Sun Fire^{TM} 15K System

Sun Microsystems, Inc.

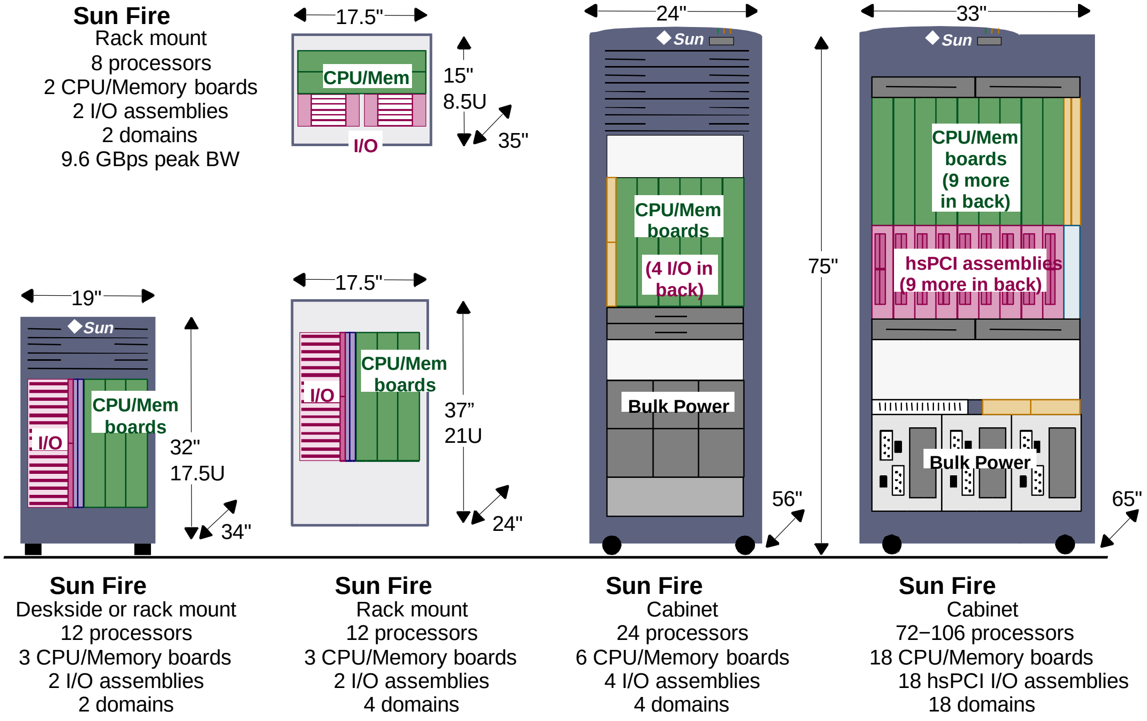
901 San Antonio Road
Palo Alto, CA 94303
U.S.A. 650−960−1300

18 SEP 2001 − REV 21−jlf

Copyrights

Table of Contents − "Sun Fire^{TM} 15K Just The Facts"

CHAPTER 1 Sun Fire 15K Server Positioning



Sun Fire
Rack mount
8 processors
2 CPU/Memory boards
2 I/O assemblies
2 domains
9.6 GBps peak BW

Sun Fire
Deskside or rack mount
12 processors
3 CPU/Memory boards
2 I/O assemblies
2 domains

Sun Fire
Rack mount
12 processors
3 CPU/Memory boards
2 I/O assemblies
4 domains

Sun Fire
Cabinet
24 processors
6 CPU/Memory boards
4 I/O assemblies
4 domains

Sun Fire
Cabinet
72−106 processors
18 CPU/Memory boards
18 hsPCI I/O assemblies
18 domains

Product Overview: Sun Fire 15K Server

The Sun Fire^{TM} 15K system is Sun's next−generation server that goes beyond the mainframe. The Sun Fire 15Kis based on the Sun Fireplane interconnect, the copper−based UltraSPARC^{TM} III Cu processors, scalable high−end symmetric multiprocessing (SMP), and the Solaris^{TM} Operating Environment. It is an ideal general purpose application and data server for host−based or client−server applications such as e−commerce, web and application hosting, online transaction processing (OLTP), decision support systems (DSS), data warehousing, communications services, or multimedia services. Additional expandability, scalability, and availability can be accomplished by clustering up to eight Sun Fire 15K systems together and employing Sun^{TM} Cluster 3.0.

The Sun Fire 15K system has true "big iron" attributes, include the most scalable and powerful server in Sun's extensive server product line. This SMP system supports up to 106 processors at 900 MHz each with 8 MB of secondary cache, over ½ TB of memory, 18 I/O hubs supporting 72 PCI slots, tested support for over 250 TB of online disk storage, and a wide range of UNIX(R) application software. The new plane interconnect implementation achieves up to 172.8 GB/sec peak (up to 43.2 GB/second sustained) with an overall I/O of up to 21.6 GB/second sustained.

In addition to unleashing unparalleled power, speed, bandwidth and capacity, Sun takes system availability to the next level. Sun Fire 15K is designed with many fault tolerant features, including automatic system controller failover, plane interconnect failure isolation, redundant communications to system controllers, error−correcting code (ECC) on datapaths, and dual power grids. The system is designed to be upgraded without disrupting users or halting the system.

The Sun Fire 15K system incorporates advanced features not found in other UNIX servers. Examples include 5G, Sun's fifth generation Dynamic System Domains, IP multi−pathing, second generation InterDomain Networking, and the ability to reboot up to all 18 domains at a time.

Dynamic System Domains partition the Sun Fire 15K into as many as 18 secure, fault−isolated domains, each running its own version of the Solaris Operating Environment.

Dynamic Reconfiguration balances resources by resizing domains on the fly. CPU/Memory and I/O can be independently and dynamically assigned and reassigned where they are most needed. For example, each of 18 CPU/memory boards and 18 I/O assemblies may be placed independently, thereby allowing the configuration of domains optimized for workloads that are either compute intensive, I/O intensive, or both. This capability allows CPU/memory boards and I/O assemblies to be serviced separately.

Target Markets and Users for the Sun Fire 15K System

The Sun Enterprise family of servers is targeted at strategic markets: manufacturing, finance, telecommunications, government, education, health care, retail, design automation, and oil and gas. The Internet adds a significant level of complexity to enterprise applications which now support all business operations, including those conducted with external business allies and customers. Applications must support business−to−business and business−to−consumer electronic commerce. These applications require a platform that provides scalability, availability, reliability, and security. The Sun Fire 15K, in combination with the Solaris Operating Environment, possesses these qualities and capabilities, positioning it as the platform of choice.

The Sun Fire 15K system is the highest performing SMP system on the market. It offers enhanced scalability and performance in a large−scale, centralized, enterprise server for parallel processing of commercial and technical applications. Commercial parallel and technical applications take advantage of the scalability of the Sun Fire 15K system along with its standard operating environment and commodity hardware components. Technical parallel applications rely heavily on the floating point performance of the Sun Fire 15K system. Commercial parallel applications include large−scale data warehousing, high−volume OLTP, server consolidation, and financial, analytical and e−commerce applications.

OLTP customers face high−volume issues associated with providing short response times and high availability for thousands of users. The Sun Fire 15K system addresses this by providing mainframe−like RAS capabilities and the ability to handle high transaction volumes and thousands of concurrent users with low response times.

Data warehousing customers appreciate the Sun Fire 15K system's ability to provide greater levels of delivered bandwidth where fast query performance is desired. SMP architecture is employed, making partitioning of data unnecessary. Additionally, the Sun Fire 15K system's large data volume, commodity RDBMS solutions, and mission−critical high−availability make it an even more attractive solution to their needs. No matter which industry or application type, customers can always find a wide range of available commercial off−the−shelf applications.

The Sun Fire 15K system supports a larger number of processors, memory, I/O and dynamic system domains than other systems in the Sun Fire Server product line. The Sun Fire 15K system is recommended for those customers who require 24 or more processors at the time of purchase, or within 18 months of the time of purchase. It is also recommended for customers that expect to have a large number of simultaneous I/Os, or plan to partition their system into more than 4 domains.

Technical Computing Customers

Technical computing customers seeking high performance compute servers are typically not divided by markets or applications, but by response time, room for growth, and cost.

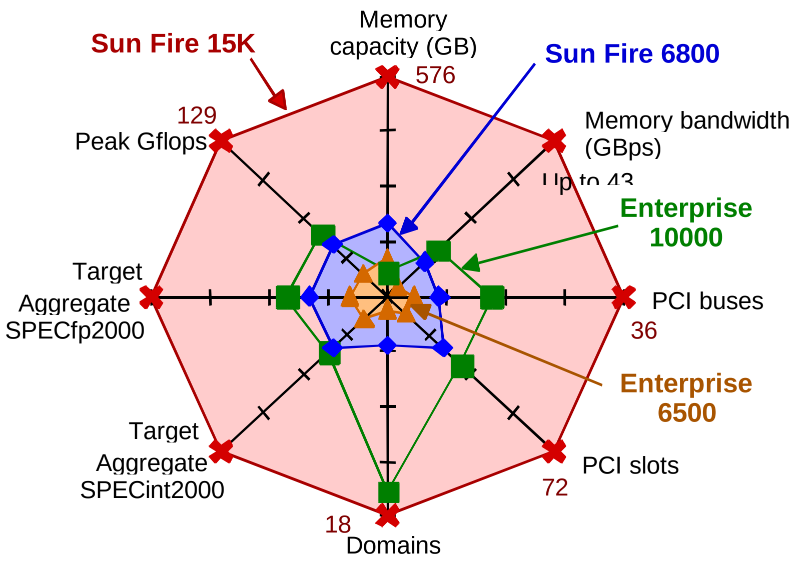
The Sun Fire 15K system has a peak performance of up to 106 CPUs x 900 MHz x 2 flops/clock = 190.8. Computationally intensive applications, where the Sun Fire 15K system is preferred, are those that are highly parallelized or those where large numbers of users are accessing particular applications. Common technical vertical markets include CAD, EDA, petroleum, and computational chemistry.

For technical customers, the Sun server family is positioned as listed below. Refer to the High Performance Computing 3.0 Just The Facts for further information.

- Sun Fire^{TM} 3800 server: A flexible workgroup server delivering exceptional performance at an affordable price.
- Sun Fire^{TM} 4800 server: A powerful mid−range server with exceptional availability.
- Sun Fire^{TM} 4810 server: A highly expandable mid−range server with a compact design.
- Sun Fire^{TM} 6800 server: Expandable, high−performance systems with mission−critical availability and integrated storage.
- Sun Enterprise^{TM} 10000 server: Sun's most powerful and highly available server for high−performance computing which incorporates additional RAS capabilities, like dynamic system domains, dramatically increasing system availability for compute−intensive applications.

Performance

The Sun Fire 15K architecture is designed to offer balanced system performance. These systems feature outstanding integer and floating−point performance, supporting up to 106 processors at 900 MHz each with 8 MB of secondary cache, 900 MHz UltraSPARC^{TM} CPUs with 8 MB of external cache each. The Fireplane interconnect, the coherent shared−memory protocol used by the UltraSPARC^{TM}−III/IV processor generation, supports a sustained plane interconnect throughput bandwidth of 43.2GB/second. High−speed networking is supported by 10/100/1000 MB Ethernet. Fast I/O capability is supported through upto thirty−six 33 MHz and upto thirty−six 66 MHz hot swappable PCI adapters including fast/wide SCSI, UltraSCSI, and fibre channel arbitrated loop (FC−AL) interfaces. SBus adapters are not supported on the Sun Fire family of servers.


Sun Fire 15K
Memory capacity (GB)
576
Sun Fire 6800
Memory bandwidth (GBps)
Up to 43
Enterprise 10000
129
Peak Gflops
Target Aggregate SPECfp2000
PCI buses
36
Enterprise 6500
Target Aggregate SPECint2000
PCI slots
72
18
Domains

Performance Type	Sun Fire 6800 Server	Sun Fire 15K System
Processor performance • SPECint_rate2000 • SPECfp_rate2000	● 96.1 Base 101 Peak ● 71.4 Base 77.1 Peak	18,844(Estimate) 30,065(Estimate)
TPC−H benchmark (1000 GB)	4735.7 QphH @ 1000GB $581 $/Qph @ 1000GB DB2 with UtralSPARC III 750MHz processors.	
Sustained system bus throughput	9.6 GB per second	43.2 GB per second
Pin−to−pin latency	approximately 200 ns	approximately 313 ns
Networking performance	Up to 1000 Mb per second	Up to 1000 Mb per second
I/O performance	up to 8 33 MHz PCI buses + 8 66 MHz PCI buses	up to 36 33 MHz PCI buses + 36 66 MHz PCI buses

Markets and Applications

The following chart illustrates how the Sun Fire 15K system fits into the current line of Sun server products.

Product	Positioning	Applications	Markets
Sun Fire 15K System	Enhanced scalability, availability, and performance in a large−scale, mission−critical, centralized, enterprise server for commercial and technical parallel processing applications.	E−commerce Data warehousing Data mining Business applications Customer management systems High−volume OLTP Engineering Design automation Analytics/commercial compute intensive Inter/Intranet Server consolidation	Manufacturing Finance Telecommunications Government Education Health care Retail Oil and gas Pharmaceuticals Chemical Internet commerce
Sun Fire 6800	High−end scalable and expandable Sun server, offering the performance and availability required for mainframe−class, mission−critical applications	Data warehousing Data mining Business applications Customer management systems OLTP NFStm software Design automation Analysis and simulation Video	Manufacturing Finance Telecommunications Government Education Health care Retail Oil and gas Pharmaceuticals Chemical Internet commerce

Sun Fire 4800	Excellent 12−way server for databases, business applications, high−performance compurting or Server consolidations.	High performance computing, Internet Data Centers	Telecommunications Government Education Retail Internet commerce
Sun Fire 4810	Affordable data center system designed to deliver high performance and high availability for enterprise−wide applications supporting thousands of users. Front access to all components makes this ideal for situations where rear or side access is not possible.	Internet Data Centers	Telecommunications Government Education Retail Internet commerce
Sun Fire 3800	Workgroup	•Internet applications Small application Server	Telecommunications Internet commerce

Specifications: Sun Fire 3800, 4800, 6800 and 15K Systems

	Sun Fire 3800	**Sun Fire 4800**	**Sun Fire 4810**	**Sun Fire 6800**	**Sun Fire 15K**
CPU/Memory boards	2	3		6	18
Maximum CPUs	8	12		24	72 w/o I/O tradeoff 106 with I/O tradeoff
Processor speed	750−900 MHz	750−900 MHz		750−900 MHz	900 MHz
Number of DIMMs	64	96		192	576
Memory capacity (with 1 GB DIMMs)	64 GB	96 GB		192 GB	576 GB
Centerplane	Active	Passive	Passive	Passive	Active
Repeater boards	0	2		4	NA
Expander boards	NA				18
Domains	2			4	18
I/O assemblies (assemblies)	2			4	18
PCI assembly types	hot−swap CompactPCI	PCI and hot−swap CompactPCI			hot−swap PCI
PCI slots/assembly	6	8 per PCI, 4 per cPCI			4
Max total PCI slot	12	16		32	72
Bulk power supplies	2	3		6	6
Power requirements	100−120 or 220−240 VAC		220−240 VAC		
System controller boards	2				

Redundant cooling	Yes				
Redundant AC input	No			Yes	
Enclosure	Rackmount	Deskside	Rackmount	Sun Fire 6800 cabinet	Sun Fire 15K cabinet
Room in rack for peripherals?	Yes				No

CHAPTER 2 Selling Highlights

Channels and Support

The Sun Fire™ 15K system uses the same selling channels as the rest of the Sun server line: direct and indirect worldwide. The principal support provider for warranty or for a SunSpectrum℠ support contract is Sun Enterprise Services. The Sun Fire 15K system warranty is one year for the hardware and software which includes 7 x 24 x 365 hardware on−site and telephone support (including holidays) with a 4 hour average hardware response time. Customers are also highly encouraged to obtain a SunSpectrum℠ Support Program service contract. This contract goes beyond the Sun Fire 15K warranty and provides for a flexible level service that allows customers to choose the right amount of service based on their specific needs. Sun™ Remote Services (SRS) monitoring is included as part of the Sun Fire 15K warranty and as part of a Gold or Platinum level service contract. Installation of the ServerStart℠ system is included in the purchase price.

