High-Reliability Technology of Mission-Critical IA Server PRIMEQUEST

● Ohsai Hamada

(Manuscript received May 20, 2005)

The mission-critical IA server PRIMEQUEST has various Reliability, Availability, and Serviceability (RAS) functions — from the Application Specific Integrated Circuit (ASIC) and unit levels up to the system level. These RAS functions provide the high reliability and high availability required for mission-critical operations on these servers. This paper gives an overview of PRIMEQUEST. It then discusses the scalability and the speed-enhancement, high-reliability, and high-availability technologies of PRIMEQUEST, focusing on the high-reliability technology implemented in the system hierarchies.

1. Introduction

The new PRIMEQUEST IA servers for mission-critical tasks incorporate application specific integrated circuits (ASICs) developed by Fujitsu and Intel's 64-bit Itanium 2 processors.

This paper gives an overview of PRIMEQUEST's hardware. It then discusses PRIMEQUEST's speed-enhancement, high-reliability, and high-availability technologies and its scalability and high-maintainability design.

This paper also outlines the following.

1) The crossbar system, which interconnects the main units of the system.
2) The partitioning function used to partition the system.
3) The Flexible I/O, which enables flexible combination of the CPU/memory and I/O resources.
4) The System Mirror function, which keeps the system running when hardware failures occur.
5) The cableless design, which improves the reliability and maintainability.
6) The various redundancy and hot-swapping functions that improve system availability.

2. Hardware outline

The PRIMEQUEST models are the PRIMEQUEST 480, which can mount up to 32 CPUs, and the PRIMEQUEST 440, which can mount up to 16 CPUs. The following OSs are supported.

1) Red Hat Enterprise Linux AS (v.4 for Itanium)
2) Novell SUSE LINUX Enterprise Server 9 for Itanium Processor Family
3) Windows Server 2003, Enterprise Edition for Itanium-Based Systems
4) Windows Server 2003, Datacenter Edition for Itanium-Based Systems

Figure 1 shows the physical layout of PRIMEQUEST 440 and 480, and **Figure 2** shows their block diagram.

PRIMEQUEST 440 and 480 are structurally identical. The only difference between these two models is the maximum number of system boards (SBs), which contain the CPUs and memories, and the number of I/O units (IOUs), which contain the hard disk drives (HDDs) and PCI slots. PRIMEQUEST 440 can have up to four SBs and four IOUs, and PRIMEQUEST 480 can have up to eight SBs and eight IOUs.

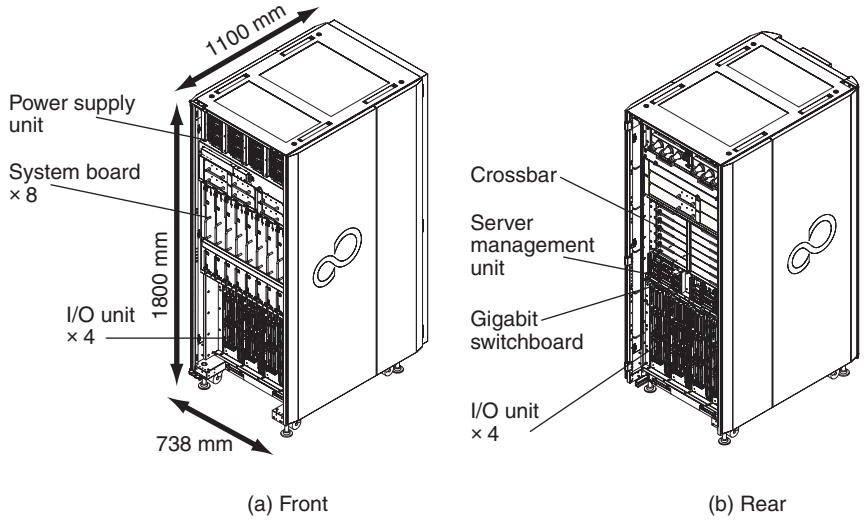

Figure 1
Physical layout of PRIMEQUEST.

