Sun Fire™ Link

Just the Facts

May 22nd 2003

Copyrights

Table of Contents

Positioning

Introduction

The Sun Fire™ Link product is Sun's highest performing cluster interconnect. It is available on the Sun Fire 6800 and Sun Fire 12K/15K servers and is supported by Sun™ Cluster and Sun HPC ClusterTools™. Sun Fire Link therefore serves both our commercial cluster market and our high performance technical computing market. The high data rate and low latency of Sun Fire Link should improve the performance of cluster applications.

Sun Fire Link can also be used as a high performance pipe between two Sun Fire servers for file transfers using TCP/IP.

Product Family Placement

Clustering of servers requires three components

- A cluster-aware operating system. This is the Solaris™ Operating System.
- Clustering software as the interface between Solaris and our customer's application. Customers have a choice here depending on the nature of the application: Sun Cluster or HPC ClusterTools.
- An interconnect between the servers for heartbeat messages, data and application coordination information.

We offer our customers a choice of cluster interconnects at various performance levels and price points as shown in the following table. The goal is to match the performance of the interconnect to the requirements of the customer's application.

Table 1

Type	***Usage***	***Max. # nodes***	***Data rate per link (hardware)***	***Latency (software) Microseconds***	***Price***
Sun Fire Link	Sun Cluster & HPC	8	1200 MBps	<4	$$$
SCI	Sun Cluster	4	200 MBps	<10	$$
Myrinet (3rd party)	HPC	16	140 MBps	16	$
Gigabit Ethernet	Sun Cluster & HPC	Large	100 MBps	100	$

Key Messages

- Sun can supply our customers with a choice of cluster interconnects. Sun Fire Link is the highest performing interconnect based on its high data rate and low latency.
- Sun Fire Link is available for the Sun Fire 6800 and Sun Fire 12K/15K servers. Up to eight of these servers may be clustered in any mix.
- For commercial applications, Sun Fire Link is supported by Sun Cluster software. For technical applications, customers can use Sun Fire Link with Sun's HPC ClusterTools software.

Availability

- The delivery dates for Sun Fire Link are platform dependent.
- For the Sun Fire 6800, GA was in February 2003.
- For the Sun Fire 12K/15K, RR is expected to be in June 2003 with GA in July 2003.
- At RR the software and documentation will be localized but Operations will not have reached volume production.

Target Users

In the Sun Cluster environment, Sun Fire Link will be selected by customers interested in exploiting the SunPlex^{tm} platform benefits of Sun Cluster 3.0 – i.e., scaling the cluster horizontally.

For HPC and technical computing, Sun Fire Link should allow clustered applications to scale better compared to the use of a slower speed interconnect.

Target Markets

Sun Fire Link is not industry specific. It is a horizontal product that will be used across all industries.

Selling Highlights

Market Value Proposition

Sun Fire Link provides our customers with a high performance, fault resilient cluster interconnect that should allow many clustered applications to run faster.

Applications

The middleware applications for Sun Fire Link are Sun Cluster 3.0 (update 3, the 5/02 release, onwards) and HPC ClusterTools 4.0 (onwards). In the commercial clustering space, the primary higher level application is Oracle[R]. For HPC clustering, any technical application that supports MPI (Message Passing Interface - an industry standard for HPC) can run in a cluster that uses Sun Fire Link as the interconnect.

Compatibility

Sun Fire Link is qualified to run on Sun Fire 6800 and 12K/15K servers.

Enabling Technology

Technology Overview

Sun Fire Link is a cluster interconnect. Its high data rate and low latency is designed to allow clusters of Sun Fire servers to perform better. Key technology factors are:

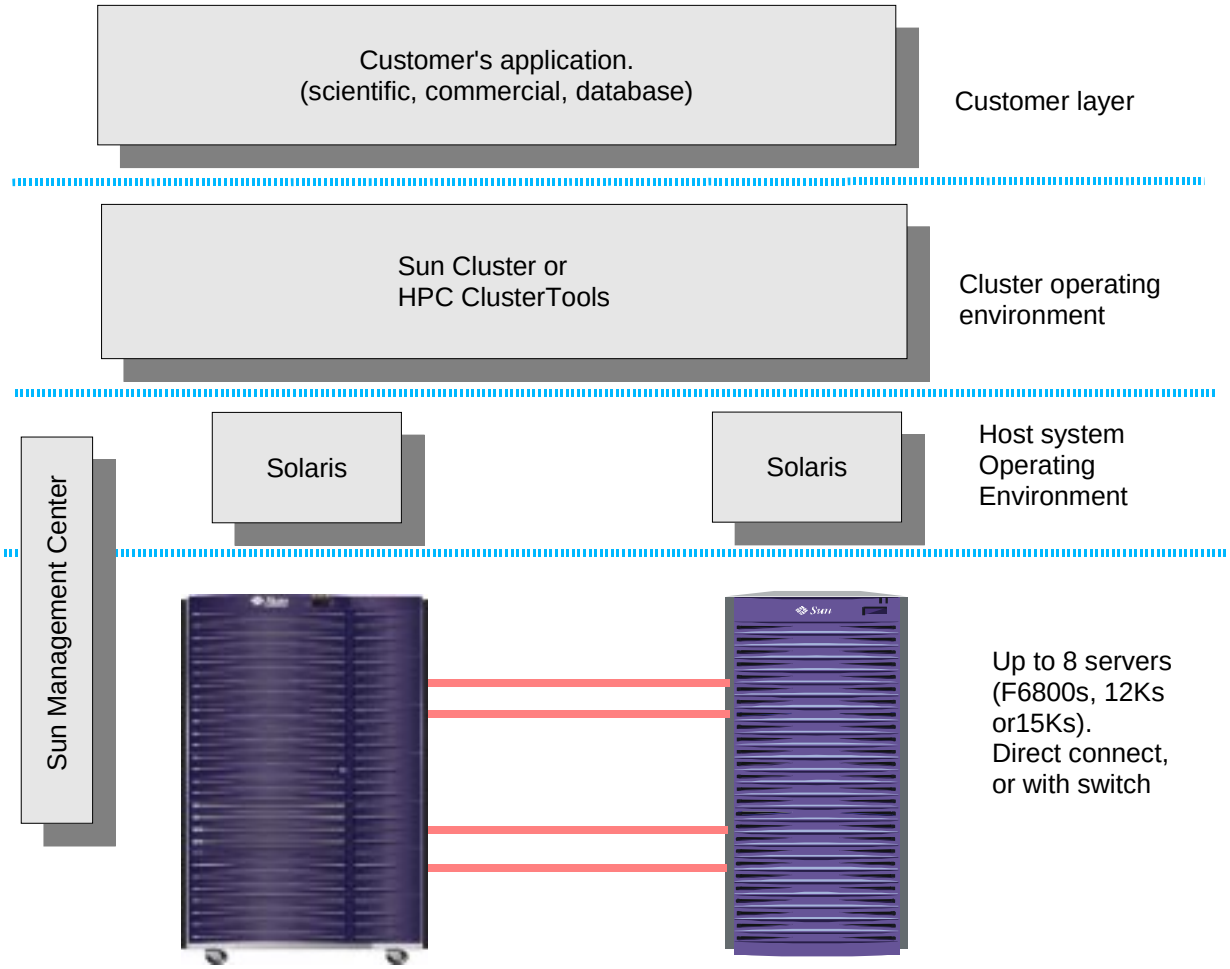
- Remote Shared Memory (RSM) software technology. This speeds up cluster transactions by by-passing the traditional Solaris OS TCP/IP communications protocol overhead. Our customers can take advantage of RSM by using the applications programming interface (API) that we have published.
- The optical connections. Information is passed between servers over optical links that carry 12 bits in parallel at Gigabit data rates. This required the development of an optical ASIC that can simultaneously convert 12 optical channels to electrical signals, and vice-versa. The Sun Fire Link optical technology is far superior to Fiber Channel used for storage as the latter has a single optical channel.
- The Sun Fire Link ASIC. This is a high speed complex networking chip that implements the cluster protocols. There is also a derivative chip, the Sun Fire Link switch ASIC, that allows us to build an efficiently packaged high speed switch.

System Architecture

Overview

Sun Fire Link is clustering technology which can be used to scale the upper end of the Sun Fire family - specifically the Sun Fire 6800 and the Sun Fire 12K/15K servers. Figure 1 shows how Sun Fire Link fits into the bigger picture that includes the Solaris OS, Sun's clustering software (Sun Cluster or HPC ClusterTools) and the customer's application.