Performance − Key Selling Factors

- Expandability

  The Sun Fire 15K server expands from entry−level configurations to system configurations that can handle terabytes of data and thousands of users. The Sun Fire 15K system is configured from 3800 to 15K UltraSPARC−III processors clocked at 900 MHz, 1 to 576 GB of main memory, and to over 250 TB of online disk storage. The Sun Fire 15K system is designed around an all new high speed interconnect with a bandwidth of 43.2 GB/sec and an I/O subsystem that incorporates 64−bit PCI technology clocked at 66 Mhz.

- Scalability

  The Sun Fire 15K system is highly modular. Customers can easily configure these systems to meet their application and performance requirements by simply adding UltraSPARC™ modules, memory, or I/O assemblies. The high−throughput Fireplane interconnect technologies and I/O architecture helps eliminate system bottlenecks and provides balanced system performance, even in systems with the maximum number of UltraSPARC™ modules and I/O devices. CPUs and I/Os scale independently.

- Security

  System controllers have been moved into the Sun Fire 15K. Comprehensive security is accomplished through improved integration between that Solaris 8 Operating Environment and the Sun Fire 15K's dual system controllers which can log all operations to a designated log host. All administration is designed to accommodate multi−tiered access, with a clear separation of responsibilities between platform and domain, operators and administrators. Dedicated communication paths on the fireplane between the system controller and every domain make the Sun Fire 15K more secure.

- Investment protection

  The CPU/memory board, UltraSPARC™ processor, dual inline memory modules (DIMMs), and industry standard PCI cards used in the Sun Fire 3800 through Sun Fire 6800 servers are also common to the Sun Fire 15K system. Therefore, when upgrading to the larger Sun Fire 15K system, customers can move these components from an existing chassis to the new chassis, provided that the CPU speed is 900 MHz or faster, protecting their investment. The Sun Fire 15K system uses the same peripherals in the same expansion cabinets as the rest of the family. Mixed speed UltraSPARC−III processors are allowed on a per board basis. The Sun Fire 15K server runs the Solaris Operating Environment, and protects the customer's software investment. Sun provides full support for existing 32−bit applications; in particular, Sun helps ensure that the 32−bit applications will run without recompilation and therefore will interoperate with 64−bit applications.

For example: One day the Sun Fire 15K is upgraded with 2 new 1 GHz processor boards. A 900 MHz board is pulled out of the Sun Fire 15K and put in a 6800. A board is pulled out of the 6800 and put into a 4800 with only 1 board already. In this case, 3 systems were upgraded with just a board purchase.

- Solaris^{TM} Operating Environment Applications

  The Sun Fire 15K requires the Solaris^{TM} 8 Operating Environment with a portfolio of over 12,000 available Commercial Off−the−Shelf (COTS) applications.

- Upgrade Program

  There is a trade−in program available to move customers to the Sun Fire 15K system from Sun's other servers, and from selected servers from Sun's competitors.

- Upgradability

  The modular design of the Sun Fire 15K system simplifies the task of upgrading to new technologies that will improve performance. The Sun Fire 15K system was designed to support future versions of the UltraSPARC^{TM} processor and standard PCI I/O interface technology.

- Reliability, Availability and Serviceability Features (RAS)

  Reliability, availability, and serviceability (RAS) are critical requirements of enterprise customers that deploy mission−critical applications. The reliability, availability and serviceability (RAS) goals for the Sun Fire 15K system are to protect the integrity of the customer's data while providing uninterrupted service to the end user.

  The focus is on three areas:

  - Problem detection and isolation− knowing what went wrong and ensuring the problem is not propagated
  - Tolerance and recovery− absorbing abnormal system behavior and fixing it, or dynamically circumventing it
  - Redundancy −−replicating critical components

To help ensure data integrity at the hardware level, all data is error correcting code (ECC) protected and control buses are protected by parity. These checks help ensure that errors are managed appropriately and data is never corrupted.

For tolerance to errors, resilience capabilities are designed into the Sun Fire 15K system to help ensure that the system continues to operate, even in a degraded mode. A symmetrical multiprocessing system, the Sun Fire 15K system can function with one or more processors disabled. In recovering from a problem, the system is checked quickly to determine the fault to allow minimum downtime.

The system can be configured with redundant hardware to reduce downtime. Major components of the Sun Fire 15K can be repaired while the system is online and in use.

Details of Sun Fire 15K RAS features follow.

• Manageability

- System Management Services, running on the system controller in combination with Sun Management Center, provides an advanced graphical interface for management, monitoring and control of multiple Sun Fire 15K servers and other servers in the Sun Fire family.
- SRS supports the customer's systems availability needs by providing the ability for Sun Enterprise Services to react quickly to problems. The service proactively detects error conditions based on sets of defined thresholds. If errors are detected, a corrective approach can be taken in partnership with the customer.
- Solaris WebStart is a browser−based tool used to install a single image of the Solaris Operating Environment and co−packages.
- Solaris Live Upgrade supports installation and reconfiguration of new versions of the operating system while the current system is still running.
- Solaris Resource Manager helps enable the consolidation of multiple applications onto a single Solaris server. It provides the ability to allocate and control major system resources, ensuring service availability for critical enterprise applications, IT−defined groups, and individual users.
- Solaris Bandwidth Manager can be used to manage the bandwidth used by IP traffic. It allows both incoming and outgoing network traffic to be prioritized by different classes of service.
- Automated Dynamic Reconfiguration (ADR) allows one to invoke scripts for tasks such as adding or deleting a system board to or from a domain, moving a board between domains, or for determining the status of a system board.
- RAS features dual system controllers. If one system controller fails, the other automatically takes over. The power supplies are designed to do dual grid power input. This means that the Sun Fire 15K can run even if one power grid fails. Also, the power supplies are N+1 redundant, so even after a power grid failure one of the six power supplies could fail and the system would be unaffected.
- Fans are installed as push−pull partners, so should one fan tray fail, its partner would provide all the cooling necessary to keep the system running. In addition, the fan tray fans are also N+1 redundant.
- All major parts in the Sun Fire 15K have a feature called FRUID (FRU = Field Replaceable Unit). FRUID is a small bit of storage that contains information about a part, like MAX temperature or failure imformation, serial numbers, etc. This allows Sun to diagnose problems much faster when a complex failure analysis must be performed.

CHAPTER 3 Reliability, Availability, Serviceability (RAS)

Reliability, Availability, and Serviceability assess and measure a system’s ability to operate continuously and to minimize service interruption. A system’s *reliability* reduces failures and insures customer data integrity. A system’s *serviceability* provides for short service cycles when component upgrades are necessary or failures occur. High reliability, to avoid failures, and quick serviceability, to recover rapidly from failures together lead to enhanced availability. A system’s Availability goes beyond reliability and serviceability though by also including the software that runs on the system. The availability of a system defines continuous accessibility to the functions and applications supported by the system.

Definitions

Reliability

The reliability of the system is characterized by the frequency of system outages. This frequency is typically measured by the Mean Time between System Interruptions (MTBSI). During a system interruption the system is unavailable to run customer applications.

When computers had one of everything, and everything needed to work for the computer to function the reliability of the system initially depended on the reliability of every individual component. Now that systems are constructed from parts that are redundant, an individual component may fail, however, the system will continue to operate. Sun’s focus is primarily on providing available systems, and clearly availability is impacted by component reliability, but major gains can be made in availability by pairing up individually reliable components. Clearly we need to build the computer out of reliable individual components.

Availability

Availability is the quality of being accessible and ready for use. The availability of a machine depends not only on the technology, but is also on the environment surrounding the machine. Obtaining optimal levels of availability starts at the core system design and extends to the overall data processing or application architecture.

An environment that promotes availability includes the people who are running the computing platform and the processes surrounding the computing environment. Without taking care of the people and processes, the features built into the Sun Fire family of computers will not lead to the availability a customer can expect. The product dimension of availability is impacted by both the system reliability and the reduction of downtime.

System Availability can be measured in minutes per year of unavailability. The “nine” is also often used. A system with “five nines” availability is expected to be available for 99.999% of the time, which means expected unavailability of 0.001% of the year or a little over five minutes per year.

Serviceability

Serviceability is characterized by the effectiveness of maintenance and repair of the system. The major servicability metric which attributes to the system availability is the Mean Down Time (MDT) which reflects the average time the system is down due to a system interruption. The Mean Down Time is influenced by the Mean Time to Repair (MTTR) of the individual Field Replaceable Units (FRUs), and the amount of time it takes to reboot the system following a repair.

Given that computer components do break from time to time despite all the efforts made to make them 100% reliable. No matter how much technology improves the reliability, and redundancy techniques improve availability — serviceability is a key component. The ease with which the computer is mended and made functional directly impacts the availability of the machine.

Since its inception, Sun has established a consistent trend in delivering increasingly modular and serviceable systems. Improving everything ranging from the reduction in slot dependencies, to tighter integration and greater environmental tolerances, Sun has made the time required for the replacement of a failed module much shorter. These enhancements, coupled with improved diagnostic capabilities, have significantly reduced the service cycle on systems, and simultaneously increased their reliability and availability.

Highly−available vs fault−tolerant

Fault−tolerant operation means that a system will continue to operate and provide service, despite any single fault. Fault tolerant machines are considerably more complex to design and produce than highly available machines.

The Sun Fire family is designed to be highly available — there will be unavailability during a diagnosis and reconfiguration following the failure of a limited range of hardware component. This can be contrasted with a fault tolerant machine, which should maintain service throughout a fault. Sun Fire will withdraw service during the fault and restart service automatically after deconfiguration of the faulty component from the machine.

SPARC processor error protection

The processor has ECC protection on its external cache SRAM, and parity protection on the major internal SRAM structures — as shown in FIGURE 3−1.

Data cache error protection

The on−chip data cache is a 64 KB, 4−way associative cache with 32−byte lines. The cache contents, physical tags, and snoop tags are separately parity protected. Errors in the data or in the physical tags are corrected in software by invalidating the cache, and retrying the load instruction. Errors in the snoop tags cause the line to be invalidated.

System interface error protection

The system address bus connection between the processor and the address repeater is protected by parity. The processor generates both parity and ECC for all outgoing data blocks. The parity is checked by the receiving dual processor data switch. The ECC is checked by all data switch units in the path of a transfer. ECC is checked and corrected by the processor when it receives a data block.

Address interconnect error protection

The address and response crossbars on the Sun Fire 15K plane interconnect have ECC protection for address transactions across the plane interconnect. The ECC corrects single−bit address errors on the fly, and detect double−bit errors. An address parity or uncorrectable ECC error stops execution in the affected domain.

Data interconnect error isolation

The ECC checks done by the Data Switches can identify the source of ECC errors in most cases. A particularly hard case for ECC errors is when a devices writes bad ECC into memory. These get detected much later by other devices reading these locations. Since the bad writer may have written bad ECC to many locations and these may be read by many devices, the errors appear to be in many memory locations, while the real culprit may be a single bad writer.
Since the Data Switch ASICs check ECC for all data entering or leaving each device from other devices, the original source of errors can be isolated. For example, a bad writer writing bad ECC to a memory on a different board will result in ECC errors being detected in two Data Switches: the first on the bad writer’s board, and the next on the write target’s board. The direction and transaction tag information can identify which CPU pair was the source of the error and which device is the target of a bad ECC write.

Console bus error protection

The console bus is a secondary bus to allow access by the System Controller to the inner working of the machine without having to rely on the integrity of the primary data and address busses. This allows the System Controller to operate even when there is a fault preventing main operation to continue. It is common to all domains and is protected by Parity.

Redundant Components

System availability is greatly enhanced by the ability to configure redundant components. All hot−replaceable (swappable) components in the system are, or can be, configured redundantly if the customer desires. Each system board is capable of independent operation. The Sun Fire family is built with multiple system boards and is inherently capable of operating with a subset of the configured boards functional. Redundant system components include:

- CPU/memory boards
- Expander boards
- hsPCI assemblies
- PCI cards
- Max CPU boards
- System controller boards
- System clock boards
- Bulk power supplies
- Fan Trays

Redundant CPU/Memory boards

A Sun Fire 15K system can be configured with up to 18 CPU/Memory boards. Each board contains up to four processors and their associated memory banks. Each CPU/Memory board is capable of independent operation. They can be hot−swapped out of running systems, and moved between system domains. The system is inherently capable of operating with a subset of the configured boards functional.

Redundant expander boards

A Sun Fire 15K system can be configured with up to 18 Expander boards. Expander boards may be hot swapped out and into the system only if the CPU/Memory board (Slot 0 board) and the bottom hsPCI assembly or MaxCPU board has been removed first.

Redundant hsPCI assemblies

A Sun Fire 15K system can configure up to 18 hsPCI assemblies. Each board supports up to four PCI cards. The hsPCI assemblies can be hot−swapped out of running systems, and moved between system domains.

Redundant PCI cards

Standard PCI card are mounted on the Sun Fire 15K PCI I/O board using a special cassette, which allows them to be safely hot swapped. Systems can be configured with multiple connections to the peripheral devices, enabling redundant controllers and channels. Software maintains the multiple paths and can switch to an alternate path on the failure of the primary.

Redundant MaxCPU boards

A Sun Fire 15K system can be configured with up to 17 MaxCPU boards. Each MaxCPU board contains two processors and their associated Ecache. Each MaxCPU board is capable of independent operation. They can be hot−swapped out of running systems, and moved between system domains.

Redundant system controllers

There are two system controller boards on the Sun Fire 15K. The system controller software, SMS which runs over Solaris^{TM} in each embedded processor keep check on each other and copy state information between them to allow automatic failover should the active system controller board fail. Two system controllers are standard with every Sun Fire 15K.

There is a main system controller and a hot−spare system controller The main system controller is responsible for providing all system−controller resources for the system.

If failures of the hardware or software occur on the main system controller, or failures on any hardware control path (e.g. console bus interface, ethernet interface) from the main system controller to other system devices, then upon detection of these failures, the system controller failover software will automatically trigger a failover to the spare system controller. The spare system controller will assume the role as the main system controller, and take over all the main system controller’s responsibilities. The system−controller data, configuration, and log files are replicated on the both system controllers.

Redundant system clocks

Sun Fire family systems have redundant system clocks. Should the clock system on one system controller board, the consumers of the clock lines will continue drawing their resources from the other system controller until such time as downtime can be arranged to replace the failed system controller board 0.

Each system controller board generates 75 MHz clocks which are separately distributed to the boards and the other system controller. Each board contains circuitry that selects which incoming clock will be buffered and fanned out to local clock loads. Each system board receives a 75 MHz clock, doubles the frequency to 150 MHz, and sends phase−aligned 75 MHz and 150 MHz clocks to local consumers. One master system controller board is selected to supply clocks. The phase detector on the slave system controller board compares its clock against the clock from the master board., and feeds back the error signal to speed up or slow down the slave system controller board.

Redundant power

The processor cabinet uses six dual−input 4 KW bulk power supplies. All six are standard feature. Two power cables go to each supply, so they can connect to separate power grids. These supplies convert the input power to 48 volt DC. All power in the system is commoned together. The power supplies are also N+1 redundant, such that the system can keep on running with a failed power supply. This means that if the Sun Fire 15K was in its maximum configuration that only 5 of the 12 power connections would be required to be active in order to provide all the power necessary to run the machine. The power supplies can be replaced while the system is in operation.

