SMP Interconnect @ Sun

Alan Charlesworth
Sun Microsystems, Inc.
Alan.Charlesworth@sun.com

Abstract
System interconnect is the heart and soul of a computer. Processor performance doubles about every eighteen months from the combined effects of faster clocks, better compilers, multi-threaded CPUs, and multiple CPUs on a die. The system interconnect strives to keep up.

This paper describes the evolution of four generations of shared-memory interconnect at Sun Microsystems. It then details the Sun Fireplane™ interconnect that is the foundation of today’s Sun Fire™ servers that use the UltraSPARC™ III processor.

1 Introduction
Section 1 provides an introduction to system-interconnect issues.

1.1 Types of interconnect
In Figure 1, we categorize two dimensions of multiprocessor interconnect scaling:

- **Shared-memory multiprocessors.** This dimension is called *vertical* scaling. The processors and I/O interfaces have load/store access to a large, cache-coherent memory space. The interconnect transfers cache blocks (~64-bytes) between processors and memory.

  It is common usage to label these as SMP systems (meaning either Shared-Memory Processing, or Symmetric Multi-Processing). For large SMP systems, the interconnect is usually built from a proprietary technology, since industry-standard solutions do not scale large enough. The SMP interconnect protocol is built into the interface of the processor chip. One instance of an operating system, such as Solaris™, controls the SMP.

- **Cluster (blade) multiprocessors.** This dimension is called *horizontal* scaling. Multiple computers are connected together by a network switch that routes network packets between the computers. The cluster interconnect is usually based on open-standard technology. These systems are more cost- and power efficient, because their interconnect is simpler than in large SMPs, and they are built in higher volumes. Each computer runs its own instance of the operating system, and the whole ensemble is orchestrated by cluster-management software.

Figure 1. Two dimensions of interconnect scaling
The two dimensions of interconnect scaling often work together in an overall solution. For example, a web-serving frontend may be implemented with horizontal scaling, and the backend database be implemented with vertical scaling. SMP interconnect is the subject of the rest of this paper.

**Figure 2. Server revenue by CPU capacity**
1.2 SMP system-revenue distribution

According to IDC estimates [1], the total worldwide server market revenue was $61 billion in 2000, and $50 billion in 2001. About 88% of this revenue comes from vertically-scaled SMP systems that support two or more CPUs. The remaining 12% of server revenue is from unprocessors that cannot hold multiple CPUs, or from massively parallel processors (MPP) that do not shared memory.

Figure 2 shows the distribution of server revenue as a function of system CPU capacity over the six years of available IDC revenue data. The fraction of revenue coming from larger SMP systems has increased noticeably since 1996. Systems that can hold more than eight CPUs now account for 26% of the total server revenue, 37% of Unix server revenue, and 54% of Sun’s server revenue.

The fastest revenue growth has been in the largest SMP systems, with have room for more than 32 CPUs. This revenue is up 5-fold from 1996 to an annual sales of over $3 billion. Sun accounts for more than half of this size’s total revenue.


1.3 Introduction to cache coherency

A shared-memory system interconnect has to send memory addresses, maintain cache coherency, and transfer cache-line sized-blocks of data. Each processor has a cache in which to keep frequently accessed data blocks, so that they are quickly available. A typical cache block size is 32, 64, or 128 bytes. These blocks are also called cache lines.

When a processor cannot find a needed data item in its cache (called a cache miss), it asks the interconnect to supply that block. Requests are satisfied from memory unless some system device (a processor or an I/O controller) currently has a modified copy of that block.

To get permission to modify a block, a processor has to become its owner. All other devices then invalidate any copies they have, and the old owner supplies the data. Henceforth, when other processors request to share a read-only copy of the data, the owning device, not memory, will supply it. Memory becomes the owner again when the owning processor needs to make room in its cache for new data, and it victimizes the cache block by writing it back to memory.

This process of finding the up-to-date copy of a cache block is called cache coherency. System designers use two main methods to keep each processor’s view of memory consistent.

