Fujitsu M10 Server Architecture

Fujitsu M10-1/M10-4/M10-4S
Featuring the SPARC64™ X
and the SPARC64™ X+ Processors

White Paper

FUJITSU

ORACLE

Manual Code: C120-E690-05EN
March 2016

any product or technology described herein are those expressly set forth in the license agreement pursuant to which the product or technology is provided.

EXCEPT AS EXPRESSLY SET FORTH IN SUCH AGREEMENT, ORACLE OR FUJITSU LIMITED, AND/OR THEIR AFFILIATES MAKE NO REPRESENTATIONS OR WARRANTIES OF ANY KIND (EXPRESS OR IMPLIED) REGARDING SUCH PRODUCT OR TECHNOLOGY OR THIS DOCUMENT, WHICH ARE ALL PROVIDED AS IS, AND ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING WITHOUT LIMITATION ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT, ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT SUCH DISCLAIMERS ARE HELD TO BE LEGALLY INVALID. Unless otherwise expressly set forth in such agreement, to the extent allowed by applicable law, in no event shall Oracle or Fujitsu Limited, and/or any of their affiliates have any liability to any third party under any legal theory for any loss of revenues or profits, loss of use or data, or business interruptions, or for any indirect, special, incidental or consequential damages, even if advised of the possibility of such damages.

DOCUMENTATION IS PROVIDED "AS IS" AND ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT SUCH DISCLAIMERS ARE HELD TO BE LEGALLY INVALID.

Fujitsu M10 is sold as SPARC M10 Systems by Fujitsu in Japan.

Fujitsu M10 and SPARC M10 Systems are identical products.

Avis de non-responsabilite : les seules garanties octroyees par Oracle et Fujitsu Limited et/ou toute societe affiliee de l'une ou l'autre entite en rapport avec ce document ou tout produit ou toute technologie decrits dans les presentes correspondent aux garanties expressement stipulees dans le contrat de licence regissant le produit ou la technologie fournis.

SAUF MENTION CONTRAIRE EXPRESSEMENT STIPULEE AU DIT CONTRAT, ORACLE OU FUJITSU LIMITED ET/OU LES SOCIETES AFFILIEES A L'UNE OU L'AUTRE ENTITE DECLINENT TOUT ENGAGEMENT OU GARANTIE, QUELLE QU'EN SOIT LA NATURE (EXPRESSE OU IMPLICITE) CONCERNANT CE PRODUIT, CETTE TECHNOLOGIE OU CE DOCUMENT, LESQUELS SONT FOURNIS EN L'ETAT. EN OUTRE, TOUTES LES CONDITIONS, DECLARATIONS ET GARANTIES EXPRESSES OU TACITES, Y COMPRIS NOTAMMENT TOUTE GARANTIE IMPLICITE RELATIVE A LA QUALITE MARCHANDE, A L'APTITUDE A UNE UTILISATION PARTICULIERE OU A L'ABSENCE DE CONTREFACON, SONT EXCLUES, DANS LA MESURE AUTORISEE PAR LA LOI APPLICABLE. Sauf mention contraire expressement stipulee dans ce contrat, dans la mesure autorisee par la loi applicable, en aucun cas Oracle ou Fujitsu Limited et/ou l'une ou l'autre de leurs societes affiliees ne sauraient etre tenues responsables envers une quelconque partie tierce, sous quelque theorie juridique que ce soit, de tout manque a gagner ou de perte de profit, de problemes d'utilisation ou de perte de donnees, ou d'interruptions d'activites, ou de tout dommage indirect, special, secondaire ou consecutif, meme si ces entites ont ete prealablement informees d'une telle eventualite.

LA DOCUMENTATION EST FOURNIE "EN L'ETAT" ET TOUTE AUTRE CONDITION, DECLARATION ET GARANTIE, EXPRESSE OU TACITE, EST FORMELLEMENT EXCLUE, DANS LA MESURE AUTORISEE PAR LA LOI EN VIGUEUR, Y COMPRIS NOTAMMENT TOUTE GARANTIE IMPLICITE RELATIVE A LA QUALITE MARCHANDE, A L'APTITUDE A UNE UTILISATION PARTICULIERE OU A L'ABSENCE DE CONTREFACON.

Contents

Introduction

Fujitsu has developed the SPARC64™ X (ten) and the new SPARC64™ X+ (ten plus) processor to combine high performance and high reliability. The Fujitsu M10 systems which surround the SPARC64™ X/SPARC64™ X+ processor merge numerous hardware and software technologies to provide customers with the most appropriate solution for their ever-growing IT infrastructure.

With the SPARC64™ X/SPARC64™ X+ processor developed for UNIX servers, the peripheral ASIC functions have been consolidated into the processor. Many processing functions, which were traditionally handled by software, have been built in to the processor hardware by adding multiple, dedicated instructions; thus achieving significant improvement in processing speed (throughput) and overall performance.

SPARC64™ X/SPARC64™ X+ processors are connected together in Fujitsu M10 systems using a cutting-edge fast interconnect technology. Moreover, the Fujitsu M10-4S model adopts the Building Block method of expansion which can reduce customers' initial investment and achieve linear performance as the system grows to meet increasing demand. Up to 16 Fujitsu M10-4S chassis can be interconnected to build a single, large system with up to 64 CPUs, delivering the highest performance in an extremely flexible and scalable system. These features make the Fujitsu M10 systems the most appropriate servers for datacenters in the cloud computing era.

Fujitsu M10 systems are offered with two kinds of processors, SPARC64™ X+ and SPARC64™ X. With these two processors customers are able to select the most appropriate engine depending on the work load required. Furthermore, the Fujitsu M10-4S server supports combined SPARC64™ X and SPARC64™ X+ chassis in a singlessystem, protecting previous IT investments.

Fujitsu M10 systems also benefit from improved power-saving features and increased ease of installation. Together SPARC64™ X/SPARC64™ X+ processors, with consolidated peripheral ASIC functions, and Fujitsu M10 systems, with high-efficiency power supplies and novel new cooling technology, lead directly to a very densely packaged server delivering substantial space savings and reduction in power usage.

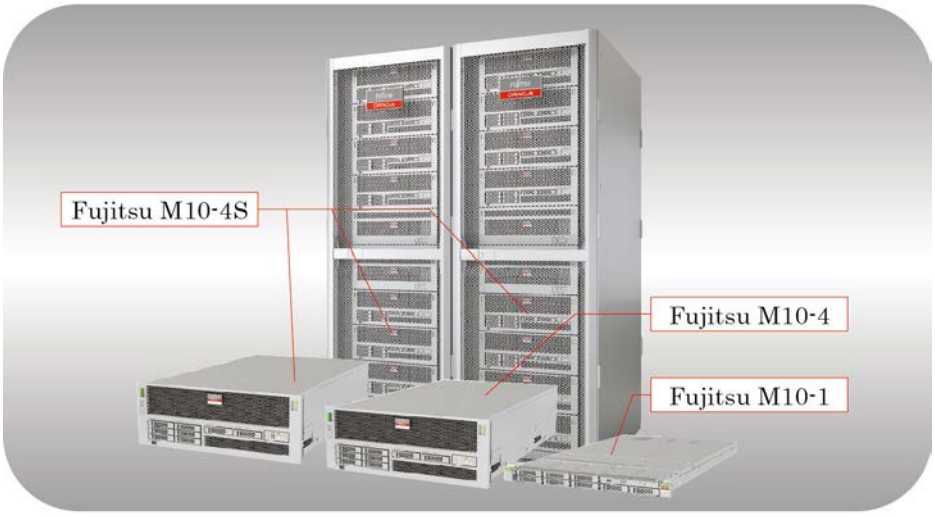
1.Instead of expanding a system through individual add-on modules or costly chassis upgrades, the Building Block method allows common 4-socket chassis (blocks) containing CPUs, memory and I/O expansion slots to be connected as if stacking up children's blocks.

1. Fujitsu M10 Systems

Fujitsu M10 systems provide the total platform required in modern IT infrastructure with leading technologies including a high-performance processor, cutting-edge semiconductor advancements, and the high reliability and high quality inherited from Fujitsu's mainframe and SPARC Enterprise M-series systems. Fujitsu M10 systems have further enhanced system reliability, flexibility, and scalability to meet the demands of cloud computing environments.

Product Lineup

Fujitsu M10 systems cover a product lineup which is applicable to a broad range of industries. As an entry-level server, Fujitsu M10-1 provides a system architecture that can be configured with up to 16 processor cores and large-capacity memory and disk in a space-saving one rack-unit (RU) chassis. The CPU resources can start very small and then be expanded in stages using the CPU Activation feature. For the midrange, with high performance and reliability, Fujitsu M10-4 can start from two physical CPU sockets and expand to four sockets by adding a CPU module. Fujitsu M10-4 also implements a new Liquid Loop Cooling technology that leads to high compute density in a 4RU chassis. To alleviate the financial burden of initial capital investment that customers face, and to achieve a flexible configuration which is suitable for datacenters, Fujitsu M10-4S introduces the Building Block (BB) expansion methodology. The Building Block concept treats a single 4-socket enclosure as one "block", and allows the machine to expand in units of this block as if stacking up children's blocks. Fujitsu M10-4S supports the connection of up to 16 blocks using a fast (14 Gbps (uni-directional)) interconnect. The maximum 16 Building Block system is installed in two 19-inch racks. Scalability from 4-sockets all the way up to 64-sockets in a mission-critical, high-end model delivers the flexibility that customers demand. Customers can start with four sockets and expand in-step with the growth of their business, thus significantly reducing the high initial costs of traditional high-end models.



Fujitsu M10 Product Lineup

Fujitsu M10-1, M10-4 and M10-4S System Specifications

		Fujitsu M10-1	Fujitsu M10-4	Fujitsu M10-4S (1BB)	Fujitsu M10-4S (16BB)
SPARC64™ X Processor (*1)	Processor	2.8 GHz	2.8 GHz	3.0 GHz	3.0 GHz
	Secondary cache	22 MB	24 MB	24 MB	24 MB
	Maximum number of CPUs	1	2 or 4	2 or 4	Up to 64
	Number of cores	16	16 per Socket	16 per Socket	Up to 1024
	Number of threads	32	32 per Socket	32 per Socket	Up to 2048
SPARC64™ X+ Processor (*1)	Processor	3.2 GHz/3.7 GHz	3.4 GHz/3.7 GHz	3.7 GHz	3.7 GHz
	Secondary cache	22 MB/24 MB	24 MB	24 MB	24 MB
	Maximum number of CPUs	1	2 or 4	2 or 4	Up to 64
	Number of cores	16/8	16/8 per Socket	16 per Socket	Up to 1024
	Number of threads	32/16	32/16 per Socket	32 per Socket	Up to 2048
Maximum Memory	Maximum size	1 TB	4 TB	4 TB	64 TB
	Maximum number of mounted memory modules	16	64	64	1024
Internal Storage	Interface	SAS	SAS	SAS	SAS
	Built-in disk	8	8	8	128

I/O Slots	Interface	PCI Express Gen3	PCI Express Gen3	PCI Express Gen3	PCI Express Gen3
	PCIe slot	3	11	8	128
Onboard Interface	1 Gb Ethernet port	4	4	4	64
	SAS port	1	1	1	16
	USB port	2	2	2	32
Form Factor		1 RU	4 RU	4 RU	40 RU x2 (including XB Box)
Virtualization Functions	Number of partitions	1	1	1 (one partition per BB)	Up to 16
	Maximum number of domains	32	128	128	256 per physical partition

*1 Either the SPARC64™ X or SPARC64™ X+ processor can be mounted in an enclosure. A SPARC64™ X M10-4S Building Block can be connected with a SPARC64™ X+ M10-4S Building Block.

Fujitsu M10-1

Fujitsu M10-1 employs a space-saving 1RU chassis, while also providing the high reliability required of a mission-critical system. Fujitsu M10-1 is a high-performance, highly reliable and space-saving entry-level server appropriate for datacenter consolidation and virtualization.

Fujitsu M10-1 supports a single high-performance CPU with up to 16 cores , up to 1 TB of memory, and up to 8 internal storage slots (supporting hard disk drives and/or SSD). The Fujitsu M10-1 also includes three PCI Express slots which support PCI Express Generation 3.0. To support further I/O expansion, external PCI Expansion Units can be connected to provide mid-range class scalability with up to 23 PCI Express Generation 3.0 slots. Two power supplies and seven fan units power and cool the server with built-in redundancy.

The Fujitsu M10-1 model scales from entry-level up to the mid-range class. The system can at first be purchased with the minimum required CPU resources. Then, as needed, the CPU resources can be expanded in a step-by-step manner by purchasing CPU core activations (Rights to Use CPU core). This model reduces initial investment and supports step-by-step processing capability growth to better match a customer's business expansion. In addition, server virtualization and system consolidation are realized through both Oracle Solaris Zones and Oracle VM Server for SPARC. Fujitsu M10-1 builds on the initial investment reductions and step-by-step processing expansion, and raises the maximum capacity of CPU cores, memory, and I/O expansion significantly over existing entry-level servers.

Fujitsu M10-4

Fujitsu M10-4 is a high-performance and highly reliable midrange server which is appropriate for datacenter consolidation and virtualization tasks requiring more processor, memory, and I/O capacity than are available in the Fujitsu M10-1 model. It provides the flexibility, scalability and reliability required to support customers' mission-critical business.

Fujitsu M10-4 occupies 4RU and supports up to four CPUs (up to 64 cores total), up to 4 TB of memory, and up to 8 internal storage slots (supporting hard disk drives and/or SSD). The Fujitsu M10-4 also includes 11 PCI Express Generation 3.0 slots. To support further I/O expansion, external PCI Expansion Units can be connected to provide up to 71 PCI Express Generation 3.0 slots. Two power supply units provide redundant power to the server. The Liquid Loop Cooling (LLC) system, a newly introduced cooling technology, and five fan units cool the server.

The system can at first be purchased with the minimum required CPU resources. Then, as needed, the CPU resources can be expanded in a step-by-step manner by purchasing CPU core activations (Rights to Use CPU core). In addition, server virtualization and system consolidation are realized through both Oracle Solaris Zones and Oracle VM Server for SPARC. Fujitsu M10-4 builds on the initial investment reductions and step-by-step processing expansion found in Fujitsu M10-1, and expands CPU socket capacity from an initial two sockets up to four sockets. This will offer maximum utilization of memory and I/O. In this way, Fujitsu M10-4 redefines the scalability, flexibility and processing capacity of midrange servers.

