SPARC Enterprise M-series Servers Architecture

Architecture Flexible, Mainframe-Class Computer Power

White Paper

Manual Code: C120-E636-02EN
April 2011, Revision A

Disclaimer: The only warranties granted by Oracle and Fujitsu Limited, and/or any affiliate of either of them in connection with this document or 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 NON-INFRINGEMENT, 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.

Table of Contents

1. Introduction

Organizations now rely on technology more than ever before. Today, compute systems play a critical role in every function from product design to customer order fulfillment. In many cases, business success is dependent on continuous availability of IT services. Once only required in pockets of the datacenter, mainframe-class reliability and serviceability are now essential for systems throughout the enterprise. In addition, powering datacenter servers and keeping services running through a power outage are significant concerns.

While availability is a top priority, costs must also remain in budget and operational familiarity maintained. To deliver networked services as efficiently and economically as possible, organizations look to maximize use of every IT asset through consolidation and virtualization strategies. As a result, modern IT system requirements reach far beyond simple measures of compute capacity. Highly flexible servers are required with built-in virtualization capabilities and associated tools, technologies, and processes that work to optimize sever utilization. New computing infrastructures must also help protect current investments in technology and training.

Fujitsu SPARC Enterprise M-series are highly reliable, easy to manage, vertically-scalable systems with many of the benefits of traditional mainframes — without the associated cost, complexity, or vendor lock-in. In fact, these servers deliver a mainframe-class system architecture at open systems prices. With symmetric multiprocessing (SMP) scalability from one to 64 processors, memory subsystems as large as 4 TB, and high-throughput I/O architectures, SPARC Enterprise M-series easily perform the heavy lifting required by consolidated workloads. Furthermore, these servers run the powerful Oracle Solaris 10 Operating System (Oracle Solaris 10 OS) and include leading virtualization technologies. By offering Dynamic Domains, eXtended system boards, Dynamic Reconfiguration, and Oracle Solaris Containers technology, SPARC Enterprise M-series bring sophisticated mainframe-class resource control to an open systems compute platform[*1].

[*1] SPARC Enterprise M3000 server do not support Dynamic Domains, eXtended system boards, and Dynamic Reconfiguration capabilities.

2. SPARC Enterprise M-series

Massive compute power, a resilient system architecture, flexible resource control features, and the advanced capabilities of Oracle Solaris 10 combine in SPARC Enterprise servers to provide organizations a best-in-class enterprise platform. As an added benefit, SPARC Enterprise M-series (Figure 2-1) also offer improved performance over previous generations of Fujitsu servers, with a clear upgrade path that protects existing investments in software, training, and datacenter practices. By taking advantage of SPARC Enterprise servers, IT organizations can create a more powerful infrastructure, optimize hardware utilization, and increase application availability — resulting in lower operational cost and risk.

Figure 2-1. The SPARC Enterprise M-series provide organizations with scalable power, reliability, and flexibility.

2.1 Capabilities Overview

The members of the SPARC Enterprise server family share many of the same characteristics that foster power, reliability, and flexibility. SPARC Enterprise servers feature a balanced, highly scalable SMP design that utilizes the latest generation of SPARC64 processors connected to memory and I/O by a new high-speed, low latency system interconnect that delivers exceptional throughput to applications. Also architected to reduce planned and unplanned downtime, these servers include stellar reliability, availability, and serviceability capabilities to avoid outages and reduce recovery time. Design features such as advanced CPU integration and data path integrity, memory extended ECC and memory mirroring, end-to-end data protection, hot-swappable components, fault resilient power options, and hardware redundancy boost the reliability of these servers.

SPARC Enterprise M4000, M5000, M8000, and M9000 servers also provide unmatched configuration flexibility. As in other Fujitsu high-end servers, administrators can use Dynamic Domains to physically divide a single SPARC Enterprise M4000, M5000, M8000, and M9000 servers into multiple electrically isolated partitions, each running independent instances of Oracle Solaris. Hardware or software failures in one Dynamic Domain do not affect applications running in other domains, unless the failed resource is shared across both domains.

Dynamic Reconfiguration can then reallocate hardware resources among Dynamic Domains — without interrupting critical systems. SPARC Enterprise servers advance resource control one-step further with eXtended System Board (XSB) technology, supporting the allocation of sub-system board resources such as CPUs, memory, and I/O components to Dynamic Domains. The fine-grain resource control provided by eXtended System Board technology helps enterprises to further optimize resource utilization.

Adding even more value, the range of compute power offered by the SPARC Enterprise server family provides the levels of vertical scalability required for many deployment classes, letting organizations match the right system to the task. Rackmount SPARC Enterprise M3000 server is the entry-level server that has many characteristics of SPARC Enterprise servers, and shares benefits such as operability and manageability common to the servers. The server combines high performance, high quality, and ecological sustainability with a resilient system architecture, the advanced functions of the Solaris 10 OS, a compact form factor (2U in a rack cabinet), and the top CPU power in the entry-level class of servers.
Rackmount SPARC Enterprise M4000 and SPARC Enterprise M5000 are economical, powerful, and reliable servers well-suited for entry-level and mid-range system requirements (Table 2-1). SPARC Enterprise M8000 and SPARC Enterprise M9000 deliver the massive processing power needed for high-end computing (Table 2-2).

Table 2-1. Characteristics of SPARC Enterprise M3000, M4000, and M5000 servers

	SPARC Enterprise M3000 server	SPARC Enterprise M4000 server	SPARC Enterprise M5000 server
Enclosure	• 2 rack units	• 6 rack units	• 10 rack units
SPARC64 VI Processors	• N/A	• 2.15 GHz • 5 MB L2 cache • Up to four dual-core chips	• 2.15 GHz • 5 MB L2 cache • Up to eight dual-core chips
SPARC64 VII Processors	• 2.52 GHz/2.75 GHz • 5 MB L2 cache • One CPU chip (quad-core or dual-core)	• 2.4 GHz with 5 MB L2 cache • 2.53 GHz with 5.5 MB L2 cache • Up to four quad-core chips	• 2.4 GHz with 5 MB L2 cache • 2.53 GHz with 5.5 MB L2 cache • Up to eight quad-core chips
SPARC64 VII+ Processors	• 2.86 GHz with 5.5 MB L2 cache • One CPU chip (quad-core or dual-core)	• 2.66 GHz with 11 MB L2 cache • Up to four quad-core chips	• 2.66 GHz with 11 MB L2 cache • Up to eight quad-core chips
Memory	• Up to 64 GB • 8 DIMM slots	• Up to 256 GB • 32 DIMM slots	• Up to 512 GB • 64 DIMM slots
Internal I/O Slots	• Four PCI Express	• Four PCI Express • One PCI eXtended	• Eight PCI Express • Two PCI eXtended
External I/O Chassis	• N/A	• Up to two units	• Up to four units
External Onboard Interface	• 4 Gigabit Ethernet ports • SAS port	• 2 Gigabit Ethernet ports	• 2 Gigabit Ethernet ports per IOU
Internal Storage	• Serial Attached SCSI • Up to four drives	• Serial Attached SCSI • Up to two drives	• Serial Attached SCSI • Up to four drives
Dynamic Domains	• One	• Up to two	• Up to four

Table 2-2. Characteristics of SPARC Enterprise M8000, M9000-32, and M9000-64 servers

	SPARC Enterprise M8000 server	SPARC Enterprise M9000 server (32 CPU configuration)	SPARC Enterprise M9000 server (64 CPU configuration)
Enclosure	• One cabinet	• One cabinet	• Two cabinets
SPARC64 VI Processors	• 2.28 GHz with 5 MB L2 cache • 2.4 GHz with 6 MB L2 cache • Up to 16 dual-core chips	• 2.28 GHz with 5 MB L2 cache • 2.4 GHz with 6 MB L2 cache • Up to 32 dual-core chips	• 2.28 GHz with 5 MB L2 cache • 2.4 GHz with 6 MB L2 cache • Up to 64 dual-core chips
SPARC64 VII Processors	• 2.52 GHz/2.88 GHz • 6 MB L2 cache • Up to 16 quad-core chips	• 2.52 GHz/2.88 GHz • 6 MB L2 cache • Up to 32 quad-core chips	• 2.52 GHz/2.88 GHz • 6 MB L2 cache • Up to 64 quad-core chips
SPARC64 VII+ Processors	• 3.0 GHz with 12 MB L2 cache • Up to 16 quad-core chips	• 3.0 GHz with 12 MB L2 cache • Up to 32 quad-core chips	• 3.0 GHz with 12 MB L2 cache • Up to 64 quad-core chips
Memory	• Up to 1 TB • 128 DIMM slots	• Up to 2 TB • 256 DIMM slots	• Up to 4 TB • 512 DIMM slots
Internal I/O Slots	• 32 PCI Express	• 64 PCI Express	• 128 PCI Express
External I/O Chassis	• Up to 8 units	• Up to 16 units	• Up to 16 units
External Onboard Interface	• Gigabit Ethernet ports per IOUA[*a]	• Gigabit Ethernet ports per IOUA[*a]	• 2 Gigabit Ethernet ports per IOUA[*a]
Internal Storage	• Serial Attached SCSI • Up to 16 drives	• Serial Attached SCSI • Up to 32 drives	• Serial Attached SCSI • Up to 64 drives
Dynamic Domains	• Up to 16	• Up to 24	• Up to 24

*a:IOUA is an optional card.

2.2 Entry-Level System — SPARC Enterprise M3000 Server

The SPARC Enterprise M3000 server enclosure measures two rack-units (RU) and supports one processor chip and up to 64 GB of memory. The SPARC64 VII/VII+ (dual-core or quad-core) processor chip is mounted on the motherboard. The I/O subsystem of the SPARC Enterprise M3000 server features four short internal PCI Express slots, four internal disk drives, one internal DVD drive, and an external SAS port for attaching external strage or tape device units. Two power supplies and two fan units power and cool the server. Front and rear views of the SPARC Enterprise M3000 server are found in Figure 2-2.

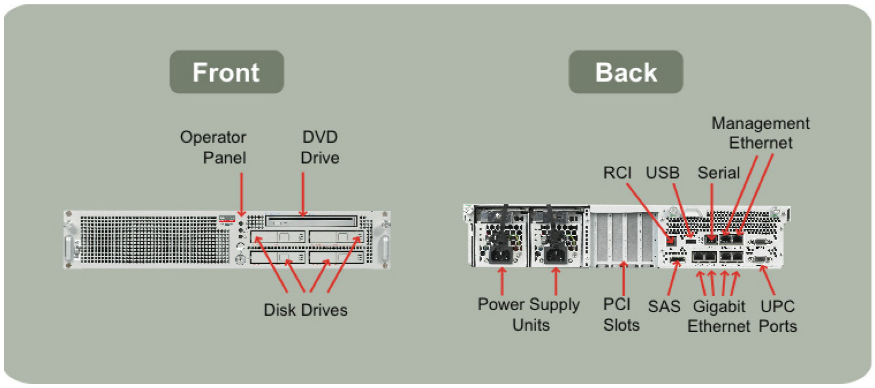

Figure 2-2. SPARC Enterprise M3000 server enclosure diagram.

2.3 Midrange Systems — SPARC Enterprise M4000 and M5000 Servers

SPARC Enterprise M4000 and M5000 servers are economical, high-power compute platforms with enterprise-class features. These midrange servers are designed to reliably carry datacenter workloads that support core business operations.

2.3.1 SPARC Enterprise M4000 Server

The SPARC Enterprise M4000 server enclosure measures six rack-units (RU) and supports up to four processor chips, 256 GB of memory, and up to two Dynamic Domains. In addition, the SPARC Enterprise M4000 server features four short internal PCI Express slots and one short internal PCI-X slot, as well as two disk drives, one DVD drive, and an optional DAT tape drive. Two power supplies and four fan units power and cool the SPARC Enterprise M4000 server. Front and rear views of the SPARC Enterprise M4000 server are found in Figure 2-3.

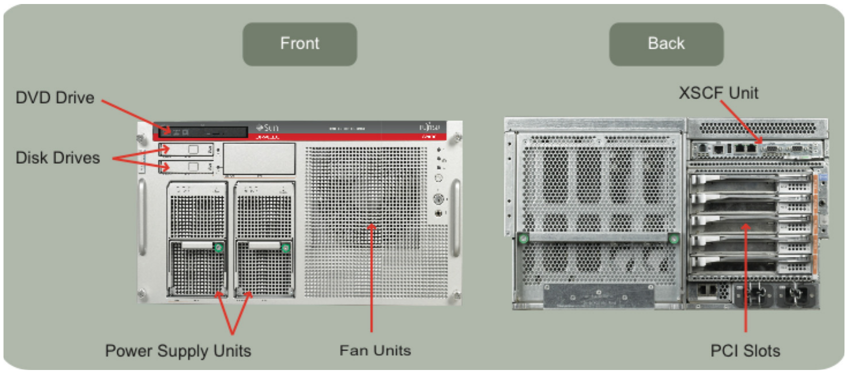

Figure 2-3. SPARC Enterprise M4000 server enclosure diagram.

2.3.2 SPARC Enterprise M5000 Server

The SPARC Enterprise M5000 server enclosure measures 10 RU and supports up to eight processor chips, 512 GB of memory, and up to four Dynamic Domains. In addition, the SPARC Enterprise M5000 server features eight short internal PCI Express and two short internal PCI-X slots, as well as four disk drives, one DVD drive, and an optional DAT tape drive. Four power supplies and four fan units power and cool the SPARC Enterprise M5000 server. Front and rear views of the SPARC Enterprise M5000 server are found in Figure 2-4.

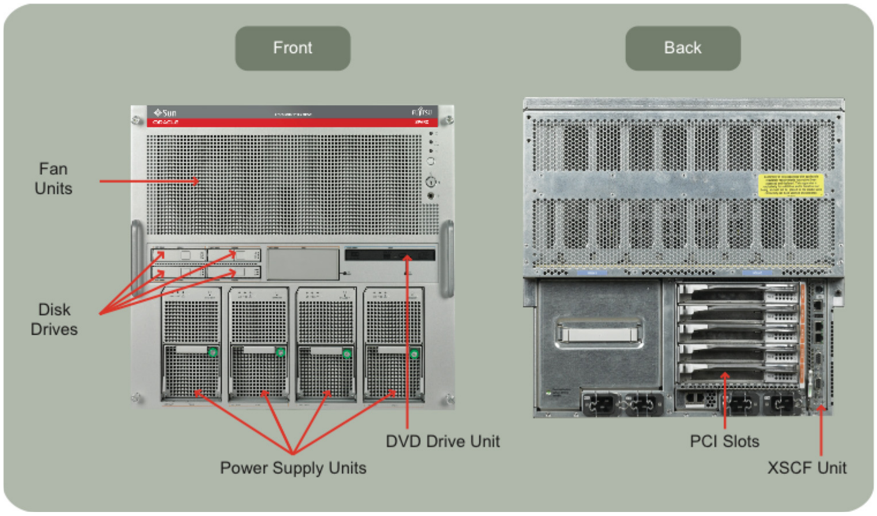

Figure 2-4. SPARC Enterprise M5000 server enclosure diagram.

2.4 High-End Systems — SPARC Enterprise M8000 and M9000 Servers

High-end SPARC Enterprise servers are designed to deliver outstanding performance for even the most challenging workloads. By merging mainframe reliability, advanced performance technology often used in supercomputers, and an open systems environment, these servers help organizations create reliable, high-throughput, flexible solutions.

2.4.1 SPARC Enterprise M8000 Server

The SPARC Enterprise M8000 server is mounted in an enterprise system cabinet and supports up to four CPU Memory Units (CMU) and four I/O Units (IOU). Fully configured, the SPARC Enterprise M8000 server houses 16 processor chips, 1 TB of memory, 32 short internal PCI Express slots, and can be divided into 16 Dynamic Domains. In addition, the SPARC Enterprise M8000 server supports up to 16 disk drives, one DVD drive, and an optional DAT tape drive. Nine power supplies and 12 fan units power and cool the SPARC Enterprise M8000 server. Front and rear views of the SPARC Enterprise M8000 server are found in Figure 2-5.

