UDC 681.32

Overview of PRIMEPOWER 2000/1000/800 Hardware

●Naoki Izuta ●Toru Watabe ●Toshiyuki Shimizu ●Tetsuhiko Ichihashi

(Manuscript received October 7, 2000)

Servers, especially highly available and scalable servers, are becoming increasingly important to the Internet. Fujitsu previously developed the GP7000F Model 2000, which uses the SPARC architecture and the Solaris operating system. This model provides scalability to 64 CPUs. Now, Fujitsu has developed the PRIMEPOWER 2000, which has extended the Model 2000's scalability to 128 CPUs. The new SPARC64 GP processor's performance has also been improved by incorporating the latest technology. To achieve symmetric multi-processing (SMP) with up to 128 CPUs, the PRIMEPOWER 2000 has an enhanced two-layered crossbar feature that provides a maximum snoop performance of 57.6 gigabytes per second. As well as this extended scalability, the PRIMEPOWER 2000 provides high-availability (HA) features such as redundant components, hot-swap functions for various components, and a partitioning function for computing resources. These features give the PRIMEPOWER 2000/1000/800 the availability demanded by users of servers that must operate 24 hours a day throughout the year. This paper describes the hardware of the PRIMEPOWER 2000/1000/800.

1. Introduction

The PRIMEPOWER 2000, 1000, and 800 greatly extend the range of the PRIMEPOWER family. These PRIMEPOWER models are designed to provide the robust server functions required for today's and tomorrow's Internet generation. PRIMEPOWER is based on a Symmetric Multi-Processing (SMP) architecture that supports up to 128 CPUs and 512 gigabytes of memory, and up to 192 PCI cards. The SMP architecture provides scalable performance without requiring user applications to be modified. To implement this scalability, Fujitsu has developed a large crossbar switch with a bandwidth of up to 57.6 gigabytes/s (operating at 225 MHz). The architecture also takes into account the demands of the next generation of the SPARC64 GP processor.

The partitioning feature of these models enables the processor, memory and PCI cards to be divided into several resource groups, each of which runs its own instance of the operating system and application. Moreover, the PRIMEPOWER 2000, 1000, and 800 support the Dynamic Reconfiguration (DR) feature, which enables the partition configuration to be changed while the system is running.

2. Features of the hardware components

The PRIMEPOWER 2000, 1000, and 800 are designed not only to achieve the scalability described above but also to satisfy the demands made of an enterprise server that is expected to operate 24 hours per day and 365 days per year.

2.1 High scalability

The PRIMEPOWER 2000 connects up to four cabinets (nodes), each with a capacity of up to 32 CPUs. The newly developed high-performance crossbar switch technology employed on the

