Twenty-port, 10-Gigabit Ethernet Switch LSI and Its Applications

● Takeshi Shimizu ● Yasuo Hidaka ● Katsuhiko Nishikawa

(Manuscript received April 19, 2007)

The next-generation Ethernet technology — 10-Gigabit Ethernet (10 GbE) — has a transmission rate of 10 Gb/s. Given the recent growth of server system performance and larger amounts of data to be processed, 10 GbE has been attracting attention in the IT industry as a high-speed network technology for connecting servers and storage at reasonable cost. To widely apply 10 GbE to IT systems at even lower prices, Fujitsu has been developing a single-chip 10 GbE switch LSI since 2001 and released the world's first 12-port 10 GbE switch LSI in 2003. Fujitsu has now developed the MB8AA3020, which is a new 10 GbE switch LSI intended to further expand the application ranges of 10 GbE switches. The MB8AA3020 supports 20 ports and has built-in 10 Gb/s serial interfaces. This paper describes the features of the MB8AA3020 and the high-speed IO circuit of its built-in interfaces. It also describes the XG2000, which is a small switch box powered by the MB8AA3020.

1. Introduction

In recent years, 10-Gigabit Ethernet (10 GbE) has been attracting attention in the IT industry as a high-speed, large-capacity network between servers and storage. Fujitsu released a 12-port, single-chip 10 GbE switch LSI in 2003 and has been promoting the research and development of 10 GbE solutions in parallel with the development of high-performance LSIs.[1),2)] As part of these activities, Fujitsu has developed the MB8AA3020, which is a next-generation 10 GbE switch LSI with significantly improved functions and performance.

This paper describes the features and configuration of the MB8AA3020 and its high-speed IO circuit, which can transmit 10 Gb/s serial signals and is a major feature of this LSI. It also describes the XG2000, a compact and sophisticated 10 GbE switch box equipped with an MB8AA3020.

2. Features of MB8AA3020 switch LSI

The MB8AA3020 is a single-chip 10 GbE switch LSI that is the successor to the MB87Q3140.[2)] It offers vastly improved functions and performance over existing switch LSIs, while supporting the basic functions of the layer-2 switch (**Table 1**). The main improvements are described below.

1) Increased number of ports and enhanced performance

The MB8AA3020 is fabricated using a 90 nm CMOS technology to significantly improve the performance of the switch core section so the number of ports can be increased from 12 to 20. To support the 20 ports, the switch has a shared memory with a switching capacity about 1.7 times greater than the 240 Gb/s of existing switches. Moreover, the fall-through latency, which is the time delay between the input and output pins, has been reduced from about 450 to 300 ns or about a third.

2) Built-in 10 Gb/s serial interface

Each 10 GbE interface has a 10 Gb/s serial interface, which facilitates direct connection to compact optical modules such as the XFP (10-gigabit small form factor pluggable), which is expected to become popular. These 10 Gb/s serial interfaces allow users to dramatically reduce cost, delay time, and power consumption. They also conform to the XAUI/CX4 (a 10 GbE electrical interface the same as existing ones) to enable connections to copper cables, backplanes, and optical modules via a standard electrical interface.

3) Integrated advanced priority control mechanisms and congestion management functions

The MB8AA3020 increases the number of internal priority classes from four to eight and has more priority queue control options. It also supports advanced congestion management functions for data centers such as a flow control function for each priority (priority PAUSE) and a message function for sending congestion information about the output ports to the transmitting terminal via backward congestion notification (BCN).

4) Improved management interface

The MB8AA3020's management interface has been changed from the existing general-purpose processor bus to two Gigabit Ethernet interfaces (GMII and MII). In addition, the MB8AA3020 has a built-in proprietary microcontroller that enables real-time on-chip processing and provides a transaction-based application program interface (API) at a high abstraction level for external management processors. These Ethernet interfaces allow users to flexibly design the physical layouts and connections between the management processors and the MB8AA3020.

Table 1
Comparison between MB8AA3020 and MB87Q3140.

Item	MB8AA3020	MB87Q3140
Number of 10 GbE ports	20	12
Interface	XAUI 10 GBASE-CX4 10 Gb/s serial	XAUI 10 GBASE-CX4
Switching capacity	400 Gb/s or broader	240 Gb/s
Fall-through latency	300 ns (for no load, cut through mode)	450 ns (for no load, cut through mode)
Number of MAC addresses	16 K entries	8 K entries
Built-in data buffer	2.9 MB	600 KB
Number of priority levels	8 levels	4 levels
Maximum frame size	16 KB	15 KB
Flow control methods	IEEE 802.3x Priority PAUSE BCN	IEEE 802.3x
Management interface	GMII/MII × 2	MPC860 bus
Package	FCBGA 1156 (35 mm × 35 mm)	FCBGA 728 (35 mm × 35 mm)
Power consumption	18.6 W (typical)	16 W (typical)
Technology	90 nm CMOS	110 nm CMOS
Others	Enhanced priority queue control Enhanced classification	

3. Configuration and performance of MB8AA3020 switch LSI

Figure 1 shows a die photograph of the MB8AA3020 chip. The chip is 262 mm^2 and uses a 90 nm CMOS technology to integrate about 15 million gates and a built-in 2.9 MB data buffer, giving it more gates and a larger buffer than any other 10 GbE single-chip switch in the industry.

