The Sun™ Enterprise™ 3500 - 6500 Server Family: Architecture and Implementation

Technical White Paper

Printed in the United States of America.
901 San Antonio Road, Palo Alto, California 94303 U.S.A.

Please
Recycle

Adobe PostScript

Contents

Introduction 1

The new Sun™ Enterprise™ 3500 - 6500 server family is a generational refinement of the extremely successful Enterprise 3000 - 6000 server product line from Sun Microsystems™. The new systems retain their simplicity and reliability, while extending the flexibility, configurability, and high performance of their predecessors.

Like its predecessor, the Enterprise X500 architecture is a large-scale shared memory symmetric multiprocessor system based on the high performance UltraSPARC™ processor. Packaged and priced to suit a wide range of computing environments, Enterprise X500 systems also offer enhanced reliability, availability, and serviceability (RAS) along with higher performance, expandability, investment protection, and upgradability.

Sun's Server Product Line

Sun has a broad, binary compatible server product line, allowing customers to upgrade servers as computing needs grow. In addition, a binary compatible product line allows customers to run the same Solaris application with no modifications as they move from one product to another in the family. This saves time and money by allowing the continued use of existing applications, even when upgrading to a different server.

Sun's server offering scales from the uniprocessor Enterprise 5S all the way up to the 64 processor Enterprise 10000 (Figure 1-1). This scalability enables customers to grow functionality and capacity on demand without impacting applications or training. The remainder of this paper discusses the Sun Enterprise 3500 - 6500 (X500) Server family.

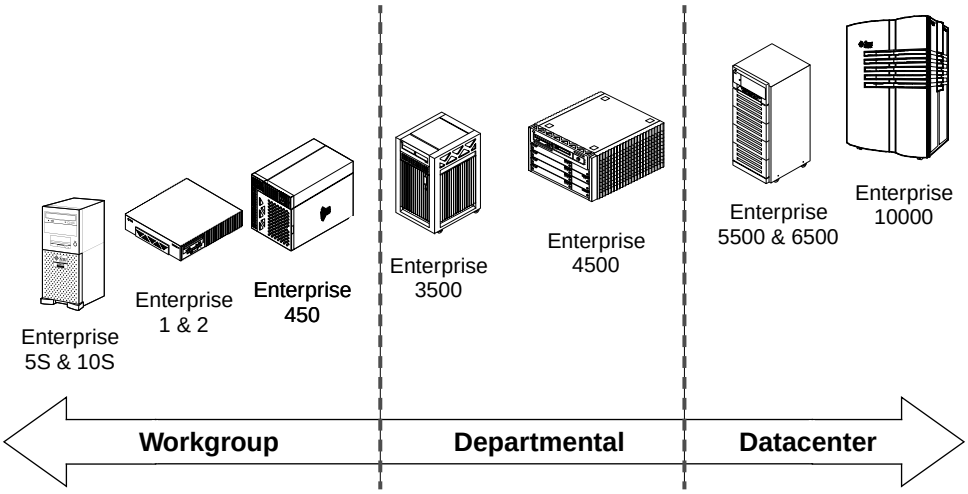

Figure 1-1 Sun's server product line offers superior scalability.

Enterprise Product Overview

The Enterprise X500 servers offer superior performance, expandability, and flexible configurations, along with high reliability, availability, and serviceability. The product line is highly scalable, starting with the eight processor Enterprise 3500 and continuing up in performance and expandability to the 30 processor Enterprise 6500 (Figure 1-2).

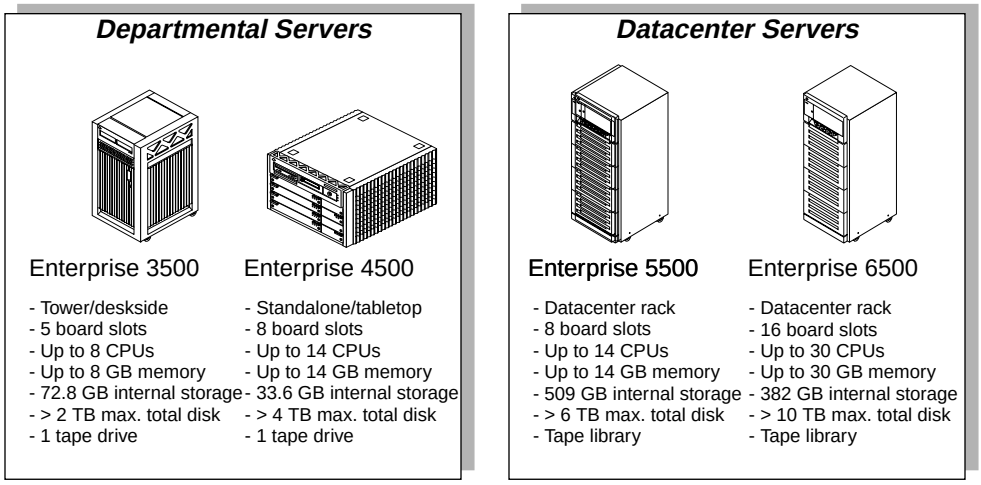

Figure 1-2 The Sun Enterprise X500 product line scales from the department-level Enterprise 3500 to the highly expandable datacenter server, the Enterprise 6500.

- *Enterprise 3500 Server*

  The Enterprise 3500 server is an entry-level, departmental system in a tower design that includes integrated storage. The system has a total of five board slots which can be used for either CPU/memory boards or I/O boards. The Enterprise 3500 supports up to eight UltraSPARC processor modules and eight GB of memory, a CD-ROM drive, and an optional half-height tape drive. The Enterprise 3500 also includes disk bays for eight dual-ported Fibre-Channel Arbitrated Loop (FC-AL) disk drives.

- *Enterprise 4500 Server*

  The Enterprise 4500 is the high-end departmental server in a compact deskside or tabletop enclosure. This server features eight system board slots and supports up to 14 CPUs. A disk card with two 3.5-inch disk drives is available for building affordable, entry level configurations. However, external Sun StorEdge™ products are typically used to build more comprehensive storage solutions for this server.

- *Enterprise 5500 Server*

  The Enterprise 5500 server is an entry-level datacenter server with eight board slots. It is packaged in a datacenter cabinet, making efficient use of limited datacenter real estate and integrating well into existing installations.

  The Enterprise 5500 provides the same system features and expandability as the Enterprise 4500. However, the integrated datacenter rack provides additional capacity for disk storage and tape options. This system can support over half a terabyte internal disk capacity together with a tape library.

- *Enterprise 6500 Server*

  The Enterprise 6500 server, a high-end datacenter server, is the most expandable system in the Enterprise X500 product family. It features 16 system board slots, supporting up to 30 CPUs and 30 GB of memory. The system rack provides support for multiple internal disk subsystems and tape options.

The Enterprise X500 Design Philosophy

Reliability, availability, and serviceability (RAS) — essential for mission-critical environments — are central to the Enterprise X500 server design. In addition to capabilities such as hot swap redundant power and cooling and hot swap disks, the Enterprise X500 servers also feature hot-plug CPU / memory and I / O boards, dynamic reconfiguration and alternate pathing software which enable on-line repair and expansion. With features not found in the same class of competitive systems, the X500 server line provides unequalled levels of system availability.

The Enterprise X500 also incorporates a modular design for easy upgradability and serviceability. Major components are modular, enabling customers to upgrade by simply unplugging previous generation components and plugging in the new ones. All other components of the system remain the same, providing outstanding investment protection. Sun's design philosophy also includes the use of common components across products, minimizing the need for a large number of costly spares.

When designing the X500 server, Sun sought a balanced system architecture. By ensuring that the throughput and performance of processors, backplane, I/O and memory are carefully matched, Sun delivers mainframe-class I/O capabilities and impressive performance scalability.

Highly Leveraged Design

The Enterprise X500 family builds on the success and technology pioneered in earlier generations of Sun servers, ensuring products that are more economical and reliable. The Enterprise X500 design shares the UltraSPARC processor and much of its interconnection technology with the current Ultra desktop systems. As a result, many of the hardware-dependent portions of the operating system and the ASICs (application specific integrated circuits) used to implement most of the system core are shared with the desktop systems. Shared ASICs include the SBus™-to-UPA interface (SYSIO), the integrated Fast/Wide SCSI-2 and Fast Ethernet (HME), and the SBus interrupt processor (RlSC).

In Sun's earlier generation systems, desktop and datacenter systems used different chipsets to implement buses, I/O, and memory interfaces. Because the underlying machine architectures were different, desktop and server systems required different operating system microkernels. Although the SPARCcenter 2000 and SPARCserver 1000 family are among the most successful multiprocessor UNIX systems in history, the number of deployed systems is small when compared to the number of UNIX-based desktop systems sold. By adopting a unified kernel architecture for all UltraSPARC-based systems, server users now benefit from the test and development cycles for both platforms.

Nevertheless, servers and desktop systems do have different requirements, and the Enterprise X500 design differs from Ultra desktops in two key areas crucial to servers: a high-bandwidth system bus to support large symmetric multiprocessor configurations, and I/O configurations which are more flexible and much higher in performance than those of their desktop counterparts.

Modular Design

Earlier generation SPARCcenter™ 2000 and SPARCserver™ 1000 servers shared crucial components, notably memory, peripheral boards, and especially the ASICs which implemented the bus interfaces and memory controllers. This architecture is collectively known as the *sun4d*. Solaris™ treated the

SPARCserver 1000 as if it were a (small) SPARCcenter 2000 running with just a single XDBus™. Although software-compatible, the system boards for the SPARCserver 1000 and SPARCcenter 2000 were not interchangeable.

The Enterprise X000 and X500 families extend the notion of family interoperability to physical field-replaceable units, further simplifying the product family. Every member of the Enterprise X500 family uses common components, including processor modules, CPU/memory boards, I/O boards, clock board, power supplies, power/cooling modules, and peripherals. Only the physical packaging is different, and only in form factor. Components can also be freely interchanged among all Enterprise X500 systems, simplifying the logistics of upgrades and maintenance. In addition, Enterprise X500 and X000 servers share the same components, and the Enterprise X500 enhanced system boards are supported in Enterprise X000 servers, and vice versa.

Enhancements over Enterprise 3000 - 6000 Server Family

The primary differences between Sun's new generation of Enterprise X500 servers and the existing Enterprise X000 server family are summarized below. This section is intended as a quick summary for those readers already familiar with the Enterprise X000 architecture. More detailed information is included in the following chapters.

- *All Enterprise X500 platforms*

  All Enterprise X500 servers include an enhanced Gigaplane™ system interconnect that is able to run at a rate of up to 100 MHz (90 MHz in the Enterprise 6500), providing better support for faster processors. At introduction, both 250 MHz and 336 MHz UltraSPARC processors can be ordered with the system. The clock board, CPU/memory board, and all I/O boards are enhanced to support the faster Gigaplane. In addition, the graphics I/O board has been revised to provide support for 100 MB/sec Fibre Channel Arbitrated Loop (FC-AL) disk subsystems. These servers also include a faster 32X CD-ROM drive.

  Software enhancements provide increased availability and serviceability. New to the Enterprise X000 and X500 server families are CPU power control, dynamic reconfiguration, and alternate pathing.

- *Sun Enterprise 3500*

  The new Sun Enterprise 3500 servers contain one more system slot than the Enterprise 3000, supporting more CPU/memory or I/O expandability. All Enterprise 3500 servers also include internal disk bays for 8 FC-AL disk drives with dual port connections, providing higher throughput and increased redundancy.

- *Sun Enterprise 5500*

  The Enterprise 5500 servers are packaged in a taller system cabinet than the Enterprise 5000. This new 68-inch cabinet provides 12 additional inches (approximately 31 cm) of space for storage devices. The vertical SCSI tray for tape drives has also been enhanced to support new tape options.

- *Sun Enterprise 6500*

  Like the Enterprise 5500, the Enterprise 6500 is packaged in the taller 68-inch system cabinet with a new vertical SCSI tape tray. The Enterprise 6500 also includes the revised Gigaplane system interconnect. However, because of the physical length of the Gigaplane system bus in the Enterprise 6500, it will have a maximum speed of 90 MHz. Furthermore, when running at 90 MHz, Enterprise 6500 systems can utilize only 12 of the 16 total system slots. Disk boards, however, are not included in the 12 slot limitation because they do not connect logically to the Gigaplane.

  To ensure maximum configuration flexibility customers can choose whether to use all 16 slots and run at up to 84 MHz, or use 12 slots (excluding disk boards) with higher Gigaplane throughput and lower latency in the system bus.

System Architecture 2

Three Packages, One Design

All members of the Enterprise X000 and X500 families share one basic physical design. Processors and memory are configured on circuit boards similar to those found in the previous generation systems, while a variety of I/O facilities are offered on separate I/O boards. CPU/memory boards and all I/O boards have the same form factor and share a common system bus interface, permitting any centerplane slot to be used for any function.

One of the goals of the design team was to provide larger configurations in less physical space than previous systems. Previous systems configured processor, memory, and I/O Boards on one side of a traditional backplane, but configuring as many as sixteen boards on a traditional backplane proved to be a packaging problem.

Enterprise 4X00, 5X00 and 6X00 systems solve this problem by moving the backplane into the physical center of the card cage and configuring boards on both sides. The centerplane configuration results in a physically shorter bus distance for a given configuration, greatly improving packaging density and permitting the bus to be run at higher speeds. The largest members of the families, the Enterprise 6000 and 6500 systems, offer sixteen centerplane slots, eight facing the front of the system and eight facing the rear. The Enterprise 4X00 and 5X00 offer eight centerplane slots, four facing the front of the system and four facing the rear.

The smallest member of the Enterprise X500 family, the Enterprise 3500, is not subject to the same physical limitations as its larger siblings. Because it has only five slots, there is little reason for a centerplane arrangement. Intended as a deskside system, Enterprise 3500 uses a traditional backplane configuration with all five boards plugged into the one side of the bus.

Expansion by Replication

The Enterprise X500 architecture features separate CPU/memory boards and I/O boards. CPU/memory boards contain zero, one, or two processors and as much as 2 GB of memory. Three types of I/O boards — SBus, Graphics, and PCI — are available, providing a wide range of I/O connectivity options.

A minimum system consists of a clock board, a CPU/memory board and an I/O board. The system is expanded by adding CPU/memory boards and/or I/O boards populating functional units until the required capacity is achieved. Almost any combination of boards is allowed, as long as the required minimum of one CPU/memory board and one I/O board is met. One I/O board must be inserted into slot 1 to provide the SCSI connection to the CD-ROM drive and the optional internal tape drive.

The X500 was designed with separate CPU/memory and I/O boards for several reasons. Most importantly, this design provides maximum configuration flexibility. Servers with compute-intensive applications can be configured with more CPU/memory boards, while servers with I/O intensive workloads can be configured with a greater number of I/O boards.

Having separate CPU/memory boards and I/O boards also allows new features to be added to one board without affecting the overall technology of the server. New features can be added more quickly and at a lower cost than with an integrated system board design.

CPU/memory boards are discussed in more detail in chapter 4; I/O boards are discussed in chapter 5.