Figure 1. Conceptual view



Sun Fire Link Hardware

Sun Fire 6800 and 12K/15K servers that participate in a Sun Fire Link network are configured with Sun Fire Link optical ports. The hardware that provides these ports is the Sun Fire Link assembly. Servers may be directly connected with Sun Fire Link cables for a cluster of two or three nodes. Beyond this, the Sun Fire Link switch is used. This section describes this hardware.

Sun Fire Link Assembly

The Sun Fire Link assembly is the interface between the system interconnect (the Sun Fireplane[tm]) and the Sun Fire Link network. Unlike a PCI adapter, Sun Fire Link taps directly into the system interconnect and runs at system interconnect speeds. The standard I/O assemblies for the F6800 and F12K/15K servers are replaced by a Sun Fire Link assembly. For packaging reasons there is one assembly for the F6800 and another for the F12K/15K although the logic on both is similar. Each Sun Fire Link assembly includes one standard PCI bridge chip which provides two adapter slots (one 33 MHz and one 66 MHz, both 32 or 64 bits wide) for storage or networking adapters. In addition, each Sun Fire Link assembly has a Sun Fire Link ASIC which supports two optical links. Figure 2 shows the Sun Fire Link assembly for the F6800 on the left and for the F12K/15K on the right.

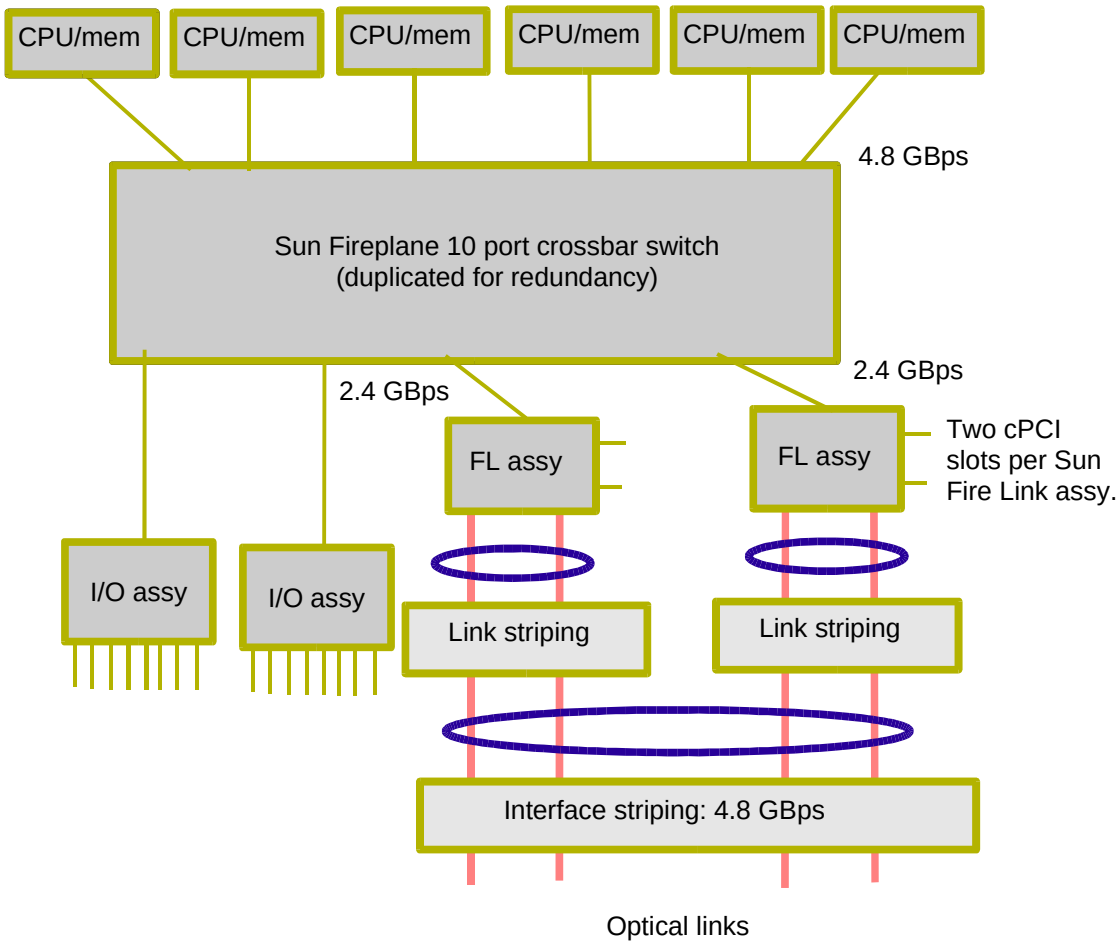
Figure 2. Sun Fire Link Assembly for the Sun Fire 6800 and Sun Fire 12K/15K servers

The Sun Fire Link assembly for the F6800 server is derived from the compact PCI (cPCI) I/O assembly for this server. Hence the two additional general purpose adapter slots only accept

cPCI cards. For the F12K/15K servers, these two additional adapter slots are for PCI cards in hot plug cassettes.

Each Sun Fire Link assembly has a pair of optical connections. Dual optical connections double the data transfer rate and also provide redundancy should an optical interface fail. In addition, Sun Fire Link assemblies are themselves used in pairs to again double the data transfer rate and provide a higher level of availability. Figure 3 below shows the hardware architecture of a Sun Fire 6800 server equipped with Sun Fire Link. The Sun Fire 12K/15K servers are similar.

Figure 3. Sun Fire 6800 architecture



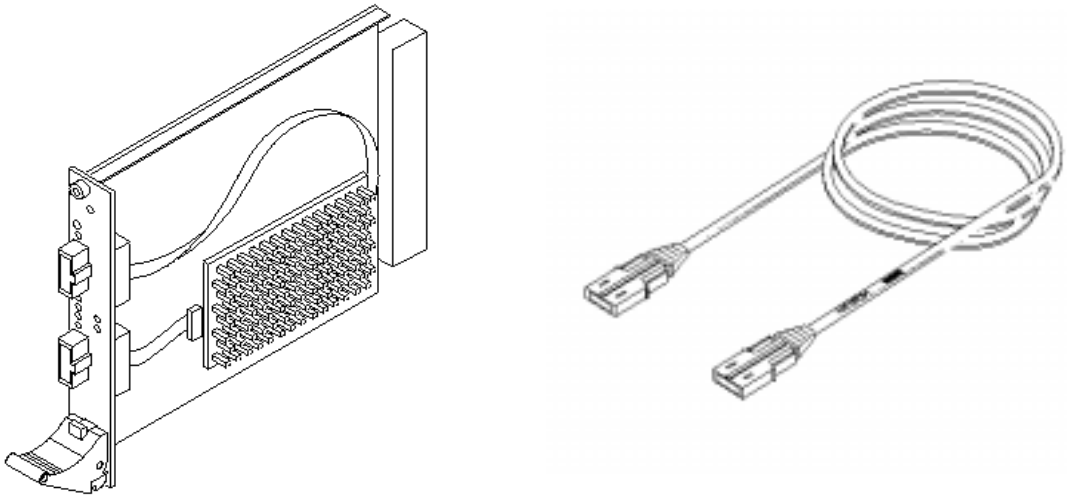
In the F6800 server, the Sun Fireplane switch has ten ports (although it is packaged as two five port boards). It is also duplicated for redundancy, not shown in figure 3. In the F6800 without Sun Fire Link, there can be up to six CPU/memory boards and up to four I/O assemblies. When Sun Fire Link is used, two of these I/O assemblies are replaced with the Sun Fire Link assemblies with their two optical ports and two general purpose cPCI adapter slots.

With the two optical ports from each Sun Fire Link assembly, we have "link striping". With two Sun Fire Link assemblies (and hence four optical links) we have "interface striping". Striping achieves redundancy and increases the data rate. The graphic shows the raw hardware data transfer rates for Sun Fire Link. Each Sun Fire Link assembly can consume the total data bandwidth (2.4 Gbytes per second) available at each I/O port.

Interface striping requires that both Sun Fire Link assemblies in the Sun Fire 6800 be configured in the same domain. A maximum of two additional domains (for a total of three) can be configured. If additional domains are used, the domain containing Sun Fire Link will only have access to the four cPCI adapter slots on the Sun Fire Link assemblies.