The cell switching method, whereby a packet is divided into fixed-length blocks (cells), is generally used for switch design. However, using this method to design a 20-port 10 GbE switch will result in the operating frequency of the core section becoming too high and will impose large constraints on implementation. We therefore developed a shared memory method that employs the multiple memory banks and multi-stage networks used by the existing MB87Q3140 instead of the cell switching method. This proprietary method is called multi-port stream memory. For the MB8AA3020, the method has been further improved to enhance performance while reducing the core operating frequency to 312.5 MHz, which is as fast as that of existing switches. This reduced frequency reduces the latency due to the use of a single clock and reduces the power consumption. Also, just like existing switch LSIs, the die is encapsulated in a flip chip ball grid array (FCBGA) package measuring 35 mm × 35 mm to facilitate built-in applications.

In terms of performance, a full-mesh RFC2889 throughput test using all 20 ports confirmed that no frames were dropped with frame sizes ranging from 64 bytes to 16 kilobytes.

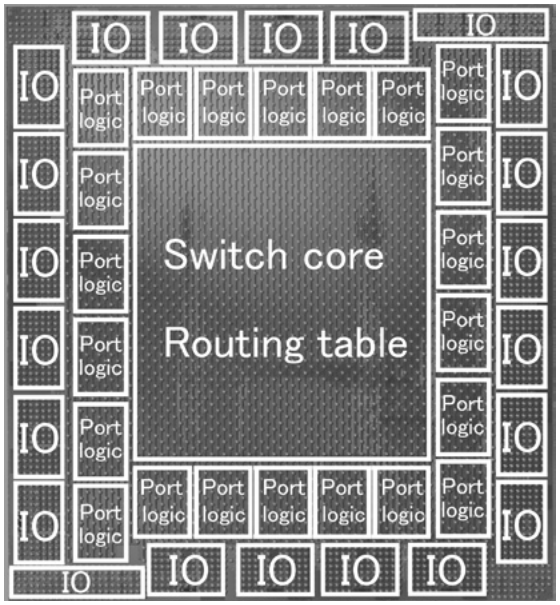


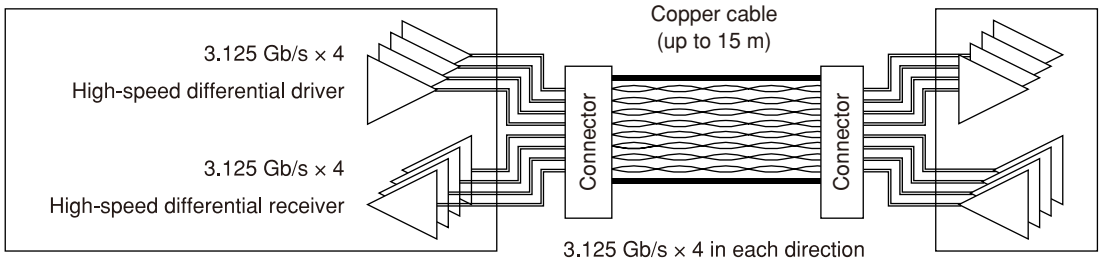
Figure 1
Die photograph of MB8AA3020 LSI.

4. High-speed IO circuit

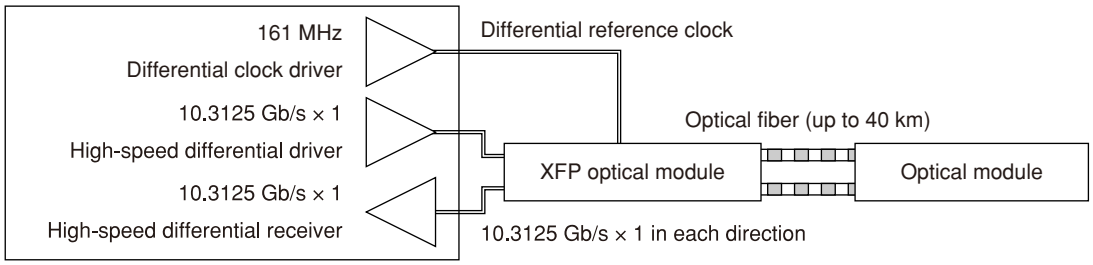
The high-speed IO circuit of the MB8AA3020 operates in the conventional XAUI/CX4 mode or a new 10 Gb/s serial mode. The MB8AA3020 is the world's first 10 GbE switch LSI equipped with a built-in 10 Gb/s serial interface.