1. In broadcast (snoopy) coherency, all addresses are sent to all system devices. Each device examines (snoops) the state of the requested cache line in its local caches. The system determines the overall snoop result a few cycles later. Broadcast coherency provides the lowest possible latency. Sun has pushed the rate of broadcast coherency, since this is the method used in Sun’s mid-sized servers.

2. In directory (point-to-point) coherency, addresses are sent only to those system devices that are known to be
interested in that cache block. The hardware keeps a directory in memory or in special RAM to keep track of which system devices share or own each cache block. Since all addresses are not sent everywhere, the total system bandwidth can be much higher than with broadcast coherency. However, the latency is longer and more variable, due to the more complicated protocol.

For more on cache coherency, see chapters six and eight of [7].

2 Sun SMP generations

Section 2 provides an overview of Sun’s four generations of SMP systems.

Approximately every four years, Sun has introduced an improved system-interconnect architecture to go with each new SPARC processor core. The Sun Fireplane interconnect described later in this paper is Sun’s fourth generation of shared-memory interconnect.

So far, each of these SMP interconnects have been used across Sun’s entire system-size spectrum, from two CPUs on up. Table 1 and Figure 3 summarize Sun’s four SMP generations. The following four subsections briefly describe each generation.

2.1 Mbus (1990)

Sun’s first multiprocessor interconnect was a 40 MHz, 32-bit wide bus, which could support up to four processors (see Figure 3). Both addresses and data were transferred on the same wires. The peak MBus bandwidth was around 0.08 gigabytes per second (GBps). Sun was among the first system vendors to implement cache-coherency logic onto a single Application Specific Integrated Circuit (ASIC).
2.2 XDBus (1993)
For its second multiprocessor generation, Sun expanded the interconnect to have up to four buses (see Figure 3). These buses ran at a faster 55 MHz clock. The protocol was improved to do split transactions, meaning that multiple requests could be outstanding on the bus, rather than having each tie up the interconnect while waiting for completion. The cache line width was doubled from 32 to 64 bytes. These improvements increased the peak bandwidth of the largest XDBus system by 16-fold to 1.28 GBps. The maximum system size increased from four to 64 processors.

2.3 Ultra Port Architecture (1996)
For its third generation, Sun split the interconnect into separate addresses and data paths (see Figure 3). This allowed Sun to use a topology that was optimized for each type of traffic, and permitted both the full address rate and the full data rate to be sustained at the same time.

These improvements increased the sustainable address rate from one address every 11 clocks to one address every two clocks. The system clock rate was doubled to 100 MHz. The bandwidth of a single-bus, mid-sized system (such as the Sun Enterprise 4500) was thus increased by ten-fold over the previous generation to 3.2 GBps.

Table 1. Sun system-interconnect generations

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor</td>
<td>Cypress SPARC</td>
<td>SuperSPARC</td>
<td>UltraSPARC-I/II</td>
<td>UltraSPARC-III</td>
</tr>
<tr>
<td>Processor clock</td>
<td>40 MHz</td>
<td>40–60 MHz</td>
<td>167–400 MHz</td>
<td>750–1050 MHz</td>
</tr>
<tr>
<td>Interconnect clock</td>
<td>40 MHz</td>
<td>40–55 MHz</td>
<td>83–100 MHz</td>
<td>150 MHz</td>
</tr>
<tr>
<td>System sizes</td>
<td>1–4 CPUs</td>
<td>1–64 CPUs</td>
<td>1–64 CPUs</td>
<td>1–106 CPUs</td>
</tr>
<tr>
<td>Cache-coherency mechanism</td>
<td>Broadcast</td>
<td></td>
<td>Broadcast + point-to-point</td>
<td></td>
</tr>
<tr>
<td>Packet protocol</td>
<td>Circuit switched</td>
<td></td>
<td>Packet switched</td>
<td></td>
</tr>
<tr>
<td>Address and data</td>
<td>Multiplexed on same wires</td>
<td>Separate wires</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cache coherency line size</td>
<td>32 bytes</td>
<td></td>
<td>64 bytes</td>
<td></td>
</tr>
<tr>
<td>System clocks per address</td>
<td>16</td>
<td>11</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Address rate per address bus</td>
<td>2.5 million/sec</td>
<td>4.5 million/sec</td>
<td>50 million/sec</td>
<td>150 million/sec</td>
</tr>
<tr>
<td>Data bandwidth per address bus</td>
<td>0.08 GBps</td>
<td>0.29 GBps</td>
<td>3.2 GBps</td>
<td>9.6 GBps</td>
</tr>
<tr>
<td>Maximum number of address buses</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>Maximum address-limited data bandwidth</td>
<td>0.08 GBps</td>
<td>1.28 GBps</td>
<td>12.8 GBps</td>
<td>172 GBps</td>
</tr>
<tr>
<td>Datapath width</td>
<td>8 bytes</td>
<td></td>
<td>16 bytes</td>
<td>32 bytes</td>
</tr>
<tr>
<td>Electrical implementation</td>
<td>Bused</td>
<td>Bused</td>
<td>Bused &amp; switched</td>
<td>Switched</td>
</tr>
</tbody>
</table>

