FUJITSU

SPARC ENTERPRISE T5440 SERVER ARCHITECTURE

Unleashing UltraSPARC T2 Plus Processors
with Innovative Multi-core Multi-thread Technology

White Paper

July 2009

TABLE OF CONTENTS

Chapter 1

The UltraSPARC T2 Plus Processors

The UltraSPARC T2 and UltraSPARC T2 Plus processors are the industry's first system on a chip (SoC), supplying the most cores and threads of any general-purpose processor available, and integrating all key system functions.

The World's First Massively Threaded System on a Chip (SoC)

The UltraSPARC T2 Plus processor eliminates the need for expensive custom hardware and software development by integrating computing, security, and I/O on to a single chip. Binary compatible with earlier UltraSPARC processors, no other processor delivers so much performance in so little space and with such small power requirements letting organizations rapidly scale the delivery of new network services with maximum efficiency and predictability. The UltraSPARC T2 Plus processor is shown in Figure 1.

Figure 1. The UltraSPARC T2 Plus processor with CoolThreads technology

Table 1 provides a comparison between the UltraSPARC T1, UltraSPARC T2, and UltraSPARC T2 Plus processors.

Table 1. UltraSPARC T1 ,T2 and T2 Plus processor features

Feature	UltraSPARC T1 Processor	UltraSPARC T2 processor	UltraSPARC T2 Plus processor
Cores per processor	Up to 8	Up to 8	Up to 8
Threads per core	4	8	8
Threads per processor	32	64	64
Hypervisor	Yes	Yes	Yes
Sockets Supported	1	1	2 or 4[a]
Memory	4 memory controllers, 4 DIMMs per controller	4 memory controllers, up to 16 FB-DIMMs	2 memory controllers, up to 16 or 32 FB-DIMMs
Caches	16 KB instruction cache, 8 KB data cache, 3 MB L2 cache (4 banks, 12-way associative)	16 KB instruction cache, 8 KB data cache, 4 MB L2 cache (8 banks, 16-way associative)	16 KB instruction cache, 8 KB data cache, 4 MB L2 cache (8 banks, 16-way associative)
Technology	9-layer Cu metal, CMOS process, 90 nm technology	65 nm technology	65 nm technology
Floating point	1 FPU per chip	1 FPU per core, 8 FPUs per chip	1 FPU per core, 8 FPUs per chip
Integer resources	Single execution unit	2 integer execution units per core	2 integer execution units per core
Cryptography	Accelerated modular arithmetic operations (RSA)	Stream processing unit per core, support for the 10 most popular ciphers	Stream processing unit per core, support for the 10 most popular ciphers
Additional on-chip resources	-	Dual 10 Gb Ethernet interfaces, PCI Express interface (x8)	PCI Express interface (x8), Coherency logic and links (4.8 Gb/secong)

a. An External Coherency Hub is used for four-socket implementations.

Taking Chip Multithreaded Design to the Next Level

When the next-generation of multi-core multi-thread processors set to be designed, they started with key goals in mind:

• Increasing computational capabilities to meet the growing demand from Web applications by providing twice the throughput of the UltraSPARC T1 processor

• Supporting larger and more diverse workloads with greater floating point performance

• Powering faster networking to serve new network-intensive content

• Providing end-to-end datacenter encryption

• Increasing service levels and reducing downtime

• Improving datacenter capacities while reducing costs

Multi-core multi-thread architecture is ultimately very flexible, allowing different modular combinations of processors, cores, and integrated components. The considerations listed above drove an internal engineering effort that compared

different approaches with regard to making improvements on the successful UltraSPARC T1 architecture. For example, simply increasing the number of cores would have gained additional throughput, but would have resulted in consuming extra die area, leaving no room for integrated components such as floating point processors.

The final UltraSPARC T2 and UltraSPARC T2 Plus processor designs recognize that memory latency is truly the bottleneck to improving performance. By increasing the number of threads supported by each core, and by further increasing network bandwidth, these processors are able provide approximately twice the throughput of the UltraSPARC T1 processor. Figure 2 provides a simplified high-level illustration of the thread model supported by an eight-core UltraSPARC T2 Plus processor.

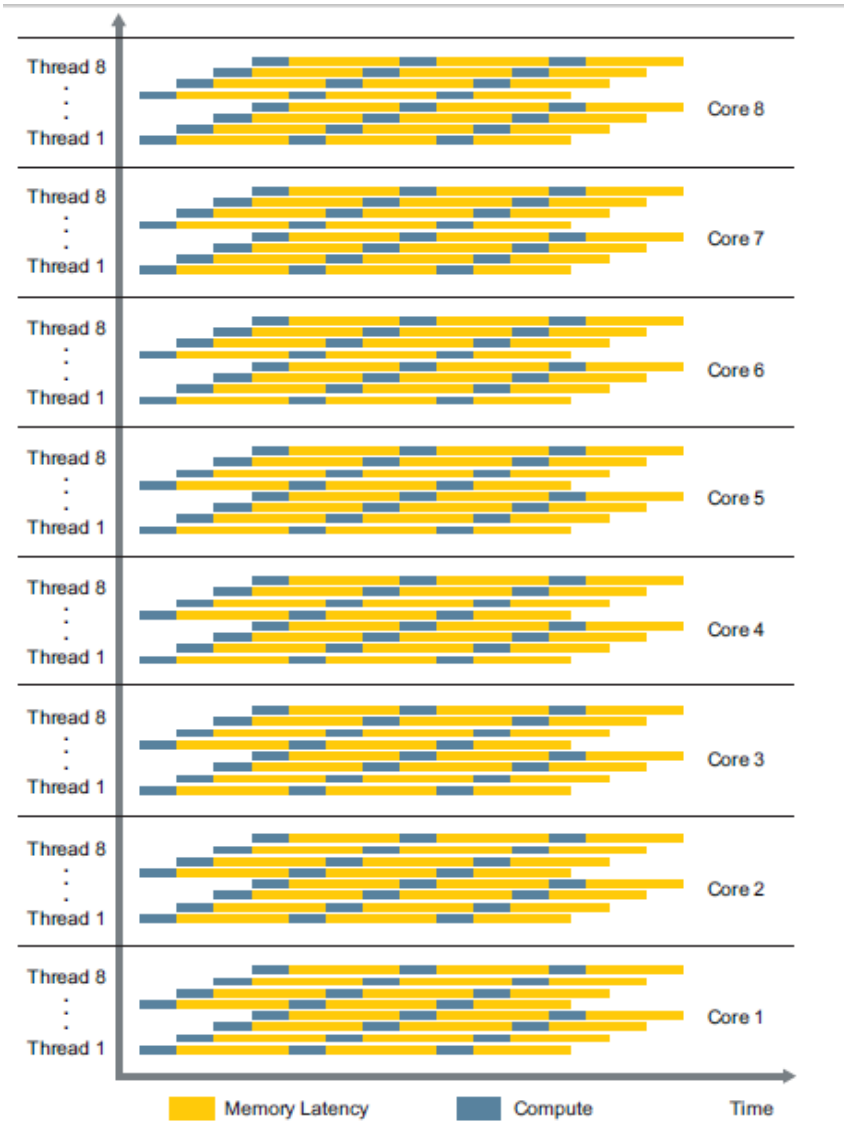


Figure 2. A single eight-core UltraSPARC T2 or UltraSPARC T2 Plus processor supports up to 64 threads, with up to two threads running in each core simultaneously

Each UltraSPARC T2 and UltraSPARC T2 Plus processor provides up to eight cores, with each core able to switch between up to eight threads (64 threads per processor). In addition, each core provides two integer execution units, so that a single UltraSPARC core is capable of executing two threads at a time.

UltraSPARC T2 Plus Processor Architecture

The UltraSPARC T2 Plus processor extend Throughput Computing initiative with an elegant and robust architecture that delivers real performance to applications through multisocket implementations. A high-level block diagram of the UltraSPARC T2 Plus processor is shown in Figure 3.

