TACC 'RANGER' INFINIBAND ARCHITECTURE WITH SUN TECHNOLOGY

John Fragalla
Principle Engineer
High Performance Computing
Sun Microsystems, Inc.
The Sun Constellation System

- The world’s largest general purpose compute cluster based on Sun Constellation System
  > 82 racks of Sun Blade 6048 Modular System
  > 2 Sun Sun Datacenter Switch 3456 Infiniband switches
  > 72 Sun X4500 “Thumpers” storage servers
- Sun is the sole HW supplier
- AMD Opteron Based
- Single Fabric based on InfiniBand with Mellanox Chipsets
- Expandable configuration
Configuration Summary

Two-tier InfiniBand topology

- A 24-port IB-NEM leaf switch on each 12-blade shelf
- Two Magnum central switches: 16 line cards each
- One 12x IB cable for every three nodes, 1,384 IB cables total
- Total Bisection BW: 46.08 Tbps (5.76 GB/s)

82 C48 Compute racks, (Sun Blade 6048)

- 3,936 Pegasus Blades w/ 2.3 GHz AMD, 4-socket, quad-core, 32 GB
  - 15,744 Sockets, 62,691 Cores, 125.95 TB of Memory
- 579.4 TFLOP/s Rpeak
- 433 TFLOP/s (#6) Rmax

Sun Lustre InfiniBand Storage

- 72 X4500 bulk storage nodes with 24-TB each
- Total Capacity is 1.7 PB
- Peak Measured BW: 50GB/s (Approx. 0.1GB/s : 1TFLOP/s)
Sun Blade 6048 Modular System
Massive Horizontal Scale

Unibody Chassis/Rack Design
- 48 server modules per rack
- 4-socket AMD Blades with 32 DIMMs
- 192 sockets per rack
- 768 cores per rack
- Weighs ~500 lbs. less than conventional rack/chassis combo
- TFLOP/s per Rack @ TACC: 7.07 (2.3GHz AMD Barcelona)

Modular Design
- 100 percent of active components hot-plug and redundant (power, fans, I/O, management)

Power and Cooling
- Eight 8,400 W (redundant in an N+N configuration may be configured to 5,600W)
- Common power and cooling

I/O
- Industry-standard PCI Express midplane
- Four x8 PCIe links per server module
- Two PCIe EM per server module, eight NEM per rack
Sun Blade 6048 InfiniBand DDR Switched Network Express Module

Supports 13,824 Node Clusters

Industry’s Only Chassis Integrated Leaf Switch
- Dual port DDR HCA per server module
- Pass thru GigE per server module
- Leverages native x8 PCIe I/O
- Compact design maximizes chassis utilization

Designed for Highly Resilient Fabrics
- Two onboard DDR IB switches
- Increased resiliency allowing connectivity up to four Ultra-dense switches

3:1 Reduction in Cabling Simplifies Cable Management
- 24 ports of 4x DDR connectivity realized with only 8 physical 12x connectors
Inside the SB6048 IB DDR Switched NEM

- **x8 PCIe to 12 Server Modules**

- **24 Port 384Gbps IB Switch**
  - Connects to Sun Data Center Switch 3456

- **24 Port 384Gbps IB Switch**
  - GigE Ports for Server Administration

- Contains 2 x 24-port Mellanox 384 Gbps Infiniscale III switches
- Supports clusters sizes of 13K server modules
  - Routes around failed 12x links
Sun Datacenter Switch 3456
Worlds Highest Scale & Density InfiniBand Switch

I/O
- 3456 IB 4x SRD/DDR connections
- 110 Terabits per second
- 1 usec latency

Fabric
- 720 24-port IB switch chips
- 5-stage switching

Availability
- Hot-swap power supplies, fans, CMC's
- Multi-Failover Subnet Manager Servers
- Multi-Redundant switching nodes

Management
- IPMI, CLI, HTTP, SNMP
- Software
- Host based subnet Management software
InfiniBand Cables

- Magnum switch & IB NEM uses Mollex iPass connectors for three 4x connections
  - 3x improvement in density over standard IB connectors
  - Three 4x connections in one 12x connector
- 12x to 12x cable
  - Magnum to C48 NEM
- 12x to triple 4x standard (splitter) cable
  - Magnum to three servers with standard IB HCA's

