Highly Reliable System Mirror Function of Mission-Critical IA Server PRIMEQUEST

● Ohsai Hamada

(Manuscript received May 19, 2005)

The System Mirror function is an option for the PRIMEQUEST mission-critical IA servers, which duplicates their hardware and operates it in clock synchronization. When this function is used and a hardware failure occurs on one of the duplicate sides, this function continues processing on the other side and therefore, considerably enhances system availability without the need for software intervention. Because this function is implemented solely through the server's hardware, commercial software can be executed with high reliability without modification. This paper gives an overview of the System Mirror function and discusses the benefits it provides.

1. Introduction

Ultra high reliability and availability are achieved in the PRIMEQUEST mission-critical IA servers mainly through the use of an optional System Mirror function. This function duplicates the hardware of the server and operates the two parts in clock synchronization; its main purpose is to keep the customer's applications running. When a hardware component fails, this function enables operation to be continued using the mirror component without having to stop the application services. This function therefore greatly increases server reliability and availability and can be used as a platform for systems that enterprises need to operate continuously.

Although mirror functions have been implemented on conventional systems, they have mainly been installed as error detection functions. Systems with these functions have two processors; however, these processors have insufficient or no internal error detection capabilities. The outputs of these processors are compared sequentially using an external comparator, and if a mismatch is detected in the outputs, the processor pair and therefore the OS and applications are stopped to prevent data corruption. In more sophisticated systems, to keep applications running, various methods are employed such as installing a third processor and using majority decision processing or providing two sets of processor pairs.

PRIMEQUEST has functions that provide complete error protection for each mirror side of the system. Therefore, when the hardware is duplicated, even if an unrecoverable error (e.g., a multi-bit memory error) occurs on one side, operation can be continued normally using the other side.

This paper describes PRIMEQUEST's System Mirror function.[1)]

2. Outline of System Mirror function

Generally, server systems are configured using a hierarchical structure — from the hardware at the lowest level up to the firmware, OS, middleware, and applications. In cluster systems, the higher levels such as the middleware and OS can also help minimize application stops. However, to achieve this, applications must be modified to operate on a cluster configuration. In contrast,

PRIMEQUEST uses a System Mirror function that duplicates the hardware components (memory, crossbars, etc.) and executes synchronized, mirrored processing on the dual configurations to increase system reliability and availability when hardware failures occur.

Because the entire hardware layer is duplicated, there is no need to modify applications or the OS in any manner. Simply by installing the System Mirror function, customers can obtain a higher level of system reliability and availability with their current software.

In addition, the System Mirror function can be used in combination with cluster systems.

3. System Mirror function design

PRIMEQUEST is configured from system boards (SBs) containing CPUs and memories, I/O units (IOUs) containing hard disk drives and PCI slots, and crossbars that interconnect these units.

In addition to the SBs and IOUs, gigabit switchboards and server management units called management boards (MMBs) are also installed. However, they are not directly related to the System Mirror's function. **Figure 1** shows the System Mirror function of PRIMEQUEST.

When the System Mirror function is applied, the entire system, including the memory section, memory controllers on the SBs, chipsets on the SBs, and crossbars, are completely duplicated. The following describes the dual operation of these components.

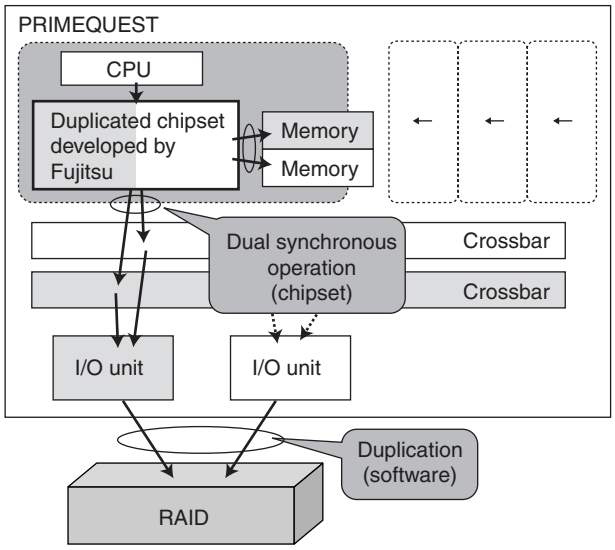

Figure 1
System Mirror function of PRIMEQUEST 400 series.