Note: 1 GBps (gigabyte per second) = 10⁹ bytes per second
The interconnect of the largest system (the Sun Enterprise™ 10000) was implemented with point-to-point electrical signals, rather than bussed signals. This improved fault-isolation, since one chip could not lock-up a bus that was shared by many processors. Dynamic System Domains were introduced to allow subdivision of the system into multiple fault-isolated computers (horizontal scaling within one box). Each domain runs its own copy of the Solaris operating environment.

The data interconnect was changed from a bus to a multi-level crossbar. These improvements increased the peak bandwidth of a large system by 10-fold over the previous generation to 12.8 GBps. The maximum processor count stayed at 64 CPUs. For more on the Sun Enterprise 10000, see [11].

2.4 Sun Fireplane (2000)
For its fourth generation, Sun has put the snooping logic and cache tags inside the system devices, so they can snoop a new address every system clock. The system clock has been increased to 150 MHz. These improvements tripled the data bandwidth of a single-address bus, mid-sized system to 9.6 GBps (see Figure 3).

All systems have been implemented with point-to-point or three-drop electrical signals. All systems use a data crossbar rather than a data bus. Dynamic System Domains have been extended down to mid-range servers. There are more common components between mid and large sized systems.

The protocol is improved so that address and data packets need only to traverse the shortest path between source and destination. In the previous generation, packets always had to go all the way to the outermost level of the interconnect, even if they were going to a destination on the same board. This enhancement lowers the latency for “close” transfers.

The coherency protocol was extended to do directory coherency between multiple snooping coherency domains — allowing the interconnect bandwidth to increase beyond the limit imposed by broadcast coherency.

The global address interconnect of the largest implementation, the Sun Fire 15K, was implemented using an address crossbar and a response crossbar to connect 18 snooping coherency domains. This increased the peak interconnect bandwidth over the previous generation by 13-fold to 172 GBps, and increased the centerplane bandwidth by 3.3x to 43 GBps. The maximum processor count was increased to 106 CPUs.

3 Sun generation scaling
Section 3 shows how Sun microprocessor and system performance have scaled over Sun’s four SMP generations.

3.1 SPARC performance scaling
Figure 4 shows processor performance scaling over the span of two generations of the SPEC CPU benchmark: CPU95 [12], and CPU2000 [13]. Also shown are CPU clock rates, peak integer operation rates, and peak floating-point operation rates.

SPARC clockrate has scaled at approximately 1.36x per year. The peak integer operation rate has scaled the same as the clock rate, since all these SPARC processors can do a peak of two integer operations per clock. The peak floating point operation rate has tracked clock scaling, except for a factor of two jump with the UltraSPARC-I. SPEC benchmark performance has scaled faster than clock rate, due primarily to compiler improvements.
Figure 4. SPARC performance scaling

Figure 5. System interconnect scaling
Figure 6. System integer performance scaling

Figure 7. System floating-point performance scaling
3.2 System interconnect scaling

Figure 5 shows system interconnect scaling across all four Sun SMP generations, for both medium-sized servers (blue diamonds), and large servers (red x’s). The number in parenthesis for each model is the number of processors.