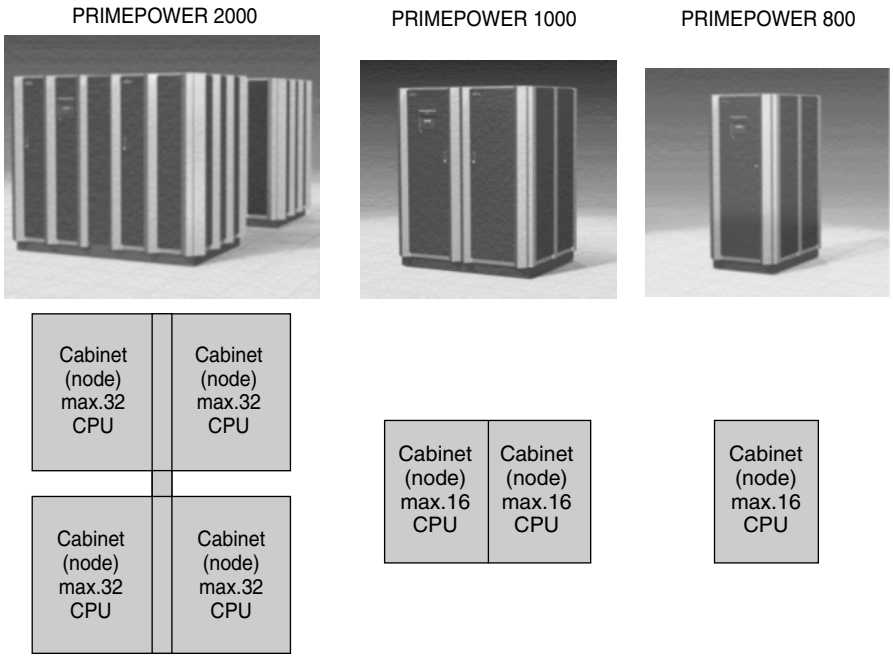

Figure 1
PRIMEPOWER 2000, 1000, and 800 node structure.

PRIMEPOWER 2000 permits an SMP configuration with a maximum of 128 CPUs, the largest in the industry.

2.2 Memory and I/O sub-system with large capacity and high performance

The PRIMEPOWER 2000 supports up to 512 gigabytes of memory with up to 128-way memory interleaving. High-speed I/O connections are supported by a high-performance I/O interface with three 64-bit/66 MHz PCI slots and three 64-bit/33MHz PCI slots per system board. The PRIMEPOWER 2000, the top of the range, can be configured with a maximum of 192 PCI slots, providing sufficient capacity to implement large-scale systems supporting many external interfaces.

2.3 High-performance SPARC64 GP processor

The SPARC64 GP processor used in the PRIMEPOWER family conforms to the SPARC International V9 architecture. The processor features an advanced micro-architecture that implements out-of-order execution for all instructions, large-capacity cache memory, register-renaming and advanced instruction branch prediction. Cache reliability is provided by ECC (error checking and correction) on both internal and external caches. The PRIMEPOWER 2000, 1000, and 800 employ the 450 MHz SPARC64 GP processor with a large-capacity secondary cache of 8 Mbytes. These processors can be upgraded to future higher-frequency processors.

2.4 HA (High Availability) features

Major hardware components can be configured redundantly and are hot-replaceable. This applies to system boards (Processors, memory, and PCI cards), power supply units, fan trays, system control boards and disks. This is a key feature in achieving 365 days per year continuous operation and ensures continuity of system operation or at worst minimal down-time due to hardware failure.

2.5 Partitioning achieves flexible system configuration

The PRIMEPOWER models support a partitioning feature that enables processors, memories and PCI cards to be divided into groups. This feature allows users to divide the PRIMEPOWER into a production system and a development system, for example, and allows integrated management of multiple servers. The feature not only reduces the load on the system administrator, but also provides flexibility in responding to various operational requirements.

3. System configuration

The PRIMEPOWER 2000 consists of up to four interconnected cabinets (nodes), each with a capacity of 32 CPUs, for a total of up to 128 CPUs. The PRIMEPOWER 1000 supports two cabinets (nodes), each of 16 CPU capacity, for a total of 32 CPUs. The PRIMEPOWER 800 has up to 16 CPUs in its single cabinet (**Figure 1**).

Nodes are connected using the second level crossbar (L2 crossbar) switch, which not only allows every processor to access all memory but also guarantees coherency. The switch is synchronized and driven by redundant clock sources. This enables every processor to check the access requests from others. This behavior is called "snoop"; snoop performance is a significant factor in the ultimate performance of an SMP architecture machine. The PRIMEPOWER 2000 has a snoop rate of 57.6 gigabytes/s, which is sufficient to support scaleable expansion to the maximum configuration of 128 processors.

The components connected to the system-control network, such as the system boards, power supply units and fan trays, are monitored and controlled by a processor independent of the main processor. **Figure 2** shows the hardware structure of the PRIMEPOWER 2000. The following section describes the major hardware components.

4. Node configuration

Each node (cabinet) of the PRIMEPOWER 2000 contains up to 8 system boards. In the PRIMEPOWER 1000 and 800, each node contains up to four system boards. **Figure 3** shows the structure of the system and crossbar boards in a node. To perform high-speed transmission with two-level crossbar switches, a back plane that connects system boards and crossbar boards is wired in about 8000 signal lines of equal length. Each node has redundant power supply units and fan trays.