Dual 12 socket on Magnum line card or C48 NEM

12x Cable connector

Sun InfiniBand Splitter cable

Connects three non-compete-node HCA cards to a Magnum line card

<table>
<thead>
<tr>
<th>Product Code</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>74582-1011</td>
<td>11 meters</td>
</tr>
<tr>
<td>74582-1013</td>
<td>13 meters</td>
</tr>
<tr>
<td>74582-1015</td>
<td>15 meters</td>
</tr>
<tr>
<td>74582-1018</td>
<td>18 meters</td>
</tr>
</tbody>
</table>
A Lustre Cluster

- MDS disk storage containing metadata targets (MDT)
- Pool of clustered MDS servers 1-100
- Lustre Clients 1 - 100,000
- Simultaneous support of multiple network types
- Elan Myrinet InfiniBand
- Router
- OSS servers 1-1000s
- OSS storage with object storage targets (OST)
- Commodity Storage
- Shared storage enables failover OSS
- Enterprise-Class Storage Arrays and SAN Fabric

= failover
Sun Fire X4500
Dual Socket Data Server

Compute
- 2x AMD Opteron Sockets
- Dual-core
- 8x DDR1 DIMM slots – Up to 32GB memory

IO
- 2 PCI-X slots
- 4x Gigabit Ethernet ports
- 48x SATA 3.5” disk drives

Availability
- Hot-swap disk
- Hot-plug fans
- Hot-plug PSU, N+N

- Lustre OSS Building Block at TACC
- Lustre OSS BW: 0.8 GB/s
- 24 TB Raw Capacity
**TACC Configuration**

**3,936 Compute nodes**
- 328 blade shelves
- 504 Tflops
- 123 TiB memory size

**Bisection Bandwidth:**
3,936 * 1 GBps = 3.9 TBps (at current SDR IB rate)

**7 N1GE, N1SM, ROCKS & IB subnet manager nodes**
- 4S x 2C
- 2.6 GHz
- 16 GB mem X4600

**8 Gateway & Datamover nodes**
- 4S x 2C
- 2.6 GHz
- 16 GB mem X4600

**4 Login nodes**
- 4S x 4C
- 2.1 GHz
- 32 GB mem X4600

**2S x 2C X4500 (Thumper)**
- 48 500 GB drives

**72 Bulk Storage nodes**
- 1.7 PBytes

**48 splitter cables**

**48 splitter cables**

**6 splitter cables**

**4S x 2C 2.6 GHz 16 GB mem X4600**

**32-port FC4 switch**

**STK 6540 FC-RAID 9 TB**

**Metadata RAID**

**Magnum Switch 16-line cards (2,304-ports)**

**Magnum Switch 16-line cards (2,304-ports)**

**NEM with 12 IB HCAs & 24-port leaf switch**
- 1,312 cables from 82 blade racks

**Bisection Bandwidth:**
3,936 * 1 GBps = 3.9 TBps (at current SDR IB rate)
Blades and Shelf as Used at TACC

• Per blade:
  > 4 quad-core CPUs
  > 16 2 GB DIMMs
  > 8 GB flash for booting
  > One PCIe 8x connection
  > One Mellanox IB HCA (on NEM)

• Per shelf:
  > One 24-port IB leaf switch
  > Four 12x cables, each to a different line card
  > CMM connection to 100 MbE

NEM GigE ports, second NEM switch chip, and PEM slots are not used
Magnum Switches as Configured at TACC

- Two Magnum switches
  - Each with 16 line cards
  - A total of 4,608 4x ports
  - Line cards cabled in pairs, with empty slots left for cable access

- Largest IB network so far
  - Equivalent to 42 conventional 288-port IB switches
  - “Only” 1,400 cables needed — conventional IB switches would require 8,000 cables
Each blade shelf connects to two different line cards in both switches.

Two bundles of eight cables from each rack to each switch.

Four separate paths from each shelf.
TACC Floorplan

Size: approximately half a basketball court

- 2 Magnum switches
  - 16 line cards each
  - (2,304 4x IB ports each)

- 82 blade compute racks
  - (3,936 4S blades)

- 12 IO racks
  - (25 X4600 4 RU
    - 72 X4500 4 RU)

- 112 APC Row coolers

- 1,312 12x cables
  - (16 per rack)
  - 16 km total length

- 72 splitter cables
  - 6 per IO rack

12x cable lengths: 171 9m, 679 11m, 406 13m, 56 15m

Splitter cable lengths: 54 14m, 18 16m
TACC 'RANGER' INFINIBAND ARCHITECTURE WITH SUN TECHNOLOGY

Thank you
John.Fragalla@Sun.COM