Partitioning Technologies and Dynamic Reconfiguration of Mission-Critical IA Server PRIMEQUEST

● Yasufumi Honda ● Hironori Kobayashi

(Manuscript received June 1, 2007)

The PRIMEQUEST series of mission-critical IA servers provides the high performance, reliability, availability, and flexibility required by mission-critical systems on an open platform. In particular, these servers adopt Fujitsu-specific Dual Synchronous System Architecture (DSSA) of the complete independent recovery type, to afford mainframe-class availability. The series also adopts open-standard operating systems (like Linux and Windows), open-standard processors (such as Intel Itanium 2), and optimizers for data centers (to achieve flexible partitioning, multi-OS scaling out, and scaling up). This paper describes some basic technologies of the PRIMEQUEST series, and then explains Extended Partitioning (XPAR) and Dynamic Reconfiguration (DR). XPAR provides up to 16 partitions operating as independent systems by separately allocating the PRIMEQUEST resources using DR as the mechanism for reconfiguring these partitions without having to reset the system.

1. Introduction

Fujitsu's PRIMEQUEST series[1)-4)] of mission-critical IA servers offers various functions for the high performance, reliability, availability, and flexibility required by mission-critical systems. These servers must specifically provide high availability and flexibility, as well as high performance and reliability to ensure continuous business processing 24 hours a day, 365 days a year and efficient, low-cost system operation. Given the background of improved CPU and IO device performance, customers have gradually demanded more minutely built systems, more effective use of resources, and the reconfiguring of systems without having to reset the systems.

This paper outlines the basic technologies underlying the various functions of the PRIMEQUEST series, and describes the partitioning technology employed for higher system availability and improved flexibility to be enhanced in the future. In particular, this paper describes the higher-level portioning technology known as "Extended Partitioning" (XPAR), which is used to build a system with more minute partitions than made possible by using the existing technology known as "Physical Partitioning" (PPAR). Then this paper describes the Dynamic Reconfiguration (DR) technology used to reconfigure already assigned partitions when necessary without having to reset the system.

2. Basic technologies of PRIMEQUEST

The PRIMEQUEST 580/540/520 models[5)] can contain up to 32, 16, and 8 CPUs, up to 2 TB, 1 TB, and 256 GB of memory, and up to 32, 16, and 8 PCI slots on a chassis, respectively. **Figure 1** shows the system configuration of the PRIMEQUEST 580 model. One system board (SB) can contain up to four CPUs and 32 dual inline memory modules (DIMMs). One I/O unit

(IOU) can contain up to four PCI slots for various peripheral units such as hard disk drives (HDDs), as well as LAN.

Each SB and IOU are mutually connected via a crossbar. The chassis contains six types of Fujitsu's core chip sets (Application Specific Integrated Circuits or ASICs) to control the SB, IOU, and crossbar, respectively.

PRIMEQUEST is supported by technologies employed for high performance, reliability, availability, and flexibility as described below.

2.1 High performance technology

PRIMEQUEST uses a high-speed synchronous parallel bus that operates at 1.3 GHz. Two types of ASICs are mounted on each SB. The first ASIC (called a north bridge) controls the CPU and the second ASIC controls memory. Two types of ASICs are also mounted on the crossbar. The first ASIC controls addresses and the second ASIC controls data. Two types of ASICs are also mounted on each IOU. The first ASIC (called a south bridge) controls IO units and the second ASIC controls the PCI-express bus and interface. Respective ASICs are interconnected via a high-speed synchronous parallel bus to provide a high bandwidth.

2.2 High reliability and availability technologies

PRIMEQUEST adopts Dual Synchronous System Architecture (DSSA) of the complete independent recovery type. DSSA includes a system mirror function that synchronously operates almost all units such as memory chip sets (ASIC) on the high-speed synchronous bus described above in dual (mirror) mode. Should system failure occur on one side of the mirror system, processing would continue normally on the other side to ensure high reliability.