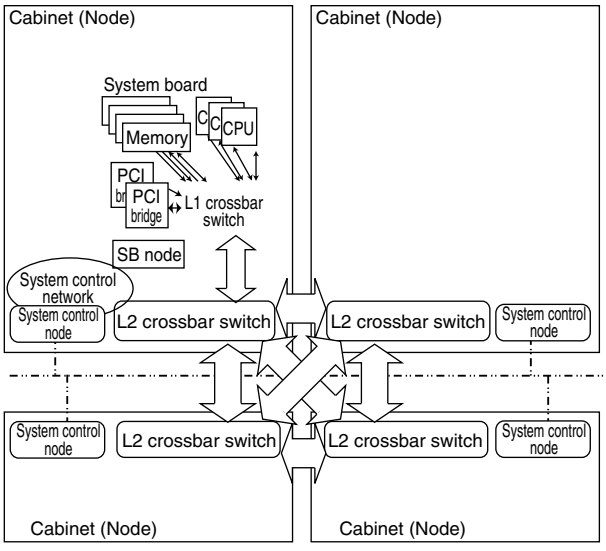

Figure 2
Maximum configuration of PRIMEPOWER 2000.

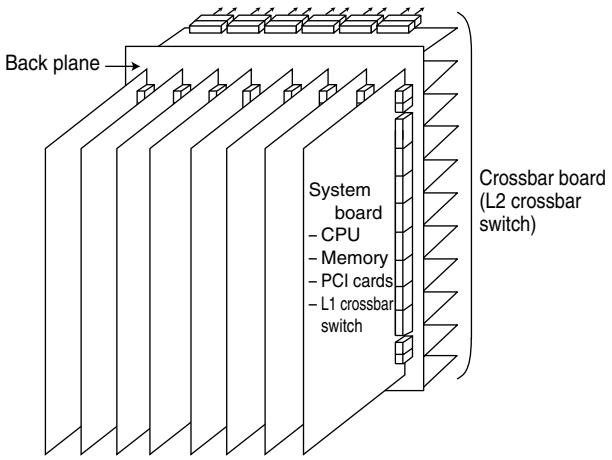

Figure 3
Node configuration of PRIMEPOWER 2000.

5. System board

A maximum of four CPUs, 16 gigabytes of memory, six PCI cards and a First Level (L1) crossbar switch are mounted on each system board. This implementation allows optimum size selection from a minimum system (4 CPUs) through the largest system (128 CPUs), system board partitioning, and system board hot replacement.

When these components are mounted on a small area of a system board, they must be laid out in good balance at high density. This high-density layout has been achieved by employing the newest implementation technology to reduce the size of the parts.

The L1 crossbar switch, which is implemented in the system board, maintains memory coherency among the processors, PCI devices and memory. The performance of the snoop plays an important role by maintaining cache coherence in the SMP architecture. A maximum of 4 address requests in one cycle are processed in parallel to maintain high throughput for each system board.

The memory subsystem implemented on a system board has four-way memory address interleaving. It controls a maximum of four requests simultaneously and uses SDRAM internal banks efficiently. This memory subsystem achieves a peak memory throughput of 8.52 gigabytes/s per system board. The memory, processor caches and system data-bus are protected by ECC, enabling continuous operation, even with a one-bit error. This is very important because this system supports up to 512 gigabytes of physical memory space. Moreover, memory will still function if one memory element on the memory module has failed completely. To meet these requirements, Fujitsu has developed three ASICs based on the newest 0.18 micrometer copper CMOS LSI technology (**Table 1**) and implemented as an MCM (Multi-Chip Module) to achieve high speed and small package size.

Table 1
ASICs mounted on the system board.

ASIC	Circuit size (K gates)	Number of I/O pins*	Number of chips
Memory & coherency control	1096	4732	1
Cache control	481	1762	4
Data crossbar	362	1681	8

*Includes power supply with GND.

6. The L2 crossbar switch

To support an SMP configuration of up to 128 CPUs, L2 crossbar switches are mounted on crossbar boards in each cabinet (node). The L2 crossbar interconnects the system boards within and between nodes.