Fujitsu M10-4S

Fujitsu M10-4S provides the world's highest scalability and flexibility, and covers a broad range of computing needs, from midrange to high-end. Fujitsu M10-4S is the most appropriate for use in cloud computing and big data processing infrastructures that are currently large scale or have the potential to grow significantly over time.

In addition to the high performance and high reliability found in all Fujitsu M10 models, Fujitsu M10-4S provides flexible scalability by virtue of the Building Block expansion methodology. With an Expansion Rack installed, a customer can freely combine scale-up configurations with scale-out configurations best suited for distributed parallel processing without interrupting working partitions. With two Expansion Racks, customer can scale-up to 16 interconnected Building Blocks without shutting down the entire system.

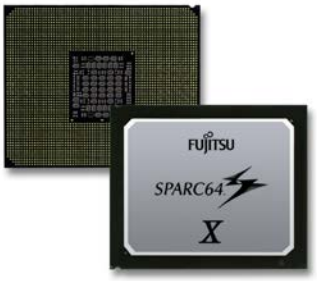
The Fujitsu proprietary interconnect technology that provides the connection between Building Blocks achieves linear performance improvement up to the maximum 16BB configuration. The system can at first be purchased with the minimum required CPU, memory, PCI Express slots and disk bay resources. As needed, additional Building Blocks can be added to increase capacity. CPU resources can also be expanded in a step-by-step manner by purchasing CPU

core activations (Rights to Use CPU core) to achieve finer granularity expansion than adding whole CPU chips.

Furthermore, additional SPARC64™ X mounted Fujitsu M10-4S Building Blocks can be connected to not only a SPARC64™ X mounted Fujitsu M10-4S Building Block but also a SPARC64™ X+ mounted Fujitsu M10-4S Building Block.

2. SPARC64™ X/SPARC64™ X+ Processor

SPARC64™ X and SPARC64™ X+ are the latest processors developed by Fujitsu to combine high performance and high reliability for UNIX servers. The processor features high operating frequencies, multi-core and multithreading features, and high memory throughput. Furthermore, up to 64 CPU chips can be connected in a single system, offering extreme high performance and scalability. Fujitsu's heritage of mainframe-class high reliability technology is found throughout the processor and corresponding systems.



SPARC64™ X/SPARC64™ X+ inherit the robust high-reliability technologies found in previous SPARC64™ processors, and at the same time offer multiple functional enhancements. Enhancements extend beyond the microarchitecture improvements, such as greatly increased core counts and the consolidation of peripheral chip functions into the processor, to include significant instruction set enhancements such as register extensions, SIMD functionality, and cryptographic processing performance improvements.

SPARC64™ X/SPARC64™ X+ consolidate the peripheral ASIC functions into the processor. This leads directly to faster processing and improvement in performance per Watt. Compared to SPARC64™ VII+, SPARC64™ X delivers approximately twice the performance per CPU core, and a 7.5-times enhancement in per socket performance (as measured with SPECint_rate). SPARC64™ X+ achieves a further improvement of 2.4 times the performance per CPU core, and a 9.5 times the performance per socket.

The significant technological enhancements of SPARC64™ X/SPARC64™ X+ will be explained in detail in the last portion of this document.

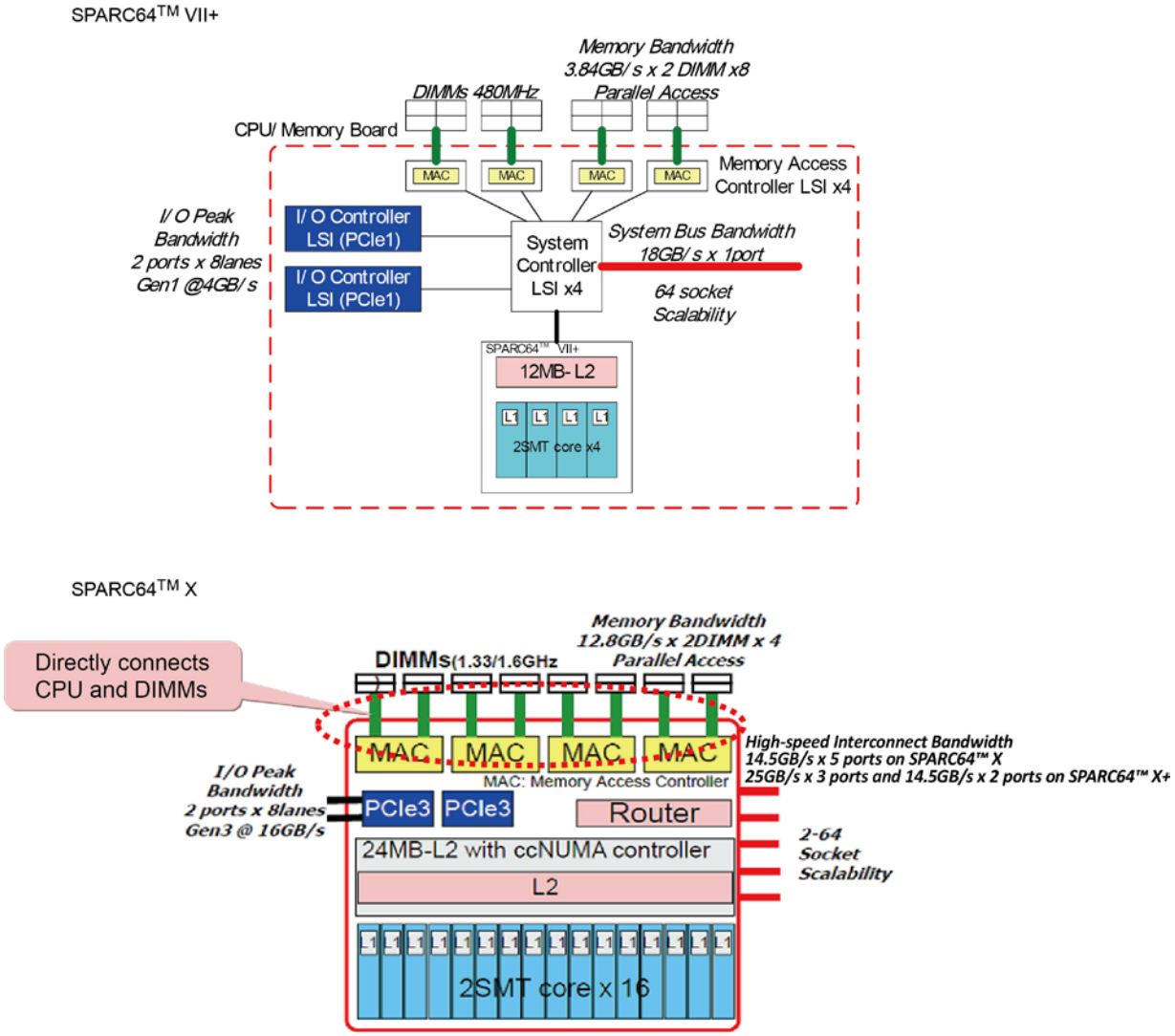


Figure 2-1. SPARC64™ X/SPARC64™ X+ Function Enhancement

3. System Architecture

In Fujitsu M10 systems, processors, system interconnects, and the memory and I/O subsystems work in concert to create a scalable, high-performance platform ready to address a wide range of workloads, from consolidation of general purpose enterprise computing to the fastest, largest, and most secure database processing.

The design of the Fujitsu M10 systems focuses on high reliability, and places emphasis on maximizing the merits of memory locality in a ccNUMA architecture to deliver outstanding performance. The characteristics and capabilities of every subsystem within the Fujitsu M10 systems work toward this goal. A high-bandwidth system bus, powerful SPARC64™ X and SPARC64™ X+ processors, dense memory support, and fast PCI Express combine within the Fujitsu M10 systems to deliver the highest levels of uptime and throughput, as well as dependable scaling for enterprise applications.

System Interconnect

The system interconnect underpins the highest levels of performance, scalability and reliability for the Fujitsu M10 systems. Multiple system controllers and crossbar units within the Fujitsu M10 systems provide point-to-point connections between CPU, memory, and I/O subsystems. Offering more than one bus route between components enhances performance and allows system operation to continue in the event of a fault. The system interconnect used in the Fujitsu M10 systems delivers as much as 6,553 GB/second of peak bandwidth, offering nine times more system throughput than Fujitsu's previous generation of high-end servers.

1. System Bus

High-end systems containing dozens of CPUs only provide scalability if all processors are able to actually contribute to the performance of the application. The ability to deliver near-linear scalability and fast, predictable performance for a broad set of applications rests largely on the capabilities of the system bus. The Fujitsu M10 systems utilize a system interconnect designed to deliver massive bandwidth and consistent, low latency between components. The system bus benefits IT operations by delivering balanced and predictable performance to application workloads.

The interconnect design maximizes the overall performance of the Fujitsu M10 systems. Implemented as point-to-point connections that utilize packet-switched technology, this system

bus provides fast response times by transmitting multiple data streams. Packet-switching allows the interconnect to operate at much higher system-wide throughput by eliminating "dead" cycles on the bus. All routes are uni-directional, non-contentious paths with multiplexed address, data, and control in each direction.

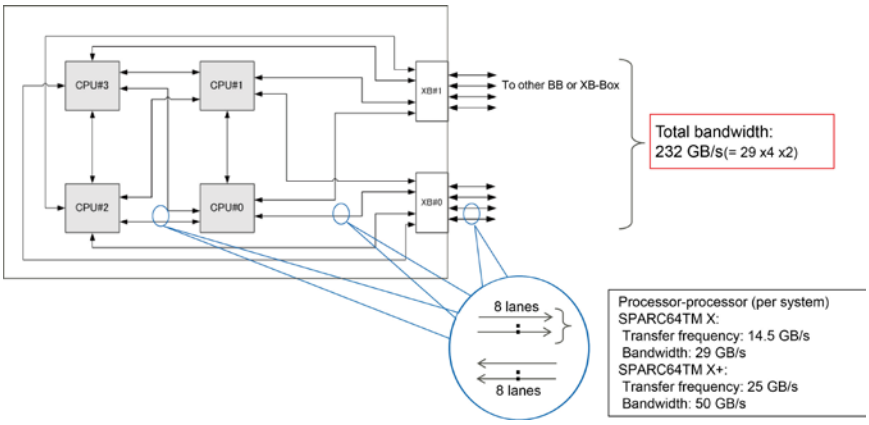


Figure 3-1. Bandwidth and Data Transfer Rate

System controllers within the interconnect architecture of Fujitsu M10-1, M10-4, and M10-4S servers direct traffic between CPUs in each chassis, memory, I/O subsystems, and interconnect paths.

2. Fujitsu M10-1 Interconnect Architecture

The Fujitsu M10-1 system is implemented on a single motherboard. Within the architecture, SPARC64™ X, or SPARC64™ X+, is a single CPU that performs all functions of memory access controller, I/O controller and system controller. SPARC64™ X/SPARC64™ X+ is connected to DIMMs and PCI Express switches. An architecture diagram of the Fujitsu M10-1 server is shown in Figure 3-2.

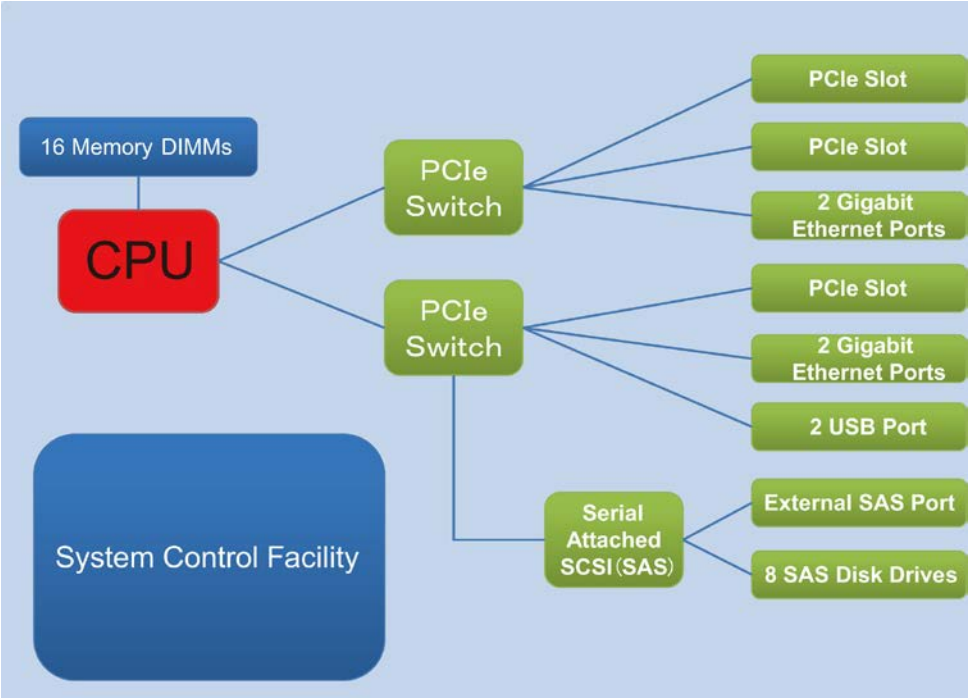


Figure 3-2. Fujitsu M10-1 Interconnect Architecture

3. Fujitsu M10-4 Interconnect Architecture

The Fujitsu M10-4 system is implemented on two motherboards, which function as a single logical system board. Fujitsu M10-4 supports up to either four SPARC64™ X or four SPARC64™ X+ processors, and as with Fujitsu M10-1, the memory access, I/O and system controllers are embedded with the CPU. Each controller connects to DIMMs and PCI Express switches, and all four SPARC64™ X/SPARC64™ X+ processors are interconnected. To increase I/O bus bandwidth and protect against processor failure, each PCI Express switch connects to two SPARC64™ X/SPARC64™ X+ processors (Figure 3-3).