Large-server bandwidth has scaled at 1.57x per year, and mid-sized-server bandwidth at 1.45x per year. In each new generation, the mid-sized servers come to market before the large servers, and so usually leap-frog the large server of the previous generation.

The latencies are pin-to-pin latencies, which count from the address request going out on a CPU’s address pins through the first piece of the requested data coming in on the CPU’s data pins.

Latency came down dramatically in the UPA generation, and then somewhat more in the Fireplane generation. The latency gap between medium and large servers was relatively wide in the UPA generation, and less so in the Fireplane generation.

The Fireplane interconnect is optimized to use the least amount of interconnect necessary to satisfy a request. Previous generations always used the entire interconnect, even if the request could have stayed inside a particular board. Thus, on the Fireplane generation, the latency for requests of data located on the same board (labeled local on the chart) is lower than for requests to other boards (labeled remote on the chart).

On the large Fireplane servers, this optimization allows for a higher peak bandwidth when most of the accesses are local to a board.

3.3 System performance scaling

Figure 6 shows peak integer performance scaling, and Figure 7 shows peak floating-point performance scaling over Sun’s system generations. The number in parenthesis for each model are the maximum number of processors that can be configured together with a full complement of I/O slots.

System CPU count has remained pretty constant over the last three generations, so performance increase has been mostly through clock scaling. Bandwidth per operation has remained relatively constant over the system generations, meaning that Sun’s SMP interconnect bandwidth has kept up with SPARC performance scaling.

In the future, CPU multi-threading and multiple CPU cores on a die will both multiply the CPU count, increasing the pressure on interconnect improvements.

4 Sun Fireplane implementation

Section 4 introduces the implementation of the Sun Fireplane interconnect, and describes the system devices. The Fireplane interconnect protocol is used in all UltraSPARC-III based systems, which range from 2 to 106 processors.

4.1 Number of interconnect levels

The Fireplane interconnect is implemented with up to four levels of interconnect logic. The number of interconnect levels required is approximated by $\log_4 n$, where $n$ is the total number of system devices (CPUs, memory units, and I/O controllers). The total amount of interconnect logic needed for $n$ components is approximately $n \log_4 n$. The Fireplane-based servers can be divided into four categories by the number of interconnect levels required:
1. **Small server.** Four system devices (two CPUs, one memory unit, and one I/O controller) require only one level of data switch interconnect.

2. **Workgroup server.** 18 system devices (eight CPUs, eight memory units, and two I/O controllers) require two level of interconnect.

3. **Mid-size server.** 56 system devices (24 CPUs, 24 memory units, and eight I/O controllers) require three levels of interconnect.

4. **Large server.** 180 system devices (72 CPUs, 72 memory units, and 36 I/O controllers) require four levels of interconnect.

The increased number of levels needed for larger systems explains why they have longer average latencies than smaller systems.

4.2 **Address interconnect**

Levels one and two of the Fireplane address interconnect perform snoopy coherency using a tree-structure of Address Repeater ASICs. A new address can be broadcast at a peak rate of one every system clock (6.67 ns). All system devices get a particular address at the same time — six clocks after it was originated.

The memory access is started as soon as the address is received by the CPU that controls the requested location. About half the memory access time is overlapped with the nine clocks required to compute the snoop result and broadcast it to the devices. If a system device owns the requested block, then it supplies the data to the requester, and the data from memory is not sent.

4.3 **Data interconnect**

The Fireplane data interconnect is implemented from a hierarchy of crossbar switches. Board-level connections are 32 bytes wide to CPU boards, and 16-bytes wide to I/O boards. All datapaths are bidirectional, and can move data in one direction or the other every system clock.

4.4 **Active system devices**

Active system devices originate transactions. Currently there are two such system devices: the UltraSPARC-III processor, and the PCI I/O controller. These are the building blocks of all Fireplane-based systems. Both CPUs and I/O controllers are interconnected in pairs, as shown in Figure 8.