Disk Boards

The current Enterprise X500 disk board contains up to two 4.2 GB SCSI disk drives, and is intended for entry-level configurations or environments that require internal SCSI disks as system disks. The Enterprise 4500 and 5500 servers both support a maximum of four disk boards, while the Enterprise 6500

server supports a maximum of two disk boards (in slots 14 and 15). The disk board is not supported in the Enterprise 3500 server, which already contains up to eight internal FC-AL disks.

Clock Board

The centerplane is operated from the Clock Board, which also includes several miscellaneous functions such as the keyboard and mouse port, two RS232 serial ports, time-of-day clock, and the system's IDPROM. The IDPROM provides the system's base network MAC address as well as its unique hostid and serial number. This IDPROM is socketed and can be removed from one clock board and inserted in another, maintaining ID information when a system is upgraded or when a clock board is replaced.

The frequency of the clock, the processor clock, and the system clock are all programmable through firmware.

Gigaplane System Bus

The system interconnect for the Enterprise X500 servers is an enhanced Gigaplane bus, capable of running at 100 MHz. This faster clock speed provides higher throughput and lower latency and therefore better supports faster processors. Currently, 250 MHz and 336 MHz UltraSPARC processors are shipped with Enterprise X500 servers. Previous generation 167 MHz UltraSPARC processors are also supported when upgrading from existing Enterprise X000 systems.

More detailed information on the Gigaplane is included in chapter 3.

Power and Cooling

Enterprise X500 servers offer a highly flexible, highly reliable power and cooling environment.

Redundant Hot-Plug Power Supplies

Enterprise X500 systems have two basically different types of power supplies: a power/cooling module (PCM) and a peripheral power supply (PPS). Each PCM provides power (300 watt) for 2 boards occupying 2 adjacent board slots, supplying 5v DC, 3.3v and 2.0v to the system boards. The peripheral power

supply (184 or 195 watt) provides power for the clock board, internal disk drives (in Enterprise 3500 only), CD-ROM, and optional tape drive (Figure 2-1). The peripheral power supply also provides necessary pre-charge for the Gigaplane board slots enabling hot-plugging of system boards.

Rather than relying upon a single power supply, separate power/cooling modules are provided for each pair of centerplane slots. Additional availability is achieved through the current share design of the system. Any board slot can be powered from any PCM in the system, and current sharing is used to permit the system to tolerate the failure of a power supply and continue operation. Failure of a power supply generates a high-priority system interrupt and the operating system is able to notify the system administrator of the situation.

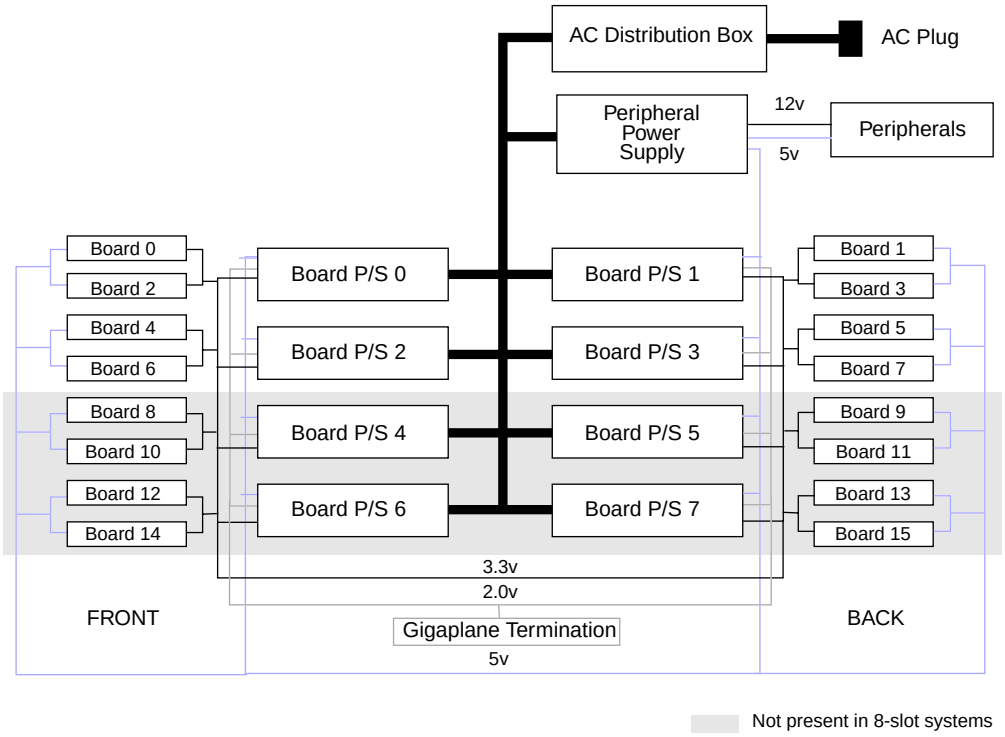

Figure 2-1 Current-sharing architecture of the Sun Enterprise X500 power supply system

In Enterprise 4500 to 6500 servers, all power supplies are fully hot swappable and failed units can be replaced at any time. For higher availability requirements, additional power supplies can be configured to provide a greater margin for error.

The 5-slot Enterprise 3500 can support three PCMs and a redundant PPS to protect against a power supply failure taking the system or disk drives down. The 8-slot Enterprise 4500 and Enterprise 5500 servers fully configured with 4 PCMs, and the 16-slot Enterprise 6500 system fully configured with 8 PCMs, have enough extra current to continue running should any one power supply fail.

Redundant Cooling

Each power/cooling module contains two cooling fans, which provide cooling for the two adjacent board slots. Because the only way to provide cooling to the system boards is through an adjacent PCM, it is necessary that any slot with a system board installed have an adjacent PCM installed. As with power supplies, distributed cooling permits a lower entry cost and a pay-as-you-go approach to system expansion.

The fans are not dependent upon their partner power supplies; instead they obtain power directly from the system-wide passive shared AC distribution circuit. Thus, the fans will stay up and running even if the co-located power supply fails.

Fan speed is thermally controlled; should a fan fail, its neighbor increases speed to account for the difference. The output of the temperature sensors on the CPU/memory and I/O boards is also fed back to the adjacent PCM to control the speed of the fans. Hotter boards will cause the fans to spin faster, and if two boards are present, the warmer of the two temperatures is used to set the fan speed.

Resilient Power Supply

Industry studies have shown that power fluctuations are the most common causes of unscheduled service interruption. The Enterprise X500 power supplies are resistant to most kinds of power fluctuation. Enterprise 5500 and Enterprise 6500 systems can continue functioning during a brownout of 160 Volts AC (VAC) for at least 15 minutes (recommended voltage is 230V in the US and 240V elsewhere). When a brownout continues for excessive periods, or

when internal voltages exceed pre-defined limits, a non-maskable powerfail interrupt is delivered to all processors and the system begins an orderly shutdown.

Complete AC line dropouts are tolerated if they are limited to a single cycle. If the electrical service is sufficiently unreliable that multi-cycle dropouts are frequent, provisions are made for mounting an uninterruptible power supply.

Sun Enterprise 3500

The Sun Enterprise 3500 is an enhanced Enterprise 3000 with five system board slots (one more than the Enterprise 3000), a faster 32X CD-ROM, an optional tape drive, a 100 MHz capable Gigaplane system interconnect, and support for eight internal FC-AL disk drives. The optional second peripheral power supply (184 watt) found in the Enterprise 3000 has been redesigned and moved from the back of the system to the front. Although the dimensions of the power supply are changed to fit its new physical space, its functionality remains unchanged and offers even greater capacity (195 watt).

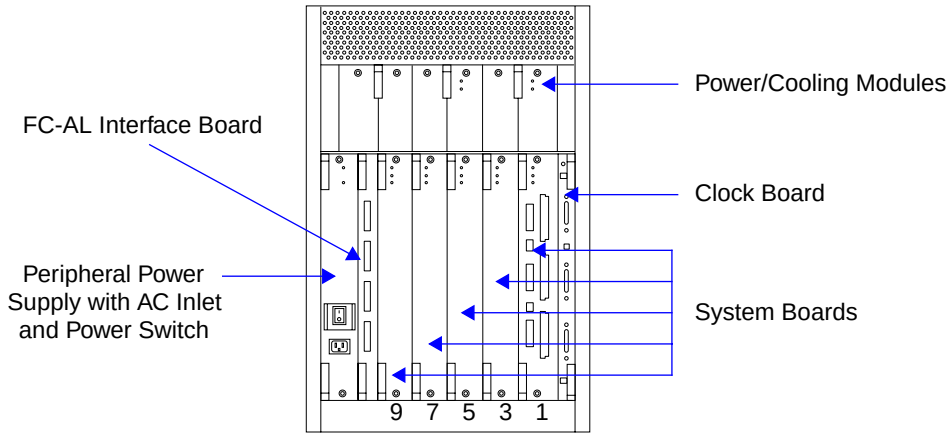

Figure 2-2 Sun Enterprise 3500 rear view

Accessible from the rear of the system are the five system board slots, the clock board, and the FC-AL interface board (Figure 2-2). Each system board slot can contain a CPU/memory board or an I/O Board. The FC-AL interface board contains up to four Gigabit Interface Converter (GBIC) modules, providing

connectivity from a host adapter on an I/O board or SBus card to the internal disk drives. The clock board has its own slot and does not use one of the five slots for the CPU/memory or I/O boards.

The peripheral power supply (PPS) and power/cooling modules (PCMs) are also accessible from the rear of the system: The PPS with AC inlet and power switch is located at the lower left of the cabinet, and the PCMs are located above the system board slots. Entry-level 3500 configurations contain one PCM located above slots 1 and 3. If a second PCM is required, it is installed above slots 5 and 7, to the left of the first PCM. A fan try, located above the peripheral power supply, is also included an any entry-level configuration. A third PCM can be used for redundant power in a fully loaded system. To install this third PCM, the existing fan tray (located to the left of the second PCM) is removed, and the third PCM is inserted in its place.

The second peripheral power supply, CD-ROM drive, tape drive and FC-AL disk drives are all located on the front of the system (Figure 2-3). Internal disk bays accommodate up to eight (currently 9.1 GB) FC-AL disk drives divided into two banks of four disk bays each. All internal drives in the Enterprise 3500 are individually hot-swappable.

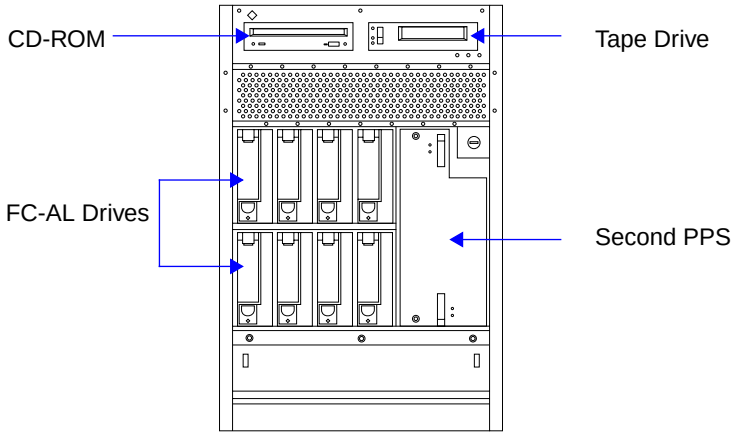

Figure 2-3 Sun Enterprise 3500 front view

Sun Enterprise 3500 Internal FC-AL Architecture

The Enterprise 3500 internal FC-AL architecture includes two separate banks of disks, each containing four FC-AL disk drives (Figure 2-4). A FC-AL Interface Board containing up to four GBIC modules provides an external interface to the host adapter. These GBIC Modules are identical to the ones used in the StorEdge A5000, the FC-AL SBus host adapter, and on the Enterprise X500 SBus I/O Board.

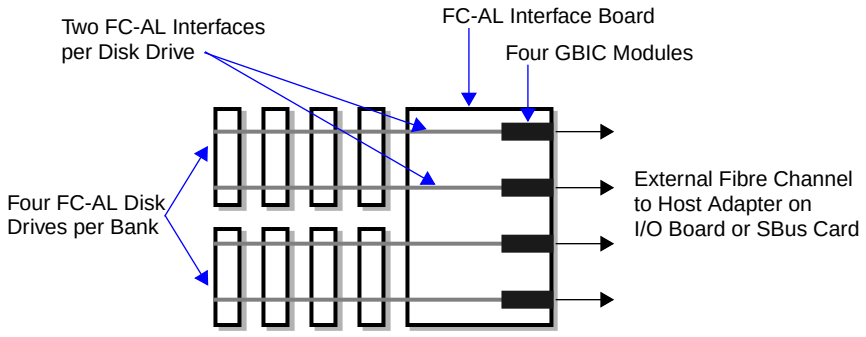

Figure 2-4 Sun Enterprise 3500 internal FC-AL block diagram

Each of the two disk banks can have one or two FC-AL loops connecting to the installed drives, providing highly available *dual loop* configurations. Since each of the banks can be configured with dual loops, it is possible to configure the system with as many as four FC-AL loops to the disk storage.

The FC-AL connections on the Interface Board are logically independent. The components do get their power through a single connection. However, the power to the Interface Board comes from the backplane which can be supported by redundant power supplies, eliminating the single point of failure.

Sun Enterprise 4500

The Sun Enterprise 4500 offers eight centerplane slots in a compact enclosure. The eight slots can be used for CPU/Memory, I/O or Disk Boards, with four boards installed from the front of the cabinet, and four boards installed from the rear (Figure 2-5). Typically, the I/O boards and the disk boards are installed from the rear to facilitate cabling.

Slots for the power/cooling modules are located in the card cage next to the board slots. Up to four power/cooling modules can be installed, two from the front and two from the rear. Power/cooling modules must be installed adjacent to CPU/Memory, I/O, and disk boards.

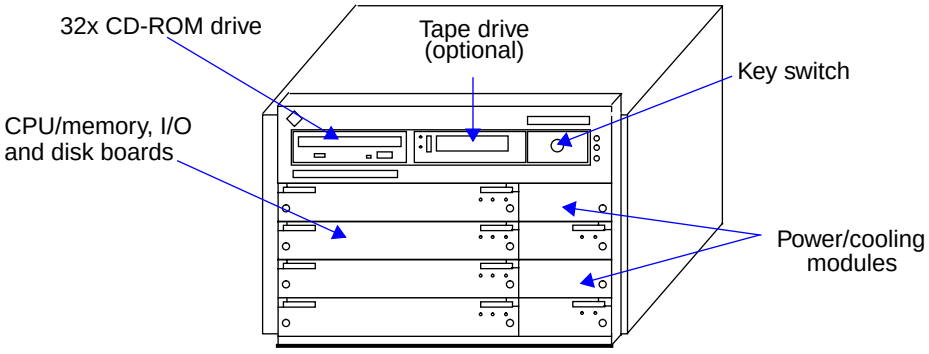

Figure 2-5 Sun Enterprise 4500 system cabinet

The 32X CD-ROM drive is located at the upper left front of the enclosure. Next to the CD is a slot for an optional half-height tape drive, plus the key switch and system LEDs.