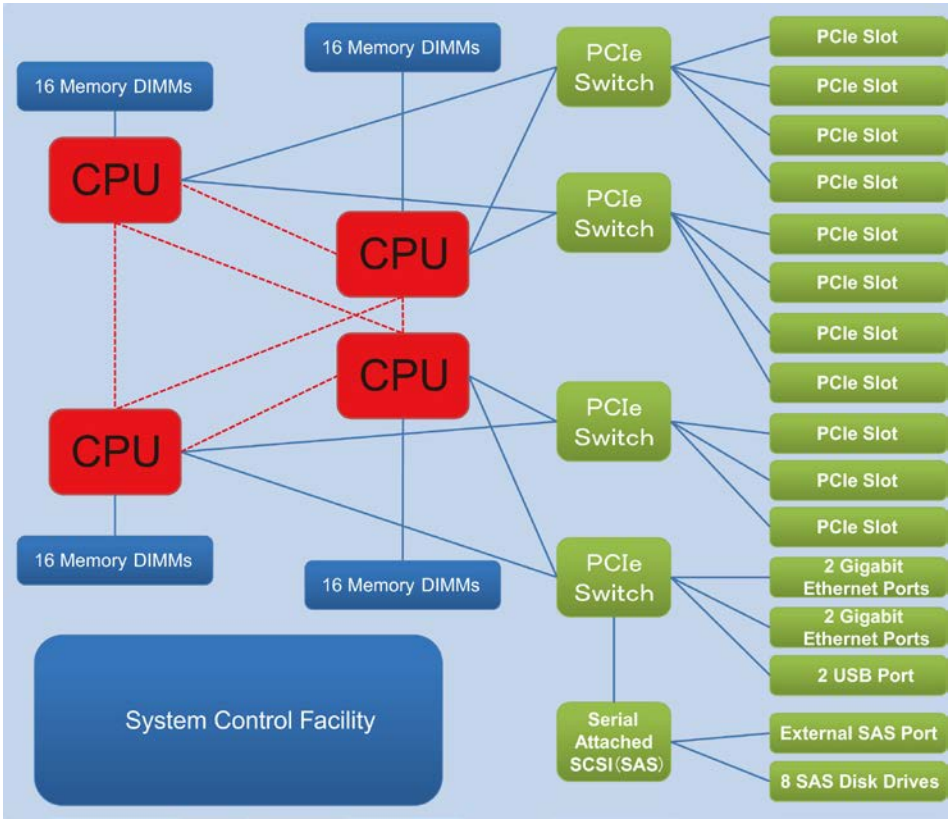


Figure 3-3. Fujitsu M10-4 Interconnect Architecture

4. Fujitsu M10-4S Interconnect Architecture

Fujitsu M10-4S implements crossbar units between multiple chassis to support the Building Block methodology. In each chassis, as in Fujitsu M10-4, two motherboards are implemented and function as a single logical system board. Each Fujitsu M10-4S chassis supports up to either four SPARC64™ X or four SPARC64™ X+ processors, and as with Fujitsu M10-1 and Fujitsu M10-4, the memory access, I/O and system controllers are embedded with the CPU. Each controller connects to DIMMs and PCI Express switches, and all four SPARC64™ X/SPARC64™ X+ processors are interconnected. To increase I/O bus bandwidth and protect against processor failure, each PCI Express switch connects to two SPARC64™ X/SPARC64™ X+ processors (Figure 3-4).

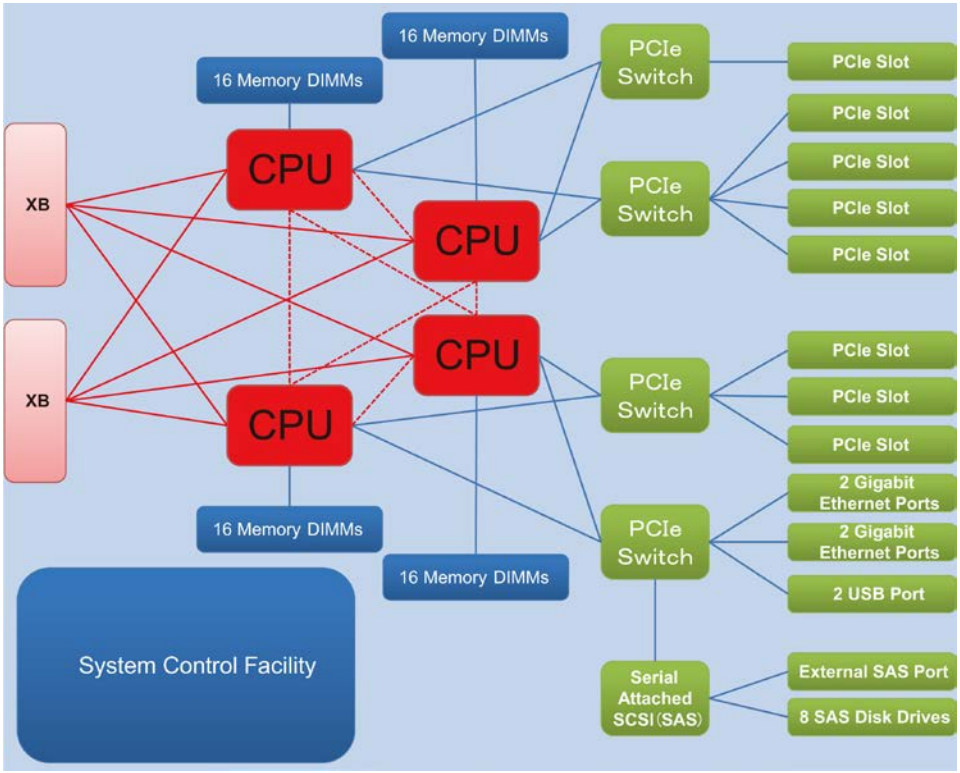


Figure 3-4. Fujitsu M10-4S Interconnect Architecture

In Fujitsu M10-4S, the system bus between all Building Block chassis is implemented as a crossbar switch. The system bus supports low-latency and high-throughput data transfers. To improve performance, the physical addressing of memory in the Building Block is evenly spread out across all system controllers in each Building Block.

The SPARC64™ X/SPARC64™ X+ processors have connections to the crossbar unit, supporting data transfer to other Building Blocks. Up to four Building Blocks can be directly connected using the crossbar unit built-in to each Fujitsu M10-4S chassis (Figure 3-5). With system configurations scalable up to 16 Building Blocks, each chassis is connected to all other chassis via the crossbar box (Figure 3-6).

The Fujitsu M10-4S Building Block mounted with SPARC64™ X processors can be connected to both the Fujitsu M10-4S Building Block with SPARC64™ X processors, and the Fujitsu M10-4S Building Block with SPARC64™ X+ processors.

This ensures the existing investment in the Fujitsu M10-4S Building Block with SPARC64™ X processors is protected.

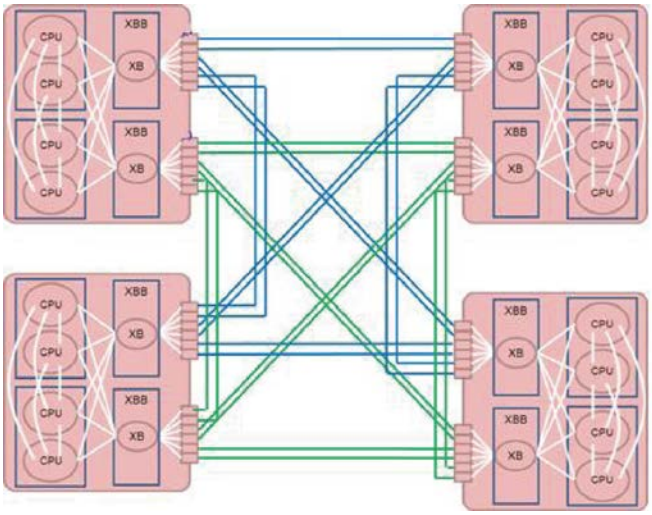


Figure 3-5. Direct BB Connection System Configuration (4BB)

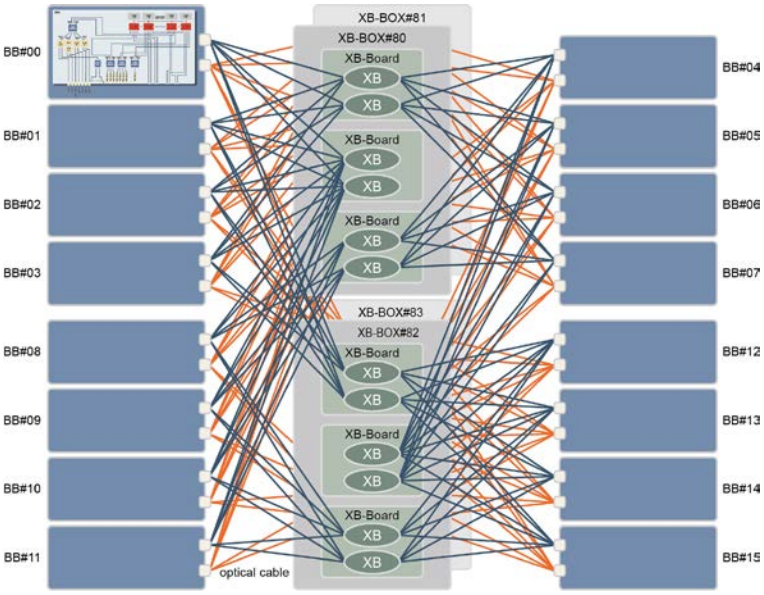


Figure 3-6. System Configuration of Connections Via the Crossbar Box (16BB)

5. System Interconnect Reliability Features

Built-in redundancy and reliability features of the Fujitsu M10 system interconnect enhance the stability of these servers. The interconnect is protected against loss or corruption of both transaction address and data with ECC or CRC protection on all system buses. When a single-bit data error is detected in a CPU, memory access controller, or I/O controller, hardware corrects the data and proceeds with the transfer. Also, when a multi-bit data error is detected by CRC on a bus which is connected via Fujitsu's high-speed interconnect technology (all crossbar-crossbar, crossbar-processor, and processor-processor busses), the hardware automatically resends the data. When the error cannot be recovered by resending the data, the specific bus is degraded. In the rare event of a hardware failure within the system interconnect, the system uses the surviving bus route on restart, isolating the faulty crossbar bus and facilitating the resumption of operation.

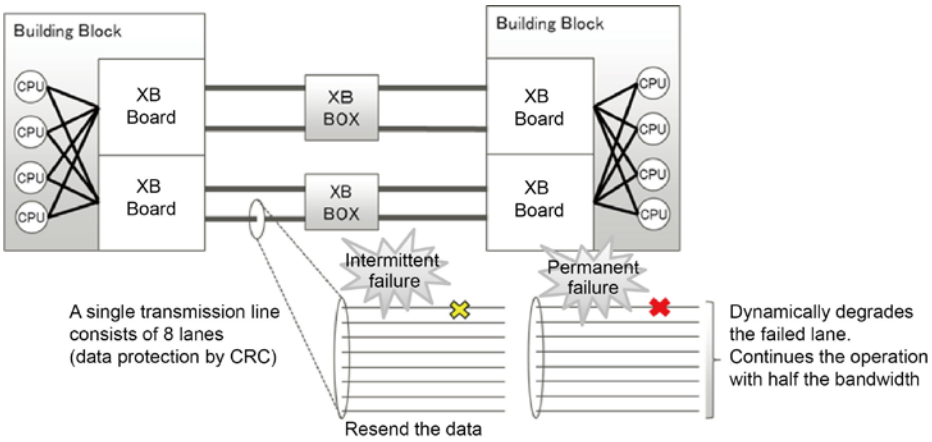


Figure 3-7. System Interconnect Reliability Features

Memory

The memory subsystem in Fujitsu M10 increases system scalability and throughput. The 16 Building Block configuration accommodates up to 64 TB of memory. All Fujitsu M10 systems support DDR3 DIMMs with 4-way memory interleaving to enhance system performance. Available DIMM capacities include 8 GB, 16 GB, 32 GB, and 64GB. Further details about the memory subsystem are described in Table 3-1.

Table 3-1. Fujitsu M10 Memory Subsystem Specifications

	SPARC M10-1	SPARC M10-4	Fujitsu M10-4S (1BB)	Fujitsu M10-4S (16BB)
Maximum Memory Capacity	1 TB	4 TB	4 TB	64 TB
Maximum DIMM Slots	16	64	64	1024
Unit of Memory Interleave	4 or 8 DIMMs	8 DIMMs	8 DIMMs	8 DIMMs
Maximum Number of Interleaves	4	8	8	128

Beyond performance, the Fujitsu M10 memory subsystem is built with reliability in mind. ECC protection is implemented for all data stored in main memory, and the following advanced features foster early diagnosis and fault isolation that preserve system integrity and raise application availability.

- Memory Patrol
  Memory patrol periodically scans memory for errors. This proactive function prevents the use of faulty areas of memory before they can cause system or application errors, improving system reliability.
- Memory Extended-ECC
  The memory Extended-ECC function in Fujitsu M10 provides single-bit error correction, supporting continuous operation even when events caused by memory device failures like burst read errors occur. This feature is similar to IBM's Chipkill technology.

1. Memory Mirroring

The Fujitsu M10 systems support memory mirroring capabilities. Memory mirroring is a high-availability feature appropriate when running applications with the most stringent availability requirements. When memory mirroring mode is enabled on the Fujitsu M10 systems, the memory subsystem duplicates the data on write and compares the data on read to each side of the memory mirror. In the event that errors occur at the bus or DIMM level, normal data processing continues through the other memory bus and alternate DIMM set. In Fujitsu M10 systems, memory can be mirrored between memory modules, using the memory access controller (MAC) built-in to the SPARC64™ X/SPARC64™ X+ processor (Figure 3-8).

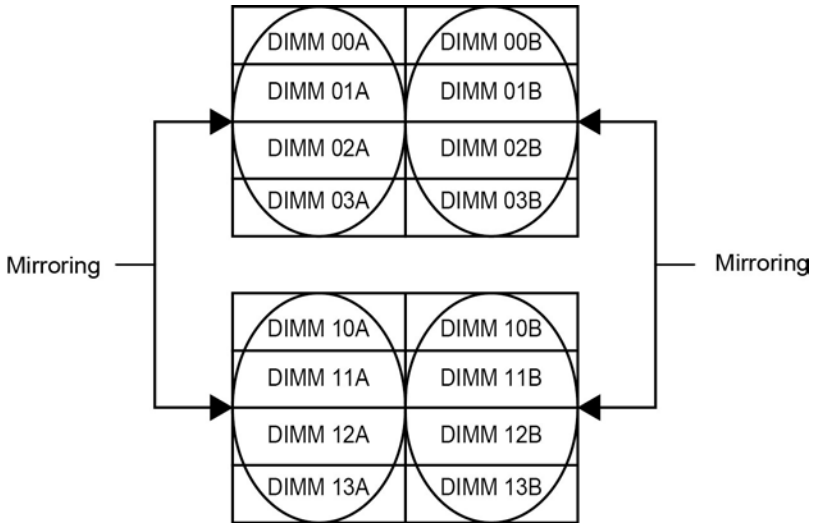


Figure 3-8. Fujitsu M10 Memory Mirroring Architecture

System Clock

In Fujitsu M10 systems the system clock is implemented independently for each CPU socket. Therefore, regardless of the Building Block configuration (single or multiple BB), even after a system clock failure, the system can be restarted by degrading only the failed CPU socket.