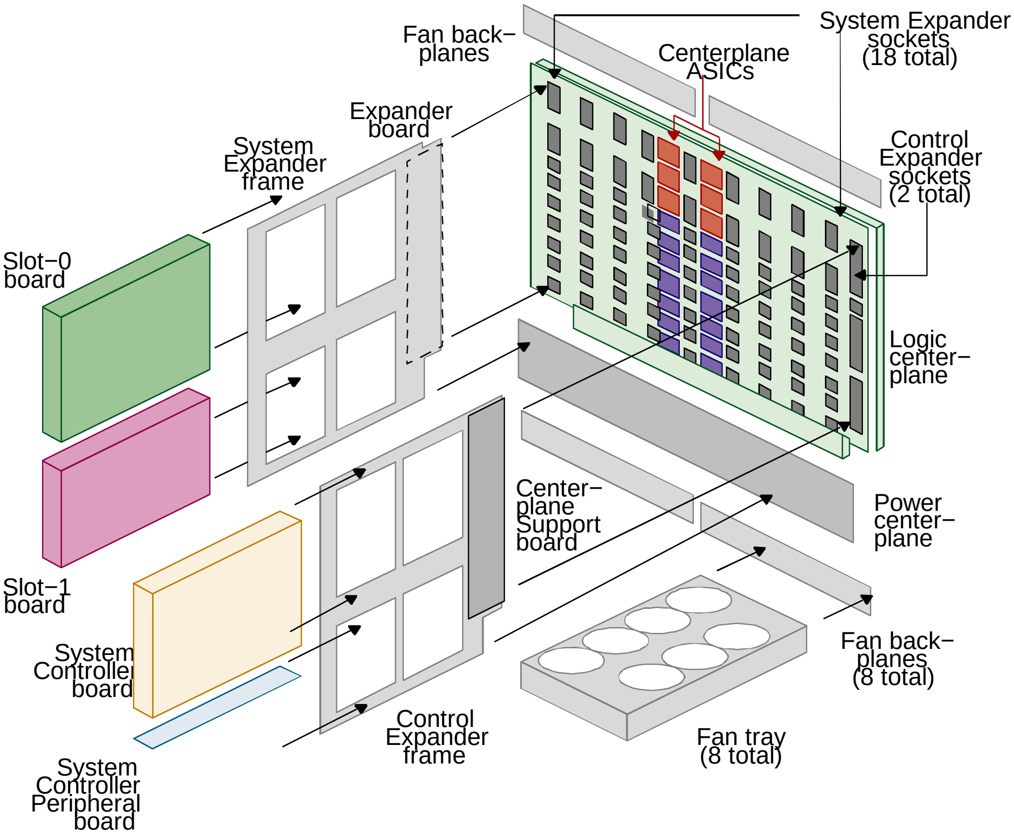
Power is distributed to the individual system boardsets through separate DC circuit breakers. The boardsets each have their own on−board voltage converters, which transform 48 VDC to the levels required by the on−board logic components. Failure of a DC−to−DC converter will affect only that particular system board.

Redundant fans

There are four fan trays above, and four below the system boards. The fans have two speeds, and normally run at low speed. If any of the sensed components get too hot, then the fans are switched to high speed. The fans in the fan tray are N+1 redundant, such that the system can run with a failed fan. The fan trays can be swapped while the system is running. In addition the fan trays are installed as "push−pull" partners with one on top, and one on the bottom. Should either partner fail, the other partner can provide the full cooling required by the machine from the partner trays.

Reconfigurable Logic Centerplane

The Sun Fire 15K has three independent crossbars implemented on the logic plane interconnect: one for addresses, one for responses, and one for data. The plane interconnect contains 20 ASICs, and is the only non−hot−swappable logic component in the system. Since a failed plane interconnect ASIC cannot be removed from a running system, each of the three plane interconnect crossbars (Address, Response, Data) can be independently be moved in and out of degraded mode on the fly. Degraded mode is separately configurable for each system domain. A failed ASIC will result in a all Domain reboot. When this occurs, the machine will reconfigure the centerplane to not use the half that was using the affected ASIC. This means that the machine will go into "double−pump" mode which effectively reduces the machines centerplane speed to ½ performance. Since the Centerplane has a very large amount of bandwidth for both sustained (43.2GB/sec) and peak loads (172.8 GB/sec) many workloads could be minimally impacted if the system comes back up in double pump mode.



Automatic System Recovery

A suitably configured system will always reboot after a failure. The System Controller locates the fault, reconfigures the system excluding the failed processor, memory, or interconnect component, and reboots the operating system.

The System Controller will only configure in parts which have the Fatal Error FRUID Bit clear. Field Replaceable Units (FRUs) that have been already detected as faulty by this or another machine will not be used.

System Controller

The heart of Sun's availability technology is the system controller. It is an Ultra Sparc stand−alone processor running System Management Services (SMS) and the Solaris operating system.

The system controller has access via JTAG to registers in each significant chip in the machine, and continuously monitors the state of the machine. Should there be a problem detected, it is the job of the System Controller to attempt to determine what hardware has misbehaved and then take steps to prevent that hardware from being used until it has been replaced.

The system controller performs the following main functions:

- Configures the system
- Set up the system and coordinate the boot process
- Set up the system partitions and domains
- Generate system clocks
- Monitor the environmental sensors throughout the system
- Handles errors: detection, diagnosis, and recovery
- Provide the platform console functionality and the domain consoles
- Provide routing via syslog of messages to a syslog host

The system controller (SC) in the Sun Fire 15K is a multifunction, Nordica−based printed circuit board (PCB) which provides critical services and resourcesrequired for the operation and control of the Sun Fire system. System boards within the platform can be logically grouped together into separately bootable systems called dynamic system domains, or simply domains. Up to 18 domains can exist simultaneously on a single platform.

The SMS software lets you control and monitor domains, as well as the platform itself. SMS software packages are installed on the SC. In addition, SMS communicates with every Expander board over an Ethernet connection.

SMS helps enables the platform administrator to perform the following tasks:

- Administrate domains by logically grouping domain configurable units (DCU) together.
- Dynamically reconfigure a domain
- Perform Automated Dynamic Reconfiguration (ADR)
- Monitor temperatures, currents, and voltage levels
- Monitor and control power to the components
- Execute diagnostic programs
- Warns of impending problems
- Notifies when a software error or failure has occurred.
- Monitors a dual SC configuration for single points of failure.
- Automatically reboots a domain after a system software failure
- Keeps logs of interactions between the SC environment and the domains.
- Provides support for the Sun Fire 15K system dual grid power.
- Configurable Administrative Privileges

You perform SMS operations by entering commands on the SC console after remotely logging in to the SC from another workstation on the local area network, or directly via a ASCII terminal session directly with the serial port on the SC. You must log in as a user with the appropriate platform or domain privileges if you want to perform SMS operations (such as monitoring and controlling the platform). Administration tasks on the Sun Fire 15K system are secured by group privilege requirements. Upon installation, SMS installs 39 UNIX groups to the /etc/group file.

Console Bus

The Console bus is a secondary bus which allows the System Controller to access the inner working of the machine without having to rely on the integrity of the system address and data busses. This allows the System Controller to operate even when there is a fault preventing system operation to continue. It is protected by Parity.

Environmental Monitoring

The System Controller is regularly monitoring the system environmental sensors in order to have enough advance warning of a potential condition so that the machine can be brought gracefully to a halt, thus avoiding physical damage to the system and possible corruption of data.

The environmental items monitored include:

- power state
- voltages
- fan speed
- temperatures
- device failure
- device presence

Temperature

The internal temperature of the system is monitored at key locations as a fail−safe mechanism. Based on temperature readings, the system can notify the administrator of a potential problem, begin an orderly shutdown, or power the system off immediately.

Power Subsystem

DC voltages are monitored at key points within the Sun Fire 15K. DC current from each power supply is monitored and reported to the System Controller.

Field Replaceable Unit Identification

The Field Replaceable Unit IDentification (FRUID) feature on the Sun Fire 15K is a new and powerful technique Sun has developed to help maximize customer availability.

Each component has an 8 KB SEEPROM chip or area on the component on which is stored 2KB of static, hardware write protected data recording such information as Manufacturing Part Number, Serial Number, Vendor Name, Ethernet Address and Bootbus Timing. The 6 KB of dynamic data holds such information as power−on hours, fatal bit error, hardware level, repair history, temperature log and error log.

The benefits to system availability are:

- The fatal error bit in each FRUID is asserted if the System Controller diagnoses that the FRU is misbehaving. This Fatal Error bit is be set until the part has been returned to the repair vendor for investigation or is reset in the field.
- Failure logged twice in succession can be logged, showing that the failure mode was not fixed the first time.
- Trend analysis on serial numbered FRU's allows identification of manufacturing defects.
- Trend analysis on power−on hours, temperature logs etc. enables identification of wearout phenomenon.
- Trend analysis on the vendor that supplied the part to SUN identifies any vendor−specific weaknesses.
- The enhanced configuration information allows selective patching of specific hardware issues by enabling the patching technology to identify the hardware components in the machine and provide only the necessary patches for that configuration.

Concurrent Serviceability

The most significant serviceability feature of the Sun Fire 15K is to replace system boards online as a concurrent service. Concurrent service is defined as the ability to service various parts of the machine without interfering with a running system. Failing components are identified in the failure logs in such a way that the field replaceable unit (FRU) is clearly identified. With the exception of the plane interconnect, all boards and power supplies in the system can be removed and replaced during system operation without scheduled down time. Replacing the System Controller that is currently active, or switching control to the redundant System Controller can also be done without causing a disruption in the main system operation.

The ability to repair these items without an occurrence of downtime is a significant contributor in achieving higher availability. A by−product of this online repairability of the Sun Fire family concerns upgrades to the on−site hardware. Customers may wish to have additional memory or an extra I/O controller. These operations can be accomplished online with users suffering only a brief (and minor) loss of performance while the system board affected is temporarily taken out of service.

Concurrent service is a function of the following hardware facilities:

- All plane interconnect connections are point−to−point making it possible to logically isolate system boards by dynamically reconfiguring the system.
- The Sun Fire family uses a distributed DC power system, that is, each system board has its own power supply. This type of power system enables each system board to be powered on/off individually.
- All ASICs that connect off−board plane interconnect have a loopback mode that enables the system board to be verified before it is dynamically reconfigured into the system.

Dynamic Reconfiguration of System Boards

NOTE:

> *DR is a post−release feature of the Sun Fire 15K. It is a new implementation of DR that builds on the "tried−and−true" DR functions first implemented on the Starfire. Rollout of these new functions is currently planned in stages, as Quality testing and Customer trials are completed over approximately the next 3−6 months. DR must be of unquestionable reliability for Data Center use. Although fully implemented in the software code, DR has been "turned off" in the current release of software, but can be easily activated for customer evaluations. These new functions are immediately available for Customer Beta testing. As we complete our field trials, the full release of each of these new functions will be separately communicated. Details of these new features are also*

available under non−disclosure. If you have Beta site candidates, please contact us for details.

The online removal and replacement of a system board is called dynamic reconfiguration (DR). Dynamic reconfiguration can be used to remove a troubled board from a running system. For example, the board can be configured in the system even though one of its processors failed. In order to replace the module without incurring down time, dynamic reconfiguration can isolate the board from the system, hot swap it out, and then allow replacement of the failing hardware parts. Therefore, the dynamic reconfiguration operation has three distinct steps:

- Dynamic detach
- Hot replace
- Dynamic attach

Dynamic reconfiguration enables a board that is not currently being used by the system to provide resources to the system. It can be used in conjunction with hot replace to upgrade a customer system without incurring any down time or to move resources from one domain to another domain on the fly. It can also be used to replace a defective module that was deconfigured by the system and subsequently hot removed and repaired or replaced.

Dynamic deconfiguration and reconfiguration is accomplished by the system administrator (or service provider) working through the system controller or from a domain.

- The first step is the logical detachment of the system board. The Solaris operating system's scheduler is informed, for the board in question, that no new processes should start. Meanwhile, any running processes and I/O operations are enabled to complete, and memory contents are rewritten into other memory banks.
- A switchover to alternate I/O paths then takes place so that when the I/O assembly is removed the system continues to have access to the data.
- The next step in the process is for the system administrator to manually remove the now inert system board from the system —a *hot replace* operation. The removal sequences are controlled by the system controller, and the system administrator follows instructions given by software.
- The removed system board is then repaired, exchanged, or upgraded.
- The second half of hot replace is employed to reinsert the new board into the system.
- Finally, the replaced system board is dynamically configured in by the operating system. The I/O can be switched back, the scheduler assigns new processes and the memory starts to fill.

So with a combination of dynamic reconfiguration and hot replace, the system family can be repaired (or upgraded) with minimal user inconvenience. Hot replace minimizes this interval to minutes by on−site exchange of system boards.

An interesting additional advantage of dynamic reconfiguration and hot replace is that online system upgrades can be performed. For instance, when a customer purchases an additional system board, it too can be added to the system without disturbing operation.

System controller removal and replacement

The hot−spare system controller board can be removed from a running system.

Remote Service

The optional capability exists for automatic reporting (to customer service headquarter sites) of unplanned reboots and error log information via email.

Every System Controller has remote access capability that enables remote login to the System Controller. Via this remote connection, all System Controller diagnostics are accessible. Diagnostics can be run remotely or locally on deconfigured system boards while the Solaris is running on the other system boards.

Reliability Features	**Availability Features**	**Serviceability Features**
• ECC−protected data path • ECC protected plane interconnect address path	• Highly configurable to increase redundancy (I/O, CPU, system boards, system controller, memory, etc.)	• Hot −swap components: • System boards, I/O, power supplies, fan assemblies, system controller boards
• Parity−protected address and control signals within a boardset	• Hot−plug components: CPU, I/O cards, power supplies, and fans	• Keyed connectors on hot−swappable components
• Highly reliable distributed power system	• Redundant power subsystem • Redundant power supplies	• No slot dependencies except for system controllers
• Environmental monitors and controls	• Redundant cooling subsystem	• Point−to−point plane interconnect connection enables component isolation, making DR possible
• Connectors, cables, and guides all designed for robustness	• Auto SC failover	• Failed components clearly identified and logged
• Point−to−point routers to maintain bus integrity over multi−drop buses.	• Cluster capability to meet high level availability requirements	• Hot−swap power/cooling modules
• ECC error logging	• Multiple operating system support using dynamic system domains	• SunVTS software

Reliability Features	Availability Features	Serviceability Features
• POST • BIST logic in all ASICS PCI Bus parity protected	• Compatible with commercial battery backup systems • ECC−protected data path • ECC protected plane interconnect address path • ECC−protected interconnect • Redundant system controllers can be configured for auto−failover high availability • Auto reboot: POST isolation of failed components prior to boot	• Several internal self−tests for error reporting • Dynamic reconfiguration for trouble isolation and online repair • Remote Monitoring available and included in purchase price • Intelligent system controller continuously monitors state of the machine
	• IP Multi−pathing increases network availability	• Console bus provides secondary path for system controller to access hardware health information
	• System controller monitored via Sun Management Center (SunMC) and can be done remotely	• Electronic serial numbers on all active boards
	Factory−configured hot spares: CPU/memory, I/O assembly, expander board	• SunMC simplifies configuration and management for the system and also aids in detecting, identifying, isolating, and correcting problems before they impact the system.
	• Redundant power cables/connections	• Sun Validation Test Suite to perform system−level diagnostics
	• Distributed dual power grid capability	
	• Dynamic Reconfiguration • Automatic System Recover • Dynamic Multi−Pathing	
	• CPU Power Control	

CHAPTER 4 Enabling Technology

Technology

Four principal areas of technology used in the design of the Sun Fire^{TM} 15K system give Sun a significant competitive advantage. They are:

- The UltraSPARC^{TM} Microprocessor Family

  The Sun Fire 15K server features the SPARC Version 9 compliant, 64−bit UltraSPARC^{TM} III Cu 900 processor with up to 8 MB of external cache and operating at 900 MHz. The UltraSPARC III processor provides very high integer and floating−point performance to address the needs of the most computationally demanding applications. With 64−bit data and addressing, the UltraSPARC III Cu processors have a number of important features to improve operating system and application performance:

  - Ultra−dense manufacturing process – 16 million transistor design (including cache) implemented using 0.25 micron, 6−layer metal CMOS technology operating at 1.8 volts. Package using a 1200−pin (800 signal) ceramic land gate array
  - 4−way associative on−chip 64 KB Data and 32 DB instruction cache, with up to 8 MB of external level−two cache through an integrated controller
  - Integrated DRAM controller with support for up to 8 GB of memory that can transfer data at up to 2.4 GB/sec
  - Interface to the new Fireplane system interconnect supports peak data rates of 2.4 GB/sec.
  - Six−way superscalar issue, no−stall 14−stage pipeline
  - Larger cache, improved branch prediction, lower cache latency, and higher clock rates combining to double the performance of the UltraSPARC III Cu.
  - High efficiency trap management
  - 16 K−entry branch prediction array
  - Enhanced error isolation and fault diagnosability
  - Enhanced VIS^{TM} instructions set with three new instructions for high performance on multimedia and networking applications
  - Binary compatibility across UltraSPARC processor line

- Enormous System Bandwidth

  The logic plane interconnect is the heart of the Sun Fire 15K system. It provides a sustained data bandwidth of 43.2 GB/s among 18 boardsets. The plane interconnect contains three 18 X 18 crossbar routers. System scalability and low latency are a function of having sufficient internal bandwidth between processors, memory and I/O and the crossbar technology, versus bus technology, was clearly proven in Sun's previous generation Enterprise 10000 server. Sun Fire 15K employs separate crossbar routers for address, response and data. Bandwidth scales up as system hardware is added.