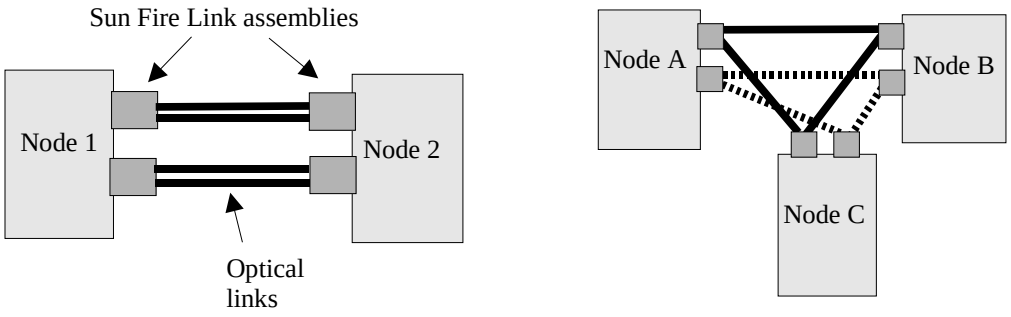
The Sun Fire Link optical interface, the Paroli, is mounted on a cPCI form factor card that plugs into the Sun Fire Link assembly. These cards, known as "Sun Fire Link optical modules" are cPCI *form factor* cards but electrically not actual cPCI cards. These optical cards can be hot swapped in the event of a failure. Optical cables plug into the modules. Figure 4 shows the optical modules and cables. A Sun Fire Link cable is actually a transmit cable and a receive cable clipped together.

Figure 4. Sun Fire Link Optical modules and cables



Note that Sun Fire Link is a *data center* cluster interconnect. Therefore the distances are limited. Cables are supplied in lengths of 5m, 12m and 20m (respectively 16, 39 and 65 feet). Sun Fire Link configurations of two or three nodes can use point to point direct connect optical links as shown in figure 5.

Figure 5. Direct connect, two and three nodes



In the two node case shown in figure 5, there is link redundancy as well as Sun Fire Link assembly redundancy. For the three node direct connect, There is still redundancy at the Sun Fire Link assembly level. But as there are half the number of links, data bandwidth is halved.

Sun Fire Link Switch

The Sun Fire Link eight port switch is used when clustering more than three nodes. Its function is to take signals from a node and route them, with minimal latency, to any one of the seven other nodes. At the heart of the switch is a cross-bar ASIC that concurrently deals with the data traffic from all eight ports.

The Sun Fire Link switch, shown in figure 6 below, is packaged in a chassis that can mount in a standard-size 19 inch wide rack. Key features are:

- Redundant AC line cords to allow the switch to be powered from two independent sources.
- Redundant cooling trays which may be hot swapped in the event of the failure of a fan.
- Redundant power supplies that can also be hot swapped.
- A switching assembly board for the cross-bar ASIC. If this assembly fails, the switch will not function. It may be replaced with switch power on.
- A switch system controller (SSC) to allow management of the switch using Sun Management Center. The SSC has both a serial connection and an Ethernet connection for management and servicing of the switch. Should the SSC fail the switch will usually continue to function (but can not be administered). Manual intervention is required to bring the replacement SSC on line.
- Slots for up to eight Sun Fire Link optical modules. These optical modules are the same as those used with the Sun Fire Link assemblies.

Figure 6. Sun Fire Link switch

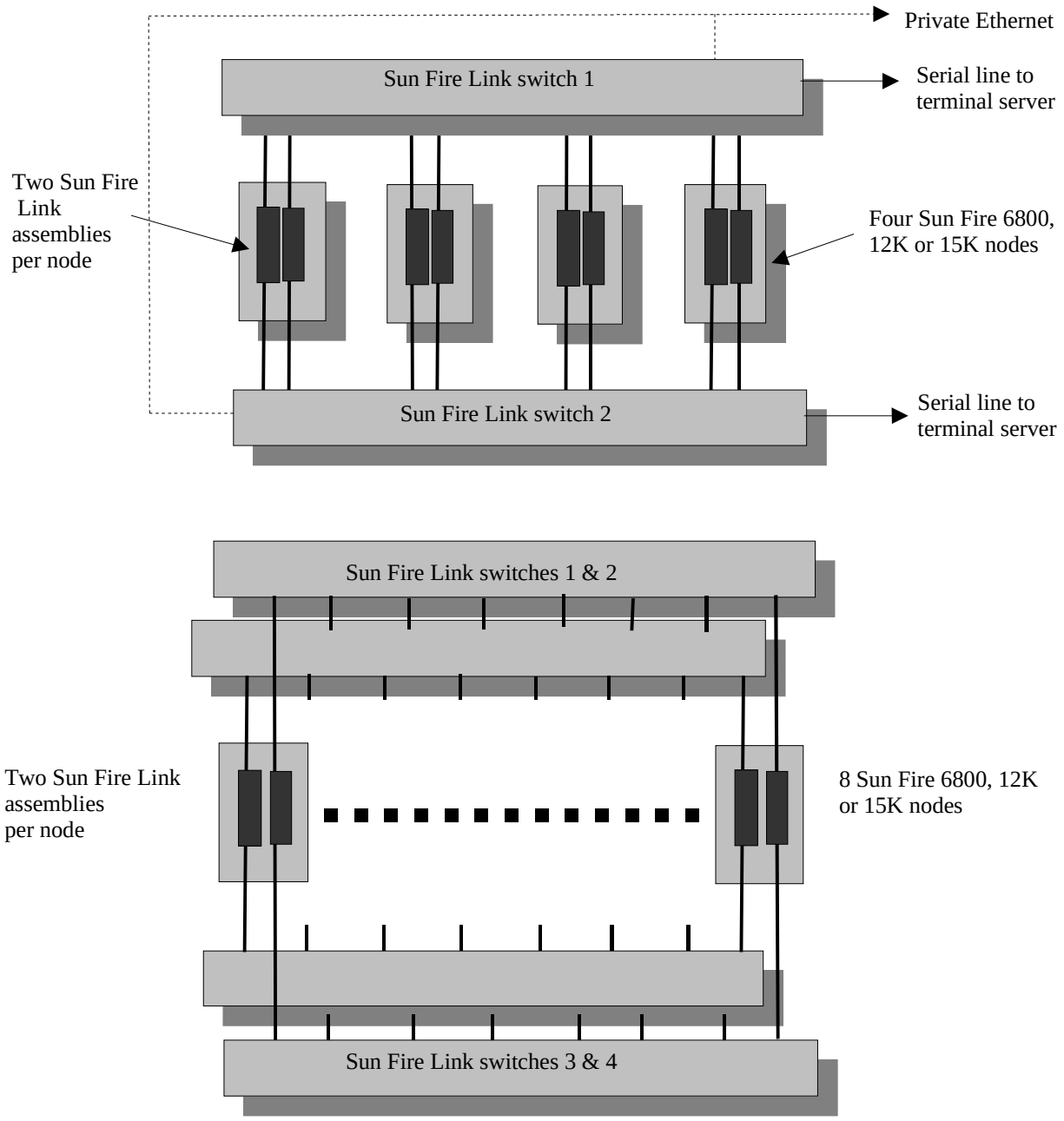
The Sun Fire Link switch is an actively managed entity. The SSC runs an embedded operating system out of flash memory. Administration of the switch is primarily via the Ethernet port. The RS-232 serial port is used for low level error messages and as a backup should the Ethernet network fail.

Sun Fire Link switches are always used in pairs to protect against failure of the switching assembly. Customers with two or three node clusters will often choose to use the switch instead of direct-connect Sun Fire Link cabling. A key advantage in favor of the switch for the three node cluster is that link redundancy and full bandwidth are restored. Another advantage of using the switch is that it is possible to fully configure the cabling for additional nodes, and verify communications capability between existing nodes and newly added nodes, without disturbing the online cluster traffic. The cluster application software can then be reconfigured at a convenient time to take advantage of the additional resources.

Should a Sun Fire Link switch fail, it can be replaced without bringing down the cluster.

The figure below show how the switches are used - for two to four nodes - and for five to eight nodes.

