# SUN™ DATACENTER SWITCH 3456 SYSTEM ARCHITECTURE

Massively Scalable InfiniBand Switch Architecture for Petascale Computing

White Paper
November 2007

# Table of Contents

# Executive Summary

Clusters or grids constructed from open and standard commercial off the shelf (COTS) systems now dominate the Top500 index of supercomputer sites[1], providing an attractive way to rapidly (and affordably) construct supercomputer-class systems of interconnected nodes. The largest of these supercomputers are now driving toward petascale deployments, delivering petaflops of computational capacity and petabytes of storage capacity. In spite of ongoing successes, those designing and building the world's largest and fastest new supercomputers face significant challenges, including:

- Rapidly building and growing the computational capacity of networked clusters to serve key needs before individual computing systems become obsolete
- Increasing levels of computational density while staying within essentially constrained envelopes of power and cooling
- Reducing complexity and cost for physical infrastructure and management
- Choosing interconnect technology that can connect hundreds or thousands of processors without introducing unacceptable levels of latency

In fact, interconnect technology plays a vital role in addressing all of these issues. InfiniBand has emerged as a compelling interconnect technology, and now provides more scalability and significantly better cost-performance than any other known fabric. In spite of its ability to provide essential high-speed connectivity and low latency, connecting and cabling thousands of compute nodes with smaller discrete InfiniBand switches remains problematic. With traditional approaches, the largest supercomputing clusters can require hundreds of switches as well as thousands of ports and cables for inter-switch connectivity alone. The result can be significant added cost and complexity, not to mention energy and space consumption.

To address these challenges, the Sun™ Datacenter Switch 3456 system provides the world's largest standards-based DDR (dual data-rate) InfiniBand switch, with direct capacity to host up to 3,456 server nodes. Only slightly larger than two conventional datacenter racks, the system drastically reduces the cost, power, and footprint of deploying very large-scale standards based high performance computing (HPC) fabrics. The Sun Datacenter Switch 3456 is tightly integrated with the Sun Blade™ 6048 modular system, facilitating deployment of very large and dense supercomputing grids with a minimum of complexity. Together these technologies offer lower acquisition costs, lower latency, very high compute density, greatly reduced cabling and management complexity, reduced real estate needs, and more teraflops per dollar with lower power consumption than with other solutions.

1.www.top500.org

Chapter 1

# Deploying Massive InfiniBand Connectivity

The largest HPC deployments are helping to push back the limits of understanding, letting engineers and scientists explore and visualize the answers to the world's most challenging and complex problems. At the same time, these large deployments often lead the way, with approaches and technologies eventually adopted by smaller-scale scientific and commercial users of HPC in both scientific and commercial pursuits. In fact, some of the world's largest supercomputers are now available to large groups of non-exclusive users, helping to answer complex problems in a host of business, medical, engineering, and scientific arenas.

## The Drive for Performance and Scale

Since the late 1950s, every decade has had its unanswerable computational problems, and its own technological limitations. While the basic quest — getting more answers more quickly — has remained the same, approaches have differed greatly. It is only relatively recently that the dream of building standards-based supercomputers has become a tangible reality. Now that reality is taking supercomputing toward true petascale deployments.

### The Rise of Interconnected Clusters and Grids

Since its origins in the late 1950s and early 1960s, supercomputing technology has taken a wide range of forms. Vector supercomputers led the way starting with Seymour Cray's activities at CDC, and then by Cray Research and others. Still available today, these systems offer a powerful system for computational codes that are not generally partitionable. Proprietary massively parallel processors (MPPs), shared memory multiple vector processors (SMPs), and distributed shared memory (DSM) systems followed. While these approaches created considerable architectural diversity, they failed to develop into a compelling industry-wide paradigm that rewarded both their inventors or their clients. Even the Cray moniker itself has moved from company to company, with Sun Microsystems eventually purchasing the SPARC® based multiprocessor technology that would eventually become the Sun Enterprise™ 10000 server.

Starting near the end of the 20th century, clusters and grids of commercial off-the-shelf x84 and x64 servers began to replace "conventional" supercomputers and their more exotic architectures. With increasingly open software such as OpenMPI and a basis on open and standard distributed resource management (DRM) technology, these systems have made made supercomputing technology available to more applications and larger numbers of industries and people. In fact, clusters and grids now easily dominate the Top500 List of SuperComputer sites (*www.top500.org/stats/list/29/archtype*).