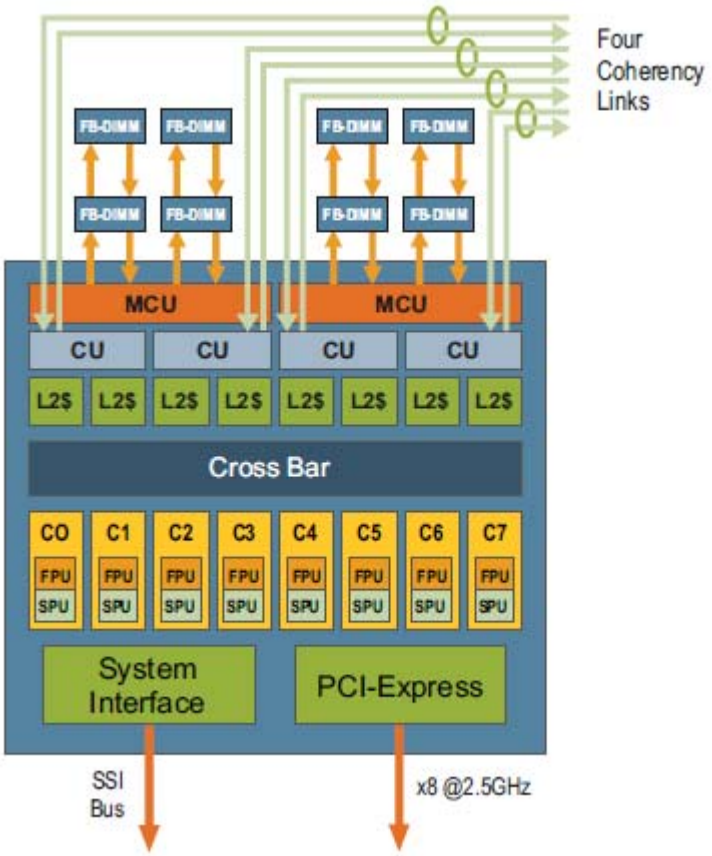


Figure 3. The UltraSPARC T2 Plus processor provides four Coherency Links to connect to up to four other processors

The eight cores on the UltraSPARC T2 Plus processor are interconnected with a full on-chip non-blocking 8 x 9 crossbar switch. The crossbar connects each core to the eight banks of L2 cache, and to the system interface unit for IO. The crossbar provides approximately 300 GB/second of bandwidth and supports 8-byte writes from a core to a bank and 16-byte reads from a bank to a core. The system interface unit connects networking and I/O directly to memory through the individual cache banks. Using FB-DIMM memory supports dedicated northbound and southbound lanes to and from the caches to accelerate performance and reduce latency. This

approach provides higher bandwidth than with DDR2 memory.

Each core provides its own fully-pipelined floating point and graphics unit (FGU or FPU), as well as a stream processing unit (SPU). The FGUs greatly enhance floating point performance over that of the UltraSPARC T1, while the SPUs provide wire-speed cryptographic acceleration with over 10 popular ciphers supported, including DES, 3DES, AES, RC4, SHA-1, SHA-256, MD5, RSA to 2048 key, ECC, and CRC32. Embedding hardware cryptographic acceleration for these ciphers allows end-to-end encryption with no penalty in either performance or cost.

The UltraSPARC T2 Plus architecture also provides four Coherency Units (CUs) to support multisocket system implementations. Four Coherence Channels (or coherence links) are provided — one associated with each Coherency Unit. These links run a cache coherence (snoopy) protocol over an FB-DIMM interface to provide up to 4.8 Gigatransfers per port, providing 204 Gb per second in each diretion. The memory link speed of the UltraSPARC T2 Plus processor was increased to 4.8 Gbps over the 4.0 Gbps of the UltraSPARC T2 processor.

The UltraSPARC T2 Plus processor can support both two- and four-socket implementations. Dual-socket UltraSPARC T2 Plus implementations — such as those provided by SPARC Enterprise T5140 and T5240 servers — interconnect the processors' four coherence links, and require no additional circuity. Four-socket implementations require additional logic in the form of four External Coherency Hubs. A typical four-socket implementation is shown in Figure 4, with more information on the details of the External Coherency Hub provided later in this document.

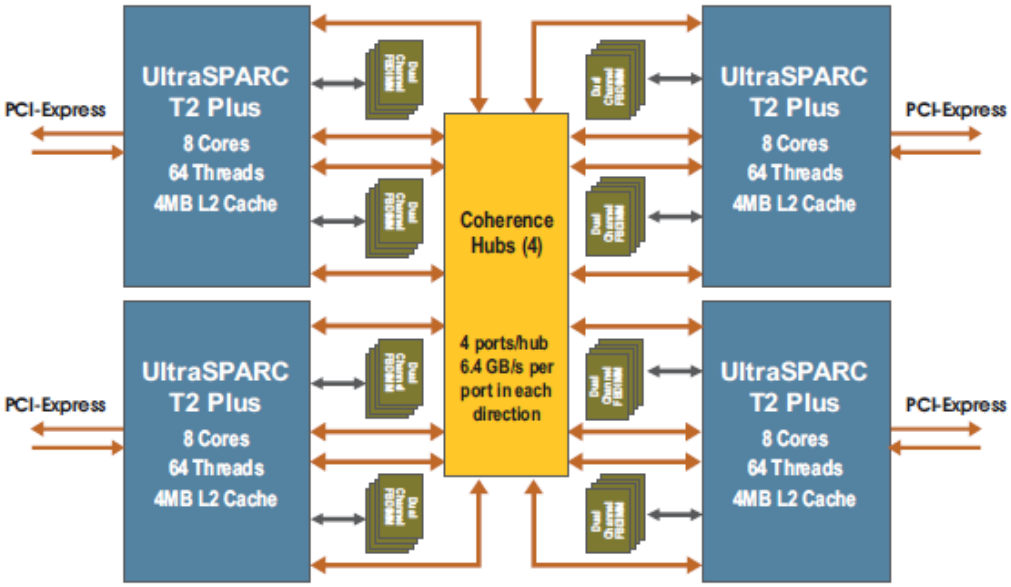


Figure 4. Quad-socket UltraSPARC T2 Plus configuration in the SPARC Enterprise T5440 server

Core Architecture and Pipelines

Both the UltraSPARC T2 and UltraSPARC T2 Plus processors share the same core design. Figure 5 provides a block-level diagram representing a single UltraSPARC cores on the UltraSPARC T2 processor (up to eight are supported per processor).

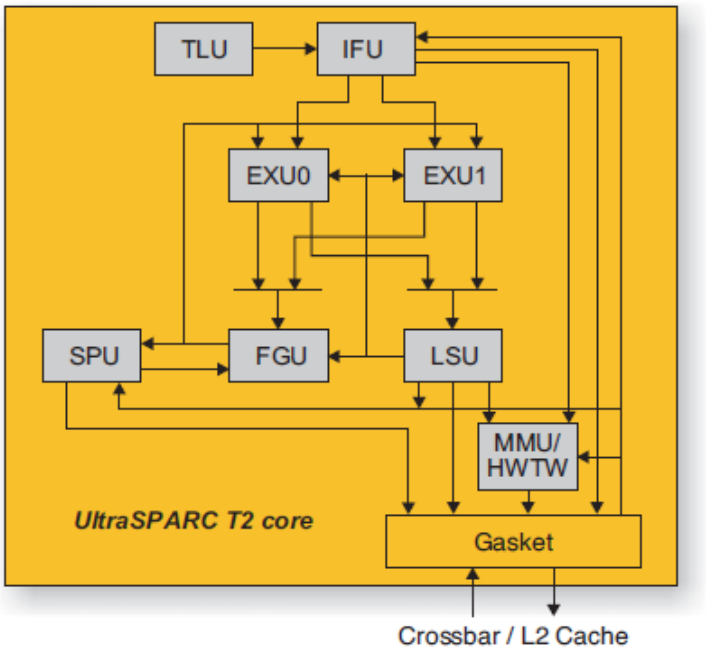


Figure5. UltraSPARC T2 Plus core block diagram

Components implemented in each core include:

- Trap Logic Unit (TLU)— The trap logic unit updates the machine state as well as handling exceptions and interrupts.• Instruction Fetch Unit (IFU)— The instruction fetch unit includes the 16KB instruction cache (32-byte lines, 8-way set associative) and a 64-entry fully-associative instruction translation lookup buffer (ITLB).
- Integer Execution Units (EXU)— Dual integer execution units are provided per core with four threads sharing each unit. Eight register windows are provided per thread, with 160 integer register file (IRF) entries per thread.
- Floating Point/Graphics Unit (FGU)— A floating point/graphics unit is provided within each core and it is shared by all eight threads assigned to the core. 32 floating point register file entries are provided per thread.
- Stream Processing Unit (SPU)— Each core contains a stream processing unit that provides cryptographic coprocessing.
- Memory Management Unit (MMU) — The memory management unit provides a hardware table walk (HWTW) and supports 8 KB, 64 KB, 4 MB, and 256 MB pages.

An eight-stage integer pipeline and a 12-stage floating-point pipeline are provided by each UltraSPARC T2 and UltraSPARC T2 Plus processor core (Figure6). A new

“pick” pipeline stage has been added to choose two threads (out of the eight possible per core) to execute each cycle.

