

Fujitsu's Chipset Development for High-Performance, High-Reliability Mission-Critical IA Servers PRIMEQUEST

● Yasuhide Shibata

(Manuscript received May 20, 2005)

Fujitsu has developed a new mission-critical IA server in close collaboration with Intel Corp. The new server, PRIMEQUEST, uses Intel's latest high-performance, high-reliability Itanium 2 processor. PRIMEQUEST offers linear scalability from a single CPU to 32 CPUs (64 CPUs in the second generation) and represents a highly reliable technology equivalent to that of a mainframe. We have also developed six new chipsets using our cutting-edge CS101 ASIC technology and employed new technologies such as a high-speed interconnection between chipsets, address and system mirroring (including mirroring of ASIC internal blocks), and a new standard high-speed I/O interface called PCI-Express. This paper gives an overview of the new chipsets.

1. Introduction

Fujitsu has developed a new series of mission-critical IA servers in close collaboration with Intel Corp. The new series, called PRIMEQUEST, use Intel's latest 64-bit high-performance, high reliability Itanium 2 processor and incorporate the high-reliability technologies we have acquired from our long experience in developing mainframes. PRIMEQUEST offers large-scale scalability, from a single CPU to 32 CPUs (64 CPUs in the second generation). In order to provide our customers with the high reliability required for mission-critical tasks, we developed six new ASIC chipsets using our cutting-edge CS101 ASIC technology.

This paper gives an overview of the new chipsets and the high-reliability technologies used in their development.

2. PRIMEQUEST and chipset configurations

2.1 PRIMEQUEST configuration

To provide functions that meet customer demands for their large-scale mission-critical

tasks, PRIMEQUEST uses the latest Itanium 2 processor. Symmetric Multiple Processors (SMPs) are used for the basic architecture to enable high-performance scalability from a small number of CPUs up to the maximum of 32 (64 CPUs in the second generation) and uniform access to system resources from any of these CPUs. The PRIMEQUEST system block diagram is shown in **Figure 1**. Up to eight system boards (SBs) can be installed, and up to four CPUs and 32 Dual Inline Memory Modules (DIMMs) can be installed on each SB. Also, up to eight I/O units (IOUs) can be installed for connecting peripheral devices such as LANs and hard disks. Each SB is connected to an IOU via the crossbars shown in the center of the diagram.

Figure 2 shows photographs of the system and component units. Figure 2 (a) shows a view from the front of the system and the component units. A power supply unit and cooling fan unit are installed in the upper part of the system, and up to eight SBs can be mounted vertically in the center part. Four IOUs are mounted from the front panel and four from the rear panel in the

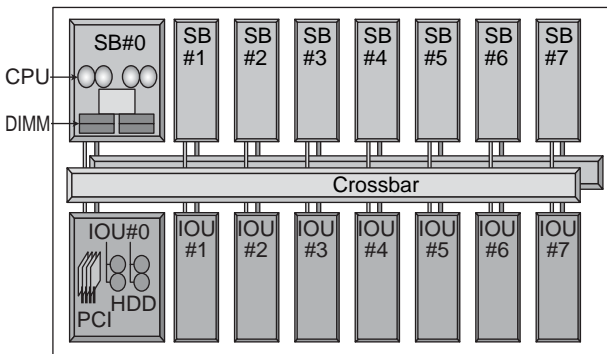


Figure 1
System block diagram.

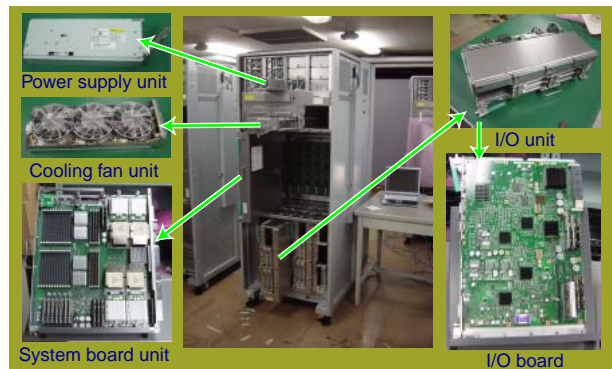
lower part of the cabinet [Figure 2 (b)]. The address crossbars (XAI crossbars) and data crossbars (XDI crossbars) are mounted in the center part of the rear panel.