I/O Bus

Fujitsu M10 systems use a PCI Express bus to provide high-speed data transfer within the I/O subsystem. To provide optimal I/O performance for current and future Fibre Channel, InfiniBand, Gigabit Ethernet, and Flash PCI Express expansion cards, the Fujitsu M10 systems implement the PCI Express Generation 3.0 protocol in each processor. The peak data transfer rate of PCI Express Generation 3.0 reaches 8 GB/second of throughput.

1. I/O Subsystem Architecture

The use of PCI Express Generation 3.0 technology is a key to the performance of the I/O subsystem within the Fujitsu M10 systems. PCI Express Generation 3.0 switches provide the connection between the SPARC64™ X/SPARC64™ X+ processor and PCI Express Generation 3.0 slots, onboard devices, and internal drives. The PCI Express Generation 3.0 bus can also be extended to external PCI Expansion Units to provide a significant increase in PCI Express Generation 3.0 slots. The Fujitsu M10 I/O architecture has been designed to provide the scalability (small to very large slot count) and performance (latest PCI Express Generation 3.0

high throughput) required for server consolidation and virtualization.

To facilitate hot-plug of PCI Express cards, the Fujitsu M10-4 and M10-4S servers and the PCI Expansion Unit utilize PCI Express cassettes. PCI Express cards which support PCI hot-plug can be mounted in PCI Express cassettes and then inserted into on-board or PCI Expansion Unit PCI Express slots of a running server.

2. Fujitsu M10-1 Server I/O Subsystem

A depiction of the I/O subsystem of the Fujitsu M10-1 server is shown in Figure 3-9. Two PCI Express Generation 3.0 switches mounted on the motherboard of the Fujitsu M10-1 server connect all I/O components to the I/O controller in the SPARC64™ X/SPARC64™ X+ processor. The I/O subsystem supports the connection of external I/O devices by providing three PCI Express Generation 3.0 slots, one external SAS port, and two USB 2.0 ports. The external SAS port can be used to connect SAS tape or storage devices, and the USB 2.0 ports to connect a supported DVD device. In addition, external PCI Expansion Units increase the number of available PCI Express Generation 3.0 slots.

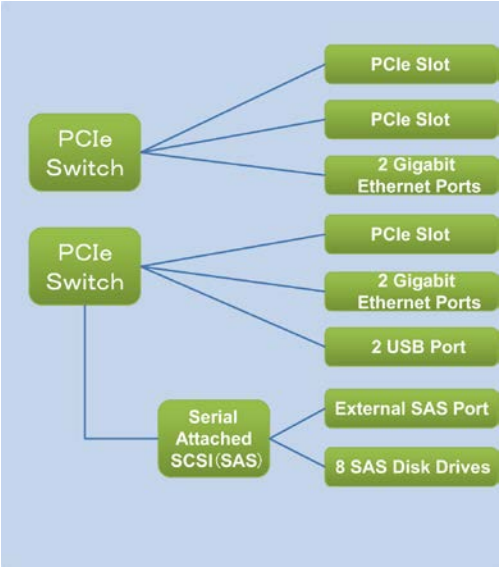


Figure 3-9. Fujitsu M10-1 I/O Subsystem Architecture

3. Fujitsu M10-4 Server I/O Subsystem

Four PCI Express Generation 3.0 switches mounted on the motherboard of the Fujitsu M10-4 server connect all I/O components to the I/O controller in the SPARC64™ X/SPARC64™ X+ processor. The PCI Express switches and the I/O controller in the CPU connect one-to-one or

one-to-two, depending on the number of CPU chips mounted. The I/O subsystem supports the connection of external I/O devices by providing 11 PCI Express Generation 3.0 slots, one external SAS port, and two USB 2.0 ports. The external SAS port can be used to connect SAS tape or storage devices, and the USB 2.0 ports to connect a supported DVD device. In addition, external PCI Expansion Units increase the number of available PCI Express Generation 3.0 slots.

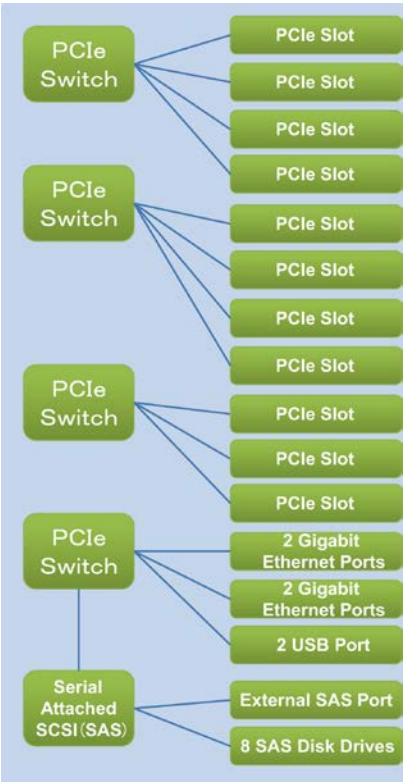


Figure 3-10. Fujitsu M10-4 I/O Subsystem Architecture

4. Fujitsu M10-4S Server I/O Subsystem

Four PCI Express Generation 3.0 switches mounted in a single Building Block of the Fujitsu M10-4S server connect all I/O components to the I/O controller in the SPARC64™ X/SPARC64™ X+ processor. The PCI Express switches and the I/O controller in the CPU connect one-to-one or one-to-two, depending on the number of CPU chips mounted in each Building Block. The I/O subsystem supports the connection of external I/O devices by providing 8 PCI Express Generation 3.0 slots, one external SAS port, and two USB 2.0 ports per Building Block. The external SAS port can be used to connect SAS tape or storage devices, and the USB 2.0 ports to connect a supported DVD device. In addition, external PCI Expansion

Units increase the number of available PCI Express Generation 3.0 slots.

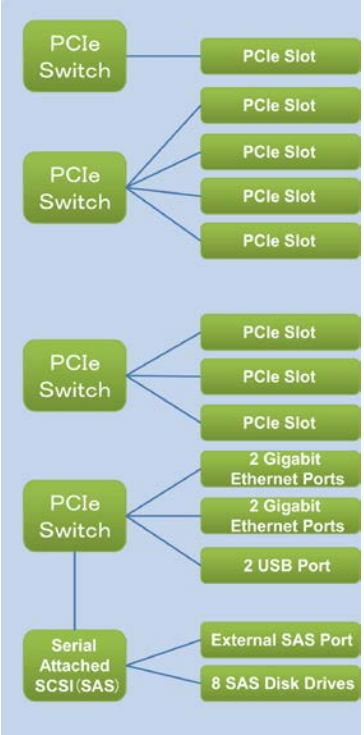


Figure 3-11. Fujitsu M10-4S I/O Subsystem Architecture

5. Internal Drives and Peripherals

The Fujitsu M10 systems support eight hot-swappable internal Serial Attached SCSI (SAS) 2.5-inch hard disk drives and 2.5-inch solid state drives (SSD). The Fujitsu M10 systems also provide one external SAS port which can be connected to SAS tape or storage device. The SAS port offers 4-lanes, supporting up to 24 Gb/second total bandwidth. The Fujitsu M10 systems also support two USB 2.0 ports that may be used to connect a supported DVD drive.

Fujitsu M10 server's on-board SAS controller supports RAID 0, 1 and 1E (striping, mirroring, and enhanced mirroring) volumes, and hot-swap drives using the LSI sas2ircu utility.

6. PCI Expansion Unit

The Fujitsu M10 systems support the attachment of optional, external PCI Expansion Units to provide additional I/O connectivity. The PCI Expansion Unit is a 2-RU rack mountable device which accommodates up to 11 PCI Express Generation 3.0 slots. By using PCI Express

cassettes, the external I/O chassis supports active replacement of hot-plug PCI Express cards.

A Link Card mounted in a PCI Express slot in the server provides connectivity to the PCI Expansion Unit and supports management control via sideband signals. The link card is a low profile card and includes an 8-lane PCI Express bus with 8 GB/second bandwidth. The architecture of the Fujitsu M10 PCI Expansion Unit provides high-throughput I/O performance, supporting maximum data rates for current and future Fibre Channel, InfiniBand, Gigabit Ethernet, and Flash PCI Express expansion cards.

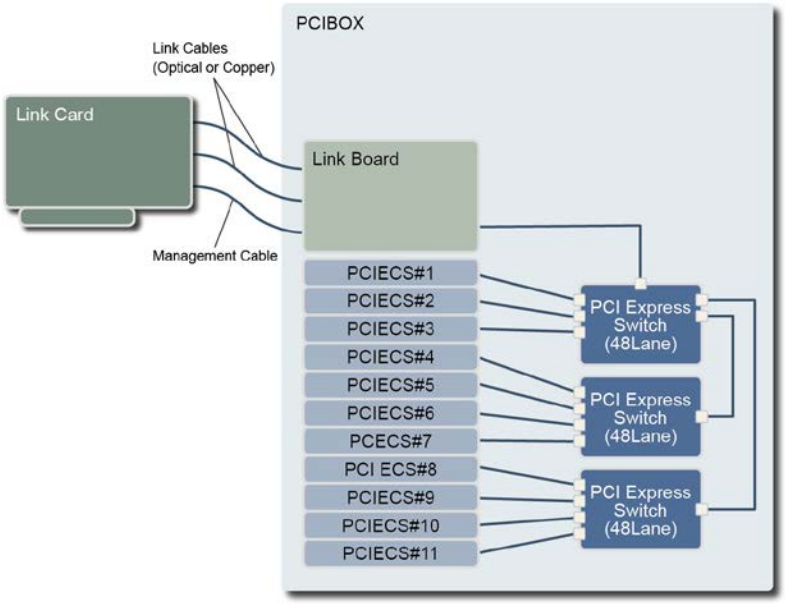


Figure 3-12. PCI Expansion Unit Architecture

PCI Expansion Units connect from the Link Board in the PCI Expansion Unit to the Link Card mounted in a PCI Express slot in the server using a cable. Cable options include a copper cable kit or a fibre cable kit. The Fujitsu M10 systems support the connection of multiple PCI Expansion Units as shown in Table 3-2.

Table 3-2. Maximum Number of PCI Expansion Units

	Maximum Number of PCI Expansion Units
Fujitsu M10-1	2
Fujitsu M10-4	3 (2CPU)
	6 (4CPU)
Fujitsu M10-4S (per Building Block Chassis)	3 (2CPU)
	5 (4CPU)

Cooling

As an entry-level system, the Fujitsu M10-1 uses traditional fans to provide air cooling. The Fujitsu M10-4 and Fujitsu M10-4S servers employ Liquid Loop Cooling (LLC), a new cooling technology. This new cooling method combines the advantages of liquid cooling technology (high cooling performance) with the advantages of air cooling technology (ease of installation and operation). LLC consists of the following three functional blocks.

1. Heat Sending Block: Piping and Pumps
2. Heat Receiving Block: Cooling Plate
3. Heat Dissipation Block: Radiator

Refrigerant is circulated in the LLC system by pumps in the heat sending block. In the heat receiving block the refrigerant absorbs heat dissipated from the cooling plate mounted to the CPU and power-supply components. The refrigerant flows through the piping of the heat sending block, and is sent into the radiator where it is cooled by forced air from fans mounted in the front of the chassis. The refrigerant returns through piping and the pumps back to the cooling plate.

By circulating the refrigerant entirely within the chassis, there is no need to supply liquid from outside or perform maintenance on the liquid cooling mechanism. LLC enhances the ease of server installation and serviceability.

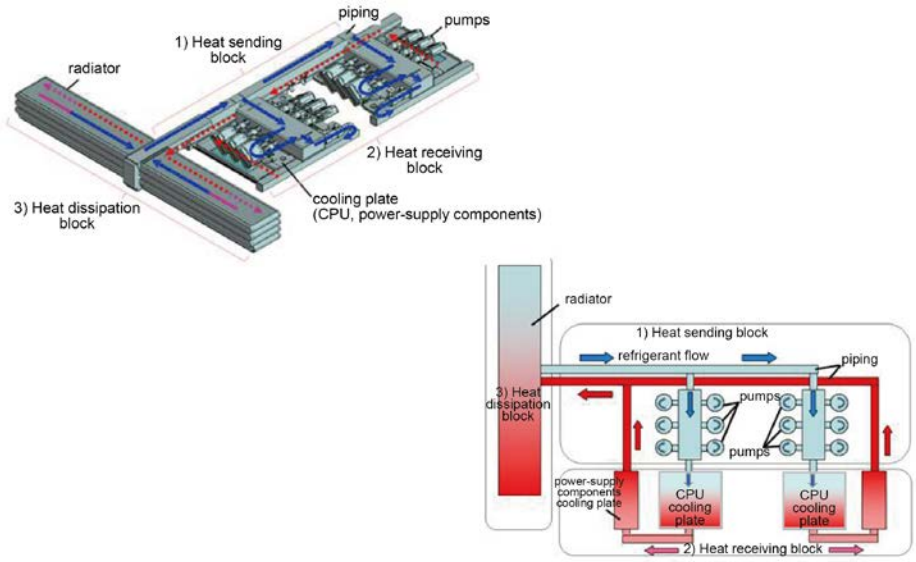


Figure 3-13. Liquid Loop Cooling

High-Efficiency Power Supply

The Fujitsu M10-1 server employs power supplies which comply with 80 PLUS? Gold, the standard concerning power conversion efficiency of power supply units for computers. Fujitsu M10-4 and Fujitsu M10-4S servers employ power supply units which comply with 80 PLUS? Platinum.

Fujitsu M10-4 and Fujitsu M10-4S power supply units feature the following three high-efficiency technologies:

1. High-efficiency components: SiC diodes and Super Junction FETs
   Traditional silicon-based MOS-FETs are commonly used in switching power supply circuits, MOS-FET output impedance continues to decrease year by year due to continuous technological development, but the material limits are approaching. New devices such as SiC diodes and Super Junction FETs have been proactively adopted in Fujitsu M10 power supply units to lower the power supply switch loss.
   High-efficiency circuits: Bridge-less rectifier and standby-less power circuits
2. In conventional power supply units, bridge diodes are often used for rectification in the power input section. To eliminate power loss from the bridge diodes, Fujitsu M10 power supply units have adopted bridge-less rectifier circuits which control rectifier and PFC (Power Factor Correction) circuits in an integrated fashion. In addition, Fujitsu's unique power technology has achieved a true standby-less power feed circuit which no longer

uses low-efficiency standby power supplies. Instead, power is sourced solely from the main power supply.