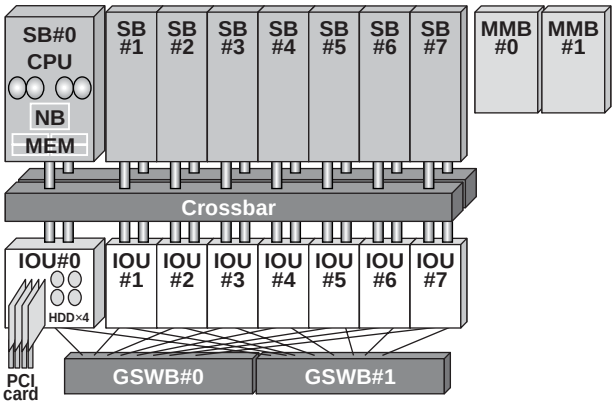

Figure 2
Block diagram of PRIMEQUEST.

As shown in Figure 2, PRIMEQUEST is configured using a crossbar that interconnects the SBs and IOUs. The crossbar is constructed on a midplane (described later) and printed circuit board assemblies containing the ASICs used for path control and transfer of address/data information.

The midplane is mounted in the center of the cabinet and can accommodate units on both sides. The various high-speed interfaces, including the crossbar, are wired on the midplane to achieve a cableless design.

In addition to the SBs and IOUs, a server management unit called the Management Board (MMB) that handles system management within the cabinet is mounted on the midplane. Firmware is installed on the MMB for system-management tasks such as controlling the power supply of each unit in the cabinet and monitoring the temperature and voltage within the cabinet and other environmental factors. The system administrator manages the system using a Web server that provides operation views. The Web server is operated using a Web browser running on a general-purpose PC that is connected via a LAN.

For details about system management, see the paper, "Operation Management of Mission-Critical IA Server PRIMEQUEST for TCO Reduction," presented elsewhere in this special edition.

Moreover, an optional Gigabit switchboard (GSWB) can be installed internally. The LAN interface from the Gigabit Ethernet (GbE) ports installed on each IOU is connected to the GSWB via the midplane and output externally from the cabinet.

Table 1 lists PRIMEQUEST's specifications.

3. Dual Synchronous System Architecture

A new architecture called the Dual Synchronous System Architecture has been developed for PRIMEQUEST. "Dual Synchronous System Architecture" is the generic name of a technology that mirrors the system by configuring a high-speed synchronous parallel bus and enables construction of a high-performance, large-scale multiprocessor system. (The System Mirror function duplicates the hardware components of the server and enables synchronous operations.)

The Dual Synchronous System Architecture enables mirror operations to be applied to a PRIMEQUEST-class, large-scale multiprocessor system for the first time in the world.

4. Speed-enhancement technology

As described above, PRIMEQUEST uses a midplane structure in which the units are densely packed on both sides of the midplane and the high-speed synchronous parallel bus that configures the crossbar is wired on the midplane.

4.1 High-speed synchronous parallel bus

PRIMEQUEST's main bus is a specially developed high-speed synchronous parallel bus. This bus operates at a very fast clock speed of 800 MHz or 1.33 GHz, which makes it the world's fastest synchronous parallel bus. In addition, the bus employs a single-ended transmission method. Compared to the differential transmission method, this method requires only half the signaling wires to transfer the same number of data bits. As a result, when printed circuit boards of the same cost are used, a bus having double the

Table 1
PRIMEQUEST 480/440 specifications.