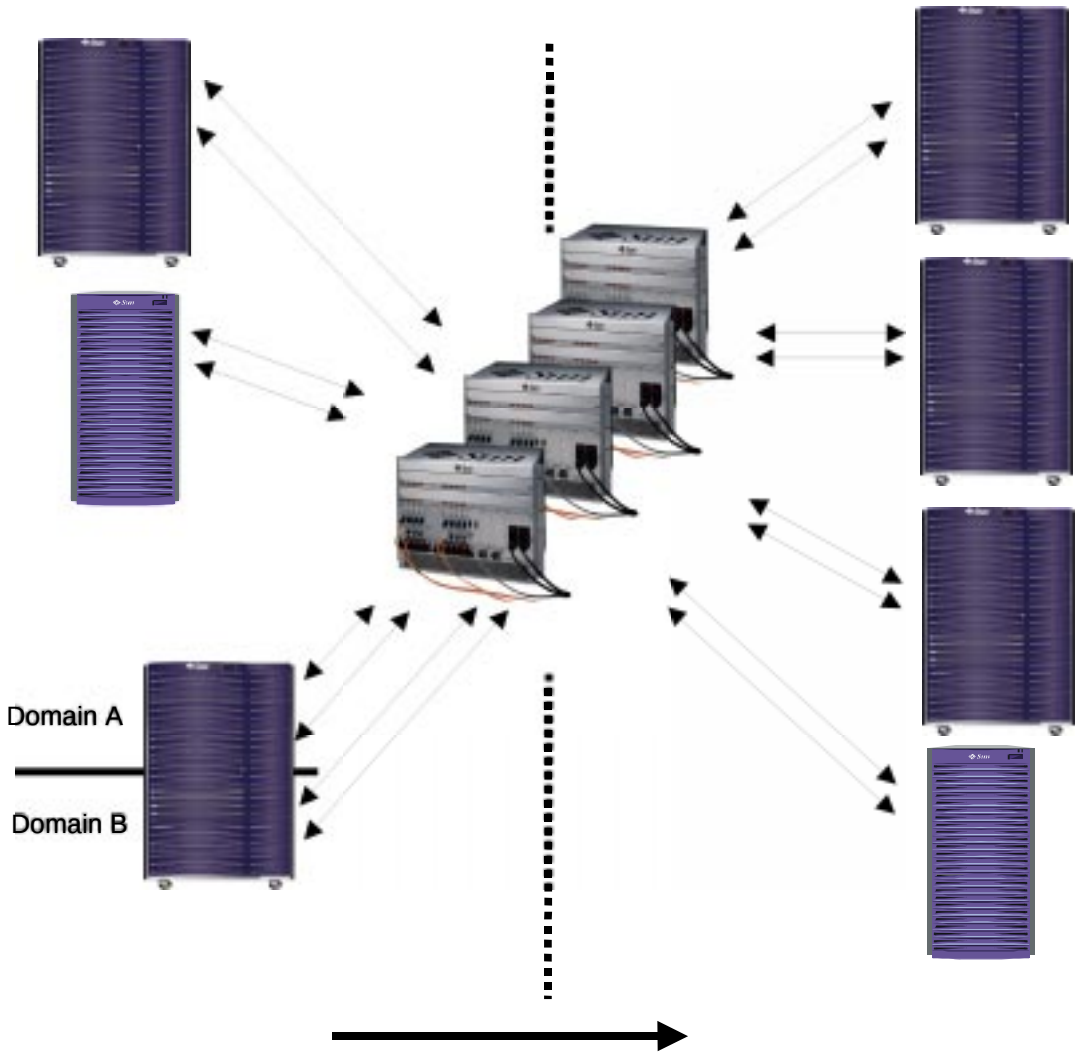
Figure 7. Two to eight node clusters



The Big Picture

Using the Sun Fire Link hardware described above, we can build an eight node cluster from F6800, 12K and 15K servers in any mix. For the F12K and 15K servers, we can cluster "in the box" so this becomes eight nodes, or a lesser number of nodes with domains clustered within the node. Figure 8 shows five Sun Fire 12K or 15K servers clustered with two Sun Fire 6800 servers. One of the Sun Fire 15K servers is split into two domains which are part of the cluster.

Figure 8. Eight clustered Sun Fire servers (or Domains)



Bi-section bandwidth of 19.2 Gigabytes/second
(4 servers each transferring at 4.8 Gigabytes/second)

Each Sun Fire Link assembly has two optical links. Each can transmit data at 800 Megabytes per second and receive data at the same rate. This gives a theoretical Sun Fire Link assembly bi-directional data transfer rate of 3.2 Gigabytes per second. However, the Sun Fireplane interconnect constrains this to 2.4 Gigabytes per second. As there are two Sun Fire Link assemblies used per node (or domain), this doubles to 4.8 Gigabytes per second.

For eight clustered nodes (or domains) the bi-sectional bandwidth is 19.2 Gigabytes per second.

Sun Fire Link Software

Sun Fire Link has the following software components:

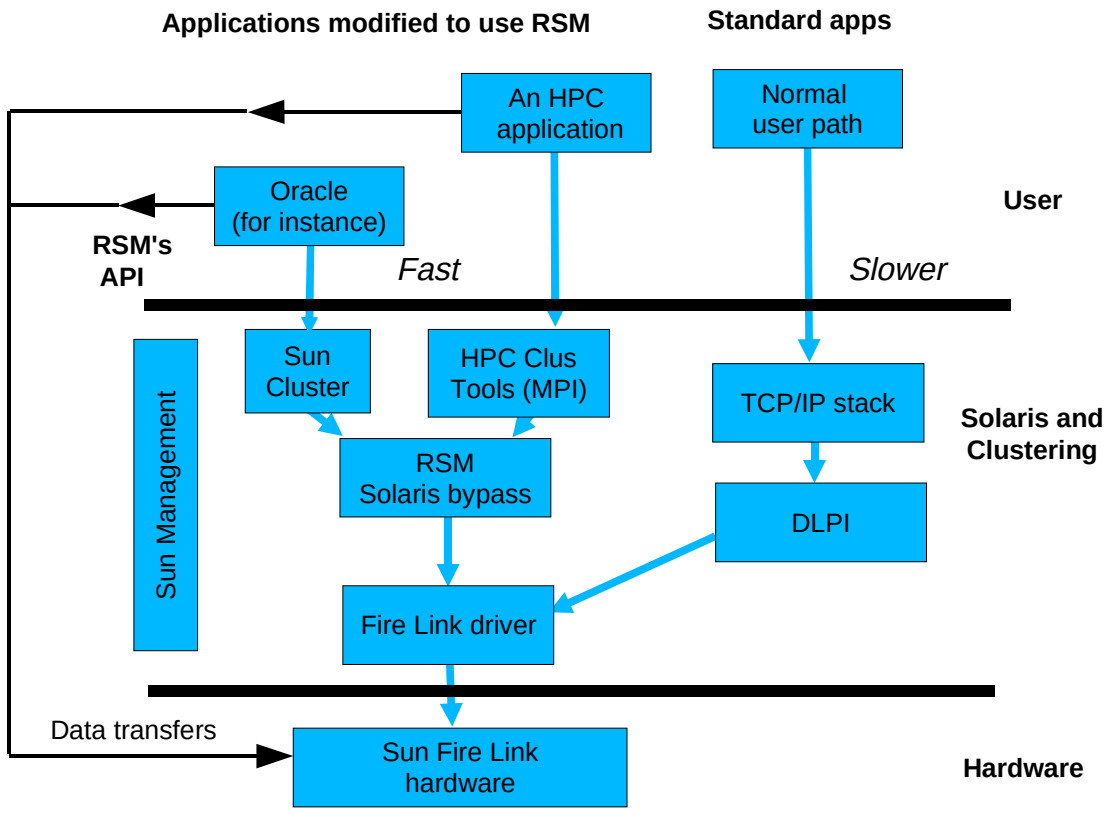
- Solaris OS support - included with the Solaris 8 OS (version) 2/2002 and the Solaris 9 OS (version) 04/2003 onwards.
- Sun Fire 6800 System Controller firmware upgraded to support Sun Fire Link. For the Sun Fire 12K/15K this would be the System Management Software that runs on the System Controller.
- Sun Fire Link switch firmware.
- Support for Sun Fire Link by Sun Management Center (Sun MC) software version 3.0 onwards. This support is principally supplied by the Sun Fire Link Fabric Manager that can be viewed as a Sun MC application.

The Sun Fire Link hardware interface is managed by a device driver that is part of the Solaris OS. The driver implements Remote Shared Memory (RSM). This is a performance feature that allows applications to directly perform operations on remote memory as if it were local memory.

Remote Shared Memory

Figure 9 below uses the layered approach to show what RSM is and does - the layers being the hardware at the bottom, the Solaris OS and Sun's clustering software in the middle and the user application at the top.

Figure 9. Software layers



The right side ("normal user path") shows the way an application traditionally accesses the Solaris OS and the Sun Fire Link interconnect. First the TCP/IP stack, then DLPI and finally the Sun Fire Link driver. RSM is a way of by-passing TCP/IP and DLPI, and hence achieving considerably less software delay. The figure has been somewhat simplified with respect to RSM because Sun Cluster or HPC ClusterTools only have to communicate with RSM for control purposes. After that, all client data is transferred directly into the memory of the remote node bypassing system calls into the kernel, thereby significantly lowering the memory latency.