2.2 Chipset configuration

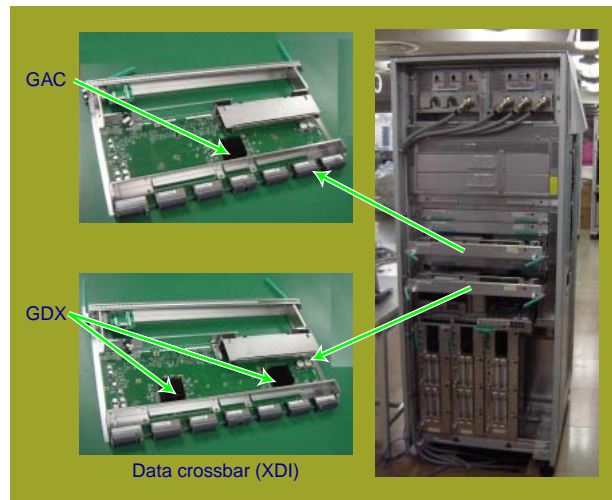
Chipsets are used to link the CPUs, DIMMs, and commercially-available I/O controllers and perform server control tasks. Each SB has two types of ASICs: the FLN ASIC for controlling the CPU and the LDX ASIC for controlling the memory (Figure 2 and **Figure 3**). In addition, two ASICs (FLI and FLP) that control the high-speed PCI-Express I/O control interface are mounted on the I/O board (IOB). ASICs are also mounted on the crossbars interconnecting the SBs and IOUs: one type (GAC) on the XAI crossbars and another type (GDX) on the XDI crossbars. The system uses six types of chipsets in total. (The names of these ASICs are derived from their addresses in the system.)

3. System board (SB) and FLN/LDX ASICs

Figure 3 (a) shows the block diagram of an SB, and Figure 3 (b) shows a photograph of the SB components. Each SB contains CPUs, DIMMs, an FLN ASIC (for interconnection processing between CPUs and other SBs and IOUs) and LDX ASICs (for data linking between the DIMM



(a) Front



(b) Rear

Figure 2
System and components.

controller and XDI crossbars).

Each SB can mount a total of four CPUs: two on each side of the FLN ASIC. The interface used by these CPUs is called the Front Side Bus (FSB). We had close and detailed discussions with Intel Corp. about the protocol and electrical specifications and developed a high-speed transistor macro for the chipset's FSB in an exceedingly short period of time. Eight flash ROMs, of which four are used as a backup in the event of a malfunction, are directly connected to the FLN ASIC. The flash ROMs contain the basic firmware for driving the CPUs.

Read/write requests from a CPU to memory

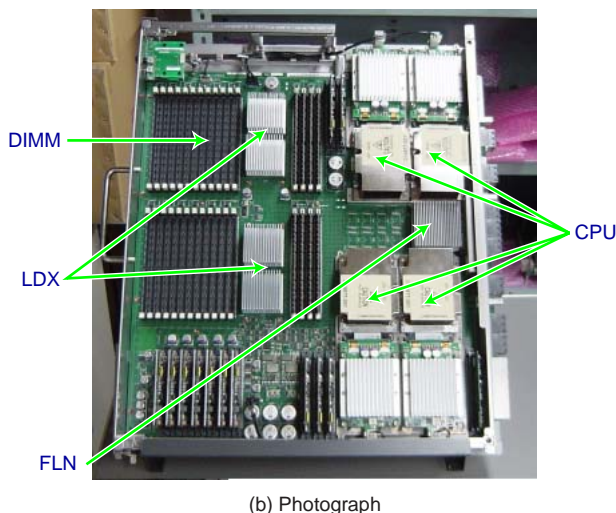
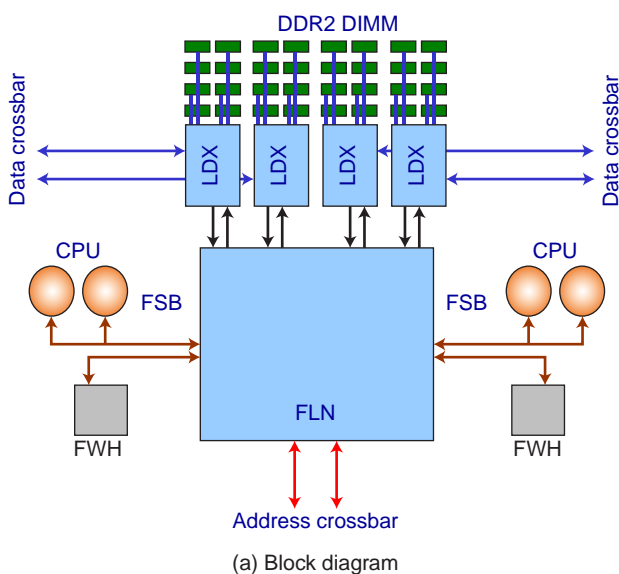