3. High-efficiency implementation: Reducing output impedance through a bus bar structure

   Since large amounts of power flow through the secondary side output section of the bus bar structure, direct-current resistance reduction is the key to enhanced efficiency. Fujitsu M10 power supply units have adopted bus bars that use copper in the connection between power components on the secondary side, which has reduced the direct-current resistance.

4. System Management

Reliability, Availability, and Serviceability

Reducing downtime - both planned and unplanned - is critical for IT services. System designs must include mechanisms that foster fault resilience, quick repair, and even rapid expansion, without impacting the availability of key services. Specifically designed to support complex, network computing solutions and stringent high-availability requirements, the Fujitsu M10 systems include redundant and hot-swappable system components, diagnostic and error recovery features throughout the design, and built-in remote management features. The advanced architecture of these reliable servers delivers high levels of application availability and rapid recovery from many types of hardware faults, simplifying system operation and lowering costs for enterprises.

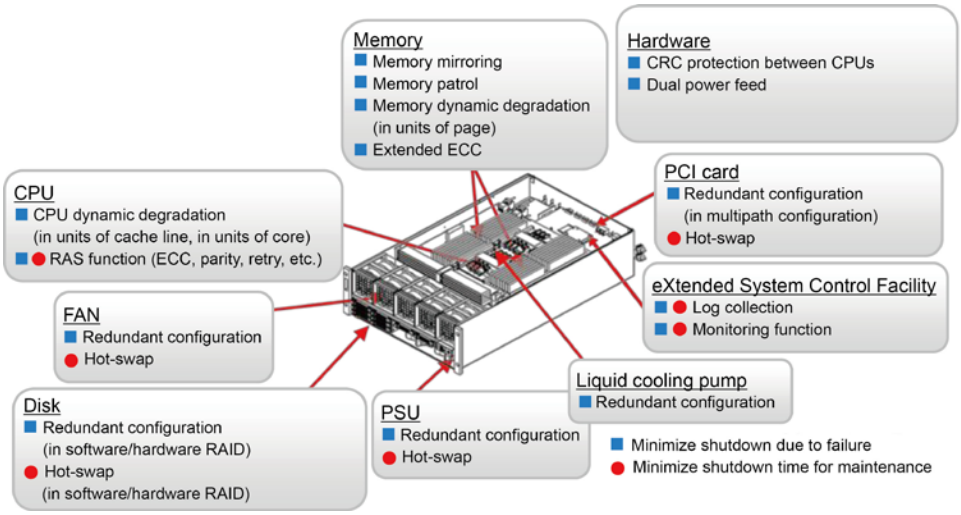


Figure 4-1. Reliability, Availability, and Serviceability

1. Redundant and Hot-Swappable Components

Today's IT organizations are challenged by the pace of non-stop business operations. This is forcing planned downtime windows to shrink, and in some cases disappear entirely. To meet these demands, the Fujitsu M10 systems employ built-in redundant and hot-swappable

hardware to help mitigate the disruptions caused by individual component failures or changes to system configurations. In fact, these systems are able to recover from hardware failures often with no impact to users or system functionality.

All Fujitsu M10 systems feature redundant, hot-swappable power supply and fan units. Fujitsu M10-4 and M10-4S servers also feature hot-swappable I/O cards. Administrators can choose to create redundant internal storage by combining hot-swappable disk drives in all Fujitsu M10 systems with either hardware RAID or disk mirroring software. Furthermore, with the Fujitsu M10-4S server, active maintenance of Building Blocks allows the CPU or memory configuration to be changed while keeping other physical partitions in operation. In multi-Building Block configurations, the Fujitsu M10-4S server features redundant service processors. The Fujitsu M10-4S server also includes degradable crossbar units. If a fault occurs, these duplicated components can support continued operation. Depending upon the component and type of error, the system can continue to operate in a degraded mode, or the system may reboot - with the failure automatically diagnosed and the relevant component automatically configured out of the system. Hot-swappable hardware within the Fujitsu M10 systems speeds service and allows for simplified replacement or addition of components, without the need to stop other physical partitions.

2. Partitioning Feature

In order to reduce costs and administrative burden, many enterprises look to server consolidation. However, organizations require tools that increase the security and effectiveness of hosting multiple applications on a single server. On the Fujitsu M10-4S server, by connecting the high-speed interconnect of multiple enclosures using the Building Block method, physical partitions can be created in units of enclosures (Building Blocks). When a physical partition is configured across multiple Building Blocks, all of the hardware resources from those Building Blocks are available to the physical partition. The physical partitioning feature provides IT organizations with the ability to construct up to 16 independent physical partitions (with 16 Building Blocks), a single large system spanning all Building Blocks, or any configuration in between. Oracle Solaris operates in each physical partition independently. Oracle VM Server for SPARC allows for the construction of multiple independent logical domains on top of the physical partitions. Oracle Solaris operates in each logical domain independently. With proper configuration, software faults in one partition or domain remain isolated and unable to impact the operation of other partitions or domains.

The Dynamic Reconfiguration (DR) feature allows Fujitsu M10-4S to expand hardware resource, such as CPU's and memory without interrupting Operating System by adding building block. The DR feature also allows moving hardware resource from one partition to the other.

By using the DR feature, great flexibility and availability can be achieved on Fujitsu M10-4S. Some of expected used cases are as follows:

- To expand hardware resource without interruption of production, when workload is increased,
- To re-assign hardware resource based on plan without interruption of production. (Ex. Move CPU's and memory to a partition, where database software for sales management, from other partition at the end of each quarter)
- Disconnect, replace and connect broken building block without interruption of production in order to repair faulted CPU/memory

3. Advanced Reliability Features

Advanced reliability features included within the components of the Fujitsu M10-4S server increase the overall stability of the platform. For example, the Fujitsu M10-4S server includes multiple degradable crossbar switches that provide redundancy within the system bus. By virtue of the consolidated peripheral ASIC functions in the processor, reduced component count and complexity within the server architecture contribute to reliability. In addition, CRC and ECC protection mechanisms guarantee data path integrity in SPARC64™ X/SPARC64™ X+ processors and provide for autonomous error recovery, reducing the time to initiate corrective action and subsequently increasing uptime. The large-capacity data path uses Fujitsu's high-speed interconnect technology to provide advanced error detection and correction features.

Oracle Solaris Predictive Self-Healing software further enhances the reliability of the Fujitsu M10 systems. Implementation of Oracle Solaris Predictive Self-Healing software provides constant monitoring of CPUs and memory. Depending on the nature of the error, persistent soft CPU errors can be resolved by automatically off-lining a single thread, one core, or even all cores in a processor. In addition, the memory page retirement function provides the ability to take memory pages offline proactively in response to multiple correctable errors in a specific memory DIMM.

4. Error Detection, Diagnosis, and Recovery

The Fujitsu M10 systems feature important technologies that correct failures early and keep marginal components from causing repeated downtime. Architectural advances which inherently increase reliability are augmented by error detection and recovery capabilities within the server hardware subsystems, as described below.

- End-to-end data protection detects and corrects errors throughout the system, ensuring complete data integrity.

- State-of-the-art fault isolation helps Fujitsu M10 servers isolate errors within component boundaries and offline only the relevant device segment instead of the entire component. Isolating errors down to the chip or down to the block in the chip improves stability and provides continued availability of maximum compute power. This feature applies to CPUs, memory, memory access controllers in the CPU, crossbar ASICs, and PCI-Express I/O controllers in the CPU.
- Constant environmental monitoring provides a historical log of all pertinent environmental and error conditions.
- The host watchdog feature regularly checks the operation of the processors to ensure operating system and application operation is proceeding normally. In the event that a watchdog timeout occurs, recovery functions are automatically triggered.
- Periodic component status checks are performed to determine the status of many system devices and detect signs of an impending fault. Recovery mechanisms are triggered to prevent system and application failure.
- Error logging, multistage alerts, electronic FRU identification information, and system fault LED indicators contribute to rapid problem resolution.

System Management

Providing hands-on, local system administration for server systems is no longer realistic for most organizations. Around the clock system operation, disaster recovery hot sites, and geographically dispersed organizations lead to a requirement for remote management of systems. One of the many benefits of Fujitsu servers is the support for lights-out datacenters, letting expensive support staff work in any location with network access. The design of the Fujitsu M10 systems combine with a powerful eXtended System Control Facility (XSCF) and XSCF Control Package to help administrators remotely execute and control nearly any task that does not involve physical access to hardware. These remote functions lower administrative burden, saving organizations time and reducing operational expenses.

1. eXtended System Control Facility

The eXtended System Control Facility (XSCF) is the heart of remote monitoring and management capabilities in Fujitsu M10 systems. The XSCF consists of a dedicated processor that is independent of the server system and runs the XSCF Control Package.

The XSCF regularly monitors environmental sensors throughout the system, provides advance warning of potential error conditions, and executes proactive system maintenance procedures as necessary. For example, the XSCF will initiate a server shutdown in response to temperature conditions which might induce physical system damage. The XSCF Control Package running

on the service processor helps administrators to remotely control and monitor partitions and domains, as well as the platform itself.

Using a network or serial connection to the XSCF, administrators can effectively administer the server from anywhere on the network. Remote connections to the service processor run separately from the operating system and provide the full control and authority of a system console.

2. Redundant XSCF

On Fujitsu M10-4S systems configured with more than one Building Block, one XSCF is configured as active and the other is configured as a standby. The XSCF network between the two service processors facilitates the exchange of system management information. In case of an XSCF failure, the service processors are already synchronized and ready to failover and change roles.

3. XSCF Control Package

The XSCF Control Package helps users to control and monitor Fujitsu M10 systems and individual partitioning functions quickly and effectively. The XSCF Control Package provides a command line interface (CLI) and Web browser user interface that give administrators and operators access to system controller functionality. Secure accounts with specific administration capabilities also provide system security for partition and primary logical domain consoles. Communication with the XSCF supports encrypted connections based on Secure Shell (SSH) and Secure Socket Layer (SSL), supporting secure, remote execution of XSCF Control Package commands.

The XSCF Control Package provides the interface for the following key server functions.

1. Physical partition administration including the assignment of Building Blocks to physical partitions.
2. Audit administration includes the logging of interactions between the XSCF and partitions and primary logical domains.
3. Monitoring and control of power to components in the server.
4. Interpretation of presented hardware information and notification of impending problems such as high temperatures or power supply problems.
5. Integration with the Oracle Solaris Fault Management Architecture to improve availability through accurate fault diagnosis and predictive fault analysis.

6. Execution and monitoring of diagnostic programs, such as the OpenBoot PROM (OBP) and power-on self-test (POST).
7. Execution of CPU Activation operations which provide the ability to stage and then later activate additional resources.
8. Monitoring of dual XSCF configurations for failures, and performing automatic failover if needed.

4. Role-Based System Management

The XSCF Control Package facilitates the independent administration of multiple autonomous physical partitions by different system administrators and operators - each utilizing portions of a single Fujitsu M10 system platform. This management software supports multiple user accounts which are organized into groups. Different privileges are assigned to each group. Privileges allow a user to perform a specific set of actions on a specific set of hardware, including physical components, physical partitions, or physical components within a physical partition. In addition, a single user can possess multiple, differing privileges on any number of physical partitions

5. Oracle Enterprise Manager Ops Center 12c

Oracle Enterprise Manager Ops Center 12c is a system management software that can manage a number of virtual and physical servers on Fujitsu M10. Ops Center can reduce cost of system management. These operations, adding hardware resources (CPU, Memory and so on), provisioning physical and virtual servers, updating system firmware and Oracle Solaris 11 patches can be done by Ops Center easily.

Eco-Friendly Computing

The power-saving functions provided in Fujitsu M10 systems reduce the power being consumed by unused or low-utilization-rate hardware. Customers who prioritize lower power consumption can select the power-saving mode (Elastic mode), whereas those who prefer performance over power-saving can select the performance preferred mode (Performance mode). The power modes are set per physical partition.

The following Fujitsu M10 server features promote power-saving.

1. Lowered hardware component power consumption

   In designing Fujitsu M10 systems the selection of hardware components was made with considerable attention to lower power consumption.

2. Reduction of power consumption from unused hardware components

   Depending on the model, physical partition configuration, and logical domain configuration, some hardware mounted in the system is unused. For example, if a CPU is not assigned to a partition or domain, the CPU's power state will automatically be lowered to save power. The same also occurs for unassigned memory.

3. Reduced power consumption from low-utilization-rate hardware

   Depending on the physical partition configuration, some controllers inside the processor are not used. Fujitsu M10 can reduce the system clock or turn on a power-saving mode for these components. The CPU core frequency can be automatically raised or lowered according to the utilization rate. The memory access controllers can also be controlled automatically according to utilization to enable or disable low-power levels such as pre-charge power-down or self-refresh. When CPU utilization rate is below a certain threshold, the number of CPUs assigned to tasks can be reduced to save power.

4. Sensor monitoring function

   Using this function, power consumption and air flow are monitored and logged. By collecting and making actual power consumption data available, data center power capacity design can be optimized. In a similar fashion, actual air flow data allows datacenter cooling facilities to be optimized.

5. Power capping function

   The power capping function allows the customer to set an upper threshold for system power consumption. The CPU frequency is automatically controlled so as not to exceed the threshold. This function provides the customer with control of system power consumption to fit the datacenter facilities.

5. Oracle Solaris 11 Operating System

Oracle Solaris is the industry-standard UNIX operating system with superiority in performance, scalability, reliability and security.

The design of Oracle Solaris is suitable for high-reliability systems. It consists of a small and compact kernel, which lessens the likelihood of operating system failure and resulting platform downtime. Furthermore, Oracle Solaris clearly distinguishes between the kernel, shared libraries, and applications; which limits the impact due of application failures. In addition, Oracle Solaris allows enterprises or organizations to install sequentially updated software without rebooting, which increases system uptime and lightens the maintenance load.

Oracle Solaris 11 offers cutting-edge new features such as IPS (Image Packaging System), BE (Boot Environments), security enhancements, network virtualization, and Oracle Solaris 10 Zones. It has achieved increases in system installation and operating efficiency, the ability to construct safe and flexible virtual environments, and robust investment protection; all of which offer a strong technological foundation for cloud computing in mission-critical systems.

