10 Gigabit Ethernet Switch Blade for Large-Scale Blade Servers

● Yoichi Koyanagi ● Tadafusa Niinomi ● Yasushi Umezawa

(Manuscript received April 8, 2009)

The introduction of blade servers for diverse applications is expanding rapidly. The switch blade—a key network component of the blade server—must provide high performance to accommodate growing data-communication demands generated by network consolidation as in Fibre Channel over Ethernet (FCoE), have a compact, low-power-consumption design, and have switch control software optimized for blade server systems. To meet these requirements, we have developed a 10 Gigabit Ethernet (10GbE) switch blade suitable for large-scale blade servers. This paper focuses on the 10GbE switch LSI MB86C69, whose features have been optimized for blade servers, the high-speed transceiver circuit integrated into this LSI, and the switch control software for easy network configuration of switch blades.

1. Introduction

Thanks to its high-performance and high-density features, the blade server is expected to be deployed in increasing numbers as servers in data centers and corporate machine rooms are consolidated. In response to this trend, we have developed a compact, low-power 10 Gigabit Ethernet (10GbE) switch blade to make possible high-performance, high-density, and easy-to-operate large-scale blade servers. A key component of a blade server is the switch blade, which provides network functions. It must satisfy three requirements: high performance, high density and low power, and easy operation.

1) High performance

The switch blade's performance must be high enough to eliminate communication bottlenecks between servers, between a server and external storage, and between networks. There is a great need for switch blades that can operate at significantly higher performance levels than existing equipment to satisfy growing data-communication demands. These demands originate in expectations for higher server operating efficiency due to the use of virtualization techniques and the spread of Fibre Channel over Ethernet (FCoE) technology for integrating storage-oriented networks with Ethernet systems.

2) High density and low power

In addition to high performance, a switch blade must have a compact configuration and low power consumption suitable for a blade server with densely installed server blades.

3) Easy operation

The switch blade needs software that provides switch functions and operating functions optimized for blade servers having a structure that integrates servers and network.

To meet these requirements, we developed a 10GbE switch LSI (MB86C69) having high-performance and low-latency features and mounted it on a switch blade. This LSI implements the performance and functions required of a blade-server switch, and it incorporates a high-speed transceiver circuit for achieving high-speed

signal transmission at 10 Gb/s. In this way, it achieves both high-density mounting and low power consumption. We also developed a function for simplifying the configuration operations of a switch blade for blade-server use and loaded it in software (firmware) resident on a control processor mounted on the switch blade.

This paper begins by outlining Fujitsu's BX900 large-scale blade server and the abovementioned 10GbE switch blade. It then describes the switch LSI that we developed, the high-speed transceiver circuit mounted on this LSI, and the switch control software.

2. Blade server and switch blade

An external view of the BX900 blade server is shown in **Figure 1**. Up to 18 server blades—each mounting a CPU, memory, and chip set—can be installed from the front of the chassis. In addition, up to 8 switch blades for handling data communications between server blades and between a server blade and external equipment can be installed from the back of the chassis. Up to 2 management blades having functions for managing the entire blade server can also be installed from the back.

A block diagram of the BX900 blade server is shown in **Figure 2**. The MB86C69 LSI on a switch blade communicates with the server blade's motherboard or with a 10GbE adaptor on an application board (mezzanine card) by high-speed signals at a bit rate of 10 Gb/s. Installing multiple switch blades in the blade server makes it possible to increase the communications performance between server blades or with external equipment, and the independent communication paths established among the multiple switch blades provide redundancy that enables the construction of a highly reliable system. Moreover, the switch blades are interconnected with the management blades, the management local area network (LAN), and the control bus, and they support reliability, availability, and serviceability (RAS) functions such as error monitoring and temperature monitoring managed centrally by the management blades.