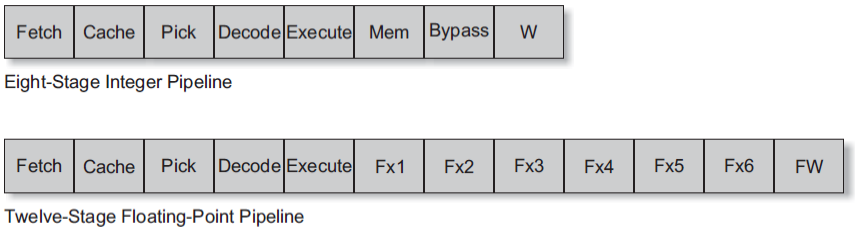


Figure6. UltraSPARC T2 Plus per-core integer and floating-point pipelines

To illustrate how the dual pipelines function, Figure 7 depicts the integer pipeline with the load store unit (LSU). The instruction cache is shared by all eight threads within the core. A least-recently-fetched algorithm is used to select the next thread to fetch. Each thread is written into a thread-specific instruction buffer (IB) and each of the eight threads is statically assigned to one of two thread groups within the core.

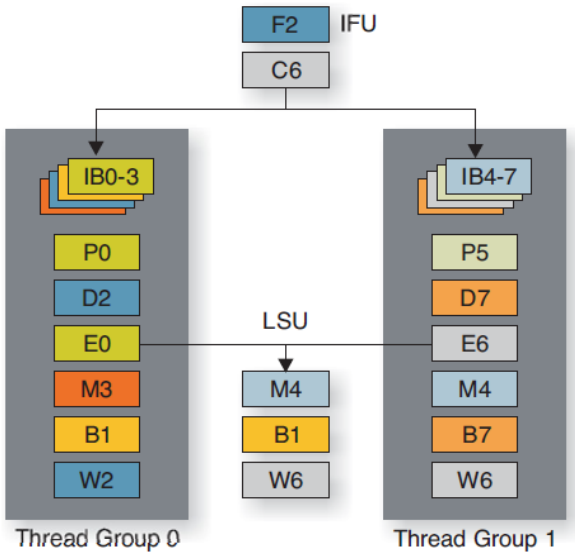


Figure 7. Threads are interleaved between pipeline stages with very few restrictions (integer pipelineshown, letters depict pipeline stages, numbers depict different scheduled threads)

The “pick” stage chooses one thread each cycle within each thread group. Picking within each thread group is independent of the other, and a least-recently-picked algorithm is used to select the next thread to execute. The decode state resolves resource conflicts that are not handled during the pick stage. As shown in the illustration, threads are interleaved between pipeline stages with very few restrictions. Any thread can be at the fetch or cache stage, before being split into either of the two thread groups. Load/store and floating point units are shared between all eight threads.

Only one thread from either thread group can be scheduled on such a shared unit.

Stream Processing Unit

The stream processing unit on each UltraSPARC T2 Plus core runs in parallel with the core at the same frequency. Two independent sub-units are provided along with a DMA engine that shares the core's crossbar port:

- A Modular Arithmetic Unit (MAU) shares the FGU multiplier, providing RSA encryption/decryption, binary and integer polynomial functions, as well as elliptic curve cryptography (ECC)
- The cipher/hash unit provides support for popular RC4, DES/3DES, AES-128/192/256, MD5, SHA-1, and SHA-256 ciphers

The SPU is designed to achieve wire-speed encryption and decryption on both of the processor's 10 GB Ethernet ports.

Integral PCI Express Support

The UltraSPARC T2 Plus processors provide an on-chip PCI Express interface that operates at 4 GB/second bidirectionally through a point-to-point dual-simplex chip interconnect. An integral IOMMU supports I/O virtualization and process device isolation by using the PCI Express BDF number. The total I/O bandwidth is 3-4 GB/second, with maximum payload sizes of 128 to 512 bytes. An x8 Serdes interface is provided for integration with off-chip PCI Express switches.

Power Management

Beyond the inherent efficiencies of multi-core multi-thread design, the UltraSPARC T2 and UltraSPARC T2 Plus processors are the first processor to incorporate unique power management features at both the core and memory levels of the processor. These features include reduced instruction rates, parking of idle threads and cores, and ability to turn off clocks in both cores and memory to reduce power consumption. Substantial innovation is present in the areas of:

- Limiting speculation such as conditional branches not taken
- Extensive clock gating in the data path, control blocks, and arrays
- Power throttling that allows extra stall cycles to be injected into the decode stage

Chapter 2

Server Architecture

The SPARC Enterprise T5440 server has been designed to provide breakthrough performance and scalability while building on the capabilities of SPARC Enterprise T5140 and T5240 servers. SPARC Enterprise T5440 servers expand on the strengths of the UltraSPARC T2 Plus processor, while providing innovative chassis design features. The result is a highly-scalable UltraSPARC T2 Plus based server with considerable CPU, memory, and I/O expansion capabilities.

System-Level Architecture

The design effort for the SPARC Enterprise T5440 server focused on providing cache coherency for up to four UltraSPARC T2 Plus processors with minimal latencies. The proven chassis design was chosen to provide computational density in a predictable thermal and power envelope. A high-level block diagram of the SPARC Enterprise server motherboard is shown in Figure 8.

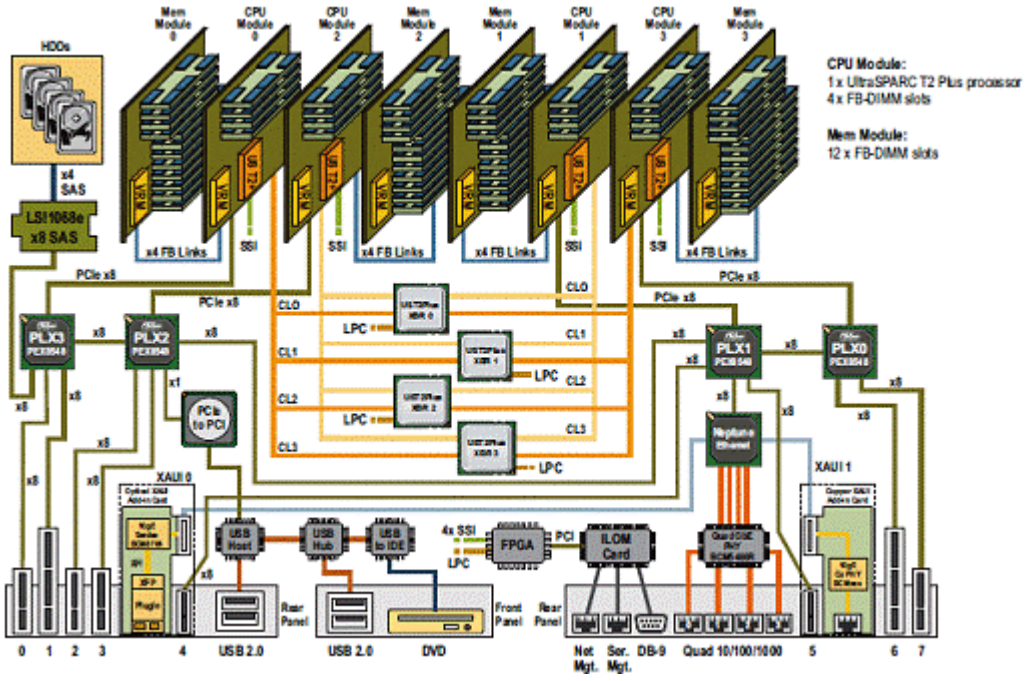


Figure 8. Block-level diagram of the SPARC Enterprise T5440 server

The SPARC Enterprise T5440 motherboard is a printed circuit board that supports all of the system capabilities, including processors and memory (on CPU and Memory Modules respectively), disk controller, and I/O subsystems. An ILOM service processor is provided on a daughter card that attaches to the motherboard. I/O options include USB, DVD control, quad Gigabit Ethernet, and four independent PCI Express buses providing sockets for a wide variety of third-party PCI Express expansion.