Oracle Solaris ZFS

Oracle Solaris ZFS is included by default to provide the storage virtualization function. Oracle Solaris ZFS manages multiple physical storage devices in a storage pool. By allocating required capacity from the storage pool, a virtualized volume can be created.

Oracle Solaris ZFS enables more efficient and optimized use of storage devices, while dramatically increasing reliability and scalability. Physical storage can be dynamically added or removed from storage pools without interrupting services, providing new levels of flexibility, availability, and performance.

Oracle Solaris ZFS protects all data with 256-bit checksums, resulting in 99.99999999999999999-percent error detection and correction. If ZFS detects an error in a storage pool with redundancy, the corrupt data is automatically repaired. This contributes to relentless availability by protecting against costly and time-consuming data loss due to hardware or software failures, and by reducing the impact of administrator errors when performing file system-related tasks.

Oracle Solaris ZFS is a 128-bit file system, which can manage an infinite capacity in practical use. Since the metadata used to manage Oracle Solaris ZFS is dynamically allocated as needed,

there are no limitations on the number of file systems and the number of files. In conventional file systems, the size of the file system was limited to the size of the physical device. However with Oracle Solaris ZFS, storage pools conceal the physical devices; and therefore, size is not limited by a specific physical device. Oracle Solaris ZFS can easily create file system layers without initialization, and automatically extends capacity as disks are allocated to the ZFS storage pool.

The Oracle Solaris ZFS storage pool is a framework which consolidates the management of physical disks. Non-redundant, mirrored, RAID-Z (single parity), RAID-Z2 (double parity), or RAID-Z3 (triple parity) redundant configurations can be selected. All data written to the storage pool is dynamically striped across all available devices. In addition, in a redundant storage pool configuration, when an illegal data block is detected, the correct data is obtained from another redundant copy for self-recovery.

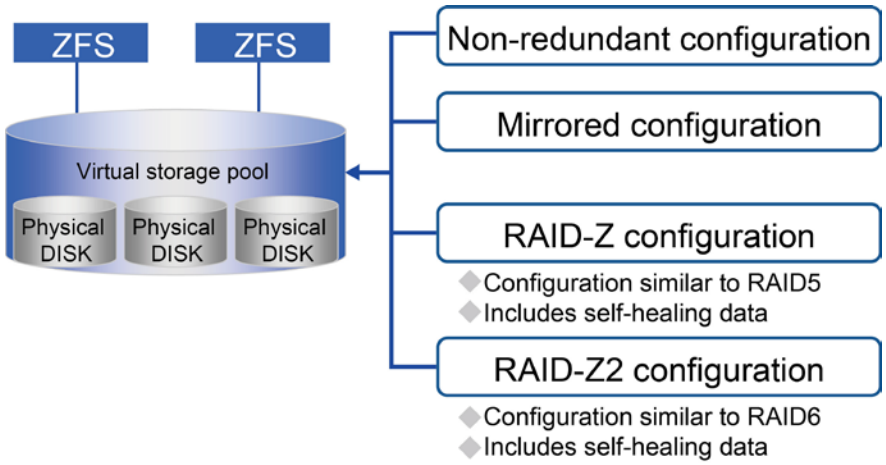


Figure 5-1. ZFS Storage Pool

ZFS deduplication allows for the sharing of duplicated data blocks in the ZFS storage pool, which reduces the amount of data stored and achieves more efficient use of storage. This deduplication function can be used in combination with compression and the cryptographic functionality.

Since the storage pool allows the disk addition and replacement online, it is not necessary to secure disk capacity for future use at the time of initial installation.

To achieve significant performance improvements and reduced power consumption, the ZFS hybrid storage pool has been developed. This combines memory, SSDs (Solid State Drives) and disks together in a single storage pool. The ZFS Intent Log (ZIL) regularly uses the storage

pool, and can deliver performance improvements for synchronized data writes by being allocated to high-speed devices such as SSD. The ZFS cache device (L2ARC: Level 2 Adaptive Replacement Cache) can improve the performance of random reads of static data by adding a high-speed device such as SSD as the cache between memory and disks.

Oracle Solaris ZFS is a transaction file system. Writing of data does not overwrite the existing data but copies the original data and updates it (Copy-on-Write). After completed a series of data updates, ZFS switches the data pointer for the old and new. This constantly keeps consistency in the file systems. Even with a sudden power loss, the file system will never be destroyed.

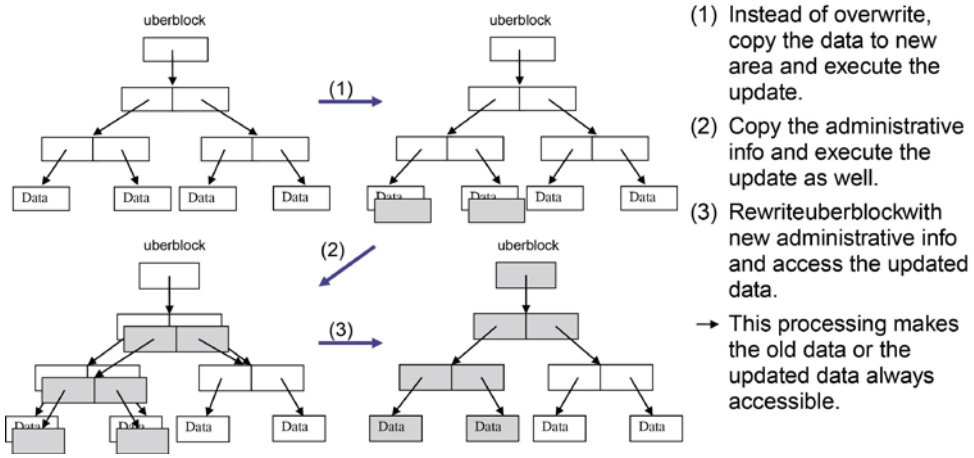


Figure 5-2. Copy-on-Write

When a file system is created, Oracle Solaris ZFS automatically generates the mount point. Since it is automatically mounted at server start, mount management is not required. When a ZFS volume is created as a block device, the initial data area size is reserved for the volume and is automatically extended as needed.

Oracle Solaris ZFS has end-to-end checksumming for metadata of all user data and administrative information. Since the checksum of the data block is retained in the parent block and iterates up to the top administrative block (uberblock) in order, self-inspection of the entire data tree is possible. When an error is detected, Oracle Solaris ZFS recovers the data from the redundant copy. In order to enhance reliability, the Oracle Solaris ZFS metadata is automatically saved across different disks (ditto blocks).

Furthermore, Oracle Solaris ZFS can save multiple copies of user data. Even if the data cannot be redundant across multiple disks, it can recover from disk block read failures.

ZFS snapshot is a read-only copy of the file system or the volume, which can be instantly created without consuming disk capacity. Although the snapshot cannot be directly referred to,

it facilitates operations such as clone and backup. ZFS clone is a writable copy of the file system or volume, which can be created from the snapshot. Just like the snapshot, it can be created instantly without consuming significant disk capacity. The clone only requires enough disk capacity to store the changed data.

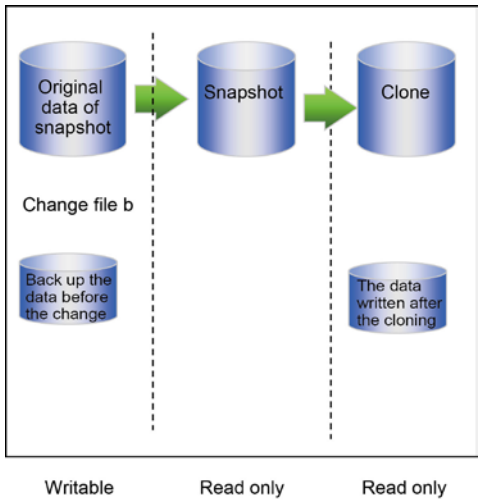


Figure 5-3. ZFS Snapshot

To protect against theft of physical storage or outside attack, as well as achieving secure deletion within the storage, data can be encrypted by each ZFS data set.

Oracle VM Server for SPARC

Oracle VM Server for SPARC can divide the physical server into virtual servers using a firmware layer Hypervisor to configure logical domains in which independent Oracle Solaris environments operate. CPU, memory, and I/O devices are flexibly allocated by the Domain Manager which can run on either Oracle Solaris 11 or 10.

The logical domain concept provides virtual servers with logically grouped CPUs, memory, and I/O devices. In each logical domain an independent Solaris OS operates. Each logical domain can be started or stopped individually, and up to 256 domains per physical partition can be created. Logical domains communicate with each other via the Hypervisor's Logical Domain Channel (LDC). Virtual devices such as disks and networks communicate with the virtual services through LDC, to access to the physical devices.

Resources such as virtual CPUs or memory can be dynamically reconfigured while the logical domain is in operation. Furthermore, the number of virtual CPUs in a logical domain can automatically be increased or decreased in accordance with the dynamic resource management

policy. This policy can be created by combining conditions such as the number of resources available, usage rate, upper and lower thresholds, and time period. In addition, CPU power management features can reduce power consumption by shutting down the power to a CPU not being used to process the current workload.

In line with changes to workload or server expansion, a guest domain can be migrated to another physical server. Live migration allows for the high-speed transfer of compressed memory content, without suspending guest domain operation.

To consolidate servers, a physical server can also be migrated to a logical domain using the Physical-to-Virtual (P2V) migration tool. This collects configuration information from a physical server and creates a file system image. Then P2V creates the logical domain based on the collected configuration information and restores the file system image to the virtual disk.

Oracle Solaris Zones

The Oracle Solaris Zones function virtually divides a single OS space and presents multiple OS spaces to the user. Oracle Solaris Zones also include the Oracle Solaris Resource Manager function which flexibly allocates hardware resources such as CPU and memory.

Oracle Solaris Zones is a virtualized OS environment that provides a safe and isolated environment suitable for application execution. The process which is executed is isolated by each zone, and does not affect other zones.

The global zone is a single zone that exists in the Oracle Solaris system. The global zone manages the entire system. Operations such as creating and managing non-global zones or allocating physical I/O device can be performed solely in the global zone.

Non-global zones are the software partitions of the virtual Solaris environment, where the applications can be executed without affecting other zones. Up to 8191 non-global zones can be created. Each zone can only use designated file system(s) and physical I/O device(s).

For system files which are used to configure the non-global zones, necessary packages can be selected and installed when the zone is created. When packages are updated in the global zone, these files are updated and all non-global zone files are also updated in synchronization.

Oracle Solaris Resource Manager periodically monitors resource usage and automatically allocates additional resources, configured for such cases, without stopping the zone. CPU resources to be allocated to the zone are managed by the resource pool. The resource upper limit daemon controls memory. Furthermore, the resource pool consists of CPUs grouped into processor sets and CPU allocating scheduling classes (in time sharing and fair allocation).

By copying an existing zone, new zones can be created easily. By using ZFS clone, a zone can

be replicated instantly, and initially does not occupy additional disk capacity.

When resources such as CPU or memory are insufficient in a physical server or if the fundamental organization or usage of zones needs to be changed, a zone can be migrated to another server. The zone is detached from the original server and attached to the destination server. Even if there are differences in the environment between the servers such as package configuration, the system file used to configure the zone will be synchronized with the global zone of the destination server at the time of attachment.

Oracle Solaris 11 provides support for Oracle Solaris 10 Zones by default.

By using the P2V (Physical-To-Virtual) function or the V2V (Virtual-To-Virtual) function, an existing Oracle Solaris 10 OS environment can be migrated onto a system with Oracle Solaris 11 OS Zones as it is. This allows for the consolidation of Oracle Solaris 10 and Oracle Solaris 11 environments, which contributes to investment protection and TCO reduction.

By virtualizing NICs (Network Interface Controllers) using the virtual network function, VNICs (Virtual Network Interface Controllers) can be allocated to multiple Oracle Solaris Zones. An independent network environment consisting of multiple Oracle Solaris Zones can be configured, which can reduce the number of required servers and NICs. In addition, since a virtual switch (ethers tab) is created, a virtual network environment which does not depend on the hardware can be achieved. The number of switches can be reduced as well. Furthermore, VNIC bandwidth limits can be configured using the resource management functions.

Image Packaging System

IPS (Image Packaging System) is a new framework which allows the management, installation, update, and deletion of the OS environment in units of packages.

IPS installs an OS environment that is optimized for basic server operation from the media and then installs additional packages via network that are required for the specific intended workload. Administrators are not required to independently prepare a network installation server but can use a repository server. IPS automatically navigates through the complicated dependencies of packages. Conventional patch application is now handled by package replacement. The replacement is done without regard to package dependencies and can avoid application failure.

As stated above, IPS aims at improving the installation and operation management efficiency, which leads directly to cost reductions.

Boot Environments

BE (Boot Environments) is a function which enables the management of multiple boot environments and simplifies online upgrade.

The root file system of Oracle Solaris 11 OS is Oracle Solaris ZFS. To add or update packages, ZFS snapshot and ZFS clone enable boot environment duplication in a short period of time. The ZFS tools copy only the data blocks which will be added or updated, reducing the required disk capacity. As well as reboot from the duplicated boot environment, it is also possible to restore to the original boot environment if any problems occur during the package update.

BE utilization enables generation management of boot environments, which drastically shortens down-time for maintenance work.

6. Technological Enhancements of the SPARC64™ X and SPARC64™ X+ Processor

Microarchitecture

1. Chip Configuration

SPARC64™ X/SPARC64™ X+ as a semiconductor adopt the 28 nm CMOS process. The chip has up to 16 cores and up to a 24 MB of 24-way shared L2 cache. SPARC64™ X has an operating frequency of up to 3 GHz while SPARC64™ X+ has up to 3.7 GHz. (See Table 6-1.)

Table 6-1. SPARC64™ X+ Specifications

	SPARC64™ X	SPARC64™ X+
Number of Cores	16	16
L2 Cache	Up to 24 MB	Up to 24 MB
Operating Frequency	Up to 3 GHz	Up to 3.7 GHz
Process Technology	28 nm CMOS	28 nm CMOS
Die Size	23.5 mm x 25.0 mm	24.0 mm x 25.0 mm
Number of Transistors	Approx. 2.95 billions	Approx. 2.99 billions
Memory Bandwidth	102 GB/s (theoretical peak value)	102 GB/s (theoretical peak value)

For lower memory access latency and higher throughput, SPARC64™ X/SPARC64™ X+ are equipped with an internal memory controller. Memory bandwidth has a theoretical peak of 102 GB/s.