6.1 Features of the L2 crossbar switch

In an SMP system, good scalability depends on the memory subsystem delivering high throughput and low access latency. The L2 crossbar switch in combination with the L1 crossbar switch fulfills the following specifications:

- Maximum snoop rate of 57.6 gigabytes/s
- Maximum data bandwidth of 81.9 gigabytes/s
- Memory access latency of about 300 ns.

The bus signal group has an error detection function. Error detection and correction by ECC is available in each data bus, and the impact of a fault in an address bus is minimized by multiple block parity checks. The cache state signals are duplicated to allow for selection of an error-free Cache State at any time. All control lines are protected by parity, important lines are duplicated and a retry feature is implemented. In addition, accurate identification of failed components aids rapid replacement and recovery.

6.2 Transmission technology

The L2 crossbar switch consists of the address crossbar switch, cache state crossbar switch and data crossbar switch.

The length of the signal lines between the system board and the crossbar board is about 900 mm. To transmit 200 - 225 MHz signals constantly, data transmission employs send/receive

clocks with phases offset from the system clock. In cable transmission between nodes, data is captured by a receive clock operating with a time delay from the synchronized clock to compensate for data transmission delay.

High-frequency operation of the L2 crossbar switch is achieved by adjusting the signal length of about 8000 total crossbar signal lines between the system boards and the crossbar boards.

Fujitsu developed three types of bus control LSI chips to implement high-speed operation and reliability for the L2 crossbar switch. **Table 2** shows the specifications of the LSI chips.

7. System control network

7.1 Overview

The PRIMEPOWER 2000, 1000, and 800 employ a newly developed system control network to support redundant configuration and the hot-swap function on all major components.

In this system control network, shown in **Figure 4**, each component in a node is connected using the serial bus method (multi-drop) to attain high reliability and high performance of data and event communication.

In previous systems, signal lines used to control and monitor the components were connected to the system control component directly. In the PRIMEPOWER 2000, 1000, and 800, this connection is implemented using a serial bus connection, drastically reducing cable cost. Control is distributed between Local Cabinet Interface (LCI) nodes; this distributed control mechanism helps to reduce the load on system control components.

Table 2
ASICs mounted on the crossbar board.

ASIC	Circuit size (K gates)	Number of I/O pins*	Number of chips per node
Address	200	900	10
Cache state	115	607	4
Data	375	900	13

*Includes the power supply with GND.

7.2 LCI nodes

The PRIMEPOWER models use 7 types of LCI nodes, as shown in **Table 3**. The system control board nodes act as a central control and monitor. Other nodes are controlled and monitored by the system control board nodes.

7.3 Network duplication

Each LCI node can switch between duplicated networks. When a permanent error occurs on one network, network operations switch to the other network. The crossbar board node mounted on the edge of the network continuously monitors the network and controls the switching circuit. When a short or open failure of a circuit is detected, the crossbar node switches the network.

7.4 Hot system maintenance function

The LCI nodes on the system control network are hot-replaceable. This function is achieved using the RS485 standard. This standard regulates the insertion and detachment of transceivers within a distributed power network. A power sequence connector prevents system instability during insertion/detachment of the nodes. Power is supplied to the network constantly by the distributed power supply system, and is implemented using an intermediate voltage distribution

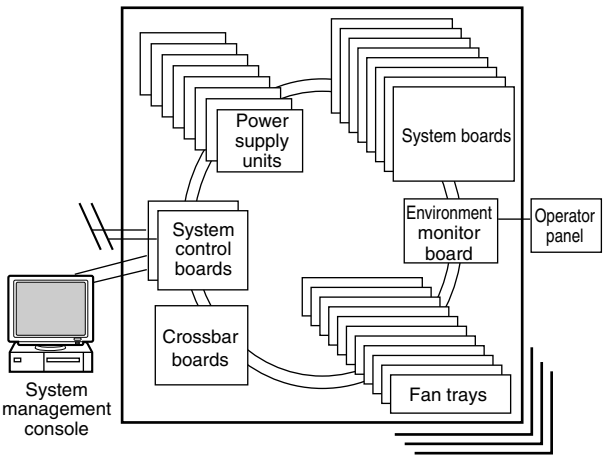

Figure 4
System control network.