4.5 **Processor**

The UltraSPARC-III processor currently has clock rates of 750, 900, and 1050 MHz. See Figure 9 for its block diagram.
4.5.1 Instruction execution. The instruction issue unit can initiate up to four in-order instructions per clock from a 32 KB, four-way associative instruction cache. Instructions may complete out of order. The integer unit can perform up to four operations per clock: two ALU ops, one load/store, and one branch. The floating-point unit can initiate a multiply-type operation and an add-type operation every clock.

4.5.2 On-chip caches. The data cache unit includes a 64 KB, four-way set associative, data cache, plus a 2 KB write cache for merging stores, and a 2 KB prefetch cache for buffering data from software.
prefetch instructions and automatic hardware prefetching from the external cache.

4.5.3 External cache. An 8 MB, two-way associative second level cache is provided using high-speed external SRAMs. The tags for the second-level cache are located inside the processor chip to make possible high-speed snooping.

4.5.4 Memory controller. To reduce chip count and latency, and to provide scalable memory bandwidth, each processor has an on-chip DRAM controller. This provides addresses and control signals for two groups of four dual-banked Synchronous Dynamic Random Access Memory Dual Inline Memory Modules (SDRAM DIMMs). Currently, the maximum DIMM size is 1 GB, so each CPU can configure up to 8 GB of memory. A memory unit has a peak bandwidth of 1.9 GBps.

4.5.5 System interface. The processor has a 150 MHz Fireplane address interface. It can snoop a new address every system clock. The processor has a 2.4 GBps data interface, which goes to the Dual CPU Data Switch. This switch is implemented from 8 bit-sliced ASICs, and connects together two processors and two memory units, and has a port to higher levels of the data interconnect. It is shown in the lower part of Figure 9.

For more information on the Ultra-SPARC-III, see [14].

4.6 PCI Controller

The PCI Controller implements two 64-bit-wide PCI buses, one running at 33 MHz and one at 66 MHz. The 33 MHz PCI bus has from one to three PCI slots, depending upon the particular server. It has a peak bandwidth of about 0.25 GBps. The 66 MHz PCI bus has one PCI slot, and has a peak bandwidth of about 0.5 GBps. The PCI Controller has an 8-byte wide (1.2 GBps peak) connection to the Fireplane data interconnect. On all but the smallest systems, the PCI Controllers are configured in pairs (see bottom of Figure 8).

5 Small server interconnect

Section 5 through Section 8 describe the Fireplane interconnect implementation for each of the four system-size ranges.

A small server like the Sun Fire 280R needs only a minimal amount of interconnect for four devices (two processors, one memory unit, and one I/O controller). See Figure 10. The two CPUs and one I/O controller connect directly together via a Fireplane address bus. A four-port data switch (implemented from six bit-sliced ASICs) connects together the four system devices.

6 Workgroup server interconnect

A workgroup server like the Sun Fire V880 needs one level of Fireplane interconnect for up to 18 devices (eight processors, eight memory units, and two I/O controllers). See Figure 10.

6.1 Dual CPU board

The Dual-CPU board used on workgroup servers holds a pair of CPUs and a pair of memory units. The two CPUs share a Fireplane address bus connection to the motherboard. The four devices connect to the motherboard via a Dual CPU Data Switch, which is implemented from 8 bit-sliced ASICs. These same ASICs are used on all servers with more than two CPUs.

6.2 Fireplane level 1: Motherboard

Level 1 is located on the system motherboard. The workgroup Address Repeater ASIC has five address bus ports, one for each pair of system devices. A peak of one address can be broadcast every system
clock (150 MHz) on the address interconnect. This address rate determines the peak data bandwidth of 9.6 GBps, since an address is required to initiate each 64-byte data transfer.

The workgroup Data Switch is implemented with six bit-sliced ASICs that provides a 6x6 crossbar between four 32-byte wide ports for the Dual-CPU boards, and two 8-byte wide ports for the I/O controllers.

7 Mid-range interconnect
Mid range servers like the Sun Fire 3800-6800 need two levels of Fireplane interconnect for up to 56 devices (24 processors, 24 memory units, and eight I/O controllers). See Figure 11.

7.1 Fireplane level 1: system boards
Level 1 is implemented on two types of boards, one type for CPUs and memory, and the other type for I/O.