Figure 2 shows an overview of signal transmission with the high-speed IO circuit. In the XAUI/CX4 mode shown in Figure 2 (a), four 3.125 Gb/s channels, high-speed differential drivers and receivers are used in parallel to send and receive a 10 GbE signal. This mode features 10 GbE signal transmission over up to 15 m using a 10 GBASE-CX4-compliant copper cable directly connected to the switch chip without any optical modules or other 10 GBASE-CX4 chips. Here, 10 GBASE-CX4 is a PHY layer of 10 GbE using a copper cable. In the XAUI/CX4 mode, as the cable becomes longer, the high-frequency signal component is drastically attenuated due to the skin effect, dielectric loss, and other cable characteristics. Accordingly, the input waveform of the receiver significantly differs from the output waveform of the driver. To solve this problem, we amplify the high-frequency component and attenuate the low-frequency component using a preemphasizer (waveform emphasizer) for the driver and a linear equalizer for the receiver.

In the 10 Gb/s serial mode shown in Figure 2 (b), a 1-channel, 10.3125 Gb/s, high-speed differential driver and receiver are used to send



(a) Transmission in XAUI/CX4 mode



(b) Transmission in 10 Gb/s serial mode

Figure 2
Signal transmission with high-speed IO circuit.

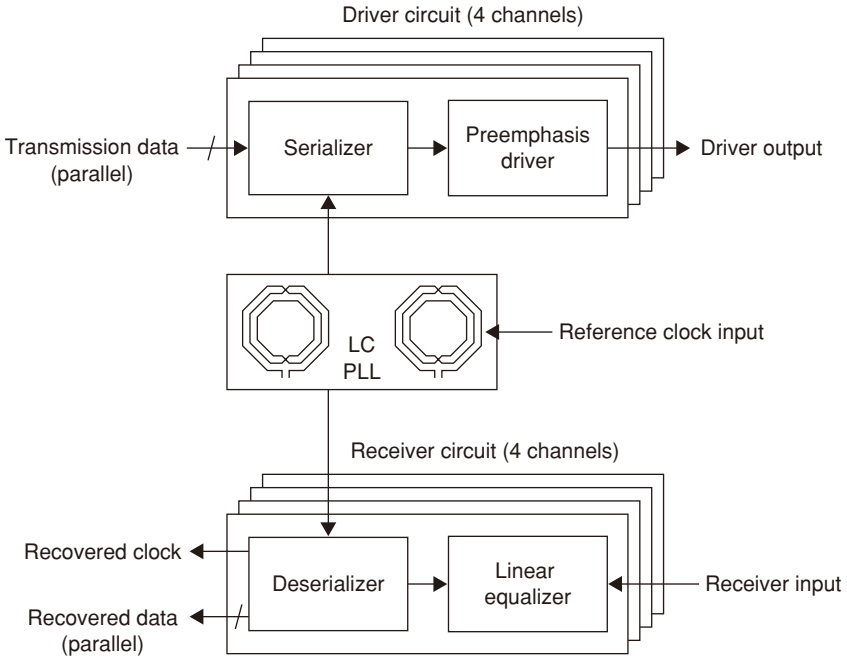


Figure 3
Block diagram of high-speed IO circuit.

and receive a 10 GbE signal. This mode features direct connection with a compact XFP optical module. For other 10 GbE switch chips that support only the XAUI/CX4 mode, an additional serializer/deserializer (SERDES) chip must be used to convert between a 4-channel, 3.125 Gb/s signal and a 1-channel, 10.3125 Gb/s signal. Conversely, the 10 Gb/s serial mode supported by the MB8AA3020 enables direct connection with an optical module without a SERDES chip and therefore can reduce the latency, mounting area, power consumption, and component cost and also improve reliability. Attenuation of the high-frequency component also poses a problem in the 10 Gb/s serial mode. However, this is a less severe problem than in the XAUI/CX4 mode owing to the short transmission distance to the optical module. A more important challenge in the 10 Gb/s serial mode is how to improve the clock quality because the transmission time per bit is as short as about 100 ps (or the time needed for light to travel 3 cm in a vacuum). To achieve the required clock quality, an LC phase locked loop (PLL) that uses low-jitter LC resonance circuit is adopted.