2.3 High flexibility technology

1) Partitioning function

PRIMEQUEST supports a partitioning

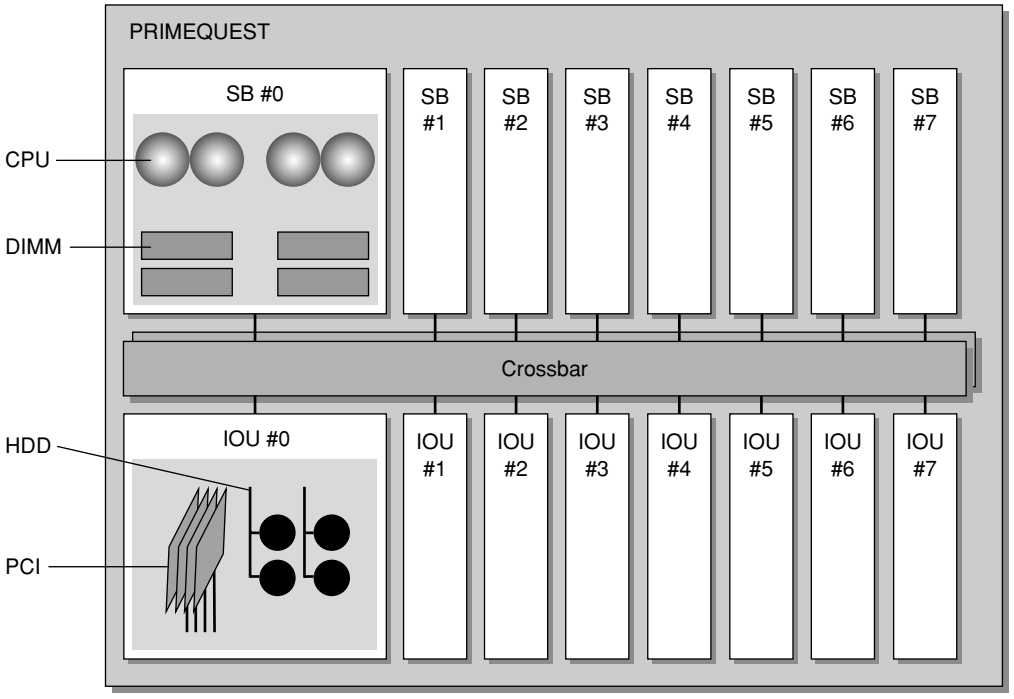


Figure 1
System configuration of PRIMEQUEST 580 model.

function that partitions the system contained on a chassis into multiple independent systems. One of the technologies used to virtualize servers, this partitioning function implemented in hardware supports two modes: PPAR and XPAR. In PPAR mode, a partition is configured in units of SBs (with one SB containing multiple CPUs and DIMMs). The highest-level model 580 can have up to eight partitions. Conversely, in XPAR mode, one SB is divided into two parts, with a partition configured in units of half-SBs. As a result, the 580 model can have up to 16 partitions.

2) Flexible IO function

The flexible IO function is used to flexibly connect the SB to an IOU via the crossbar described above. **Figure 2** shows how partitions can be configured by selecting arbitrary SBs from SB#0 to SB#7 and arbitrary IOUs from IOU#0 to IOU#7. For example, many IOUs can be assigned to partition A which requires many IO resources such as PCI cards, whereas fewer IOUs can be assigned to partition B which does not require many IO resources, but requires many CPUs and memory resources. In this way, hardware resources can be properly distributed to curtail unproductive investment.

3) Dynamic Reconfiguration (DR) function

The DR function is used to dynamically change the partition configuration without having to reset the system. Similar to the flexible IO function described above, the DR function can dynamically add many hardware resources to a partition requiring more hardware resources. Conversely, the DR function can dynamically detach hardware resources from a partition requiring fewer hardware resources. Should an error occur in a hardware resource, the DR function can detach the hardware resource in which the error was detected, and instead assign and start a standby hardware resource. Since resources are dynamically added and released, greater flexibility and higher availability are