- The Solaris^{TM} (10/01 or later) Operating Environment

  Without a stable and well−proven operating system, the best hardware in the world is useless. The Sun Fire 15K server includes the industry's leading enterprise Operating Environment, the Solaris^{TM} Operating Environment. Built on the latest UNIX technology, the Solaris^{TM} Operating Environment delivers unparalleled scalability and performance. The Solaris Operating Environment has been enhanced over the past few years to be able to address very large memories and to scale up to 106 processors, both important features for the Sun Fire 15K system. With enterprise integration by design, the Solaris Operating Environment provides easy access to a wide range of computing environments and network technologies. It delivers a competitive advantage to business through networked computing, scalability, and multi−architecture support. The Solaris Operating Environment provides an advanced, superior solution for all customer IT needs, both technical and business. The Solaris Operating Environment is an industrial−grade solution with the performance, quality, and robustness to deliver mission−critical reliability.

The Solaris Operating Environment delivers a unique advantage for mission−critical environments, providing advanced features and functionality that, combined with built−in networking, gives users a high−performance computing environment enabling faster, and more productive work.

The Solaris Operating Environment delivers the power of the Sun Enterprise servers to users through enhanced networking capabilities and performance, graphics and imaging, increased standards compliance, and key operating system management advancements.

The Sun Fire 15K system requires Solaris 8 Operating Environment.

- Custom ASICs

A number of custom integrated circuits were designed and fabricated for the Sun Fire family of servers. This enabling technology represents a huge engineering investment by Sun and promises improvements in reliability, performance, and overall cost. The following list represents some of the latest technology developed for the Sun Fire generation of compute platforms. For additional detail, please refer to the “The Sun Fire 15K System Architecture” documents.

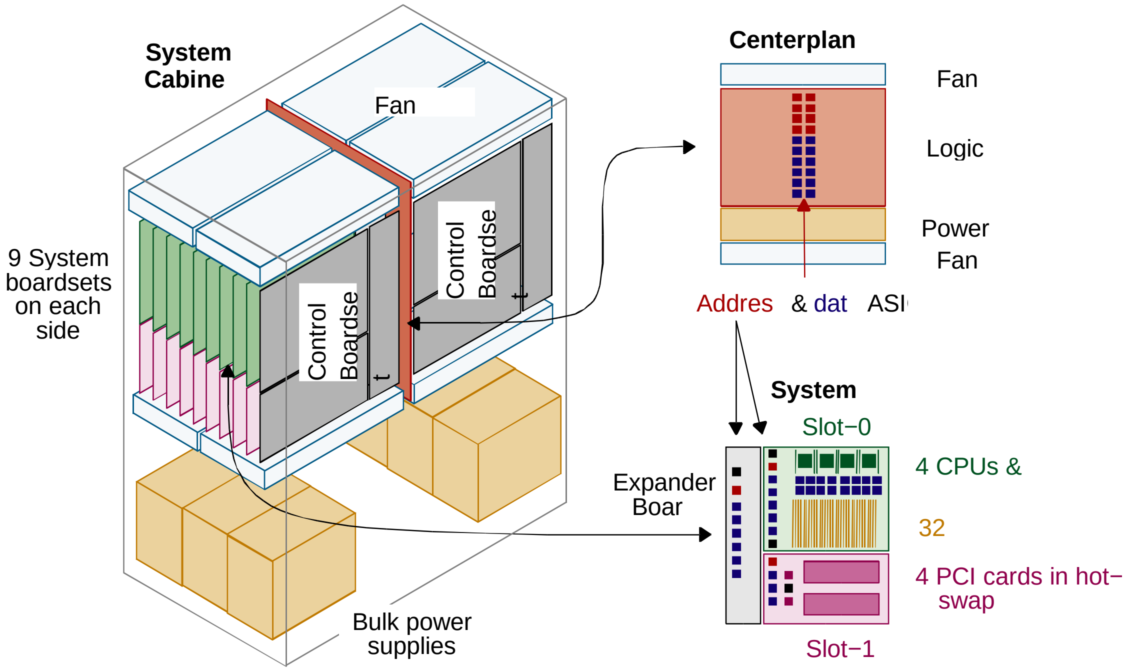
Slot−0 and Slot−1 Components
UltraSPARC^{TM} Processor
PCI Controller ASIC (Schizo)
Link Controller ASIC (WCI)
Address Repeater ASIC (AR)
Datapath Controller ASIC (SDC)
Dual−Processor Data Switch ASIC (DCDS)
Data Switch ASIC (DX)
Centerplane Components
Address Multiplexer ASIC (AMX)
Response Multiplexer ASIC (RMX)
Data Arbiter ASIC (DARB)
Data Multiplexer ASIC (DMX)
Expander Board Components
System Address Controller ASIC (AQX)
System Data Interface (SDI) ASIC
Slot−0, Slot−1, Expander Component
Boot Bus Controller ASIC (SBBC)

CHAPTER 5 System Architecture

Introduction

The Sun Fire™ 15K system is comprised of a plane interconnect, system board sets (CPU/memory board, PCI I/O board, plane interconnect expander), and Control Board sets (plane interconnect support board, Control Board, SPARC engine), peripherals, and power and cooling subsystems. These components and their relationships are illustrated in the figure below, and their functions are described in the following pages.

System Block Diagram



Logic Centerplane

With the rapid movement of processor technology and performance, bus technology has been hard−pressed to keep up. The all−new Fireplane system interconnect reverses this trend by providing superior memory and I/O bandwidth, ensuring balanced and predictable performance under the most demanding loads. The Fireplane system interconnect is the heart of the Sun Fire 15K. It provides a maximum data bandwidth of 43.2 GB per second between 18 board sets. The logic plane interconnect also delivers a console bus and an Ethernet to each board set.
The plane interconnect contains three 18x18 crossbars:

- Address crossbar
- Response crossbar
- Data crossbar

Address Crossbar

The 18x18 Address crossbar provides a path for address transactions between the System Address Controller (AQX) ASIC on each expander board. A pair of unidirectional paths connects to each expander board: one sending, one receiving. Each address transaction requires two clock cycles. The Address crossbar is one of the major changes over the Enterprise 10000 which uses a bus architecture.

Response Crossbar

The 18x18 Response crossbar provides a reply path between the System Address Controller (AQX) ASIC on each expander board. Each response message takes either one or two clock cycles, depending on the type. The response path is half the width of the address path. A pair of unidirectional paths connects to each expander board: one sending, one receiving. The response path is used for responses from home and slave agents to the original requester, and for completion messages from the requester to the home agent.

Data Crossbar

The 18x18 Data crossbar moves cache line (72−byte wide) packets between the System Data Interface ASICs (SDI) on each expander board. Each connection is a bi−directional 36−byte wide path. The bandwidth is (18 slots) x (32−byte path) x (150 MHz) divided by 2 for bi−directional paths, resulting in a maximum throughput of 43.2 GB per second. To maximize the use of these bi−directional paths, queuing occurs in the Data Multiplexer (DMX) ASICs. The data crossbar is implemented from 12 bit sliced Data Multiplexer (DMX) ASICs, which are under the control of two lockstep data arbiter (DARB) ASICs. If a failure occurs in the data crossbar, it can be put into double−pumped mode using half the datapath. It is possible to move into and out of double−pumped mode during system operation without rebooting.

Board sets (Expanders)

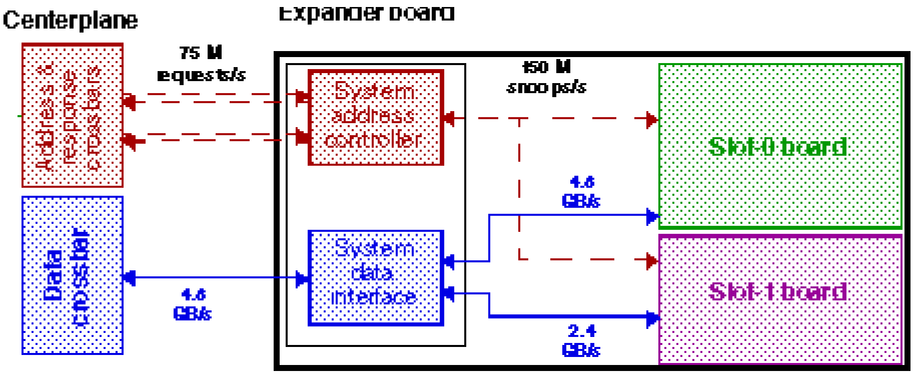
A board set is a combination of three boards that plugs into the plane interconnect.

There are two types of board sets:

1. System board set. These contains Slot 0 boards with processors and memory, and Slot 1 boards such as the hsPCI assembly or the MaxCPU board.
2. Control Board set is comprised of the Centerplane support board, the SC Peripheral board, the SC CPU board and the Control board.

System board set

A system board set is a combination of three boards, an Expander board, a slot−0 board, and a slot−1 board. The board set as a unit is hot−swappable from the plane interconnect, and the slot−0 and slot−1 board are individually hot−swappable from the Expander.



Logical View of System board set

Expander board

An expander board expands a plane interconnect slot to accommodate two boards: a slot−0 type and a slot−1 type. It provides a Level−2 Fireplane address bus that can execute 150 million snoops per second. The System Address Controller (AQX) on the expander board recognizes addresses targeted at other board sets, and transmits them across the plane interconnect.

The expander provides a three−port data switch to route data between the slot−0 board, the slot−1 board, and the plane interconnect. This dataport is 36−bytes wide to the plane interconnect and to the slot−0 board, and 18−bytes wide to the slot−1 board. A board set can transfer a maximum rate of 4.8 GB per second in each direction to other board sets.

It is possible to use an expander with only one system board (either slot−0 or slot−1). A system board may be hot plugged into the expander, tested, and configured into a running system, without disturbing the other board. The expander may be hot inserted or removed after its two system boards are removed. The expander also receives a console bus from each of the two system controllers.

Slot−0 (Top) boards

- **System board set: CPU/Memory board (Slot−0)**

The CPU/Memory board is the only slot−0 type System board and is common to the Sun Fire 3800, 4800, 4810 and 6800 servers. The CPU/Memory board accommodates four processors. Each processor has an associated memory subsystem of 2 banks of four DIMMs, so memory bandwidth and capacity are both scaled up as processors are added. The memory capacity of the board is 32 GB using 1 GB DIMMs. The maximum memory bandwidth on a board is 9.6 GB per second. The CPU/Memory board has a 4.8 GB (each direction) per second connection to the rest of the system.

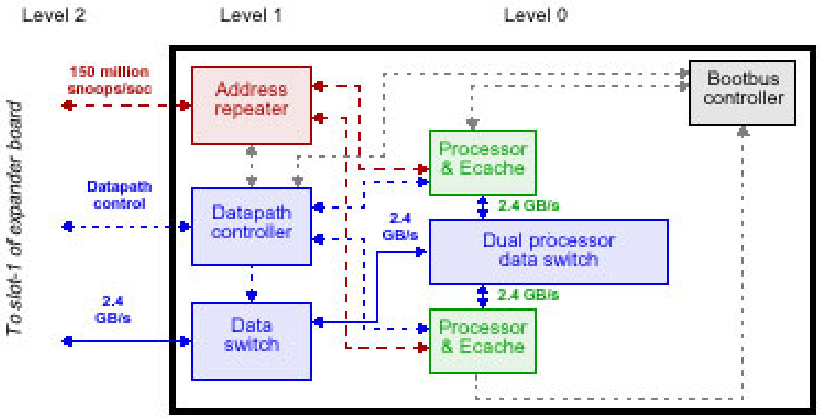
Slot−1 (Bottom) boards

- **System board set: hot swap PCI assembly (hsPCI) (Slot−1)**

The Sun Fire hsPCI assembly has two I/O controllers and is a slot−1 board. It is called the hsPCI assembly since it is hot−swappable. Each controller provides one 66 MHz PCI bus, and one 33 MHz PCI bus, for a total of two of each speed on the I/O board. Each controller has a total bandwidth of 1,200 MB per second. A Sun Fire I/O assembly has a 2.4 GB per second connection to the rest of the system. A cassette is used to provide hot−swap capabilities for industry−standard PCI cards. The cassette is a passive card carrier that adapts the standard PCI edge connector to a pin and socket connector that is suitable for hot−swapping. A PCI card is placed into a hot−swap cassette, and then the cassette is hot−swapped onto the PCI assembly. It relies on software along with the System Controller, for turning power on and off to each PCI slot.

- **System board set: MaxCPU board (Slot−1)**

The MaxCPU board is a slot−1 board. It accommodates two UltraSPARC−II processors, but does not accommodate main memory. This board allows processors to replace PCI cards in system configurations when more processing power is needed than I/O connectivity or when domains need processors that can be DR’ed in very quickly.



Logical View of the Dual−CPU board

Control Board Set

The System Control Board set is the heart of the Sun Fire family’s availability and serviceability technology. It provides the critical services and resources required for operation and control of the Sun Fire 15K System. It configures the system, coordinates the boot process, sets up the dynamic system domains, monitors the system environmental sensors, and handles error detection, diagnosis, and recovery. Two System Controller board sets can be configured into the system to provide redundancy and automatic failover in the event that one fails. The operating system used on the Sun Fire 15K is Solaris, loaded on a hard disk drive.

The Control Board set consists of three boards:

- Centerplane Support board: Plugs into a dedicated plane interconnect slot, and is the same size as an expander board. It provides power/clock/JTAG support for the plane interconnect
- Control Board: Plugs into the Centerplane Support board, and is the same size as a slot−0 board.

Control Board is implemented as a two−board combination:

- An off−the−shelf SPARC engine CP1500 6U high cPCI board with an UltraSPARC−IIi. This runs Solaris and System Management Services (SMS). The SPARC engine cPCI card mounts flat on top of the Control Board, similar to the way PCI cards are mounted onto hsPCI assemblies.
- Control Board, for the Sun Fire 15K−specific logic and connection to the Centerplane Support board.

System Controller: System Clock:

- Driven by 1 of 5 sources.
- Monitors clock input conditions and automatically switches if one is bad.
- Each board comes up running off its own system clocks by default.
- Each board in the system receives a clock from each System Controller .

System Controller: I2C Bus

- The System Controller can access every board, fan, and power supply in the system via an I2C bus.
- Exception is the power plane interconnect which has no I2C components.

System Controller: Console Bus

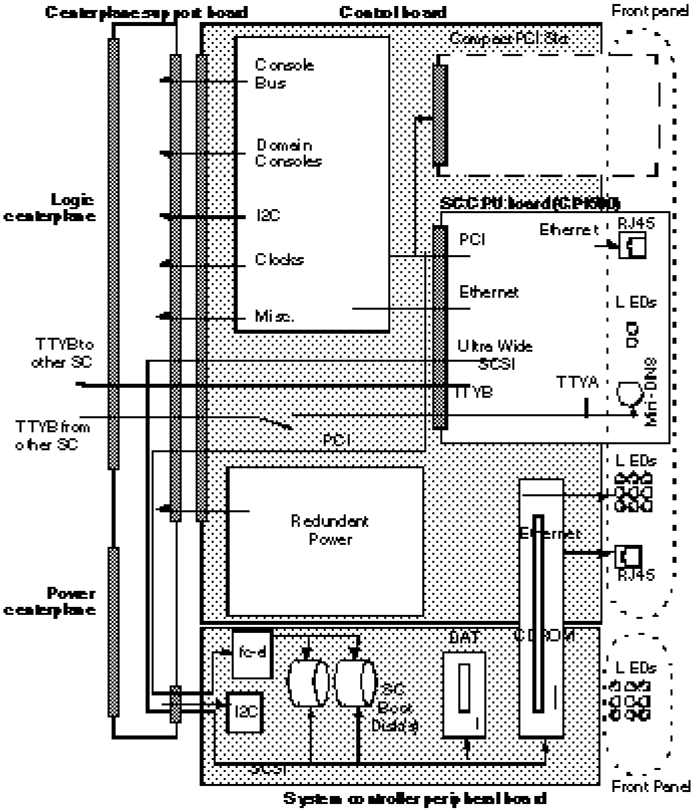
- The System Controller is the console bus master for the entire system, and either System Controller can be the master.
- Each System Controller uses the CP1500 board for its “brains”.
- The CP1500 board uses an UltraSPARC−IIi processor to run the Solaris operating environment and all associated applications needed for bring up, maintenance, and interrogation of the system.
- Two ethernet port on the CP1500 is used for the System Management Services (SMS) monitoring software and the other ethernet port is used for the System Controller−to−System Controller "heartbeat".
- One serial port is used for external access via the front panel, and the other serial port is used for back−door tip access.
- The SCSI port is used to support the boot disk, an optional hard disk drive (HDD), and a DVD drive.

