NX7700i Series Supporting "REAL IT PLATFORM"

YOKOYAMA Jun, SUZUKI Kenichi, SUZUKI Kumiko, KAWAGUCHI Shinichi

Abstract

The integrated enterprise server NX7700i/5080H-64 offers a high-performance, high-reliability and highly flexible environment using the original "A³ chipset" and enables the vision of "REAL IT PLATFORM." This paper introduces the NEC designed "A³ chipset" and the technologies incorporated in a scalable, reliable, and flexible system implemented using Dual-Core Intel Itanium2 processors.

Keywords

integration, server, enterprise, NX7700i

1. Introduction

The integrated enterprise server NX7700i series is the enterprise server that offers the performance, reliability and functionality required for IT infrastructures and large-scale mission-critical systems. The following sections introduce the NX7700i/5080H-64 (**Photo**) the core model of the NX7700i series, which is a hardware platform for enabling the vision of "REAL IT PLATFORM."

2. Hardware Architecture

2.1 Outline of the NX7700i/5080H-64 Architecture

The NX7700i/5080H-64 supports high-scalability, RAS functionality and partition functionality and offers a flexible, dependable and simple REAL IT PLATFORM environment by incorporating NEC's original "A³ chipset" as well as other unique functionalities.

Photo NX7700i/5080H-64.

NX7700i/5080H-64 is composed of up to 8 CPU/Memory modules called cell, maximum 8 I/O modules which have 8 PCI-X and 4 hard disk slots, a crossbar switch connecting the cells and I/O modules, and a service processor that manages the entire server. Each cell can accommodate a maximum of 4 of the Dual-Core Intel Itanium2 processors and a maximum of 128GB of memory (when 4GB DIMM is used) and the total system offers scalability of up to 32 CPUs, up to 1TB memory and up to 64 PCI-X slots.

This server adopts a physical partition functionality that divides itself into a minimum units composed of cells or IO modules and enables them to work as independent servers. The provision of a perfectly partitioned environment at the hardware and firmware levels helps prevent performance shortfalls. In addition, hardware is inaccessible between the physical partitions from the OS so that the security of the operating environment is ensured. A flexible and secure virtualized platform is enabled by the floating IO function, which will be described later, and a system management software for autonomic operating environment called GlobalMaster.

2.2 A³ Chipset

This section discusses the system acceleration functions provided by the "A³ chipset" as the core device to support the implementation of high scalability, reliability and flexibility for the NX7700i/5080H-64.

(1) High-Throughput Interface

The A³ chipset is composed of cell controller that connects the CPU/memory and crossbar switch, the memory controller that controls the memory access, the crossbar controller that interconnects the cells and to and from IO, and the IO controller.