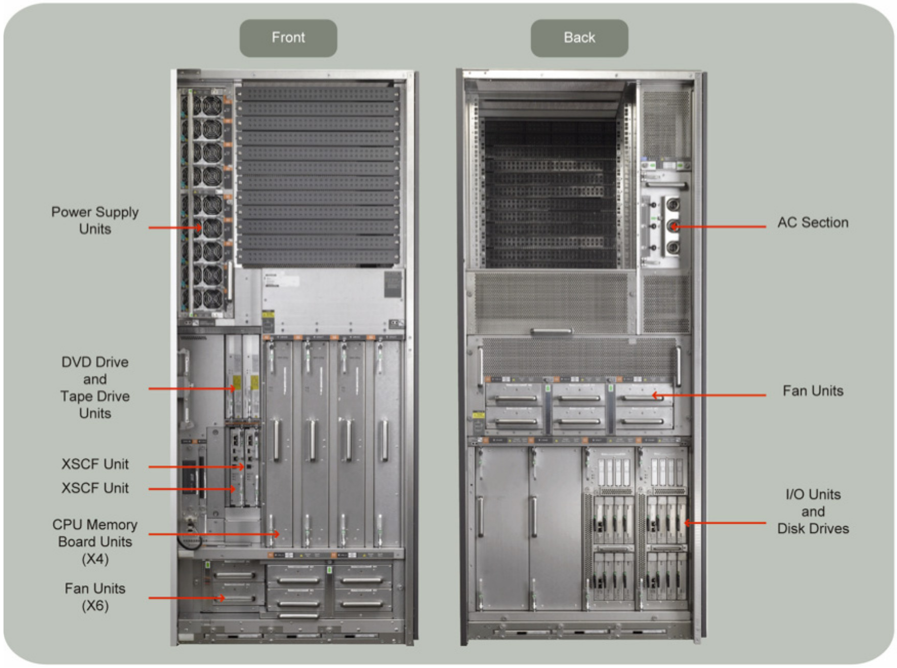
Front
Back
Power Supply Units
DVD Drive and Tape Drive Units
XSCF Unit
XSCF Unit
CPU Memory Board Units (X4)
Fan Units (X6)
AC Section
Fan Units
I/O Units and Disk Drives

Figure 2-5. SPARC Enterprise M8000 server enclosure diagram.

2.4.2 SPARC Enterprise M9000 Server (32 CPU Configuration)

The SPARC Enterprise M9000 server (32 CPU configuration) mounts in an enterprise system cabinet and supports up to eight CMUs and eight IOUs. Fully configured, the SPARC Enterprise M9000 server (32 CPU configuration) houses 32 processor chips, 2 TB of memory, 64 short internal PCI Express slots, and can be divided into 24 Dynamic

Domains. In addition, the SPARC Enterprise M9000 server (32 CPU configuration) supports up to 32 disk drives, one DVD drive, and an optional DAT tape drive. Power and cooling for the SPARC Enterprise M9000 server (32 CPU configuration) is provided by 15 power supplies and 16 fan units. Front and rear views of the SPARC Enterprise M9000 server (32 CPU configuration) are found in Figure 2-6.

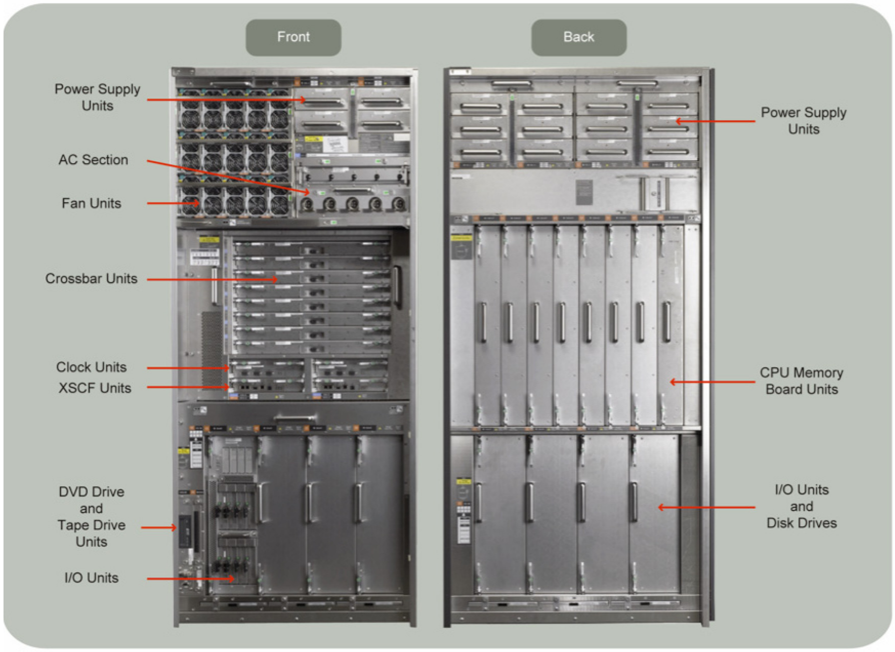

Figure 2-6. SPARC Enterprise M9000 server (32 CPU configuration) enclosure diagram.

2.4.3 SPARC Enterprise M9000 Server (64 CPU configuration)

An expansion cabinet can be added to an existing base cabinet to create the SPARC Enterprise M9000 server (64 CPU configuration), supporting up to 16 CMUs and 16 IOUs. Fully configured, the SPARC Enterprise M9000 server (64 CPU configuration) houses 64 processor chips, 4 TB of memory, 128 short internal PCI Express slots, and can be divided into 24 Dynamic Domains. In addition, the SPARC Enterprise M9000 server (64 CPU configuration) supports up to 64 disk drives two DVD drives, and two optional DAT tape drives. The SPARC Enterprise M9000 server (64 CPU configuration) utilizes 30 power supplies and 32 fan units for power and cooling. A front of the SPARC Enterprise M9000 server (64 CPU configuration) is found in Figure 2-7.

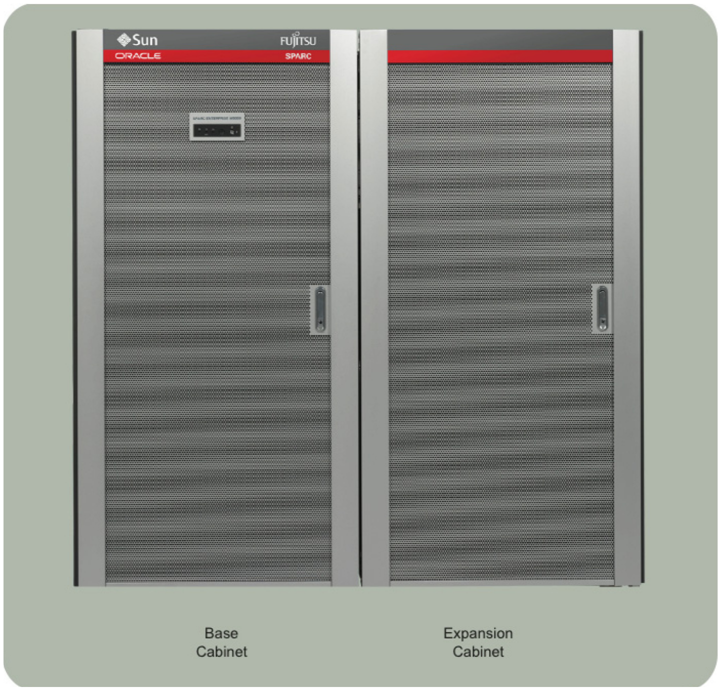

Figure 2-7. SPARC Enterprise M9000 server (64 CPU configuration) enclosure diagram.

2.5 Meeting the Needs of Commercial and Scientific Computing

Suiting a wide range of computing environments, SPARC Enterprise M-series provide the availability features needed to support commercial computing workloads along with the raw performance demanded by high performance computing (HPC) (Table 2-3).

Table 2-3. The power and flexibility of SPARC Enterprise servers benefit a broad range of enterprise applications.

SPARC Enterprise M3000, M4000, and SPARC Enterprise M5000 servers	SPARC Enterprise M8000 and SPARC Enterprise M9000 servers
• Server consolidation • Business processing (ERP, CRM, OLTP, Batch) • Database • Decision support • Datamart • Web services • System and network management • Application development • Scientific engineering	• Server consolidation • Business processing (ERP, CRM, OLTP, Batch) • Database • Decision support • Data warehouses • IT infrastructure • Application serving • Compute-intensive scientific engineering

3. System Architecture

Continually challenged by growing workloads and demands to do more with less, IT organizations realize that meeting processing requirements with fewer, more powerful systems can provide economic advantages. In SPARC Enterprise M-series the interconnect, processors, memory subsystem, and I/O subsystem work together to create a scalable, high-performance platform ready to address server consolidation needs. By taking advantage of these servers, organizations can load multiple projects onto a single platform and accelerate application execution at lower costs.

3.1 System Component Overview

The design of SPARC Enterprise servers specifically focuses on delivering high reliability, outstanding performance, and true SMP scalability. The characteristics and capabilities of every subsystem within these servers work toward this goal. A high-bandwidth system bus, powerful SPARC64 VI and SPARC64 VII/VII+ processors, dense memory option, and fast PCI Express (PCIe), and PCI eXtended (PCI-X) expansion slots combine within these servers to deliver high levels of uptime and throughput, as well as dependable scaling for enterprise applications.

3.1.1 System Interconnect

Based on mainframe technology, the Jupiter system interconnect fosters high level of performance, scalability and reliability for SPARC Enterprise M-series. A single system controller within SPARC Enterprise M3000 servers and multiple system controllers and crossbar units within SPARC Enterprise M4000, M5000, M8000, and M9000 servers 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 faulty switch. Indeed, the system interconnect used in these servers delivers as much as
737 GB/second of peak bandwidth, offering 5.5x more system throughput than Fujitsu's previous generation of high-end servers. Additional technical details for the system interconnect on each SPARC Enterprise server are found in *Chapter 3 – System Bus Architecture*.

3.1.2 The SPARC64 VI and SPARC64 VII/VII+ Processors

SPARC Enterprise M3000 servers support SPARC64 VII and VII+ processors while the SPARC Enterprise M4000, M5000, M8000, and M9000 servers can utilize SPARC64 VI and SPARC64 VII/VII+ processor developed by Fujitsu. The design of the multi-core, multithreaded SPARC64 VI and SPARC64 VII/VII+ processors are based on several decades of experience in creating mainframe systems that achieve high levels of reliability and performance. SPARC64 VI dual-core, multithreaded processor takes advantage of 90 nm technologies while the SPARC64 VII/VII+ processor provides a quad-core implementation with a faster clock speed and a reduction in size using 65 nm fabrication. Both processors execute at a power consumption level below 150 W. Moreover the SPARC Enterprise M4000, M5000, M8000, and M9000 servers increase flexibility and maintain investment protection by supporting configurations that can mix SPARC64 VI and SPARC64 VII/VII+ processors within the same system board or the same Dynamic Domain. Additional technical details about the SPARC64 processors are found in *Chapter 4 – SPARC64 VI/SPARC64 VII/SPARC64 VII+ Processors*.

3.1.3 Memory

The memory subsystem of SPARC Enterprise M-series increase the scalability and throughput of these systems. In fact, the SPARC Enterprise M9000 server accommodates up to 4 TB of memory. SPARC Enterprise M3000 server support DDR-II DIMMs with 2-way memory interleave. SPARC Enterprise M4000, M5000, M8000, and M9000 servers use DDR-II DIMM with 8-way memory interleave to enhance system performance. While multiple DIMM sizes are not supported within a single bank, DIMM capacities can vary across system boards. Available DIMM sizes include 1 GB, 2 GB, 4 GB, and 8 GB. Further details about the memory subsystem of each SPARC Enterprise server are described in Table 3-1.

Table 3-1. SPARC Enterprise server memory subsystem specifications.

	SPARC Enterprise M3000 server	SPARC Enterprise M4000 server	SPARC Enterprise M5000 server	SPARC Enterprise M8000 server	SPARC Enterprise M9000 server (32 CPU configuration)	SPARC Enterprise M9000 server (64 CPU configuration)
Maximum Memory Capacity	• 64 GB	• 256 GB	• 512 GB	• 1 TB	• 2 TB	• 4 TB
DIMM slots	• 8	• 32	• 64	• Up to 128	• Up to 256	• Up to 512
Bank Size	• 4 DIMMs	• 4 DIMMs	• 4 DIMMs	• 8 DIMMs	• 8 DIMMs	• 8 DIMMs
Number of Banks	• 2	• 8	• 16	• Up to 16	• Up to 32	• Up to 64

Beyond performance, the memory subsystem of SPARC Enterprise servers are 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 of these servers provides single-bit error correction, supporting continuous processing despite events such as burst read errors that are sometimes caused by memory device failures. This feature is similar to IBM's Chipkill technology.

3.1.4 Memory Mirroring

SPARC Enterprise M4000, M5000, M8000, and M9000 servers support memory mirroring capabilities. Memory mirroring is an optional, high-availability feature appropriate for execution of applications with the most stringent availability requirements. When memory mirroring mode is enabled on SPARC Enterprise M4000, M5000, M8000, and M9000 servers, 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 SPARC Enterprise M4000 and M5000 servers, memory is mirrored within the same memory module, using the common memory address controller (MAC) Application Specific Integrated Circuit (ASIC) (Figure 3-1 and Figure 3-2).

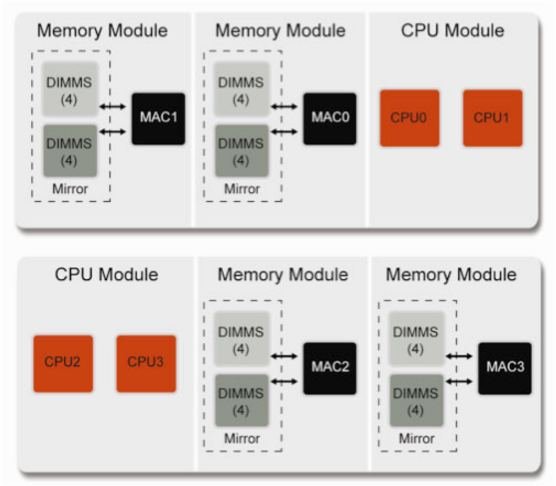

Figure 3-1. SPARC Enterprise M4000 server memory mirroring architecture.

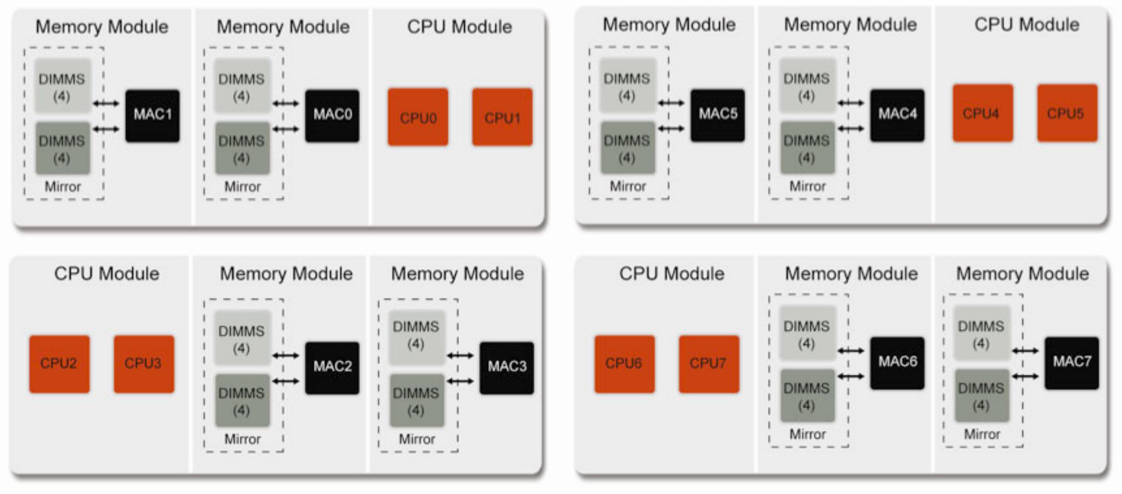

Figure 3-2. SPARC Enterprise M5000 server memory mirroring architecture.

