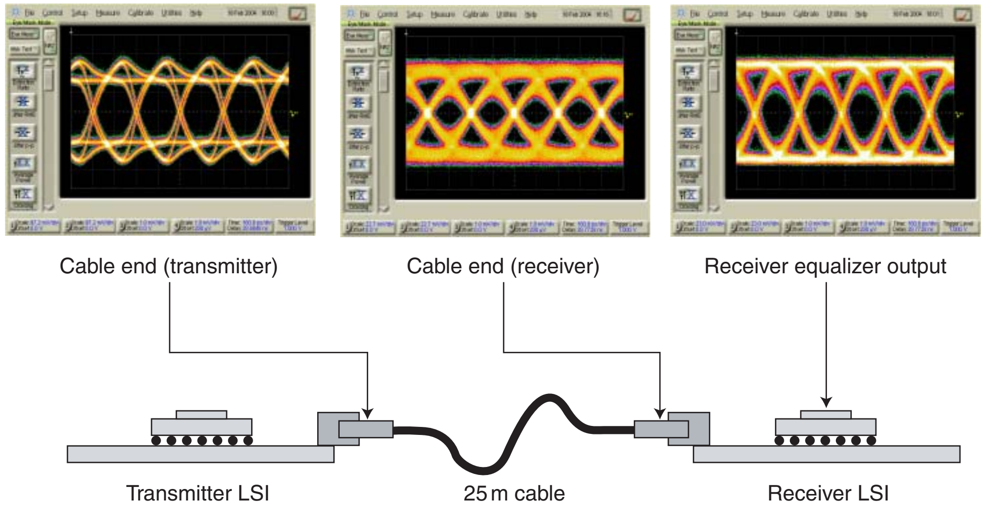

Figure 7
eXAUI evaluation.

tency measurements we made for the LSI and another manufacturer's solution. The figure shows the LSI achieved a significantly shorter latency than the competing solution, which required a multiple-chip configuration. An optical interface was used for the measurements.

We also evaluated the eXAUI. We made the transmitter output compliant with 10GBASE-CX4 using the transmitter equalization circuits and then observed whether the signal could be received through over-specification 25 m copper cables. **Figure 7** shows the measured eye patterns at the transmitter cable end (left) and receiver cable end (middle) and the output eye pattern from the receiver equalization circuit. From these results, we confirmed that the receiver equalization can open the eye.

8. Reference board

We developed a reference board to help customers evaluate the MB87Q3140 and develop MB87Q3140 systems more quickly (**Figure 8**). The board has 10-Gigabit Ethernet interfaces for the optical modules of XENPAK and XFP and for 10GBASE-CX4. It has pre-installed switch software to facilitate the testing of interconnections with equipment that supports the 10-Gigabit Ethernet interface.

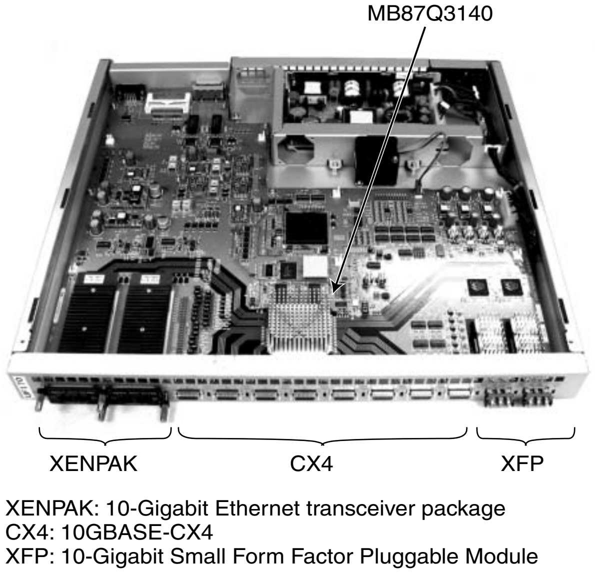

Figure 8
Reference board.

9. Conclusion

This paper described the world's first single-chip, 10-Gigabit Ethernet switch LSI, which was developed for high-speed interconnects to realize an IT systems infrastructure with high performance and reliability.

This new LSI enables 10-Gigabit Ethernet interconnection of servers and storages, which was previously difficult to do, and makes it possible to realize high-performance, low-cost, and low-power switches in blade servers in a small form factor. We have already developed 10-Gigabit Ethernet switches that incorporate this LSI, and they are now in use in data centers and high-performance computers. In the future, we will develop further-advanced switch chips with enhanced features, higher performance, and lower power consumption.

The development of this chip was partially funded by the New Energy and Industrial Technology Development Organization (NEDO) under the project name, "Research and Development of High-Reliability and Low-Power-Consumption Servers."

References

1) T. Shimizu, Y. Nakagawa, S. Pathi, Y. Umezawa, T. Miyoshi, Y. Koyanagi, T. Horie, and A. Hattori: A Single Chip Shared Memory Switch with Twelve 10 Gb Ethernet Ports. Hot Chips 15, August 2003.
2) W. Gai, Y. Hidaka, Y. Koyanagi, J. H. Jiang, H. Osone, and T. Horie: A 4-Channel 3.125 Gb/s/ch CMOS Transceiver with 30 dB Equalization. 2004 Symposium on VLSI Circuits, 2004, p.138-11.

Takeshi Horie, *Fujitsu Laboratories of America*
Mr. Horie received the B.S. degree in Electrical Engineering, M.S. degree in Electronics Engineering, and Ph.D. degree in Engineering from the University of Tokyo, Tokyo, Japan in 1984, 1986, and 2003, respectively. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1986. In 2001, he joined Fujitsu Laboratories of America, Inc., where he has been engaged in research and development of parallel computer systems and networks. He received the IEICE outstanding paper award in 1993 from the Information and Communication Engineers (IEICE) of Japan and the Sakai Memorial Award and Industrial Achievement Award from the Processing Society of Japan (IPSJ) in 1995 and 2005, respectively. He also received the OHM Technology Award from the Promotion Foundation for Electrical Science and Engineering in 2004. He is a member of the IPSJ and the Institute of Electrical and Electronics Engineers (IEEE).

Takeshi Shimizu, *Fujitsu Laboratories of America*
Mr. Shimizu received the B.E. degree in Mathematical Engineering and Information Physics and the M.E. and D.E. degrees in Information Engineering from the University of Tokyo, Tokyo, Japan in 1988, 1990, and 1993, respectively. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1993. In 2001, he joined Fujitsu Laboratories of America Inc., Sunnyvale, California, USA, where he has been engaged in research and development of high-performance processors, parallel computer systems, and interconnection networks. He received the Industrial Achievement Award from the Information Processing Society of Japan (IPSJ) in 2005. He is a member of the IPSJ and IEEE-CS.

Akira Hattori, *Fujitsu Laboratories Ltd.*
Mr. Hattori received the B.S. and M.S. degrees in Electronics Engineering from Osaka University, Osaka, Japan in 1972 and 1974, respectively. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1974, where he has been engaged in research and development of computer architecture and parallel computer systems. Since 2001, he has been engaged in research and development of high-performance interconnects. He received the OHM Technology Award from the Promotion Foundation for Electrical Science and Engineering in 2004 and the Industrial Achievement Award from the Processing Society of Japan (IPSJ) in 2005. He is a member of the Information Processing Society of Japan (IPSJ) and the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.