In addition to the MB86C69 LSI that provides switch functions, the switch blade has an embedded processor with a PowerPC core for running switch software and a microcontroller for control purposes. The microcontroller provides RAS functions in conjunction with the management blades and issues power-consumption notifications. It also provides a function for reducing power consumption by turning off the high-speed transceiver circuit and control logic circuit of any port for a server blade that does not have a 10GbE adaptor card connected.

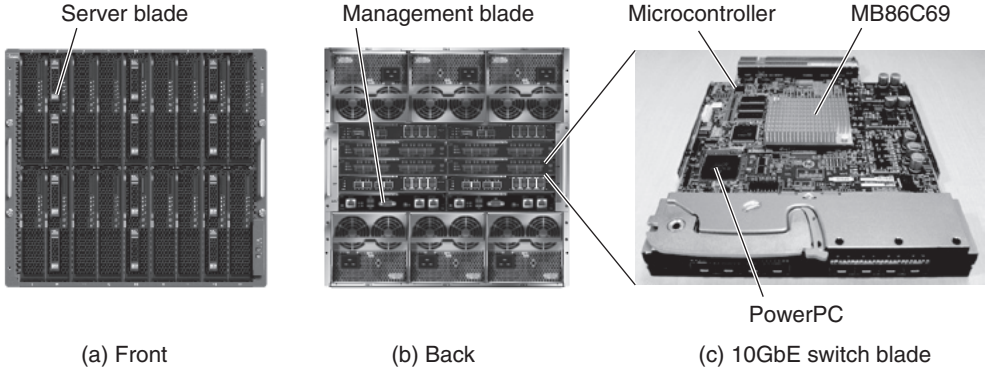
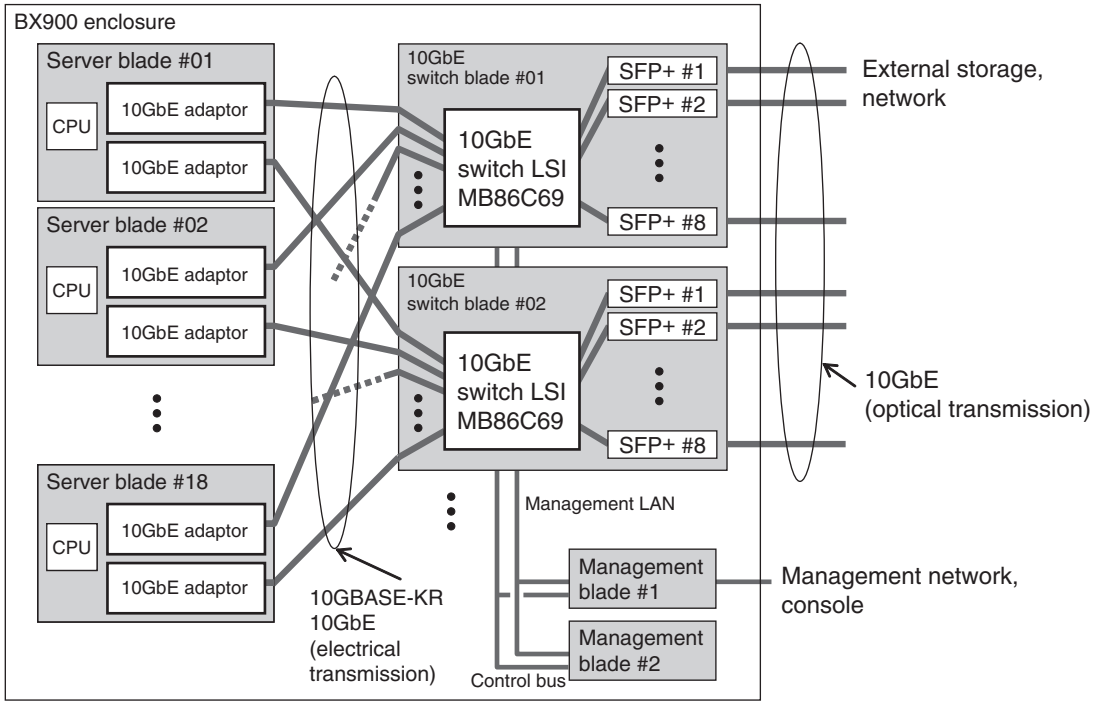


Figure 1
BX900 blade server and 10GbE switch blade.