Clusters and grids now also permeate scientific, educational, research, and commercial institutions alike due to their lower cost of acquisition, scalability, and generally accessible programming models. The ability to quickly and incrementally replace individual servers with the latest and fastest technology is a significant advantage. The ubiquity of clusters and grids is due in large part to new technologies based on open systems and standards. Key technologies are now helping clusters and grids scale while making the most of the resources involved:

- Distributed Resource Management (DRM) technology such as Sun Grid Engine software helps assign computational resources appropriately, allowing the grid or cluster to be shared effectively between multiple users and projects.
- Virtualization technology is increasingly important, across systems, storage, and networking, allowing isolated and fine-grained assignment of resources where they are needed most.
- Visualization technology is letting scientists see the patterns behind their data, scale visualization processing across multiple systems, and share visual results with others for valuable collaboration.

## Connectivity Challenges

InfiniBand technology has emerged as an attractive fabric for building large supercomputing grids and clusters. As an open standard, InfiniBand presents a compelling choice over proprietary interconnect technologies that depend on the success and innovation of a single vendor. InfiniBand also presents a number of significant technical advantages, including:

- A switched fabric offers considerable scalability, supporting large numbers of simultaneous collision-free connections with virtually no increase in latency.
- Host channel adaptors (HCAs) with RDMA support offload communications processing from the operating system, leaving more processor resources available for computation.
- Fault isolation and troubleshooting are easier in switched environments since problems can be isolated to a single connection.
- Applications that rely on bandwidth or quality or service are also well served since they each receive their own dedicated bandwidth.

Even with these advantages, building the largest InfiniBand clusters and grids remains complex and expensive — primarily because of the need to interconnect very large numbers of computational nodes. Density, consolidation, and management are not just important for computational platforms, but for InfiniBand interconnect infrastructure as well. Even with very significant accomplishments in terms of processor performance and computational density, large clusters are ultimately constrained by real estate and and the complexities and limitations of interconnect technologies. Cable length limitations constrain how many systems can be connected together in a given physical space while avoiding increased latency. Traditional large clusters require literally

thousands of cables and connections and hundreds of individual core and leaf switches, adding considerable expense, cable management complexity, and consumption of valuable datacenter rack space.

## The Sun™ Datacenter Switch 3456 System

The Sun Datacenter Switch (DS) 3456 system is designed to drastically reduce the cost and complexity of delivering large-scale HPC solutions, such as those scaled for leadership in the Top500 list of supercomputing sites, as well as HPC applications in scientific, technical, and financial markets. Providing the world's largest InfiniBand core switch (Figure 1) with capacity for connection of up to 3,456 nodes, the Sun DS 3456 allows deployment of more teraflops per dollar and lower complexity and power consumption than is possible from alternative solutions. The Sun DS 3456 is a non-blocking monolithic core switch, designed to replace as many as 300 standard switches and thousands of cables.

*Figure 1. The Sun Datacenter 3456 system offers up to 3,456 4x InfiniBand connections in a single dual-width datacenter cabinet.*

The Sun DS 3456 provides the following specifications:

- 55 Tbps bisection bandwidth
- Port-to-port latency of 700 nanoseconds
- Congestion control with Forward and Backward Explicit Congestion Notification (FECN and BECN)
- 8 data virtual lanes
- 1 management virtual lane
- 4096 byte MTU
- Ability to scale to 13,824 end nodes[1] when deployed with the Sun Blade 6048 Switched InfiniBand Network Express Module (NEM)

1.Four core switches required, with each core switch contributing 3,456 4x InfiniBand connections

## The World's Largest Standards-Based InfiniBand Switch System

InfiniBand technology is growing rapidly, as witnessed by rapid adoption by many of the top supercomputing grids in the Top500 list of supercomputing sites. Not only is InfiniBand more scalable than any known fabric, but it also delivers significantly better cost-performance than any other datacenter network. With its datacenter background and systems focus, Sun is able to make considerable contributions to InfiniBand technology.