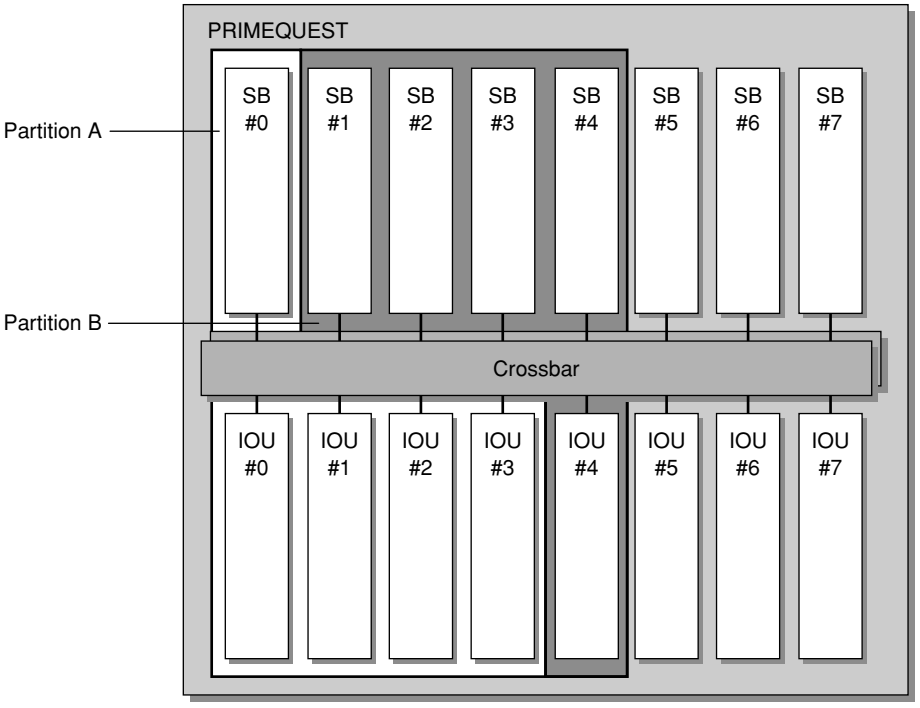


Figure 2
Example of flexible IO and partition configuration.

made possible. The DR function is described later in detail.

3. Hardware technology of XPAR

This section outlines the PPAR and XPAR hardware partitioning technologies, and describes XPAR in detail.

3.1 PPAR and XPAR technologies

The performance levels of single CPU chips, such as multi-core or multi-thread CPUs, have recently been remarkably improved. For example, Dual Core Itanium 2 (Intel's 64-bit CPU) has two logical CPU cores in one socket. The trend toward such CPU integration is expected to continue in the future, with an increasing number of cores and threads according to higher degrees of integration.

In line with the trend described above, when using PPAR technology to partition a system in units of boards as shown in **Figure 3 (a)**, each partition may contain up to four CPUs. In Dual Core Itanium 2, for example, eight logical CPU cores belong to one partition. Although this arrangement improves the performance per partition, it cannot adequately satisfy customer needs for using many small-scale applications efficiently in different partitions. For instance, when eight logical CPU cores are assigned to an application that can sufficiently operate with four CPU cores, four of the eight CPU cores may be unnecessary.

When using XPAR technology as shown in **Figure 3 (b)**, one SB can be split into two partitions, with the number of CPUs per partition reduced to two or less (as opposed to a maximum of four logical CPU cores). Depending on the scale of a given application, such SB splitting facilitates a more effective use of server resources.

3.2 XPAR provided by the core ASIC

XPAR technology is provided by the core ASIC.

1) Splitting SBs

Each SB can have two types of ASICs as explained above. **Figure 4** shows how the CPU and DIMM resources are split into two partitions by these ASICs. The hardware function that splits the internal circuit of the north bridge ASIC into two partitions is used to group and assign two CPUs to each partition. DIMMs are grouped and assigned to each partition by splitting the internal circuit of the memory controller ASIC into two partitions.

The north bridge ASIC provides the XPAR function by assigning a dual circuit for the system mirror function to different partitions. Such an XPAR function can be provided because the north bridge ASIC has a hardware function to completely separate the CPU bus interface and memory controller interface.

The memory controller ASIC has four DIMM interface channels. Assigning two of these four channels to different partitions provides the XPAR function.

2) Splitting IOUs

A south bridge ASIC, legacy IO units (for system operation), HDDs, Gigabit Ethernets (GbEs), and the PCI-express interface are mounted on each IOU. One south bridge has two legacy IO units, two HDDs, and two GbEs. In the XPAR configuration, these legacy IO units, HDDs, and GbEs in the south bridge are fixedly split and assigned to each of the two partitions. Figure 4 also shows that the PCI-express interface can be assigned to an arbitrary partition.