On SPARC Enterprise M8000 and M9000 servers, memory is mirrored across adjacent MAC ASICs to increase reliability (Figure 2-5). However, in a Quad-XSB configuration, paired DIMMs are split across different SPARC Enterprise M8000 and M9000 servers Quad-XSBs. As such, memory mirroring is not supported on a Quad-XSB configuration.

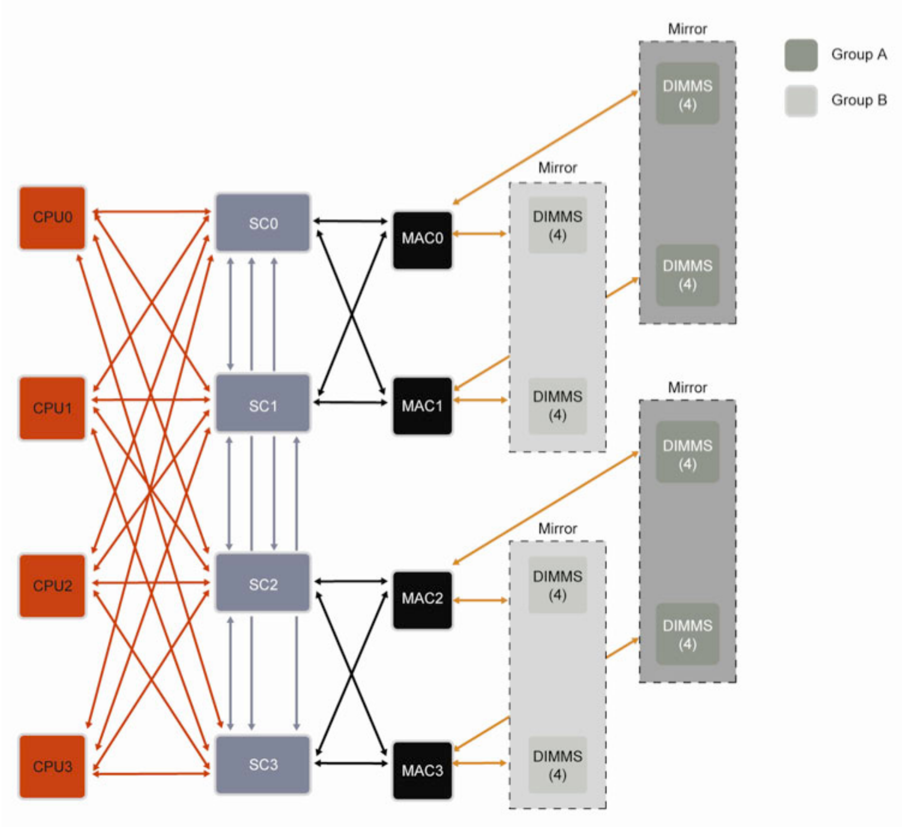

Figure 3-3. SPARC Enterprise M8000 and SPARC Enterprise M9000 memory mirroring architecture.

3.1.5 System Clock

While the implementation of the system clock varies with the member of the SPARC Enterprise M-series each server is engineered with reliability in mind. In particular, SPARC Enterprise M8000 server clock chip is built with redundant internal components. Further enhancing availability and easing maintenance, the SPARC Enterprise M9000 server also implements two sources of clock signal and a dual signal source synchronous line exists between the clock chip and the system boards. In the event one route fails, the system can be restarted via the other route.

3.1.6 PCI-Express and PCI-X Technology

SPARC Enterprise M-series use a PCI bus to provide high-speed data transfer within the I/O subsystem. In order to support PCI Express expansion cards, SPARC Enterprise M-series use a PCI Express physical layer (PCIe PHY) ASIC to manage the implementation of the PCI Express protocol. PCI Express technology doubles the peak data transfer rates of original PCI technology and reaches 2.5 Gb/second of throughput. In fact, PCI Express was developed to accommodate high-speed interconnects such as Fibre Channel, Infiniband, and Gigabit Ethernet. SPARC Enterprise M4000, M5000, M8000, and M9000 servers also support PCI-X expansion cards. PCI-X is backward compatible with existing PCI cards, but increases bandwidth enabling data transfer of up to 1 GB/second for 64-bit devices. Additional technical details for SPARC Enterprise server I/O subsystems can be found in *Chapter 5 – I/O Subsystem*.

3.1.7 Service Processor – Extended System Control Facility

Simplifying management of compute systems leads to higher availability levels for hosted applications. With this in mind, SPARC Enterprise M-series include an eXtended System Control Facility (XSCF). The XSCF consists of a dedicated processor that is independent of the server and runs the XSCF Control Package (XCP) to provide remote monitoring and management capabilities. This service processor regularly monitors environmental sensors, provides advanced warning of potential error conditions, and executes proactive system maintenance procedures as necessary. Indeed, while power is supplied to the server, the XSCF constantly monitors the platform even if the system is inactive. XCP facilitates Dynamic Domains configuration, audit administration, hardware control capabilities, hardware status monitoring, reporting, and handling, automatic diagnosis and domain recovery, capacity on demand operations, and XSCF failover services. Additional technical details about the XSCF and XCP are found in *Chapter 7 – System Management*.

3.1.8 Power and Cooling

SPARC Enterprise M-series use separate modules for power and cooling. Sensors placed throughout the system measure temperatures on processors and key ASICS as well as the ambient temperature at several location. Hardware redundancy in the power and cooling subsystems combined with environmental monitoring keep servers operating even under power or fan fault conditions.

3.1.8.1 Fan Unit

Fully redundant, hot-swap fans function as the primary cooling system for SPARC Enterprise M-series (Table 3-2 and Table 3-3). If a single fans fails, the XSCF detects the failure and switches the remaining fans to high-speed operation to compensate for the reduced airflow. SPARC Enterprise M-series can operate normally under these conditions, allowing ample time to service the failed unit. Replacement of fans units can occur without interrupting application processing.

3.1.8.2 Power Supply

The use of redundant power supplies and power cords adds to the fault resilience of SPARC Enterprise M-series
(Table 3-2 and Table 3-3). Power is supplied by redundant hot-swap power supplies, helping to support continued server operation even if a power supply fails. Since the power units are hot-swappable, removal and replacement can occur while the system continues to operate.

As an option, SPARC Enterprise M8000 and SPARC Enterprise M9000 can be ordered with a three-phase power supply unit and corresponding server cabinet. Models with a three-phase power supply permit two configurations, a star connection that connects a neutral line and each phase, and a delta connection that connects each phase.

Table 3-2. SPARC Enterprise M3000, M4000, and M5000 servers power and cooling specifications.

	SPARC Enterprise M3000 server	SPARC Enterprise M4000 server	SPARC Enterprise M5000 server
Fan Units	• Two fan units • Two 80 mm fans • Two 60 mm fans • 1+1 redundant	• Four fan units • Two 172 mm fans • Two 60 mm fans • One of each type is redundant	• Four fan units • Four 172 mm fans • Two fan groups, each containing two fan units • One redundant fan per fan group
Power Supplies	• AC: 505 W, DC: 510 W (2.86 GHz) (max) • Two units • 1+1 redundant • Single-phase AC or DC power supplies	• 1,692 W (2.66 GHz) (max) • Two units • 1+1 redundant • Single-phase	• 1,692 W (2.66 GHz) (max) • Four units • 2+2 redundant • Single-phase
Power Cords	• Two power cables • 1+1 redundant power cables	• Two power cables • 1+1 redundant power cables	• Four power cables • 2+2 redundant power cables

Table 3-3. SPARC Enterprise M8000 and M9000 servers power and cooling specifications.

	SPARC Enterprise M8000 server	SPARC Enterprise M9000 server (32 CPU configuration)	SPARC Enterprise M9000 server (64 CPU configuration)
Fan Units	• 12 fan units • Four 172 mm fans • Eight 60 mm fans • N+1 redundant	• 16 fan units • 16 172 mm fans • N+1 redundant	• 32 fan units • 32 172 mm fans • N+1 redundant
Power Supplies	• 10,500 W (3.0 GHz) (max) • 9 units • N+1 redundant	• 20,220 W (3.0 GHz) (max) • 15 units • N+1 redundant	• 40,440 W (3.0 GHz) (max) • 30 units • N+1 redundant
Options	• Single-phase • Three-phase • Dual-grid	• Single-phase • Three-phase • Dual-grid	• Single-phase • Three-phase • Dual-grid
Power Cords	• 3 power cables (single feed) • 6 power cables (dual feed) • 2 power cables (three-phase)	• 5 power cables (single feed) • 10 power cables (dual feed) • 2 power cables (three-phase)	• 10 power cables (single feed) • 20 power cables (dual feed) • 4 power cables (three-phase)

3.1.8.3 Optional Dual Power Feed

While organizations can control most factors within the datacenter, utility outages are often unexpected. The consequences of loss of electrical power can be devastating to IT operations. In order to help reduce the impact of such incidents, SPARC Enterprise M-series are dual power feed capable. The AC power subsystem in these servers is completely duplicated, providing optional reception of power from two external and independent AC power sources. The use of a dual power feed and redundant power supplies increases system availability, as server operations can remain unaffected even after a single power grid failure.

3.1.9 Operator Panel

SPARC Enterprise M-series feature an operator panel to display server status, store server identification and user setting information, change between operational and maintenance modes, and turn on power supplies for domains (Figure 3-4). During server startup, the front panel LED status indicators verify XSCF and server operation.

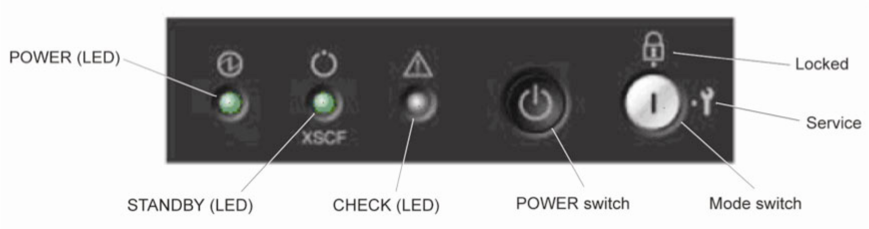

Figure 3-4. The SPARC Enterprise server operator panel.

4. System Bus Architecture — Jupiter Interconnect

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. SPARC Enterprise M-series utilize a system interconnect designed to deliver massive bandwidth and consistent, low latency between components. The Jupiter system bus benefits IT operations by delivering balanced and predictable performance to application workloads.

4.1 Interconnect Architecture

The Jupiter interconnect design maximizes the overall performance of SPARC Enterprise M-series. 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 plus ECC in each direction.

System controllers within the interconnect architecture on all SPARC Enterprise M4000, M5000, M8000, and M9000 servers direct traffic between local CPUs, memory, I/O subsystems, and interconnect paths. On SPARC Enterprise M8000 and M9000 servers, the system bus is implemented as a crossbar switch between system boards to support high-throughput data transfer with consistent latency times between all components. To improve performance, the physical addressing of memory on a motherboard of a SPARC Enterprise M4000 and M5000 servers or the CMU of a SPARC Enterprise M8000 and M9000 servers is evenly spread out across all system controllers on the same board.

4.1.1 SPARC Enterprise M3000 Server Interconnect Architecture

The SPARC Enterprise M3000 server system design is contained within a single motherboard. Within the architecture, the Jupiter System Controller (JSC) is a single ASIC that performs both the memory address controller and system controller functions. The JSC is connected to the CPU, memory DIMMs, and the I/O controller (PCIe bridge). An architecture diagram of the SPARC Enterprise M3000 server is shown in Figure 4-1.

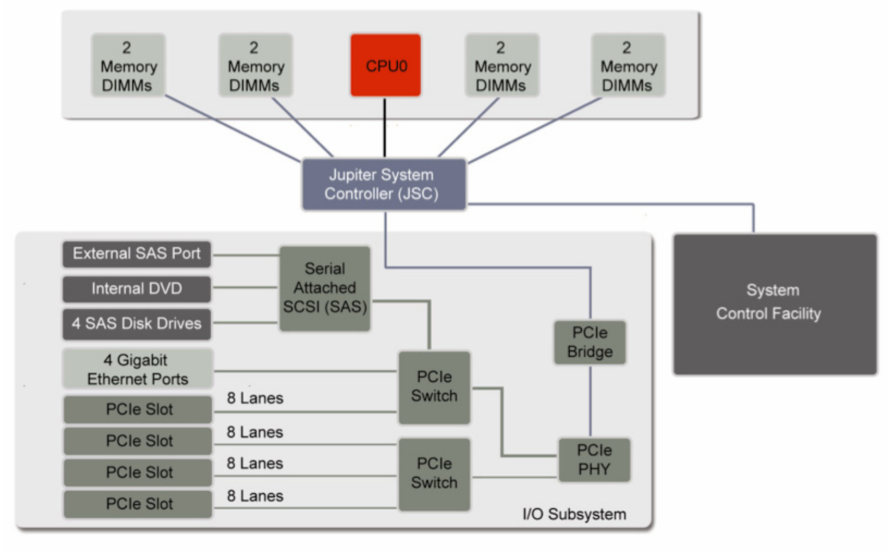

Figure 4-1. SPARC Enterprise M3000 server interconnect diagram.

4.1.2 SPARC Enterprise M4000 Server Interconnect Architecture

The SPARC Enterprise M4000 server is implemented within a single motherboard that contains two system controllers. Both system controllers connect to each other, as well as CPU modules, memory address controllers, and the IOU
(Figure 4-2).

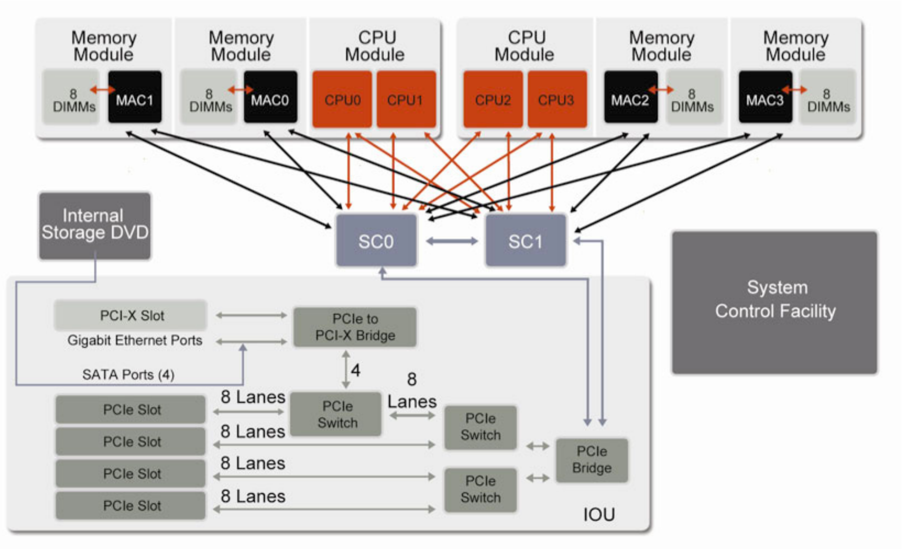

Figure 4-2. SPARC Enterprise M4000 server interconnect diagram.

4.1.3 SPARC Enterprise M5000 Server Interconnect Architecture

The SPARC Enterprise M5000 server is implemented within a single motherboard but features two logical system boards. Similar to the SPARC Enterprise M4000 design, each logical system board contains two system controllers that connect to each other, as well as CPU modules, memory access controllers, and an IOU. In addition, each system controller connects to a corresponding system controller on the other logical system board (Figure 4-3).