In addition, SPARC64™ X/SPARC64™ X+ are equipped with an internal CPU-to-CPU interface to interconnect the multiple CPUs found within the Fujitsu M10-4 and Fujitsu M10-4S Building Block chassis. This CPU-to-CPU interconnect uses a high-speed serial interface to

deliver high throughput when processing spans multiple CPU sockets.

To complete System-On-Chip features of SPARC64™ X/SPARC64™ X+, the I/O controller has also been integrated into the silicon, and provides two 8-lane PCI Express Generation 3.0, 8 Gbps ports per CPU socket.

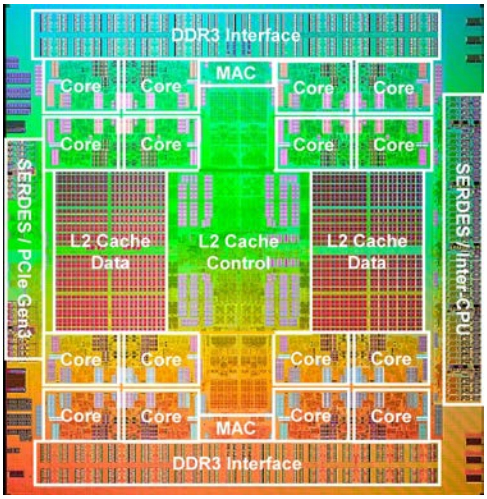


Figure 6-1. SPARC64™ X+ Block Diagram

2. Core Microarchitecture

(1) SPARC64™ X/SPARC64™ X+ Core

The SPARC64™ X core/SPARC64™ X+ core consists of an instruction fetch block and instruction execution block, as shown in Figure 6-2. The instruction fetch block contains the L1 instruction cache, and the instruction execution block contains the L1 data cache for operand, execution unit, registers, etc.

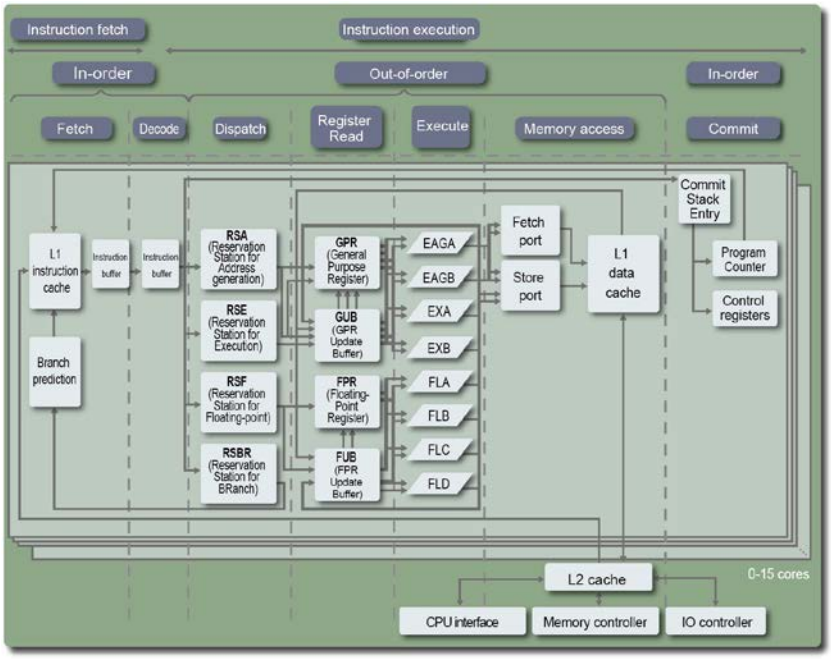


Figure 6-2. SPARC64™ X/SPARC64™ X+ Pipeline

(2) Simultaneous Multithreading (SMT)

SPARC64™ X/SPARC64™ X+ carries over the simultaneous multithreading (SMT) technology first deployed in SPARC64™ VII+ . With this technology multiple threads in a single core are simultaneously processed. From the software point of view, each thread appears as an independent CPU. Hardware shares the execution resources such as instruction buffers, reservation station, pipeline, and caches between the threads. Even when one thread is stalled in a wait for data, another thread continues processing by using the pipeline; and thus the processing performance as a core improves. Also, if one of the threads is idle, it is possible for another thread to utilize all of the execution resources and perform processing.

The following sections respectively explain the instruction fetch block and the instruction execution block.

(3) Instruction Fetch

The instruction fetch block reads out a series of instructions from memory or from the L1 instruction cache according to branch prediction. The L1 instruction cache is a 64KB 4-way configuration. For this L1 instruction cache, instruction fetch can start in every cycle, and eight instructions are fetched in parallel. The series of instructions which are read out is taken into the

instruction buffer all at once. The instruction buffer has a capacity of 256 bytes, and can store up to 64 instructions. When both threads are running, the instruction buffer is divided evenly for each thread. When the instruction buffer becomes full or when an L1 instruction cache miss occurs, the processor keeps working by executing hardware prefetch to load instructions from L2 cache or memory to L1 cache.

Instruction fetch operates independently of the instruction execution block. In the event that the instruction execution is stalled, instruction fetch continues as long as there are available instruction buffers. In contrast, with an event such as a cache miss, instruction supply from the instruction buffer to the instruction execution block continues as long as the instruction buffer includes instructions.

While the throughput of instruction execution is four instructions per cycle (a maximum of six instructions when including the prefix instruction SXAR described later), the throughput of instruction fetch is eight instructions per cycle, surpassing that of instruction execution. This helps improve system performance by concealing the access latency of the L1 instruction cache.

When using SMT, instruction fetch of a single thread is performed in the same cycle, and the threads are alternatively switched in each cycle.

Furthermore, SPARC64™ X/SPARC64™ X+ has enhanced the branch prediction mechanism. When making branch predictions, in addition to the branch history of its own instructions, the taken or not-taken history pattern from the last series of instructions are taken into consideration; and thus prediction accuracy improves.

(4) Instruction Execution

Instructions from the instruction buffer in the instruction fetch block are supplied to the instruction execution block by up to six instructions per cycle (two of which are the prefix instruction SXAR described later). Within the instruction execution block, these instructions are decoded, issued, executed, and committed.

- Instruction Decode and Issue

In the instruction decode and issue stage, when the SXAR prefix instruction is included, up to six instructions (two of which are SXAR) are decoded simultaneously. The prefix instruction SXAR is a new instruction. SXAR has the extension information of the two subsequent instructions and performs instruction extension of up to two of the subsequent instructions in the instruction decode stage. Furthermore, it performs "connection processing" with the subsequent instructions, to avoid the consumption of execution resources in the pipeline stage later on. After the following processes are performed, instructions extended by SXAR are treated as so-called "valid instructions":

- An instruction, which follows the SXAR prefix instruction for instruction extension, performs "connection processing".
- The next instruction after the above instruction, if SXAR shows it should be extended, also performs "connection processing".

In the instruction decode stage up to four valid instructions can be decoded simultaneously.

When using SMT, instruction decode of one of the two threads is performed in the same cycle, and threads are alternatively switched in each cycle.

In the instruction decode stage, resources required for execution - such as various reservation stations, fetch port and store port, and register update buffer - are determined. If the required resources can be allocated, up to 96 valid instructions can be issued. When using SMT, the maximum number of valid instructions for each thread is 48. In SPARC64™ X/SPARC64™ X+, in order to speed up the instruction processing, the resources required for execution such as reservation stations have been enhanced compared to the former SPARC64™ VII+ processor.

When performing instruction decode simultaneously, there are no restrictions on instruction type combinations. As long as there are free resources, instructions can be issued. If insufficient space exists for four instructions, as many instructions as possible are issued according to the instruction order in the program. As described above, by eliminating stall conditions of issued instructions as much as possible, a high multiplicity level is assured for any binary code.

- Instruction Execution

The SPARC64™ X/SPARC64™ X+ processor decodes instructions, and then they are registered in a reservation station. Among the fixed-point instructions, the addition and subtraction operations and the logical operations can be registered to RSE (Reservation Stations for Execution) or RSA (Reservation Stations for Address generation). Other fixed-point instructions are registered to RSE only. In addition, the address calculations of load and store instructions are registered to RSA. That is, RSA is shared by the addition and subtract operations, the logical operation of the fixed-point instructions, and the address calculation of load and store instructions. The floating-point arithmetic instructions are registered to RSF (Reservation Stations for Floating-point) and the branch instructions are registered to RSBR (Reservation Stations for BRanch).

RSE starts the arithmetic pipelines of EXA and EXB. RSA starts the pipelines of EAGA and EAGB. RSF starts the pipelines of FLA, FLB, FLC and FLD. Each instruction stored in a reservation station is dispatched to the execution unit that corresponds to that reservation station out-of-order. This prioritizes the older instruction in program order from among those in which input operands are prepared for instructions.

When using SMT, pipelines can be used by multiple threads simultaneously.

The execution block has two fixed-point arithmetic pipelines (EXA/B), two arithmetic pipelines (EAGA/B) which perform the address calculation of load and store or the addition and subtraction operations and logical operation of the fixed-point arithmetic pipelines, and four floating-point arithmetic pipelines (FLA/B/C/D).

Specifically, the floating-point arithmetic pipeline has been significantly extended compared to the former processor. It has adopted the SIMD (Single Instruction Multiple Data) function which was first introduced in the SPARC64™ VIIIfx supercomputer processor. With a single instruction, this function performs parallel processing using two pipelines (FLA and FLC, or FLB and FLD). Each pipeline has an FMA (Floating-point Multiply and Add) execution unit as before. A single FMA execution unit can execute floating-point multiplication and addition in every cycle. In each core, eight double-precision floating-point operations can be executed in every cycle.

As described later with SWoC (SoftWare on Chip) enhancements, the SPARC64™ X/SPARC64™ X+ core also includes new execution units to support cryptographic processing and IEEE754 for decimal floating-point numeric calculation in DPD (Densely Packed Decimal) and Oracle NUMBER formats. Also, an execution unit has been added to the core for the acceleration of business applications, such as databases, through the use of parallel processing.

In SPARC64™ X/SPARC64™ X+, the floating-point registers (FPR) have been increased by up to four times, which enhances instruction scheduling by software performance (such as loop unrolling or software pipelining).

By using the floating-point arithmetic pipelines for a wider variety of arithmetic operations and the four fold increase in registers, marked improvement of processing performance in various applications can be achieved.

The L1 data cache block processes the load and store instructions. Each core has a 64KB 4-way data cache. The data cache can provide the data for the subsequent load instruction without waiting for the address calculation of the store instruction which comes later in the series of instructions. Also, the L1 data cache is in a dual port configuration that is simultaneously accessible by two load instructions. Two 16-byte SIMD load instructions or a single 16-byte SIMD store instruction can be executed. As long as there are no bank conflicts, simultaneous processing of read-out by the load instruction and write by the store instruction is possible and improves the cache throughput.

- Instruction Commit

Data resulting from out-of-order executions is stored in different locations, depending on the type of data: GPR Update Buffer (GUB) for fixed-point data, FPR Update Buffer (FUB) for floating-point data, and Store Port for store data. Next, instruction commit is performed

according to the program order to update registers such as GPR (General Purpose Register) or FPR and memory.

The maximum number of valid instructions that can be committed at one time is four. SPARC64™ X/SPARC64™ X+ enables four simultaneous writes to GPR by the fixed-point arithmetic instruction or the fixed-point load instruction. The four fixed-point arithmetic and load pipelines deliver increased throughput to maximize performance.

When using SMT, an instruction commit by one of the two threads is performed in a single cycle, and threads switch on alternating cycles.

Before an instruction is committed, the execution result remains inaccessible from software. As stated above, control registers including GPR, FPR and PC (Program Counter) and memory are updated in the commit stage in order and at the same time according to the program order. By using this synchronous update method, precise interrupts are guaranteed, and in-progress processing can be cancelled at any time. This method enables the processor to implement instruction retry, which will be described later, and contributes to increased reliability.

3. Interface between Chips

SPARC64™ X/SPARC64™ X+ has an on-chip CPU interface function for the connection between CPU sockets. Up to four CPU sockets can be connected directly. Expanding beyond 4-sockets, up to 64-sockets, CPUs are connected via a crossbar (XB) chip. By virtue of the increased performance of a single CPU a fast, high-throughput serial transfer protocol has been implemented for the connection between CPU sockets. The transfer rate of a connection between CPUs is 14.5 GB/s for SPARC64™ X, and 25.0 GB/s for SPARC64™ X+.

Extended Instruction Set Architecture

SPARC64™ X/SPARC64™ X+ has adopted the virtual machine architecture which is fully compatible with the sun4v architecture supported by Oracle SPARC servers.

SPARC64™ X/SPARC64™ X+ has also introduced HPC-ACE (High Performance Computing Arithmetic Computational Extensions). This extension of the SPARC-V9 architecture instruction set was first introduced in the SPARC64 VIIIfx supercomputer processor. HPC-ACE enhances the SIMD (Single Instruction Multiple Data) function, enabling parallel processing of operations, and increases the number of floating-point registers (FPR).

New SWoC (SoftWare on Chip) functionality has also been added to SPARC64™ X/SPARC64™ X+. Using dedicated hardware, execution of cryptographic processing, which has conventionally relied on a combination of general instructions, can now be performed

significantly faster. Another significant enhancement found in SPARC64™ X/SPARC64™ X+ is the ability to perform IEEE754 decimal floating-point numeric calculations in DPD and Oracle NUMBER formats.

To address trends in business applications, such as database processing, instruction extensions to manage parallel processing of data have also been added to SPARC64™ X/SPARC64™ X+.

The following sections describe HPC-ACE, SWoC, and the acceleration of business applications through parallel data processing.

1. HPC-ACE

(1) Extension of Floating-Point Registers (FPR)

The number of floating-point registers (FPR) in SPARC-V9 is 32. This is not sufficient to fully exploit the performance of many applications. However, to increase the number of registers, the 32-bit fixed instruction length in the SPARC architecture falls short and is difficult to modify. To solve this problem, HPC-ACE has created a new prefix instruction called SXAR (Set eXtended Arithmetic Register). For up to two subsequent instructions, the SXAR instructions perform operations such as register address extension. SXAR has extended the register address with two additional bits, and the number of addressable floating-point registers (FPR) has been increased up to 128; four times as many as are defined in SPARC-V9 (see Figure 6-3). Compilers use this large-capacity register to perform, among other uses, software pipeline optimization to fully exploit instruction level parallelism in applications.