Key features of the SPARC Enterprise T5440 motherboard includes:
• Four sockets for CPU Modules, with each module containing an UltraSPARC T2 Plus processor and four FB-DIMM sockets. Each CPU Module contains the minimum memory configuration to support the UltraSPARC T2 Plus processor.
• Four sockets for Memory Modules, with each Memory Module providing expansion memory for the corresponding CPU Module through use of up to 12 FB-DIMM sockets. Each Memory Module is connected to a CPU module through a physical FB-DIMM interface via the motherboard.
• Four External Coherency Hub ASICs are provided on the motherboard. Each External Coherency Hub is connected to the Coherency Links on each UltraSPARC T2 Plus processor.
• Four PCI Express expansion hubs are provided, with each hub connected to the PCI Express interface on each UltraSPARC T2 Plus processor through the CPU Module socket. The PCI Express expansion hubs support a variety of peripheral system devices as well as eight PCI Express slots at the rear of the chassis, and are interconnected to help ensure connectivity from any CPU module to any peripheral device.
• Integration of Neptune chip provides 10 gigabit Ethernet functionality as well as four standard gigabit Ethernet ports (10/100/1000-BaseT)

The motherboard interconnect for T5440 has been greatly simplified over previous-generation systems. 12-volt power is distributed to the motherboard through a pair of metal bus bars, connected to a Power Distribution Board (PDB). A single flex-circuit connector routes all critical power control and DVD drive signaling over to the PDB. One or two mini-SAS cables connect the motherboard to the disk drive backplane, providing data access to the system hard drives.

External Coherency Hub (UltraSPARC T2 Plus Crossbar)

Dual-socket systems such as SPARC Enterprise T5140 and T5240 servers can be designed to directly connect coherency links of two UltraSPARC T2 Plus processors. For larger systems such as the SPARC Enterprise T5440 server, an External Coherency Hub is required. The External Coherency Hub extends the cache hierarchy of a single UltraSPARC T2 Plus processor across up to four sockets and processors (nodes).

The External Coherency Hub is a four-port arbiter/switch implemented in a custom ASIC that interfaces to the coherency control portion of the UltraSPARC T2 Plus processor. A high-level block diagram of the External Coherency Hub is shown in Figure 9. The External Coherency Hub performs a number of functions, including:

- Serializing requests to the same address for loads, stores, and write-backs
- Broadcasting snoops and aggregating responses
- Providing flow control read and write requests to a processor
- Providing ECC or parity protection

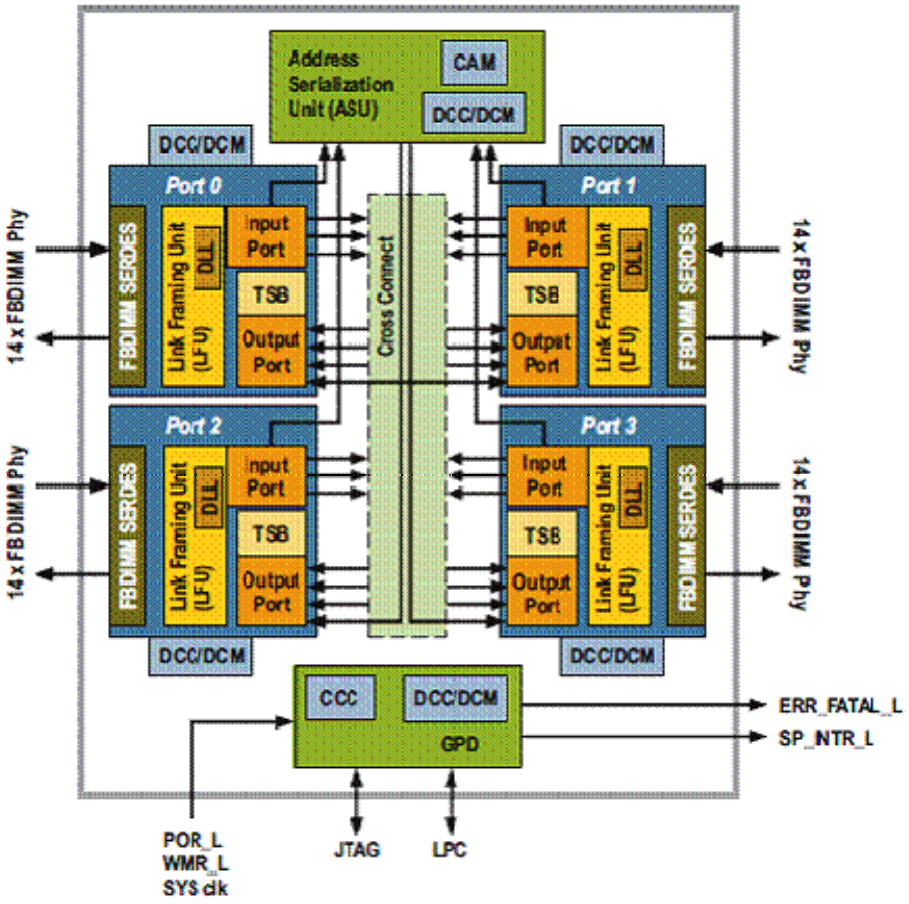


Figure 9. Block-level diagram of the External Coherency Hub

When memory requests fail to hit a processor's L2 cache, they are forwarded to the External Coherency Hub. The hub fulfills these requests through a series of messages to multiple nodes in a system. The hub also functions as the single point of serialization for addresses within a coherency plane, and enforces ordering rules necessary to comply with the Total Store Ordering (TSO) consistency model. The External Coherency Hub is also responsible for forwarding a node's noncacheable space and locations mapped to I/O devices on the PCI Express I/O fabric attached to the destination node.

Each SPARC Enterprise T5440 server provides four External Coherency Hubs, with each connected to all four CPU Module sockets. Components of the External Coherency Hub include:

•FB-DIMM Interface — The FB-DIMM interface block contains the Serdes cores and symbol alignment logic, virtually identical to that used for an FB-DIMM Northbound interface.

•Link Framing Unit — The Link Framing Unit (LFU) on each External Coherency Hub implements a protocol on top of the FB-DIMM Frame/Cell link layer that includes packetization, CRC checking, and retransmission of packets that had errors detected at the receiver.

• Input/Output Ports — The Output Port receives forwarded messages from all input ports, while processing some messages. A Transaction Score Board (TSB) tracks the state of requests, replies, and data respectively for coherent read and writeback.

• Address Serialization Unit (ASU) — The Address Serialization Unit is responsible for serializing requests to the same memory address, and for maintaining the list of requests pending on a particular address. The ASU handles requests from all four ports.

• Cross Connect —The Cross Connect consists of the set of interconnect wires that create the various port-to-port networks.

Memory Subsystem

In SPARC Enterprise T5440 server, the UltraSPARC T2 Plus processor provides on-chip memory controllers that communicate directly to FB-DIMM memory through high-speed serial links. Two dual-channel FB-DIMM memory controller units (MCUs) are provided on the UltraSPARC T2 Plus processor in addition to the four Coherency Units. Each MCU on the UltraSPARC T2 Plus processor can transfer data at an aggregate rate of 4.8 Gbps.

In addition to the UltraSPARC T2 Plus processor, each CPU Module provides four populated FB-DIMM slots. These four slots make up the "Primary" bank of FB-DIMMs, and are the minimum required to support a single processor. The CPU Module is always accompanied by its corresponding Memory Module, connected through an FB-DIMM interface via the motherboard connectors. By default, Memory

Modules are populated with four FB-DIMMs, allowing sufficient DIMM space to doubling memory capacity for each processor. Memory modules with no memory can be installed if desired, and are provided with DIMM fillers to maintain correct airflow.

Each Memory Module provides an additional 12 FB-DIMM sockets, yielding a total of 16 memory socket locations per processor. Note that the FB-DIMMs on each of the Memory Modules are physically connected to only one CPU Module, and one processor. With up to 16 667 MHz FB-DIMMs associated with each processor, the maximum system memory capacity is 128 GB using 2 GB FB-DIMMs, 256 GB using 4 GB FB-DIMMs, and 512 GB using 8 GB FB-DIMMs.

I/O Subsystem

Each UltraSPARC T2 Plus processor incorporates a single, 8-lane (x8) PCI Express port capable of operating at 4 GB/second bidirectionally. In SPARC Enterprise T5440 server, this port natively interfaces to the I/O devices through four PLX technology PCI Express expander chips(bridge chips). All of the PCI Express expander chips are themselves interconnected with x8 PCI Express interfaces — thus helping to ensure that any processor has access to any I/O device. The PCI Express expander chips connect either to PCI Express card slots, or to bridge devices that interface with PCI Express, such as those listed below.