The RSM application programming interface (RSM API) is available for use by applications programmers desiring to take full advantage of Sun Fire Link's communications capabilities.

The MPI implementation in Sun's HPC ClusterTools 4.0 software uses the RSM API to transparently provide high performance low latency messaging to applications communicating using the MPI standard.

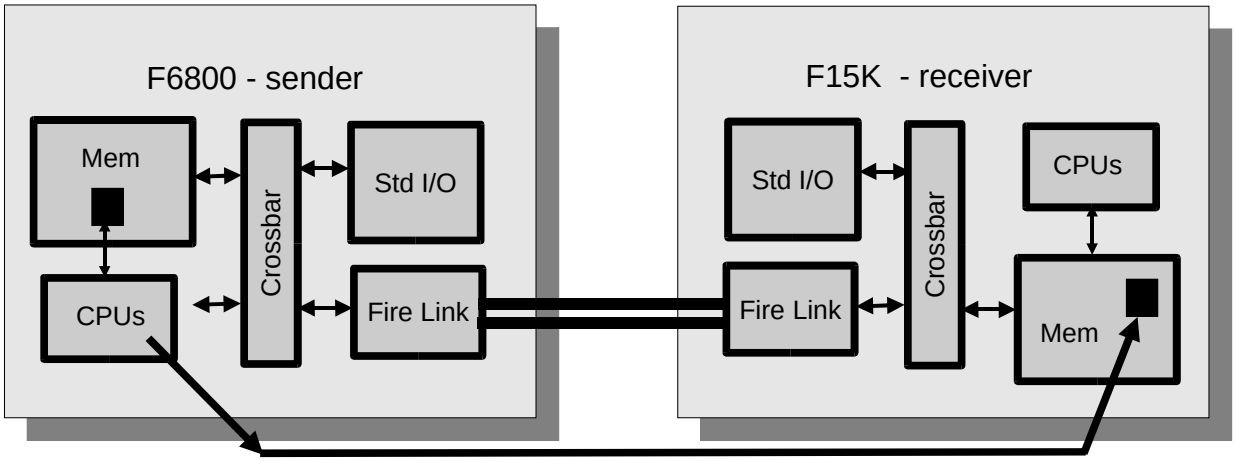
Oracle 9i RAC release 3, when used with Sun Cluster 3.0 software, makes use of the RSM API for key cluster communications operations. The RSM API is available for use by customers in their own clustered applications if they require high performance communications.

Figure 10 shows how the hardware behaves when RSM is being used. What is taking place is the movement of information from the sending server's memory to the receiving server's memory as fast as possible. With Sun Fire Link hardware and RSM, an application maps remote memory (up to 4 GB) directly into user address space. A CPU on the sending node can then simply write to remote memory without making system calls into the kernel, or involving

CPU cycles on the receiving node. This results in a much lower latency compared to using the traditional TCP/IP path. Note that there is no cache coherency between the memories of the two systems.

Although remote memory is mapped directly in to the user's address space, any faults occurring due to access of the remote memory (for instance, memory errors or link failures) will be detected and handled by the Sun Fire Link hardware. Errors are reported to the application through the RSM software.

Figure 10. RSM operations



Logical flow of information

Management and Administration

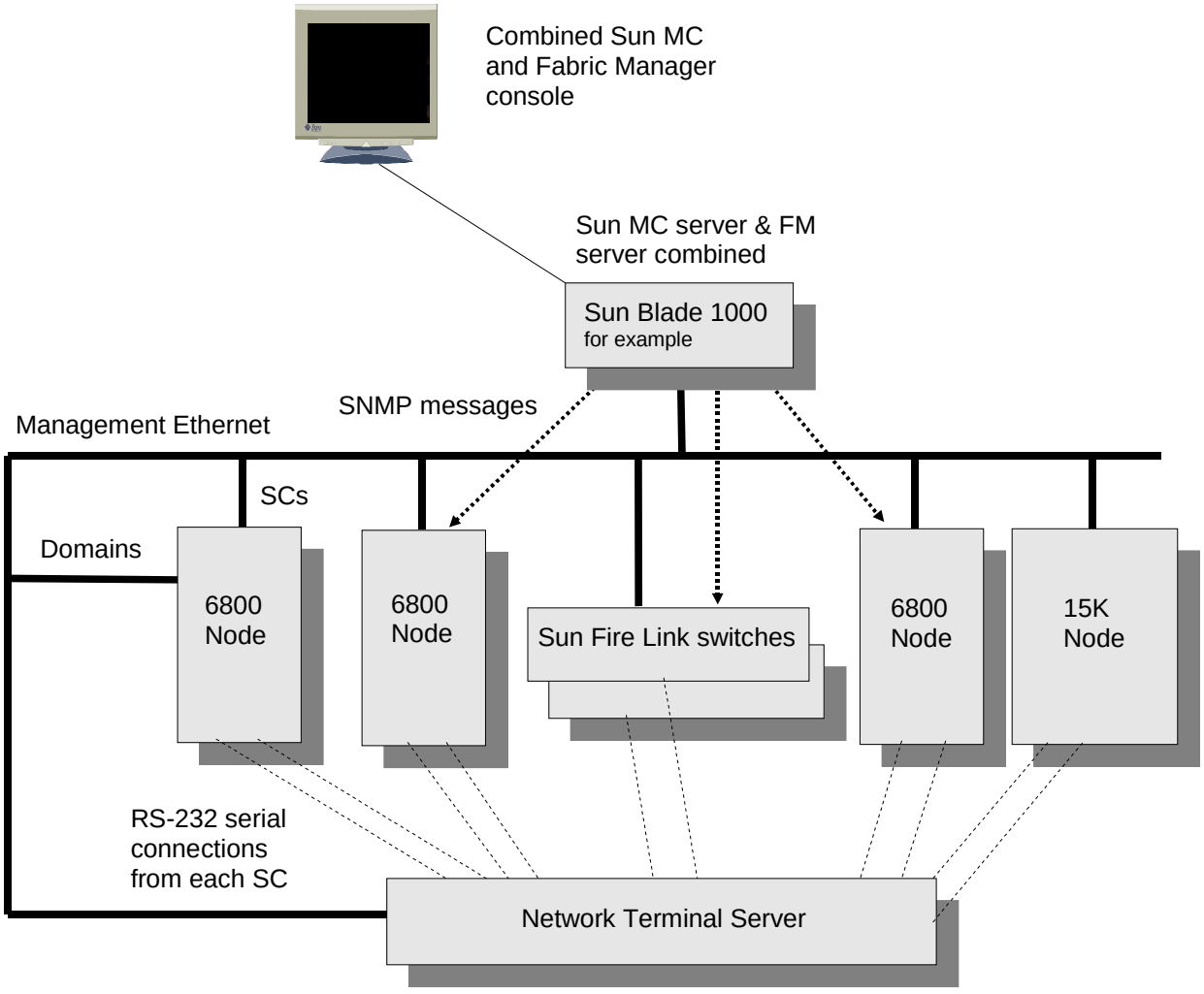
Sun Fire Link is a complex product and it requires sophisticated management tools to:

- Aid customers in configuring the available hardware to meet their requirements. Some of the tasks managed in this way are the bringing up and taking down of optical links, the addition of new links to the fabric and the configuration of links into stripe groups to increase throughput.
- Monitor for faults and problems - as well as ensure that the cluster achieves its expected performance.
- Only allow node topologies that have been tested by Sun. With a product such as Sun Fire Link, there are many ways of cabling a network. Management tools are needed to guarantee customers that they are using a Sun-tested and supported configuration.

The above tasks are handled by Sun Fire Link's fabric manager (FM). The FM is integrated into Sun Management Center (Sun MC) which is Sun's umbrella product for the management of Sun servers. Sun Fire Link may be used without Sun MC by using just the command line

interface provided by FM. However this is not recommended because Sun MC's GUI makes the task of diagnosing and correcting network problems that much easier. Figure 11 shows how the administration components relate to each other.