With a bandwidth 25.6GB/sec for the memory interface

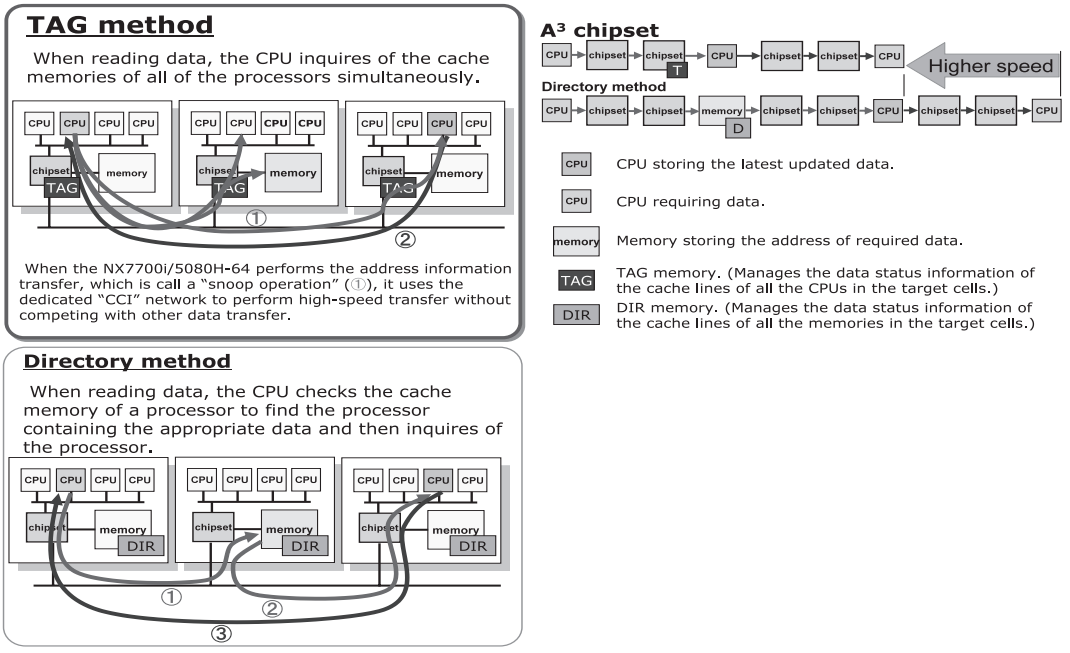

Fig. 1 Comparison of cache situation information retention method.

within a cell and 25.6GB/sec for the interface between a cell and a crossbar switch, an adequate bandwidth is available for the memory access across cells. Rapid handling of a large amount of data via the high-throughput interfaces makes it possible to fully exploit the high performance of the Itanium2 processor.

(2) Dedicated Cache Coherency Interface (CCI)

In addition to the data interfaces, the A³ chipset has an interface dedicated to the cache coherency control across cells. The CCI connects each cell with a 1:1 connection to contribute to increasing the speed of access latency of cached data across cells.

When multiple CPUs run on a single computer and one of the CPUs accesses a cache, it needs the status information that indicates which CPU has the cache storing the latest data. This is based on the cache status information that the CPU accesses the correct cache in order to read the proper data. There are two methods for retaining the cache status information which are the TAG and Directory methods shown in **Fig. 1**. The status information is held in the TAG in the chipset for the TAG method and stored in the external memory (DIR memory) with the Directory method.

With the TAG method, the CPU requiring data can make simultaneous inquiries of all TAGs and accesses the data in the appropriate cache. On the other hand, with the Directory method, the CPU requiring data, first inquires of the external memory and then accesses the appropriate cache. The NX7700i/5080H-64 adopts the TAG method for increasing the speed of cache to cache data access. In addition, its data transfer rate is increased further by "CCI" high-speed connection that connects all of the cells directly without passing through the crossbar.

3. RAS Function

3.1 Design Concept of RAS

Generally, in order to improve both the reliability and availability on an open server system, clustering would be implemented. The NX7700i/5080H-64 has succeeded in improving its reliability and availability even within a single server configuration and also in offering an improved secure environment with its dependable server technology. Continuous operations throughout failures; minimize the spread of failures; and smooth recovery after failures were goals set forth which lead to implementation of technologies such as memory mirroring, increased redundancy of intricate components, and modularization. Through these technologies a mainframe level of continuous operation was achieved.

3.2 Memory Mirroring

The memory-mirroring configuration is supported in order to continue operation even in the case of an uncorrectable fault in the memory system and also to improve data integrity. Memory mirroring consists of always writing data to two memory blocks. Even if either of the memory blocks fails, the other block will maintain data to enable continuity of operations (**Fig. 2**).

3.3 Avoidance of Multi-Partition Downtime Even in the Case of a Chipset Fault

When the system is divided into multiple servers by means of physical partitioning, problems are prevented that may potentially be caused by the shared use of a crossbar switch, such as faults spreading to multiple partitions with a consequent shutdown of multiple tasks. The NX7700i/5080H-64 avoids multi-partition downtime whenever possible by using the partial degeneration function incorporated in the chipset. The crossbar controller is composed of multiple sub-units that can be divided logically when performing partitioning. If an uncorrectable error occurs in a sub-unit, only the partition to which the faulty sub-unit belongs is shut down forcibly in order to limit the fault to within partial sub-units and only those sub-units affected by the fault will then be deteriorated. This arrangement makes it possible to prevent a fault from spreading to other partitions and to avoid a simultaneous stoppage of multiple partitions. The tasks of the shut-down partitions can be resumed by rebooting the sub-blocks in the fault path after degeneration.