SFP+: Small Form-Factor Pluggable Plus

Figure 2
Block diagram of BX900 blade server.

The main specifications of the developed switch blade are listed in **Table 1**. In terms of operating modes, this switch blade supports the usual layer 2 (L2) switch mode as well as the intelligent blade panel (IBP) mode that simplifies configuration operations. The IBP mode is described in Section 5.

3. 10GbE switch LSI MB86C69

The MB86C69 chip is an enhanced version of the MB86C68.[1] It has been optimized for blade servers. It is fabricated using 90-nm complementary metal oxide semiconductor (CMOS) technology, and it integrates a logic circuit equivalent to about 22 million gates and a 2.9-MB built-in data buffer. The chip size is 262 mm^2. The LSI is enclosed in a 35 mm × 35 mm flip chip ball grid array (FCBGA) package. The main specifications of the MB86C69 are listed in **Table 2** and the main features are described below.

Table 1
Specifications of 10GbE switch blade.

Downlink	18 ports (10GBASE-KR)
Uplink	8 ports (SFP+)
Switch throughput	520 Gb/s
Switch latency	300 ns
Power consumption	30 W (max. 40 W)
Physical size	Width: 193 mm, depth: 268 mm, height: 28 mm
Modes	Layer 2 and IBP
VLAN functions	Port, tag, protocol
Redundancy functions	STP (STP, MSTP, RSTP), link aggregation, backup port, link-down relay
Quality of service functions	• IEEE 802.1p (COS), TOS (IP Precedence), DSCP • ACL (IPv4, IPv6) • Strict, DDR
Network authentication	MAC, IEEE 802.1X, Web, RADIUS, TACACS+
Access control	Layers 2–4
Multicast	IGMP snooping function
Monitoring	Port mirroring
Network control	CLI, Web UI, SNMP/RMON, LLDP, logging

COS: Class of service
DSCP: DiffServ code point
TOS: Type of service

Table 2
Specifications of MB86C69 LSI.

No. of 10GbE ports	26
Interfaces	XAUI 10GBASE-CX4 10GBASE-KR 1000BASE-KX
Switch throughput	520 Gb/s or greater
Switch latency	300 ns
No. of MAC addresses	16K entries
Built-in data-buffer capacity	2.9 MB
Internal priority	8 levels
Maximum frame size	16 KB
Flow control	IEEE 802.3 pause Priority PAUSE BCN
Management interface	GMII/MII × 2
Security	ACL (L2–L4) DoS attack detection
Package	FCBGA1156 (35 mm × 35 mm)
Power consumption	23.1 W (typical)
Fabrication technology	90-nm CMOS

DoS: Denial of service
MAC: Media access control

1) Number of 10GbE ports

In accordance with switch-blade requirements for the BX900 blade server, a single chip provides 26 ports (uplink: 18, downlink: 8).

2) Interfaces

In addition to XAUI and 10GBASE-CX4, the LSI is capable of BX900 backplane transmission in conformance with 10GBASE-KR. It also allows for direct connection of Small Form-Factor Pluggable Plus (SFP+) modules, which are expected to be widely used in the future; this helps to reduce the cost, latency, and power consumption. The MB86C69 LSI also conforms to 1000BASE-KX so that transmission at both GbE and 10GbE speeds can be achieved at all 26 ports in accordance with customer requests.