The peripheral power supply (184 watt) is located at the top rear of the card cage, with the clock board located below. The clock board has its own slot, and does not utilize one of the eight board slots used for the CPU/Memory, I/O and disk boards.

Sun Enterprise 5500

The Sun Enterprise 5500 server is architecturally identical to the Enterprise 4500, but is rack-mounted in a 68-inch datacenter system cabinet. This enclosure provides room for additional disk storage subsystems and a tape library. Two Sun Enterprise 4000 or 4500 servers can also be mounted in the system cabinet. An optional second power sequencer is also offered to provide dual power sources and support for more devices.

A Sun StorEdge CD32 drive is installed in the upper left front of the system cabinet (Figure 2-6). Next to the CD drive is a slot for an optional half-height tape device. An optional Sun StorEdge FlexiPack removable storage tray, or a Sun StorEdge Tape Library, or one FC-AL hub tray can be installed above the card cage.

The following storage devices can be installed in the Enterprise 5500 system cabinet below the card cage:

- Up to four Sun StorEdge A5000s
- Up to six Sun StorEdge RSM disk trays
- Up to four SPARCstorage Array model 100 series subsystems

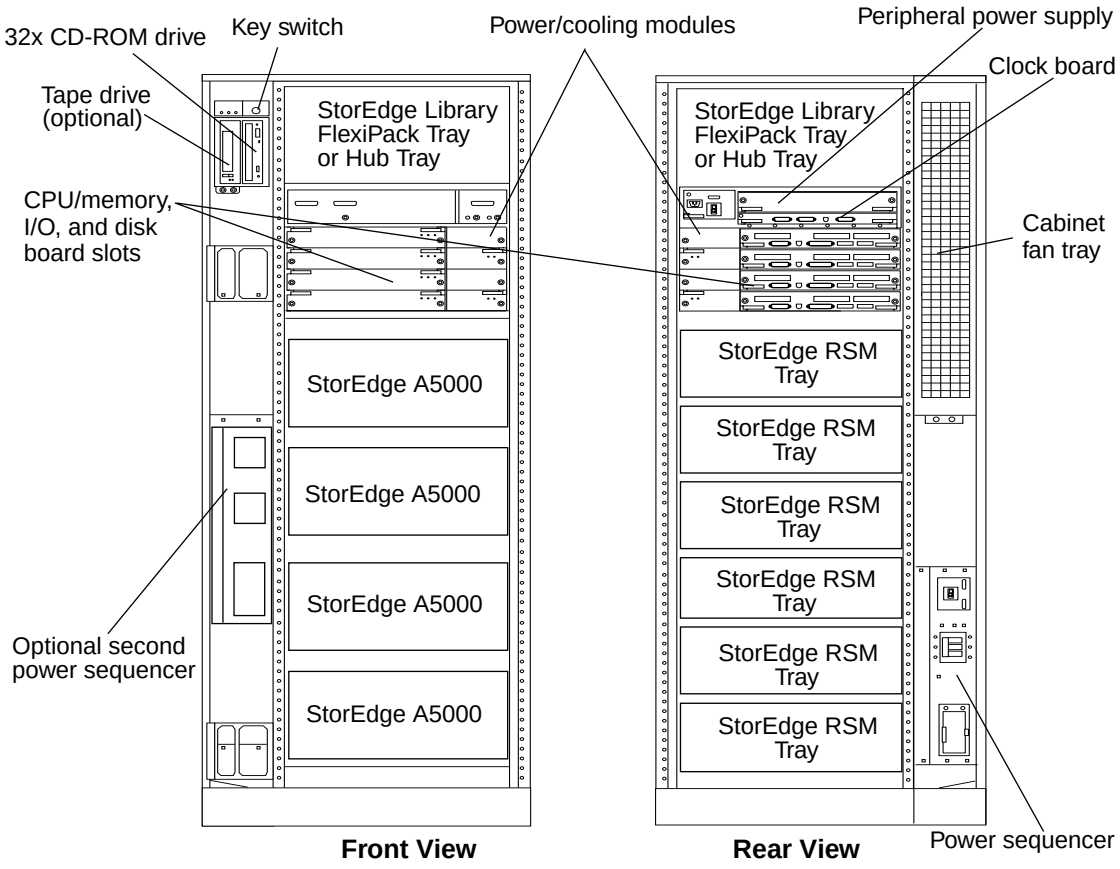

Figure 2-6 Enterprise 5500 system cabinet

Maximum Enterprise 5500 configurations containing four StorEdge A5000s and one hub tray require mounting the hub tray in the upper portion of the cabinet, where the tape library is typically installed. This maximum configuration, therefore, precludes the inclusion of a tape library.

For detailed information on configuring the Enterprise 5500 system rack, please contact your local Sun representative.

Sun Enterprise 6500

The Sun Enterprise 6500 server is housed in the same 68-inch system cabinet as the Enterprise 5500 (Figure 2-7). Inside the cabinet is a 16-slot card cage for CPU/memory boards, I/O boards, and disk boards. This card cage is essentially the same as in the eight-slot system, but with twice the number of board slots. An optional second power sequencer is also offered to provide dual power sources and support for more devices.

As with the Sun Enterprise 5500 system cabinet, a Sun StorEdge CD32 drive is installed in the upper left front of the system cabinet (Figure 2-6). Next to the CD drive is a slot for an optional half-height tape device. An optional Sun StorEdge FlexiPack removable storage tray, or a Sun StorEdge Tape Library, or one FC-AL hub tray can be installed above the card cage.

Slots for the power/cooling modules are located in the card cage next to the board slots. Up to eight power/cooling modules can be installed, four from the front and four from the rear. The peripheral power supply is located at the top rear of the card cage. The clock board is located in the card cage below the peripheral power supply. The clock board has its own slot and does not use one of the 16 board slots.

The following storage devices can be installed in the Enterprise 6500 system cabinet below the card cage:

- Up to three Sun StorEdge A5000s
- Up to five StorEdge RSM disk trays
- Up to three SPARCstorage Array model 100 series subsystems

Two Enterprise 4000 or 4500 servers can also be installed in the Enterprise 6500 system cabinet below the card cage.

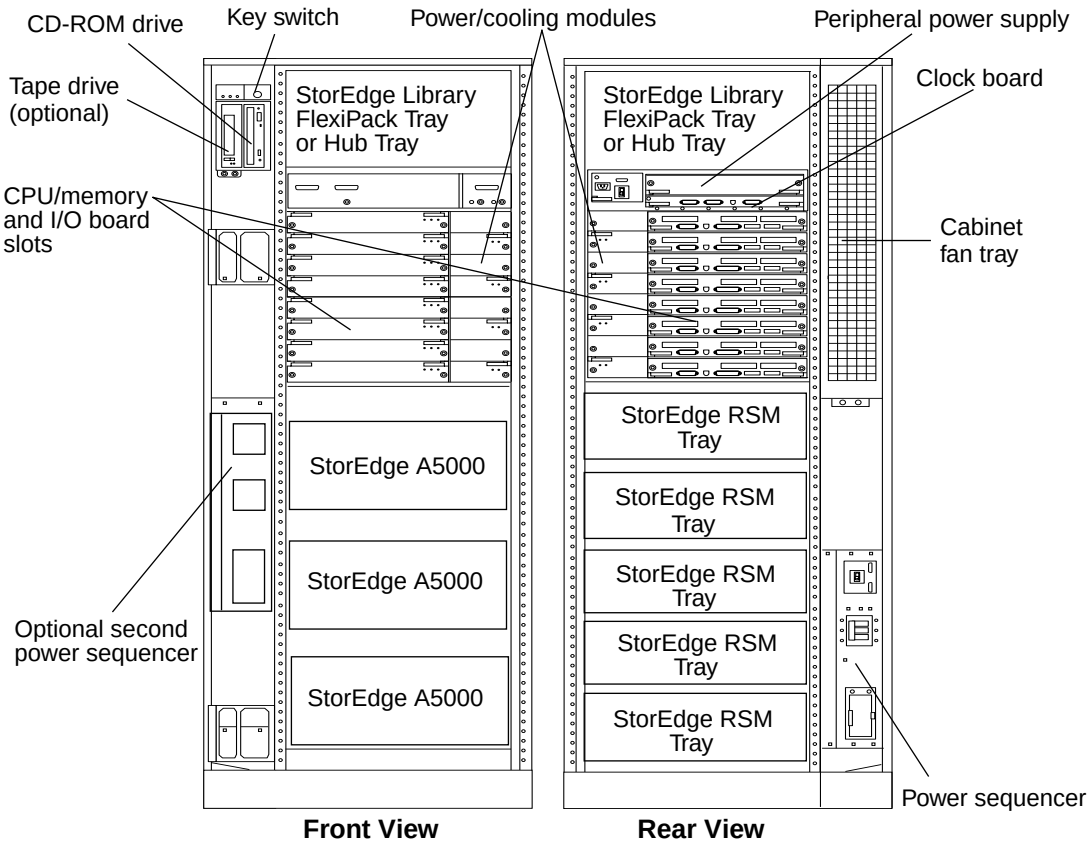

Figure 2-7 Enterprise 6500 system cabinet

Maximum Enterprise 6500 configurations containing three StorEdge A5000s and one hub tray require mounting the hub tray in the upper portion of the cabinet, where the tape library is typically installed. This maximum configuration, therefore, precludes the inclusion of a tape library.

For detailed information on configuring the Enterprise 6500 system rack, please contact your local Sun representative.

System Bus Architecture 3

To provide sustainable high-performance, the Sun Enterprise X500 servers must exchange data quickly and efficiently. One of the simplest system organizations provides a central bus, to which each subsystem is connected. In addition to being versatile, this arrangement is very simple. The primary design issue is providing sufficient system bus performance to ensure that client subsystems are adequately supplied with data. If the central bus is a bottleneck, every component of the system becomes starved, and degraded performance is the result.

Many systems use a variety of point-to-point buses and other specialized mechanisms to carry data around the system. Often, one bus connects the processor(s) to memory, while I/O is provided with a separate memory interface, and interrupts are delivered from I/O devices to the processor by yet another mechanism.

Sun's Ultra workstation systems employ such a design, using a crossbar switch called the Ultra Port Architecture (UPA) to a memory controller and an SBus-to-UPA interconnect (SYSIO) peripheral interface chip. The crossbar switch is responsible for routing data between the various components.

Enterprise X500 systems use a similar overall architecture, but the implementation is slightly different, due to the much higher bus bandwidth required to support large multiprocessor configurations — as well as the much greater physical distances involved in configuring more than forty discrete functional units. Instead of a small-scale crossbar switch, Enterprise X500 systems are designed around a system bus known as the *Gigaplane*. The Enterprise X500 centerplane carries the Gigaplane bus to all parts of the system

(Figure 3-1). CPU / memory and I / O boards are replicated and connected to the Gigaplane until a sufficiently large system is configured.

In order to accommodate as many as thirty UltraSPARC processors and multiple high-performance I / O buses, the Gigaplane has extremely high throughput — 3.2 GB / sec at 100 MHz. Experience with a previous generation server, the SPARCcenter 2000 (*sun4d*), showed that a small but significant class of I / O operations and applications are extremely sensitive to memory latency (the end-to-end period of time required to service a memory request), and substantially improved memory latency was a primary Gigaplane design criteria.

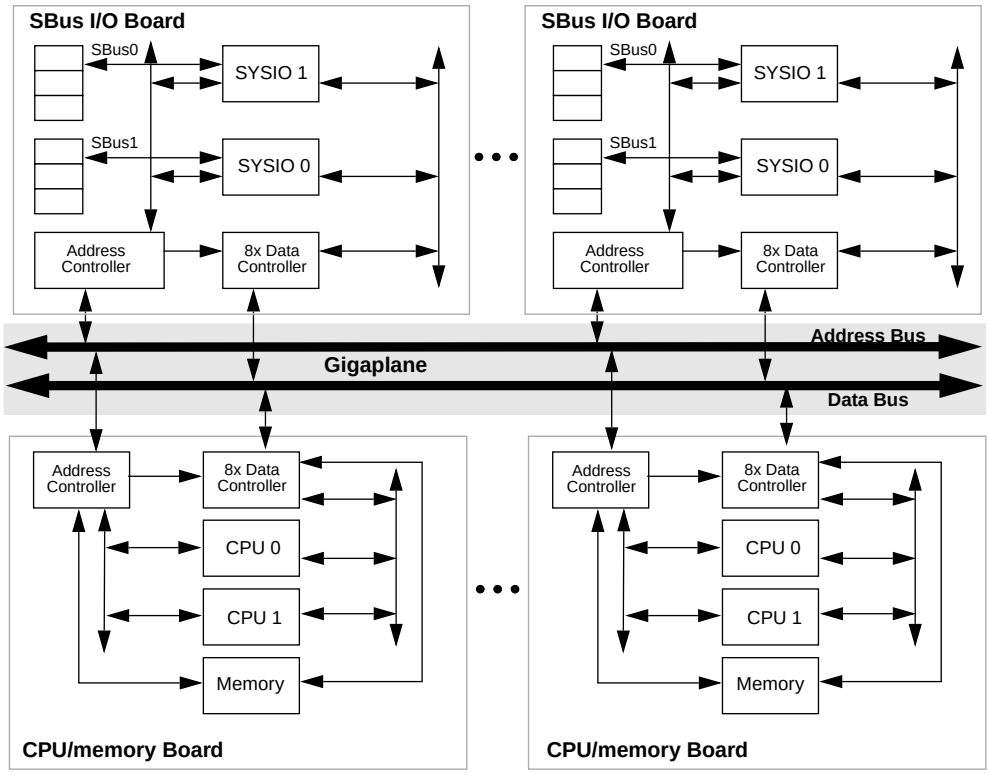

Figure 3-1 Sun Enterprise X500 system architecture. CPU/memory and I/O boards are simply replicated until a sufficiently large system is configured.

Like its predecessor *sun4d* architecture, the Enterprise X500's centerplane bus complex consists of a hierarchy of buses. Each component is designed for simplicity, high capacity, reliability, and efficiency. A single mechanism is used to transfer data between CPU and memory, between memory and the I/O buses, and to transmit interrupts all at rates which ensure that a system remains balanced as the configuration is expanded.

One of the key innovations permitting the development of the unified desktop and server architecture is the use of the UPA interconnect to attach the individual units. This arrangement permits the sharing of crucial desktop circuitry and software. The large-scale system is then constructed by connecting individual modules with the fast, high-capacity Gigaplane bus.

UPA Mezzanine Bus

Because the Enterprise X500's centerplane bus is both very wide and runs at a very high clock rate, it has an electrical limit of sixteen connections. Some mechanism is required to support the extensive configurations in Enterprise X500 systems — a full configuration can consist of 30 processors, 30 memory banks and two I/O buses, resulting in a total of 62 required connections. To address this problem, Enterprise X500 uses a variant of the two-level bus scheme found in the SPARCcenter 2000 and SPARCserver 1000. Instead of connecting processors, memory and I/O buses directly to the Gigaplane centerplane, each CPU/memory board and I/O board uses a UPA interconnect switch to *fan out* from the Gigaplane to individual units.