3.4 High-Availability Center Plane

The crossbar switch is usually mounted on the center plane, but the NX7700i/5080H-64 mounts this component in an independent module. Other electrically active components on the center plane are also eliminated in order to minimize fault probability. The crossbar switch is not only modularized but is also provided with a dynamic switching capability that is independent of the OS.

In case of a fault, the node related to the fault is rebooted once as described above. However, thanks to the modularized crossbar, maintenance and replacement after a fault does not require jobs to be shut down but the system can be recovered perfectly by simply replacing the applicable crossbar switch module (**Fig. 3**).

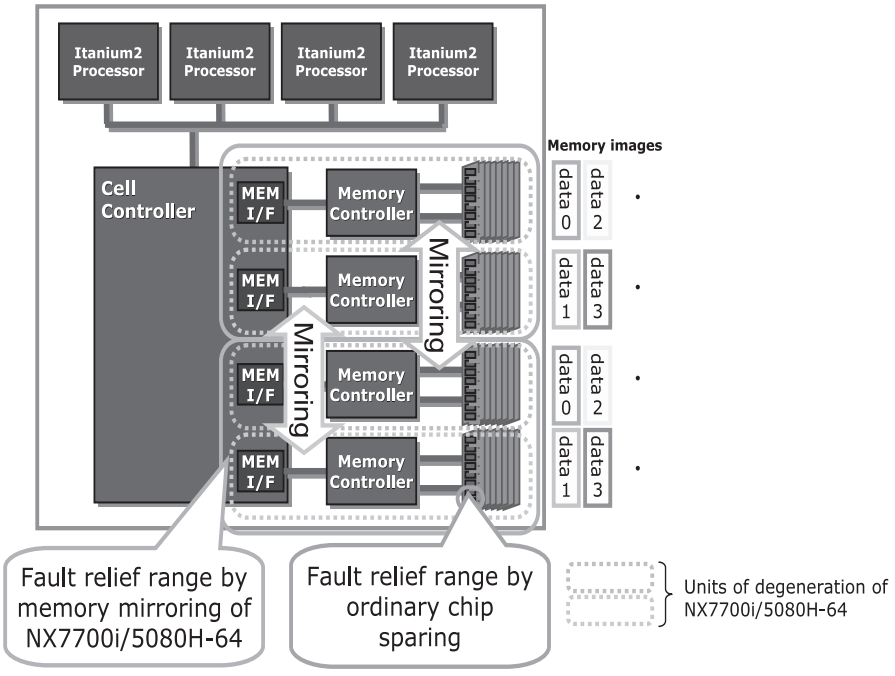

Fig. 2 Positions relieved by memory mirroring.

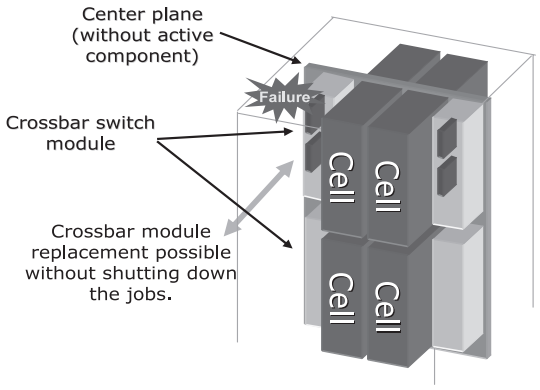

Fig. 3 High-availability center plane.