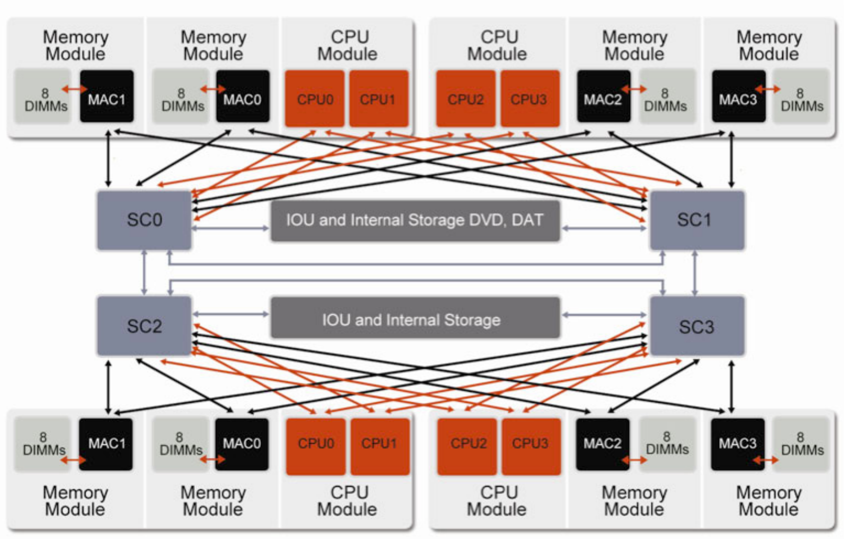

Figure 4-3. SPARC Enterprise M5000 server interconnect diagram.

4.1.4 SPARC Enterprise M8000 and SPARC Enterprise M9000 System Interconnect Architecture

SPARC Enterprise M8000 and SPARC Enterprise M9000 feature multiple system boards that connect to a common crossbar. Each system board contains four system controllers and each system controller connects to every CPU module. For improved bandwidth, every memory access controller connects to two system controllers, and each system controller connects to every other system controller within the system board. The system controllers also provide a connection to each crossbar unit, supporting data transfer to other system boards (Figure 4-4).

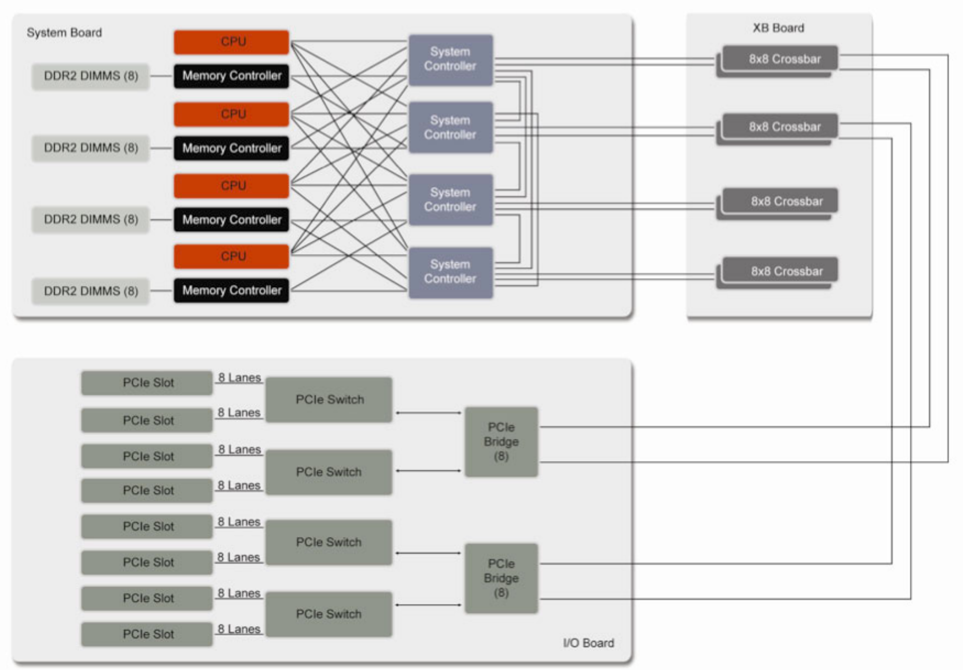

Figure 4-4. SPARC Enterprise M8000 and SPARC Enterprise M9000 system interconnect diagram.

4.2 System Interconnect Reliability Features

Built-in redundancy and reliability features of SPARC Enterprise M-series system interconnect enhance the stability of these servers. The Jupiter interconnect protects against loss or corruption of both address and data of transaction with ECC 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 performs the transfer. SPARC Enterprise M8000 and M9000 servers can automatically make the best choice of whether degrading a specific bus or degrading the crossbar switches. In the rare event of a hardware failure within the interconnect, the system uses the surviving bus route on restart, isolating the faulty crossbar and facilitating the resumption of operations.

4.3 Scalable Performance

The high bandwidth and overall design of the Jupiter system interconnect contributes to the scalable performance of SPARC Enterprise M-series. Theoretical peak system throughput, snoop bandwidth, and I/O Bandwidth numbers, as well as Stream benchmark results are found in Table 4-1.

In SPARC Enterprise M4000, M5000, M8000, and M9000 servers, the CPUs, memory address controllers, and I/O controllers are directly connected to the system controllers by a high-speed broadband switch for data transfer. As a result, a relatively even latency can be maintained between individual components. As components are added, processing capability and latency are not degraded. In fact, the crossbar interconnect implementation in SPARC Enterprise M8000 and M9000 servers results in increased interconnect bandwidth every time a system board is added to the server.

Table 4-1. Theoretical system bandwidth and theoretical I/O bandwidth at peak time, snoop bandwidth, and stream benchmark results for SPARC Enterprise.

	Theoretical system bandwidth at peak time*a (GB/second)	Snoop bandwidth (GB/second)	Stream Benchmark Triad results (GB/second)	Stream Benchmark Copy results (GB/second)	Theoretical Peak I/O bandwidth*b (GB/second)
SPARC Enterprise M3000	20	N/A	5.1	6.4	4
SPARC Enterprise M4000	32	129	12.7	12.5	8
SPARC Enterprise M5000	64	129	25.2	24.8	16
SPARC Enterprise M8000	184	245	69.6	60.3	61
SPARC Enterprise M9000 (32-CPU configuration)	368	245	134.4	114.9	122
SPARC Enterprise M9000 (64-CPU configuration)	737	245	227.1	224.4	244

*a:Theoretical Peak System Bandwidth is calculated by multiplying the bus width by the frequency of the bus between the system controller and the memory access controller.

*b:Theoretical Peak I/O Bandwidth is calculated by multiplying the bus width by the frequency of the bus between the system controller and the PCI bridge.

5. SPARC64 VI and SPARC64 VII/VII+ Processors

SPARC Enterprise M3000 server supports quad-core and dual-core of SPARC64 VII and SPARC64 VII+ processors. SPARC Enterprise M4000, M5000, M8000, and M9000 servers can utilize both the Fujitsu dual-core SPARC64 VI and quad-core SPARC64 VII/ VII+ processors. These processors incorporate innovative multi-core and multithreaded technology, and provide extensive reliability features. In addition, the SPARC64 VI and SPARC64 VII/VII+ processors are SPARC V9 level 2 compliant, helping to provide support for thousands of existing software applications. The use of SPARC64 VI and SPARC64 VII/VII+ processors within SPARC Enterprise M-series offer organizations exceptional reliability, application choice, and outstanding processing power.

5.1 Next-Generation Processor Technology

The past decade introduced major changes to processor architectures as system design engineers found that increases to CPU clock rates began exhibiting diminishing returns on performance, and creating power and heat concerns. Innovations such as Chip Multiprocessing (CMP), Vertical Multithreading (VMT), and Simultaneous Multithreading (SMT) technologies now dominate plans for improving compute capacity.

5.1.1 Chip Multiprocessing (CMT)

Chip Multiprocessing technology (CMT) is an architecture in which multiple physical cores are integrated on a single processor module. Each physical core runs a single execution thread of a multithreaded application independently from other cores at any given time. With this technology, dual-core processors often double the performance of single-core modules. The ability to process multiple instructions at each clock cycle provides the bulk of the performance advantage, but improvements also result from the short distances and fast bus speeds between chips as compared to traditional CPU to CPU communication.

5.1.2 Vertical Multithreading (VMT)

Vertical multithreading (VMT) technology lets a single physical core host multiple threads, each viewed by the operating system as a virtual CPU. Multiple threads on the same core run in a time-sliced fashion, with only one executing at any given moment. A thread does not run if it is idle, or has encountered a cache miss and is waiting for main memory. A thread switch occurs on events such as an L2 cache miss, hardware timer, interrupt, or specific instruction to control threads. In this way, VMT improves system performance by maximizing processor utilization and effectively mitigating the impact of a cache miss. VMT is enabled automatically to improve performance when the number of threads in the system exceeds the number of cores.

5.1.3 Simultaneous Multithreading (SMT)

Simultaneous multithreading (SMT) technology supports the simultaneous execution of multiple threads in a multithreaded core. From the software point of view, each thread is independent. This method of multithreading is facilitated by duplicating compute resources. A few resources remain shared between threads, such as instruction buffers, Reservation Station, and caches. With SMT, context switch time is eliminated and threads within a single core share the instruction pipeline smoothly. When both threads are ready to run, they alternate cycles for superscaler instruction issue, and share the functional units according to need.

5.2 Architecture of SPARC64 VII and SPARC64 VII/VII+ Processors

Steady enhancements and changes rather than design overhauls are key to the current success of SPARC64 series processors. Toward that end, the SPARC64 VI and SPARC64 VII/VII+ CPU modules reuse the proven and reliable SPARC64 V core with no major pipeline changes. To increase throughput, the SPARC64 VI and SPARC64 VII/VII+ processors leverage multi-core, multithreaded architectures. The SPARC64 VI processor utilizes VMT technology and offers two cores and two threads per core. The SPARC 64 VII/VII+ processors incorporates a quad-core architecture with two threads per core and takes advantage of SMT technology to further improve performance. Specifications for the SPARC64 VI and SPARC64 VII/VII+ processors are detailed in Table 5-1.

Table 5-1. SPARC64 VI and SPARC64 VII/VII+ processors specifications

	SPARC VI processor	SPARC VII processor	SPARC VII+ processor
Speed	• 2.15 GHz • 2.28 GHz • 2.4 GHz	• 2.4 GHz • 2.52 GHz • 2.53 GHz • 2.75 GHz • 2.77 GHz • 2.88 GHz	• 2.66 GHz • 2.86 GHz • 3.0 GHz
Architecture	• Dual-core • SPARC V9 • sun4u • 90 nm process technology	• Quad-core, Dual-core[*a] • SPARC V9 • sun4u • 65 nm process technology	• Quad-core, Dual-core[*a] • SPARC V9 • sun4u • 65 nm process technology
L1 cache	• 128 KB L1 I-cache per core • 128 KB L1 D-cache per core	• 64 KB L1 I-cache per core • 64 KB L1 D-cache per core	• 64 KB L1 I-cache per core • 64 KB L1 D-cache per core
L2 cache	• 5 MB (2.15 GHz and 2.28 GHz) • 6 MB (2.4 GHz) • 10-way associative (2.15 and 2.28 GHz) • 12-way associative (2.4 GHz) • 256 byte line size • ECC tag and data protection	• 5 MB (2.4 GHz, 2.52 GHz[*a], 2.75 GHz, and 2.77 GHz) • 5.5 MB (2.53 GHz) • 6 MB (2.52 GHz[*b] and 2.88 GHz) • 12-way associative • 256 byte line size • ECC tag and data protection	• 5.5 MB (2.86 GHz) • 11 MB (2.66 GHz) • 12 MB (3.0 GHz) • 12-way associative • 256 byte line size • ECC tag and data protection
Power	• 120 watts nominal • 150 watts maximum	• 135 watts nominal • 150 watts maximum	• 135 watts nominal • 160 watts maximum

*a:M3000 server
*b:M8000/M9000 servers

5.2.1 Cache System

As shown in Figure 5-1, SPARC64 VI and SPARC64 VII/VII+ processors implement a two-layer cache memory structure. Both processors include a moderate-capacity primary cache (L1 cache) and a high-capacity secondary cache (L2 cache). The L1 cache consists of a cache dedicated for instructions (L1 I-cache) and a cache dedicated for operands (L1 D-cache). The L1 D-cache is divided into eight banks on four-byte address boundaries, and two operands can be accessed at one time. The L1 cache utilizes virtual addresses for cache indexes and physical addresses for cache tags, a method known as virtually indexed physically tagged (VIPT). In order to protect against the VIPT synonym problem — indexing the same physical address to two virtual addresses — the L2 cache on SPARC64 VI and SPARC64 VII/VII+ processors prevents creation of synonym entries in the L1 cache.

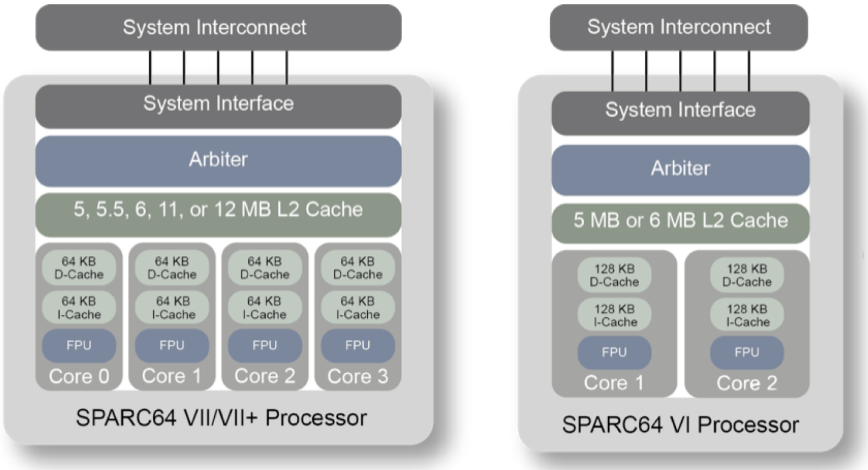

Figure 5-1. The SPARC64 VI and SPARC64 VII/VII+ processors incorporate a cache structure.

Within SPARC64 VI and SPARC64 VII/VII+ processors, the L2 cache is shared across all cores. The bus for sending data that is read from the L2 cache to the L1 cache provides a width of 32 bytes per two cores, and the bus for sending data from the L1 D-cache to L2 cache offers a width of 16 bytes per core.

The cache update policies of L1 cache and L2 cache are both write-back. Updates to cached data are written only to cache rather than system memory. As a result, store operations can complete with the update of one cache hierarchy. Data in the cache location is written-back to system memory when the cache line is reassigned to another memory location. Given the high frequency of the store operation, the write-back method can provide a performance advantage by reducing inter-cache and memory access traffic.

Since the write-back method keeps the latest data in the cache, special measures must be taken to prevent a single cache fault from impacting system wide operations. SPARC64 VI and SPARC64 VII/VII+ processors include advanced capabilities to track error conditions and force write-backs from cache to system memory when the potential for processor faults exist. More information on these features can be found in this chapter in the section titled *Reliability, Availability, and Serviceability Features*.

5.2.2 SPARC64 VI Processor Micro-Architecture Overview

The SPARC64 VI processor design retains the high performance and reliability of the SPARC64 V processor, while offering considerable throughput improvement. To achieve performance gains, the SPARC64 VI processor implements a combination of CMP and VMT technologies. This processor consists of two physical cores where each core supports two VMT threads, supporting the execution of four threads in parallel. The operating system views each thread as a virtual processor. For example, the Oracle Solaris psradm command can be used to set each virtual CPU as spare, on-line, or off-line as desired.

Two threads that belong to the same physical core share most of the core's resources, such as the ALU and instruction pipeline, while the two physical cores only share the L2 cache and system interface. Using coarse-grained multithreading techniques, a single thread occupies the full resources of the core until a long latency event. Specifically, a thread switch is triggered on an L2 cache miss or passage of a periodic time interval. This approach mitigates the effect of cache misses by scheduling an unblocked thread, while maintaining fairness so all threads make progress.

Support for multiple threads in this manner requires duplicating general-purpose registers (GPR), floating point registers (FPR), the program counter (PC), and the control (state) registers. A copy of GPR called the current window register (CWR) realizes one-cycle register file access time. In addition, a fast, high-bandwidth path from the GPR to the CWR speeds operations when a register window move or thread switch is required.

While the CMP and VMT innovations specifically enhance multithreaded performance, single-threaded throughput is not compromised. The SPARC64 VI processor delivers approximately two times the single threaded performance of SPARC64 V running at 1.35 GHz. Other enhancements to the SPARC64 VI processor include a refined core with floating point improvements and a doubled Translation Lookaside Buffer (TLB) which reduces the miss rate to improve both integer and floating-point application performance.