Model	PRIMEQUEST 480	PRIMEQUEST 440
CPU	Itanium 2 processor (1.60 GHz/9 MB L3 cache, 1.50 GHz/4 MB L3 cache)	
CPU count	Up to 32 CPUs (4 CPU/SB × 8)	Up to 16 CPUs (4 CPU/SB × 4)
Memory (*1)	Up to 512 GB (64 GB/SB × 8)	Up to 256 GB (64 GB/SB × 4)
System boards	Up to 8	Up to 4
I/O units	Up to 8	Up to 4
Crossbar	Up to 102.4 GB/s	Up to 51.2 GB/s
Disks	Up to 147 GB × 32 units	Up to 147 GB × 16 units
PCI slots	Up to 128 slots	Up to 64 slots
GbE interface	Up to 32 ports	Up to 16 ports
SCSI interface	Up to 16 ports	Up to 8 ports
Partitions	Up to 8	Up to 4
Redundant configuration	Disks, power supplies, fans, server management units (MMBs), gigabit switchboard (*2), crossbar (*2), and memories (*2)	
Basic cabinet dimensions (mm)	738 W × 1100 D × 1800 H	
Weight (kg)	720	600
OS (support schedule, etc.)	Red Hat Enterprise Linux AS（v. 4 for Itanium） (End of June 2005) Novell SUSE LINUX Enterprise Server 9 for Itanium Processor Family (End of September 2005, for overseas markets) Windows Server 2003, Enterprise Edition for Itanium-Based Systems (End of September 2005) Windows Server 2003, Datacenter Edition for Itanium-Based Systems (End of September 2005)	

*1: Standard. In mirror mode, up to 256 GB × duplication for PRIMEQUEST 480, up to 128 GB × duplication for PRIMEQUEST 440
*2: Option

bit width can be installed. The high transmission speed and wide bit-width of this bus give the PRIMEQUEST models the highest bandwidth in the world. Fujitsu's state-of-the-art CS101 semiconductor technology is used for the ultra high-speed I/O interface circuits that make this bus possible. For details about the ASICs developed for PRIMEQUEST, see the paper, "Fujitsu's Chipset Development for High-Performance, High-Reliability Mission-Critical IA Servers PRIMEQUEST," presented elsewhere in this special edition.

4.2 Crossbar

A point-to-point crossbar is used to interconnect the SBs and IOUs. Dedicated ports for each unit are provided to connect up to eight SBs and eight IOUs. For the SB ports, a bandwidth from 12.8 to 21.3 GB/s (= 16 × 0.8 GB/s to 16 × 1.33 GB/s) is achieved. Because there are eight ports, the total bandwidth is from 102.4 to 170 GB/s (= 8 × 12.8 GB/s to 8 × 21.3 GB/s).

5. Flexible operation and high scalability

To increase the flexibility of system operation, PRIMEQUEST provides a partitioning function and a Flexible I/O. In addition, PRIMEQUEST supports high scalability to handle increases in the workload after the server is installed. This section describes these functions and high scalability.

5.1 Partitioning function

PRIMEQUEST supports a partitioning function that can divide the system installed in one cabinet into multiple independent systems. Each of these independent systems is called a partition. The minimum components of a partition are one SB and one IOU. The PRIMEQUEST 480 model can have up to eight partitions, and the PRIMEQUEST 440 model can have up to four partitions.

The partitioning function enables various types of uses, for example, mixed use of different OSs or mixed use of partitions for executing production tasks and partitions for executing development tasks. As such, the partitioning function enables the construction of flexible systems. In addition, because the partition configuration can be changed on a daily basis, only the minimum number of SBs and IOUs need be prepared. That is, only the required resources are needed. Moreover, by using the Flexible I/O, which is the main feature of PRIMEQUEST, the SB and IOU resources can be assigned to the partitions in any combination.

5.2 Flexible I/O

PRIMEQUEST provides a function called the Flexible I/O that enables partitions to be constructed from arbitrary SBs and IOUs. For details about the Flexible I/O, see the paper, "Flexible I/O Improves Flexibility and Reliability of Mission-Critical IA Server PRIMEQUEST," presented elsewhere in this special edition.