3.5 Clock Cards

The clock cards are modularized and duplexed in order to minimize the range affected by a clock fault. In most cases the duplexing of clock cards is simply applied to the clock oscillators and the clock distributor containing the clock driver and amplifier are simplexed. But such duplexing cannot deal effectively with a fault in the clock distributor or in the clock distribution path. The NX7700i/5080H-64 employs the clock distributor oscillator in a single module and duplexes the modules so that, when a clock fault occurs in either module, it can be switched to the unaffected one after rebooting. This strategy makes it possible to continue operation without replacing the faulty module and to minimize the downtime. It is also possible to let each of the duplexed clock modules supply clocks to one half of the cabinet. This mode makes it possible to prevent all of the partitions from going down even in the case of a fault with one clock.

3.6 Checking Function of Fault Detector Circuitry

Every main data path of the A³ chipset has an ECC that executes data correction by hardware when a 1-bit error is detected. The interfaces of the chipset additionally support multi-bit error detection and error data retransfer functions. While the data integrity of the NX7700i series is enhanced by the excellent RAS functions of the A³ chipset, the reliability is further improved by a function that checks the fault detector circuitry of the chipset itself. This function checks the entire fault detection circuitry every time after system booting and thereby prevents a situation in which error detection becomes impossible during task execution.

3.7 Service Processors

The service processors of the NX7700i/5080H-64 feature server management and fault processing functions and can be regarded as core RAS components. Each service processor is also capable of built-in diagnosis (BID) based on the error log information collected from the chipset in the case of a fault.

The BID is capable of automatic pinpoint identification of the faulty field replaceable unit (FRU) so that the only the faulty unit can be replaced easily and the recovery time can be reduced. BID also has a function that automatically notifies the hardware detail log in case of a fault. This function helps reduce the maintenance time even further and the connection of logs including the data communication history from the internal path enhances the performance of fault analysis at the software level. In addition, the preventive diagnosis function using the diagnoses manager allows systematic maintenance to be performed before a fatal fault occurs and thereby prevents a sudden system failure.

4. Excellent Flexibility and Operability - Resource Virtualization with Floating I/O -

As mentioned in the description of the architecture, each cell that incorporates a CPU/memory is connected to the I/O module via the crossbar switch. The connections between the cells and I/O modules can be controlled arbitrarily so that the same effect as virtualization of the CPU/memory resources can be provided by combining with the physical partitioning function. For example, even when the CPU/memory resources required for certain processing are temporarily inadequate, it is possible to transfer resources from other tasks (development type, etc.). In addition, even if a fatal fault occurs with a processor memory, a temporary transfer of spare resources or resources engaged in other tasks enables an early resumption of the task. Thus, an environment in which hardware allocation and emergency replacements are executed autonomously is provided based on linkage with GlobalMaster management software.

5. Conclusion

In the above, we introduced the NX7700i/5080H-64, which enables the vision of REAL IT PLATFORM for providing a flexible, dependable and simple environment. In the future, we will continue the provision of attractive products such as those that are introduced in the present report, by advancing high-

speed and high-reliability technologies as well as by further promoting high flexibility.

* Intel and Itanium are trademarks of Intel Corp., USA.

References

1) Takahashi, et al.: "VALUMO Platform-Itanium2 32way server, NX7700/i9510," NEC GIHO, Vol. 56, No. 7, pp.18-21.
2) Senta, et al.: "Itanium2 32way Server System Architecture," NEC GIHO, Vol. 56, No. 1, pp.26-29.

Authors' Profiles

YOKOYAMA Jun
Manager,
Product Engineering Department,
Computers Division,
1st Computers Operations Unit,
NEC Corporation

SUZUKI Kenichi
Assistant Manager,
Product Engineering Department,
Computers Division,
1st Computers Operations Unit,
NEC Corporation

SUZUKI Kumiko
Assistant Manager,
Product Engineering Department,
Computers Division,
1st Computers Operations Unit,
NEC Corporation

KAWAGUCHI Shinichi
Principal Engineer,
1st Computer Engineering Department,
NEC Computertechno, Ltd.

●The details about this paper can be seen at the following.
Related URL:
http://www.sw.nec.co.jp/products/nx7700i