The same bus architecture used in Sun's UltraSPARC-based desktop systems, UPA is a physically short, 128-bit packet-switched bus which permits connection of up to three devices. The UPA implementation provides two 128-bit ports and a DRAM controller interface. On CPU/memory boards, the two 128-bit ports are used to connect two processors. The ports can be software configured to operate in 64-bit mode to match the width of I/O buses. Each I/O board connects one or two I/O buses to the Gigaplane. In some circumstances, peripherals requiring exceptionally high throughput or very low latency access to memory can connect directly to the UPA-64 slave port. This feature is used in the Graphics I/O board to ensure very high levels of graphics performance.

The UPA is fully synchronous, both internally and to the system centerplane. Because all processors and branch *buses* run synchronously, a significant reduction in design complexity for both the UPA and the Gigaplane is possible.

The simpler design results in a higher overall clock rate and consequent higher performance on both buses. In the Enterprise X500 implementation, the UPA runs at the same speed, up to 100 MHz, as the Gigaplane. This yields a UPA burst speed of 1.8 GB/sec and sustainable bandwidth of over 1.2 GB/sec[1] (CPU/memory board). The synchronous nature of the UPA is particularly effective in reducing CPU-to-memory latency. UPA delivers data to the UltraSPARC processor in 28 bus clock cycles, compared to 90+ bus clock cycles on SuperSPARC/XDBus and 40 bus clock cycles on SuperSPARC/MBus.

Gigaplane Implementation

The Gigaplane design uses a highly optimized 256-bit (32 byte) packet-switching design with separate 42-bit address lines. The primary unit of memory transfer — a single cache line — is 64 bytes, enabling a cache miss to be filled in just two bus cycles. This minimizes delays due to bus contention even in the event that the bus approaches saturation. At 100 MHz clock speed, the Gigaplane delivers a throughput of 3.2 GB/sec. In Enterprise X000 servers, the maximum clock speed is 84 MHz, delivering 2.68 GB/sec throughput. For comparison with Sun's earlier generation technology, the dual XDBuses in the SPARCcenter 2000 are 128 bits wide, with a cache subblock size of 64 bytes, and a cache line fill taking nine bus cycles. The burst transfer rate on a single 50 MHz XDBus is 400 MB/sec. Table 3-1 lists the comparative specifications of the Gigaplane and XDBus.

Specification	XDBus	84 MHz Gigaplane	100 MHZ Gigaplane
Data Width	64 Bits	256 Bits	256 Bits
Address Width	36 Bits (64 GB)	42 Bits (2 TB)	42 Bits (2 TB)
Clock Rate	50 MHz	84 MHz	100 MHz
Unit of Transfer	64 Bytes	64 Bytes	64 Bytes
Data Transfer Cycles	8 Cycles	2 Cycles	2 Cycles
Address/Arbitration Overhead	2 Cycles (Address), 1 Cycle Arbitration	0 Cycles (Separate bus)	0 Cycles (Separate bus)
Throughput	400 MB/sec (peak)	2.68 GB/sec	3.2 GB/sec

Table 3-1 Comparison of Gigaplane and XDBus specifications

1. The UPA bus width is 128-bit on the CPU/memory board but is configured to operate in 64-bit mode on the I/O boards. As a result, the peak and sustained throughput in each case is different.

The main Gigaplane clock is generated on the clock board, and each component board interfaces to the passive centerplane using a paired Address Controller (AC) and Data Controller (DC). Because of the 256-bit width of the Gigaplane bus, the DC on each board is implemented using eight data controller chips operating in a parallel bit-sliced arrangement. A small local bus called the Boot Bus connects the Address Controller with its Data Controllers; the Boot Bus connects all boards with a low-speed utility bus, in effect parallel to the high-performance Gigaplane. A simple interface called the Boot Controller generates the Boot Bus and connects it to local devices on each board such as the Flash-EPROM and a small amount of local RAM. These ten chips — Address Controller, Boot Controller, and eight Data Controllers — form the characteristic *wall* of cooling towers on the centerplane edge of each board.

Packet-Switched Design

One of the significant characteristics of the Gigaplane is that it is a packet-switched bus, rather than a circuit-switched bus. Packet-switching permits the bus to operate at much higher system-wide throughput because it eliminates "dead" cycles on the bus; the ability to complete transactions in any order (rather than the order in which they were issued) allows the connection of arbitrarily slow devices without overall performance degradation (in this context, "slow" refers to SBus I/O which moves at *only* 200 MB/sec).

A packet-switched bus permits substantially greater overall throughput than comparable circuit-switched buses by separating bus requests from their corresponding replies, a method known as *split transactions*. Most bus transactions consist of two logical parts: first the bus master specifies the target address, and then the data is actually transferred. In a circuit-switched bus, the transaction is an indivisible operation. The requestor arbitrates for the bus, places the target address on the bus and holds the bus while the request is serviced. Much of the service time of a request is spent performing non-bus activity (e.g., memory latency). In a circuit-switched bus, this time and the corresponding bus utilization is wasted.

A packet-switched bus is much more efficient: a client arbitrates for the bus, sends a request packet which specifies the target address and then releases the bus. While the request is being serviced, the bus is free to perform transfers on behalf of any other bus client, including units with outstanding requests. The bus protocol requires that clients process requests in the order in which they arrive. The transaction is said to be split because an arbitrary number of other

packets or transactions can use the bus between a transaction's request and reply. Each packet is tagged so that requests and replies can be properly associated.

High Efficiency Implementation

Most buses, even packet-switched buses, use a single set of wires for bus control and arbitration, address transmission and data transfer. The bus is used in turn for arbitration, addressing and then data transmission. Because the bus is dedicated to one function at a time, the effective data rate across the bus can be much lower than the burst transfer speed

Gigaplane uses dedicated wires for control, address, and data, permitting near-total data utilization on the bus. Control and addresses can be transferred at the same time as data. For example, requests to read data from memory can be transmitted concurrently with responses from previous read requests. Because this permits three transactions to be in-flight on the bus simultaneously, it is possible to utilize virtually the full theoretical throughput of the Gigaplane. (This is the reason why only one throughput value is specified for any given Gigaplane clock speed.) Earlier buses multiplexed data and address lines in order to save the expense associated with the independent address lines. Gigaplane uses 42 address signals, along with 256 data signals and 32 ECC signals. Despite the very large address space available on Gigaplane, the extra expense associated with dedicated control and address lines is necessary to avoid the design problems associated with running a physically long and wide bus at even higher clock rates.

Another design criteria was to maintain response from memory, and in particular to deliver data from memory with low latency. Compared to previous generation systems, the Enterprise X500 systems are nearly twice as fast to memory in terms of processor cycles, despite the fact that both the processor clock and the bus clock are additionally more than twice as fast.

Single-Bus Implementation

One of the overriding Gigaplane design criteria was the need for simplicity. Among the lessons learned from the *sun4d* architecture was that performance was significantly degraded by many of the complex features that went into its flexibility and configurability. For example, the asynchronous interface between the CPU modules and the XDBus permitted the processor to run at a

clock speed that is unrelated to the system backplane, but this feature was a significant contributor to the long memory latency associated with those platforms. In the Enterprise X500, the CPU modules run at increments of the Gigaplane speed.

Because of the requirement to handle more functional units (thirty CPUs compared to twenty, and thirty I/O buses compared to ten) — combined with the much higher performance demanded of each unit — the Gigaplane had to eliminate as much pipeline complexity as possible. One of the casualties of this drive for simplicity was the dual buses found in the SPARCcenter 2000. Coordination of the interleaved XDBuses lengthened the XDBus pipeline, limiting its clock rate. Moreover, the duplex nature of the buses required considerable extra silicon on the SPARCcenter 2000 system boards (two very large ASICs were required to attach each processor and each SBus to the dual XDBus), making those boards much more complex. Furthermore, the dual bus on the SPARCcenter 2000 meant that its field replaceable units were not interchangeable with the single-bus (and much lower cost) SPARCserver 1000. The simplified Gigaplane bus delivers the performance required to support very large configurations. About half of the improvement in speed comes from the greater bus width; the remainder derives from the higher clock rate made possible by design simplification.

Jumperless Geographical Installation

Because the Enterprise X500 centerplane is implemented with programmable device addressing, system boards may be plugged anywhere into the centerplane without physical reconfiguration. There are no jumpers or switch settings associated with any configuration. This permits easier maintenance and reduces the likelihood of installation errors.

Reliability

In addition to being less complex than its predecessor, the Gigaplane bus is protected against loss or corruption of data by a full implementation of ECC codes on the bus's data lines and parity on the control and address lines. The SEC-DED-S4D ECC is able to correct single errors, and detect two bit or four-bit nibble errors (the same capability as the predecessor SPARCcenter 2000). Although the bus's electrical environment is relatively tightly controlled under

normal circumstances, ECC provides confidence in data transfer, even in environments which involve hot pluggable power supplies, processors, and I/O boards.

Cache Coherency Protocol

All high-performance functional units found in modern microprocessor-based computers use caches to reduce bus traffic and increase system throughput. When these units are configured in a shared-memory multiprocessor, keeping all of the caches consistent requires support in the bus protocol. In Enterprise X500 systems, the processor modules and I/O buses use caches in their centerplane interfaces.

Caches store copies of part of the memory image. These blocks are noted to be in one of three states: *shared*, *unshared*, or *invalid*. Writes to unshared blocks do not cause bus traffic; data is simply updated in place. An update to one copy of a shared block requires that every other copy must either be updated or invalidated. How writes to shared blocks are handled is called the *cache coherency protocol*.

All of Sun's multiprocessor systems, like most other commercial multiprocessor systems, use a technique called *bus snooping* to maintain consistency in multiple caches. Under this scheme, each cache monitors (snoops) transactions on the bus looking for updates to shared blocks which it possesses.

The Gigaplane implementation uses a variation called *write-invalidate snooping*. When a cache performs a write to a block that is shared, it issues a broadcast on the bus indicating that other copies of the block are now invalid. The cache that performs the write retains its copy. When multiple caches update a block, each cache must reload a copy of the block after another cache invalidates it. In this case, the entire cache block must be transferred. This arrangement is most efficient when writes to blocks are issued in batches by a single cache.

CPU/Memory Board Design 4

One of the fundamental boards in the Enterprise X500 family is the CPU / memory board (Figure 4-1). Each Enterprise X500 CPU / memory board is capable of running at 100 MHz, and can contain zero, one or two UltraSPARC processors and up to 2 GB of memory. (Boards with no UltraSPARC modules can be added to systems requiring additional memory capacity.) CPU / memory boards utilize the standard Gigaplane bus interface to connect to the system centerplane.

Mixing of X000 and X500 CPU / memory boards is supported. The Enterprise X500 CPU / memory board is backwards compatible with the 84 MHz capable board of the Enterprise X000 and is supported in Enterprise X000 systems. The X000 CPU / memory board can also be used in the Enterprise 4500-6500 systems (not supported in the Enterprise 3500 server), although this naturally limits the Gigaplane speed to 84 MHz.

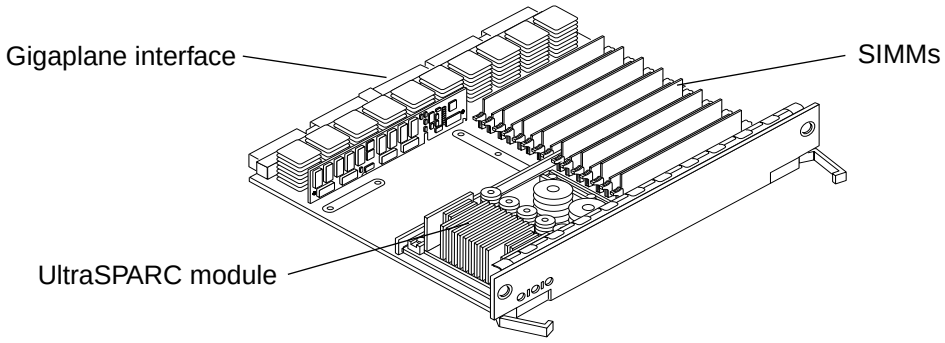

Figure 4-1 Sun Enterprise X500 CPU/Memory Board

The internal architecture of the CPU/memory board is shown in Figure 4-2. The address controller and data controllers connect directly to two banks of memory, each capable of holding eight 32 MB or 128 MB SIMMs. Zero, one or two banks may be populated on any given board, for a maximum memory capacity of two GB per board. The UPA interconnect connects the two UltraSPARC processors, with their matching 4 MB external caches, to the system centerplane.

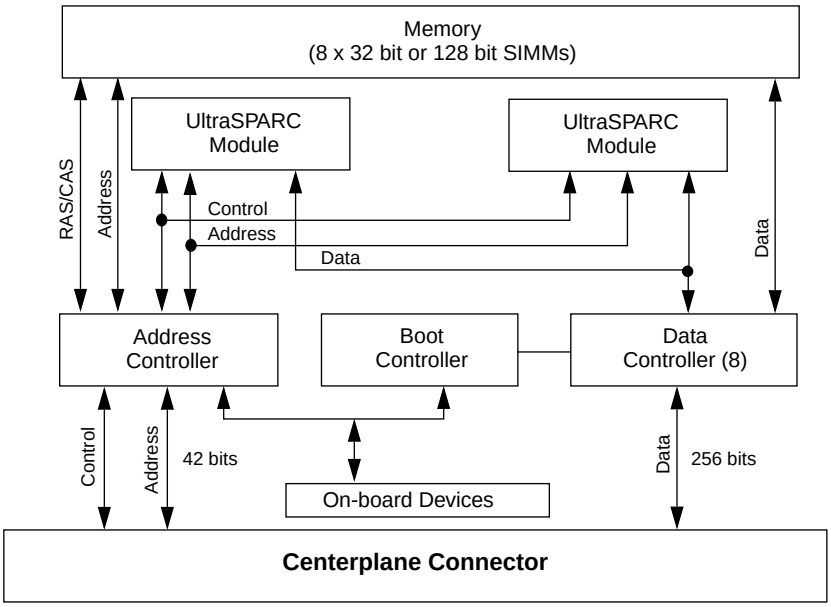

Figure 4-2 Internal architecture of the Sun Enterprise X500 CPU/Memory Board

The UltraSPARC processors are mounted on daughter card modules. This design enables processors to be quickly, easily and inexpensively upgraded, with substantial room to grow compute performance, both by adding more processors and more CPU/memory boards, and by adding faster processors over time.

The CPU/memory boards have temperature sensors located under each UltraSPARC module. These sensors control the fan speed in the adjacent power/cooling module and allow the actual temperature of the individual boards to be monitored through Enterprise SyMON™.

UltraSPARC Design

Enterprise X000 systems were the first large-scale SMP server systems to utilize the UltraSPARC processor. Introduced initially at 167 MHz, the UltraSPARC-II running at 336 MHz is the latest version of the powerful UltraSPARC processor. With a faster clock speed and a larger 4 MB external data cache, the UltraSPARC-II is functionally identical to the UltraSPARC processor.

UltraSPARC-II is the second generation in a family of high-performance superscalar 64-bit SPARC processors, designed to deliver very high performance in demanding workloads. UltraSPARC supports a 44-bit virtual address space (16 TB per process) mapped into a 41-bit physical address space (2 TB). The external bus interface is 128 bits wide, with a carefully tailored cache and memory interface designed to maintain superscalar throughput in spite of the inevitable cache misses.