5.2.3 SPARC64 VII/VII+ Processor Micro-Architecture Overview

The basic structure of the core pipeline of the SPARC64 VII/VII+ processor is the same as that of the SPARC64 VI processor. However, the SPARC64 VII/VII+ processor utilizes SMT technology instead of VMT technology to implement multithreading. As shown in Figure 5-2, the SPARC64 VI processor takes advantage of VMT technology to execute two threads in parallel — only one thread is active at any given time.

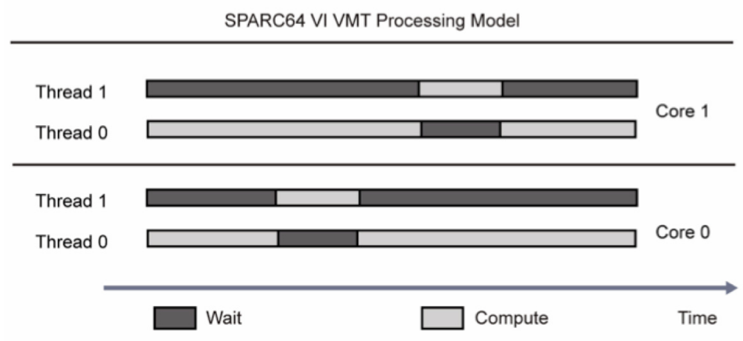

Figure 5-2. Within the VMT processing model utilized by the SPARC64 VI processor only one thread per core is active at any given time.

Within the VMT model, a specific trigger must occur for processing to switch over to the alternate thread. By implementing SMT technology, both threads within each core on the SPARC VII/VII+ processor can execute simultaneously (Figure 5-3). As a result, the SPARC VII/VII+ offers the potential to achieve greater throughput and performance.

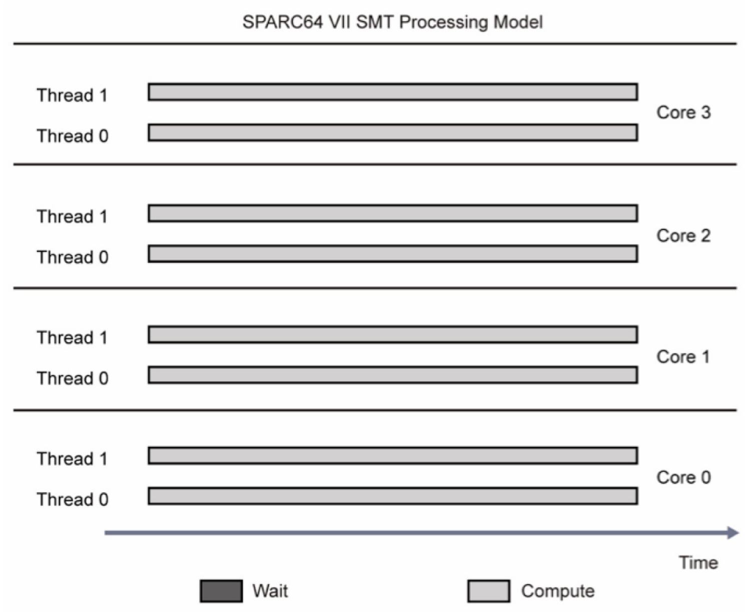

Figure 5-3. SMT technology implemented by the SPARC64 VII/VII+ processor offers the potential execute all threads within a core simultaneously.

While SMT supports execution of both threads simultaneously, these threads still share the same pipeline core. The SPARC64 VII/VII+ processor works to optimize performance by minimizing interference between threads. For example, the instruction fetch stage, instruction decoding stage, and commit stage facilitate selection of either thread within each cycle. This measure helps avoid the potential for a stalled thread to block the progress of an executing thread. In addition, while two threads are executing within a single core, the hardware resources specifically assigned to each thread are kept isolated. At the same time, when one thread reaches an idle state, the executing thread can then fully leverage shared resources — a feature that improves the performance of single-threaded operation.

The SPARC64 VII/VII+ processor also offers the following advanced features :*1

- Integer Multiply-Add Instruction
  The Integer Multiply-Add instruction performs a fused multiply and add instruction on data in double-precision floating-point registers containing unsigned 8-byte integer values. Increasing the efficiency of the multiply and add operation enhances the performance of some applications.

- Shared Context
  Shared Context is a virtual address space that can store instructions or data shared by two or more processes. Utilizing this feature, different processes can use a shared TLB entry to access memory mappings to a set of executables, Intimate Shared Memory segments (ISM), or Dynamic Intimate Shared Memory (DISM) segments — reducing overhead and TLB misses. Implementing a Shared Context can help improve the performance of virtual machines (VM), databases, and other parallelized applications.

5.2.4 SPARC64 VII/VII+ Processors Instruction Processing

Details of the micro-architecture are outlined below.

As shown in Figure 5-4, cores within the SPARC64 VII/VII+ processor include an instruction fetch block and an instruction execution block. The instruction fetch block includes the primary cache for instructions (L1 I-cache), and the instruction execution block includes the primary cache for operands (L1 D-cache).

*1:To take advantage of the advanced features of the SPARC64 VII/VII+ processor, specific compiler versions, operating system levels, and system configuration settings may be required. Please consult system documentation (product note and compiling-related documents) for further information.

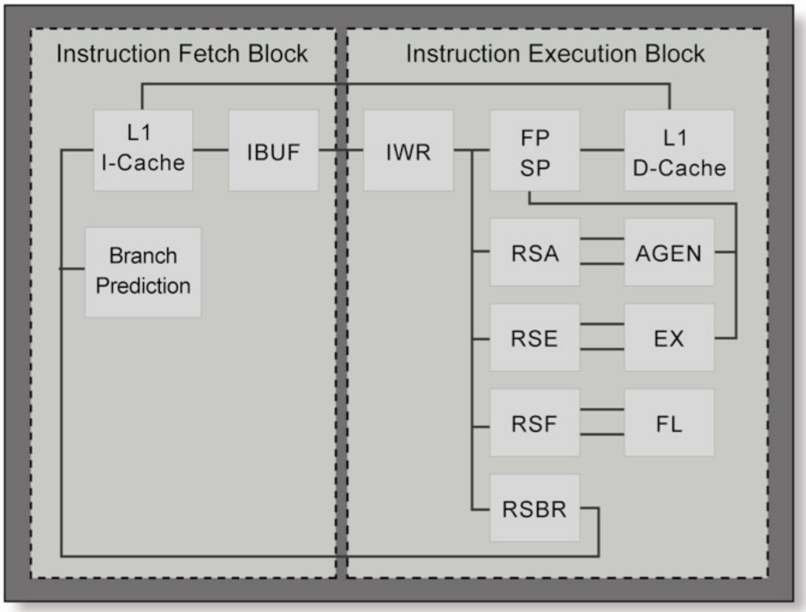

Figure 5-4. Functional diagram of the SPARC64 VII/VII+ processor core.

5.2.5 Instruction Fetch block

The instruction fetch block, which operates independently of the instruction execution block, takes a series of instructions into the instruction buffer (IBUF). The IBUF offers a capacity of 256 bytes, and can store up to 64 instructions. When both threads are running, the IBUF is divided evenly for each thread.

Instructions stored within the IBUF are ready for execution according to branch prediction. In the event that instruction execution is stalled, instruction fetch continues until the IBUF becomes full. In contrast, if the instruction fetch operation pauses for an event such as a cache error, instructions retrieval from the IBUF continues and software execution moves forward as long as the IBUF includes instructions.

Instruction fetch can start in every cycle, and eight instructions are fetched at one time. Importantly, the throughput of instruction execution is a maximum of four instructions per cycle. With twice the throughput, instruction fetch and the IBUF can help improve system performance by concealing any latency imposed by the L1 I-cache.

5.2.6 Instruction Execution Block

Instructions stored in the IBUF during the instruction fetch block are retrieved by the Instruction Word Register (IWR) at a rate of four instructions per cycle. Within the Instruction Execution Block these instructions are decoded, issued, executed, and committed.

5.2.6.1 Instruction Decode and Issue

In the Instruction decode and issue stages, the four instructions in the IWR are decoded simultaneously, and 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, instruction identifications (IID) ranging from 0 to 63 are assigned and the instruction is issued. Given this identification scheme, the maximum number of in-flight instructions is 64 and when both threads are running, the maximum number of instructions for each thread is 32. In each cycle, an instruction from either thread is decoded and threads are alternately switched.

When an instruction is issued, the current IWR is released. There are no restrictions on the allocation of resources such as reservation stations for instructions in any slot of the IWR. Also, there are no restrictions on instruction type combinations. Therefore, 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 in program order. As described above, by eliminating stall conditions of instruction issue as much as possible, a high multiplicity level is assured for any binary code.

5.2.6.2 Instruction Execution

The SPARC64 VII/VII+ processors registers decoded instructions in a reservation station. Reservation stations for integer operation (RSE) and reservation stations for floating point operation (RSF) are provided, as well as reservation stations for branch instructions (RSBR) and reservation stations for calculating addresses for load/store instructions (RSA). The RSEs and RSFs are divided into two queues for the execution unit. — RSEA, RSEB, RSFA, and RSFB. Each instruction stored in a reservation station is dispatched to the execution unit that corresponds to the reservation station in the order in which source operands are prepared for instructions. Therefore, four operations can be dispatched simultaneously.

Generally, the oldest instruction that is ready for dispatch is selected from the instructions in a reservation station. However, in cases where a load instruction is set to update a register and the register is used as a source operand for an operation, the instruction is speculatively dispatched before the result of the load instruction is obtained. Success of the speculative dispatch is determined in the execution stage (EX). Use of speculative dispatch conceals the latency of the pipeline for cache access, increasing the efficiency of the execution unit.

5.2.6.3 Instruction Commit

The maximum number of instructions that can be committed at one time is four. The instruction commit stage is shared by the two threads, and either thread is selected in each cycle to perform commit processing. Results of instructions that are executed out of order are initially stored in the GPR Update Buffer (GUB) and FPR Update Buffer (FUB) work registers and remain inaccessible to software.

To assure the instruction order in a program, registers such as GPR and FPR and memory are updated in program order in the commit stage. In addition, control registers such as the PC are also updated at the same time in the commit stage. As described above, precise interrupt is guaranteed, and processing in execution can always be canceled. Utilizing this synchronous update method, simplifies re-execution of instructions due to a branch prediction error, and helps contribute to increased reliability.

5.3 Reliability, Availability, and Serviceability Functions

The design of the SPARC64 VI and SPARC64 VII/VII+ processor modules increase system reliability by delivering improved fault avoidance and error correction capabilities. In fact, much of the area on the SPARC64 VI and SPARC64 VII/VII+ processors are dedicated to error detection and data correction within the CPU. RAM units are ECC protected or duplicated, and most latches and execution units are parity protected. The powerful RAS function of the SPARC64 processor detects the 1-bit error without fail, and performs correction or retry. For some errors, it degrades the cache in units of way or in units of core.

Other reliability features of the SPARC64 VI and SPARC64 VII/VII+ processors include support for error marking, instruction retry, and preventive maintenance. When memory read data has a multi-bit error, a special mark identifying the source of the error is written into the data and the ECC syndrome becomes a special value, providing valuable information for identifying the exact source of the fault. In addition, when a hardware error is detected, instructions that are currently in execution are cancelled, and retried automatically to prevent transient errors. Error data generated by the SPARC64 VI and SPARC64 VII/VII+ processors is also sent to the service processor to support preventive maintenance.

As described above, in the SPARC64 VII/VII+, RAS functions comparable to mainframe computers have been implemented. With these RAS functions, errors are reliably detected, their effect is kept within a limited range, recovery processing is tried, error logs are recorded, software is notified, and so forth. In other words, the basics of RAS functions are thoroughly implemented. Through the implementation of the RAS functions, the SPARC64 VII/VII+ provides high reliability, high availability, high serviceability, and high data integrity as a processor for mission-critical UNIX servers.

5.3.1 Internal RAM Reliability and Availability Features

The SPARC64 VI and SPARC64 VII/VII+ processors offer reliability and availability features that support high levels of data integrity. Table 5-2, highlights the error detection and correction capabilities of the SPARC64 VI and SPARC64 VII/VII+ processor.

Table 5-2. Detection and error correction methods to support high levels of data integrity

Type		Error detection method Protection method	Error correction method
L1 Instruction Cache	Data	Parity	Invalidation and reread
	Tag	Parity + duplication	Rewrite of duplicated data
L1 Data Cache	Data	SECDED ECC	One-bit error correction using ECC
	Tag	Parity + duplication	Rewrite of duplicated data
L2 Cache	Data	SECDED ECC	One-bit error correction using ECC
	Tag	SECDED ECC	One-bit error correction using ECC
Instruction TLB		Parity	Invalidation
Data TLB		Parity	Invalidation
Branch History		Parity	Recovery from branch prediction failure

SECDED: Single Error Correction Double Error Detection

For the L1 cache, L2 cache, and TLB, degradation can be performed separately in way units. Since the SPARC64 VII/VII+ processor implements a set-associative scheme that divides the L1 cache, L2 cache, and TLB into way units, degradation can be performed in a granular manner. Error occurrence counts are counted for each function unit. When an error occurrence count per unit time exceeds the upper limit, degradation is performed and the relevant way is not used subsequently. Hardware automatically performs degradation and initiates following operations to assure the continuity of coherency of data stored in cache. :

- Write-back to L2 cache — Write-back the dirty lines in the way of the L1D cache to be degraded
- Write-back to Memory — Write-back the dirty lines in the way of the L2 cache to be degraded. The degradation of a way is performed without adversely affecting software, and software operation is free from any effects except for a slowdown of processing speed.

5.3.2 Internal Registers and Execution Units Reliability Features

To further increase reliability, SPARC64 VI and SPARC64 VII/VII+ processors also provide error protection for registers and execution units. A summary of these capabilities is found in Table 5-3.

SPARC64 VI and SPARC64 VII/VII+ processors include ECC protection for integer architecture registers. When an error occurs, the ECC circuit corrects the error. The floating point architecture registers and other registers are protected by parity bits. Also, the parity prediction circuit, residue check circuit, and other circuits are implemented within execution unit to propagate parity information to output results. In the unlikely event that a parity error is detected, hardware automatically re-executes the instruction to attempt recovery as described below.

Table 5-3. SPARC64 VI and SPARC64 VII/VII+ processor error detection and data protection methods for internal registers and execution units

Type		Error detection method Protection method
Register	Integer register	Parity or SECDED ECC
	Floating-point register	Parity
	PC, PSTATE	Parity
	Computation input-output register	Parity
Execution unit	Addition and subtraction, division, shift, and graphic operation	Parity prediction
	Multiplication	Parity prediction + residue check

5.3.3 Synchronous Update Method and Instruction Retry

SPARC64 VI and SPARC64 VII/VII+ processors employ a synchronous update method. When an error is detected, all the instructions being executed at this time is canceled. Intermediate results before commitment can be discarded, and only results updated by instructions that have been completed without encountering any errors remain in programmable resources. Therefore, not only can the destruction of programmable resources due to errors be prevented, hardware can also perform an instruction retry after error detection. Since stalled instructions can be discarded once and then retried from the beginning, there is a possibility of recovery in the case of a hang-up.

AS shown in Figure 5-5, instruction retry is triggered by an error and is automatically started. A retry is performed instruction-by-instruction to increase the chance of normal execution. When the execution completes normally, the state automatically returns to the normal execution state. During this period, no software intervention is required, and if the instruction retry succeeds, the error does not affect software. An instruction retry is repeated until the number of retry times reaches the threshold. If the threshold is exceeded, the processor logs the source of the error and notifies the operating system of the error occurrence, to request the operating system for the subsequent processing.