7.1.1 Uniboard. A Uniboard holds two pairs of CPU and two pairs of memory units. The Uniboard block diagram is shown in Figure 11, and Uniboard physical layout is shown in Figure 13. The Uniboard is used in all medium and large Sun Fire servers.

The Address Repeater ASIC collects address requests from the four CPUs, and forwards them up to the level-2 Address Repeater. It also broadcasts addresses that come down from level-2.

Eight Dual CPU Data Switch ASICs connect together a pair of CPUs and a pair of memory units. Four Data Switch ASICs connect the two halves of the board to the
level 2 data interconnect via a 32-byte wide, 4.8 GBps path.

**Figure 11. Mid-range interconnect (Sun Fire 3800-6800)**

### 7.1.2 I/O assemblies

An I/O assembly holds a pair of I/O controllers. The Address Repeater ASIC collects address requests from the two I/O controllers, and forwards them up to the level-2 Address Repeater. It also broadcasts addresses that come down from level-2. Two Data Switch ASIC connect the two I/O controllers to the level-2 data switch via a 16-byte wide, 2.4 GBps path.

So far, there are three types of I/O assemblies for the mid range. They all have the same architecture, and provide two 33 MHz PCI buses, and two 66 MHz PCI buses. They have different numbers and types of PCI slots:

- Six compact PCI slots (Sun Fire 3800).
- Four compact PCI slots (Sun Fire 4800-6800).
- Eight standard PCI cards (Sun Fire 4800-6800).
7.2 Fireplane level 2: Switch boards
Level 2 connects multiple Uniboards and I/O assemblies. Level 2 is the top of the address-broadcast tree, and so encompasses a snooping coherency domain. The maximum data bandwidth inside a snooping coherency domain is determined by the 150 MHz address rate \( \times 64\)-byte cache block = 9.6 GBps.

7.2.1 Sun Fire 3800. Level 2 is implemented on the motherboard. The Sun Fire 3800 can hold two Uniboards (8 CPUs) and two six-card Compact PCI assemblies.

7.2.2 Sun Fire 4800-6800. Level 2 is implemented on hot-swappable Fireplane Switch boards. Two Fireplane Switch boards are used in the Sun Fire 4800-4810 to interconnect three Uniboards and two I/O assemblies, and four are used in the Sun Fire 6800 to interconnect six Uniboards and four I/O assemblies. A Fireplane Switch board contains one Address Repeater ASIC and two Data Switch ASICs.

These systems use the four-slot Compact PCI assembly, and the eight-slot regular PCI assembly.

7.2.3 Segments. The mid-range level-2 interconnect can optionally be split into two segments, where half the Fireplane Switch boards are in one segment, and half are in the other. This creates two independent systems in one box which do not share any logic components. Configuring a mid-range system into two segments doubles the address-limited bandwidth of the box to 19.2 GBps, but halves the Fireplane Switch bandwidth, since it must then be run in double-pumped mode.

When the segments are configured along the half-cabinet power-grid boundaries of a Sun Fire 6800, then there is no electrical connection between the two halves, and they function as a totally isolated, two-way cluster-in-a-box.

7.2.4 Domains. A segment is configured into one or two Dynamic System Domains. Each domain runs a separate instance of Solaris, and has its own boot disk. A domain must have at least one Uniboard and one I/O assembly, so there can be up to two domains in a Sun Fire 3800-4810, and four in a Sun Fire 6800. A domain is isolated from other domain’s Uniboard and I/O assembly hardware faults, as well as any software faults.

When there are two domains in a segment, they share the Fireplane Switch boards of that segment, and split between them the address bandwidth. A system board failure in one domain will not affect the other domain, but a switch board failure will bring down both domains in a segment.

8 High-end interconnect
A large server like the Sun Fire 15K needs three levels of interconnect for up to 180 devices (72–106 processors, 72, memory units, and 36–2 I/O controllers). See Figure 12.