Figure 3 System board.

or an I/O device are temporarily stored in the FLN ASIC and then transmitted to the GAC ASICs on the XAI crossbars in an internally determined order of priority. After processing each request, the XAI crossbars transmit the results to the FLN/FLI ASICs. The FLN/FLI ASIC with the corresponding memory or I/O device handles the requested data.

Each SB has four LDXs to perform data exchange between the FLN ASIC and the memory or XDI crossbars. Data is divided into four

groups of 128 bytes and then processed by the four LDXs.

The DDR2 interface used by the DIMMs is also a high-speed transistor macro specially designed for this application. The four LDXs enable a maximum of 32 DIMMs to be mounted on a single SB. Overall, the system can mount up to 256 (32 × 8) DIMM units. The maximum amount of memory that can be installed is 512 GB when 2 GB DIMMs are mounted and 1 TB when 4 GB DIMMs are mounted.

4. I/O board (IOB) and FLI/FLP ASICs

The IOU comprises an I/O control board (IOB), I/O hard disk slots, and PCI card slots. **Figure 4 (a)** shows the block diagram of the IOB, and **Figure 4 (b)** shows a photograph of the IOB. The FLI ASIC connects the XAI and XDI crossbars and controls access from the CPUs to the I/O devices and data transfer from the I/O devices to memory (DMA). Moreover, using the PCI-Express bus, the FLI ASIC acts as a connection between the Intel ICH6 and PXH chips that are used for bus conversion for the I/O devices. PCI-Express is the latest ultra-high-speed I/O bus; it has a 2.5 GHz transfer capability and was developed to replace the standard PCI bus. The FLI ASIC performs bus conversion from the system bus to the PCI-Express bus, while the FLP ASIC is the electrical interface for the PCI-Express bus. The ICH6 chip is a standard I/O control chip supplied by Intel Corp. and provides functions such as LAN and timer control. In the near future, I/O control chips will be directly connected to the PCI-Express bus. However, because most of the I/O control chips that are available are for the PCI bus and most of the optional cards are PCI cards, we installed a PXH chip for conversion from PCI-Express to PCI. A wide variety of I/O controllers and PCI cards that are available from Fujitsu and Independent Hardware Vendors (IHVs) can therefore be freely incorporated into the system.

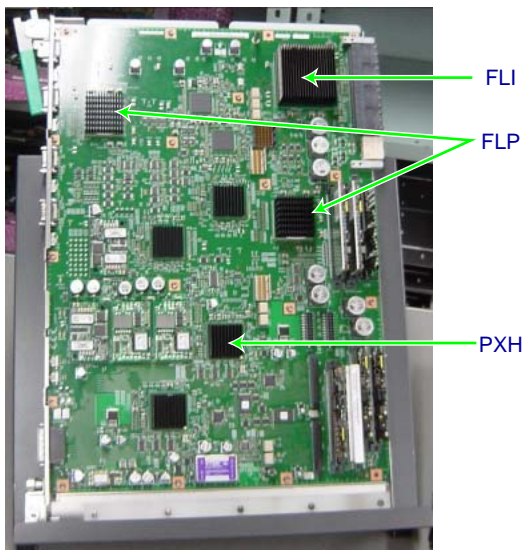
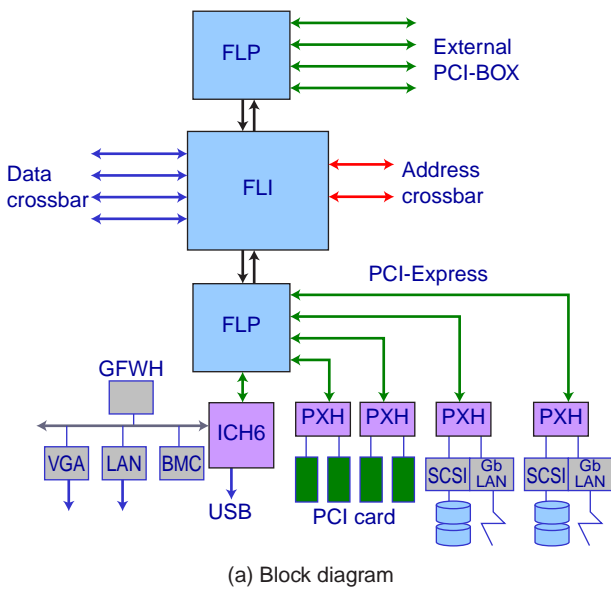