3.1 Duplicated SB components

Each PRIMEQUEST SB contains CPUs, memory, an ASIC called the North Bridge that is connected to the CPU bus, and memory controller ASICs. When the System Mirror function is applied, the memory controller ASICs, memory, address and data crossbar interfaces, and the internal blocks of the North Bridge ASIC are duplicated.

1) Duplicated memory controller ASICs and memory

Each SB has four memory controller ASICs. These ASICs are divided into two pairs, and these pairs perform the same processing as each other and at the same timing. In addition, the memories connected to each pair are duplicated; that is, the same data is read and written at the same timing. For a memory read, data is read from the same locations on both memories. Because error correcting codes (ECCs) are added to the data of both memories, errors can be independently detected and corrected on each side. As shown in **Table 1**, processing stops only if an uncorrectable error is detected on both sides. In all other cases, system processing can be continued because at least one side can obtain the correct data. As a result, availability is improved significantly.

Generally, in server systems, memory has the second highest failure rate next to hard drives. Accordingly, mirroring the memory improves the availability of the entire system.

2) Duplicated crossbar interface

The SBs are connected to the data crossbars via the memory controller ASICs and to the address crossbars via the North Bridge ASIC.

Because the memory controller ASICs are duplicated as described above, the data crossbars are also duplicated.

As described below, the North Bridge ASIC is duplicated internally, which means that the address crossbars are also duplicated.

3) Duplicated internal blocks of North Bridge ASIC

The CPU bus interface and memory controller interface are duplicated in the North Bridge ASIC, and the two sides of this ASIC execute the same processing in clock synchronization with each other.

This enables detection and recovery of failures that occur in the ASIC's internal control circuits, which is not possible for non-duplicated hardware.

3.2 Duplicated crossbars

The address and data crossbars are both duplicated. The following access operations are performed through these crossbars:

1) Memory read/write accesses initiated by the CPUs
2) Read/write accesses of IOU resources initiated by the CPUs
3) Memory read/write accesses initiated by the IOUs

For read and write operations, address information is sent to the address crossbar followed by the information for the data crossbar. The same address and data information is sent at the same timing to both sides of the duplicated crossbars, and the same processing is performed on both sides. Both sides therefore provide error protection functions such as parity and ECC within their own regions. While information is transferred to and from the crossbar of each system, error checking is performed at each phase of processing. If an uncorrectable error is detected during a transfer, processing is continued using the information of the other side. When an error is detected, the error processing is completely transparent to the software; that is, the software is not aware of it.

In addition to the SBs and crossbars, the IOUs are also key components in PRIMEQUEST. The crossbar interface of the IOUs is duplicated. The PCI buses, which are not duplicated, are connected downstream of the IOUs. Multiple PCI cards for the LANs and fibre channel (FC) are installed, and the software of the middleware layer is used to give the I/O system a redundant configuration.

For details about the ASICs that were 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 issue.

Table 1
Memory mirror read operation.

Side 0 read data	Side 1 read data	Operation
No error	No error	Data of each side is used.
	Correctable error	Data of each side is used.
	Uncorrectable error	Processing is continued with both sides using data of side 0.
Correctable error	No error	Data of each side is used.
	Correctable error	Data of each side is used.
	Uncorrectable error	Processing is continued with both sides using data of side 0.
Uncorrectable error	No error	Processing is continued with both sides using data of side 1.
	Correctable error	Processing is continued with both sides using data of side 1.
	Uncorrectable error	Emergency shutdown and system reboot.

4. Benefits of System Mirror function

The benefits to the customer's applications derived from using the System Mirror function can be summarized as follows.

1) Compared with non-mirroring systems, the error protection level is significantly improved and the availability is also improved. When a hardware error occurs, applications can keep running and the necessary maintenance work can be postponed until a suitable point in the customer's operation schedule. As a result, the customer can avoid losing business opportunities.
2) The System Mirror function is implemented purely in the hardware layer and is completely transparent to the software. No software intervention is required for the function to operate. As a result, the function improves the availability of the middleware and applications required for the customer's tasks.

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.

5. Conclusion

This paper described the System Mirror function of the PRIMEQUEST server.

To provide servers that customers can easily install in their mission-critical information systems, we will further increase the reliability and availability of PRIMEQUEST. For the future, we will continue to pursue higher reliability and availability to provide servers for the core of the IT platforms of the ubiquitous computing society, in which all businesses and individuals are connected through networks.

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).

Reference

1) O. Hamada: High-Reliability Technology of Mission-Critical IA Server PRIMEQUEST. *FUJITSU Sci. Tech. J.*, **41**, 3, p.284-290 (2005).