Figure 11. Sun Fire Link administration



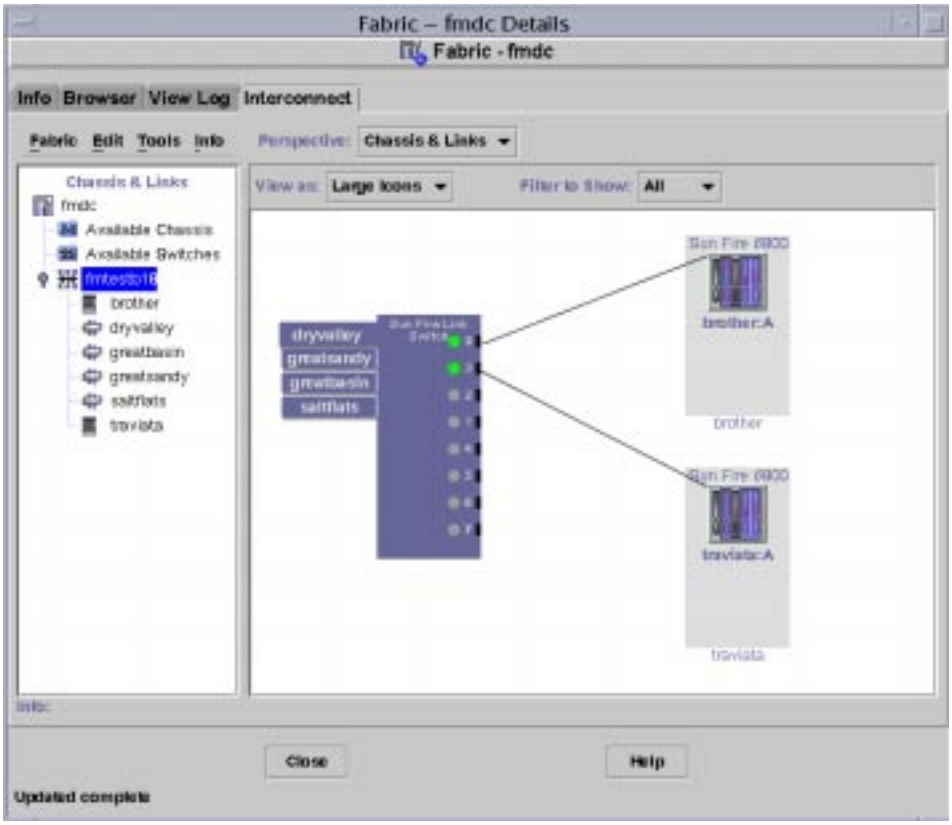
Each System Controller (SC) on the servers (F6800s, F12Ks, F15Ks) has a serial port and an Ethernet port. Normally these serial connections, and those from the Sun Fire Link switches, are aggregated using a terminal server whose output is fed to the management Ethernet. The management Ethernet is normally dedicated to this task and is not attached to the site Ethernet.

Sun Management Center requires two additional hardware components – a Sun MC server and a workstation. For smaller installations, the workstation can do double duty as the server too.

(Refer to Sun MC documentation for information on sizing the Sun MC server). The FM runs on the Sun MC server and shares the same display.

The Sun Fire Link Fabric Manager controls and monitors cluster components. A system administrator can send tasks to the Fabric Manager and this in turn sends instructions to Sun Fire Link agents that are running on the servers and the switches. These agents respond with configuration, status and error information that is sent to the FM server. These are the SNMP messages as shown in figure 11. A typical Fabric Manager screen, viewed using Sun MC, is shown in figure 12.

Figure 12: Sun Fire Link administration



Sun Fire Link and its applications

The two Sun applications used with Sun Fire Link are Sun Cluster and HPC ClusterTools. They use identical Sun Fire Link hardware but use it differently.

HPC ClusterTools uses “single controller” mode where both Sun Fire Link assemblies in a domain are joined into a single logical pipe over all four links. Failure of a link, or of a complete assembly, is managed by the driver software that automatically routes the traffic over the remaining links.

Sun Cluster uses Sun Fire Link in “dual controller” mode. In this mode each Sun Fire Link assembly is individually addressable by the Sun Cluster software. This mode is more management intensive at the application level with the Sun Cluster software performing the recovery should a Sun Fire Link assembly fail. Sun Cluster software is also responsible for the distribution of the cluster traffic between the two controllers. (One does not sit idle waiting to take over should the primary fail).

Features and Benefits

Features	Benefits
• A high performance cluster interconnect	• The potential to improve the performance of clustered applications
• Available for the Sun Fire 6800, 12K and 15K	• Improved competitiveness of these high end servers when used in a clustered environment
• Supported by Sun Cluster 3.0	• Sun Fire Link is designed to improve the horizontal scaling of commercial applications compared to the use of slower cluster interconnects
• Supported by HPC ClusterTools	• When Sun Fire Link is used as the cluster interconnect, many clustered technical and scientific applications should run faster and complete sooner
• TCP/IP transfers over Sun Fire Link	• Data can be exchanged between two nodes much faster than using alternatives such as Gigabit Ethernet

Reliability, Availability and Serviceability (RAS)

Reliability

The reliability features for Sun Fire Link include the following:

- The Sun Fire Link assembly includes a bank of memory that is protected by parity and error correction codes (ECC).

- Built-in self tests (BIST) that are run when powering up Sun Fire Link. These tests are for the Sun Fire Link ASIC, the switch ASIC and the Sun Fire Link memory. After BIST is complete POST (Power On Self Test) takes over and completes the diagnostic testing prior to Solaris loading.
- A further set of diagnostics is SunVTS™ software that can be used while the Solaris OS is running. SunVTS software supports Sun Fire Link.

Availability

The availability features for Sun Fire Link include the following:

- Redundant optical links protect against failure of the cable or the optical interface. Sun's cluster software is designed to recover from this situation without manual intervention.
- Use of redundant Sun Fire Link assemblies allows the cluster interconnect to continue to function (albeit at one half the bandwidth) should an assembly fail.
- The Sun Fire Link switches have fault tolerant line cords, power supplies and cooling trays.
- Sun Fire Link switches are used in pairs to protect against the failure of the cross-bar ASIC module or the switch system controller (SSC). If either of these components fail, Sun's cluster software should recover

Serviceability

The serviceability features for Sun Fire Link include the following:

- The optical interfaces and cables can be hot swapped.
- The optical cables are both mechanically keyed and color coded. In addition, both ends of each cable have a cable serial number marked on them for easy cable identification in the machine room environment.
- A failed Sun Fire Link assembly can be replaced online using dynamic reconfiguration (DR).
- The Sun Fire Link assemblies support adapter cards that can be hot swapped. These are cPCI cards for the Sun Fire 6800 server and PCI cards in hot swap cassettes for the Sun Fire 12K/15K servers.

Installation Data

The Sun Fire Link assemblies replace the regular I/O assemblies in the Sun Fire 6800 and Sun Fire 12K/15K servers. The hardware installation procedures are similar.

The Sun Fire Link switch is designed to be rack mounted in either a Sun cabinet or the customer's cabinet. Physical features of the switch are as follows:

	U.S.	Metric
Height	7 rack units. 12.25 inches	27.6 cm
Width	Designed for a 19" rack	
Depth	13 inches. The switch is recessed 1 inch behind the cabinet rails	29.25 cm. 2.54 cm recess
Weight, fully configured	29.1 LB (for one Sun Fire Link switch)	13.2 kg
Voltage	100 - 240 volts	
Frequency	47 - 63 Hz	
Current draw	2.25 amps	
Temperature, operating	41° F to 95° F	5° C to 35° C
Temperature, non-operating	-40° F to 149° F	-40° C to 65° C
Relative humidity, operating	20% RH to 80% RH, non-condensing, 80° F max. wet-bulb	27° C max. wet-bulb
Relative humidity, non-operating	93% RH, non-condensing, 100° F max. wet-bulb	38° C max. wet-bulb

Regulations

The Sun Fire Link switch meets or exceeds the following requirements.

Category	Specifications
Safety	UL1950, CSA950 and EN60950 with Worldwide CB Scheme Certificate
Emissions	FCC Class A, IES Class A, VCCI Class A and EN55022 Class A EN61000-3-3
Immunity	EN55024 & Sun 990-1151
Harmonics	EN61000-3-2

Requirements and Configuration

How to configure Sun Fire Link for the Sun Fire 6800 server

- Sun Fire Link assemblies are used in pairs for redundancy and performance. The maximum per F6800 server is two assemblies (one pair) supporting one domain. There may be two additional domains in the system. These Sun Fire Link assemblies will replace two of the regular I/O assemblies in the F6800 server.
- Each assembly requires two optical cables, for a total of four per F6800 server.
- Each Sun Fire Link assembly has two slots for cPCI adapters. There are no fundamental restrictions on what cPCI adapters can be used but the bandwidth consumed is subtracted from that available for Sun Fire Link. Therefore slower speed adapters are preferred if a choice is available. The list of adapter cards qualified for use in the Sun Fire Link assemblies is as follows:

System	***Function***	***Part number***
F6800 (cPCI cards)	Ultra SCSI / Ethernet	1232A
	Dual FC-AL 100 MBps	6748A
	Dual Ultra SCSI	6749A
	Quad Fast Ethernet	1234A
	SunATM 155 Fiber	1266A
	SunATM 622 Fiber	1268A
	Gigabit Ethernet	1261A

- The minimum CPU level is 900 MHz UltraSPARC III cu.