Figure 4 I/O board.

The FLP ASIC is connected to a cable-connected external PCI expansion box via the PCI-Express bus, so it can support many more PCI cards in the box. Furthermore, this solution enables smooth switching to and compatibility with the direct connection of I/O control chips to the PCI-Express bus and the use of PCI-Express cards—functions that will both be available in the near future.

5. XAI/XDI crossbars and GDX ASIC

As shown in Figure 1, the SBs and IOUs are interconnected through crossbars. The XAI and XDI crossbars are configured separately to achieve high performance and high throughput. The XAI crossbar configuration is shown in Figure 5 (a), and the XDI crossbar configuration is shown in Figure 5 (b).

To improve performance, the two XAIs each have a GAC [Figure 2 (b)]. Also, because of the high data traffic on the XDIs, eight XDI crossbar chips (GDXs) are used for the XDI crossbars (Figure 5). The crossbar chips are conceptually identical to the crossbar switchboard used in a telephone exchange. Because the sender and receiver are connected as a pair in this system, simultaneous multiple data transfer operations can be realized.

6. High-reliability technologies

PRIMEQUEST's chipsets were developed using the highly reliable technologies that Fujitsu has acquired in the production of mainframes for mission-critical tasks. This section describes some of these technologies.

6.1 High-speed links between chipsets

A link called the Morimuta-Transceiver-Logic (MTL) was specially developed to link the chipsets that were introduced in earlier sections. The MTL link achieves single-ended (unbalanced) transfer at a designed clock speed of 1.3 GHz (transfer at 1.6 GHz was achieved in a test environment). In the training phase that is conducted for the hardware prior to actual use, adjustments are made automatically for delay errors between signals in the same signal group and for manufacturing variations among the PC boards and ASICs. This eliminates the need to make manual adjustments before shipment and installation and therefore saves time and reduces costs. In addition, this high-speed link uses a reference clock sent from the clock board to each chipset to analyze the dis-

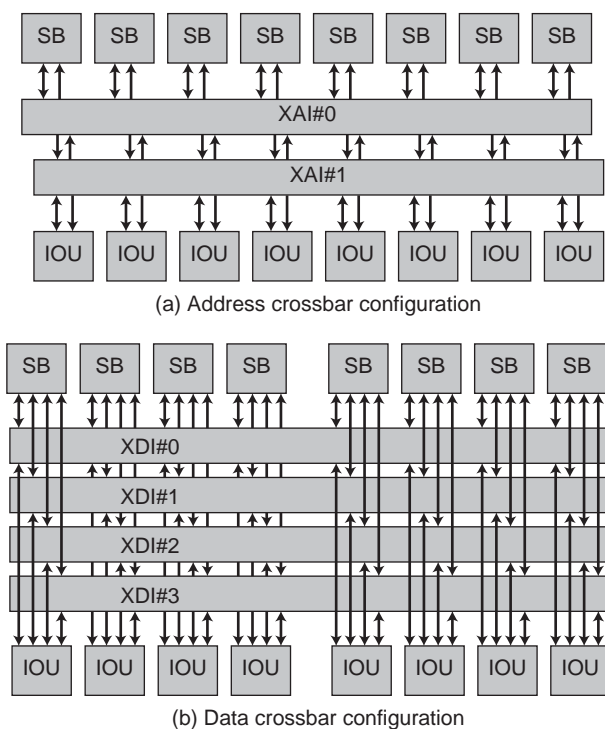


Figure 5 Address crossbar and data crossbar network diagram.

tance between chipsets and provide feedback. As a result, each chipset has the same time characteristics and SMP is achieved, even for the extremely large-scale 32-CPU (first generation) and 64-CPU (second generation) configurations.