Support for 32-bit data types is identical to previous SPARC V8 processors, permitting complete backwards compatibility. Existing SPARC V8 binaries run unmodified on UltraSPARC systems.

UltraSPARC Implementation

As with the entire Enterprise X500 design, the UltraSPARC was designed with simplicity and high performance in mind. For example, the on-chip caches were simplified from 4-way or 5-way set associative to two-way or direct-mapped, and the number of implemented bus interfaces was reduced from two to one.

The UltraSPARC design includes two 16 KB caches, a memory management unit (MMU), an external cache controller, a UPA bus interface, an extensive register file, dual translation lookaside buffers (TLBs) and nine discrete execution units. The processor core consists of two integer units, five floating-point / graphics units, a load / store unit, and a branch processor. Most integer instructions complete in one cycle, permitting back-to-back dependent instructions to be issued in consecutive cycles. Most floating-point operations complete in three cycles (the exceptions are division and square root). Any combination of two integer instructions, two floating point instructions, a load / store and a branch (up to four instructions total) can be issued each clock cycle.

The pipeline consists of nine stages. The integer portion is six stages (two more than its SuperSPARC predecessor), three of which are devoted to floating-point and graphics operations. The last stage of the pipeline commits results to the register file, but a bypass mechanism makes results available immediately (i.e., results are available while the register commit takes place) if they must be used for the next computation, avoiding the delays associated with writing to the register file.

Branch Prediction and Speculative Execution

Experience with first-generation superscalar processors has shown that conditional branches cause significant pipeline *bubbles* and correspondingly lower performance. The UltraSPARC design goes to considerable lengths to avoid this problem. The solution consists of a mechanism to predict which way a conditional branch is expected to transfer. This is matched with the ability to speculatively execute instructions based on the prediction of the branch direction.

The instruction cache stores a two-bit field for each conditional branch instruction. The processor updates the state to reflect the most recent branch taken or not taken. The most recent history of the branch is a very strong predictor of behavior the next time the branch is encountered. Two bits are used to handle the case of looping branches, which a single-bit history frequently mispredicts.

This mechanism permits the processor to continue to fetch and execute instructions along the predicted branch path. When the branch is not predicted (or if it is mispredicted), the processor stalls, waiting for the pipeline to refill. References to the external cache suffer delays of six clock cycles, and the processor can execute as many as four instructions every cycle, so every stall represents a loss of performance.

The dynamic branch predictor is able to correctly forecast branches nearly all the time (branch predictions are correct nearly 90 percent of the time on common benchmarks). Execution of instructions which might not actually be involved in the final computation is called *speculative execution*. The processor maintains internal state, remembering when it is executing speculatively. In the relatively unlikely event that the branch was mispredicted, the erroneously executed results are annulled. The UltraSPARC can speculatively execute as

many as 18 instructions (typically in 6 to 9 clock cycles) before a branch is resolved, providing compilers with ample opportunity to optimize the code to avoid pipeline stalls.

Out-of-Order Completion

The UltraSPARC pipeline enables several instructions to be dispatched each clock cycle. Although most integer operations complete in a single cycle, some do not; moreover, some of the floating point instructions can take a substantial number of clock cycles to complete (floating point square root is a notable example). Most traditional processors are organized so that instructions are required to complete in the same order they are dispatched. For scalar processors this is not an issue, since the processor can only execute one instruction at a time anyway, and some early superscalar processors opted for this simplified design.

For a superscalar processor with high performance requirements, it is unreasonable to demand in-order completion, because this wastes many cycles in the pipeline. For example, consider a clock cycle in which two integer adds, a floating point square root and a branch instruction are dispatched. If the processor requires in-order completion, none of the instructions can complete until all four have completed. This would be particularly unfortunate, because three of the instructions complete in a single cycle, while the square root might take as many twenty cycles. Out-of-order completion permits the integer and branch units to execute other instructions while a long-latency instruction completes.

UltraSPARC does not make use of the similarly-named but much more complex strategy called out-of-order dispatch. In the UltraSPARC, instructions are always dispatched in the order they appear in the instruction stream. Some processors attempt to dynamically reorder the instruction stream in an attempt to more fully optimize superscalar pipelines. The UltraSPARC architecture design team opted not to implement this feature in order to keep the processor design simple enough to run at high clock rates.

Real World Performance

Many design improvements over the SuperSPARC have been integrated into the UltraSPARC. Many of them are aimed at improving raw single-threaded performance, as represented by SPECint95 and SPECfp95, and their impact shows up in excellent SPEC performance, but others are squarely aimed at improving real-world performance.

For example, dramatic improvements in register window handling are significant to systems which handle very high I/O loads and which perform many context switches. Previous SPARC architecture specifications define a single type of exception. This meant that any nested trap — such as an interrupt occurring during a context switch or page fault — required a relatively large amount of work to resolve.

Because UltraSPARC implements five levels of traps, it permits greatly simplified trap handling. UltraSPARC provides separate traps for regular execution, system calls, common operating system routines, the page fault handler, and a emergency RED_mode handler. The additional facilities substantially streamline trap handling. For example, the register window overflow trap handler is reduced from about 300 instructions on SuperSPARC to just 20 instructions on UltraSPARC. The RED_mode handler is used to precisely define the system environment during catastrophic errors, permitting cleaner and more robust error recovery. As a result, context switches are much more efficient in machine cycles; and, perhaps more importantly, crucial system routines are greatly simplified and easier to understand and maintain, leading to higher overall reliability.

In addition to multi-level traps, UltraSPARC provides four sets of eight 64-bit global registers, used as quick scratch registers when interrupts or traps such as Memory Management Unit (MMU) traps occur. The use of scratch registers eliminates expensive register save/restore sequences at every interrupt or trap as is necessary on SPARC V8 systems. Database applications, with their very large active address spaces, are particularly sensitive to MMU traps and interrupt handling performance. Finally, each bank of registers is equipped with a dirty bit; register banks that are not modified need not be saved to memory during a context switch.

Separate VM Facilities

Virtually all modern processors use a translation lookaside buffer (TLB) to reduce the size of the page tables to a manageable size while maintaining high performance. The TLB preserves the most recent virtual-to-physical address mappings. Since memory access tends to be strongly localized, retaining the few most recent translations in a high speed cache permits the processor to maintain high performance while operating in a full virtual address environment.

The UltraSPARC design uses a variation of this scheme. Two different TLBs are provided, one dedicated to each of the on-chip caches. Experience with the SuperSPARC processor showed that the reference patterns of the two caches were sufficiently different to warrant providing a separate TLB for each. This is most advantageous to the instruction stream, where the strongly sequential access pattern is quite predictable and is easily susceptible to optimization in the TLB.

Memory Interface and Cache Design

Another important part of a processor's design is its memory interface. All processors must exchange data with memory, and because of the large gap between processor and memory speed, optimization of the CPU / Memory interface is crucial.

The UltraSPARC includes an on-chip 32 KB Harvard architecture cache, with 16 KB each dedicated to instructions and data. One of the distinguishing attributes of the UltraSPARC on-chip cache is that it is non-blocking: the load / store unit is able to continue issuing memory requests even after a cache miss is detected. This means the on-chip cache is always paired with a second-level cache. Because the secondary cache is mandatory, the cache controller is integrated into the processor chip. The non-blocking nature of the on-chip cache is important because it permits extensive buffering and careful compiler optimization to substantially reduce access latency to the large external cache, and permits access to the external cache to be heavily pipelined. This design sustains a data transfer in every clock cycle at the full processor clock rate.

Instruction Cache

The instruction cache is a 16 KB two-way set associative cache that is designed to deliver instructions to the execution units at the rate of four per cycle. Instructions are pre-fetched and loaded into a 12-entry input buffer, based upon the results of the dynamic branch prediction. The input buffer is normally able to hide the latency of an instruction cache miss, assuming that the miss can be filled from the external cache. Both on-chip caches use a line size of 32 bytes.

Data Cache

The on-chip data cache is a 16 KB direct-mapped cache that does not allocate on a write miss. The data cache uses a 32 byte line size with 16 byte subblocks; the smaller subblock corresponds to the 128-bit width of the UltraSPARC's external cache interface. The non-allocating design minimizes cache pollution, preserving precious cache memory for data which is demonstrably valuable. For example, newly-created data, which is not normally reused within a few clock cycles, is not a good candidate for cache residency.

The data cache is equipped with load and store buffers that are used to pipeline access to the external cache. When a data cache miss occurs, the data is fetched from the external cache while the next instructions are executed without stalling. The load buffer can handle up to nine outstanding loads without stalling the processor. Combined with a minimal six-cycle latency to fetch data from the external cache, the processor is able to access cached data at the rate of one load every cycle. This represents a substantial improvement over other designs when accessing large data structures such as those found in finite-element analysis and computational fluid dynamics.

Writes out of the data cache are also buffered, using an eight-entry store buffer which equals the throughput of the load rate of the instruction cache — one store per clock cycle (this is about twice the rate of most processors). Combined with the low latency of the external cache, the store buffer is able to hide most of the latency of the external cache.

The store buffer also permits loads to take priority over stores on the memory interface. As long as the store buffer is not in danger of overflowing, loads are more important than stores because the processor is unable to proceed without the data from a read. Stores can be deferred to some degree while the processor continues execution. The store buffer has priority over the load buffer after five

stores accumulate. Finally, a feature known as *store compression* collapses together back-to-back writes in the same 16-byte block, replacing the pair with a single 16-byte write. This improves latency for both operations and reduces overall memory traffic.

External Cache

Even though the UltraSPARC has on-chip caches, it is always mated with a high-performance external cache. The current UltraSPARC modules include a larger, 4 MB external cache. This external cache is organized into 64-byte cache lines with 16 byte subblocks. The secondary cache holds both instructions and data. The external cache communicates with the rest of the system via a 128-bit UPA interface.

The on-chip data cache is always operated in write-through mode (i.e., writes to the data cache always write back to the secondary cache). Regardless of their size, the secondary caches are direct-mapped and are operated in write-back mode, in order to minimize centerplane traffic.

Considerable attention was paid to minimizing the latency of processing cache data. Latency from the external cache is limited to six clock cycles for a hit in the external cache, about 260 percent faster than SuperSPARC-II. (It is worth noting that the UltraSPARC's six-cycle external cache latency is comparable to what some competing processors pay to access their internal caches!) The minimal latency means that extensive load and store buffering between the external cache and data cache / instruction cache can hide most external cache latency, resulting in a cache subsystem which can deliver a 16 byte data cache subblock every clock cycle.

The UltraSPARC external cache operates on physical addresses, avoiding the problems associated with aliasing between different virtual address spaces. Because there is no aliasing, there is no need to flush cache lines upon context switch—the physical addresses remain constant across contexts. This reduces memory traffic and complexity compared to otherwise equivalent virtually-addressed caches.

High Performance Register File

One of the most serious design problems in the earlier SuperSPARC processor family was the processor's use of a register file with relatively few read / write ports. In order to sustain superscalar performance, the register file had to be

accessed twice per clock cycle. While this design saved crucial die area, it also severely restricted the processor's overall clock rate to less than 100 MHz. With a design goal in excess of 300 MHz, the UltraSPARC team had to develop an alternative register file implementation. The UltraSPARC uses an innovative cell organization which permits a ten ported register file. Additional ports yield greater parallelism within the processor core. Because the ports are not multiplexed each cycle, the file is not an obstacle to increased clock rate.

Another key performance feature is a much larger floating point register file than previous generations. UltraSPARC has 32 double-precision floating point registers, compared to 16 double-precision FP registers on previous processors. Earlier processors also had 32 single-precision floating point registers, but floating point is now dominated by double-precision operations. Because RISC processors operate on data only when in registers, the size of the register set is crucial (CISC processors are able to operate directly on memory, but at a considerable increase in internal complexity and consequent loss of performance). Larger register files permit the processor to retain more data in registers without generating memory traffic. The greatly enlarged floating point register file contributes significantly to UltraSPARC's substantially improved floating point performance. The larger floating-point register file is one of the few UltraSPARC features that cannot be utilized without recompilation.

Instruction Set Enhancements

The UltraSPARC processor implements a proper superset of the standard SPARC V9 instruction set, and code generated for other V9 systems will run unmodified on Enterprise X500 systems and their desktop UltraSPARC counterparts. However, the V9 definition includes a block of opcodes that is reserved for implementation-specific use. The UltraSPARC design uses this feature to accommodate two additional groups of instructions designed to accelerate bulk data manipulation and graphics/visualization. These are collectively called the VIS instruction set.

Bulk Data Operations

The Enterprise X500 design is optimized for use as a server, although the design makes an excellent high-end graphics workstation when equipped with framebuffers. With one primary exception, the VIS™ instruction set is primarily of use on graphics workstations. However, the lone exception is of crucial

importance: it is the block copy accelerator. Server systems must transfer data much more frequently than typical workstations, and a reasonable first approximation of a UNIX server's performance is the speed of its `bcopy(3)` routine. When the VIS block copy instruction is used, an UltraSPARC processor running at 336 MHz in an 84 MHz system can copy data at a sustained 240 MB/sec (such a copy operation requires 480 MB/sec of memory bandwidth).

From a processor's perspective, data which is moved in large blocks is unlikely to be accessed again soon, so the UltraSPARC does not copy data into cache when it loads from memory, nor does it copy data to cache on writes with this instruction. At the extremely transfer high speed made possible by the block copy mechanism, even large caches would be quickly overwritten with relatively infrequently used data, a circumstance sometimes known as cache pollution.

User-Accessible Block Copy

Unlike previous SPARC processors, the UltraSPARC block copy accelerator is implemented as an unprivileged instruction, and is fully accessible to user applications. The system libraries in Solaris detect when they are operating on a suitably equipped processor and use the VIS acceleration when appropriate. The difference will be most noticeable in applications which are very copy-intensive, such as DBMS systems or applications which make heavy use of the socket library. (DBMS systems operated in client/server mode often use the socket library for transport, and such applications are greatly accelerated by the VIS instructions.)

VIS Instruction Set

The VIS instruction set design applies the RISC principle to graphics: build a few of the most common operations directly into the hardware, and make those operations very fast — and leave the rest to the compiler. The multimedia capabilities of UltraSPARC are designed to accelerate 2-D and 3-D graphics, video compression/decompression, and display and image manipulation. Support is provided for 8-bit and 16-bit color and alpha information, permitting easy handling of 24-bit information as well as higher resolution images such as those required for medical or color-prepress imaging. For pixel interpolation (used in scaling or rotating objects in 3-D) and alpha blending, packed integer operations permit four or eight pixels to be

processed every clock cycle. Many of these operations required fifty or more cycles on previous processors, but UltraSPARC is able to complete them in a single cycle.

Instructions are also provided to support motion estimation, the computationally-intensive portion of video compression. The process for eight pixels requires eight subtractions, eight absolute values, eight additions, a load of eight pixels, an align of eight pixels and an addition. UltraSPARC performs this entire operation for eight pixels in one single-cycle instruction, compared to 48 instructions and well over 100 cycles on other processors. UltraSPARC also enables full-speed MPEG-2 compression and decompression without add-on hardware. The VIS instruction set is designed with major video and still-image formats in mind, including H.261, MPEG-1 and MPEG-2, and JPEG.