4. Characteristics of XPAR

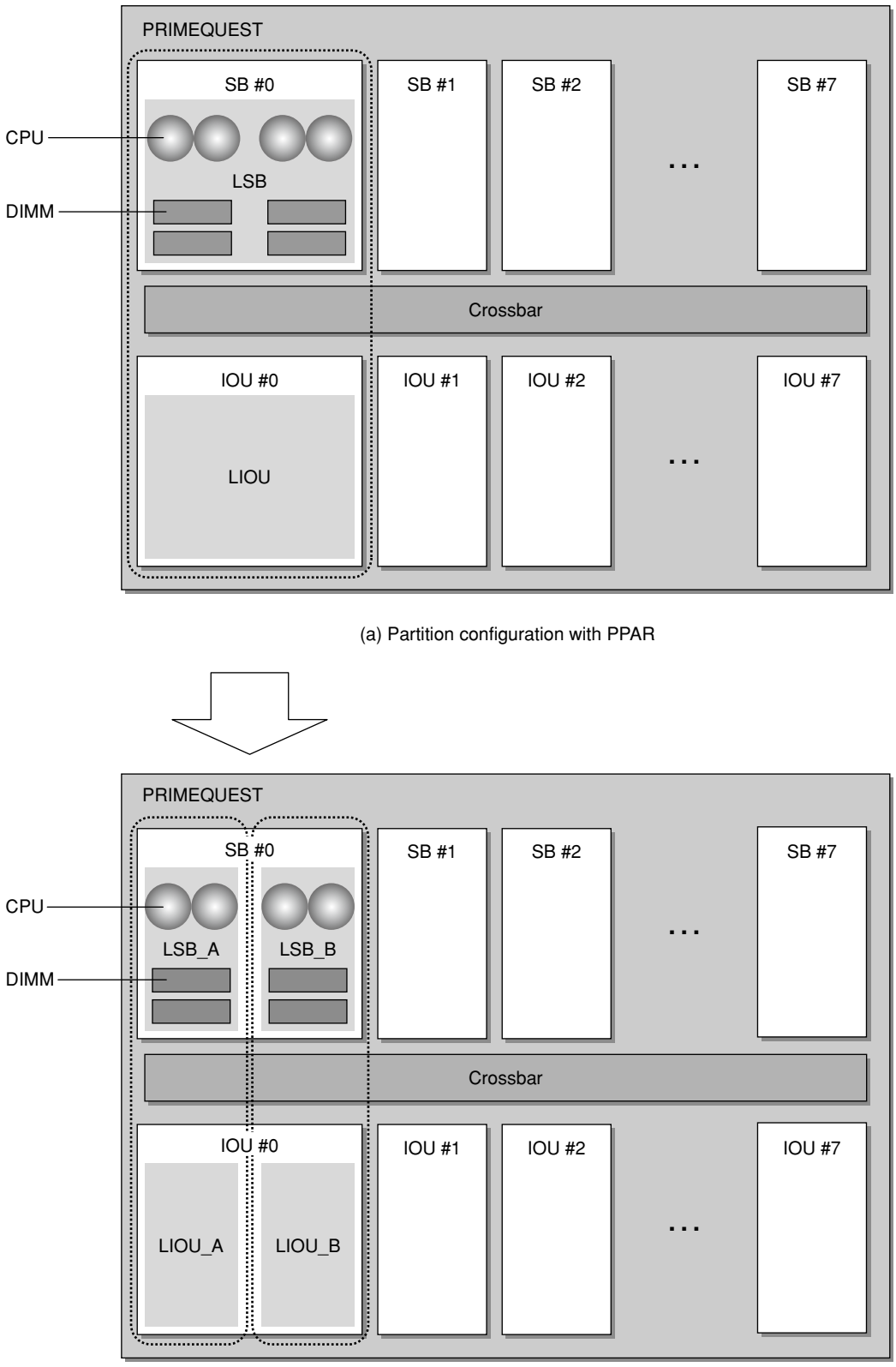
A flexible, strong system can be configured using XPAR technology.

