Express 5800/ft Server – I/A Server Offering Extremely High Availability from a Single Unit

KOSEKI Yasuharu, MATSUSHITA Junichi

Abstract

While increases in the amounts of information and diversification of the modes of information usage have tended to increase the need for more information handling capacity and higher-speed processing, information systems are increasing their social impact and are creating strong pressures to improve availability.

The ft server NEC announced in February 2006 implements the flexibility and availability that can meet the current needs of an industry-standard, open platform. This paper introduces the features and key technologies of the core LSI "GeminiEngine" component of the ft server.

Keywords

fault tolerant, availability, fast resynchronization, ft server, lock-step synchronization

1. Introduction

While recent increases in the amount of information and diversification of the modes of information usage have made more information handling capacity and higher-speed processing necessary, the information systems are increasing in social importance. This trend makes it almost impossible to manage enterprises without the stable operation of information systems. Due to these pressures and in order to meet the high-availability market needs for server products such as infrastructures for information systems, in 2001 NEC commercialized the Express 5800/ft server based on a technology tie-up with Stratus Technologies, USA. This product was acclaimed by a large variety of businesses and total shipments had exceeded 4,000 units by the end of FY2004.

Nevertheless, the dual redundancy engine of the Express 5800/ft server has not been able to trace the rapid advancements of CPU and memory due to the long time taken for its development. In order to deal with this issue, we have newly developed the dual redundancy engine "GeminiEngine" and commercialized a new Express 5800/ft server that can trace technical advancements in a similar manner to other servers and which may be used universally due to its affordable price setting.

2. The New ft Server

As the ft server has been acclaimed more and more in the marketplace, expectations for its use as the platform of the ubiquitous society have increased. In order to accelerate acceptance and by using our proprietary technologies, we began the development of the ft server about three and half years ago. Previously, we had commercialized an ft server based on the technology developed by Stratus Technologies by concluding a collaboration agreement with them. However, in response to a variety of requests from customers, we judged it necessary to develop the new server independently. We thus made use of our own highly-approved hardware development capabilities so as to make the new ft server capable of tracing the latest technological trends and aiming also to reduce its price so that it might become available to many more customers.

As this project also met the aims of the "Semiconductor Application Chip Project" of the New Energy Development Organization (NEDO), we were able to proceed to the development stage under the auspices of this organization. The newly developed ft server was announced on January 23, 2006 and more than 1,500 units have been shipped since then (**Photo**).

3. Lock-Step Synchronization and Determinism

The basic idea of the ft server is quite simple: It must duplicate every item of hardware and guarantee continuity of the overall system by running them in permanent synchronization and continuing the operation of the surviving hardware even if either of the units fails.

Specifically, since a CPU runs in synchronization with a clock signal, applying the same clock signal to two CPUs allows them to always be run in the same way. This status is referred to as lock-step synchronization, and the characteristic

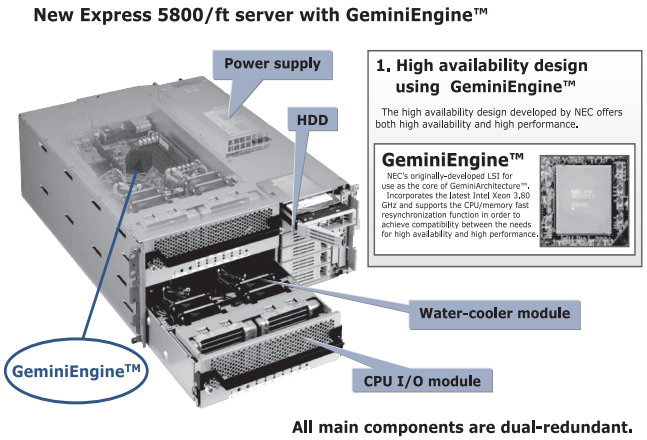

Photo New ft Server Express 5800/320Fz and GeminiEngine.

maintaining this status is referred to as determinism.

The previous ft server was advanced using the principle of determinism. However, as the speeds of all the components and interfaces have recently tended to increase, the environments surrounding the ft server based on determinism have been changing significantly. For example, increases in the CPU operating frequencies and the shifting of inter-chip interface technologies from "parallel bus/low-speed clock synchronization" such as PCI bus to "serial link/high-speed clock asynchronization" such as PCI Express. These trends have led to increases in the factors inhibiting lock-step synchronization, such as asynchronous action elements and clock fluctuations.