Finally, the block move capabilities provide the ability to store and retrieve video images at full motion speeds. When the framebuffer's display memory resides in the processor's virtual memory address space, as it does for the Creator Graphics framebuffer, the UltraSPARC is able to transfer double-buffered video data to the framebuffer at full motion speeds. For example, a 24-bit 1024x1024 display refreshed at 30 frames per second requires 120 MB/sec of copy bandwidth, far less than the bandwidth available per processor.

Although it seems logical to provide concurrent support for audio streams, no direct support for audio manipulation is provided because audio data streams require vastly less data manipulation than graphics. For example, full CD-format stereo sound requires only 176 KB/sec of data, while a 640 x 480 x 24 bit video display refreshed at 30 frames per second requires 27 MB/sec—about 135 times as much as the accompanying audio data.

Applications can exploit VIS instructions by making use of Sun graphics libraries. For example, most of the video processing operations reside in the XIL libraries. Likewise most geometric manipulation routines such as screen coordinate transforms and Z-buffering are handled in XGL™ or OpenGL libraries. As long as the application dynamically links to these libraries, VIS acceleration is made available transparently. Highly optimized block copy routines such as `bcopy(3)`, `bzero(3)` and `memcpy(3)` are also made available through dynamically linked library routines. In practice, recompilation is not necessary to obtain most of the benefits of the VIS instruction set enhancements.

Memory Subsystem

Large-scale systems must provide sufficient memory capacity to sustain high performance from the processors and I/O channels. Additionally, memory must be quickly accessible in order to avoid interfering with other subsystems' activities. Finally, the large concentration of data likely in a datacenter system necessitates a design which is highly reliable.

As with virtually every other facet of the overall design, the memory system was closely scrutinized for opportunities to provide higher performance and greater reliability through simplification. The selected design uses commodity SIMMs, organized in groups of eight to provide the requisite throughput. The complex (and expensive) memory crossbar used on sun4d SIMMs was eliminated by the simple expedient of providing a very wide data path. In the Enterprise X500 memory unit, an entire line of the external cache (64 bytes + ECC, 576 bits) is obtained in a single cycle.

The Enterprise X500 memory controller manages two banks of memory on each CPU/memory board. Each bank of memory consists of eight standard JEDEC SIMM modules, implemented in 3.3 volt (3.3v) CMOS. Unfortunately, the 64 Mbit DRAMs used to construct the 128 MB SIMMs are only available in 3.3v, so earlier parts are not compatible. Future DRAMs will also be available only in 3.3v form.

Using currently-available 64 Mbit DRAMs (i.e., 8 MBytes per chip), SIMMs contain 128 MB, while a bank contains 1 GB — each CPU/memory board thus holds 2 GB. The memory controller is capable of handling up to 1 Gbit DRAMs, ensuring future growth potential. Because the Gigaplane bus has only 41 bits of address space, the maximum configurable physical memory is limited to 2 TB.

High Performance Memory

One of the overriding design criteria of modern microprocessor-based systems is to provide sufficient access to main memory. The Enterprise X500 memory bank is very "wide" — each bank can deliver 512 bits (64 bytes) in a single cycle, meaning that a 64-byte cache miss can be filled by a single memory bank. Despite the high density of the chips in use, physical DRAM chips are organized within the SIMM module so that the failure of an entire DRAM

appears as a series of single-bit errors, rather than a large block of vacant addresses. Single-bit errors are detected and corrected by the ECC scheme, permitting the system to proceed even when entire memory chips fail.

Extensive Interleaving

Individual DRAMs are not able to provide data on a continuous basis. After each access, the chip must spend time recovering before handling the next access. Interleaved memory helps reduce the cost of memory access by permitting multiple memory components to operate in parallel. One way to increase memory performance is to arrange for one bank to supply data while other banks are recovering. In interleaved memory schemes, memory is divided into *n* banks arranged so that every *nth* byte is supplied by a different memory bank. In a two-way interleaved system, the first doubleword is supplied by bank 0 while the second is supplied by bank 1; normally the size and extent of interleave is arranged so that a single typical request is satisfied by as many banks as possible. This permits a single memory request to be fulfilled without waiting for memory recycle time.

In the Enterprise X500 architecture, memory is interleaved on 64 byte boundaries. Enterprise X500 systems permit an aggregate interleave of 16-way, assuming sufficient CPU/memory boards and memory configuration. Extensive interleave is required to support the extremely high memory access rates possible with UltraSPARC and high-throughput SBus and PCI peripheral buses. With high speed DRAM chips configured in very wide memory arrays, single memory banks deliver sustained bandwidth of about 500 MB/sec.

Enterprise X000 and X500 systems extend upon the earlier generation systems by imposing fewer restrictions on interleave configuration. Earlier systems configured interleave only when memory banks were identical. This meant that a SPARCserver 1000 system with two 32 MB banks and two 256 MB banks would use a two-way interleave between the 32 MB banks and two-way interleave between the 256 MB banks. An Enterprise X000 or X500 with two 256 MB banks and two 1 GB banks configures four-way interleave on the two 256 MB banks when combined with the first 256 MB of the 1 GB banks, and the remaining memory is two-way interleaved. This permits considerably higher interleave performance in systems that have many different memory densities.

I/O Subsystems 5

Most servers are called upon to handle massive amounts of data, and to make that data available to other systems. Accordingly, the I/O subsystems are a crucial component of high-performance servers. Enterprise X500 systems deliver superior I/O performance, while also providing excellent configurability and unequalled flexibility.

Currently, three different, interchangeable I/O boards — SBus, Graphics, and PCI — are available, offering a wide range of I/O connectivity. Servers can be configured with any combination of I/O boards. Thus, a single server can support both PCI and SBus cards.

Mixing of X000 and X500 I/O boards is supported. The Enterprise X500 I/O board is backwards compatible with the 84 MHz capable board of the Enterprise X000 and is supported in Enterprise X000 systems. The X000 I/O board can also be used in the Enterprise 4500-6500 systems (not supported in the Enterprise 3500 server) although this naturally limits the Gigaplane speed to 84 MHz.

Flexible I/O Implementation

All three I/O boards share a common basic design. Each board uses the standard Gigaplane interface to connect to the centerplane. And, as with the CPU/memory boards, the address controller provides four UPA ports. These ports connect to the ASICs which implement the SBus or PCI Bus.

All I/O boards in Enterprise X500 servers are hot-pluggable. As a result, new I/O boards can be installed and I/O boards that have been de-configured from the system can be removed while the system is on-line. System availability is further improved by dynamic reconfiguration software that eliminates the need for a server reboot.

The I/O boards also offer comprehensive environmental monitoring. Temperature sensors are located on each I/O board, allowing the actual temperature of individual boards to be monitored with Enterprise SyMON.

SBus Implementation

In addition to the integrated devices, the Enterprise X500 SBus I/O board offers three SBus slots on two independent SBuses. Each SBus is a full implementation of the industry-standard IEEE 1496 SBus. The SBus supports burst transfers in all defined sizes from 1 byte to 64 bytes. The SBus clock is fixed at 25 Mhz, yielding a burst transfer speed of 200 MB/sec per SBus. Sustained data transfers operate at up to 120 MB/sec per SBus.

Although the SBus fully implements Extended Mode (i.e., 64-bit) transfers, the SBus is fully backwards compatible with the large installed base of 32-bit SBus boards. Extended mode operation can be selected on a per-slot basis, permitting arbitrary mixing of 32-bit and 64-bit boards anywhere in the system. The SBus is able to sustain about 65 MB/sec when operating in 32-bit mode.

All three data transfer modes provided in the *sun4d* architecture — *programmed I/O*, *consistent mode direct virtual memory access* (DVMA), and *streaming mode DVMA* — are fully supported on Enterprise X500 systems. The SBus interface chip permits independent selection of operating modes for each SBus slot, allowing device drivers to select the most appropriate transfer mode for each task.

Consistent mode DVMA and streaming mode DVMA use an I/O cache to achieve maximum performance. The two differ primarily in the way they present the memory model to the SBus. In consistent mode, all stores issued by an SBus board are guaranteed to be observed in issuing order by all processors. The practical impact is that an SBus operating in consistent mode permits only one pending transaction between each SBus slot and system memory at any given moment. This mode is provided for SBus boards or drivers that require automatic maintenance of a completely consistent memory image.

Streaming mode DVMA is the most efficient form of transfer, permitting all slots on each SBus to overlap transactions to the system centerplane. This mode uses multiple streaming buffers for each SBus slot to buffer transactions into and out of the SBus. Because these buffers are outside the cache coherency domain of the Gigaplane centerplane, consistency must be managed by the device driver by invalidating buffers at the beginning of the DVMA read and flushing buffers at the end of a DVMA write.

Enterprise X500 systems protect data and addresses on each SBus with parity checksumming. I/O buses connecting to the SBus, such as SCSI-2 and Fibre Channel, are also protected by parity.

Low-Contention Implementation

Because the design criteria for the SBus include low cost and simplicity of implementation for peripheral designers, the SBus is a circuit-switched bus. As a result, SBus clients face increasing contention for access to the bus when the number of clients on the bus increases or when the bus must be shared with slow clients. To avoid these issues, internal SBus resources, such as the buffers used to implement streaming mode I/O, are dynamically allocated to SBus slots. This permits the SBus to provide better service to high-performance peripherals without wasting resources on slower or inactive devices.

I/O MMU

The SBus operates in virtual memory space, meaning that the SBus accepts virtual addresses rather than physical addresses from the I/O devices performing DMA transfers. Each SBus has a dedicated I/O Memory Management Unit (I/O MMU) to perform the requisite virtual-to-physical memory address translation. The I/O MMU is implemented in the SBus-to-UPA interconnect (SYSIO) chip, logically placed between the SBus and the I/O cache interface to the UPA.

The I/O MMUs are completely separate from the MMUs which reside in each processor. As in previous systems, the operating system is responsible for maintaining consistency between the processor's MMUs and the other SBus I/O MMUs.

Like the *sun4d* design, the I/O MMU provides each SBus with a 64 MB DVMA address space. This large DVMA address space provides the operating system with flexibility in choosing I/O buffers, simplifying I/O processing.

Additionally, because each SBus is completely independent, their DVMA spaces may be configured separately, potentially allowing over 1 GB of DVMA space in the system.

SBus I/O Board

The Enterprise X500 SBus I/O board provides a total of three SBus slots, as well as built-in 10 and 100 Mbit/sec Fast Ethernet, Fast/Wide SCSI-2 support, and two 100 MB/sec Fibre Channel Arbitrated Loop (FC-AL) sockets (Figure 5-1). The FC-AL sockets enable connection to the new Sun StorEdge A5000 disk arrays and internal disks in the Enterprise 3500.

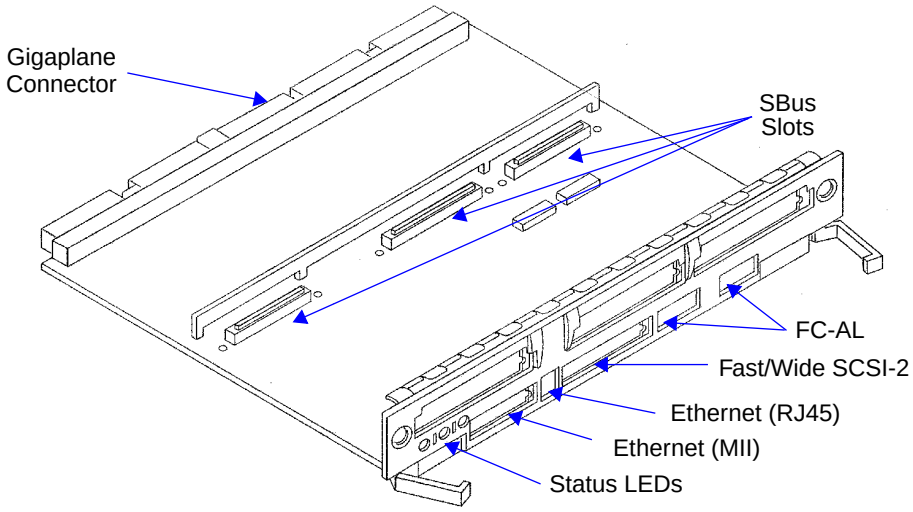

Figure 5-1 Sun Enterprise X500 SBus I/O board

The Fast Ethernet interface on the I/O board provides access to Category 5 twisted pair through an RJ45 connector. To support a wider array of cabling options, it also features access to a Media Independent Interface (MII). Accessible through a 40-pin, miniature “D” connector, the MII allows adoption to any other form of Ethernet, including ThickNet (AUI), twisted pair, ThinNet, or Fiber.

One of the significant features in the Enterprise X500 I/O subsystem is the use of dual SBus channels on each SBus I/O board. With dual 64-bit SBus channels on each SBus I/O board, the maximum bandwidth for an SBus I/O board in an Enterprise X500 server is 400 MB/sec.

The on-board Fast/Wide SCSI-2 connector of the first I/O board (in slot 1) cannot be used in the Enterprise 5500 and 6500. This SCSI host adapter supports the internal CD-ROM and tape drives, and the SCSI bus length prevents the support of additional devices. The SCSI bus length in Enterprise 3500 and 4500 systems is much shorter, allowing the use of the on-board connector.

Figure 5-2 shows the internal organization of the SBus I/O Board. The on-board devices are split across the two SBus buses— slot 0 is paired with the Fast Ethernet and SCSI-2, while slot 1 and slot 2 are on the same bus with the Fibre Channel controller.

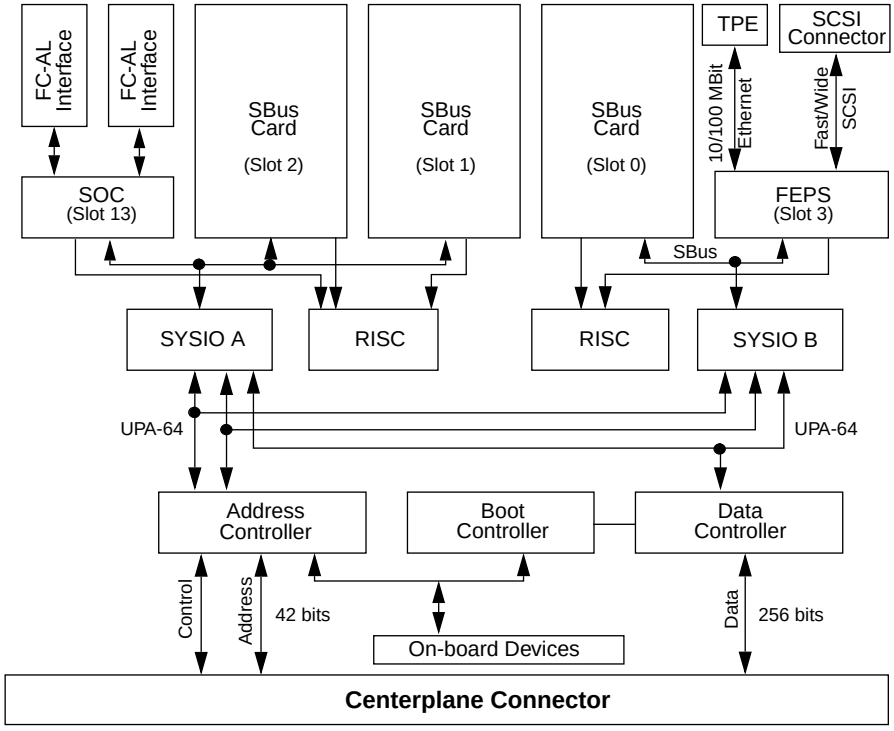

Figure 5-2 Internal organization of the SBus I/O board

The same UPA implementation is used on the I/O Boards and CPU/memory boards, except that the ports on the I/O Boards are only 64 bits wide to accommodate the 64–bit I/O buses. Thus the I/O Boards are restricted to transferring at 753 MB/sec, compared to 1.5 GB/sec on the CPU/memory boards. However, this restriction is not a bottleneck for the SBus I/O board which needs only a bandwidth of 400 MB/sec to fully support the two SBus channels.