5.3 Scalability

To handle the ever-growing mission-critical operations, up to 16 CPUs and 256 GB of memory can be mounted in the 440 model, and up to 32 CPUs and 512 GB of memory can be mounted in the 480 model. In addition, up to 16 or 32 PCI slots can be mounted in the cabinets of the 440 and 480 models, respectively. When more PCI slots are needed, external PCI boxes can be connected so up to 64 or 128 PCI slots can be added for the 440 and 480 models, respectively. Moreover, four Gigabit Ethernet ports are installed for each IOU, so up to 16 or 32 Gigabit Ethernet ports can be installed for the 440 and 480 models, respectively.

In addition, scenarios of scale-out and scale-up can be realized by using the partitioning function described above.

Figure 3 shows an example of scale-out and scale-up.

When applied to the server of the AP layer (application layer), any increase in the workload

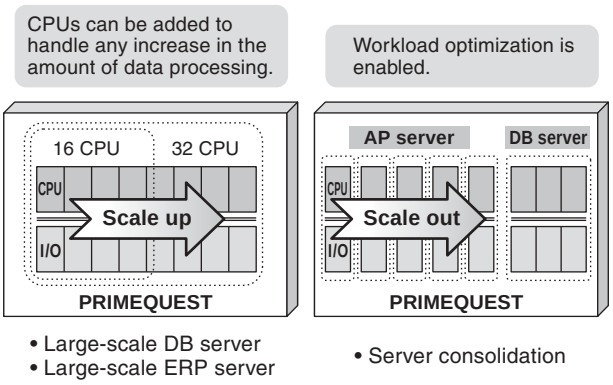

Figure 3
Scale-out and scale-up of PRIMEQUEST.

of the AP layer can be handled by increasing (scaling-out) the number of small partitions allocated to the server of the AP layer. When applied to the server of the DB layer (database layer), increases in workload can be handled by adding SBs and IOUs to increase (scale-up) the partition sizes.

6. High-reliability and high-availability technologies

High-reliability and high-availability are essential for mission-critical systems that must operate 24 hours a day, 365 days a year. For PRIMEQUEST, Fujitsu has used the various technologies it developed for its mainframes and UNIX servers and has developed zero-failure systems that guarantee data integrity.

6.1 Data integrity

To ensure that no data is corrupted during processing, PRIMEQUEST adds an Error Correcting Code (ECC) or parity protection to the RAM data and data bus. Some examples are given below. An ECC or parity is added to the embedded RAM data of each ASIC. When the RAM data is read, the ECC or parity is checked to ensure its integrity.

When transmitting data within an ASIC or between ASICs, ECC or parity is also added to ensure data integrity during transmission. In particular, for data transmission lines that include the crossbar, an ECC or parity check is performed for each stage of the line to ensure data integrity. If a data error occurs, extremely precise error analysis is performed so the location of the error can be identified.

6.2 System Mirror function

PRIMEQUEST offers a System Mirror function as an option. This function duplicates most of the units and implements synchronous operation. If a failure occurs on one side of the duplicated units, the System Mirror function enables the other side to continue operation. For details about the System Mirror function, see the paper, "Highly Reliable System Mirror Function of Mission-Critical IA Server PRIMEQUEST," presented elsewhere in this special edition.

6.3 Redundancy functions

PRIMEQUEST supports redundant configuration of most of its components, including the power supplies, fans, hard disk drives, and MMBs.

1) Power supply system

The standard power-supply configuration is $N + 1$ redundancy. Therefore, if one of the power supply units fails, power is still supplied so operation can continue. In addition, the AC power supply source can be duplicated by installing the optional dual AC power feed function.

2) Cooling system

The cabinet is divided internally into several cooling zones with $N + 1$ redundancy of the cooling fans for each zone. If a cooling fan fails, the remaining cooling fan will continue to operate and the failed fan can be hot swapped without stopping the system.

3) Server management unit

As standard, two MMBs are installed for redundancy. One of the MMBs is active, while the other is on hot standby. The MMBs check each other to ensure they are operating normally. When the hot standby MMB detects an error in the active MMB, it automatically takes over as the active MMB. The failed MMB can then be hot

swapped without stopping execution of the tasks on each partition.