System Controller: Ethernet

- The System Controller has a dedicated ethernet port for each I/O subsystem.
- Two “heartbeat” ports exist between the two System Controllers
- Two external ports are provided for the SMS terminal

System Controller: I/O board

- Holds the DVD drive and two HDD drives (one boot device and one optional)
- Provides power to the drives
- I2C will monitor the board’s power supply status
- Provides SCSI port connection from the System Controller to the drives



Logical View of the I/O board

Peripherals

The Sun Fire 3800 through Sun Fire 6800 have room in the same rack with the system enclosure for various peripherals. The Sun Fire 15K cabinet does not have room for peripherals.

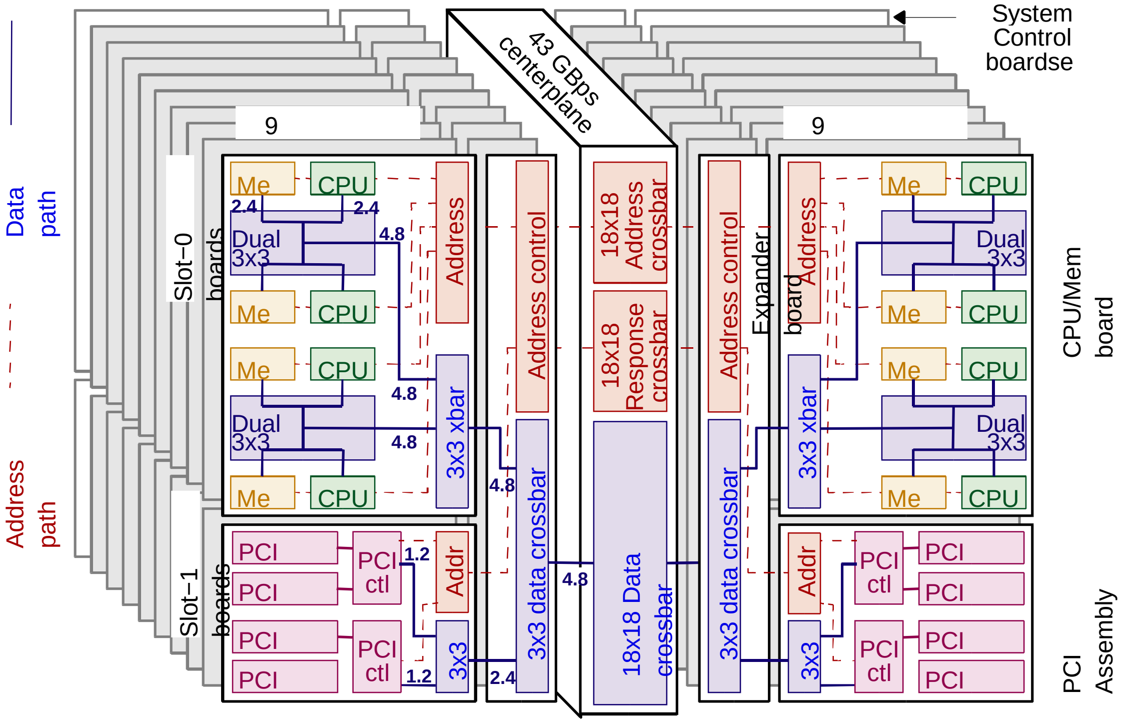
System Interconnect (Fireplane)

The Sun Fire 15K system uses the Fireplane interconnect which is the coherent shared−memory interconnect architecture used in the UltraSPARC−IIi/IV generation of systems. This is Sun Microsystems’ fourth generation of shared−memory interconnect, dating back to the early 1990s. Sun Microsystems uses an improved system interconnect with each new processor generation to keep system performance scaling with CPU performance.

Fireplane is an evolutionary improvement over the previous generation Ultra Port Architecture (UPA). The system clock rate is increased by 50% from 100 MHz to 150 MHz. The snoops per clock cycle is doubled from one half to one. Taken together, this triples the snooping bandwidth to 150 million addresses per second. The maximum data bandwidth for the Sun Fire 6800 and smaller systems is the snoop rate times the 64−byte coherency block, which is 9.6 GB per second, triple that of the previous generation Sun Enterprise 3500–6500 servers. The bandwidth for the Sun Fire 15K is (18 slots) x (32−byte path) x (150 MHz) divided by 2 for bi−directional paths, resulting in a maximum throughput of 43.2 GB per second, 3.4 times greater than that of the Enterprise 10000.

Fireplane also adds a new layer of point−to−point directory−coherency protocol, for use in systems that require more bandwidth than a single snoopy bus can provide coherency for. This facility allows coherency to be maintained between multiple snoopy buses, and is used in the Sun Fire 15K.

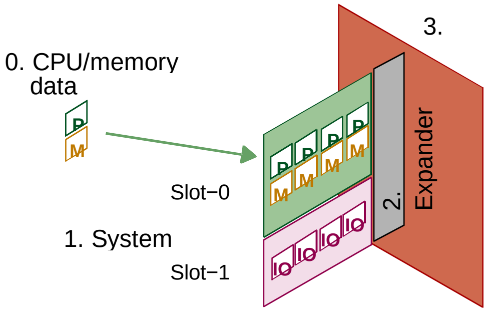
The small numbers in the block diagram below, represent peak bi−directional data bandwidths at each level of centerplane interconnect.



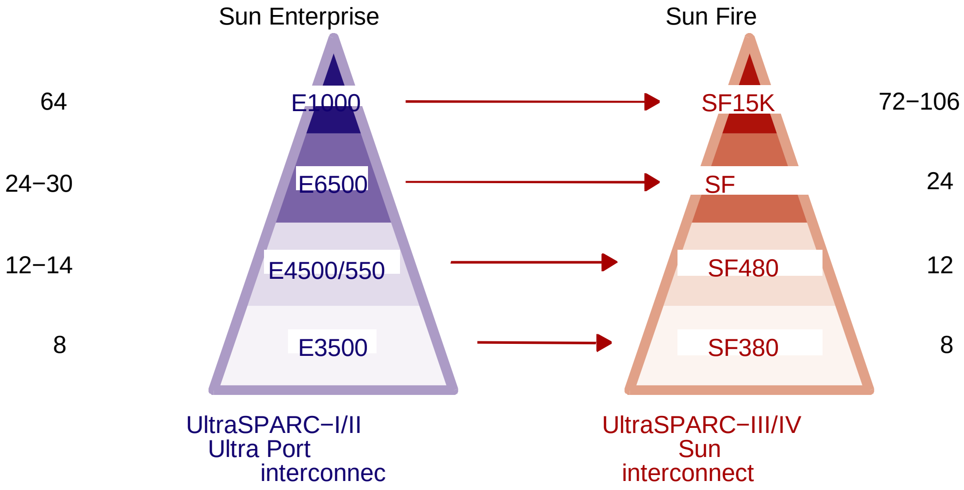
Centerplane Interconnect Levels

The Sun Fire 15K interconnect is implemented in several physical layers. The realities of physical packaging make it impractical to connect all the functional units (processors, memory units, I/O controllers) of a large server directly together. The system interconnect of a server is implemented as a hierarchy of levels: chips connect to boards, which plug into a plane interconnect. The latency is lower and the bandwidth is higher between components on the same board, because there are more connections between them than there is to off−board components.

Sun Fire™ 15K System



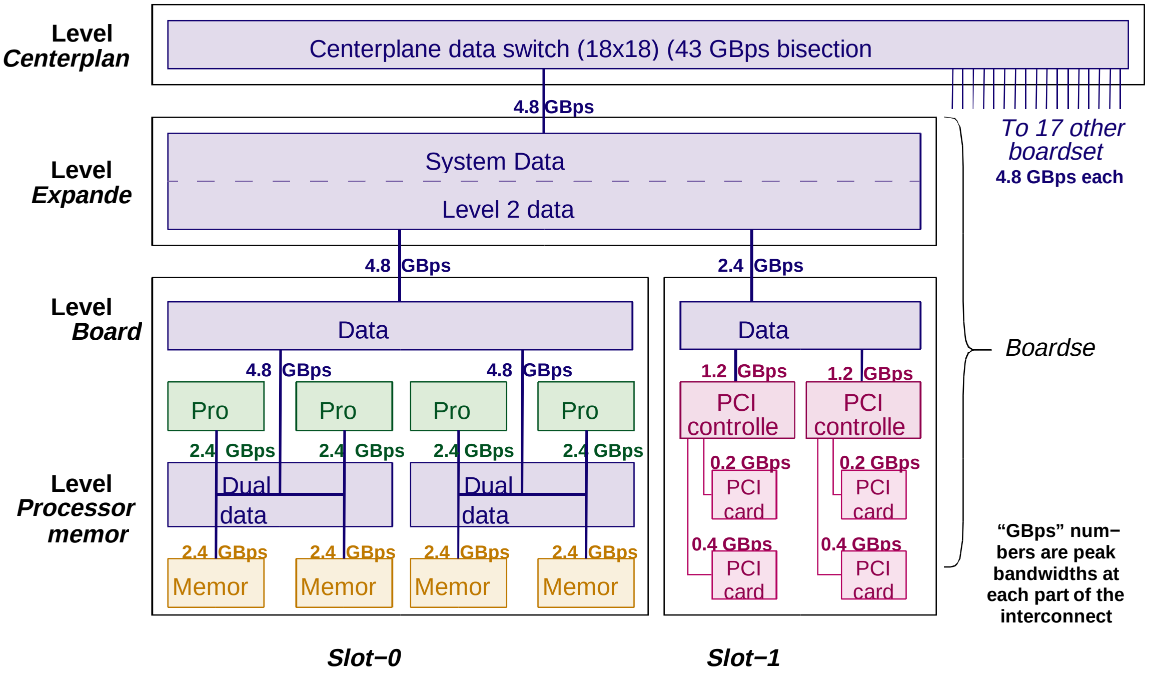
Sun Fire 15K Interconnect Architecture



Interconnect Performance

This section briefly quantifies the centerplane interconnect latency and bandwidth of the Sun Fire 15K.

Data Interconnect Levels



The numbers in the above figure refer to the peak bandwidth at each level. All datapaths are bi−directional and the bandwidth on each path is shared between traffic going into and out of a functional unit.

Bandwidth

Bandwidth is the rate at which a stream of data is delivered. These numbers are the peak memory bandwidths, as limited by the centerplane interconnect implementation. Memory is assumed to be interleaved 16−ways across the eight memory banks on one board.

All accesses to memory on:	**Sun Fire 15K**
Same CPU as requester	9.6 GB/s x number of board sets, 172 GB/s maximum for 18 board sets
Same board as requester	6.7 GB/s x number of board sets, 121 GB/s maximum for 18 board sets
Different board than requester	2.4 GB/s x number of board sets, 43 GB/s maximum for 18 board sets
Random data location	47 GB/s

Same−board Peak Bandwidth

The maximum same−board bandwidth is 9.6 GB per second per board. This occurs when: processors all access their own local memory, or all access the memory of the other processor in their pair, or two access their local memory, and two access memory on the other half of the board from themselves. The minimum same−board peak bandwidth is 4.8 GB per second per board. This occurs when all four processors access memory on the other half of the board from themselves. When memory is 16−way interleaved the peak bandwidth is 6.7 GB per second per board. In Sun Fire 3800 through Sun Fire 6800, the total same−board bandwidth is limited to 9.6 GB per second by the address bus snoop rate. Sun Fire 15K does not have this restriction.

Off−board Bandwidth

The off−board datapath is 32 bytes wide x 150 MHz = 4.8 GB per second. Since this bandwidth has to serve both outgoing requests from this board's CPUs, and incoming requests for memory from other CPUs, the per−board bisection bandwidth is halved to 2.4 GB per second. In Sun Fire 3800 through Sun Fire 6800, the total off−board bandwidth is limited to 9.6 GB per second by the address bus snoop rate.

Latency

Latency is the time for a single data item to be delivered from memory to a processor. There are several kinds of latency that can be calculated or measured. Two are presented below:

1. Pin−to−pin latency: calculated from the centerplane interconnect logic cycles. It is independent of what the processor does with the data.
2. Back−to−back load latency: measured by a kernel of the lmbench benchmark.

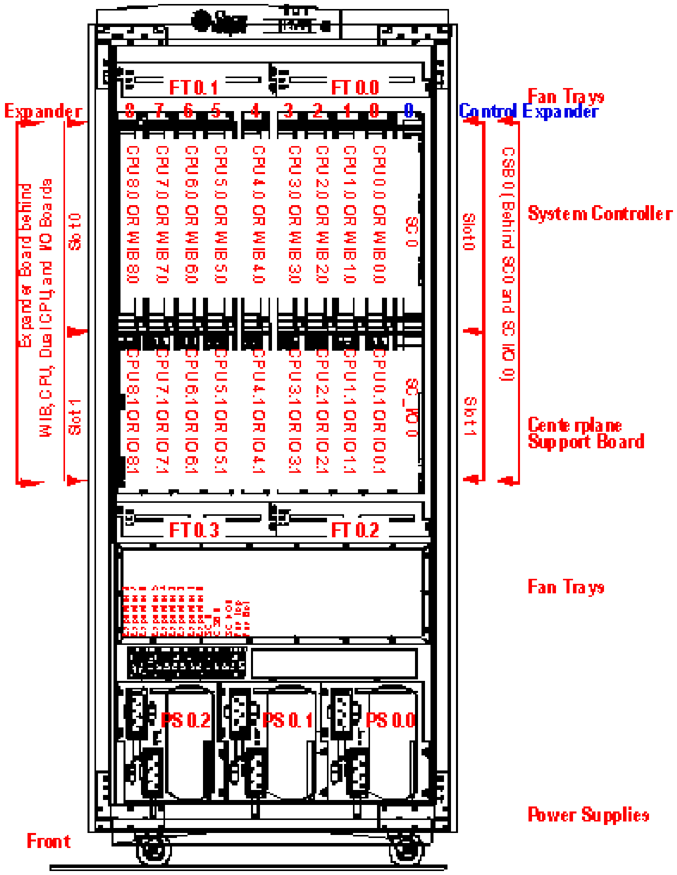
CHAPTER 6 System Components

Component	Function	Quantity
Control Board	**Set of two boards that are accessible via ethernet or serial interface:** **– SPARC™engine CP1500 6U cPCI board with an UltraSPARC™−IIi embedded system** **– Control board for Sun Fire 15K specific logic and connection to Centerplane Support Board**	**2**
SC Peripheral Board	**Accommodates a DVD, 2 disk drives, and a 4 millimeter DAT tape drive. Same form factor as a slot−1 board**	**2**
Memory	**Four banks of 8 SDRAM DIMMS per CPU/Memory board (1GB DIMMS)**	**Up to 576GB**
I/O	**Hot swappable PCI (hsPCI)** **2 PCI controllers per slot−1 type PCI Assembly** **4 PCI slots total per slot−1 type PCI Assembly** **− 2 at 33MHz** **− 2 at 66MHz**	**Up to 72**
Centerplane (Fireplane Bus)	**Fireplane Bus is the coherent shared−memory interconnect architecture which supports:** **− 18 X 18 Data Crossbar Router** **− 18 X 18 Address Crossbar Router** **− 18 X 18 Response Crossbar Router** **− 18 System Boardsets** **− 2 System Control Boardsets** **− Runs at 150 MHz and supports up to 72 processors**	**1**
System Cooling	**Four fan trays above System and Control Boards (7 fans/tray)** **Four fan trays below System and Control Boards (7 fans/tray)** **Base system configured to full fan/cooling capacity**	**8**

Component	Function	Quantity
AC power controller	**48V bulk power distribution and I^{2}C distribution**	**1**
Power supply	**6 dual−input AC−to−48 volt DC power supplies that can be split across two power girds. Base system configured with full up power**	**6**
Circuit breaker panel	**Interrupts power to various components within the system**	**1**
Processor Cabinet	**Houses plane interconnect, processors, memory, system controllers** **75” high** **33.25” wide** **65” deep**	**1**
Peripheral Cabinet	**Houses mass storage devices such as disk and tape drives** **75” high** **24” wide** **42.75” deep**	**1+**
Filters	**Cleans the air before it crosses into the fans.**	**6**