- Disk Controller — Disk control is managed by an LSI Logic SAS1068E SAS controller chip that interfaces to a four-lane (x4) PCI Express port. RAID levels 0 and 1 are provided as standard.
- Modular disk backplanes — A four-disk backplane is attached to the LSI disk controller by an x4 SAS link.
- Quad gigabit Ethernet — On SPARC EnterpriseT5440 server, Neptune Ethernet chip provides two 10/ 100/1000-BaseT ports and two 10/100/1000/10000-BaseT interfaces, exposed as four RJ-45 connectors on the rear panel.
- Dual 10 Gb Ethernet — The SPARC Enterprise T5440 server provides dual 10 Gb XAUI connections, expressed through shared XAUI/PCI Express slots. These 10 Gb Ethernet interfaces are provided by Neptune Ethernet chip. When the 10 Gb Ethernet ports are connected, the corresponding number of the gigabit Ethernet ports become unavailable for use. The same XAUI expansion cards are supported on SPARC Enterprise T5120, T5220, T5140, T5240, and T5440 servers.
- USB and DVD —A single-lane PCI Express port connects to a PCI bridge device. A second bridge chip converts the 32-bit 33MHz PCI bus into multiple USB 2.0 ports. The system’s USB interconnect is driven from those ports. In addition, the DVD is driven from a further bridge chip that interfaces one of the USB ports to IDE format.

Chassis Design Innovations

SPARC Enterprise T5120/T5220, T5140/T5240, and T5440 servers share basic chassis design innovations. This approach not only provides a consistent look and feel across the product line, but it simplifies administration through consistent component placement and shared components. Beyond mere consistency, this approach provides a datacenter design focus that places key technology where it can make a difference for the datacenter.

- **Enhanced System and Component Serviceability**

  Finding and identifying servers and components in a modern datacenter can be challenge. T5440 server is optimized for lights-out datacenter configurations with easy to identify servers and modules. Color-coded operator panels provide easy-to-understand diagnostics and systems are designed for deployment in hot-isle / cold-isle multi-racked deployments with both front and rear diagnostic LEDs to pinpoint faulty components. Fault Remind features identify failed components.

  Consistent connector layouts for power, networking, and management make moving between systems straightforward. All hot-plug components are tool-less and easily available for serviceability. For instance, an integral hinged lid provides access to dual fan modules so that fans can be serviced without exposing sensitive components or causing unnecessary downtime.

- **Robust Chassis, Component, and Subassembly Design**

  SPARC Enterprise T5120/T5220, T5140/T5240 and T5440 servers share chassis that are carefully designed to provide reliability and cool operation. Even features such as the hexagonal chassis ventilation holes are designed to provide the best compromise for high strength, maximum air flow, and electronic attenuation.

  Fan assemblies direct airflow to efficiently cool the system. Fan modules are isolated from chassis to avoid transfer of rotational vibration to other system components. In addition, fans are monitored for vibration and actively controlled to avoid the effects of sympathetic vibrations being passed on to disk drives. Active monitoring and control of fan speeds is employed to reduce aural noise and fan power draw from system to a minimum while maintaining sufficient system cooling.

  In spite of their computational, I/O, and storage density, Fujitsu’s servers are able to maintain adequate cooling using conventional technologies. Minimized DC-DC power conversions also contribute to overall system efficiency. By providing 12 volt power to the motherboard, multiple power conversion stages can be eliminated. This approach reduces generated heat, and introduces further efficiencies to the system.

- **Minimized Cabling for Maximized Airflow**

  To minimize cabling and increase reliability, a variety of smaller boards and connectors are employed, appropriate to each chassis. These infrastructure boards

serve various functions in the SPARC Enterprise T5440 server.

- A power distribution boards distribute system power from the dual power supplies to the motherboard and to the disk backplane (via a connector board)
- Connector boards eliminate the need for many discrete cables, providing a direct card plug-in interconnect to distribute control and most data signals to the disk backplane, fan boards, and the PDB.
- PCI Express cards plug directly into the motherboard.
- Two XAUI slots provide access to the 10 Gb Ethernet interfaces on the UltraSPARC T2 Processor or Neptune Ethernet chip. Alternately, these slots can provide access to PCI Express interfaces. Each slot can either accept an optical/copper XAUI card, or an industry standard low-profile PCI Express card with up to an x8 form factor edge connector.
- The disk backplane mounts to the disk cages in the chassis, delivering disk data through 4-channel discrete mini-SAS cables from the motherboard. A 4-disk backplane is offered for the SPARC Enterprise T5440 server.
- Also provided by the disk backplane, a DVD-RW and two USB connections route to the front of the system.

SPARC Enterprise T5440 Server Overview

The SPARC Enterprise T5440 server provides breakthrough computational power and scalability in a space-efficient 4U rackmount package. By extending the proven benefits of multi-core, multi-thread architecture to the most mission-critical and OLTP database workloads supporting Web services, these systems redefine mid-range computing. The server is also designed to address the challenges of modern datacenters with greatly reduced power consumption and a small physical footprint. Depending on the model selected, the SPARC Enterprise T5440 server features one to four eight-core UltraSPARC T2 Plus processors, and up to 512 GB of memory.

System Motherboard and Chassis Perspective

Figure 10 provides a top-down perspective of the SPARC Enterprise T5440 chassis with the top cover removed and a full complement of CPU Modules and Memory Modules installed. As discussed, each CPU Module is paired with a corresponding Memory Module. A CPU Module is not required to have its corresponding Memory Module installed, and all CPU Modules are not required to be installed in order for the system to operate. If any CPU or Memory Module is not installed, a filler module is installed to help ensure proper airflow and front-to-back cooling.

All eight PCI Express slots are low profile, and are wired to x8 PCI Express interfaces. Two slots provide x16 physical connectors to support x16 PCI Express cards such as lowprofile graphics accelerators. Two PCI Express slots share the rear opening with optional XAUI cards that plug in behind the PCI Express connectors.

When an XAUI card is installed, no PCI Express card can be used in the corresponding PCI Express slot. Also, when an XAUI card is installed, one of the gigabit Ethernet ports attached to the Neptune chip becomes unavailable.

Four system fans insert from the top of the chassis, and four power supplies insert from the rear of the chassis. The four power supplies provide 2+2 redundancy. The system can continue to operate at full capacity with any combination of two of the four power supplies, thereby covering a failure in an individual power supply, or an entire circuit within the datacenter.

Enclosure

The 4U SPARC Enterprise T5440 server enclosure is designed for use in a standard 19-inch rack (Table 2).

Table 2. Dimensions and weight of the SPARC Enterprise T5440 server

Dimension	U.S.	International
Height	6.92 inches (4 RU)	176 millimeters
Width	17.5 inches	445 millimeters
Depth	24.9 inches	633 millimeters
Weight (approximate maximum, without PCI cards, rackmounts)	88 pounds	40 kilograms

The SPARC Enterprise T5440 server includes the following major components:

• One to four UltraSPARC T2 Plus processors with eight cores at speeds of 1.2 GHz , 1.4 GHz or 1.6 GHz

• Up to 512 GB of memory in 16 Fully Buffered Dual Inline Memory Module (FB-DIMM) slots per processor (2 GB , 4 GB, and 8 GB FB-DIMMs supported)

• Four on-board 10/100/1000 Mbps Ethernet ports

• Six dedicated low-profile PCI Express slots (x8 electrically, two x16 physical slots)

• Two combination XAUI or low-profile PCI Express x8 slots

• Four USB 2.0 ports (2 forward, 2 rear facing)

• Four available disk drives supporting SAS disk drives

• Integrated Lights out Management (ILOM) system controller

• Four hot-swappable high-efficiency 1,120 watt power supply units, providing 2+2 redundancy when distributed over two independent power circuits

• Four (N+1) cooling fans under environmental monitoring and control, accessed directly from the top of the chassis

Front and Rear Perspectives

Figure 10 illustrates the front and rear panels of the SPARC Enterprise T5440 server.

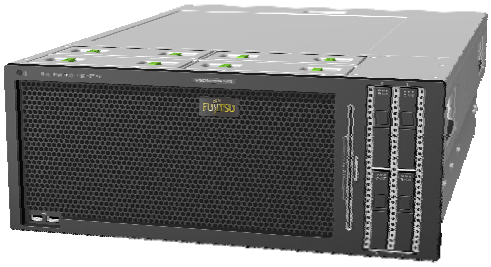


Figure 10. SPARC Enterprise T5440 server perspective

External features of the SPARC Enterprise T5440 server include:

• Front and rear system and component status indicator lights provide locator (white), service required (amber), and activity status (green) for the system.

• Four hot-plug SAS disk drives insert through the front panel of the system.

• One slimline, slot-accessible DVD-RW is accessed through the front panel.

• Four USB 2.0 ports are provided, two on the front panel, and two on the rear.