Graphics I/O Board

One of the most demanding I/O applications is graphics: framebuffers demand higher I/O bandwidth than almost any other peripheral. Although the performance of the SBus has improved dramatically in recent implementations, the most suitable high-bandwidth, low-latency architectural location for a framebuffer remains directly in the memory subsystem.

The Graphics I/O board is similar in design to the SBus I/O board. Like the SBus I/O board, it contains two 100 MB/sec FC-AL sockets, a Fast/Wide SCSI-2 port, and 10/100 Mbit/sec FastEthernet support (Figure 5-3). It differs from the SBus I/O board by providing one UPA slot, one SBus channel and two SBus slots.

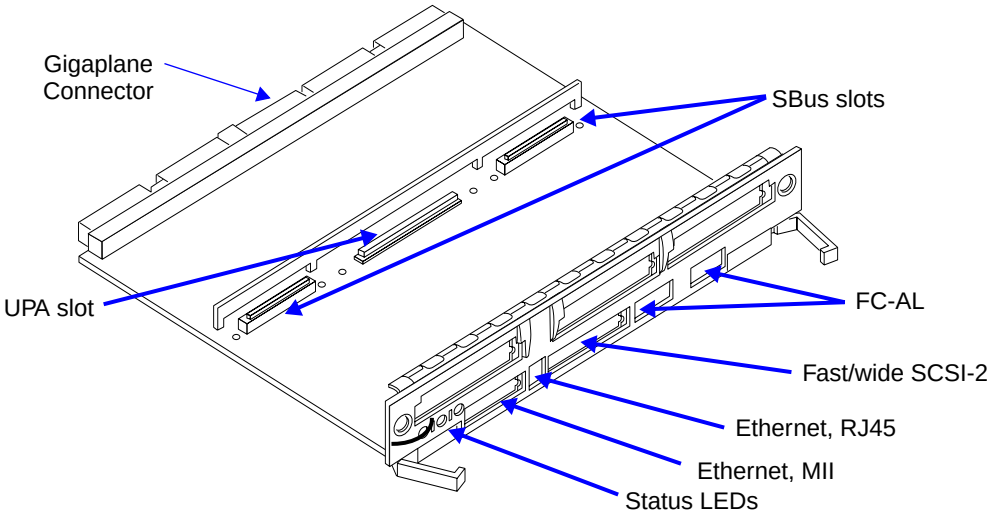

Figure 5-3 Sun Enterprise X500 Graphics I/O board

The internal organization of the Graphics I/O board is illustrated in Figure 5-4. As seen in the figure, one of the UPA ports connects to the same SYSIO SBus interface used on the SBus I/O board. The other UPA port is directly terminated in a standard UPA-64 connector and is externally visible as the center slot. The Creator and Creator3D framebuffers are supported in this UPA slot. Framebuffers have full access to the UPA interfaces if necessary, although even the current Creator3D requires far less bandwidth.

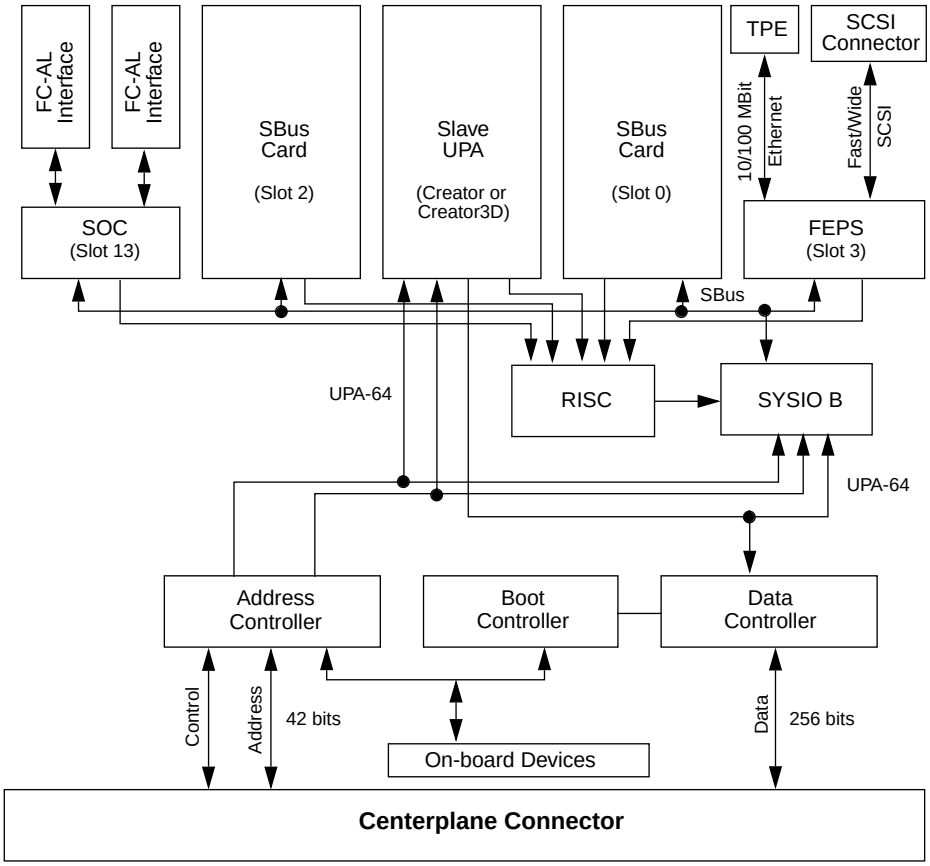

Figure 5-4 Internal organization of the Sun Enterprise X500 Graphics I/O board

Although performance is usually the first reason cited for placing the framebuffer in memory, this architecture also provides graphics subsystem designers with access to much larger memory systems.

Texture mapping applications benefit most from this technique, as very large texture maps can be made available to geometry processing units without imposing transfers across peripheral buses. Large geometry models also benefit, because physical space need not be physically reserved on the frame buffer for memory adequate to hold a large model

Creator and Creator3D series 3 Frame Buffer

The Creator3D frame buffer features a third-generation high-performance design that is quite different from other products which deliver similar performance. The Creator3D provides extremely competitive performance (3.7 million 3-D vectors / second, 1.2 million triangles / sec) with a very simple design, its major components being the Frame Buffer Controller (FBC) ASIC, frame buffer memory, the UPA interface, and RAMDAC memory-to-analog converter (Figure 5-5).

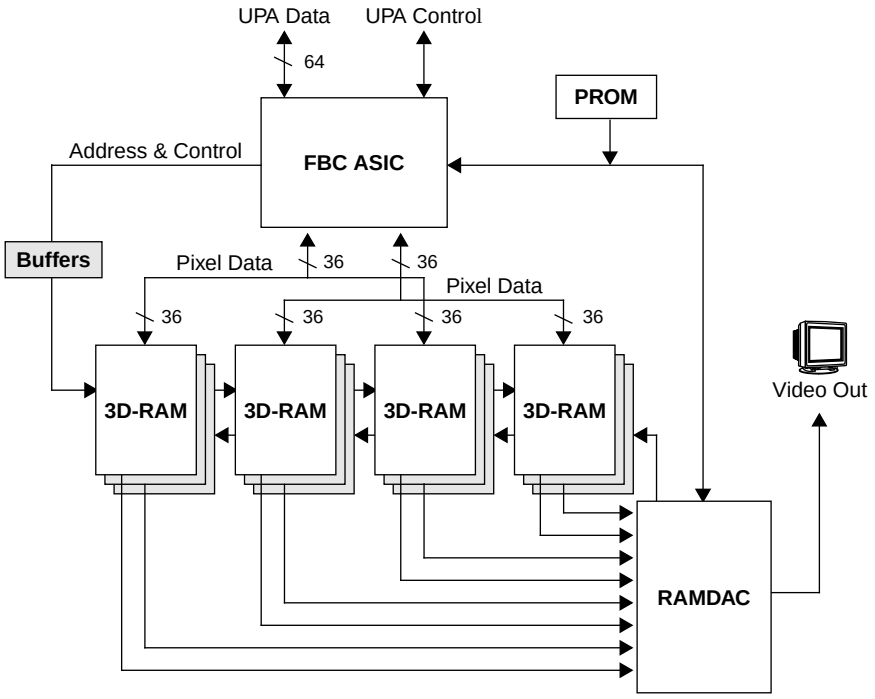

Figure 5-5 The Creator/Creator3D frame buffer

The Creator series 3 graphics subsystems provide the third generation and an unparalleled level of acceleration and scalability for common graphics operations required by windowing, imaging, video, and 2D and 3D graphics applications.

Sun has long recognized that graphical performance cannot be considered independently from system design issues such as system throughput, processor speed, and memory bandwidth. As a result, Sun took a new approach with Creator Graphics by treating graphics as a fundamental part of the platform architecture, and as an integral part of the overall system design.

Creator Graphics replaces separate, specialized frame buffers with a single high performance, low cost architecture. The Creator Graphics architecture represents an innovative combination of modular design and provides very high levels of system integration.

Creator Graphics' design leverages powerful system functionality provided by high performance UltraSPARC processors, fast system memory, and high speed system interconnects. This approach maximizes performance while avoiding expensive duplication of resources.

With continually increasing graphics rendering and manipulation requirements, it is essential that graphics subsystems deliver scalable performance. Creator Graphics systems address scalability in a number of ways which provide transparent acceleration to existing applications.

As a function of their integrated design, Creator Graphics systems take advantage of high performance integer and floating point capabilities and the unique VIS instruction set of UltraSPARC processors for many graphics operations. As newer, faster versions of the UltraSPARC processor become available, graphics performance in Creator Graphics systems simply increases.

For additional information, see the *Creator Graphics Architecture White Paper.*

5

PCI Connectivity

In addition to its commitment to expand the capacity and performance of all of its systems, Sun is continually looking for ways to increase their openness and standards compliance. Sun has chosen to support PCI on the Enterprise X500 servers, and other systems, for a variety of reasons:

- *PCI is an open, architecture-independent bus*

  Because PCI, like SBus, is open and shipping in volume, it has been adopted by both consumers and producers of computer hardware. As a result, the potential exists for a large number of platform-independent peripherals to be supported.

- *PCI is fast*

  The PCI bus architecture is designed to provide high performance, with its I/O performance a key differentiator from other bus architectures. Running at 33 MHz and 66 MHz, PCI offers configurations that meet a variety of developer and user needs.

- *PCI is standardized*

  PCI is a standard bus architecture that has been adopted by the high volume personal computer industry. Because of its wide acceptance, PCI promises that compliant adapter cards will be available from more sources than ever before.

PCI is a performance-oriented I/O bus optimized for high speed data transfers. Used as an interconnect between highly integrated components and subsystems, such as peripherals, add-on boards, and memory systems, PCI ensures that the Enterprise X500 systems will have the industry leading flexibility, compatibility, performance, and investment protection expected of products from Sun.

PCI I/O Board

The Enterprise X500 PCI implementation is based on the industry standard PCI specification version 2.1, and supports the following specifications:

- 33 MHz (standard) and 66 MHz buses
- 32-bit or 64-bit cards
- 5 volt cards (33 MHz bus)
- 3.3 volt cards (33 & 66 MHz bus)
- 6.875 inch (short) cards, maximum width 4.2 inches

The Sun Enterprise X500 PCI I/O board provides built-in 10/100 Mbit/sec Fast Ethernet and 20 MB/sec Fast/Wide SCSI-2 support, as well as providing two PCI slots for I/O expansion (Figure 5-6).

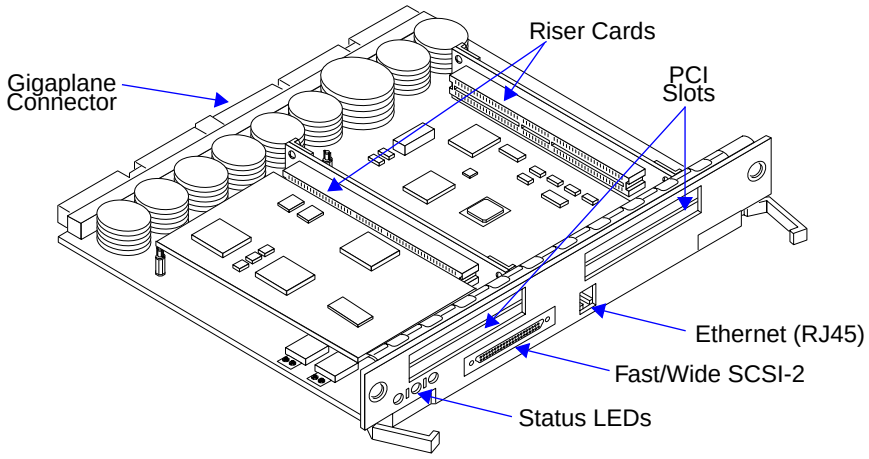

Figure 5-6 Sun Enterprise X500 PCI I/O board

The internal organization of the X500 PCI I/O board is shown in Figure 5-7. Each PCI I/O board has dual 528-MB/sec PCI-66 (66-MHz, 64-bit) channels, each leading to a PCI slot. There are also two 264-MB/sec standard PCI channels. One standard channel leads to the on-board 10/100 Mbit/sec Ethernet, and the other leads to the on-board Fast/Wide SCSI-2. The 100 MHz

UPA interface that connects the PCI I/O board to the Gigaplane has a throughput of 774 MB/sec. When run at 84 MHz, the UPA interface has a throughput of 668 MB/sec.

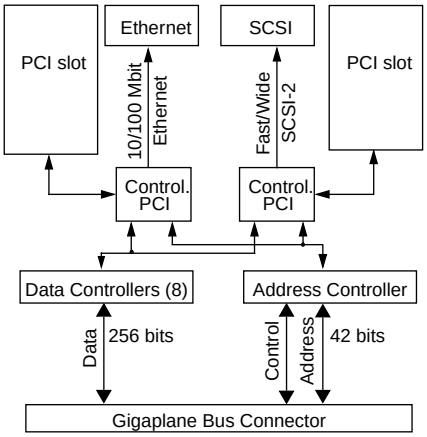

Figure 5-7 Internal organization of the Sun Enterprise X500 PCI I/O board

Reliability, Availability, Serviceability 6

Reliability, availability, and serviceability (RAS) continue to be highly important goals of the Enterprise X500 design. These goals are met in a number of ways: Environmental sensors, hardware monitoring, and ECC circuitry increase system reliability. Automatic System Recovery, dynamic reconfiguration, alternate pathing, redundant power and cooling, and hot-swappable components provide high levels of system availability. The Enterprise SyMON system monitoring tool, remote control capabilities, and a modular system design increase overall serviceability. This powerful combination of features enables the Enterprise X500 servers to provide superior levels of RAS, unmatched by the current generation of competitive servers.