This section describes the following characteristics of PRIMEQUEST XPAR:

- Efficiently distributing server resources
- Maintaining high performance and reliability
- Providing high maintainability

1) Efficiently distributing server resources

Since partitions can be minutely config-



(a) Partition configuration with PPAR

(b) Partition configuration with XPAR

Figure 3
Partition configurations.

ured, many small-scale server functions can be installed in one PRIMEQUEST model. In the PRIMEQUEST 580 model, up to 16 partitions can be configured. As a result, server resources can be efficiently distributed according to the load imposed by customer applications. This enables a significant reduction in operation management costs and the manager's workload necessary to individually manage multiple servers in a conventional system.

2) Maintaining high performance and reliability

XPAR is a hardware function. Therefore, if a hardware or software error occurs in a partition, XPAR can protect the other partitions against the error. The partition where an error is detected stops, but the other partitions can continue normal operation with high performance and reliability, without being affected by the error.

3) Providing high maintainability

Like the PPAR configuration, the XPAR configuration also enables the simultaneous use of different operating system versions on one chassis. In the XPAR configuration, up to 16 OS versions can be used at the same time. This multi-version environment enables a smooth switchover from an older operating system to a newer version, and thus ensures high maintainability. For example, a partition can be used for a production run using an older OS version, while using another partition at the same time for development and evaluation under a new OS version. Should an error occur, these partitions are then isolated and individually protected. Consequently, if the system is reset during software debugging in the partition used for development under the new OS version, the system reset does not affect the production run in the other partition. In this way, the system can be smoothly switched from development to

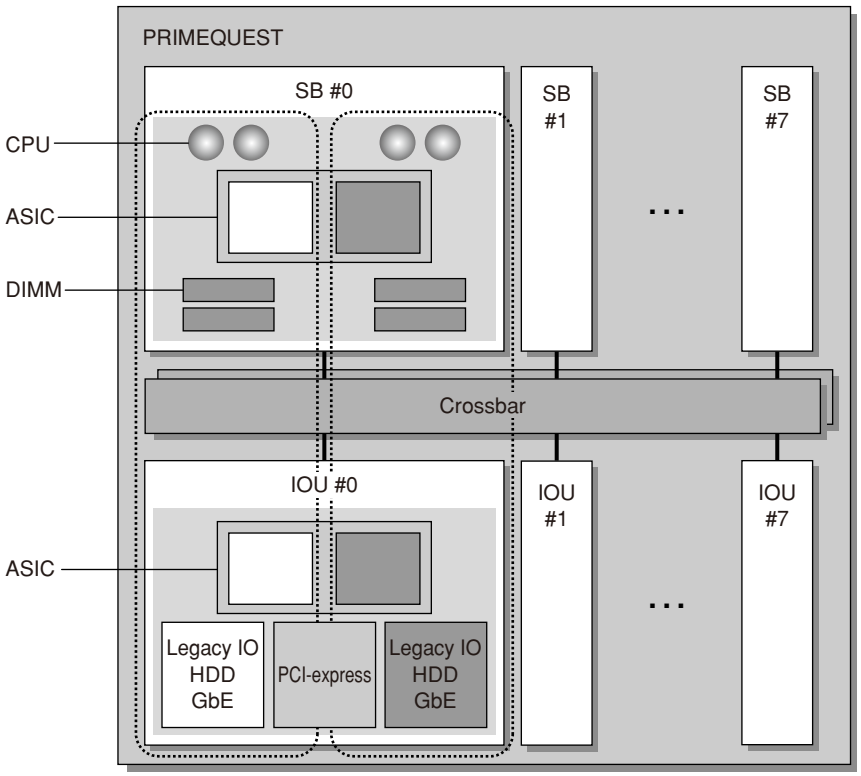


Figure 4
Splitting ASICs in XPAR configuration.

a production run. When the XPAR function is combined with the DR function to move partitions as described in the next section, periodic hardware maintenance is possible without having to stop normal applications.

5. DR function and its necessity

The DR function is used to dynamically add, delete, or replace hardware resources in a partition in which an operating system is running.

PRIMEQUEST provides the DR function for dynamically changing the partition configuration.

The DR function described in this paper refers to a function for changing the configuration of a partition without having to reboot in the partition.

This section describes the necessity for the PRIMEQUEST DR function from the following standpoints:

- Improving availability
- Improving fault tolerance
- Improving maintainability

1) Improving availability

Without the DR function being supported, applications would always have to be stopped on the server when changing the server configuration to separate or build hardware resources according to changes in application load. In other words, the partition power must be turned off in order to change the hardware configuration in the server, and this stops server operation (in what is called a static configuration change).