How to configure Sun Fire Link for the Sun Fire 12K or 15K servers

- The maximum number of domains per F15K server that can have a Sun Fire Link interface is four. For the F12K server, it is two.
- Each domain with a Sun Fire Link interface will have two Sun Fire Link assemblies. As there are two optical cables per assembly, there will be four per Sun Fire Link domain.
- Each F12K/15K Sun Fire Link assembly has two slot positions for PCI adapters in hot plug cassettes. As above, there are no fundamental restrictions on what PCI adapters can be used but the bandwidth consumed is subtracted from that available for Sun Fire Link. Therefore slower speed adapters are preferred if a choice is available. (Qualification of adapter cards is expected to complete by September 2003).
- A Sun Fire Link cluster within-the-box may be configured from a single F12K or 15K server.

How to configure the Sun Fire Link switch

- The Sun Fire Link 8 port switch is required when clustering more than three nodes. Some customers will prefer to use the switch even for two or three node clustering to allow the cluster to be more easily expanded in the future.
- The switch includes rack mounting hardware and consumes 7U of rack space plus 1U for cable access. It may be mounted in the Sun Fire expansion cabinet, the Sun StorEdge™ cabinet or a customer's cabinet. Cooling is from lower front to upper rear. For redundancy, switches are always used in pairs. The switch has redundant power supplies, cooling trays and AC power cords.
- Each switch has eight ports and each port that is to be used requires a Sun Fire Link optical module. The number of optical modules to order will equal twice the number of Sun Fire Link assemblies forming the cluster.

Sun Fire Link cables

These are supplied in standard lengths of 5, 12 and 20 meters (respectively 16, 39 and 65 feet). Lengths used should be the minimum that will satisfy the system site layout. A cluster will require the following quantity of cables:

- 2 node direct connect - 4 cables
- 3 node direct connect - 6 cables
- Connections using the switch - the number of cables equals twice the number of Sun Fire Link assemblies used in the cluster.

Sun Fire Link software and documentation

Sun Fire Link software is supplied on a CD. One copy of the software may be ordered for every server that uses Sun Fire Link. This CD will contain a soft copy of all Sun Fire Link documentation plus Sun Fire Link software and firmware not yet available with the Solaris OS or Sun MC.

Sun Fire Link hard copy hardware and software documentation is also available.

Ordering Information

The table below shows the Sun Fire Link part numbers to support the configuration information given above.

Part number	***Description***
(C) 4121A	Sun Fire Link assembly for the Sun Fire 12K or 15K servers. Includes one Sun Fire Link ASIC, 20 MB of Sun Fire Link memory, one PCI bridge with two PCI adapter slots. Includes two optical links on two optical cards (4175As).
(X) 4141A	Sun Fire Link assembly for the Sun Fire 6800 server. Includes one Sun Fire Link ASIC, 20 MB of Sun Fire Link memory, one PCI bridge with two cPCI adapter slots. Includes two optical links on two optical cards (4175As).
(X) 4170A	Sun Fire Link switch. Eight ports. Includes rack mounting hardware, power and cooling and System Controller. Excludes optical modules and cables. May be mounted in the Sun StorEdge cabinet.
(X) 4175A	Sun Fire Link optical module. One module mounted on cPCI card. For the Sun Fire Link switch.
X4180A	Sun Fire Link optical cable, 5m
X4181A	Sun Fire Link optical cable, 12m
X4182A	Sun Fire Link optical cable, 20m
LNKSS-101-E999	Sun Fire Link software on CD - version 1.0
LNKSS-111-E999	Sun Fire Link software on CD - version 1.1 (available June 2003)
X4190A	Sun Fire Link documentation set

When a F6800 or 12K/15K system order includes Sun Fire Link, the preferred ordering mechanism is to use Sun's ATO (assemble to order) model. With ATO, part numbers without the initial "X" or “C” are used. The Sun Fire Link assemblies are configured into the servers in the factory and tested using loop-back.

"X option" part numbers are used when Sun Fire Link is to be added to a system already installed at a customer's site. The C4121A is a “configure to order” part. If PCI adapter cards are ordered with a C4121A, the assembly will be tested with the cards and shipped to the customer in one package.

Upgrades

Upgrade Paths

Sun Fire Link is a cluster interconnect for the F6800 and the F12K/15K servers. For the F6800 server this requires the installation of two Sun Fire Link assemblies. For the F12K/15K servers, this would be up to eight Sun Fire Link assemblies.

The F6800 server is sold fully configured with four I/O assemblies - either PCI (4050A) or cPCI (4030A) in any mix. To use Sun Fire Link, two of these I/O assemblies must be replaced with Sun Fire Link assemblies. For customers that do not wish to keep I/O assemblies that would be removed in order to install the Sun Fire Link assemblies, there is an upgrade program to provide a trade-in allowance on their I/O assemblies.

The F12K and F15K servers are configured with at least one hot swap I/O assembly out of a possible 9 (F12K) or 18 (F15K). Therefore in most cases there will be I/O space to install Sun Fire Link assemblies. However, to gain flexibility, these customers are also eligible to participate in the upgrade program.

The following table summarizes the upgrade paths for F6800, Sun Fire 12K and 15K server customers. Note that there is no upgrade allowance to trade-in PCI adapter cards for cPCI adapter cards.

Upgrade from	***Upgrade to***	***Return***	***Upgrade Allowance***
PCI I/O assembly for the Sun Fire 6800 (4050A) Has 8 PCI slots	Sun Fire Link assembly for the Sun Fire 6800 (X)4141A Has a dual Sun Fire Link interface and 2 cPCI slots	4050A	ALW-25-S-IO-SER
A cPCI I/O assembly for the Sun Fire 6800 (4030A) Has 4 cPCI slots	Sun Fire Link assembly for the Sun Fire 6800 (X)4141A Has a dual Sun Fire Link interface and 2 PCI slots	4030A	ALW-25-S-IO-SER
Hot swap PCI I/O assembly for the Sun Fire 12K and 15K. (4575A) Has 4 hot swap PCI slots	Sun Fire Link assembly for the Sun Fire 12K/15K (C)4121A Has a dual Sun Fire Link interface and two hot swap PCI slots	4575A	ALW-25-S-IO-SER

Service and Support

Service, support and warranty

Sun Fire Link, a cluster interconnect, assumes the same service, support and warranty policies of the underlying server into which it is installed. These service factors are common for the Sun Fire 6800, 12K and 15K servers and so the total cluster will have a uniform support level.

Installation of Sun Fire Link assemblies into these servers is included in the price of the assemblies. This applies to assemblies sold at the time of the server sale or installed later at the customer's site.

Installation of the Sun Fire Link switch into Sun-provided cabinets or the customer's cabinets is included in the price of the switch. Included also is under-floor installation of cables from the switch to the Sun Fire Link assemblies.

Installation of Sun Fire Link includes testing of the cluster network using SunVTS software to help ensure that all hardware components are functioning.

Education

Two existing courses are expected to be updated to include Sun Fire Link. They are:

- ES333. Sun Cluster 3.0 administration
- ES320. High Performance Computing administration

Professional Services

The Sun Professional services organization offers Sun Cluster Implementation Services. These are designed to deliver the level of implementation that cluster customers require. This service, which can be customized from basic to complex, is available for customers that have Sun Fire Link as their cluster interconnect.

Glossary

Cluster	A way of aggregating multiple servers to achieve greater power than that achieved by a single server. Clustering is also used to provide fail-over between servers. (Usually two).
DLPI	Data Link Provider Interface. A lower level standardized communications interface.
Sun Fire Link Fabric Manager	The Sun Fire Link administration software to allow a system administrator to manage the Sun Fire Link-based cluster using a command line interface.
Sun Fireplane	The internal system interconnect of Sun's mid range, and above, Sun Fire servers. This is the cross bar between CPUs, memory and I/O.
HPC ClusterTools	Sun's software for technical and scientific applications. A development environment as well as HPC clustering software.
Remote Shared Memory	A mechanism for reducing the latency of Sun Fire Link data transfers between two servers. Shortened to RSM.
SNMP	Simple Network Management Protocol. An industry standard way of sending messages - usually administration messages - between servers.
Sun Cluster	Sun's clustering software for commercial applications
TCP/IP	Industry-standard communications protocols for data traffic.
U	A measure of rack space consumed – where 1U is equivalent to 1.75 inches.

Materials Abstract

All materials will be available on SunWIN except where noted otherwise.