4) Clock circuits

Two system clock oscillators are installed in the cabinet. When the system is powered, initial diagnosis checks that the selected system clock oscillator is operating normally. If an error is detected in this oscillator, the system switches to the other oscillator to start operation.

6.4 Hot-swapping function

PRIMEQUEST supports a hot plug function for the PCI cards, hard disk drives, and cooling fans.

In addition, the SBs, IOUs, and GSWBs can be hot swapped under specific conditions. Therefore, because the main units can be hot swapped, faulty units can be replaced without stopping the whole system.

6.5 Cableless design for high maintainability

A complete plug-in method is employed by which all high-speed interfaces are installed on the midplane in the center of the cabinet and all units are plugged in on both sides of the midplane. The following interfaces are installed on the midplane without the use of cables.

1) Crossbars between the SBs and IOUs
2) A Gigabit Ethernet interface between the IOUs and GSWBs
3) A management LAN (100 megabit Ethernet interface) between the IOUs and MMBs
4) Video, keyboard, and mouse interfaces between the IOUs and KVM (Keyboard, Video, and Mouse) interface units
5) Various management interfaces between the MMBs and units
6) Power lines

By making all wiring cableless, maintenance work such as installing the servers and adding optional units can be done more easily and with higher reliability. Human errors such as incorrect cable connections can also be avoided. In addition, by reducing the cabling costs incurred when adding and changing the cable wiring and achieving more reliable signal connections, the reliability and maintainability are improved.

The following gives an outline of the major interfaces installed on the midplane.

1) Built-in GSWB

Using a GSWB provides several benefits. It significantly reduces the amount of work needed to connect the cables, reduces the operation costs, improves reliability because of the reduced number of cable connections, and improves maintainability and operability.

Changes to the partition configuration within the cabinet, including additions of new partitions, can be done simply by changing the Virtual LAN (VLAN) definitions of the GSWB. There is no need to physically change or add any LAN cable connections, which reduces the operation costs.

The GSWB significantly reduces the number of required external cables, which greatly improves reliability.

2) Management LAN and built-in LAN switch

In addition to the wiring of the gigabit LAN, which is mainly used for tasks, the wiring of the management LAN used for operation management is also arranged on the midplane. Moreover, the LAN switch of the management LAN is built-in. Therefore, the LAN can immediately be used as a management LAN after the server is installed. This not only reduces the number of setup processes at the installation site, but can also reduce the number of connections and settings, which helps to reduce the Total Cost of Ownership (TCO). For example, when a new SB or partition is added, only the IP address of the Ethernet interface within the partition connected to the management LAN need be changed. As a result, the amount of work required when expanding the system can be kept to a minimum.

3) KVM interface unit

The video interface and three USB interfac-

es installed on each IOU are connected to the KVM interface unit via the midplane. MMB firmware controls selection of the video and USB interfaces of one of the IOU's interfaces for connection to the display unit, keyboard, and mouse via the external connector of the KVM interface unit.

One of the USB interfaces is connected to the DVD-ROM drive installed in the cabinet. This interface is also controlled by the MMB firmware. The interface can be switched to a USB interface of one of the IOUs and shared from the individual IOUs.

7. Conclusion

This paper discussed the speed-enhancement technology, scalability, high-reliability technology, and high-availability technology of PRIMEQUEST.

To handle mission-critical operations in the current ubiquitous computing era, in which nearly every business operation involves the use of a computer network, we will continue our efforts to expand the high-reliability technology and speed-enhancement technology of the PRIMEQUEST series to develop and deliver servers that meet our customers' requirements.

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).

Ohsai Hamada received the B.S. and M.S. degrees in Electronics Engineering from Hokkaido University, Sapporo, Japan in 1981 and 1983, respectively. He joined Fujitsu Ltd., Kawasaki, Japan in 1983, where he has been developing server systems, including mainframes and IA servers. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan and the IEEE.