In a system supporting the DR function, the hardware configuration can be changed without stopping server operation (in what is called a dynamic configuration change). If the running software does not support the DR function, however, the software must be stopped even for a dynamic configuration change.

In order to improve availability, the DR function must therefore be supported for moving, adding, or deleting resources without having to stop the application.

2) Improving fault tolerance

A fault in an operating server degrades system performance. A fault occurring in an important part could adversely affect performance significantly.

To solve this problem, the DR function must be supported for moving, adding, or deleting resources from a fault-detected partition without having to stop the application.

3) Improving maintainability

System recovery from a server failure may entail a long time. For instance, using the cold standby method for failure recovery to set the environment may take much time, and thereby cause the application to be stopped for a long period.

A high-end server having a larger partition configuration may also require a longer reboot time. Therefore, to improve maintainability, the DR function must be supported for moving, adding, or deleting resources without having to stop the application.

6. Characteristics of DR in PRIMEQUEST

This section describes the characteristics of the DR function in PRIMEQUEST from the following standpoints:

- Distributing operation load, improving availability, and reducing cost
- Maintaining high reliability and high fault tolerance
- Ensuring high maintainability

1) Distributing operation load, improving availability, and reducing cost

In PRIMEQUEST, hardware resources can be detached or built according to variations in application load.

For example, when CPU performance in a partition becomes inadequate due to an increasing application load in the partition, the DR function can be used to assign a standby SB to the partition and thereby increase CPU performance up a sufficient level.

Conversely, in case of excessively high CPU

performance in a partition due to a decreasing application load in a partition, the DR function can be used to detach an active SB from the partition and thereby lower CPU performance to an appropriate level.

2) Maintaining high reliability and high fault tolerance

The DR function can be combined with the "Pre-failure detection analysis function" that prevents failure from occurring by monitoring fatal hardware errors, in order to conduct preventive maintenance.

For instance, should memory correctable errors frequently occur in a partition, the SB containing the error-detected memory can be detached from the partition before uncorrectable errors occur. In this way, serious partition failure can be prevented.

3) Ensuring high maintainability

DR is characterized by its function to detach resources from an active operating system or add new resources to the operating system. The resources that can be detached or added by DR are Logical System Boards (LSBs) and Logical IO Units (LIOUs). The partitioning function and DR function ensure high maintainability.

In other words, if a hardware error occurs in a partition, the "Hot-swapping function" can be used to detach the SB containing the faulty hardware resource. Then the Hot-swapping function can be used to assign a new hardware resource to replace the faulty hardware resource in the operating partition.

The "Hot-add function" can be used to assign a newly added hardware resource to the operating partition. Moreover, the "Inter-partition hardware resource movement function" can be used to detach a hardware resource from one partition and assign it to another partition.

7. Implementing the DR function in PRIMEQUEST

This section describes the specific functions under DR implemented in PRIMEQUEST, and gives examples of application.

1) SB Hot-Add

The SB Hot-Add function adds a SB to the partition where an operating system is running without having to reset the system.

This function is used, for example, to add an already prepared, idle SB to a specified partition requiring enhanced resources due to an increasing amount of jobs. **Figure 5** shows an example.

This solves the problem of insufficient CPU/DIMM resources.

2) SB Hot-Replace

The SB Hot-Replace function replaces a busy SB in the partition where an operating system is running without having to reset the system.

This function builds a standby SB into the operating system, copies the CPU/DIMM resources from the busy SB to the standby SB, and then detaches the busy SB from the operating system.

Thus, if a hardware error occurs in a SB, the faulty SB can be replaced with a standby SB without having to reset the system.

3) SB Hot-Remove

The SB Hot-Remove function removes a SB from the partition where an operating system is running without having to reset the system.

Thus, if intermittent errors occur in a hardware resource, the error-detected hardware resource can be detached from the partition without having to reset the system.