6.2 High reliability within each chipset

PRIMEQUEST is designed for mission-critical tasks, and mainframe technology has also been incorporated into the chipsets. An Error Correcting Code (ECC), by which 1-bit errors are automatically corrected by the hardware to enable continued operation, has been added to all the memory, buffers, queues, and other storage mechanisms for data and control information inside the chipset. When two-bit errors occur, they are readily detected and the system can be safely halted. Triplex (triple module redundancy), duplex (mirroring), and parity protection functions are also provided to suit the control system's degree of criticality.

When a fault occurs, the built-in error

notification mechanism provides software or firmware processing at three levels to suit the degree of fault. If a more serious problem occurs, a system management system that is completely independent of the operating system and application programs performs error notification via a dedicated interface at one of three levels according to the degree of fault. This error notification makes it possible to locate the malfunction and minimizes system downtime.

6.3 System mirroring

Figure 6 shows the configuration in the high-reliability, duplicated-address mode. In normal use, the two XAI crossbar chipsets (GAC#0 and GAC#1) operate independently, while in the high-reliability mode, the crossbar chipsets perform simultaneous control of the same address. The FLN and FLI ASICs issue the same address request to the GACs simultaneously and then check and compare the address responses sent from the two GACs. If an ECC error larger than two bits occurs in one of the address responses, the response without the error is used and operation is continued.

Figure 7 shows a further development in which not only the addresses but also the data is mirrored. Usually, data is mirrored simply by duplicating the memory module; however, in PRIMEQUEST, both the memory module and XDI crossbars are duplicated. Because the whole system is duplicated, this mechanism is referred to as system mirroring.

When PRIMEQUEST is in the high-reliability mode, the ASICs themselves are configured internally as mirrored systems, and the operations on one side of the ASICs are compared with those on the other side. Furthermore, various kinds of error detection circuits at each critical point ensure error-free transfer of information and also ensure that an error at one location does not reduce the entire system to non-duplicated operation.

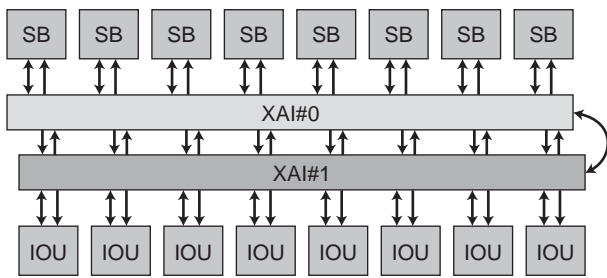
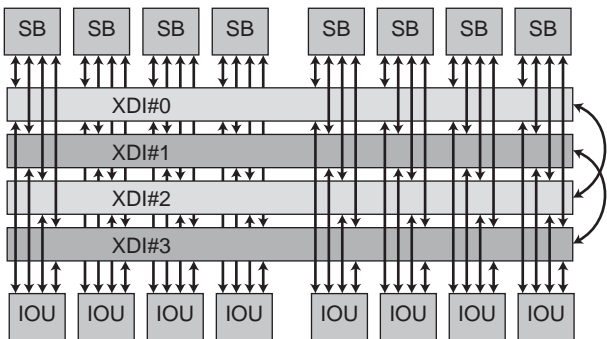
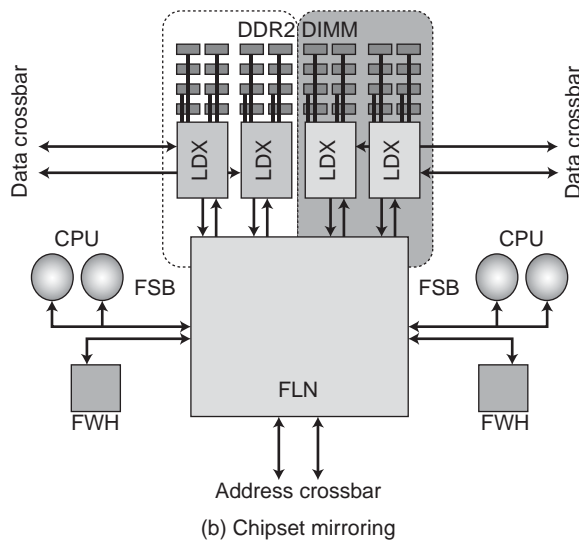


Figure 6 High-reliability, duplicated address mode.