• Four hot-plug/hot-swap (2+2) power supplies with integral fans insert from the rear.

• Rear power-supply indicator lights convey the status of each power supply.

• A single AC plug is provided on each hot-plug/hot-swap power supply.

• Four 10/100/1000Base-T autosensing Ethernet ports are provided.

• A DB-9 TTYA serial port is provided for serial devices (not connected to the ILOM system controller serial port).

• A total of eight PCI Express card slots are provided, two of which can alternately support XAUI cards connected to the Neptune 10 Gb Ethernet chip.

• Two management ports are provided for use with the ILOM system controller. The RJ-45 serial management port provides the default connection to the ILOM controller.

PCI Express Expansion Unit

With four sockets for UltraSPARC T2 Plus processors, SPARC Enterprise T5440 servers are ideally suited for mission-critical applications and databases, applications that often require fast access to considerable near-line storage. With each of up to four UltraSPARC T2 Plus processors providing a PCI Express root complex, the system also has considerable I/O capacity and bandwidth available for external expansion. To scale I/O expansion beyond the constraints of the 4U system chassis, SPARC Enterprise T5440 servers support the attachment of up to two optional External I/O Expansion Units to provide additional I/O connectivity. As a result, a maximally configured SPARC Enterprise T5440 server can provide up to 28 PCI Express slots[1].

1.One I/O boat per installed CPU module represents the maximum configuration.

The External I/O Expansion Unit is a four RU rack-mountable device which

accommodates up to two 12 additional PCI Express slots — connected to, and managed by, the SPARC Enterprise T5440 server. By using cassettes, the external I/O chassis supports active replacement of hot-plug cards. An I/O Link card mounted in the host provides connectivity to the External I/O Expansion Unit and enables host management control via sideband signals. The I/O link card is available as a low height copper card, and includes a single 8-lane PCI Express bus with 4GB/second bandwidth. The architecture of the External 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.
Each External I/O Expansion Unit contains either one or two “I/O boats”, with each boat providing six external x8 PCI Express slots. Individual I/O boats connect to the SPARC Enterprise T5440 server via a link card that is installed in one of the system’s PCI Express slots. Each installed I/O boat requires an installed CPU module, and one I/O boat per installed CPU module is the maximum configuration. The External I/O Expansion Unit includes several key technologies, including:
• Link card side-band communication technology along with I/O manager capabilities built into the system software that allow seamless remote management and integration with the host server
• Redundancy and hot-plug capabilities for power supply units, fans, I/O boats, and I/O cards
• Thermal monitoring and remote diagnostic capabilities.

Figure 11 provides a block-level diagram of the External I/O Expansion Unit.

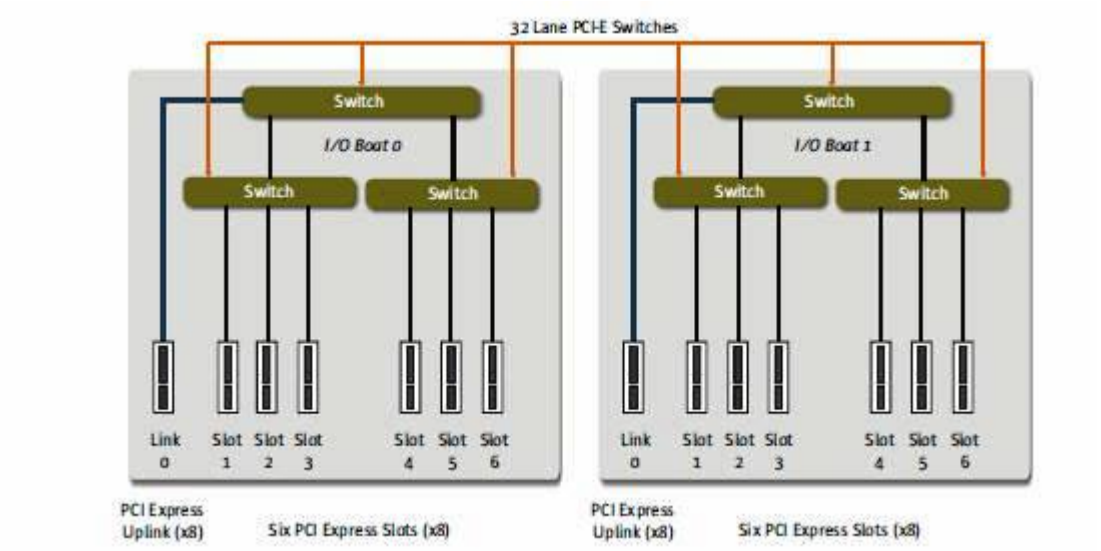


Figure 11. External I/O Expansion Unit architecture

The External I/O Expansion Unit consists of an advanced network of switches, bridges, and controllers to allow I/O data communication along with system control information via the PCI Express link card connection to the host server. The x8 PCI

Express connection from the host is forwarded over the link card cable to a switch in each I/O boat. Two additional switches then connect to the six x8 PCI Express slots in the I/O boat. The switches support the side-band management function of the link card, by providing the following functionality:

- Gathering and communicating I2C diagnostic and environmental information to the host server's ILOM service processor
- Executing system instructions to modify fan speeds or selectively power down components within the unit
- Performing locate/warning information via the front-panel LEDs
- Providing support for the online maintenance and replacement of power supply units, I/O boats, or I/O cards during system operation, and also supporting addition and deletion of active I/O cards

Chapter 3

Enterprise-Class Management and Software

While new technology often requires time for tools and applications to arrive, delivering agile and highly-available services that take advantage of available resources requires stable development tools, operating systems, middleware and management software. Fortunately, in spite of the breakthrough UltraSPARC T2 Plus processor technology, SPARC Enterprise T5440 server provide full binary compatibility with earlier SPARC systems and are delivered ready to run with pre-loaded tools and the solid foundation of the Solaris OS. Moreover, these systems are provided with a wealth of sophisticated tools that let organizations develop and tune applications as they consolidate and manage workloads while effectively utilizing the resources of UltraSPARC T2 and UltraSPARC T2 Plus processors.

System Management Technology

As the number of systems grows in any organization, the complexities of managing the infrastructure throughout its lifecycle become increasingly difficult. Effective system management requires both integrated hardware that can sense and modify the behavior of key system elements, as well as advanced tools that can automate key administrative tasks.

Integrated Lights-Out Management (ILOM) System Controller

Provided across SPARC Enterprise T5120 and T5220, the Integrated Lights Out Management (ILOM) service processor acts as a system controller, facilitating remote management and administration of SPARC Enterprise T5440 servers. As a result, these servers integrate easily with existing management infrastructure. In the SPARC Enterprise T5440 server, the ILOM service processor is provided on a daughter card that connects to the system motherboard.

Critical to effective system management, the ILOM service processor:

- Implements an IPMI 2.0 compliant services processor, providing IPMI management functions to the server's firmware, OS and applications, and to IPMI-based management tools accessing the service processor via the ILOM Ethernet management interface, providing visibility to the environmental sensors (both on the server module, and elsewhere in the chassis)
- Manages inventory and environmental controls for the server, including CPUs, DIMMs, and power supplies, and provides HTTPS/CLI/SNMP access to this data
- Supplies remote textual console interfaces
- Provides a means to download upgrades to all system firmware

The ILOM service processor also allows the administrator to remotely manage the server, independent of the operating system running on the platform and without

interfering with any system activity. ILOM can also send e-mail alerts of hardware failures and warnings, as well as other events related to each server. The ILOM circuitry runs independently from the server, using the server’s standby power. As a result, ILOM firmware and software continue to function when the server operating system goes offline, or when the server is powered off. ILOM monitors the following SPARC Enterprise T5440 server conditions:

• CPU temperature conditions
• Hard drive presence
• Enclosure thermal conditions
• Fan speed and status
• Power supply status
• Voltage conditions
• Solaris watchdog, boot time-outs, and automatic server restart events

Server Management Software

Monitoring and control software is essential in managing today's server infrastructure and the complications of IT device distribution. Fujitsu's Server System Manager (SSM) enables management of all Solaris, Linux, and Windows servers. Plus, in combination with Systemwalker Fujitsu's integrated systems management software, SSM provides autonomic operation and ensures you maintain business continuity as your IT environment changes to match your business.

Server System Manager (SSM)

SSM software provides a common server management environment for PRIMEQUEST mission-critical open servers, industry standard PRIMERGY servers, and SPARC Enterprise UNIX servers. Graphical server management views and functions for hardware configuration and hardware monitoring make SSM the easy option in a heterogeneous server environment.