The key Enterprise X500 RAS features are summarized in the remainder of this chapter. For more detailed information, see the technical white paper *Availability Features in the Sun Enterprise 3500 to 6500 Server Family.*

Designed for Reliability

The Enterprise X500 server family is inherently designed to be reliable. The system has fewer active components than previous-generation servers, increasing the mean time between failures (MTBF) of the system. MTBF is further increased by a simple, elegant centerplane bus design that reduces the amount of bus circuitry required, and the passive nature of the centerplane/backplane design.

Environmental Sensors

Environmental sensors have been built in to the servers to increase system reliability. Each CPU/memory board and each I/O Board contains a software-readable thermal sensor which detects overtemperature conditions. These thermal sensors are located directly below the processor modules on CPU/memory boards, and near the SYSIO ASIC on I/O Boards. In addition to the thermal sensors, there are also built-in rotational speed sensors. If a fan fails, or is detected to be operating at too slow a speed, the system is notified and the remaining fans increase their speed as necessary to provide adequate air flow.

If the thermal sensors indicate an overtemperature condition, they trigger a system response that makes the fans run faster — cooling the boards and allowing the system to continue to run uninterrupted. A notice is also sent to the administrator, allowing them to take proper actions. These environmental sensors are also used in conjunction with CPU power control, enabling CPUs to be taken off-line automatically when thermal overload conditions are detected.

By default, the system issues an interrupt if the temperature near the critical components exceeds normal operating range, and the fans are running at maximum speed. The default action taken by the operating system upon receipt of a temperature interrupt is to shut down the system with one minute of advance warning (system administrators may override this if necessary). This is sufficient to bring the system to an orderly halt without risking physical damage. In addition to active monitoring for trouble, software can also record an environmental history.

ECC Circuitry and Parity Checks

Large-scale systems such as Enterprise X500 manipulate so much data that electrical errors will inevitably occur, even with the extremely reliable components. Accordingly, Enterprise X500 systems are designed throughout to provide assurance that data is not lost or corrupted. The Gigaplane centerplane and the UPA interconnect are fully protected by ECC circuitry. Of course, main memory is also ECC protected. Even within the tightly controlled environments within chips such as the UltraSPARC processor, cache SRAMs, and the various bus interface ASICs, data paths are fully protected by parity checks. Address lines on the Gigaplane and the SBus peripheral bus are also protected by parity.

Hardware Monitoring

The Enterprise X500 implementation also features a large number of hardware monitors. Counters are located in strategic points in the system to monitor important events. They can be used to monitor the effectiveness and performance of many components which are normally unavailable, such as cache miss rate, centerplane statistics, and various internal statistics for the UltraSPARC, UPA, and SBus interfaces. In most competitive systems, information such as this must be obtained with external test equipment such as logic analyzers or custom monitoring boards.

The information provided by the monitoring hardware is sufficiently detailed to permit problem isolation and fine tuning of the operating system and applications.

JTAG Scan Coverage

All of the Enterprise X500's characteristic chips and boards include extensive support for JTAG scan testing. JTAG is a widely accepted methodology developed by IEEE to test electrical circuitry. Test logic is incorporated into the chips and boards during design and manufacture; this logic is later utilized by diagnostic processors. Although it is common in the industry to provide test coverage only on critical paths, all ASICs used in the Enterprise X500 provide 100 percent coverage. JTAG scan is used by the diagnostic routines at power-up, by the automatic reconfiguration software, and by Sun during the manufacturing process.

Unmatched Availability

Automatic System Recovery

Automatic System Recovery (ASR) improves the availability of the system. If a component failure causes a system to go down, as in the event of a CPU failure, the server will reboot without the failed component. This is particularly important when system uptime is a critical factor.

The Enterprise X500 family uses an extensive set of diagnostics to determine which units are functional. Units which are not functioning may have failed, or they may simply not be populated. This software is collectively known as the Power-On Self Test (POST), although it is run before each reboot and when

recovering from errors. POST runs in several phases. First, each processor checks itself, its cache, and its centerplane interface using the built-in JTAG scan support. This initial phase is executed from memory located on the Boot Bus, rather than from main memory as the centerplane is not yet enabled. Next all the processors in the system elect one of their number as a Master. Voting for Mastership takes place on the system scan bus. Next, the bus interfaces are tested. Finally, the centerplane is turned on and the system is tested as a whole, and inoperative units are noted. Finally, the Master configures the system, including assigning addresses to memory banks to maximize interleaving. Once the system has been configured, all processors have equal status; in particular, once UNIX is booted, no processor is designated as special in any way (i.e., the system is completely symmetrical).

The POST/ASR found in the Enterprise X000 and X500 families has been considerably enhanced over the similar programs found in the earlier sun4d systems. The new POST contains a configurable rule-based expert system that is used to work around component failures. Given the substantially richer configurations made possible by the new system design, the increased ASR capabilities are required. For example, the POST can be configured to use alternative boot disks to work around the failure of a disk controller or the I/O Board to which it is attached (obviously the contents of the boot disk would have to be mirrored in such cases). This capability can also be used by OEMs and other providers of sophisticated turn-key solutions to handle customized devices.

Because the POST is designed to handle many eventualities, combined with the increased flexibility made possible by numerous types of I/O Boards, the POST is delivered on each board using a flash EPROM. This makes field upgrade of the POST a trivial matter — new PROM code can be installed by a simple `tftp` process. When combined with hot plug capability, the Enterprise X500 design minimizes downtime due to failed components without incurring the much greater expense of fully redundant components.

Dynamic Reconfiguration

Dynamic Reconfiguration is the ability to change the configuration of a running system by bringing components on-line or taking them off-line during normal operation. With the latest Solaris 2.6 HW 5/98 release, I/O boards can be logically and physically removed or added to a live system — without halting the operating system or interrupting any of the user programs. This

simplifies administration and increases availability, as a system reboot is not required. Dynamic reconfiguration of the CPU/memory boards will be enabled with future Solaris releases.

Dynamic Reconfiguration includes both *dynamic detach* and *dynamic attach*. With dynamic detach, components are logically (rather than physically) removed from a configuration. This includes taking the components off-line and powering them down, thus making them ready for physical removal. With dynamic attach, components can be logically (rather than physically) added to a system and made available for use.

Alternate Pathing

Alternate Pathing manages I/O and network controllers, and is the foundation of Dynamic Reconfiguration. With Alternate Pathing, I/O operations can be redirected to an alternate path if the system board serving the primary path must be removed from the configuration. These actions take place dynamically, without disrupting network or storage connections.

For example, assume the active Ethernet controller is on the I/O board which is to be removed. The Ethernet card must be taken off-line, but this would cause the network connection to go down. Alternate pathing enables swapping the logical Ethernet connection to another stand-by Ethernet controller on a different I/O board, eliminating network disruption and increasing availability.

CPU Power Control

CPU power control is an availability feature which allows processors to be automatically taken off-line, powered down and later manually brought back on-line without system interruptions. If the temperature sensor on a particular CPU/memory board senses that the board is overheating and increasing the fan speed does not help, the CPU power control feature will automatically power off the associated processors, thereby removing the heat source. Without CPU power control, the whole system would have to be shut off to protect it from being damaged.

The CPU/memory board and memory remain on-line, however, even after the processors are shut off. Once the situation is resolved, the processors can be seamlessly reintegrated into the system.

CPU power control is used primarily to maintain availability in over-heated systems. But this same capability will be used when performing dynamic reconfiguration with the processors. Availability increases, as all actions occur without any interruption to running processes, and without requiring a system reboot.

Redundant Power and Cooling

Redundant power and cooling improves the availability of the system. The Enterprise X500 servers can be configured with N+1 redundant power supplies that, combined with current-sharing power circuitry, keep the system running should a single power supply fail. Dual variable-speed fans in each power/cooling module also increase availability by keeping the system cool in the event that one fan within a module fails.

Hot Swap Components

Most major components in the system are hot swappable and can be added or removed from the configuration at any time, even while the system is operational. This increases system availability by eliminating system down time, and improves serviceability by making maintenance and repairs more convenient for administrators.

The hot swappable components include power/cooling modules and peripheral power supplies, CPU/memory boards and I/O Boards, and the internal disk drives in the Enterprise 3500. Full support for live insertion and removal of these components eliminates one of the significant remaining motivations for system reboots.

Increased Serviceability and Manageability

Modular Design

The modular design of the Enterprise X500 servers dramatically improves serviceability. Boards are interchangeable among the different servers in the Enterprise X500 family, making it easier for customers to maintain smaller replacement parts inventories. Hot swap components enable administrators to perform maintenance when it is convenient, without affecting system availability. No special tools are required to service the system, making adding

and removing components quick and easy. In addition, system boards plug in both the front and rear of the system, enabling flexible and convenient configurations. For example, all I/O Boards can be grouped at the rear of the system for easier cable management.

Solstice SyMON and Enterprise SyMON

Enterprise X500 systems will initially be shipped with Solstice™ SyMON 1.6, a server monitoring tool that helps increase system availability and serviceability. Leveraging built-in environmental sensors and hardware monitoring counters, SyMON provides sophisticated diagnosis capability. In addition to monitoring the hardware, SyMON uses system logs, operating system counters, the hardware diagnostic log, and other information to provide extensive fault management and system health diagnosis. SyMON also provides a physical picture of the system, highlighting any components that need attention. This increases serviceability by allowing administrators to quickly locate problem areas.

SyMON can be used to pro-actively manage the system to prevent failures and system bottlenecks from occurring. The graphical interface shows how busy each system resource is, helping administrators without extensive backgrounds in computer architecture analyze their systems for bottlenecks, from either a physical or logical point-of-view.

Fault detection and analysis is carried to a much more elaborate degree than in previous systems. The GUI permits users to “drill down” on a diagram to locate faults in specific components — for example to an individual SIMM module — by clicking on icons which depict the actual components.

While Solstice SyMON 1.6 is a valuable system monitoring tool, there exist other critical functions that a robust system management solution must provide. These functions are mandatory for enterprise-class computing environments where scalability, ease-of-management, and interoperability requirements exist. These capabilities will be included in the next generation of SyMON — Enterprise SyMON 2.0.

When Enterprise SyMON 2.0 becomes available[1], it will be shipped with all Enterprise X500 systems, providing additional system management functionality. Key new features in Enterprise SyMON include Java-based GUI support enabling remote management from any computer platform, active management controls (dynamic reconfiguration and dynamic system domains), full SNMP connectivity, end-to-end security, historical data storage and management capabilities, integration with many system management platform vendors, integration with firmware and patch management, and application / database fault management.

Remote Control

The Remote Administration Control feature on Enterprise X500 servers allows an administrator to reboot, power-cycle, and even perform extensive operating system debugging remotely via a modem or other remote serial connection. This remote console functionality allows for true “lights out” management of a system, giving system administrators more flexibility in point of management and control. This also enhances the serviceability aspect of the systems by providing a method of retrieving the system state on “hung” systems. If Remote Administration Control is not desired for any reason, it can easily be disabled from the front panel keyswitch.

The remote control capability is handled via a special serial processor on the system clock board. This processor scans the system console for specialized system commands. The commands utilize relatively odd combinations of keystrokes and must arrive on the console port with specific timing (more than 0.5 seconds and less than 2.0 seconds between the keystrokes) in order to avoid accidental confusion with arbitrary line noise.

The console monitor is implemented in hardware which is not under the control of the operating system. This design allows the console monitor to circumvent a hard fault in the operating system.

In addition to remote administration commands, the PROM also has the ability to perform relatively extensive debugging of the operating system. The system state can be investigated, including a complete stack traceback and active thread lists, without requiring recourse to a kernel debugger.

1. Enterprise SyMON 2.0 is planned to be available in the third quarter of 1998.

References

Sun Microsystems Computer Company posts product information in the form of data sheets, specifications, and white papers on its Internet World Wide Web Home page at: *http://www.sun.com*.

Look for abstracts on these and other Sun technology white papers:

Solstice SyMON System Monitor, Technical White Paper, Sun Microsystems Computer Company, 1996.

Ultra Enterprise Server Performance Brief, April 1996, Sun Microsystems Computer Company, 1996.

Availability Features in the Sun Enterprise 3500 to 6500 Server Family, Technical White Paper, Sun Microsystems Computer Company, 1998.

Creator Graphics Technology, Technical White Paper, Sun Microsystems Computer Company, 1997.

Sun Microsystems, Incorporated.
2550 Garcia Avenue
Mountain View, CA 94043 USA
650 960-1300
FAX 415 969-9131
http://www.sun.com

Sales Offices

Argentina: +54-1-311-0700
Australia: +61-2-9844-5000
Austria: +43-1-60563-0
Belgium: +32-2-716-7911
Brazil: +55-11-524-8988
Canada: +905-477-6745
Chile: +56-2-638-6364
Colombia: +571-622-1717
Commonwealth of Independent States:
+7-502-935-8411
Czech/Slovak Republics:
+42-2-205-102-33
Denmark: +45-44-89-49-89
Estonia: +372-6-308-900
Finland: +358-0-525-561
France: +33-01-30-67-50-00
Germany: +49-89-46008-0
Greece: +30-1-680-6676
Hong Kong: +852-2802-4188
Hungary: +36-1-202-4415
Iceland: +354-563-3010
India: +91-80-559-9595
Ireland: +353-1-8055-666
Israel: +972-9-956-9250
Italy: +39-39-60551
Japan: +81-3-5717-5000
Korea: +822-3469-0114
Latin America/Caribbean:
+1-415-688-9464
Latvia: +371-755-11-33
Lithuania: +370-729-8468
Luxembourg: +352-491-1331
Malaysia: +603-264-9988
Mexico: +52-5-258-6100
Netherlands: +31-33-450-1234
New Zealand: +64-4-499-2344
Norway: +47-2218-5800
People's Republic of China:
Beijing: +86-10-6849-2828
Chengdu: +86-28-678-0121
Guangzhou: +86-20-8777-9913
Shanghai: +86-21-6247-4068
Poland: +48-22-658-4535
Portugal: +351-1-412-7710
Russia: +7-502-935-8411
Singapore: +65-224-3388
South Africa: +2711-805-4305
Spain: +34-1-596-9900
Sweden: +46-8-623-90-00
Switzerland: +41-1-825-7111
Taiwan: +886-2-514-0567
Thailand: +662-636-1555
Turkey: +90-212-236-3300
United Arab Emirates:
+971-4-366-333
United Kingdom: +44-1-276-20444
United States: +1-800-821-4643
Venezuela: +58-2-286-1044
Worldwide Headquarters:
+1-415-960-1300

Printed in USA