3) Performance

The recent trends toward server virtualization and multicore processors mean that switch blades must be able to process input/output (I/O) requests from many user tasks. The switching bandwidth of the MB86C69 LSI is the widest in the industry for a single chip at 520 Gb/s or greater, and its latency between input and output is the shortest in the industry at 300 ns. These performance levels eliminate communications bottlenecks.

4) Ethernet enhancements for data center needs

In terms of functional requirements placed on a switch blade, there are great demands to reduce costs by integrating local and storage area networks (LAN/SAN integration) and for supporting FCoE so that operations can be simplified. To achieve Converged Enhanced Ethernet, an elemental technology essential to FCoE, the MB86C69 LSI supports Backward Congestion Notification (BCN), a message mechanism for passing output-port-side congestion information to the transmission-side terminal, and Priority PAUSE, a flow-control mechanism for each priority level.

5) Security

The LSI has an access control list (ACL) on its input side to achieve the high level of security required of switch blades. The ACL performs filtering to extract standard fields in layers 2–4 and user-defined fields from an input frame. If a match is found for an input frame, the frame is processed using the appropriate action for the type of filter, such as allow, dispose, copy to CPU, or forward to CPU.

4. 10-Gb/s backplane-supporting transceiver

As shown in **Figure 3**, a switch blade is electrically connected to server blades via connectors on a printed wiring board called a backplane (or midplane). In older equipment, signals passed through these lines at 1–5 Gb/s, but as the performance of switch blades increased, it became necessary to increase the number of communication lines on the backplane to handle the increase in communication capacity. However, limitations on the backplane wiring

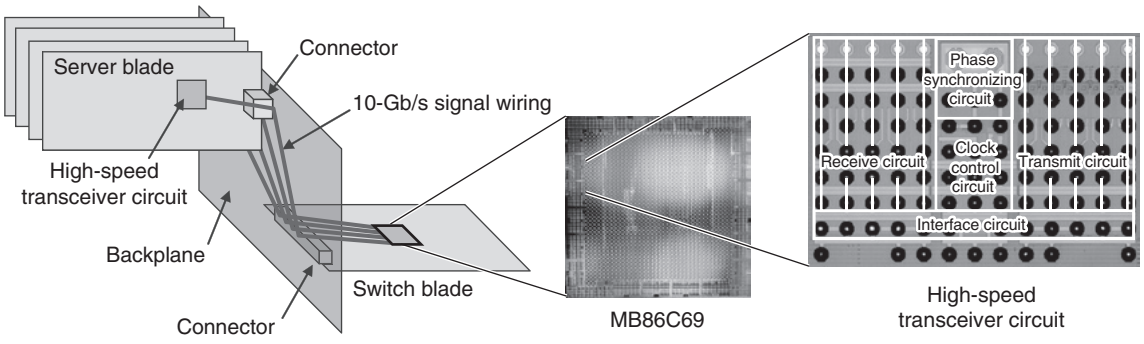


Figure 3
Backplane and high-speed transceiver circuit.

density made it difficult to achieve a high-density, high-performance server overall. To solve this problem, it was essential to find a technology that could support an increase in transmission speed per signal line to 10 Gb/s.

However, transmitting signals at 10 Gb/s on a printed wiring board results in significant waveform distortion called inter-symbol interference caused by signal attenuation. This distortion makes it difficult to transmit data without error. One means of solving this problem is to apply circuit technology called an equalizer that compensates for the signal attenuation. Although a specialized signal transceiver IC having such an equalizer function exists, mounting an equalizer IC on each channel of a switch blade connected to many server blades would be physically difficult. The large amount of power such a configuration would consume would also present a problem. Needless to say, this approach would not be realistic for achieving a high-performance, high-density, low-power switch blade.

In response, we developed a compact, low-power, high-speed transceiver circuit[2)] having an equalizer function for achieving a transmission speed of 10 Gb/s on the blade server backplane, and we succeeded in embedding it in the MB86C69 switch LSI. In this way, we achieved 10-Gb/s Ethernet communications with server blades without having to increase the number of communication lines while achieving a switch blade with one switch LSI without using external equalizer ICs. As a result, we attained a more compact, power-saving configuration. This high-speed transceiver circuit has two main features.