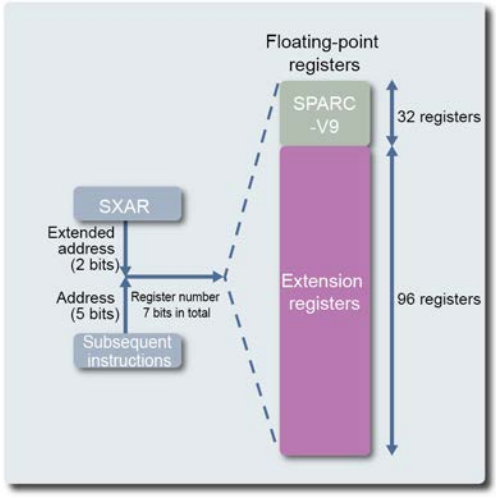


Figure 6-3. Register Address Extension by SXAR Instruction

(2) SIMD (Single Instruction Multiple Data)

The SIMD enhancement processes data using multiple pipelines in parallel, all with a single instruction. By adopting SIMD, HPC-ACE enables the use of two FMA (Floating-point Multiply and Add) execution units with a single arithmetic instruction.

In conventional scalar operation there is a 1:1 ratio of data to instruction. Execution of four add operation instructions are required to process four results. With SIMD functionality, up to 32 sets of 8-bit data can be compared simultaneously, or four sets of 64-bit data can be processed simultaneously with a single arithmetic instruction on SPARC64™ X+.

SPARC64™ X/SPARC64™ X+ leverages this SIMD enhancement to acceleratecryptographic processing and business applications, such as databases.

In addition, load and store instructions can use SIMD. For the load instruction especially, processing is done without any penalties even if it steps over the cache line.

By utilizing this function for massive data retrieval and data compression/decompression (encryption/decryption) loads, faster processing of massive data or in-memory databases is achieved.

2. SWoC (Software on Chip)

In SPARC64™ X/SPARC64™ X+, multiple instruction combinations can be replaced by dedicated hardware, leading to much faster computing. This is called "SWoC (Software on Chip) and accelerates processing as follows:

- Cryptographic Processing
  SPARC64™ X/SPARC64™ X+ implements two cryptographic arithmetic units per core, shared by each core's two threads.

  A dedicated cryptographic instruction executes cryptographic processing (encryption/decryption) at high-speed without using an add-on adapter. Supported encryption modes are AES, DES, 3DES, RSA and SHA.

  Since the encryption processing is executed by hardware, additional cost and performance deterioration are avoided. Also full database encryption allows the construction of a secure environment.

  In addition, OpenSSL and encrypt/decrypt commands are already compliant with the SPARC64™ X/SPARC64™ X+ cryptographic arithmetic processing functions. Standard libraries such as libpkcs11 also benefit from the cryptographic arithmetic processing enhancements.

- IEEE754 for Decimal Floating-Point Numeric Calculations in DPD-Format
  SPARC64™ X/SPARC64™ X+ inherits this high-speed computing technique from the mainframe.

  A decimal floating-point arithmetic unit has been added to execute decimal floating-point arithmetic processing directly by hardware at high speeds. Conventionally, this was processed by software.

  In a conventional decimal floating-point arithmetic unit, the decimal number data is converted to binary by software, and then arithmetic processing is performed by hardware. Computational results are then converted from binary back to decimal number data by software.

  In the SPARC64™ X/SPARC64™ X+ processor, arithmetic processing of the decimal number data can be performed as it is by hardware, without conversion to and then from binary. This SPARC64™ X/SPARC64™ X+ feature will accelerate computing across many industries such as distribution, manufacturing, finance, etc., by significantly accelerating common processing tasks like sales accounting, cost accounting, rebate accounting, and compound interest calculation.

  This SPARC64™ X/SPARC64™ X+ calculation feature complies with the standard for decimal floating-point arithmetic processing, IEEE754-2008.

- Compare and Copy Operations
  With SPARC64™ X/SPARC64™ X+ the width of the memory access bus has been increased. Extended SIMD instructions have also been implemented, allowing for multiple blocks of data to be loaded from memory, compared, and then stored to memory. Extended SIMD instructions allow for 100% utilization of the memory bus capacity, maximizing computation power of SPARC64™ X/SPARC64™ X+. In this way, memory processing functions such as memcpy(3C) in the standard libc library are accelerated automatically.

Processing which is conventionally executed by application software will now be processed by SPARC64™ X/SPARC64™ X+ hardware, leading directly to improved application performance. Oracle Solaris 11 has been optimized to take full advantage of the SWoC functions in SPARC64™ X. This translates to performance gain without requiring changes to the application.

Reliability, Availability, and Serviceability Features

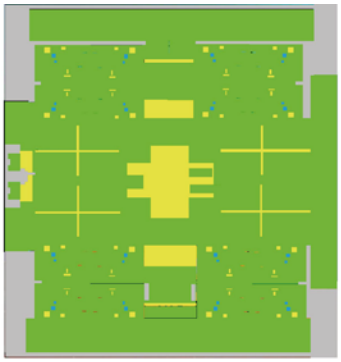
The SPARC64™ X/SPARC64™ X+ processor increases system reliability by delivering advanced error detection and correction capabilities. In fact, 99% of the SPARC64™

X/SPARC64™ X+ processor circuitry is protected by error detection and/or data correction mechanisms. RAM units are ECC protected or duplicated, and all 1-bit errors can be corrected. In addition, all latches and execution units are parity protected, and when the SPARC64™ X/SPARC64™ X+ processor detects a 1-bit error, it performs instruction retry. Cache is dynamically degraded either in units of cache-way or units of CPU core, and isolation after reboot is performed for some types of errors. Other reliability features of the SPARC64™ X/SPARC64™ X+ processor include support for error marking, instruction retry (re-execution) by hardware, and preventive maintenance, which will be mentioned later.

As described above, in SPARC64™ X/SPARC64™ X+, RAS functions comparable to mainframe computers have been implemented. With these RAS functions, errors are reliably detected, their impacts are kept within a limited range, recovery processing is attempted, error logs are recorded and notified to software, and so forth. Throughout SPARC64™ X and SPARC64™ X+ development, RAS functions are robustly implemented to provide high reliability, high availability, high serviceability, and high data integrity, making the processor a perfect match for mission-critical UNIX servers.

Table 6-2. Error Detection Mechanisms

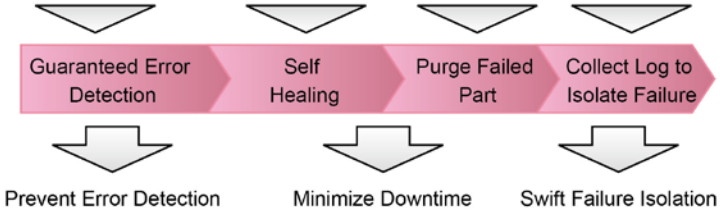
Unit	Error Detection and Correction Method
Cache (tag)	ECC Duplication + Parity
Cache (data)	ECC Parity
Register	ECC (integer/floating-point) Parity (others)
Operation Unit	Parity/Residue Check
Cache Dynamic Degradation	Implemented
Hardware Instruction Retry	Implemented
Processor History Logging	Implemented



Green: Error detection and self-healing available
Yellow: Error detection available
Gray: Error has no impact on actual operation

Figure 6-4. RAS Coverage Map

	Error Detection and Correction Scheme	Error Handling		Logging
		Correction	Deconfiguration	
L1 Cache	Duplication + Parity, ECC	Retry, ECC	Dynamic Way Deconfiguration*2	History
L2 Cache	ECC	ECC	Dynamic Way Deconfiguration*2	
ALU	Parity*1, ECC	ECC, Hardware Instruction Retry	Core Deconfiguration	
Register				



*1: The processor recovers from parity errors using hardware instruction retry.
*2: "Way" is a unit of cache.

Figure 6-5. Error Detection and Recovery Scheme

1. Error Marking

When data read from main memory or cache is found to contain a multi-bit error, a special value is written into the data by the error marking function to identify the location of the error. In combination with ECC syndrome reporting, the error marking function identifies the original

error location that caused the failure, thus aiding accurate degradation and parts replacement.

2. Internal RAM Reliability and Availability Features

The SPARC64™ X/SPARC64™ X+ processor offers reliability and availability features that support high levels of data integrity. Table 6-3 shows the error detection and correction capabilities of the SPARC64™ X/SPARC64™ X+ processor.

Table 6-3. Error Protection of SPARC64™ X/SPARC64™ X+ Internal RAM

RAM Type	Error Detection / Protection Method	Error Correction Method
L1 Instruction Cache Data	Parity	Invalidation and Reread
L1 Instruction Cache Tag	Parity + Duplication	Reread of Duplicated Data
L1 Data Cache Data	SECDED* ECC	1-bit Error Correction Using ECC
L1 Data Cache Tag	Parity + Duplication	Reread of Duplicated Data
L2 Cache Data	SECDED* ECC	1-bit Error Correction Using ECC
L2 Cache Tag	SECDED* ECC	1-bit Error Correction Using ECC
Instruction TLB	Parity	Invalidation
Data TLB	Parity	Invalidation
Branch History	Parity	Recovery from Branch Predication Failure

* SECDED: Single Error Correction Double Error Detect

For the L1 cache, L2 cache, and TLB dynamic degradation can be performed in units of ways. The SPARC64™ X/SPARC64™ X+ processor employs a set-associative scheme that divides the L1 cache, L2 cache, and TLB into way units. Error occurrence counts are tabulated for each way unit. When an error occurrence count exceeds an upper threshold, degradation is performed and the relevant way taken out of service. Hardware performs the dynamic degradation of the way, and software is unaffected except for a decrease in processing speed due to the degradation of the way.

3. Internal Registers and Execution Units Reliability Features

To further increase reliability, the SPARC64™ X/SPARC64™ X+ processor also provides error protection for registers and execution units. The general purpose register (GPR) and the floating-point register (FPR) have been enlarged and include ECC protection. When an error occurs, the ECC circuit corrects the error. Other registers are protected by parity. Parity prediction circuitry is implemented within execution units. In multiplication and division execution units a residue check circuit has been implemented to perform further error detection in output results. In the unlikely event that an error is detected, the hardware automatically re-executes the instruction to attempt recovery, as described below.

4. Synchronous Update Method and Instruction Retry

Similar to the previous processor generation, the SPARC64™ X/SPARC64™ X+ processor employs a synchronous update method at the time of commit. Only instructions that have been committed without detecting an error update the execution result in programmable resources such as GPR, FPR, other registers, memory, etc. When an error is detected, all the processing that occurred before the commit is canceled, preventing data integrity issues from occurring. After an error is detected and the processing cancelled, hardware can also re-execute the instruction. This is called instruction retry.

As shown in Figure 6-6, instruction retry is triggered by an error and starts automatically. The retry is performed instruction-by-instruction to increase the chance of successful execution. When the execution completes normally, the state automatically returns to the normal execution state. During this period, no software intervention is required. If the instruction retry succeeds the error has no effect on software. An instruction retry is repeated until the number of retry attempts reaches a threshold. If the threshold is exceeded, the processor logs the source of the error and notifies the operating system of the error occurrence for subsequent operating system processing.

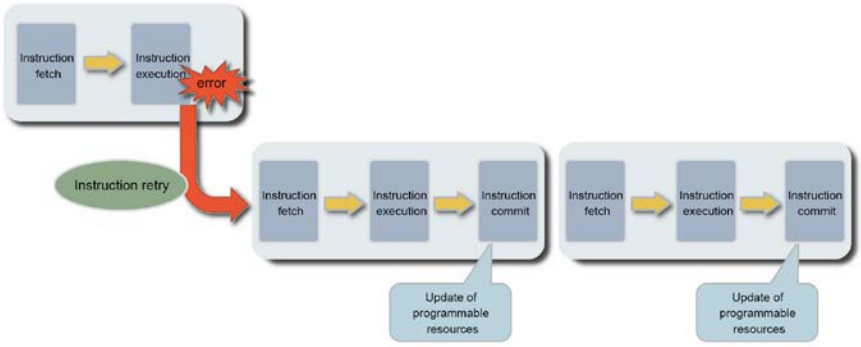


Figure 6-6. Instruction Retry

5. Increased Serviceability

The SPARC64™ X/SPARC64™ X+ processor, as stated above, provides a variety of error checking mechanisms. The processor monitors for errors and sends information to the eXtended System Control Facility (XSCF) if an error occurs. On receipt of this notification, the XSCF firmware collects and analyzes the error logs. By taking advantage of the SPARC64™ X/SPARC64™ X+ processor's extensive error notification features, the Fujitsu M10 system can identify the location and type of fault quickly and accurately while continuing operation. The system then provides information useful for preventive maintenance, increasing serviceability.

7. Conclusion

As the sheer volume of data explodes in this era of big data and clouds, customers from across the processing spectrum require new levels of scalability and flexibility, reliability and manageability, and above all, high performance. On all counts, Fujitsu M10 systems deliver on these needs.

The product of a long, stable history in high performance supercomputing and mission-critical processing; Fujitsu's tenth generation SPARC64™ X and SPARC64™ X+ processors provide superior performance and a break-through new concept in business computing: Software on Chip. This architecture takes a bold step forward to accelerate key database functions and foreshadow a new blurring of the line between hardware and software.

The Fujitsu M10 model line wraps the processors in highly scalable and flexible systems that boast a unique new Building Block architecture, full support of Oracle VM for SPARC, and further flexibility from Oracle Solaris Zones. These technologies, together with per-core activation of resources, allow new levels of granularity in how Fujitsu M10 grows with you. The mainframe-class RAS functionality that Fujitsu prides itself on is found throughout the Fujitsu M10 systems and extends to new levels of protection. Scaling from the entry-priced, but mid-range performing, M10-1 single socket model all the way to the largest stack of M10-4S Building Blocks (64 sockets), Fujitsu M10 delivers a consistent, easy-to-use management interface in the form of the XSCF. Furthermore, the Fujitsu M10-4S server supports combined SPARC64™ X and SPARC64™ X+ chassis in a single system, protecting previous IT investments.

Advanced technical features like Liquid Loop Cooling, a heritage of reliability, world-beating performance, and a trusted partnership for hardware-software collaboration all combine to make Fujitsu M10 systems the best answer to the ever-expanding needs and growth of you, our customers and partners.

Contact
FUJITSU LIMITED
Shiodome City Center, 5-2, Higashi-shimbashi 1-chome, Minato-ku,
Tokyo 105-7123 Japan
Website: http://www.fujitsu.com/global/products/computing/servers/unix/sparc/