Figure 3 shows the block diagram of the high-speed IO circuit. The driver circuit receives parallel transmission data from the core logic section and converts it to serial data using the serializer. Then, the serialized data is sent to the transmission line using a driver with a preemphasis function. The receiver circuit uses a linear equalizer to restore the input signal to the original waveform and uses a deserializer to recover the parallel data. It then passes the recovered parallel data to the core logic section along with the recovered clock. The parameters for the linear equalizer are automatically adjusted according to the attenuation characteristics of the cable being used based on a newly developed adaptive control scheme.[3] This scheme requires no particular training pattern or encoding scheme. The LC PLL, which uses on-chip helical inductors, provides a low jitter clock to the serializer and deserializer. The high-speed IO circuit features a small size and low power consumption. Each port is 2218.72 µm × 1585.36 µm and has a power consumption of 600 mW in the 10 Gb/s serial mode and 545 mW in the XAUI/CX4 mode.

5. Compact and high-performance 10 GbE switch box: XG2000

The XG2000 is a compact, high-performance 10 GbE switch box based on the MB8AA3020 (**Figure 4**). It is ideally suited for connection between servers or between a server and storage.

The XG2000 provides the following three features to enable full use of the MB8AA3020:

1) High performance with a bandwidth of 400 Gb/s and 300 ns delay
2) 20 XFP ports
3) Fits into a standard 19-inch rack (1.75-inch or 1U height)

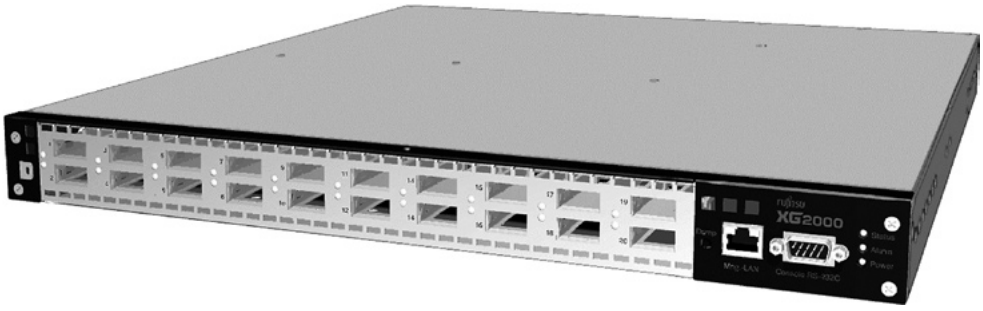


Figure 4
XG2000 switch box.

We expect that the XG2000 will be used in high-speed, large-capacity networks of enterprise IT systems and also in various broadband applications such as real-time transmission of uncompressed high-definition video and applications in video production system networks.

6. Conclusion

This paper described the features of the MB8AA3020, which is a single-chip 10 GbE switch LSI having 20 ports, and the MB8AA3020's high-speed IO circuit, which is one of its major features. It then described the XG2000, which is a compact, high-performance 10 GbE switch box based on the MB8AA3020. Fujitsu will continue developing new functions for 10 GbE switch LSIs as part of its efforts to promote research and development of high-performance IT systems.

The New Energy and Industrial Technology Development Organization (NEDO) funded part of these studies under the theme of "Research and Development of High-Reliability and Low Power Consumption Servers."

References

1) T. Shimizu et al.: A Single Chip Shared Memory Switch with Twelve 10 Gb Ethernet Ports. *Hot Chips 15*, 2003.
2) T. Horie et al.: Single-Chip, 10-Gigabit Ethernet Switch LSI. *FUJITSU Sci. Tech. J*, **42**, 2, p.206-213 (2006).
3) Y. Hidaka et al.: A 4-Channel 3.1/10.3 Gb/s Transceiver Macro with a Pattern-Tolerant Adaptive Equalizer. *ISSCC Dig. Tech. Papers*, p.442-443 (2007).

Takeshi Shimizu, *Fujitsu Laboratories Ltd.*
Dr. 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 moved to 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.

Yasuo Hidaka, *Fujitsu Laboratories of America, Inc.*
Dr. Hidaka received the B.E degree in Precision Machinery and the M.E. and Ph.D. degrees in Information Engineering from the University of Tokyo in 1989, 1991, and 1995, respectively. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1996. He was dispatched to HAL Computer Systems, Campbell, California and Fujitsu Laboratories of America, Inc., Sunnyvale, California in 1996 and 2001, respectively. He has been engaged in research and development of media processors and IA-server chip sets. Also, his current research interest is high-speed I/O circuits for 10 Gb Ethernet. He is a member of IEEE, ACM, IEICE, and IPSJ.

Katsuhiko Nishikawa, *Fujitsu Laboratories Ltd.*
Mr. Nishikawa received the B.E degree in Electronics Engineering from the University of Tokyo in 1982. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1982, where he has been engaged in research and development of information processing, computer graphics accelerators, video servers, and 10 GbE switches. Also, his current research interest is virtualization in IT systems. He is a member of the Institute of Television Engineers (ITE) of Japan.