1) Two types of equalizer circuits to correct waveforms that have been distorted by signal attenuation: a linear equalizer and a decision feedback equalizer. This minimizes the power consumed for correcting 10-Gb/s signals in the backplane, which has traditionally been difficult to achieve.
2) A newly developed control system for dynamically controlling the intensity of equalizer compensation (adaptive equalizer control). This system can be achieved by a digital circuit requiring a small amount of computation, which means that it can be implemented in a small area and integrated in this multiport switch LSI.

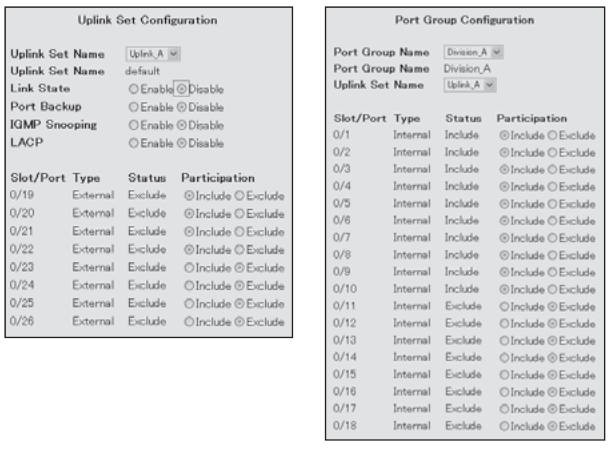
5. Switch blade software

Switch blades in a blade server must suit situations managed by a server administrator rather than ones managed by a network administrator. Thus, one requirement of switch-blade software functions is that they must be easy for server administrators to use.

The main desire of a server administrator is to connect blade servers to the network, whereas performing communications among external

network devices via switches within the blade servers is thought to be incidental. In addition, there are times when a server administrator may not be accustomed to configuring equipment that requires knowledge of network-related matters as in cases involving a virtual LAN (VLAN) or the Spanning Tree Protocol (STP). Such a server administrator might have trouble configuring an L2 switch, whose settings could have a big impact on the network. With this in mind, we considered that another switch blade requirement should be to provide a user interface for making patch panel-like settings for server blades that is easier to use than the usual L2-switch user interface. Therefore, we developed IBP software to provide simple settings in addition to software for the usual L2-switch functions. The customer can select either the L2-switch or IBP operating mode as required.

A screenshot of the configuration operation in IBP mode is shown in **Figure 4**. The uplink configuration screen (a) enables the user to assign a name (Uplink Set Name) to external network ports and to specify which ports are to belong to that uplink set using "Include" buttons. Ports with the same uplink name are bundled by link aggregation and treated as a single logical port.



(a) Uplink configuration (b) Port group configuration

Figure 4
Configuration screen shot of IBP mode.

Next, the port group configuration screen (b) enables the user to specify which network ports to server blades (downlink) connect to which uplink. In Figure 4 (a), Uplink_A is set as the Uplink Set Name and the four ports 19–22 are set as a single logical port. In Figure 4 (b), port group Division_A is set and downlink ports 1–10 are set to connect to Uplink_A. This graphical user interface simplifies connection settings.

The IBP mode has been designed so that multiple logical uplinks cannot be set for a port group and so that communications cannot be performed between two uplink sets. This prevents a loop from being generated when uplinks are being connected to the external network. As a result, the administrator has no need to set a loop-prevention function as in STP, so the STP function itself is excluded from IBP. Since IBP enables port groups to be created by simple operations, the allocation of server uplinks on a business or department basis can be easily and safely performed without having to make L2-switch-like settings as in the case of a VLAN or STP.