4. GeminiEngine

The previous ft server used to identify synchronism between two pieces of hardware by comparing the operations of their PCI buses. However, with the latest servers that use the asynchronous PCI Express interface, it is extremely difficult for the reasons described above to maintain the same lock-step synchronization as before. In addition, certain measures are also essential in order to deal with the clock fluctuations accompanying increases in the CPU operating frequency.

Considering the difficulty of attaining perfect synchronization, we decided to deal with these issues by preparing a mechanism that permits a certain degree of sync deviation but corrects it instantly whenever the error amount becomes large.

The result of all this was the development of the original GeminiEngine LSI. As shown in **Fig. 1**, GeminiEngine contains the north bridge function including memory buses and system buses (FSBs), connects the two duplexed systems through an asynchronous cross-link and has a mechanism that monitors and controls the behaviors of both the system buses (FSBs) of the CPUs and the I/O I/F.

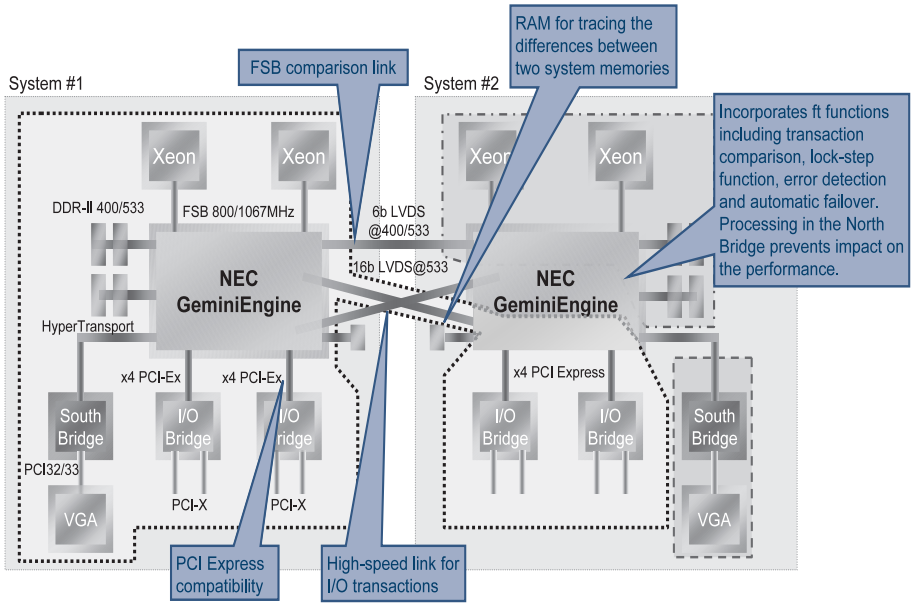

Fig. 1 System configuration using GeminiEngine.

The two systems are compared to both the system buses (FSBs) and the I/O I/F, and the fast resynchronization described later on is made possible by detecting signs of CPU sync deviation on the system buses (FSBs).

The cross-link connecting the two systems is an asynchronous interface, but the problems are avoided by applying synchronization inside the GeminiEngine. The integration of the 2-chip configuration of the previous ft server into the 1-chip configuration of GeminiEngine has contributed to a reduction in device size, reduction of cost and improved reliability.

5. Fast Resynchronization Function

The two systems of GeminiEngine are compared based on both the FSBs and I/O I/F, but it is mostly in the FSB that sync deviations including those due to CPU fault or clock fluctuations are initiated. However, the resynchronization is not started immediately when sync deviation is detected in an FSB.

The possible causes of CPU sync deviation include a fatal failure inside the CPU as well as an operation timing variation within the normal operation range due to variance of the asynchronous circuitry inside the CPUs. As a result, it is extremely difficult to identify the cause and to pinpoint the faulty CPU immediately after detection of a sync deviation in the FSBs.

Therefore, we decide to maintain the duplexed status of the two CPUs in a condition that is "fluctuated" momentarily after detection of an FSB sync deviation. At the same time, we also decided to retain the main memory updating information in the trace memory of the GeminiEngine to enable operation in a condition that is fluctuated to a certain degree. After these processes, the fast resynchronization mechanism activates only after confirming that no abnormality (error) is observed.

The fast resynchronization mechanism copies only the main memory updating area using the trace memory after sync deviation detection. It can thus complete resynchronization in a period of about 200 milliseconds, without either the software or the user being aware of it (**Fig. 2**).

The GeminiEngine maintains the duplexed status of CPUs in the period from the moment of FSB sync error detection to the start of fast resynchronization. This indicates that, even if the real fault causing the sync deviation is identified during this period, operations may be continued simply by isolating the module that is detected to be faulty and degenerating to the normal CPU.