8.1 Fireplane level 1: system boards

8.1.1 Uniboard. This is the same Uniboard as on all mid and high-end servers.

8.1.2 I/O assembly. The hot swap PCI I/O assembly has the same architecture as for the mid-range systems, but is different physically to fit into the slot-1 form factor of the expander board. It accommodates up to four standard PCI cards, which are mounted in hot-swappable cassettes.

8.1.3 MaxCPU board. When not all of the 18 type-1 slots are needed for I/O con-
nectivity or I/O bandwidth, then the remainder of the type-1 slots can be populated with MaxCPU boards.

A MaxCPU board has two CPUs, but no on-board memory. These CPUs use the memory on other boards. The off-board bandwidth per CPU is 1.2 GBps, the same as for the Uniboard.

---

**Figure 12. High-end interconnect (Sun Fire 15K)**
8.2 Fireplane level 2: Boardset

Level 2 is implemented on an Expander board. The Expander board has one type-0 slot for a Uniboard, and one type-1 slot for an I/O assembly or MaxCPU board. The three-board combination of an Expander board, a Uniboard, and slot-1 board is called a *boardset*. Each of these boards is hot swapable.

For transfers within a boardset, the System Address Controller acts as a level-2 Address Repeater, and the six System Data Interface ASICs act as a level-2 data switch. A boardset by itself with one Uniboard and one I/O assembly behaves identically to a similarly configured Sun Fire 3800-6800.

Figure 13. Sun Fire 15K boardset
8.3 Fireplane level 3: Centerplane

Level 3 connects together the 18 boardsets. Each boardset has separate input and output connections to the address crossbar and to the response crossbar. Address requests are sent across the address crossbar, and take two clocks. Replies are sent back on the response crossbar, and take one or two clocks. Each boardset has a 4.8 GBps peak data connection to the 18x18 centerplane data crossbar.

The centerplane address crossbar is implemented by four ASICs, the response crossbar by two ASICs, and the data crossbar by 14 ASICs. In the event of an ASIC failure, each of these crossbars can be individually reconfigured into double-pumped mode which operate at half the usual rate.

8.3.1 Directory coherency. Each boardset is a separate snooping coherency domain. When a processor requests an address in a different snooping coherency domain, the scalable-shared memory (SSM) agent in the system address controller ASIC on the requester’s expander board notices the request, and sends it across the address crossbar to the home SSM agent in the system address controller ASIC on the home boardset. Home is the boardset where a page of memory is physically located. The home agent keeps track of which snooping coherency domain any sharers are in, and where the current owner is.

The home agent broadcasts the request on the home address bus, and if the block is not owned by a cache on the home boardset, then memory supplies the data. The data is moved through the home boardset data switch, the centerplane data crossbar, and the requesting boardset’s data switch to reach the requester.

If the requested location is currently owned by a system device, then the home agent sends the address over the address crossbar to the SSM agent on the owning boardset. The owning agent broadcasts the address on its address bus, and the owning device supplies the data.

In all cases, the requesting agent reruns the address transaction on the requester’s address bus, to establish the request’s place in the global memory order.

The SSM protocol uses three tag bits stored in memory (the Mtags) to specify the global coherency state of a memory block. The Mtag state is in addition to the snoopy Modified, Owned, Exclusive, Shared, Invalid (MOESI) state of a line.

The Mtags are stored in the eight-bytes of error-correcting code (ECC) information that goes with each 64-byte memory block. Memory is used because the number of snoop tags that would have been required to represent all the data cached in other snooping coherency domains would have been too large to fit in high speed SRAM. Instead, each SSM agent has an SRAM Coherency Directory Cache (CDC) to give it quick access to the most recent coherency information.

For more detail on the Fireplane SSM coherency protocol, see [10].

8.3.2 Peak bandwidth. When all accesses in a Sun Fire 15K are local to each boardset, the peak interconnect bandwidth is the aggregate local interconnect bandwidth of each boardset: 18 boardsets × 9.6 GBps = 172 GBps. The peak local memory bandwidth is about 7.2 GBps per boardset, which gives a peak memory bandwidth of about 129 GBps. When all accesses are to a different boardset than the requester, the peak interconnect bandwidth is the bisection bandwidth of
the centerplane: 18 boardsets × 2.4 GBps = 43 GBps.