(a) Data crossbar mirroring



(b) Chipset mirroring

Figure 7 System mirroring mode.

7. Transistor technology

New, high-speed interfaces (described in earlier sections) were used in the development of these chipsets. These interfaces enabled us to design high-speed I/O macros at the transistor level, develop test chips, and quickly verify the test chips' operation. We were therefore able to realize a high-reliability design without any problems occurring after the chipsets were mounted.

Moreover, the six chipsets were developed simultaneously. To enable GHz operation of the high-speed I/O section, we used Fujitsu's cutting-edge 90 nm CS101 CMOS transistor, which has nine copper layers and forms the basis of our transistor technology. We developed special design tools for the clock circuit, power supply ground wiring, and other details. We also developed a basic gate called a standard cell for these chipsets.

Figure 8 shows the die of the large-scale FLN ASIC, which is the most complex of the six chipsets. The die is 16.5 × 17.0 mm, and there are 9024 power, ground, and signal output terminals (bumps). The black squares in the upper part of the die are RAM units; the die contains about 200 RAM units of various sizes.

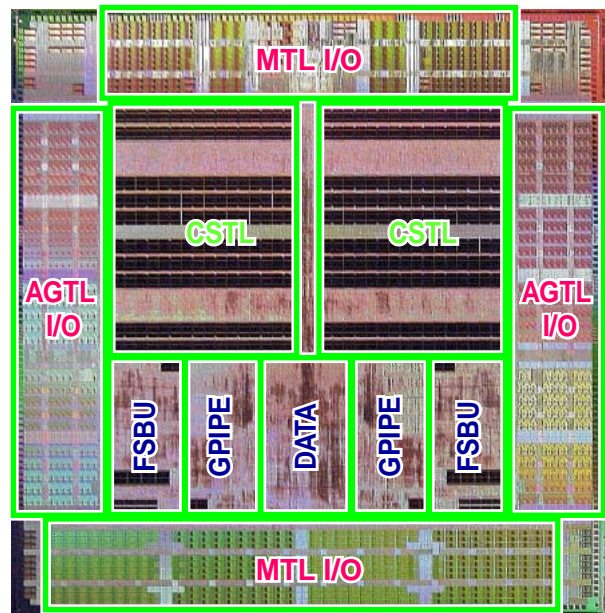


Figure 8 Die of large-scale FLN ASIC.

Excluding the RAM units, this ASIC has about five million gates (2-input NAND conversion). I/O macros are positioned around the edges. The AGTL+ I/O macros, which are the interfaces to the CPUs, are positioned at the left and right edges. The MTL I/O macros, which are the inter-chip interfaces, are positioned at the top and bottom edges (the top ones interface between the LDX ASICs, and the bottom ones interface between the GACs).

8. Conclusion

This paper gave an overview of the six chipsets we developed for use in PRIMEQUEST,

including the new and highly reliable technologies that were implemented in these chipsets.

Our next step is to develop a second-generation device with a maximum of 64 CPUs and to ensure that it supports the next-generation Intel Itanium 2 processors. Our mid-term goals are to meet the challenges of new ideas and technologies and continue developing chipsets so our PRIMEQUEST servers have even better performance and reliability.

This research has been partially funded by the Ministry of Economy, Trade and Industry (METI) and the New Energy and Industrial Technology Development Organization (NEDO).



Yasuhide Shibata received the B.S. degree in Physics from Nagoya University, Nagoya, Japan in 1984. He joined Fujitsu Ltd., Kawasaki, Japan in 1984, where he has been developing SPARC processors and ASICs for servers.

E-mail: yshibata@jp.fujitsu.com