Collateral	Description	Purpose	Distribution	Token # or COMAC Order #
Product Literature				
– *Sun Fire Link: Just the Facts*	Sun Fire Link Just the Facts	Training Sales Tool	SunWIN, Reseller Web	Token 360070
– *Sun Fire Link Customer Presentation*	Sun Fire Link Customer Presentation with Notes	Sales Tool	SunWIN, Reseller Web	Token 361831
– *Intro*	Intro for Sun Fire Link	Sales Tool	SunWIN, Reseller Web	Token 362357
References				
– *Data Sheet*	Sun Fire Link datasheet	Sales Tool	SunWIN, Reseller Web, COMAC	COMAC 1775-0 Token 363768
– *White Paper*	Sun Fire Link – a High Speed Interconnect for SunPlex systems	Sales Tool	SunWIN, Reseller Web COMAC	Token 360770
– *White Paper*	Sun Fire Supercluster Solutions (targeted to HPTC markets)	Sales Tool	SunWIN, Reseller Web	Token 374443
External Web Sites – *Sun Web Site* –	`http://www.sun.com/servers/cluster_interconnects`			
Internal Web Sites – –				

Sun Proprietary—Confidential: Internal Use Only

More on positioning

The clustered server purchasing decision is unlikely to be won or lost on the capability of our cluster interconnect as there are other factors that are so much more important – the Solaris OS and our cluster software, the price/performance of the servers, service considerations, etc. However it is interesting to review the offerings from our system competitors to know how Sun Fire Link compares to their cluster interconnects.

A customer who has decided to purchase a cluster from Sun has a choice of interconnects to consider as shown in Table 1. We offer our customers choices that allow them to match the interconnect to the demands of their applications. The following should be considered:

- Simple Sun Cluster failover is served well by an Ethernet interconnect.
- HPC applications that run in clustered mode stress the interconnect and will usually benefit from Sun Fire Link – especially for the F12K and F15K servers.
- Sun Cluster applications intended to scale horizontally can stress the interconnect - especially for a cluster of four or more nodes. But this is very application dependent and difficult to calculate ahead of time. Sun Fire Link should not be over sold when a slower cluster interconnect may be quite satisfactory.

Competitive Information

Sun Fire Link's competition can be presented two ways. A Sun server customer has a choice of cluster interconnects. So Sun Fire Link will have to compete against our own internal solutions, SCI and Gigabit Ethernet, as well as third party Solaris OS interconnect products such as Myrinet and Quadrics. Sun Fire Link will also compete against interconnects offered by our system competitors (IBM, HP and the former Compaq).

Cluster interconnects (Solaris solutions)

SCI from Dolphin

Sun has been using SCI as our high end interconnect since 1997 - first an SBus solution for Sun Enterprise servers and now PCI. With the advent of Sun Fire Link, this now becomes a mid-range interconnect. We are using the Dolphin D320 PCI adapter and their 4 port switch. Data rate is 200 MBps. Latency is 70 microseconds using TCP/IP. Latency, using the same RSM as Sun Fire Link, is under 10 microseconds. Adapter price (Sun's price) is $4,500. The

switch price is $17,500. SCI is only available for use with Sun Cluster – not HPC ClusterTools. SCI is an excellent mid performance, mid price cluster interconnect.

Gigabit Ethernet

Gigabit Ethernet, as a cluster interconnect, was made available with Sun Cluster 2.2 in April 2000. Its advantage is that it uses commodity components and wiring. So a customer could re-deploy his cluster interconnect and use it as a general purpose network. Transfer rate is 100 MBps and the latency is about 100 microseconds. Gigabit Ethernet is not hardware-capable to allow RSM operations so the latency is much greater than that of SCI or Myrinet. Adapter price (from Sun) is $2095. We do not sell Gigabit Ethernet switches but a typical price is $2,400 for 8 ports.

Myricom's Myrinet.

Myricom is a California company established in 1994. The hardware is a PCI board, a 16 port switch and copper links. The performance is 140 MBps one way, with 16 microseconds of MPI latency using their own OS bypass protocol called GM. Myrinet is a Sun-promoted cluster interconnect for use with HPC ClusterTools. It is sold directly to our customers by Myricom and they provide support. Pricing is commodity: $1400 for each network adapter and $6000 for a 16 port switch. Myrinet is not available for use with Sun Cluster software.

Quadrics QsNet.

Quadrics is a UK company with Italian ownership. It was established in 1996. The network is called QsNet and it uses a PCI adapter. A 16 port switch is available. The links are copper and delivered bandwidth is about 200 MBps (in one direction). MPI latency is 5 microseconds. The product is available for Solaris and it is used by HP/Compaq for high end clustering for their AlphaServers. Compared to Sun Fire Link, QsNet is low performing but also lower priced.

Cluster interconnects (our system competitors)

HP/Compaq's Memory Channel (Version II)

This is Reflective Memory technology that Compaq (originally Digital) purchased from Encore. It is available for AlphaServers – which are expected to be phased out by HP. Memory Channel is quite similar to Sun Fire Link with hardware transfers of memory pages from one system to the other. The host interface is a PCI adapter with a data transfer rate of 90 MB/sec. Direct connection is possible for two nodes - with an 8 port switch used beyond that. Switches and adapters are duplicated for availability. Link cables are copper up to 12 meters or fiber up to 3 km. The switch bandwidth is 800 MB/sec. End to end latency (over copper) is 6 microseconds. The adapter's list price is $2990 and a 4 port switch is $9793. Memory Channel has the elegance of Sun Fire Link but it is implemented with mid-90s hardware components.

HP's HyperFabric/2

This is a PCI-based solution for HP's Enterprise servers. Supports direct connect to 10 nodes (with multiple adapters per node). There is a 16 port switch that can be cascaded to increase connectivity to 64 nodes. The adapter bandwidth per port is 320 MB/sec one way which

matches that of the switch. The software end-to-end latency is 22 microseconds using Hyper Messaging Protocol - their OS bypass mechanism. Cables are copper to 18 meters with 200 meters distances possible with fiber. The adapter price is $4590. An 8 port fiber switch is $30,600. This is a mid price / mid performance cluster interconnect which competes against our SCI product.

IBM SP Switch2

IBM has been clustering RS/6000s since the mid 1990s and thereby creating the SP MPP systems. Each node has an adapter and there is a central SP switch. The switching hardware has been upgraded in speed over the years and IBM now has the SP Switch2. Aggregate data rate of the switch is 1 GB/sec bi-directional. Data rate of the adapters for a p690 server is 150 MBps with one byte wide transfers over copper cables. The switch has 16 ports and switches can be cascaded to create a cluster of 128 nodes. The switch hardware latency is 1 microsecond. MPI latency is 18.5 microseconds. The SP switch adapter (for the p690) has a list price of $10,000. The switch lists for $56,000. This IBM cluster interconnect is not in the same performance category as Sun Fire Link but is high priced. Expected from IBM, mid calendar 2003 (and shown at the SC2002 conference), is a new interconnect code named Federation. Its specifications are expected to be comparable to those of Sun Fire Link.

Summary of competition

Product	***Capability***	***Link data rate (Hardware) MB/sec***	***Latency (MPI) Microseconds OS bypass?***
Sun Fire Link	Direct attach to Fireplane. 8 port switch	1200+1200 dual links	<4 RSM OS bypass
SCI	PCI adapter, 4 port switch	200	<10 RSM OS bypass
Gigabit Ethernet	PCI adapter, commodity switch, high node count	100	100 No OS bypass
Myrinet	PCI adapter, 16 port switch	140	16 OS bypass
QsNet	PCI adapter, 16 port switch	200	5 OS bypass
HP/Compaq Memory Channel	PCI adapter, 8 port switch	90	6 Mem. Channel s/w OS bypass
HP HyperFabric2	PCI adapter, 16 port cascadable switch	320	22 HMP OS bypass
IBM SP Switch2 (used with pSeries 690)	PCI adapter, 16 port cascadable switch	150	18.5 OS bypass ?

Future/Roadmap and Caveats

Roadmap

There will be no major enhancements to Sun Fire Link following GA. The product will remain as an eight node cluster interconnect for the upper end of the Sun Fire series. We expect to make one configuration change to allow optional use of single Sun Fire Link assemblies per interconnect (instead of dual). This will make Sun Fire Link more affordable at the expense of availability.

Adapter cards in Sun Fire Link assemblies

Qualification of adapter cards for the Sun Fire Link assemblies for the F12K and F15K servers is expected to complete in Q3 CY2003. The goal is that all cards available for the hsPCI assemblies be also available for the Sun Fire Link assemblies.