Not only are you able to monitor the status of multiple servers you can also control server power from the same single consistent display. Ease of use continues as all server configurations are viewed using a single tree structure and common formats. This lets you check server component status a glance. Racked servers and related items are shown relative to their rack positions using life-like server images. This further simplifies status monitoring of server resources such as CPU and memory. Importantly, the networked nature of SSM means you can perform power on/off and other operations from any location with a management display.

With SSM server operation management is greatly simplified, letting you concentrate on hardware problem resolution and other critical work. As a result your work and the work of other administrators will be both reduced and more productive.

Enhanced Support Facility

Enhanced Support Facility is specific software that improves the operation

management and the maintainability of SPARC Enterprise servers. Working in combination with ILOM, server configuration, status and error messages can all be displayed. If a problem occurs, the information reported to ILOM 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.

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

Scalability and Support for Innovative Multithreading Technology

The Solaris 10 Operating System is specifically designed to deliver the considerable resources of UltraSPARC T2 Plus processor based systems. In fact, the Solaris 10 OS provides key functionality for virtualization, optimal utilization, high availability, unparalleled security, and extreme performance for both vertically and horizontally scaled environments. The Solaris 10 OS runs on a broad range of SPARC systems and compatibility with existing applications is guaranteed.

One of the most attractive features of systems based on the UltraSPARC T2 Plus processor is that they appear as a familiar Symmetric Multiprocessing (SMP) system to the Solaris OS and the applications it supports. In addition, the Solaris 10 OS has incorporated many features to improve application performance on mutithreading architectures:

- ***System Awareness***

  The Solaris 10 OS is aware of the UltraSPARC T2 and UltraSPARC T2 Plus processor hierarchy so that the scheduler can effectively balance the load across all the available pipelines. Even though it exposes each of these processors as 64 logical processors, the Solaris OS understands the correlation between cores and the threads they support, and provides a fast and efficient thread implementation.

- ***Fine-Granularity Manageability***

  For the UltraSPARC T2 Plus processor, the Solaris 10 OS has the ability to enable or disable individual cores and threads (logical processors). In addition, standard

Solaris OS features such as processor sets provide the ability to define a group of logical processors and schedule processes or threads on them.

- ***Binding Interfaces***

  The Solaris OS allows considerable flexibility in that processes and individual threads can be bound to either a processor or a processor set, if required or desired.

- ***Support for Virtualized Networking and I/O, and Accelerated Cryptography***

  The Solaris OS contains technology to support and virtualize components and subsystems on the UltraSPARC T2 Plus processor, including support for the on-chip PCI Express interface(10 Gb Ethernet ports). As a part of a high-performance network architecture, system-aware device drivers are provided so that applications running within virtualization frameworks can effectively share I/O and network devices. Accelerated cryptography is supported through the Solaris Cryptographic framework.

NUMA Optimization in the Solaris OS

With memory managed by each UltraSPARC T2 Plus processor on SPARC Enterprise T5440 server, the implementation represents a non-uniform memory access (NUMA) architecture. In NUMA architectures, the speed needed for a processor to access its own memory is slightly different than that required to access memory managed by another processor. The Solaris OS provides technology that can specifically help applications improve performance on NUMA architectures.

- Memory Placement Optimization (MPO) — The Solaris 10 OS uses MPO to improve the placement of memory across the physical memory of a server, resulting in increased performance. Through MPO, the Solaris 10 OS works to help ensure that memory is as close as possible to the processors that access it, while still maintaining enough balance within the system. As a result, many database applications are able to run considerably faster with MPO.
- Hierarchical lgroup support (HLS) — HLS improves the MPO feature in the Solaris OS. HLS helps the Solaris OS optimize performance for systems with more complex memory latency hierarchies. HLS lets the Solaris OS distinguish between the degrees of memory remoteness, allocating resources with the lowest possible latency for applications. If local resources are not available by default for a given application, HLS helps the Solaris OS allocate the nearest remote resources.

- ***Solaris ZFS File System***

  Solaris ZFS offers a dramatic advance in data management, automating and consolidating complicated storage administration concepts and providing unlimited scalability with the world’s first 128-bit file system. Solaris ZFS is based on a transactional object model that removes most of the traditional constraints on I/O issue order, resulting in dramatic performance gains. Solaris ZFS also provides data

integrity, protecting all data with 64-bit checksums that detect and correct silent data corruption.

- ***A Secure and Robust Enterprise-Class Environment***
  Best of all, the Solaris OS doesn't require arbitrary sacrifices. The Solaris Binary Compatibility Guarantee helps ensure that existing SPARC applications continue to run unchanged on UltraSPARC T2 Plus platforms, protecting investments. Certified multi-level security protects Solaris environments from intrusion. Comprehensive Fault Management Architecture means that elements such as Solaris Predictive Self Healing can communicate directly with the hardware to help reduce both planned and unplanned downtime. Effective tools such as DTrace help organizations tune their applications to get the most of the system's resources.

End-to-End Virtualization Technology

Virtualization technology is increasingly popular as organizations strive to consolidate disparate workloads onto fewer more powerful systems, while increasing utilization. SPARC Enterprise T5440 server is specifically designed for virtualization, providing very fine-grained division of multiple resources — from processing to virtualized networking and I/O. Most importantly, virtualization technology is provided as a part of the system, not an expensive add-on.

A Multithreaded Hypervisor

Like the UltraSPARC T1 and the UltraSPARC T2 processors, the UltraSPARC T2 Plus processor offers a multithreaded hypervisor — a small firmware layer that provides a stable virtual machine architecture that is tightly integrated with the processor. Multi-core Multi-thread is crucial, since the hypervisor interacts directly with the underlying chip-multithreaded processor. This architecture is able to context switch between multiple threads in a single core, a task that would require additional software and considerable overhead in competing architectures.

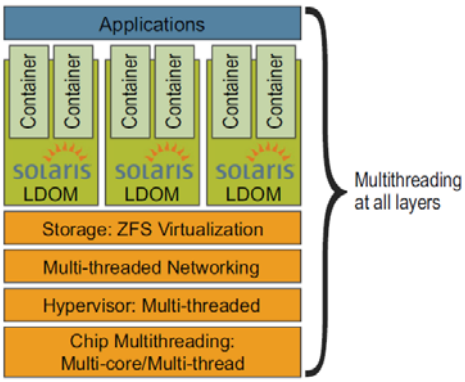


Figure12. Fujitsu provides parallelization and virtualization at every level of the technology stack

Corresponding layers of virtualization technology are built on top of the hypervisor as

shown in Figure12. The strength of Fujitsu's approach is that all of the layers of the architecture are fully multithreaded - from the processor up through applications that use the fully threaded Java application model. Far from new technology, the Solaris OS has provided multi-core multi-thread support since 1992. This considerabe experience has helped to inform technology decisions at other levels, ultimately resulting in a system that parallelizes and virtualizes at every level. In addition to the processor and hypervisor, Fujitsu provides fully multithreaded networking and the fully multithreaded Solaris ZFS file system. Logical Domains (LDOMs), Solaris Containers, and multithreaded applications are able to receive exactly the resources they need.

Logical Domains

Supported in all servers utilizing multi-core multi-thread technology, Logical Domains provide full virtual machines that run an independent operating system instance, and contain virtualized CPU, memory, storage, console, and cryptographic devices. Within the Logical Domains architecture, operating systems such as the Solaris 10 OS are written to the hypervisor, which provides a stable, idealized, and virtualizable representation of the underlying server hardware to the operating system in each Logical Domain. Each Logical Domain is completely isolated, and the maximum number of virtual machines created on a single platform relies upon the capabilities of the hypervisor, rather than the number of physical hardware devices installed in the system. For example, the SPARC Enterprise T5440 s supports up to 128 logical domains[1], and each individual logical domain can run a unique OS instance.

1.Though possible, this practice is not a generally recommended.

By taking advantage of Logical Domains, organizations gain the flexibility to deploy multiple operating systems simultaneously on a single platform. In addition, administrators can leverage virtual device capabilities to transport an entire software stack hosted on a Logical Domain from one physical machine to another. Logical Domains can also host Solaris Containers to capture the isolation, flexibility, and manageability features of both technologies. Deeply integrating Logical Domains with the UltraSPARC T2 Plus processors and the Solaris 10 OS increases flexibility, isolates workload processing, and improves the potential for maximum server utilization.

The Logical Domains architecture includes underlying server hardware, hypervisor firmware, virtualized devices, and guest, control, and service domains. The hypervisor firmware provides an interface between each hosted operating system and the server hardware. An operating system instance controlled and supported by the hypervisor is called a *guest domain*. Communication to the hypervisor, hardware platform, and other domains for creation and control of guest domains is handled by the *control domain*. Guest domains are granted virtual device access via a *service*

domain which controls both the system and hypervisor, and also assigns I/O.