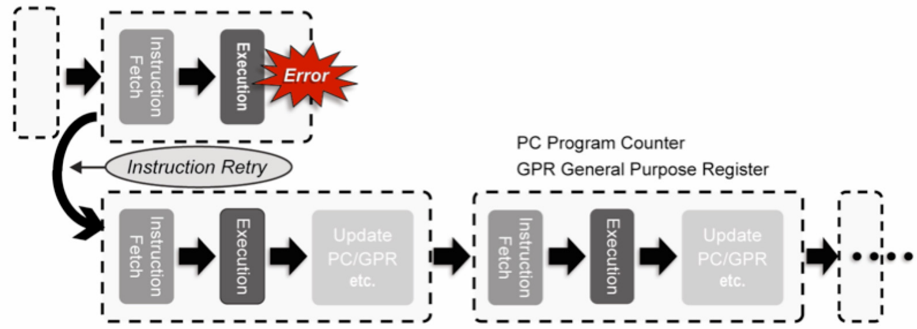

Figure 5-5. The SPARC64 VI and SPARC64 VII/VII+ processors implement an automated instruction retry process to increase availability.

5.3.4 Increased Serviceability

The SPARC64 VI and SPARC64 VII/VII+ processors provide a variety of error checking mechanisms. These processors monitor for errors and send the information to the eXtended System Control Facility (XSCF) if an error occures. On receipt of this notification, the XSCF firmware corrects and analyzes error logs. By taking advantage of SPARC64 VI and SPARC64 VII/VII+ processor error notification features, systems can identify the location and type of a failure quickly and accurately while continuing the operation. As a result, systems provide information useful for preventive maintenance to increase serviceability.

6. I/O Subsystem

A growing reliance on compute systems for every aspect of business operations brings along the need to store and process ever-increasing amounts of information. Powerful I/O subsystems are crucial to effectively moving and manipulating these large data sets. SPARC Enterprise M3000, M4000, M5000, M8000 and M9000 servers deliver exceptional I/O expansion and performance, helping organizations readily scale systems and accommodate evolving data storage needs.

6.1 I/O Subsystem Architecture

The use of PCI technology is key to the performance of the I/O subsystem within SPARC Enterprise M-series. A PCI Express bridge supplies the connection between the main system and components of the I/O unit, such as PCI-X slots, PCI Express slots, and internal drives. The PCI Express bus also supports the connection of external I/O devices by using internal PCI slots or connecting an External I/O Expansion Unit.

In order to facilitate hot-plug of PCI Express and PCI-X adapter cards, SPARC Enterprise M4000, M5000, M8000, and M9000 servers utilize PCI cassettes. PCI cards which support PCI Hot Plug can be mounted by administrators into a PCI cassette and inserted into an internal PCI slot or External I/O Expansion Unit of a running server.

6.1.1 SPARC Enterprise M3000 Server I/O Subsystem

A depiction of the I/O subsystem of the SPARC Enterprise M3000 server is shown in Figure 6-1. A single PCI Express bridge mounted on the motherboard of the SPARC Enterprise M3000 server connects all I/O components to the system controller. The I/O subsystem supports the addition of external I/O devices by providing four PCI Express slots and one external SAS port. The external SAS port can be utilized to connect any SAS tape or strage device.

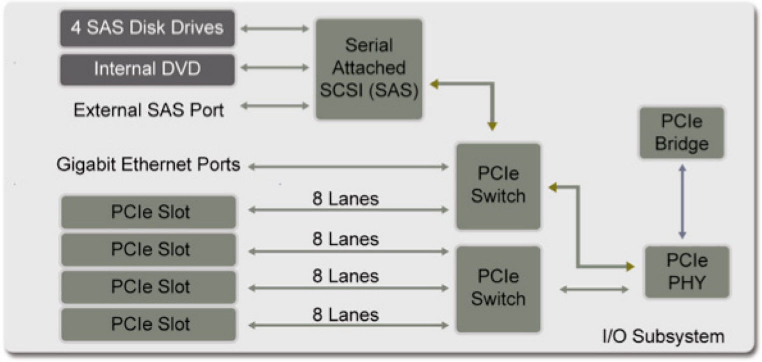

Figure 6-1. SPARC Enterprise M3000 server I/O subsystem architecture.

The SPARC Enterprise M3000 server's on-board SAS controller supports RAID 1 (mirroring) volumes using the Oracle Solaris OS *raidctl utility.*

6.1.2 SPARC Enterprise Midrange M4000 and M5000 Servers I/O Subsystem

The SPARC Enterprise M4000 server supports one IOU, while the SPARC Enterprise M5000 server supports two IOUs. A single PCIe bridge connects each IOU to the system controllers and a PCI Express to PCI-X bridge supports inclusion of on-board PCI-X slots on SPARC Enterprise midrange servers. The single IOU on the SPARC Enterprise M4000 contains four PCI Express slots and one PCI-X slot (Figure 6-2). The two IOUs on the SPARC Enterprise M5000 contain a total of eight PCI Express slots and two PCI-X slots (Figure 6-3). In addition, an External I/O Expansion Unit increases the number of available PCI slots on midrange SPARC Enterprise servers.

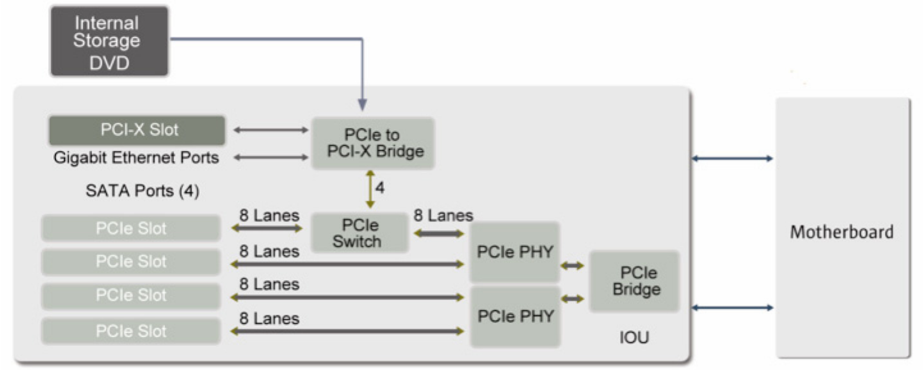

Figure 6-2. SPARC Enterprise M4000 server I/O subsystem architecture.

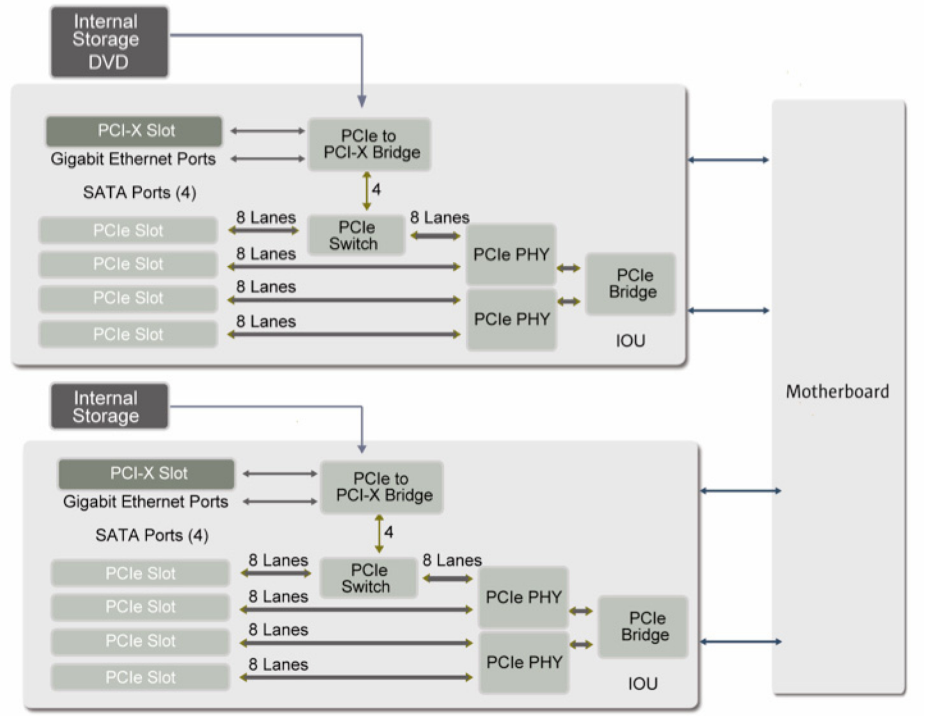

Figure 6-3. SPARC Enterprise M5000 server I/O subsystem architecture.

6.1.3 SPARC Enterprise M8000 and M9000 Servers I/O Subsystem

On the SPARC Enterprise M8000 and M9000 servers, you can use one system board and one IOU in combination. Two PCIe bridges connect the IOU on each system board to a crossbar switch. Each PCI Express bridge also controls communications to four PCI Express slots on the system board (Figure 6-4).

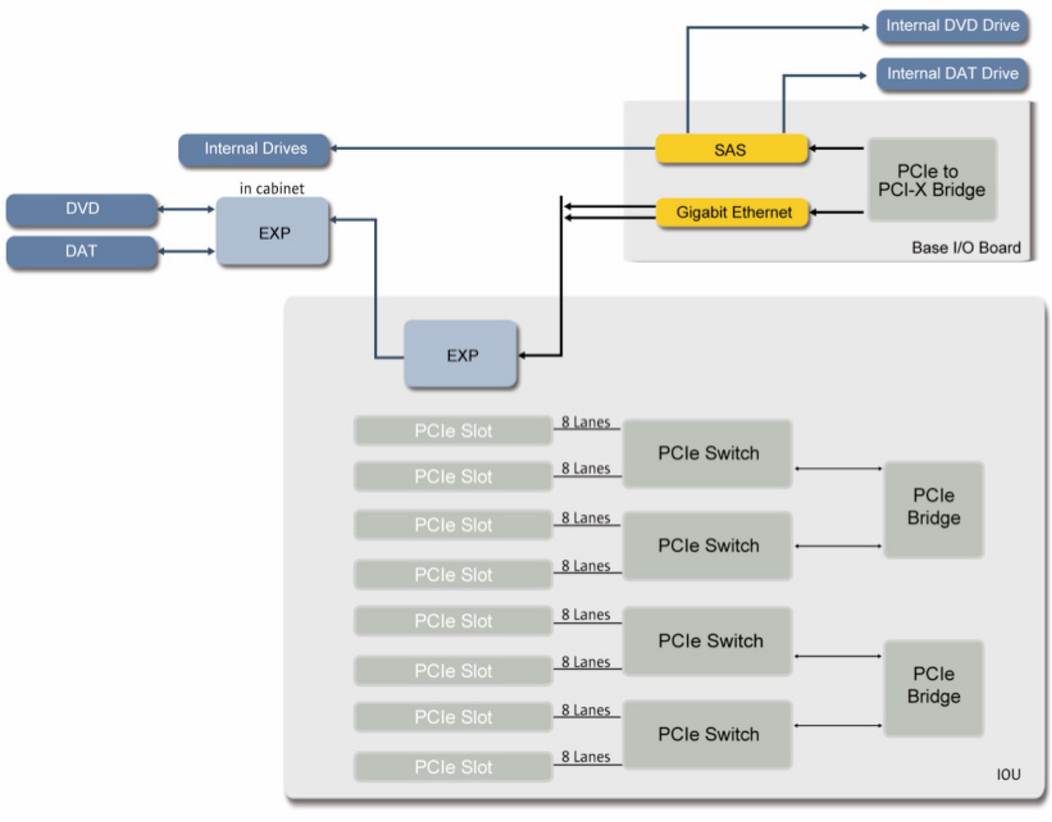

Figure 6-4. SPARC Enterprise M8000 and SPARC Enterprise M9000 I/O subsystem.

A SPARC Enterprise M8000 and SPARC Enterprise M9000 IOU contains eight PCI Express slots with the total number of PCI slots for these servers dependent upon the number of mounted system boards. The maximum number of internal PCI Express slots for SPARC Enterprise high-end servers is listed in Table 6-1. In addition, an External I/O Expansion Unit can be added to a SPARC Enterprise server in order to increase the total number of available PCI slots.

Table 6-1. SPARC Enterprise high-end server internal PCI slot counts.

SPARC Enterprise Model	Maximum Number of Internal PCI Express slots
M8000	32
M9000 (32 CPU configuration)	64
M9000 (64 CPU configuration)	128

6.2 Internal Peripherals

The SPARC Enterprise M3000 server supports one internal DVD drive, four internal Serial Attached SCSI (SAS) 2.5-inch hard disk drives, and 2.5-inch solid state drive (SSD). The SPARC Enterprise M3000 server also supports one external SAS port, which can be connected to any SAS tape or additional file unit. The SAS port offers 2-lanes, supporting up to 6 Gb/second total bandwidth.

While disk and tape devices are directly integrated into SPARC Enterprise M4000 and M5000 servers, an add-on base I/O card provides access to internal devices on SPARC Enterprise M8000 and M9000 servers. SPARC Enterprise M4000, M5000, M8000, and base cabinet of M9000 servers support one internal DVD drive and an optional DAT tape drive. A SPARC Enterprise M9000 with an expansion cabinet supports two internal DVD drives and the option for two internal DAT tape drives. SPARC Enterprise M4000 and M5000 servers support multiple internal SAS 2.5-inch hard disk drives. SPARC Enterprise M8000 and M9000 servers each support multiple internal SAS 2.5-inch hard disk drives (or 2.5-inch SSD).

6.3 External I/O Expansion Unit

SPARC Enterprise M4000, M5000,M8000, and M9000 servers support the attachment of an optional External I/O Expansion Unit to provide additional I/O connectivity. The External I/O Expansion Unit is a four RU rack mountable device which accommodates up to two IOUs with six PCI Express or PCI-X slots. By using PCI cassettes, the external I/O chassis supports active replacement of hot-plug PCI cards.

An I/O Link card mounted in the host provides connectivity to the SPARC Enterprise External I/O Expansion Unit and supports host management control via sideband signals. The I/O link card is available as a low height copper or full height fibre card and includes a single 8-lane PCI Express bus with 4GB/second bandwidth. The architecture of the SPARC Enterprise Expansion unit provides high-throughput I/O performance, supporting maximum data rates for many types of PCI Express cards and bursty traffic from additional PCI Express cards (Figure 6-5).

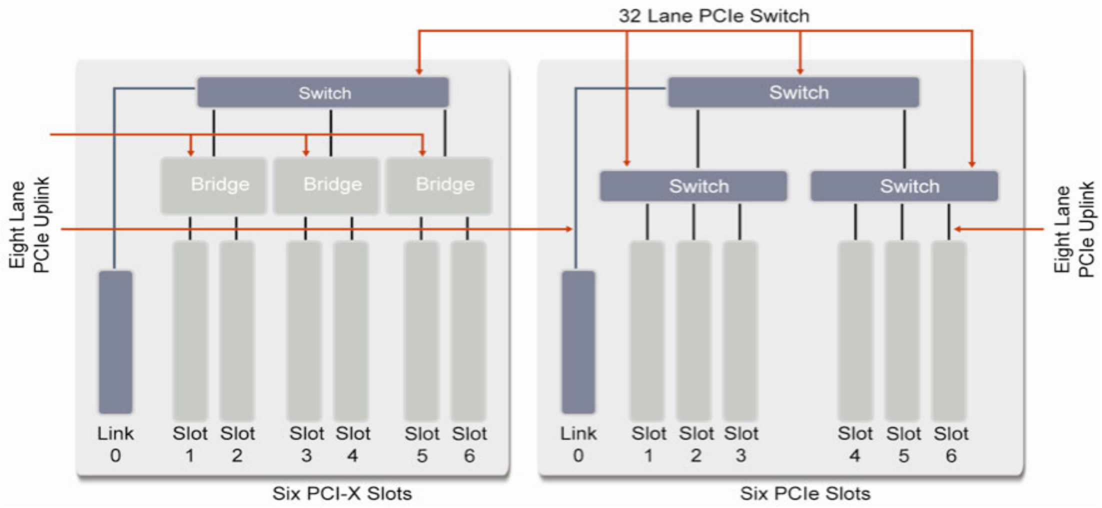

Figure 6-5. External I/O Expansion Unit architecture diagram.