4) IOU Hot-Add

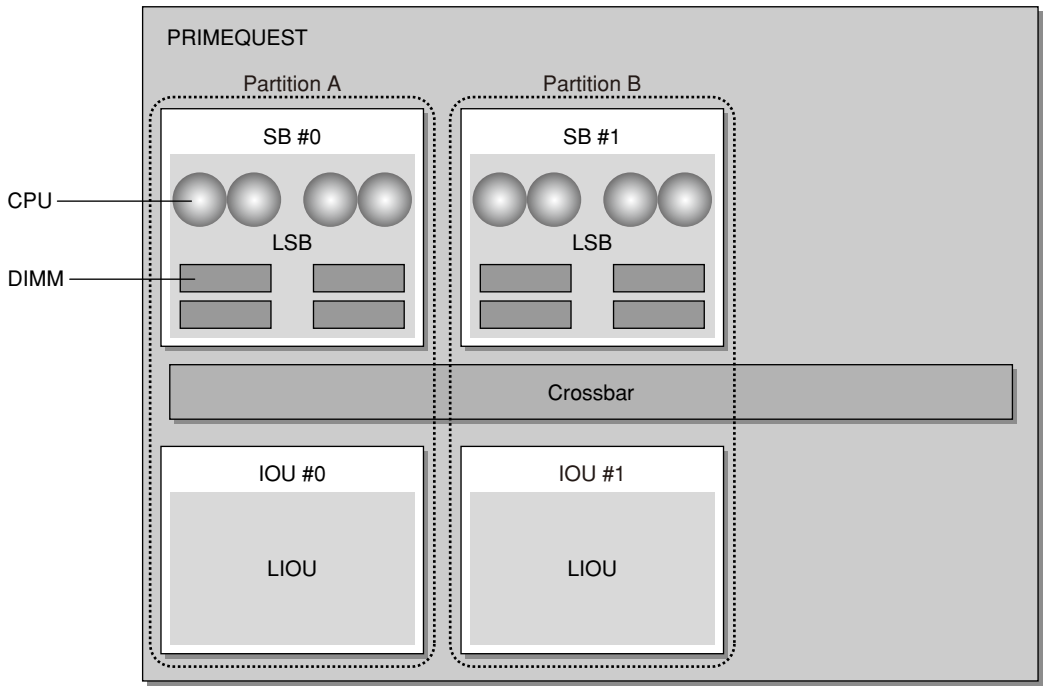
The IOU Hot-Add function adds an IOU to the partition where an operating system is running without having to reset the system.

This solves the problem of IOU shortage.