The port group function is achieved using the partitioning function (extended VLAN) of the MB86C69 LSI. Extended VLAN divides a switch into even lower levels than VLAN, which means that VLAN can be used within a port group if necessary.

The IBP mode can also be used along with Fujitsu's ServerView server management system and its Virtual I/O Manager (VIOM) virtualization software. Linking IBP with VIOM simplifies the deployment, operation, and maintenance of blade server systems.

While the use of IBP simplifies configuration operations, as described above, some administrators may wish to use a switch blade as an ordinary L2 switch and L2-switch associated functions including STP. To accommodate this need, the developed switch blade allows the user to switch to L2-switch mode. As shown in Table 1, this switch blade has general L2-switch

functions needed by an administrator familiar with network management.

6. Conclusion

This paper described a 10GbE switch blade featuring high performance, a compact configuration, and low power consumption applicable to a large-scale blade server. Looking forward, we plan to enhance the switch software on the switch blade to support FCoE, which is already implemented as a switch-LSI function, and to continue our R&D efforts toward optimizing the switch blade for Ethernet enhancements for data center needs and integrating switch-blade communications with storage communications.

References

1) T. Shimizu et al.: A 20-port 10 Gigabit Ethernet Switch LSI and Its Application. (in Japanese), *FUJITSU*, Vol. 58, No. 3, pp. 246–250 (2007).
2) Y. Hidaka et al.: A 4-Channel 10.3 Gb/s Backplane Transceiver Macro with 35 dB Equalizer and Sign-Based Zero-Forcing Adaptive Control. ISSCC Dig. Tech. Papers, paper 10.5 (February 2009).

Yoichi Koyanagi
Fujitsu Laboratories Ltd.
Mr. Koyanagi received B.E. and M.E. degrees in Computer Science from Tokyo Institute of Technology, Tokyo, Japan, in 1990 and 1992, respectively. In 1992, he joined Fujitsu Laboratories Ltd., Kawasaki, Japan, where he developed the AP1000+ highly parallel computer. In 1994, he joined Fujitsu Ltd, where he developed AP3000 scalar parallel servers. In 1997, he joined HaL Computer Systems Inc., Campbell, CA, where he developed the Synfinity II high-speed interconnect switch LSI. In 2001, he joined Fujitsu Laboratories of America, Sunnyvale, CA, where he worked on the development of 10GbE switch LSIs. Since 2007, he has been with Fujitsu Laboratories Ltd., Kawasaki, Japan, where he is currently a Senior Researcher in the Server Technologies Laboratory. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.

Tadafusa Niinomi
Fujitsu Ltd.
Mr. Niinomi received B.E. and M.E. degrees in Electronics Engineering from Keio University, Yokohama, Japan in 1988 and 1990, respectively. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1990 and was engaged in research and development of high-speed networks and communication technologies. From August 1997 to August 1998, he was a Visiting Industrial Fellow at the University of California, Berkeley, in the department of Electrical Engineering and Computer Sciences. From 2005 to 2009, he was a Senior Researcher at Fujitsu Laboratories Ltd. In October 2009, he moved to the Field Innovator Development Promotion Office, Fujitsu Ltd. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan and the Association for Computing Machinery (ACM).

Yasushi Umezawa
Fujitsu Laboratories Ltd.
Mr. Umezawa received B.S. and M.S. degrees in Physics from Toho University, Chiba, Japan in 1990 and 1992, respectively. In 1992, he joined Fujitsu Laboratories Ltd., Kawasaki, Japan, where he developed an architecture simulator for SPARC V9. In 1994, he joined HaL Computer Systems Inc., Campbell, CA, where he developed interconnect LSIs for IA servers. In 2001, he joined Fujitsu Laboratories of America, Sunnyvale, CA, where he worked on the development of 10GbE switch LSIs. Since 2007, he has been with Fujitsu Laboratories Ltd., Kawasaki, Japan, where he is currently a Senior Researcher in the Server Technologies Laboratory.