Single-Chip, 10-Gigabit Ethernet Switch LSI

Takeshi Horie  Takeshi Shimizu  Akira Hattori
(Manuscript received October 1, 2005)

1. Introduction
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.

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 sub-system 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-
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/MLD snooping significantly reduces multicast traffic without generating additional network traffic.

Table 1
Principal specifications.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of ports</td>
<td>12</td>
</tr>
<tr>
<td>Interface</td>
<td>IEEE802.3ae XAUI and IEEE802.3ak CX4</td>
</tr>
<tr>
<td>Switch throughput</td>
<td>240 Gb/s</td>
</tr>
<tr>
<td>Fall-through latency</td>
<td>450 ns (under unloaded condition)</td>
</tr>
<tr>
<td>Switching system</td>
<td>Cut-through and store-and-forward</td>
</tr>
<tr>
<td>MAC address</td>
<td>8 K entry</td>
</tr>
<tr>
<td>VLAN</td>
<td>IEEE802.1Q</td>
</tr>
<tr>
<td>Stack VLAN</td>
<td></td>
</tr>
<tr>
<td>Link aggregation</td>
<td>IEEE802.3ad</td>
</tr>
<tr>
<td>Jumbo frame</td>
<td>Maximum 15 KB</td>
</tr>
<tr>
<td>Spanning tree</td>
<td>IEEE802.1D, IEEE802.1s (MSTP), IEEE802.1w (RSTP)</td>
</tr>
<tr>
<td>Layer-3 snooping</td>
<td>IGMP and MLD snooping</td>
</tr>
<tr>
<td>Quality of service</td>
<td>IEEE802.1Q/p 4 priority</td>
</tr>
<tr>
<td></td>
<td>DiffServ for IPv4 and IPv6</td>
</tr>
<tr>
<td></td>
<td>Shaper (CIR: committed information rate)</td>
</tr>
<tr>
<td></td>
<td>Meter (PIR: peak information rate)</td>
</tr>
<tr>
<td>Port security and control</td>
<td>Source MAC address based filtering</td>
</tr>
<tr>
<td></td>
<td>Broadcast storm control</td>
</tr>
<tr>
<td>Flow control</td>
<td>IEEE802.3ae full duplex</td>
</tr>
<tr>
<td>Management</td>
<td>RMON, SMON statistical information</td>
</tr>
<tr>
<td></td>
<td>sFlow (RFC1276, wire-speed traffic monitoring)</td>
</tr>
<tr>
<td></td>
<td>Port mirroring</td>
</tr>
<tr>
<td>CPU interface</td>
<td>MPC860 and MPC8260</td>
</tr>
</tbody>
</table>
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...
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-Gigabit Ethernet IEEE 802.3ae standard and the 10GBASE-CX4 IEEE 802.3ak 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.

![Diagram of eXAUI](image-url)

Figure 3

eXAUI configuration.
6. LSI implementation

Figure 4 shows a photograph of the MB87Q3140 chip.

The LSI uses 0.11\mu m CMOS technology and operates at 312.5 MHz with a 1.2 V and 2.5 V power supply. The chip is 16 mm \times 16 mm, the package is a 728-pin Flip Chip Ball Grid Array (FCBGA, 35 mm \times 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 lat-

![Figure 4](image-url)
**Figure 4**
MB87Q3140 chip photograph.

![Figure 5](image-url)
**Figure 5**
MB87Q3140 throughput evaluation.

![Figure 6](image-url)
**Figure 6**
MB87Q3140 latency 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 soft-

ware to facilitate the testing of interconnections with equipment that supports the 10-Gigabit Ethernet interface.

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