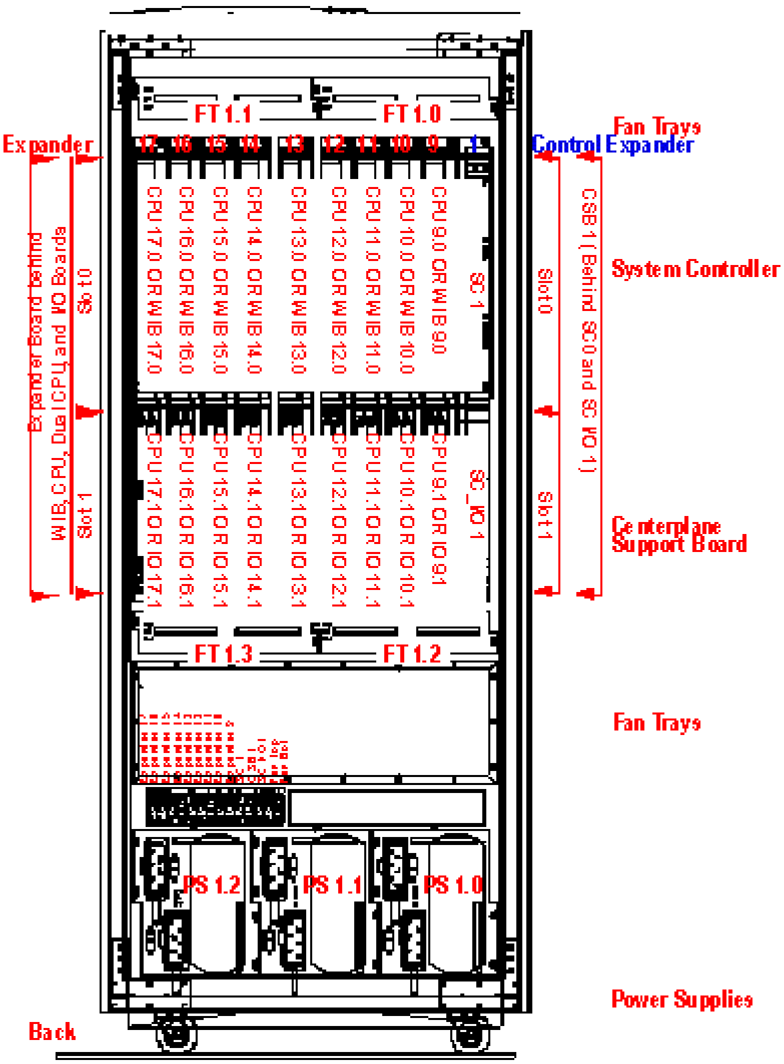
Component Numbering − FRONT

- Numbered right to left, top to bottom
- Fans and power supplies: 0.x front; 1.x back

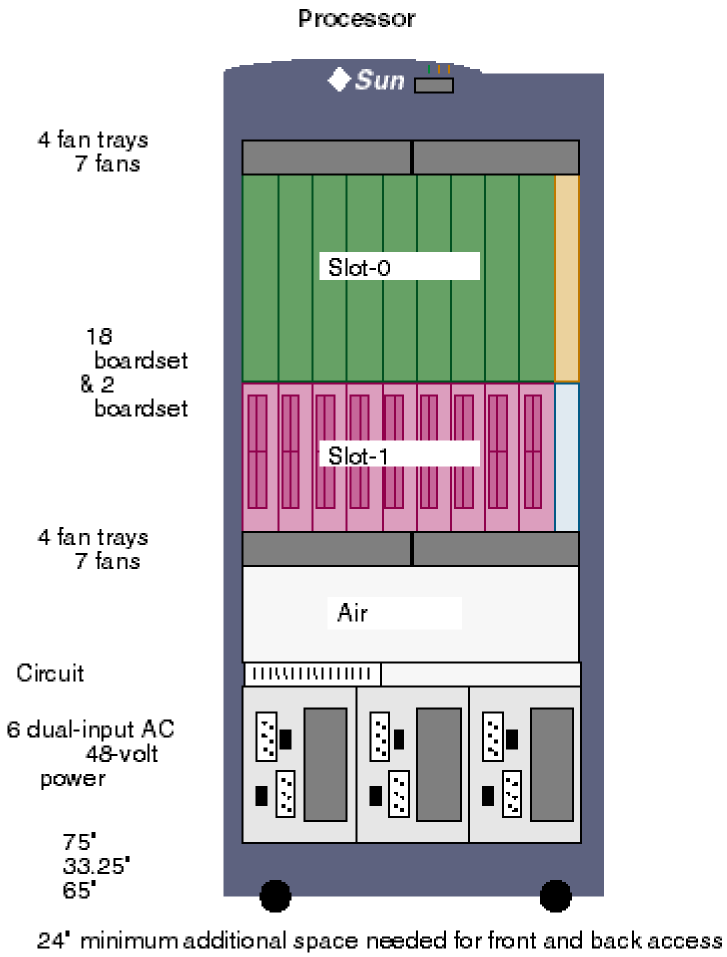


Component Numbering – BACK

- Numbered right to left, top to bottom
- Fans and power supplies: 0.x front; 1.x back


FT 1.1
FT 1.0
Fan Trays
Expander
17
16
15
14
13
12
11
10
9
1
Control Expander
Expander Board behind WIB, CPU, Dual CPU, and I/O Boards
Slot 0
Slot 1
CPU 17.0 OR WIB 17.0
CPU 16.0 OR WIB 16.0
CPU 15.0 OR WIB 15.0
CPU 14.0 OR WIB 14.0
CPU 13.0 OR WIB 13.0
CPU 12.0 OR WIB 12.0
CPU 11.0 OR WIB 11.0
CPU 10.0 OR WIB 10.0
CPU 9.0 OR WIB 9.0
SC 1
Slot 0
CSB 1 (Behind SC0 and SC I/O 1)
System Controller
CPU 17.1 OR IO 17.1
CPU 16.1 OR IO 16.1
CPU 15.1 OR IO 15.1
CPU 14.1 OR IO 14.1
CPU 13.1 OR IO 13.1
CPU 12.1 OR IO 12.1
CPU 11.1 OR IO 11.1
CPU 10.1 OR IO 10.1
CPU 9.1 OR IO 9.1
SC_I/O 1
Slot 1
Centerplane Support Board
FT 1.3
FT 1.2
Fan Trays
PS 1.2
PS 1.1
PS 1.0
Power Supplies
Back

Sun Fire 15K System Cabinet



- The Sun Fire 15K system enclosure is a 74.75−inch high, 33.25−inch wide, and 65−inch deep, data center cabinet that is symmetric except for the cap on the top front.
- The processor cabinet is configured with a full complement of eight fan trays, six bulk power supplies, and two system control boardsets which perform RAS services.
- A fully−loaded processor cabinet weighs 2,200 pounds (1000 kg)
- Inside the cabinet is an 18−slot card cage for system boards, control boards, and plane interconnect support boards.
- Directly above and below the card cage are the fan trays which draw air up through the cabinet and filters to exhaust out the top. Four trays containing 7 fans are located above the system boards and 4 trays containing 7 fans are located below the system boards.
- The cabinet will house 6 4KW power modules, each with dual AC inputs in order to support dual power grids. 12 power cords will be included to help ensure redundancy.

System Power

Site Power

The system runs from 200/240 VAC, single phase power, with a frequency of 47 to 63 Hz. The processor cabinet requires twelve 30−amp circuits, which are usually hooked up to two separate power grids. In North America and Japan, the site power receptacles are NEMA #L6−30R, otherwise they are IEC 309. The power cables that go between the system and the site's power receptacles are supplied with the system.

Processor Cabinet Power System

The processor cabinet uses six dual−input 4 KW bulk power supplies. Two power cables go to each supply. These supplies convert the input power to 48 volt DC, which is commoned together. The system can keep on running with a failed power supply. The power supplies can be replaced while the system is in operation. Power is distributed to the loads through individual DC circuit breakers. The boards each have their own on−board voltage converters, which transform 48 VDC to the levels required by the on−board logic components. Hot−swappable power and cooling components are hot−swappable.

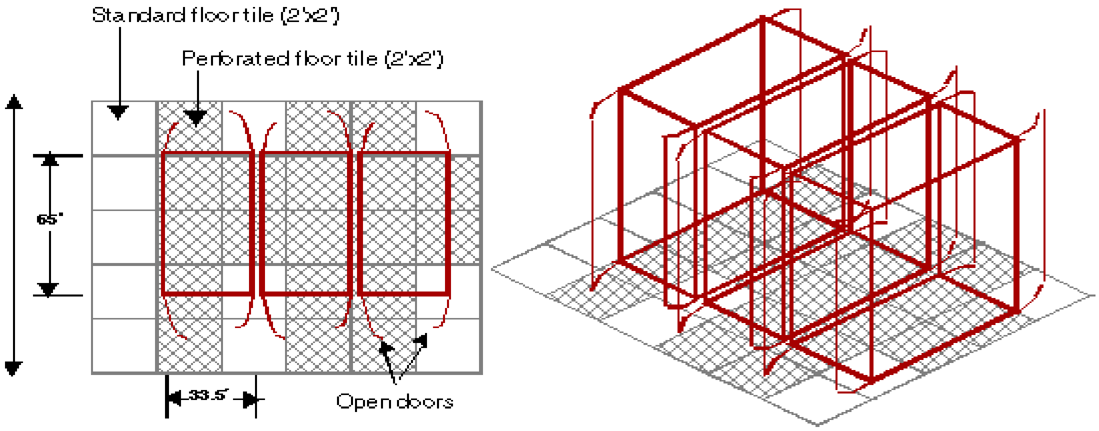
System Cooling

Environmental Requirements

The operating environment limits are − temperature: from 10° to 35° C (50° to 90° F); humidity: from 20% to 80%; and altitude: up to 3,048 m (10,000'). A fully−loaded system draws 24 KW of power, and has an air conditioning load of approximately 80,000 BTU/hour. Smaller configurations draw less power.

Raised Floor Requirements

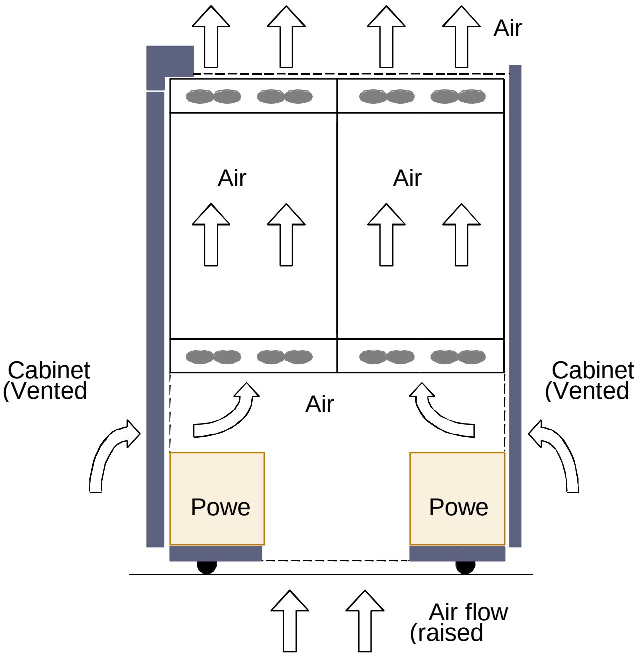
While Sun does not require the Sun Fire 15K to be installed on a raised floor, it does strongly urge its customers to do so. Raised floor enviroments are more likely to be less subject to air flow blockages, and are generally easier to work with to ensure that hot spots do not occur. When used on a raised floor, a fully−loaded processor cabinet needs seven 2'x2' perforated tiles under it (assuming that the tiles are capable of delivering 600 CFM cooling air at 0.07" of water). Rows of fully−loaded processor cabinets can be located adjacent to each other, with 5' of space per row, as shown below.



Processor Cabinet Cooling System

- Air flows from bottom to top: in through air inlets in the bottom, front, and back of the processor cabinet, and out through the top.
- There are four fan trays above and four fan trays below the boards.
- The two−speed fans run at medium speed normally. If any of the sensed components exceed temperature thresholds, the fans will switch to high−speed to provide additional cooling.
- The system can run with a failed fan or fan tray
- The fan trays can be swapped while the system is running.

System Cooling



Sun Fire^{TM} 15K System

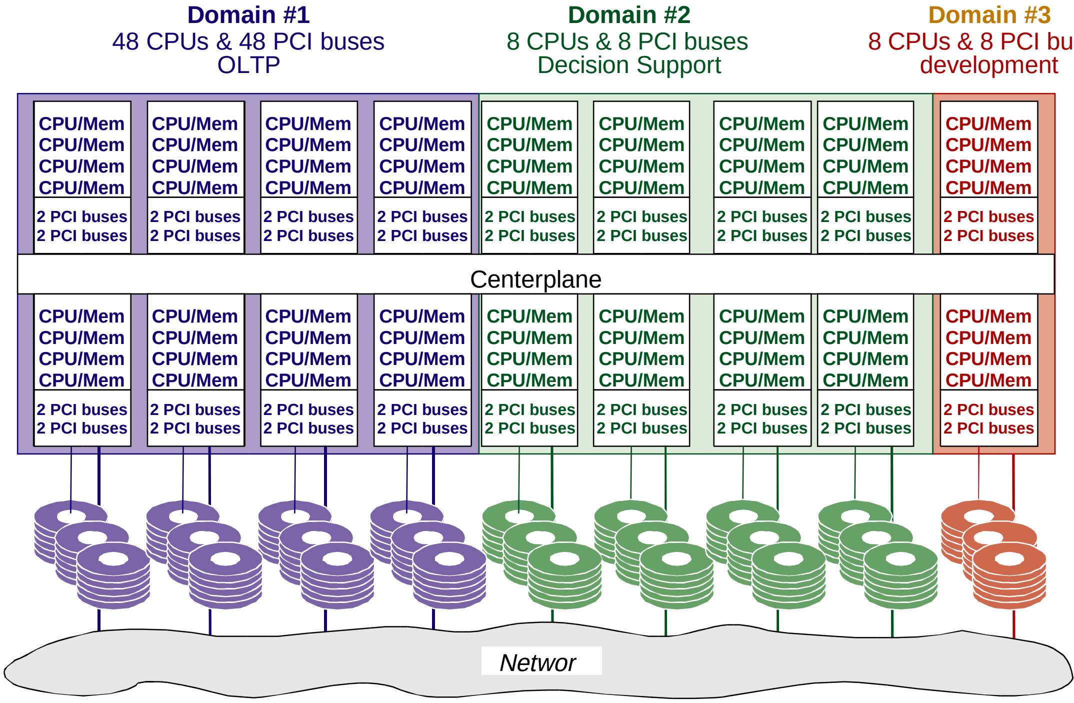
The Sun Fire 15K requires boot storage. Sun recommends either the Sun StorEdge S1, Sun StorEdge T3−A, Sun StorEdge T3−B or the Sun StorEdge 5200. These devices need to be installed into a data center rack. The Other storage devices can be used with the Sun Fire 15K, though the ones listed above are the only ones that can be used to boot it.

CHAPTER 7 System Management

Dynamic System Domains

This section describes the dynamic system domains feature of the Sun Fire 15K system interconnect. A Sun Fire 15K system may be dynamically subdivided up into as many as 18 domains. Each domain has a separate boot disk to execute a private instance of the Solaris™ Operating Environment, as well as separate disk storage, network interfaces, and I/O interfaces. CPU boards and hsPCI assemblies may be separately added and removed on the fly from running domains.

Domains are used for server consolidation and to run separate parts of a solution, such as application server, web server, and database server. The domains are hardware−protected from hardware or software faults in other domains. InterDomain Networking (IDN) provides high−bandwidth communication between domains via the plane interconnect, without compromising the isolation between domains. IDN support will be provided in a later release of SMS software.

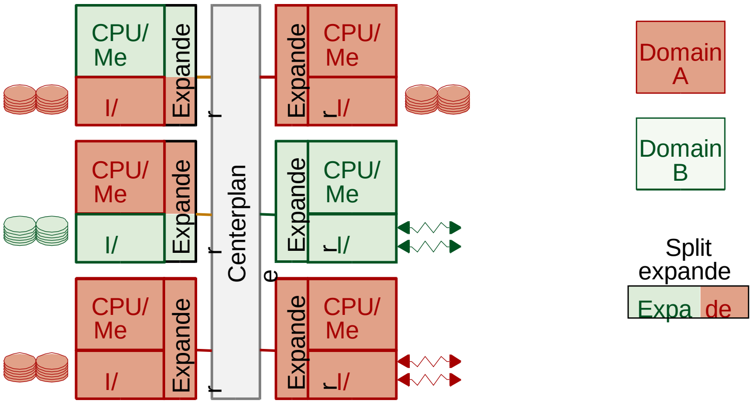


Domain Configurability

Each of the 36 system boards (18 slot−0 boards and 18 slot−1 boards) may be independently added to (or removed from) a running domain. This allows processor and memory resources to be moved from one domain to another, without disturbing the disk storage and network connections. Each domain must have an hsPCI assembly, so the maximum number is 18 domains.