External I/O Expansion Units are added to SPARC Enterprise M4000, M5000, M8000, and M9000 servers by inserting a link card into an internal PCI Express slot and using a cable to connect the link card. The link card options include a low height copper link card kit or full height fibre link card kit. SPARC Enterprise servers support the connection of multiple External I/O Expansion Units as shown in Table 6-2.

Table 6-2. SPARC Enterprise M4000, M5000, M8000, and M9000 servers support massive expansion using the optional External I/O Expansion Unit.

SPARC Enterprise Model	Maximum Number of External I/O Expansion Units	Maximum Number of PCI Slots
M4000	2	25
M5000	4	50
M8000	8	112
M9000 (32 CPU configuration)	16	224
M9000 (64 CPU configuration)	16	288

To ease management of the External I/O Expansion Unit, I/O Manager, software is included with SPARC Enterprise M4000, M5000, M8000, and M9000 servers and provides the following command line accessible functions.

- Discovers External I/O Expansion Units and FRUs when PCI Express slots are powered on
- Collects environmental, voltage, status information
- Logs External I/O Expansion Unit error data

7. 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 systems in the SPARC Enterprise M-series include redundant and hot-swap system components, diagnostic and error recovery features throughout the design, and built-in remote management features. The advanced architecture of these reliable servers fosters high levels of application availability and rapid recovery from many types of hardware faults, simplifying system operation and lowering costs for enterprises.

7.1 Redundant and Hot-Swap Components

Today's IT organizations are challenged by the pace of non-stop business operations. In a networked global economy revenue opportunities remain available around the clock, forcing planned downtime windows to shrink and in some cases disappear entirely. To meet these demands, SPARC Enterprise M-series employ built-in redundant and hot-swap 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.

The SPARC Enterprise M3000, M4000, M5000, M8000, and M9000 servers feature redundant, hot-swap power supply and fan units. In addition to that, the SPARC Enterprise M4000, M5000, M8000, and M9000 servers feature hot-swap I/O cards by PCI hot plugging. Furthermore, the SPARC Enterprise M8000 and M9000 servers feature hot-plug of CMU and IOU, as well as the option to configure multiple CPUs and memory DIMM. Administrators can create redundant internal storage by combining hot-swap disk drives with disk mirroring software. SPARC Enterprise M8000 and M9000 servers also support redundant, hot-swap service processors, and degradable Crossbar Units and SPARC Enterprise M9000 servers also include redundant Clock Control Units. If a fault occurs, these duplicated components can support continued operation. Depending upon the component and type of error, the system may continue to operate in a degraded mode or may reboot — with the failure automatically diagnosed and the relevant component automatically configured out of the system. In addition, hot-swap hardware within the SPARC Enterprise servers speeds service and allows for simplified replacement or addition of components, without a need to stop the system.

7.2 Dynamic Domains

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. Dynamic Domains (= partitioning feature) on SPARC Enterprise M4000, M5000, M8000, and M9000 servers provide IT organizations with the ability to divide a single large system into multiple, fault-isolated servers each running independent instances of the Oracle Solaris operating system. With proper configuration, hardware or software faults in one domain remain isolated and unable to impact the operation of other domains. Each domain within a single server platform can even run a different version of the Oracle Solaris OS, making this technology extremely useful for pre-production testing of new or modified applications. The maximum number of Dynamic Domains by server is itemized in Table 7-1.

Table 7-1. Dynamic Domains limits for SPARC Enterprise M4000, M5000, M8000, and M9000 servers.

Server	Maximum Number of Domains
SPARC Enterprise M4000	2
SPARC Enterprise M5000	4
SPARC Enterprise M8000	16
SPARC Enterprise M9000	24

7.2.1 eXtended System Board (XSB)

Dynamic Domains provide a very effective tool for consolidation and enable the ideal separation of resources. In order to achieve this high level of isolation, previous generations of Fujitsu servers designated entire system boards as the smallest unit assignable to a domain. However, some organizations do not require complete hardware isolation and can benefit from the ability to create a higher number of domains with compute power that more precisely matches current workloads. To meet these needs, the SPARC Enterprise M4000, M5000, M8000, and M9000 servers introduce support for eXtended System Boards (XSB).

To use a physical system board, the hardware resources on the board are divided, reconfigured as eXtended System Boards, and assigned to a Dynamic Domains. There are two types of eXtended System Boards. An XSB that consists of an entire system board is called a Uni-XSB. Alternatively, a system board or motherboard that is logically divided into four parts is called a Quad-XSB. The following diagrams depict the logical division lines within each type of SPARC Enterprise M4000, M5000, M8000, and M9000 servers (Figure 7-1, Figure 7-2, Figure 7-3, Figure 7-4, and Figure 7-5).

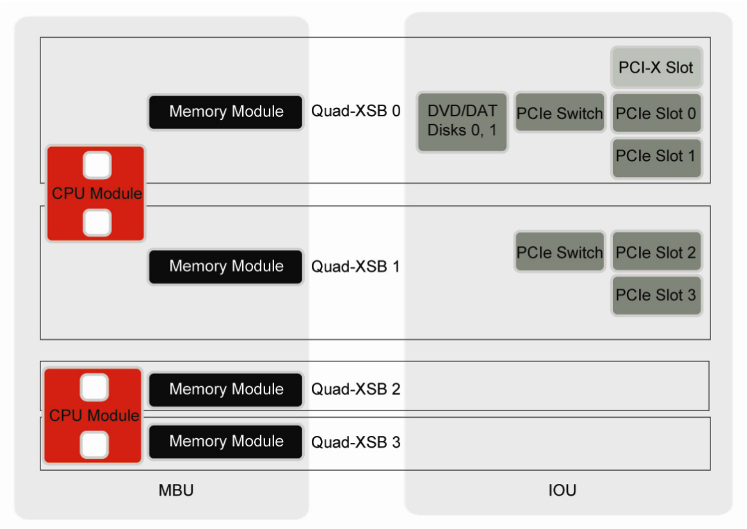

Figure 7-1. SPARC Enterprise M4000 server Quad-XSB configuration.

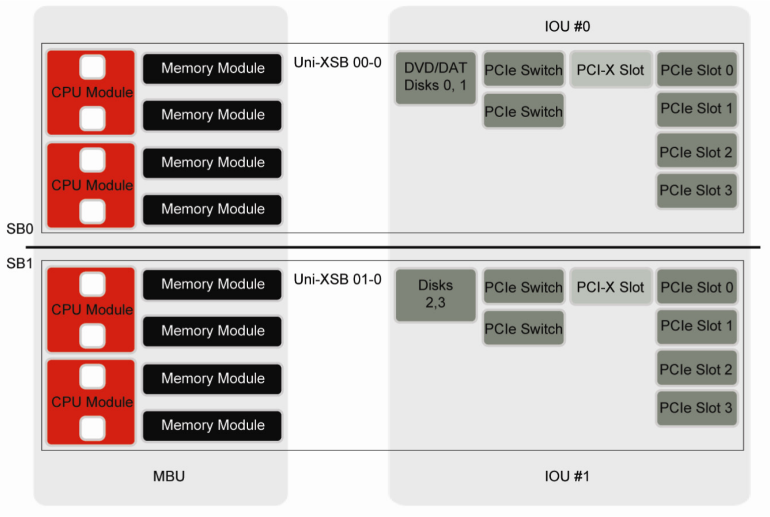

Figure 7-2. SPARC Enterprise M5000 server Uni-XSB configuration.

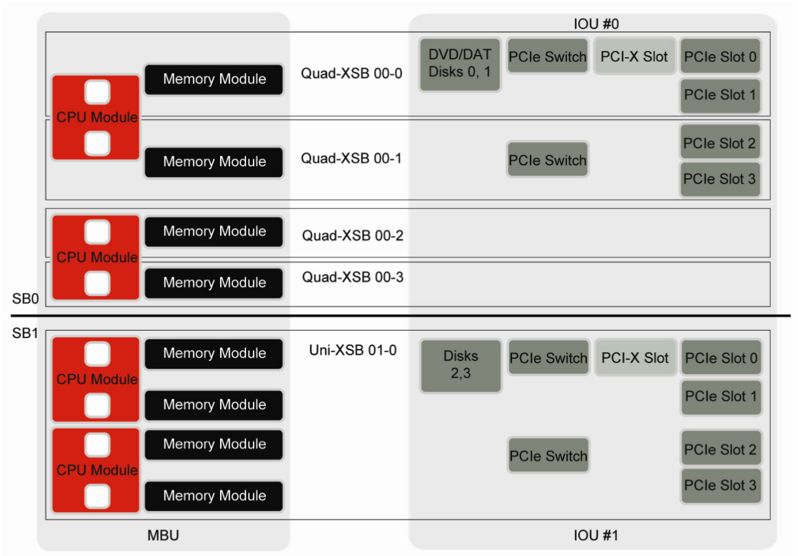

Figure 7-3. SPARC Enterprise M5000 server Uni-XSB and Quad-XSB configuration.

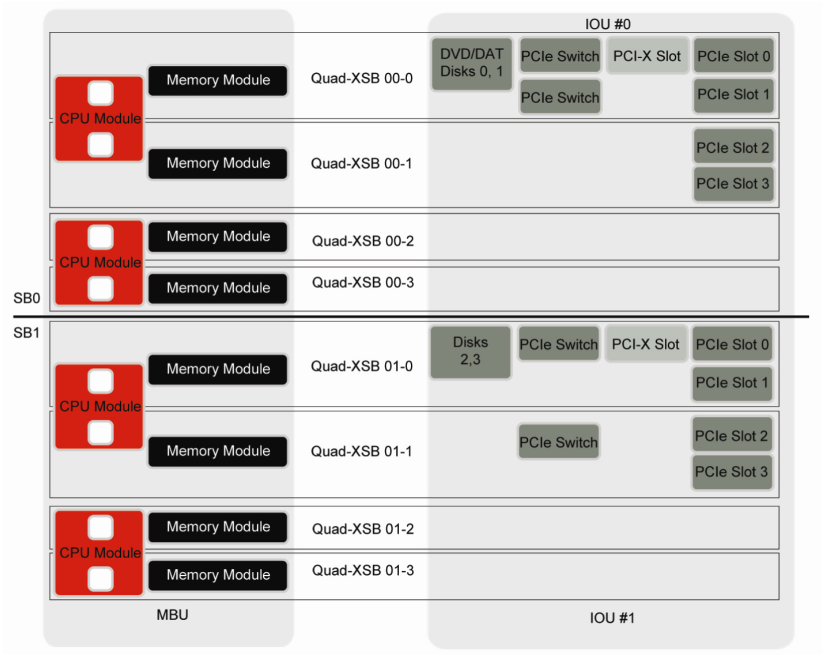

Figure 7-4. SPARC Enterprise M5000 server Quad-XSB configuration.

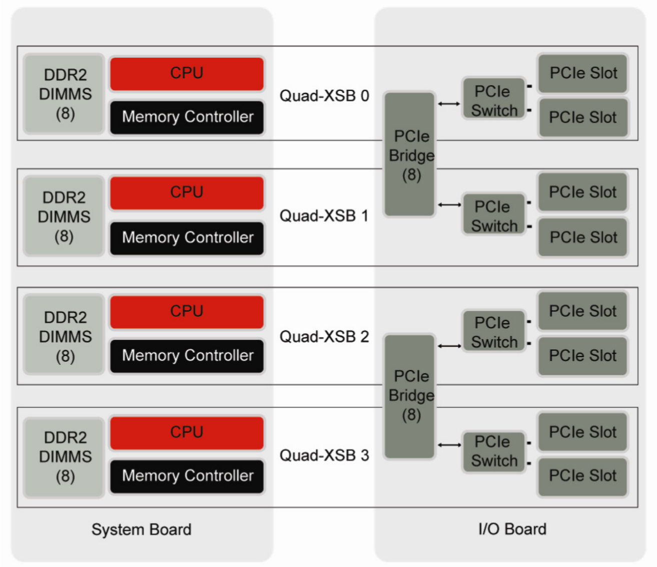

Figure 7-5. SPARC Enterprise M8000 and M9000 servers system board Quad-XSB configuration.

Using eXtended system boards facilitates granular, sub-system board assignment of compute resources to Dynamic Domains. A Dynamic Domains can consist of any combination of Uni-XSBs and Quad-XSBs, providing enterprises the ability to perform sophisticated asset allocation. Determining the exact number and type of XSBs for inclusion in a domain requires balancing the need for fault isolation against the desire to maximize resource utilization. In additions to XSBs, DVD and DAT devices connected to an I/O unit are also assignable to Dynamic Domains. By using Dynamic Domains and XSBs, enterprises can better optimize the use of hardware resources while still providing isolated and secure data and programs to customers.

7.3 Mixed CPU Configurations

Support for mixing SPARC64 VI and SPARC64 VII/VII+ processors within SPARC Enterprise M4000, M5000, M8000, and M9000 server configurations provides an added level of investment protection and further extends solution flexibility. As shown in Figure 7-6, SPARC64 VI and SPARC64 VII/VII+ processors can co-exist within physical system boards, individual Dynamic Domains.

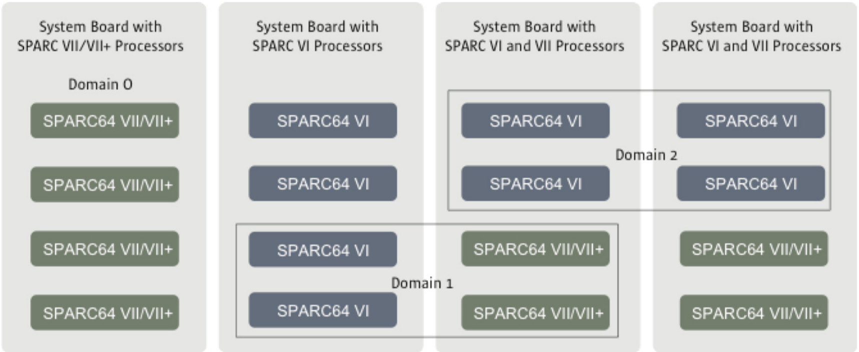

Figure 7-6. SPARC Enterprise M4000, M5000, M8000, and M9000 servers support mixing SPARC64 VI and SPARC64 VII/VII+ processors within system configurations.

The SPARC64 VII+ processors for the M4000 and M5000 come with 11 MB of L2$, while the SPARC64 VII+ for the M8000 and M9000 come with 12 MB of L2$. To achieve this L2$ capacity, two conditions must be met:

1. All four processors on the system board must be SPARC64 VII+ processor. None of the four can be either SPARC64 VI or SPARC64 VII processors.
2. The motherboard on the SPARC Enterprise M4000 and M5000 servers must be MBU_B or later, and the CMU on the SPARC Enterprise M8000 and M9000 servers must be CMU_C or later.

The reason for a new MBU or CMU board is because the SC ASIC on previous versions could only access a maximum of 6 MB of L2$. The newer version of the boards have an updated SC ASIC that can access up to 12 MB of L2$. The SPARC64 VII+ processors can be installed on earlier versions of the boards, however, their L2$ will be cut in half (5.5 MB of L2$ on the SPARC Enterprise M4000 and M5000 servers and 6 MB of L2$ on the SPARC Enterprise M8000 and M9000 servers). If a SPARC64 VI or VII processors is installed on the newer MBU_B or CMU_C boards where two SPARC64 VII+ processors are already installed, then the L2$ of the SPARC64 VII+ processors will be reduced in half (5.5 MB of L2$ on the SPARC Enterprise M4000 and M5000 servers and 6 MB of L2$ on the SPARC Enterprise M8000 and M9000 servers). These rules only apply to the physical installation of the processors, not the logical assignment of XSBs to domains.

7.4 Dynamic Reconfiguration

Dynamic Reconfiguration technology provides added value to Dynamic Domains by providing administrators with the ability to shift resources without taking the SPARC Enterprise M4000, M5000, M8000, and M9000 servers offline. This technology helps administrators perform maintenance, live upgrades, and physical changes to system hardware resources, while the server continues to execute applications. Dynamic Reconfiguration even supports multiple simultaneous changes to hardware configurations without interrupting critical systems.

The ability to remove and add components such as CPUs, memory, and I/O subsystems from a running system helps reduce system downtime. Using Dynamic Reconfiguration simplifies maintenance and upgrades by eliminating the need for system reboots after hardware configuration changes.