**8.3.3 Domains.** The Sun Fire 15K can be divided into as many as 18 Dynamic System Domains, on a board-level granularity. A Uniboard from a given boardset can be in one domain, and the I/O assembly can be in another, which makes it possible to move CPUs from one domain to another without affecting the I/O connections.

**9 Sun Fire server cabinets**

Section 9 provides an overview of the family packaging. Figure 14 shows an overall picture of the Fireplane interconnect. Figure 15 shows the family cabinets.

---

**Figure 14. Sun Fire family interconnect overview**
10 Error protection

Section 10 discusses the error protection specific to the interconnect. Other facilities that improve reliability, availability, and serviceability — like the System Controller — are beyond the scope of this paper.
10.1 Processor error protection
The location of the various processor error protection mechanisms is shown in Figure 9.

10.1.1 On-chip SRAM protection. The processor does error correction on its seven major on-chip SRAM structures:
- Instruction cache physical tags
- Instruction cache snoop tags
- Instruction cache contents (32 KB)
- Data cache physical tags
- Data cache snoop tags
- Data cache contents (64 KB)
- Tags for the external cache (90 KB).

The first six of these are protected by parity bits. On an error, a retry is done by a combination of hardware and software to reload the correct value. The external cache tags are protected by ECC.

10.1.2 Off-chip SRAM protection. The 8 MB contents of the external cache is located on off-chip SRAM chips, and is protected by hardware ECC.

10.2 Interconnect error protection
The location of the various interconnect error protection mechanisms is shown in Figure 16.

10.2.1 Address path protection. Levels 1 and 2 of the Fireplane address interconnect are protected by three parity error bits. The Address Repeater ASICs generate parity on outgoing transfers, and check the parity of incoming transfers.

Level 3 of the address interconnect, which goes across the centerplane of the Sun Fire 15K, is protected by ECC.

10.2.2 Datapath protection. The data path is protected by end-to-end ECC. An active system device (CPU or I/O controller) generates the ECC bits when it does a store, and the bits are carried all the way through the datapath to memory. The bits are later checked by an active system device when the data is read from memory. Single-bit errors are corrected, and double-bit errors are detected.

10.3 Fault isolation
Two additional checks are made to help isolate the cause of datapath errors:
- The level-1 Data Switches check the ECC of data that passes through them.
- Individual chip-to-chip connections are covered by parity. ASICs generate parity for outgoing transmissions, and check incoming data for errors.

The ECC checks that are done by the Level-1 Data Switches can identify the source of ECC errors in most cases. A particularly hard case for detecting the cause of ECC errors is when a device writes bad ECC into memory. Without extra checks, these errors would get detected only much later when some other device reads the location. The bad writer may have written bad ECC to many locations, which will later be read by many devices. Thus the errors will appear to be in many memory locations, when the real culprit was in fact a single bad writer. Since a level-1 Data Switch checks ECC for all data entering or leaving each device from other devices, the original source of errors can be isolated.
11 Future trends

CPU clock rates show no sign of stopping their steady upward progress. In addition, the next few years will see the wide deployment across the industry of multi-threaded CPUs and multiple CPUs per chip. The multiplicative effects of these three trends will produce a major upward step function in the load on system interconnects.

Like its predecessors, the Sun Fireplane is not the end of Sun’s interconnect evolu-
tion. The next-generation UltraSPARC-V processor will have both a new CPU core and an improved SMP interconnect. Sun will continue to push the speed of snoopy coherency for mid-sized systems.

The requirements of high and low end systems have diverged enough that Sun has announced a new low-end SMP interconnect, the JBus, for systems with four or fewer processors.

Several new industry-standard interconnects are contending for acceptance, among them PCI-X, 10 Gb Ethernet, Hyper-Transport, Rapid I/O, 3GIO, and InfiniBand. Some of these interconnects will be used for I/O card connectivity, some for small SMPs, and some for clustering and networking. InfiniBand looks promising to improve high-end horizontal scaling. For an overview of InfiniBand and a comparison to other standards, see [15] and [16].

12 References