Table 3
LCI nodes and their functions.

Node name	Target function
system control board	– Controls and monitors the system control networks
environment monitor board	– Controls the operator panel and the temperature monitor – Maintains various ID info
crossbar board	– Controls the clocks – Controls and monitors crossbar boards, and the duplicated LCI buses
system board	– Controls and monitors the system boards – Resets control – Controls and monitors temperature/current/voltage
power supply unit (2 types)	– Controls and monitors DC output – Monitors temperature and fan rotation
fan tray	– Controls and monitors fan rotation

method (+8.0 V) which uses a local regulator mounted on each node.

Each LCI node has READY lamp and a non-lock push button. Service personnel can detach a failed node from the live network and replace it with a new component by using the button and checking the lamp.

8. Operational functions

By using the partition feature and the dynamic reconfiguration function, the PRIMEPOWER 2000, 1000, and 800 meet the requirements of advanced operations such as server consolidation. Multiple partitions are controlled and operated through the system management console.

8.1 System management console

The system management console is installed alongside the PRIMEPOWER system, and is used by system administrators to define and control partitions. System administrators can use the system management console to provide a console window for each partition, as an NTP server for partitions, and as an install server for partitions. The console provides various status indications for the system and for network components, and also provides event notification functions including failure event notification. This allows the system administrator to recognize failures immediately and take action to replace failed components.

8.2 Dynamic reconfiguration

The PRIMEPOWER 2000, 1000, and 800 support the "Dynamic Reconfiguration" (DR) feature, which permits system administrators to change the partition configuration while the system is running. The unit of change is a system board. In order to use the dynamic reconfiguration feature, system administrators need to configure the system to meet DR conditions.

9. Summary

This paper is an overview of the features and functions of the PRIMEPOWER 2000, 1000, and 800 models. The PRIMEPOWER 2000 is a SMP server providing scalability to a maximum of 128 CPUs and 512 gigabytes of memory. It responds to the most demanding requirements by interconnecting up to four 64-CPU nodes. This connection is implemented using the L2 crossbar developed specifically for this purpose.

Each PRIMEPOWER model is designed for constant operation, 24 hours per day, 365 days per year. Most major components can be configured redundantly and hot replaced. As well, the various status indications and events notification functions support mission critical operations.

These high-scalability and high-availability features enhance high-performance UNIX servers, offering more power and greater flexibility. Fujitsu will continue to enhance the performance of these servers and to enrich their high availability features.

References

1) SPARC64 III, Microprocessor Report, December 8, 1997.
http://www.hal.com/home/sparc643_mda.html

Naoki Izuta received the B.S. degree in Precision Engineering from the University of Tokyo, Tokyo, Japan in 1985. He joined Fujitsu Ltd., Kawasaki, Japan in 1985 and has been engaged in development of operating system, planning of server products and designing of system architecture. He is currently in charge of system architecture of high-end UNIX server products.

Toshiyuki Shimizu received the B.S. degree in Electro-Physics and the M.S. degree in Computer Science both from Tokyo Institute of Technology, Tokyo, Japan in 1986 and 1988, respectively. He joined Fujitsu Laboratories Ltd., Kawasaki, Japan in 1988 and has been engaged in research and development of the architecture of parallel computers. He currently belongs to Fujitsu Ltd., Kawasaki, Japan and is designing the UNIX server products.

Toru Watabe received the B.S. degree in Industrial Engineering from Aoyamagakuin University, Tokyo, Japan in 1985. He joined Fujitsu Ltd., Kawasaki, Japan in 1985 and has been engaged in development of system ASIC on server products.

Tetsuhiko Ichihashi received the B.S. degree in Electronics and Communications Engineering from Meiji University, Tokyo, Japan in 1988.
He joined Fujitsu Ltd., Kawasaki, Japan in 1988, where he has been engaged in development of UNIX server products.