When the two system boards in a boardset are in different domains, it is termed a split expander, since the expander board keeps the transactions separate for each system board. The figure below shows a sample configuration with some of the boardsets split between the two domains. No physical proximity is needed between boards in a domain.

Since split expander hardware is shared between two domains, its failure will bring down both domains. Also, memory accesses that go through a split expander take two system clocks (13 ns) longer. If all expanders were split, the load−use latency for accesses to other boardsets would go up about 6%.



Drawing 1Split Expander

Domain Protection

Primary domain protection is done in the system address controller (AXQ) ASICs, by checking each transaction for domain validity when it is first seen. The system data interface chips can also screen data transfer requests for valid destinations (to a granularity of the 36 system boards). In addition, each plane interconnect arbiter (data, address, response) screens requests to a granularity of the 18 expanders. This is a double−check on the other domain protection mechanisms which are in the system address controller and system data interface chips.

If a transgression error is detected in the system address controller, it treats the operation like a request to nonexistent memory. A transgression error in the plane interconnect will cause a domainstop of the transgressing domains, since this must indicate a failure of the primary protection mechanism.

Domain Fault Isolation

Domains are protected against software or hardware faults in other domains. Failures in hardware shared between domains will cause failures only in the domains that share it. Shared hardware includes plane interconnect ASICs, clocks, etc., and the expander board logic serving a CPU/memory board and an I/O board in two different domains.

The steering signals from the system address controllers on the expander boards to the address and response mux chips on the plane interconnect are parity protected. If there is a parity error, the multiple copies of the plane interconnect arbiter could get out of lockstep with each other. So, for a normal transaction, this will cause an immediate domainstop of the domain.

Domain Errors

Nonfatal errors, such as correctable single−bit errors in a datapath, cause a recordstop. History buffers in the ASICs are frozen so that information about the failure can be scanned out via JTAG while the domain continues to run.

Generally, a domainstop shuts down cleanly and quickly only the domain that has encountered the fatal error. When the hardware detects an unrecoverable error, domainstop operates by shutting down the paths in and out of the system address controller (AXQ) and system data interface ASICs. This shutdown is intended to prevent further corruption of data, and to facilitate debugging by not allowing the failure to be masked by continued operation.

The only case where multiple domains are stopped by the same error is when the error is in hardware shared by several domains.

- An error on an expander board shared by two domains will shut down both domains.
- An error in a plane interconnect ASIC, or plane interconnect wiring (connections between two plane interconnect ASICs) may shut down any or all domains.
- Connections between plane interconnect ASICs and expander ASICs are considered a part of that expander's wiring. An error here will shut down only the one or two domains that share the failed expander.

A parity error in the steering signals from an expander board to an arbiter. This error cannot cause the arbiters to so lose lockstep that another domain cannot correctly continue.

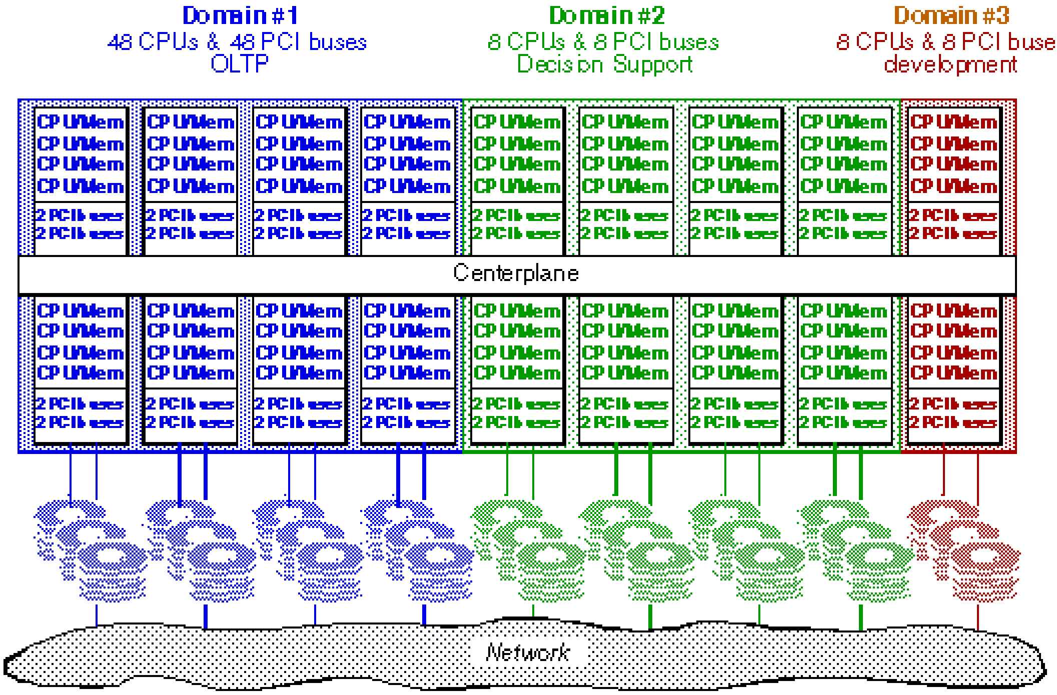
System Domain Implementations

- **LAN consolidation:** One Sun Fire 15K system can replace two or more smaller servers. It is easier to administer (uses a single system controller), more robust (more RAS features), and offers the flexibility to shift resources freely from one "server" to another. This is a benefit as applications grow, or when demand reaches peak levels requiring rapid deployment of additional computing resources.
- **Development, production, and test environments:** In a production environment, most sites require separate development and test facilities. Isolating facilities enables the development work to continue on a regular schedule without impacting production. With the Sun Fire 15K system, development and test functions can safely coexist on the same platform.

- **Software migration:** Dynamic system domains may be used as a means of migrating systems or application software to updated versions. This applies to the Solaris^{TM} Operating Environment, database applications, new administrative environments, or any type of application.
- **Special I/O or network functions:** A system domain may be established to deal with specific I/O devices or functions. For example, a high−end tape device could be attached to a dedicated system domain, which is alternately merged into other system domains which need to make use of the device for backup or other purposes.
- **Departmental systems:** A single Sun Fire 15K system may be shared by multiple projects or departments, simplifying cost justification and cost accounting requirements.
- **Configuring for special resource requirements or limitations:** Projects which have resource requirements that might overflow onto other applications may be isolated to their own system domain. For applications that cannot take advantage of all resources (that is, they lack scalability), multiple instances of the application may be run in separate system domains.
- **Data warehouse applications:** Many data warehouses use multiple systems to tier data. The Sun Fire 15K system can tier data on the same system and dynamically allocate more resources to individual tiers as needed.
- **Hardware repair and/or upgrade**

  The Sun Enterprise 10000 pioneered dynamic system domains (DSD) in the Unix server marketplace. In this system generation, domains have been extended to the entire Sun Fire family. Each domain is composed of one or more CPU/memory boards and one or more hsPCI assemblies. Each domain runs its own instance of the Solaris^{TM} Operating Environment, and has its own peripherals and network connections. Domains can be reconfigured without interrupting the operation of other domains. Domains can be used for:
  – Testing new applications
  – Operating system updates
  – Supporting different departments
  – Removing and reinstalling boards for repair or upgrade

Example of a Sun Fire 15K divided into three domains.



System Controller (SC)

The Sun Fire System Controller (SC) is an embedded computer that resides within each Sun Fire 15K server. The SC is responsible for many housekeeping and monitoring functions, playing a key role in keeping system availability high and in streamlining management tasks. The SC provides system console access to all domains within the server and the system as a whole. It manages hardware configuration during booting and during dynamic reconfiguration. It is also responsible for system error detection, status logging and reporting, environmental monitoring, identification services, and reporting.

Dual SCs are standard in all Sun Fire 15K systems, with one configured as a master and one as as slave controller. The slave controller monitors the "heartbeat" of the master to ensure its proper functioning. Should the master controller fail, the slave controller assumes the role of master, with all functions, including the system clock, moving seamlessly from the master controller to the slave. Like other boards in the Sun Fire 15K server, the SC boards are hot−swappable, allowing system controllers to be repaired and replaced without requiring any system downtime.

Communication with other system boards of the Sun Fire 15K system is handled by the console bus, a dedicated bus used for control and management operations. Communication external to the Sun Fire 15K server is available via a serial or Ethernet connection.

The SC is an integral part of the operation of the Sun Fire 15K server. SC functions include the following:

- **Virtual system clock**: The SC provides time of day to all domains within the system, in addition to providing the system clock signals to system boards.

- **Virtual system console:** The SC provides the system console interface to each domain. A connection to the console of a particular domain may be placed in "advise mode", allowing other read−only domain console connections to view all console activity.
- **Virtual key switch:** Domains do not have physical system key switches. Instead, the system console implements a "virtual key switch" for each domain. The virtual key switch allows the system administrator to secure the domain in the same way a physical key would be used. The virtual key switch supports the same four settings used on all Sun servers: off, normal operation, system diagnostics, and secure operations.
- **Power control:** The SC is used to control the 48V power supplies within the Sun Fire 15K system and to turn individual boards and field−replaceable units (FRUs) on and off.
- **Environmental monitoring and reporting:** Temperature sensors are placed adjacent to all critical components in the Sun Fire 15K system, with their measurements transmitted to the SC. The SC monitors these readings and takes appropriate action if an over−temperature condition occurs. Environmental conditions are available via SNMP and vital command−line queries.
- **Error management:** The SC acts on errors reported to it from system components via the console bus. Using information from components along the error path, the SC identifies the component(s) generating the error and marks the failed component for system removal at the next reboot.
- **Hardware configuration management:**
  - Automatic system recovery (ASR) management: In the ASR process, the SC checks each component for errors. If the part is faulty, either by failing a diagnostic test or by a notation of a fault in its EEPROM, the part is removed from the system during boot time.
  - Dynamic reconfiguration (DR) and domain creation responsibilities: The SC is responsible for the configuration within the Fireplane system interconnect to identify which boards will make up each domain. Identification can be done when a domain is created or when resources are added to or deleted from a domain. In either case, the system interconnect must be told which boards make up each domain so that it can isolate its activity to one domain at a time.
- **POST Management:** Like previous Sun servers, the Sun Fire 15K servers run a series of diagnostic tests when they are powered on. These power−on self−tests (POSTs) are more extensive and complete than in previous generations, since testing of domains and boards is scheduled in parallel whenever possible. POST management features and activity include the following:
  - Exercises the Sun Fire 15K system logic below the FRU level; with a high degree of accuracy, finds failing components and enables isolation to the FRU
  - Provides a highly−available platform for customer applications, even in the face of hardware failures
  - Provides low−level, start of day configuration services, including detailed interaction with specific hardware components
  - Records sufficient information about failed and marginal components so that both field replacement and subsequent factory repairs are expedited
  - Remembers which components passed the tests, and will configure only those components into the final system configuration. This is possible by using the JTAG access to each of the key Sun Fire 15K system ASICs.
  - Can be directed to ignore certain components by looking them up in the blacklist. In this way, components scheduled for service, of questionable functionality, or at a specified revision level can be kept out of a system configuration.
  - Has the responsibility for establishing the final system hardware configuration. If there

is a failed or blacklisted component, there are usually a variety of ways in which the final system may be configured.

- **Sun Management Center (SunMC) software:** In combination with the optional software, SunMC, the networked SC gives administrators a powerful tool for system and resource management. The SunMC software provides features such as domain management, a GUI interface to dynamic reconfiguration, alternate pathing, and SC commands. SunMC also provides hardware information, environmental monitoring, and propagation of alarms to associated devices.
- **Sun Fire 15K system Capacity on Demand software**: The capacity on Demand sofware for the Sun Fire 15K is being developed to run on the SC.
- **Remote service:** All functions on the SC that can be performed remotely are set up for terminal execution via dial−up connections.
- **SunVTS^{TM} software:** SunVTS^{TM} software (Sun Validation Test Suite) is the replacement product for SunDiag^{TM} software. Like SunDiag software, SunVTS software is run at the UNIX® level and is designed to exercise the entire system. It supports either a graphical or TTY user interface and provides error and information logging.

  The key features of SunVTS software are:

  - UNIX−level diagnostics **−** System tests execute real UNIX code under the Solaris^{TM} Operating Environment.
  - Automatic system probing − The system configuration is displayed through the user interface.
  - Two user interfaces − A graphical−based interface and a character−based interface are both available. The SunVTS kernel is cleanly separable from the user interface, such that multiple−user interfaces can communicate with the same SunVTS kernel. The character−based interface permits the writing of shell scripts to control SunVTS software.
  - Application Programming Interface (API) **−** The API provides a defined interface into the SunVTS kernel from other processes, as well as the user interfaces. A SunVTS execution could be initiated in a cron−like fashion, with no direct user interface at all.
  - Advanced configuration and execution control − Tests can be grouped together based on user requirements, with fine−grained execution control for status and logging information.
- **Network Console (netcon):** The SC provides a service called Network Console (netcon). Netcon provides a “console” for single−user operations. Normally, a SC must be on the same subnetwork as the Sun Fire 15K system; with netcon, they can be anywhere in the world that has a network attachment. The service is provided such that sessions, similar to rlogin sessions, can be provided to X−Windows clients on the same network as the SC. This helps enable system administrators to access the SC from any location on the same network as the SC.

  Netcon:

  - Helps enable access to the UNIX console, as with the Sun Enterprise 10000 −− shell prompt, ASCII session
  - Requires domain administrator authority
  - Private internal Ethernet when the Solaris Operating System is running
  - IOSRAM "tunnel" protocol when in OpenBoot PROM (OBP)
  - Like the Sun Enterprise 10000, has the ability to log all traffic to a file

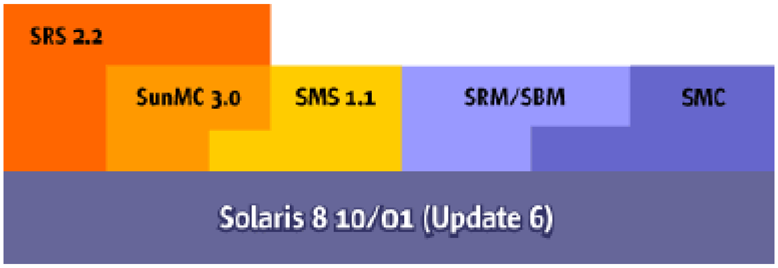
CHAPTER 8 Software

Sun Fire™ 15K Server Software

Introduction

The software market is constantly evolving, creating new products and services to meet the explosive growth and changing demands of information technology. To manage Sun's products in an enterprise environment, Sun provides software solutions that improve network efficiency, reduce congestion, and manage thousands of system components under a single management solution − satisfying the demands of large corporations.

To operate a Sun Fire 15K server, users must run the Solaris 8 Operating Environment and SMS software. It is also recommended that customers use SunMC and other management services software to enhance server administration and monitoring. The software information is presented as it specifically relates to the Sun Fire 15K server environment. The following graphic depicts the software stack for the Sun Fire 15K.



Sun's Sun Fire 15Ksoftware includes:

- **Solaris 8** − provides advanced features and functionality that give users the high−performance operating environment needed for running mission−critical applications.
- **System Management Services** − helps enable users to control and monitor each domain as well as the Sun Fire 15Ksystem.
- **Sun Management Center** − helps enable users to monitor and manage thousands of Sun systems from a single source. For the Starcat, the software package provides support for the platform and domains.
- **Sun Remote Services** − remote management service that provides proactive problem detection and prompt resolution of system events. This reduces system downtime and provides health and utilization reports that can be an effective IT business planning tool.
- **Solaris Management Console** − client/server application used to manage one or more Solaris domains. It can launch any UNIX application on any Solaris server in the network.
- **Solaris Resource Manager** − helps enable the consolidation of multiple applications onto a single Solaris server. It provides the ability to allocate and control major system resources, allowing service availability for critical enterprise applications, IT−defined groups, and individual users.
- **Solaris Bandwidth Manager** − helps enable control of bandwidth assigned to particular applications, users, and departments that share the same intranet or Internet link. It efficiently distributes network bandwidth and ensures service for mission−critical applications.

