Alan Charlesworth
Sun Microsystems

The Sun Fireplane Interconnect in the Mid-Range Sun Fire Servers

Vertical & Horizontal Scaling

Many CPUs in one box
- Cache-coherent shared memory (SMP)
- Usually proprietary interconnect
- Can be dynamically partitioned
- High bandwidth & low latency
- Good performance on most parallel apps
- Can be more costly for bigger boxes

Many systems
- Separate systems, communicate by network APIs
- Usually commodity interconnect & boxes
- Needs partitionable parallel apps
- Databases typically can’t be horizontally scaled
Shared-Memory Server Sales

Source: IDC, March 2001

Datacenter Tiers

Users

Web Tier

Application Tier

Database Tier

OLTP Database Server

DSS Database Server

Horizontal scaling

Vertical scaling

Lots of Small sys

10’s small to big

Few Big Sys
Big Database Examples

SF 6800: 1000 GB TPC-H
1 Server 3 Storage racks
24 CPUs 352 drives
80 GB mem 6 TB

5-yr Millions KW Sq ft
Server $1.1 5.4 8.8
Storage $0.6 7.5 24.4
Software $0.5
HW Maint $0.9
Total $3.1 12.9 33.2

E10K: 3000 GB TPC-H
2 Servers 11 Storage racks
128 CPUs 1,392 drives
128 GB mem 24 TB

5-yr Millions KW Sq ft
Server $4.6 13.5 27.1
Storage $2.5 31.8 89.4
Software $1.4
HW Maint $4.6
Total $13.2 45.3 116.5

Cache Coherency

CPU 1  CPU 2  CPU 3  CPU 4

1. Read to Share
2. Read to Share
3. Read to Own
4. Read to Share
5. Writeback

Coherency blocks
(Aligned 32, 64, or 128 bytes)
Broadcast & Point-to-Point

1. Broadcast (Snoopy)
   - All addresses sent everywhere
   - Snoop result computed in a few cycles
   - Lowest possible latency, especially for cache-to-cache transfers
   - Data bandwidth limited by snoop bandwidth

2. Point-to-point (Directory)
   - Directory keeps track of who is “interested” in each block
   - Addresses sent only to “interested” parties
   - Latency usually longer
   - Bandwidth can be much greater

Sun Interconnect Timeline

- UltraSPARC-I / UPA (1)
- SuperSPARC / XDBus (2)
- Cypress SPARC / MBus (3)
- UltraSPARC-III / Fireplane (4)
- UltraSPARC-V (5)

Timeline:
- Development
- Production
- ‘90
- ‘95
- ‘00
- Now

CPU core / Interconnect
### Sun Interconnect Generations

<table>
<thead>
<tr>
<th>Year (in mid-size servers)</th>
<th>MBus</th>
<th>XDBus</th>
<th>UPA</th>
<th>Fireplane</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>40</td>
<td>50–55</td>
<td>83–100</td>
<td>150</td>
</tr>
</tbody>
</table>

| System clock (MHz)       | 40   | 50–55 | 83–100 | 150       |

<table>
<thead>
<tr>
<th>Coherency type</th>
<th>Broadcast</th>
<th>Broadcast point-to-point</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Packet switching</th>
<th>Circuit</th>
<th>Packet switched</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Address &amp; Data</th>
<th>Together</th>
<th>Separate</th>
</tr>
</thead>
</table>

| Coherency block (bytes)   | 32       | 64       |

<table>
<thead>
<tr>
<th>Sys clocks/snoop</th>
<th>16</th>
<th>11</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
</table>

| Address B/W (GBps)        | 0.08     | 0.3      | 3.0     | 9.6       |

| # Address buses           | 1        | 4        | 4       | >16       |

| Max data B/W (GBps)       | 0.08     | 1.3      | 12.8    | >150      |

| Datapath width (bytes)    | 8        | 16       | 32      |

| Wiring                    | Bused    | Mid: Bused | High: Switched | Switched |

### Sun Snooping Progress

- **1. MBus**
- **2. XDBus**
- **3. UPA**
- **4. Fireplane**

- Doubling every 18 months trend line

**Year of first shipment in medium-sized servers**

- **1990**
- **1992**
- **1994**
- **1996**
- **1998**
- **2000**
- **2002**
- **2004**
Fireplane Cache Coherency

1. Broadcast (snoopy) coherency *inside* a snooping coherency domain

2. Point-to-point (directory) coherency *between* snooping coherency domains

Address Bus Implementation
Snoopy Coherence Domain

- Address transaction
- Data transfer
- Snoop
- Memory cycle
- Processor
- Memory
- I/O interface
- Address Repeater
- Data Switch