To support virtualized networking, Logical Domains implement a virtual Layer 2 switch(vswitch), to which guest domains can be connected. East guest domain can be connected to multiple vswitches and multiple guest domains can also be connected to the same vswitch. Vswitches can either be associated with a real physical network port, or they may exist without an associated port, in which case the vswitch provides only communications between domains within the same server. This approach also gives guest domains a direct communication channel to the network (Figure13). Each guest domain believes it owns the entire NIC and the bandwidth it provides, yet in practice only a portion of the total bandwidth is allotted to the domain. As a result, every NIC can be configured as demand dictates, with each domain receiving bandwidth on an asneeded basis. Dedicated bandwidth can be made available by tying a vswitch device to a dedicated physical Ethernet port.

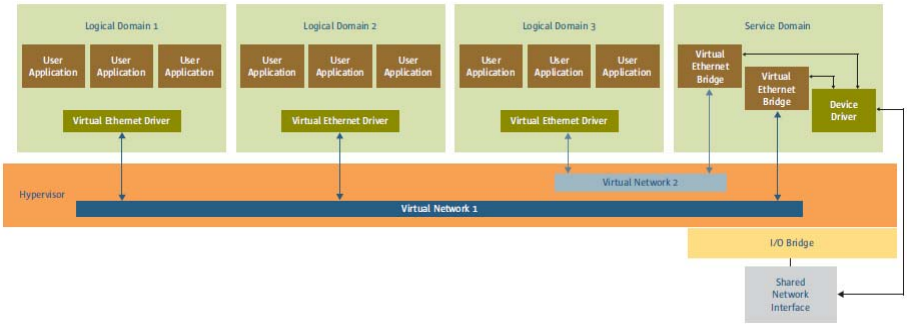


Figure13. Data moves directly between a Logical Domain and a virtualized device

Solaris™Containers

Providing virtualization at the OS level, Solaris Containers consist of a group of technologies that work together to efficiently manage system resources, virtualize the environment, and provide a complete, isolated, and secure runtime environment for applications. Solaris containers include important technologies that work together with the fair-share scheduler:

- ***Solaris Zones***

  The Solaris 10 OS provides a unique partitioning technology called Solaris Zones that can be used to create an isolated and secure environment for running applications. A zone is a virtualized operating system environment created within a single instance of the Solaris OS. Zones can be used to isolate applications and processes from the rest of the system. This isolation helps enhance security and reliability since processes in one zone are prevented from interfering with processes running in another zone.

- ***Resource Management***

  Resource management tools provided with the Solaris OS help allocate resources such as CPUs to specific applications. CPUs in a multiprocessor system (or threads

in the UltraSPARC T2 Plus processors) can be logically partitioned into processor sets and bound to a resource pool, which in turn can be assigned to a Solaris zone. Resource pools provide the capability to separate workloads so that consumption of CPU resources do not overlap, and also provide a persistent configuration mechanism for processor sets and scheduling-class assignment. In addition, the dynamic features of resource pools enable administrators to adjust system resources in response to changing workload demands.

Fault Management and Predictive Self Healing

The Solaris 10 OS introduced a new architecture for building and deploying systems and services capable of fault management and predictive self-healing. Predictive Self Healing is an innovative capability in the Solaris 10 OS that automatically diagnoses, isolates, and recovers from many hardware and application faults. As a result, business critical applications and essential system services can continue uninterrupted in the event of software failures, major hardware component failures, and even software miss-configuration problems.

- ***Solaris Fault Manager***

  The Solaris Fault Manager facility collects data relating to hardware and software errors. This facility automatically and silently detects and diagnoses the underlying problem, with an extensible set of agents that automatically respond by taking the faulty component offline. Easy-to-understand diagnostic messages link to articles in knowledge base to clearly guide administrators through corrective tasks that require human intervention. The open design of the Solaris Fault Manager facility also permits administrators and field personnel to observe the activities of the diagnostic system. With Solaris Fault Manager, the overall time from a fault condition, to automated diagnosis, to any necessary human intervention is greatly reduced, increasing application uptime.

- ***Solaris Service Manager***

  The Solaris Service Manager facility creates a standardized control mechanism for application services by turning them into first-class objects that administrators can observe and manage in a uniform way. These services can then be automatically restarted if they are accidentally terminated by an administrator, if they are aborted as the result of a software programming error, or if they are interrupted by an underlying hardware problem. In addition, the Solaris Service Manager software reduces system boot time by as much as 75 percent by starting services in parallel according to their dependencies. An “undo” feature helps safeguard against human errors by permitting easy change rollback. The Solaris Service Manager is also simple to deploy; developers can convert most existing applications to take full advantage of Solaris Service Manager features by simply adding a simple XML file to each application.

Predictive self healing and fault management provide the following specific

capabilities on SPARC Enterprise T5440 servers:

- *CPU Offlining* takes a core or threads offline that has been deemed faulty. Offlined CPUs are stored in the resource cache and stay offline on reboot unless the processor has been replaced, in which case the CPU is cleared from the resource cache.
- *Memory Page Retirement* retires pages of memory that have been marked as faulty. Pages are stored in the resource cache and stay retired on reboot unless the offending DIMM has been replaced, in which case affected pages are cleared from the resource cache.
- *I/O Retirement* logs errors and faults.
- *fmlog* logs faults detected by the system.

Chapter 4

Conclusion

Delivering on the demands of Web 2.0 applications and virtualized, eco-efficient data centers requires a comprehensive approach that includes innovative processors, system platforms, and operating systems, along with leading application, middleware, and management technology. With its strong technology positions and R&D investments in all of these areas, Fujitsu is in a unique position to deliver on this vision. Far from futuristic, Fujitsu has effective solutions today that can help organizations cope with the need for performance and capacity while effectively managing space, power and heat.

Building on the successful UltraSPARC T1 and the UltraSPARC T2 processor, the UltraSPARC T2 Plus processor extends the proven benefits of multi-core/multi-thread architecture into the most mission-critical applications and OLTP database workloads supporting Web services.
With up to 256 threads, up to 512 GB of memory, and massive I/O bandwidth, these systems represent ideal consolidation and virtualization platforms for essential applications and services. Providing 64 threads per processor, on-chip memory management, and cache coherency for multisocket support, PCI express, and on-chip cryptographic acceleration, the UltraSPARC T2 Plus processor fundamentally redefine the capabilities of a modern processor. By incorporating cache coherency for multiprocessor support, the SPARC Enterprise T5440 server take advantage of these strengths to provide powerful and highly-scalable server platforms while delivering new levels of performance and performance-per-watt in a compact rackmount chassis. The result is datacenter infrastructure that can literally replace racks of existing equipment, lowering power and space requirements.

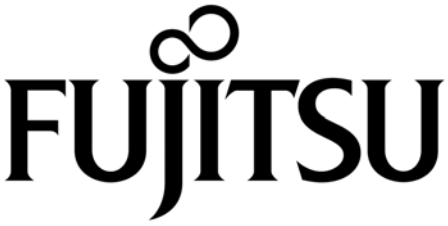
The SPARC Enterprise T5440 server provides the computational, networking, and I/O resources needed by the most demanding Web, application, and database applications - facilitating highly-effective consolidation efforts. With end-to-end support for multi-core multi-thread technology and virtualization, these systems can consolidate workloads and effectively utilize system resources even as they preserve investments in SPARC/Solaris technology and provide tools for open-source software environments. With innovations such as Logical Domains, Solaris Containers, and Java technology, organizations can adopt these radical new systems for their most important projects —acting responsibly toward the environment and the bottom line.

For More Information

To learn more about Sun products and the benefits of Sun SPARC Enterprise T5440 server, contact a Sun sales representative, or consult the related documents and Web sites listed in Table 3.

Table 3. Related Websites

Web Site URL	Description
sun.com/coolthreads	Sun SPARC Enterprise T5120/T5220, T5140/ T5240, and T5440 Servers
sun.com/processors/UltraSPARC-T2	Sun UltraSPARC T2 and UltraSPARC T2 Plus Processors
opensparc.net/opensparc-t2	OpenSPARC T2
sun.com/processors/throughput	Throughput Computing
sun.com/servers/coolthreads/overview	Sun Servers with CoolThreads Technology
sun.com/servers/coolthreads/ldoms	Sun Logical Domains
sun.com/solaris	The Solaris Operating System


FUJITSU