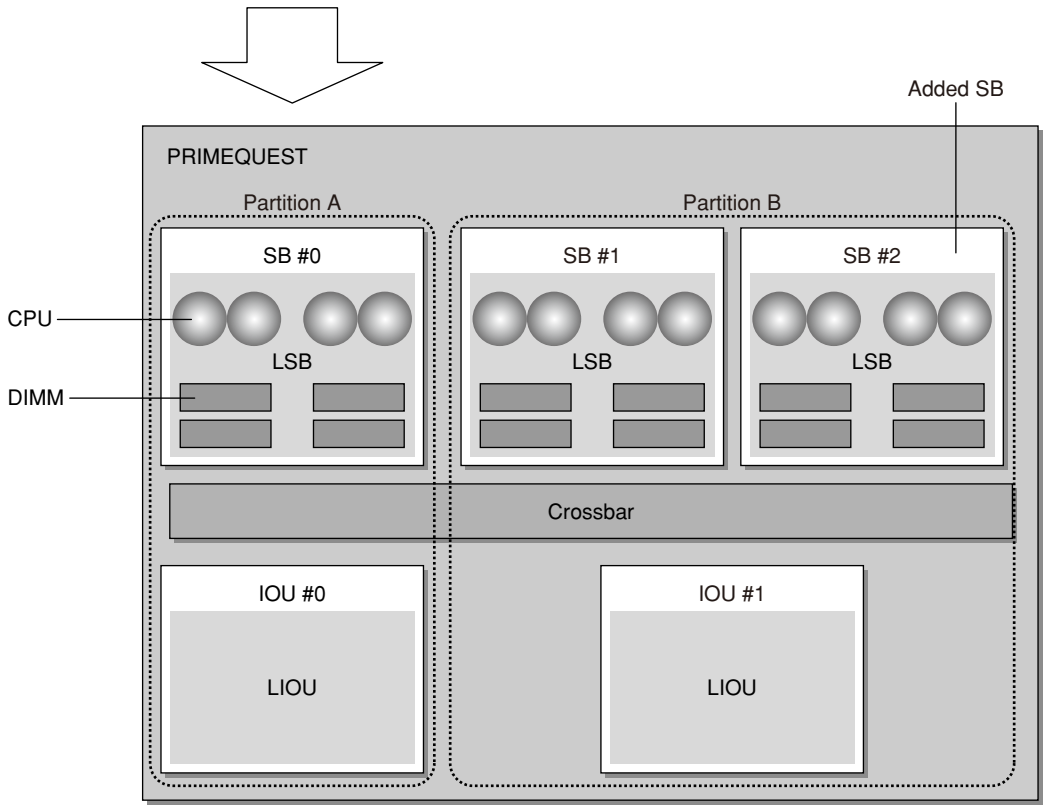
5) IOU Hot-Remove

The IOU Hot-Remove function removes an IOU from the partition where an operating system is running without having to reset the system.

Thus, if intermittent errors occur in a hardware resource, the error-detected hardware resource can be detached from the partition



(a) Before SB addition



(b) After SB addition

Figure 5
Example of SB Hot-Add in DR.

without having to reset the system.

8. Conclusion

This paper described the XPAR and DR technologies that characterize PRIMEQUEST.

The XPAR function provides a flexible partition configuration depending on the user's needs. The DR function provides a flexible allotment of resources depending on variations in load. These functions ensure efficient system operation.

We will continue enhancing the XPAR and DR functions in the future in order to provide customers with systems that are easier to operate.

References

1) O. Hamada: High-Reliability Technology of Mission-Critical IA Server: PRIMEQUEST. (in Japanese), *FUJITSU*, **56**, 3, p.194-200 (2005).
2) Y. Shibata: Fujitsu's Chipset Development for High-Performance and High-Reliability Mission-Critical IA Servers. (in Japanese), *FUJITSU*, **56**, 3, p.201-206 (2005).
3) O. Hamada: Highly Reliable System Mirror Function of Mission-Critical IA Server: PRIMEQUEST. (in Japanese), *FUJITSU*, **56**, 3, p.207-210 (2005).
4) O. Hamada: Flexible IO Improving Flexibility and Reliability of Mission-Critical IA Server "PRIMEQUEST." (in Japanese), *FUJITSU*, **56**, 3, p.211-215 (2005).
5) Fujitsu: PRIMEQUEST Servers. *http://www.computers.us.fujitsu.com/www/products_primequest.shtml?products/servers/primequest/index*

Yasufumi Honda, *Fujitsu Ltd.*
Mr. Honda received the B.S. degree in Electronics Engineering from Chuo University, Tokyo, Japan in 1991. He joined Fujitsu Ltd., Kawasaki, Japan in 1991, where he has been engaged in development of ASICs for high-end server systems.

Hironori Kobayashi, *Fujitsu Ltd.*
Mr. Kobayashi joined Fujitsu Ltd., Kawasaki, Japan in 1989, where he has been engaged in development of ASICs for high-end server systems and the logical verification of high-end servers.

April 24, 2008

Addendum to the article on FSTJ Vol.44, No.1, January 2008

The following is the addendum to the article that describes "Partitioning Technologies and Dynamic Reconfiguration of Mission-Critical IA Server PRIMEQUEST" on FSTJ Vol.44, No.1, January 2008.

Additional contents: The current environment for supported DR function

- Supported functionality depends on OSs.
- Table 1 shows the available combinations of DR functions and OSs.

Table 1 : Dynamic Reconfiguration(DR [*1]) availability table

OS	SB			IOU	
	Hot-Add	Hot-Replace	Hot-Remove	Hot-Add	Hot-Remove
Windows	Yes [*2]	Yes(W2K8) [*2]	n/a	n/a	n/a
RHEL	Yes [*3]	n/a	n/a	n/a	n/a
SLES	n/a	n/a	n/a	n/a	n/a

n/a: not available now

[*1] : We call “DP” (Dynamic Partition) what we support as Dynamic Reconfiguration feature functional in PRIMEQUEST500A series.

[*2] : Microsoft® Windows Server® 2008 for Itanium®-Based Systems or later
Available date is 3Q in 2008.

[*3] : Red Hat® Enterprise Linux® 5.1 (for Intel® Itanium®) or later
Available date is 3Q in 2008.