1. Permanent comparison of FSB operations:
Packet exchanges of addresses, commands and timing information.

2. Start of write address recording on detection of sync deviation:
A sync deviation may cause the CPUs to execute different operations that result in the production of a discordance between the contents of the memories of the two systems. The Tracer Memory records the write addresses in which the above occurs.

3. Confirmation of non-erroneous sync deviations and copying of only the differences:
The DMA engines of the two systems read the Tracer Memory and copy the applicable addresses.

4. Sync reset of the CPUs of the two systems after completion of copy:
After resetting, the context immediately before the CPU shutdown is recalled to complete the fast resynchronization process. The series of operations above are performed instantaneously, without stopping the operating services.

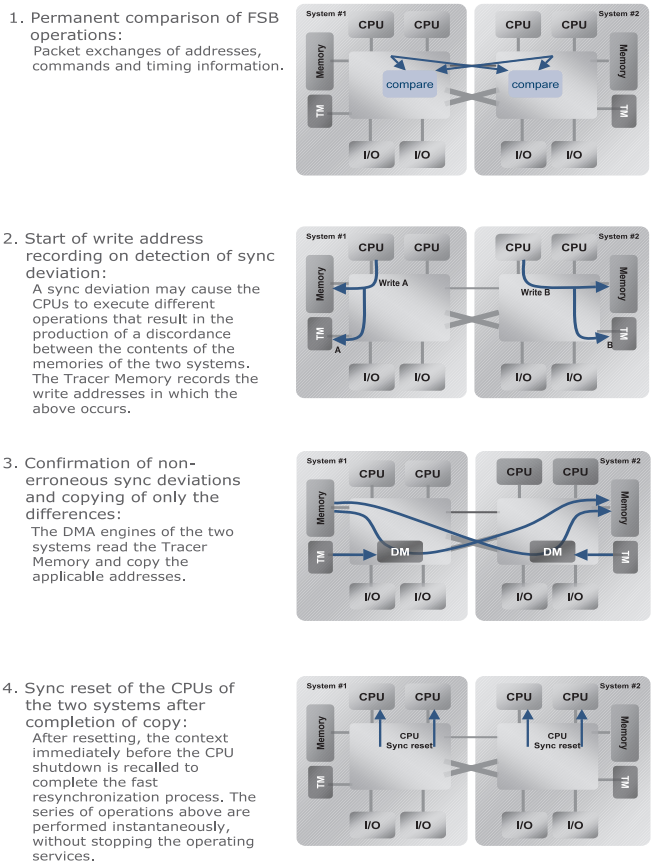

Fig. 2 Flow of the fast resynchronization processing.

6. Virtualization of I/O Devices

With each pair of duplexed I/O devices, only one device is run as the active device and the other functions as the standby device. Therefore, they are switched over only when a fault occurs in the active device. This process is implemented by linking the hardware and the device driver to construct each pair of devices as a virtually single device. With general servers, it often occurs that notification of a fatal fault in the I/O device or the path to I/O including PCI-X and PCI Express to the OS immediately leads to a system down. To prevent this, GeminiEngine hides all of the fatal faults in the I/O system from the OS based on the surprised removal scheme of the hot plug supported by the OS. Namely, even when a fatal fault occurs, the OS is not immediately notified of it but it is disguised as if the I/O device was suddenly disconnected. In this case, when the OS or driver detects the loss of the I/O device, it

performs failover by shifting the control to the standby device.

This series of procedures is executed within the range of the device driver and without the application software being aware of either the fault occurrence or the failover. This makes it possible to use existing applications without any modifications.

7. Conclusion

In the above, we introduced the features of the GeminiEngine LSI, which forms the core of the new ft server that enables the economical construction of a high-availability system aimed at providing "security," which is one of the key concepts of the "REAL IT PLATFORM."

Although we developed GeminiEngine without external assistance, in 2006 we resumed collaboration on the ft server development with Stratus Technologies in order to continue commercialization of Express 5800/ft server products that trace the technological trends and are usable widely within an affordable price range.

NEC is one of the few manufacturers that has access to both of the two high-availability technologies, which are the clustering and ft server technologies and to products that apply these technologies in order to significantly improve the availability of general-purpose servers. We believe that we can respond to the need for a secure and safe society by making full use of the advantages of these technologies and by offering solutions to meet a variety of customer needs.

Authors' Profiles

KOSEKI Yasuharu
Assistant Manager,
1st Engineering Department,
Client and Server Division,
2nd Computers Operations Unit,
NEC Corporation

MATSUSHITA Junichi
Assistant Manager,
1st Engineering Department,
Client and Server Division,
2nd Computers Operations Unit,
NEC Corporation