7.5 Advanced Reliability Features

Advanced reliability features included within the components of SPARC Enterprise M4000, M5000, M8000, and M9000 servers increase the overall stability of these platforms. For example, SPARC Enterprise M4000/M5000/M8000/M9000 servers include multiple system controllers, and high-end servers include degradable crossbar switches to provide redundancy within the system bus. Reduced component count and complexity within the server architecture contributes to reliability. In addition, advanced CPU integration and guaranteed data path integrity provide for autonomous error recovery by the SPARC64 VI processor and SPARC64 VII/VII+ processors, reducing the time to initiate corrective action and subsequently increasing uptime.

Oracle Solaris Predictive Self Healing software further enhances the reliability of SPARC Enterprise M-series. The implementation of Oracle Solaris Predictive Self Healing software for SPARC Enterprise M-series provides constant monitoring of CPUs and memory. Depending upon the nature of the error, persistent CPU soft errors can be resolved by automatically offlining either a thread, core, or entire CPU. In addition, the memory page retirement function supports the ability to take memory pages offline proactively in response to multiple corrections to data access for a specific memory DIMM.

7.6 Error Detection, Diagnosis, and Recovery

SPARC Enterprise M-series 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. Ultimately, the following features work together to raise application availability.

- End-to-end data protection detects and corrects errors throughout the system, ensuring complete data integrity.
- State-of-the-art fault isolation helps these servers isolate errors within component boundaries and offline only the relevant chips instead of the entire component. Isolating errors down to the chip improves stability and provides continued availability of maximum compute power. This feature applies to CPUs, memory access controllers, crossbar ASICs, system controllers, and I/O ASICs.
- Constant environmental monitoring provides a historical log of pertinent environmental and error conditions.
- The host watchdog feature periodically checks for operation of software, including the domain operating system. This feature also uses the XSCF firmware to trigger error notification and recovery functions.
- Dynamic CPU resource degradation provides processor fault detection, isolation, and recovery. This feature dynamically reallocates CPU resources into an operational system using Dynamic Reconfiguration without interrupting the applications that are running. All servers support dynamic CPU degradation; however, dynamic replacement is supported on the SPARC Enterprise M8000 and M9000 servers. And the Dynamic Reconfiguration process to the domain is supported on the SPARC Enterprise M4000, M5000, M8000 and M9000 servers.
- Periodic component status checks are performed to determine the status of many system devices to 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.

8. 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 requirements for remote management of systems. One of the many benefits of Fujitsu servers is the support for *lights-out* datacenters, letting expensive support staff to work in any location with network access. The design of SPARC Enterprise M3000, M4000, M5000, M8000, and M9000, combine with a powerful eXtended System Control Facility (XSCF), XSCF Control Package, and Fujitsu's system management software to help administrators to remotely execute and control nearly any task that does not involve physical access to hardware. These management tools and remote functions lower administrative burden, saving organizations time and reducing operational expenses.

8.1 Extended System Control Facility

The eXtended System Control Facility provides the heart of remote monitoring and management capabilities in SPARC Enterprise M-series. The XSCF consists of a dedicated processor that is independent of the server system and runs the XSCF Control Package. The Domain to Service Processor Communication Protocol (DSCP) is used for communication between the XSCF and the server. The DSCP protocol runs on a private TCP/IP-based or PPP-based communication link between the service processor and each domain. While input power is supplied to the server, the XSCF constantly monitors the system even if all domains are inactive.

The XSCF regularly monitors the environmental sensors, provides advance warning of potential error conditions, and executes proactive system maintenance procedures as necessary. For example, the XSCF can 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 domains, as well as the platform itself.

Using a network or serial connection to the XSCF, operators 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.

8.1.1 Redundant XSCF

On SPARC Enterprise M8000 and M9000 servers, one XSCF is configured as active and the other is configured as a standby. The XSCF network between the two service processors facilitate the exchange of system management information. In case of failover, the service processors are already synchronized and ready to change roles.

8.1.2 DSCP Network

The Domain to Service Processor Communication Protocol service provides a secure TCP/IP and PPP-based communications link between the service processor and each domain. Without this link, the XSCF cannot communicate with the domains. The service processor requires one IP address dedicated to the DSCP service on its side of the link, and one IP address on each domain's side of the link. In a system with more than one XSCF, the standby XSCF does not communicate with the domains. In the event of a failover of the XSCF, the newly active XSCF assumes the IP address of the failed-over service processor.

8.1.3 XSCF Control Package

The XSCF Control Package helps users to control and monitor SPARC Enterprise M3000, M4000, M5000, M8000, and M9000 platforms and individual Dynamic Domains quickly and effectively. The XSCF Control Package provides a command line interface (CLI) and Web browser user interface that gives administrators and operators access to system controller functionality. Password-protected accounts with specific administration capabilities also provide system security for domain consoles. Communication between the XSCF and individual domains uses an encrypted connection based on Secure Shell (SSH) and Secure Socket Layer (SSL), supporting secure, remote execution of commands provided with the XSCF Control Package.

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

- Execution of Dynamic Reconfiguration tasks to logically attach or detach installed system boards from the operating system while the domain continues to run applications without interruption.
- Domain administration which consist of creating logical system boards comprised of Uni-XSB and Quad-XSB units.
- Audit administration includes the logging of interactions between the XSCF and the domains.
- Monitor and control of power to the components in all SPARC Enterprise M-series.
- Interpretation of hardware information presented, and notification of impending problems such as high temperatures or power supply problems, as well as access to the system administration interface.
- Integration with the Fault Management Architecture of the Oracle Solaris to improve availability through accurate fault diagnosis and predictive fault analysis.

- Execution and monitoring of diagnostic programs, such as the Open Boot Prom (OBP) and power-on self-test (POST).
- Execution of Fujitsu Capacity on Demand operations which provide the ability to stage and then later activate additional processing resources.
- Monitoring of the dual XSCF configuration on SPARC Enterprise M8000 and SPARC Enterprise M9000 for failure and performing an automatic failover if needed.

8.1.4 Role Based System Management

The XSCF Control Package facilitates the independent administration of several autonomous domains by different system administrators and operators — each cooperating within a single SPARC Enterprise 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, domains, or physical components within a domain. In addition, a user can possess multiple, different privileges on any number of domains.

8.1.5 Enhanced Support Facility

Enhanced Support Facility is specific software that improves the operation management and the maintainability of SPARC Enterprise M-series. Working in combination with XSCF, server configuration, status and error messages can all be displayed. If a problem occurs, the information reported to XSCF ensures the status, of disks, power, PCI cards and OS, is always monitored. It also enables you to display other system information including batch collections, /etc/system file settings, server power on/off scheduling and disk hot swap procedures.

8.1.6 Systemwalker Centric Manager

Centric Manager lets you follow the system operation lifecycle (installation/setup, monitoring, fault recovery, assessment), making it possible for you to create highly-reliable systems. It reduces the workload required for operations management and provides high-value functions for life-cycle tasks. These include the remote distribution of software resources, central monitoring of systems and networks, and prompt resolution of problems from any location. It performs integrated management, operational process standardization (ITIL), while enabling security control of the latest business IT technology such as multi-platform and intranet/Internet environments.

9. The Oracle Solaris 10

With mission-critical business objectives on the line, enterprises need a robust operating environment with the ability to optimize the performance, availability, security, and utilization of hardware assets. In a class by itself, the Oracle Solaris 10 offers many innovative technologies to help IT organizations improve operations and realize the full potential of SPARC Enterprise M-series.

9.1 Observability and Performance

Organizations need to make effective use of the power of hardware platforms. Oracle Solaris facilities near linear scalability as processor count increases. In addition, Oracle Solaris supports memory addressability that reaches well beyond the physical memory limits of even Fujitsu's largest server. The following advanced features of Oracle Solaris provide IT organizations with the ability to identify potential software tuning opportunities and maximize raw system throughput.

- Oracle Solaris Dynamic Tracing framework (DTrace) is a powerful tool that provides a true, system-level view of application and kernel activities, even those running in a Java virtual machine (JVM). DTrace software safely instruments the running operating system kernel and active applications without rebooting the kernel or recompiling — or even restarting — software. By using this feature, administrators can view accurate and concise information in real time and highlight patterns and trends in application execution. The dynamic instrumentation provided by DTrace enables organizations to reduce the time to diagnose problems from days and weeks to minutes and hours, resulting in faster data-driven fixes.
- The highly scalable, optimized TCP/IP stack in Oracle Solaris lowers overhead by reducing the number of instructions required to process packets. This technology also provides support for large numbers of connections and helps server network throughput to grow linearly with the number of CPUs and network interface cards (NICs). By taking advantage of the Oracle Solaris 10 network stack, organizations can significantly improve application efficiency and performance.
- The memory handling system of Oracle Solaris 10 provides multiple page size support in order to help applications to access virtual memory more efficiently, improving performance for applications that use large memory intensively. In addition, Oracle Solaris 10 Memory Placement Optimization (MPO) works to ensure that data is stored in memory as close as possible to the processors that accesses it while still

maintaining enough balance within the system. MPO can boost performance in business workloads by as much as 20 % and as much as 50 % in some High Performance Computing workloads.

- The Oracle Solaris multithreaded execution model plays an important role in helping Fujitsu servers to deliver scalable performance. Improvements to the threading capabilities in the Oracle Solaris occur with every release, resulting in performance and stability improvements for existing applications without recompiliation.

9.2 Availability

The ability to rapidly diagnose, isolate, and recover from hardware and application faults is paramount for meeting the needs of non-stop business operations. A long standing features of Oracle Solaris provide for system self-healing. For example, the kernel memory scrubber constantly scans physical memory, correcting any single-bit errors in order to reduce the likelihood of those problems turning into un-correctable double-bit errors. Oracle Solaris 10 takes a big leap forward in self-healing with the introduction of Oracle Solaris Fault Manager and Oracle Solaris Service Manager technology. With this software, business-critical applications and essential system services can continue uninterrupted in the event of software failures, major hardware component breakdowns, and software misconfiguration problems.

Oracle Solaris Fault Manager software reduces complexity by automatically diagnosing faults in the system and initiating self-healing actions to help prevent service interruptions. The Oracle Solaris Fault Manager diagnosis engine produces a fault diagnosis once discernible patterns are observed from a stream of incoming errors. Following error identification, the Oracle Solaris Fault Manager provides information to agents that know how to respond to specific faults. Problem components can be configured out of a system before a failure occurs — and in the event of a failure, this feature initiates automatic recovery and application re-start. For example, an agent designed to respond to a memory error might determine the memory addresses affected by a specific chip failure and remove the affected locations from the available memory pool.

- Oracle Solaris Service Manager software converts the core set of services packaged with the operating system into first-class objects that administrators can manipulate with a consistent set of administration commands. Using Oracle Solaris Service Manager, administrators can take actions on services including start, stop, restart, enable, disable, view status, and snapshot. Service snapshots save a service’s complete configuration, giving administrators a way to roll back any erroneous changes. Snapshots are taken automatically whenever a service starts to help reduce risk by guarding against erroneous errors. The Oracle Solaris Service Manager is integrated with Oracle Solaris Fault Manager. When a low-level fault is found to impact a higher-level component of a running service, Oracle Solaris Fault Manager can direct Oracle Solaris Service Manager to take appropriate action.

In addition to handling error conditions, efficiently managing planned downtime greatly enhances availability levels. Tools included with the Oracle Solaris, such as Oracle Solaris Flash and Oracle Solaris Live Upgrade software, can help enterprises achieve more rapid and consistent installation of software, upgrades, and patches, leading to improved uptime.

- The Oracle Solaris Flash facility enables IT organizations to quickly install and update systems with an operating system configuration tailored to enterprise needs. This technology provides tools to system administrators for building custom rapid-install images—including applications, patches, and parameters—that can be installed at a data rate close to the full speed of the hardware.
- The Oracle Solaris Live Upgrade facility provides mechanisms to upgrade and manage multiple on-disk instances of Oracle Solaris. This technology provides system administrators to install a new operating system on a running production system without taking it offline, with the only downtime for the application being the time necessary to reboot the new configuration.

9.3 Security

Today's increasingly connected systems create benefits and challenges. While the global network offers greater revenue opportunities, enterprises must pay close attention to security concerns. The most secure operating system on the planet, the Oracle Solaris 10 provides features previously only found in military-grade Trusted Oracle Solaris. These capabilities support the strong controls required by governments and financial institutions but also benefit all enterprises focused on security concerns and requirements for auditing capabilities.

- User Rights Management and Process Rights Management work in conjunction with Oracle Solaris Container virtualization technology to let multiple applications to securely share the same domain. Security risks are reduced by granting users and applications only the minimum capabilities needed to perform assigned duties. Best yet, unlike other solutions on the market, no application changes are required to take advantage of these security enhancements.
- Oracle Solaris Trusted Extensions extend the existing Oracle Solaris 10 security policy with labeling features previously only available in highly specialized operating systems or appliances. These extensions deliver true multilevel security within a commercial-grade operating system, beneficial to civilian organizations with specific regulatory or information protection requirements.
- Core to the Oracle Solaris 10 OS are features which fortify platforms against compromise. Firewall protection technology included within the Oracle Solaris 10 distribution protects individual systems against attack. In addition, file integrity checking and digitally signed binaries within Oracle Solaris 10 help administrators to verify that platforms remain untouched by hackers. Secure remote access capabilities also increase security by centralizing the administration of system access across multiple operating systems.

9.4 Virtualization and Resource Management

The economic need to maximize the use of every IT asset often necessitates consolidating multiple applications onto single server platforms. Virtualization techniques enhance consolidation strategies one step further by helping organizations to create administrative and resource boundaries between applications within each domain on a server. Oracle Solaris Containers technology provides a breakthrough approach to virtualization and software partitioning, supporting the creation of many private execution environments within a single instance of the Oracle Solaris OS. Using this technology, IT organizations can quickly harness and provision idle compute power into a secure, isolated runtime environment for a new deployments without increasing the number of operating system instances to manage. In addition, hosting applications within individual Oracle Solaris Containers provides administrators the ability to exert fine-grained control over rights and resources within a consolidated server.

In addition, Oracle Solaris Resource Manager software supports the allocation of computing resources within Oracle Solaris Containers and among individual tasks and users in a structured, policy-driven fashion. Using Oracle Solaris resource management facilities to proactively allocate, control, and monitor system resources —such as CPU time, processes, virtual memory, connect time, and logins— on a fine-grained basis helps organizations obtain more predictable service levels. As business needs change, Oracle Solaris Resource Manager software helps enterprises to regularly set new priorities for the use of compute resources. By taking advantage of Oracle Solaris Containers and Solaris Resource Manager software, organizations can improve resource utilization, reduce downtime, and lower solution costs.

10. Summary

To support demands for greater levels of scalability, reliability, and manageability in the datacenter, infrastructures need to provide ever-increasing performance and capacity while becoming simpler to deploy, adjust, and manage. SPARC Enterprise M-series outfitted with SPARC64 VI processors or SPARC64 VI and SPARC64 VII/VII+ processors, large memory capacity, an inherently reliable architecture, and an eXtended system control facility deliver new levels of power, availability, and ease-of-use to enterprises. The sophisticated resource control provided by Dynamic Domains, eXtended System Boards, and Dynamic Reconfiguration further increase the value of these servers by helping enterprises to optimize the use of these hardware assets. By deploying fast, scalable SPARC Enterprise M-series from Fujitsu, organizations gain extraordinary power and flexibility — a strategic asset in the quest to gain a competitive business advantage.

For More Information

To learn more about innovative Fujitsu products and the benefits of Fujitsu SPARC Enterprise servers and the Oracle Solaris, contact a Fujitsu sales representative or consult the Fujitsu Web site listed in the Table below.

Related Web sites

Web Site URL	Description
http://www.fujitsu.com/sparcenterprise/	Fujitsu SPARC Enterprise servers
http://www.fujitsu.com/global/services/computing/	Fujitsu Computing products
http://www.fujitsu.com/global/	Fujitsu Global portal

(Reference material)

Processor for a UNIX Server "SPARC64 V", Aug 2004, Fujitsu Limited