Broadcast address bus
Top-level Address Repeater

CPU/Mem 1 2 3 4 5
IO 1 2 3 4 5

Data path

UltraSPARC-III / Fireplane — 2000
- Ultra-III Processor
- External cache controller
- External cache tags
- Coherency controller
- Memory controller

UltraSPARC-I / UPA — 1996
- Ultra-I Processor
- External cache controller
- External cache tags
- Coherency controller
- Memory controller

SuperSPARC / XDBus — 1993
- SuperSPARC Processor
- External cache controller
- External cache tags
- Coherency controller
- Memory controller

Cypress SPARC / MBus — 1990
- Cypress FPU
- Cypress IU
- Cache controller and Coherency controller
- Cache tags
- Memory controller

Increasing CPU Integration
UltraSPARC-III Processor

Processor chip
- 9 instructions/clock
- Instruction Issue Unit
  - 10 instruction queue
  - 16 K branch predictor
  - Instruction TLB
- Integer Unit
  - 16 registers
  - ALU pipe 0
  - ALU pipe 1
  - Load/Store pipe
  - Branch pipe
- Floating-Point Unit
  - 32 registers
  - FP multiply/graphics/div pipe
  - FP add/graphics pipe
- Instruction Cache
  - 32 KB
- Data Cache
  - 64 KB
- Data TLB
- Prefetch Cache
  - 2 KB
- Write cache
  - 2 KB
- External Cache
  - Control Tags (90 KB)

Fireplane system interface & memory control

External Cache Control
- 8 MB

SRAM DIMM
- 8 MB External Cache data

P = Parity generate & check
E = ECC generate, check & correct

Sun Fire System Board

Address repeater
P = Parity generate & check
C = ECC check
E = ECC generate, check & correct

Datapath controller
- 150 million snoops/sec

SDRAM DIMM
- 4.8 GBps

Memory
- 2.4 GBps
Sun Fire IO Assembly

- Address repeater
- Datapath controller
- PCI controller
- PCI card (33 MHz)
- PCI card (66 MHz)
- Address repeater
- Datapath controller
- PCI controller
- PCI card (33 MHz)
- PCI card (66 MHz)

- 150 million snoops/sec
- 2.4 Gbps (16 bytes)
- 1.2 Gbps
- ~200 MBps
- ~400 MBps
- P = Parity generate & check
- C = ECC check
- E = ECC generate, check & correct

Fireplane Switch Boards

- Address repeater
- Datapath controller
- Data switch
- 5 pairs of address ports
- Six 32-byte ports & four 16-byte ports
**System Board Picture**

- Two sets of 8 Dual CPU Data Switch ASICs
- Four banks of 8 SDRAM DIMMs

**I/O Assembly Pictures**

- 6 slot cPCI
- 8 slot PCI
### Sun Fire Server Cabinets

<table>
<thead>
<tr>
<th>Model</th>
<th>Type</th>
<th>Height (in)</th>
<th>Width (in)</th>
<th>Depth (in)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun Fire 3800</td>
<td>Rack mount</td>
<td>32&quot;</td>
<td>19&quot;</td>
<td>17.5&quot;</td>
</tr>
<tr>
<td>Sun Fire 4800</td>
<td>Rack mount</td>
<td>32&quot;</td>
<td>19&quot;</td>
<td>17.5&quot;</td>
</tr>
<tr>
<td>Sun Fire 4810</td>
<td>Rack mount</td>
<td>32&quot;</td>
<td>19&quot;</td>
<td>17.5&quot;</td>
</tr>
<tr>
<td>Sun Fire 4800/4810</td>
<td>Rack mount</td>
<td>32&quot;</td>
<td>24&quot;</td>
<td>17.5&quot;</td>
</tr>
<tr>
<td>Sun Fire 6800</td>
<td>Rack mount</td>
<td>32&quot;</td>
<td>24&quot;</td>
<td>17.5&quot;</td>
</tr>
<tr>
<td>Large Server</td>
<td>Cabinet processors</td>
<td>&gt;64&quot;</td>
<td>&gt;64&quot;</td>
<td>&gt;64&quot;</td>
</tr>
</tbody>
</table>
A Micro Benchmark

Parallel pointer-chasing

Benchmark Record

- **SpecWeb99**
  - 12 CPUs: Web serving

- **SpecJBB**
  - 24 CPUs: 32-bit OLTP app-tier perf

- **TPC-H 1 TB Decision Support**
  - 24 CPUs: Price/perf & perf/CPU

- **Oracle Apps**
  - 24 CPUs: OLTP performance

- **PeopleSoft**
  - 24 CPUs: General Ledger
  - 24 CPUs: Financials