The Solaris™ 8 (10/01 or later) Operating Environment

The Solaris™ (10/01 or later) Operating Environment provides an advanced, industrial−grade solution for all customer IT needs, both technical and business; it has the performance, quality, and robustness to deliver mission−critical reliability. The Solaris Operating Environment, with enterprise integration by design, provides easy access to a wide range of computing environments and network technologies while delivering a competitive advantage to business through networked computing, unparalleled scalability, and multi−architecture support. It combines new levels of availability and reliability − to help enable continuous, 24x7 uptime − with massive scalability, sophisticated manageability, and advanced security. With the Solaris Operating Environment, enterprises adapting to the Internet age, as well as dot−com businesses adopting the disciplines of the data center, can increase service levels while at the same time reducing IT risks and lowering service costs.

The Sun Fire 15K server includes the Solaris Operating Environment 8, which has been enhanced to address very large memories and scale to support hundreds of processors. The Sun Fire 15K system will run the Solaris Operating Environment 8 with features and benefits that include:

- Scalability enhancements
- Improved ease−of−use features
- Improved reliability, availability, and serviceability (RAS) features
- Sun Management Center 3.x
- Integrated directory services
- Integrated security
- Hot relief
- Clustering
- True 64 bit computing to address large memory sets
- 32 and 64 bit support
- 12000 ISV applications
- Live upgrade
  – Installing and reconfiguring new versions of the Operating System while the current system is still running.

Sun Cluster 3.0 and the Solaris 8 Operating Environment

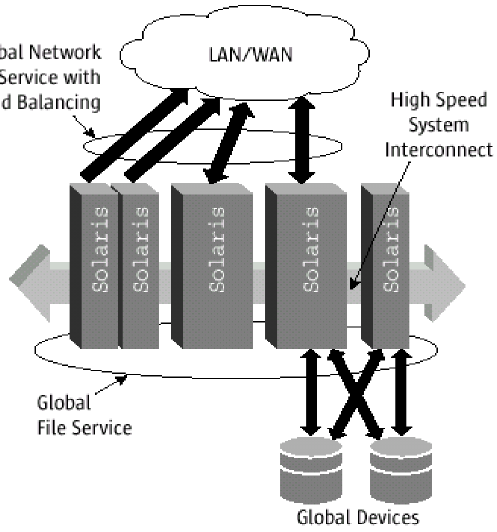
"Sun Cluster 3.0 software is a new approach to creating a cluster computing environment for the networked data center. Based on abstracting applications and services, such as data storage and network connectivity, from the physical hardware. Sun Cluster 3.0 extends a high−availability (HA) environment to provide a single, logical view of a commercial computing environment from both an application services and administrative perspective."

Sun Cluster 3.0 extends the Solaris 8 Operating Environment with Sun’s cluster framework; it helps enable core Solaris services − devices, file systems, and networks − to operate seamlessly across a tightly coupled cluster while maintaining full Solaris compatibility with existing applications. Sun Cluster 3.0 offers a platform that provides high availability (HA) and scalability to everyday Solaris applications through continuous network and data availability. Services that are written to the simple−to−use Sun Cluster 3.0 API can achieve even higher availability as well as scalability.

Summary of Features and Benefits

- Global devices
- Global file service
- Global network services
- Scalable Services
- Failover services
- Faster failover
- Diskless failover
- Simplified, centralized cluster management
- Cluster−enabling of applications
- Eight−node support

Cluster 3.0 Features



Sun Management Center

In combination with the Sun Management Center (SunMC), the networked system controller (SC) gives administrators a powerful tool for system and resource management. SMC software provides features such as domain management, a GUI interface to dynamic reconfiguration, alternate pathing, and SC commands. SunMC also provides hardware information, environmental monitoring and propagation of alarms to associated devices.

- SunMC helps enable flexible integration with the major enterprise management packages

- Critical to SRS (Sun Remote Services)
- Global DR across domains
- Local DR within a domain
- Logical view per domain; tree hierarchy including all hardware and OS components
- Alarms and traps
- Monitor master and standby SCs
- Solaris objects in the SCs
- Environmentals of the system controllers themselves
- Platform Management Information Bus (MIB)

- Platform view and domain view
- Hardware monitoring (presence, state)
- Environmental state and events
- Domain management
- System controller monitoring
- SMS processes on the SC
- Sun Fire 15K specific physical views "...displaying photo−realistic images of hardware components and pointing to components with an associated event, helping to enable administrators unfamiliar with a particular Sun platform to quickly determine which components need to be replaced"
- Support controlled SC

Console

The System Controller provides a service called console. Console provides a “console” for single−user operations. Normally, a System Controller must be on the same subnetwork as the Sun Fire 15K system; with console, they can be anywhere in the world that has a network attachment. The service is provided such that sessions, similar to rlogin sessions, can be provided to X−Windows clients on the same network as the System Controller. This enables system administrators to access the System Controller from any location on the same network as the System Controller.

Console enables:

- Access to UNIX console
- Requires domain administrator authority
- Private internal ethernet when the Solaris^{TM} Operating System is running
- IOSRAM "tunnel" protocol when in OBP

Automated Dynamic Reconfiguration (ADR)

Dynamic Reconfiguration (DR) and Automated Dynamic Reconfiguration (ADR) allow resources to be dynamically reallocated between domains. DR helps enable a physical or logical restructuring of the hardware components of a system while the system is running. This high degree of resource flexibility allows administrators to reconfigure the system easily in order to meet load or schedule demands. As a system’s resources and requirements change from hour to hour or from month to month, DR optimizes the time and money needed to reallocate those resources.
ADR might be used during those times when it is preferable to reconfigure domains on the Sun Fire 15K server automatically based on predetermined schedules, resource or user load, or other system events. Although domains can still be configured manually as needed, ADR ’s high−level, automated UNIX scripts allow fine−grained domain management.

ADR raises domain management resource planning and management, and system−wide management to a new level of flexibility and convenience. Additionally, because ADR reduces the degree of human interaction necessary for domain reconfiguration, errors and time can be significantly reduced −in turn, reducing the overall total cost of operations for the system.

InterDomain Networking (IDN)

Note: IDN is a post−release feature of the Sun Fire 15K
InterDomain Networking has been completely re−written from the original IDN software now offered on the E10000. When IDN is ready for release, there will be an announcement and a detailed description of how IDN works.

Solaris^{TM} Resource Manager (SRM)

Sun’s Solaris Resource Manager software offers the means to allocate, control and monitor system resource usage. Previously, this capability was only available in the mainframe environment. With SRM software, system resources such as CPU, virtual memory, and a number of processes can be allocated to users, groups, and applications. SRM software is a Sun key enabler for server consolidation and increased system resource utilization. Critical application availability is increased because priority can be established over less important applications or functions. It is an effective tool for creating and managing shared service environments. Beyond simple time−sharing schemes, it provides fine−grained, hierarchical control of system resources for users, groups, and applications, enabling an equitable distribution of computational resources within a given Solaris system and promoting server consolidation. Solaris Resource Manager software is particularly effective for use in enterprise servers since it can prevent server resources from being usurped by rogue processes, abusive users, and large computational loads.

Features and benefits include:

– Major system resource controls
– Fair−share allocation
– Hierarchical control model
– Policy−based resource administration
– Resource reporting

Solaris^{TM} Bandwidth Manager (SBM)

Solaris^{TM} Bandwidth Manager is a policy−based, directory−enabled software solution regulating IP networks’ bandwidth usage in LANs and WANs. SBM is a Sun key enabler for providing differentiated classes of services. It helps enable administrators ability to provide high−quality network service by controlling the bandwidth assigned to applications and users, prioritizing traffic, and building advanced bandwidth management policies. These policies make it possible to mirror the business organization into the information system. The policies can then be stored in a directory to be enforced throughout the organization as required.

Solaris Bandwidth Manager allows monitoring of network traffic, receiving detailed network traffic statistics and collecting of accounting information to feed to the organization’s billing system. SBM is customizable and extensible; it provides Java and C programming APIs, which simplifies customization and makes possible the integration of SBM with other applications.

Key features and benefits include:

– Monitoring and allocation of IP traffic priorities and bandwidth
– Directory−enabled
– Policy−based management
– Granular usage reporting

Glossary – Technical Terminology

ACL

Access Control List. In order to assign a board to a domain, the board name must be listed in the Access Control List (ACL). When a board is listed in the ACL, the system controller software is allowed to process addboard or deleteboard requests for that board.

ADR

Automated Dynamic Reconfiguration. Allows one to invoke scripts for tasks such as adding or deleting a system board to or from a domain, moving a board between domains, or for determining the status of a system board.

alternate pathing

A software feature which gives you the ability to have multiple paths to the same device from one domain. This provides an extra degree of fault tolerance. For disks, this is MPxIO. For networks it is IPMP.

Automatic System Recovery (ASR)

Provids system operation in the event of a hardware failure. Identifies and isolatse a failing hardware component, and builds a bootable system configuration without the failed hardware component.

boardset

In the Sun Fire 15K, one expander board plus its attached System board and/or I/O board.

CDC

Coherence Directory Cache. A cache of information associated with each home agent, located in SRAMs on the expander board, which may record a cache line where its owner is or which expanders may have a shared copy.

Cheetah

The internal name for the UltraSPARC−III processor.

CLI

Command Line Interface. A user interface to a computer’s operating system or an application in which the user responds to a visual prompt by typing in a command on a specified line, receives a response back from the system, and then enters another command, and so forth.

Coherent

A somewhat overused (and therefore ambiguous) term. *Coherent transactions* may refer only to RTO, RTS, and WB type transactions; or is sometimes used to refer to any transactions to cacheable address space.

COMA

Cache−Only Memory Architecture. An architecture in which main memory is treated like a large cache, and memory pages can be present in more than one place. WildCat uses *Simple COMA*, renamed Coherent Memory Replication (see).

control boardset

Plugs into one of two control slots on the plane interconnect. Consists of a plane interconnect support board, a system controller, and a peripheral board.

cPCI

Compact PCI. A form of PCI that allows an individual PCI card to be removed and replaced while the system is running.

CPU

Central Processing Unit. An UltraSPARC III processor in this context.

Domain

A logical grouping of system boards. Each partition contains up to two domains. Domains do not interact with each other. Domains differ from partitions in that they share Repeater boards. A domain is able to run its own copy of the Solaris operating environment and has its own host ID. A domain must contain at least one system board and one I/O board.

domain isolation

Mechanism where by the AXQ confirms that the address transaction destination is in the same domain as the source. Enabling technology for "split slot" configurations.

Double−pumped

A degraded mode in which the information is sent over half the wires using twice the normal number of cycles, thus avoiding a broken wire.

DR

Dynamic Reconfiguration. Adding or removing resources to a domain, such as removing a CPU board from one domain and giving it to another.

DSD

Dynamic System Domain. Mainframe−style partitioning allows a single UNIX server to be divided logically into multiple servers, creating "systems within a system."

dynamic reconfiguration

Software support for the hardware configuration changes made to a domain running the Solaris operating environment. Dynamic Reconfiguration handles the software aspects of dynamically removing a defective board from the system and installing a replacement board without bringing the system down and with minimum disruption to user processes running in the domain.

ecache

External cache. In Cheetah, this is the L2 cache, with the data kept in external SRAM chips and the tags kept in the Cheetah chip.

ECC

Error Correcting Code. Extra bits that go along with data bits to detect or correct errors. See SECDED.

expander board

The Sun Fire 15K board that interposes between the CPU and IO boards and the plane interconnect. The expander board contains the AXQ and SDI ASICs.

fault tolerance

The ability of a system to continue to perform its functions, even when one or more components have failed.

fireplane bus

The interconnect architecture used by the Sun Fire servers when communicating between L1 and L2 devices.

FRU

Field replaceable unit or replacement part.

HA

High Availability.

Hot Plug

Physically connecting hardware to a powered system.

Hot Swappable

The ability to physically add or remove a component from the system whiule the system is up and allowing that hardware to be configured into the system without interruption to system operation. Hot swap is combination of Hot plug and Dynamic Reconfiguration.

IDN

Inter−Domain Networking.

IPMP

IP Multipathing. Increases network availability and performance.

keyswitch

See virtual domain keyswitch.

Latency

The time between initiating a request from data and the beginning of the actual data transfer.

LVD

Low Voltage Differential. (Ultra2 and beyond) LVD reduces the amount of power needed to drive the SCSI bus, and increases the cable length from 3 meters to 12 meters.

MaxCPU board

A Dual Proc board, which plugs into slot 1 (the “I/O” slot) of the Expander. Originally called the MacCat board.

MPxIO

MultipathI/O. I/O framework allowing Solaris to represent and manage storage devices which are accessible through multiple host controller interfaces within a single instance of Solaris.

MTBF

Mean Time Between Failure. Is a measure of how reliable a hardware product or component is.

MTTR

Mean Time to Repair. The average time it takes to repair a Failed component.

NAFO

Network Adapter Failover. Exclusively a feature of Sun Cluster.

NetCon

Network Console. System controller service that provides a "console" for single−user operations.

No Single Point of Failure

Implies that for any one component failure (which may cause a system crash), the system can still be rebooted with full data access, though perhaps with degraded performance. When a failure occurs, Sun systems automatically reboot (fast reboot) and reconfigure to recovery.

NUMA

Non−Uniform Memory Access. An architecture in which memory and processors are in clumps, and accesses to memory in another clump is possible but slower.

NVRAM

Non−Volatile RAM. Any semiconductor memory device that does not lose its contents when power is turned off.

OLTP

Online Transaction Processing. Is a class of programs that facilitates and manages transaction−oriented applications, typically for data entry and retrieval transactions.

partition

A hardware feature that provides very good hardware failure isolation between two groups of system boards and their associated Repeater boards. Each partition has its own set of Repeater boards. When a system is divided into two partitions, the system logically behaves as two separate systems. Now called a segment.

RAS

Reliability, Availability, and Serviceability.

RCM

Reconfiguration Coordination Manager. A generic framework which allows dynamic reconfiguration (DR) to interact with system management software. The framework enables automated DR removal operations on platforms with proper software and hardware configuration and provides enhanced error reporting on DR operation failures.

SBM

Solaris^{TM} Bandwidth Manager. A policy−based, directory−enabled software solution regulating IP networks’ bandwidth usage in LANs and WANs.

SC

System Controller (see).

Scalable Shared Memory (SSM)

A mode of the Fireplane interconnect which allows multiple Fireplane buses to be connected together.

Shared Resource Domain (SRD)

A Sun Fire 15K domain that is permitted to send and receive a restricted set of transactions to/from other domains (its client domains), thus allowing it to facilitate and control inter−domain communication.

shared expander

In the Sun Fire 15K, an expander which is either an SRD expander (see) and/or a split expander (see).

Slot−0 board

A board that has an off−board bandwidth of 4.8 GBps. One type of slot−0 board is used in the Sun Fire: the System board.

Slot−1 board

A board that has an off−board bandwidth of 2.4 GBps. Three slot−1 board types are used in the in Sun FIre servers: PCI, cPCI, and 3800 cPCI.

SMP

Symmetric Multi−Processor. The Sun Fire servers are SMPs.

SNMP

Simple Network Management Protocol. SNMP is any system listening to SNMP events. This is usually the system with the Sun Management Center software installed.

split expander

When the two system boards in a boardset are in different domains, since the expander board keeps the transactions separate for each system board.

SRD

Shared Resource domain. A Sun Fire domain that is permitted to send and receive a restricted set of transactions to/from other domains (its client domains), thus allowing it to facilitate and control inter−domain communication.

SRS

Sun Remote Services. Continuous system monitoring that utilizes an intelligent agent based architecture to monitor key systems variables 7x24. When a problem is detected, an alarm is automatically generated to notify SRS engineers and initiate resolution.

SSM

Scalable Shared Memory. That part of the Safari architecture that enables point−to−point coherency, thus minimizing system addressing bottlenecks. Coherency messages only get sent to components that need to be involved in a particular transaction.

System Controller

The system controller consists of the System Controller board and the system controller software. The system controller provides communication pathways for console traffic and other data that needs to be passed between the system controller and the system. The system controller software monitors and controls the system, manages hardware, and configures domains.

System Controller board

A board containing a microSPARC™ processor, which oversees operation of the system and provides clocks and the console bus.

UltraSPARC−III

The SPARC V9 processor used in all systems comprising the Sun Fire family of systems. Internally known as the Cheetah.

US−III

See UltraSPARC−III.

UltraSPARC IV

Second generation processor model of the Safari−bus generation. Pin compatible with UltraSPARC III.

UPA

Ultra Port Architecture.