The Sun DS 3456 changes the model for deploying large-scale InfiniBand networking. In its design, Sun took a systemic approach that goes beyond individual networking components to deliver a system that is designed to deliver significant gains in scalability, simplification, and cost. The Sun DS 3456 is compliant with the InfiniBand specification, based on the PCI Industrial Computer Manufacturers Group (PICMG) shelf management specification, fully compliant with the OpenMPI and Open Fabrics host stacks, and (optionally) with the OpenSM subnet manager. The Sun DS 3456 is comprised of the following key components:

- An ultra scalable InfiniBand switching fabric
- Ultra high-performance host adapters
- A new high-density cabling system
- An open management platform

With a bisection bandwidth of 55 Tbps and hosting up to 3,456 connections, the Sun DS 3456 is deployed a single chassis that is slightly larger than two full datacenter racks (when accounting for integrated cable management). Combined with a custom-designed leaf switch, a single switch can connect up to 72 Sun Blade 6048 modular system chassis. Multiple switches can also be combined for redundancy, or to connect even larger numbers of servers. Sun’s systemic focus brings multiple benefits. The Sun DS 3456:

- Reduces cable count 6:1 over legacy InfiniBand Clos (Fat Tree) topologies
- Reduces complexity by reducing management points while adding enormous scale to InfiniBand networking
- Provides a vastly more reliable connector and cabling system
- Reduces datacenter footprint by consolidating as many as 300 discrete InfiniBand switches into a relatively compact 48” x 62” x 30” space
- Reduces network latency while increasing reliability and resiliency by replacing as many as 3456 cables with internal circuit traces
- Facilitates cluster scaling up to 13,824 servers (as many as 288 Sun Blade 6048 modular system chassis with four Sun DS 3456 core switches)
- Offers considerable cost savings per server node port compared with competitive solutions

## Key Technical Innovations

Building the largest supercomputing and HPC configurations demands a new approach, one that rapidly helps bring new technology to bear on the most important problems and questions. Not content to accept the status quo in terms of available InfiniBand switching, cabling, and host adapters, Sun engineers used their considerable networking and datacenter experience to view InfiniBand technology from a systemic perspective.

More than just a switch, the Sun DS 3456 represents a complete system that is based on multiple technical innovations:

- The Sun DS 3456 chassis implements a five-stage Clos fabric with up to 720 24-port Mellanox InfiniScale III switching elements, integrated into a single mechanical modular enclosure.
- Sun's custom-designed miniature (12x) connector[1] consolidates three discrete InfiniBand 4x connectors, resulting in the ability to host 144 4x ports through 48 physical 12x connectors on a single line card.
- Complementing the 12x connector, a 12x trunking cable carries signals from three servers to a single switch connector, offering a 3:1 cable reduction when used for server trunking, and reducing the number of cables needed to support 3,456 servers to 1,152. A splitter cable that converts one 12x connection to three 4x connections is provided for connectivity to legacy InfiniBand equipment.
- Integrated cable signal conditioner circuitry is provided that optimizes 4x signal integrity to support long cable runs, effectively doubling the supported cable length for dual data rate (DDR) to 15 meters, and facilitating the use less expensive, thinner and more flexible copper cables.
- The cable and connector system provides status information such as cable serial number and events such as local and remote cable insertion, and remote node power. This information is used to keep track of presence, state and connectivity by the Fabric Director, and can assist the system administrator during construction and fault isolation.
- A custom-designed double-height Network Express Module (NEM) for the Sun Blade 6048 modular system provides seamless connectivity to the Sun DS 3456. Using the same 12x connectors, the Sun Blade 6048 InfiniBand Switched NEM can trunk 12 Sun Blade 6000 server modules to the Sun DS 3456 using only four 12x cables. The switched NEM together with the 12x cable facilitates connectivity of up to 13,824 servers in a 7-stage Clos topology.
- Fabric topology for forwarding InfiniBand traffic is established by a redundant host-based Subnet Manager. A host-based solution allows the Subnet Manager to take advantage of the full resources of a general-purpose multicore server.

1.Sun's 12x connector has been proposed to the InfiniBand Working Group in response to a call for a new 12x connector standard.

## Driving Down Cost and Complexity

Given the scale involved with building supercomputing clusters and grids, cost and complexity figure large. Regrettably, traditional approaches to using InfiniBand for massive connectivity have required very large numbers of conventional switches and cables. In these configurations, many cables and ports are consumed redundantly connecting core and leaf switches together, making advertised per-port switch costs relatively meaningless, and reducing reliability through extra cabling.

In contrast, the very dense InfiniBand fabric provided by the the Sun DS 3456 is able to eliminate hundreds of switches and thousands of cables — and millions of dollars in acquisition costs. Table 1 illustrates Sun estimates for deploying a 3,456-node HPC supercomputing cluster using the Sun DS 3456 and a leading competitor. The Sun DS 3456 can easily provide significant savings in terms of cost of acquisition for switches and cables, not to mention the added savings from simplified deployment and long-term management. With simplified deployment, organizations can also build large supercomputing clusters more quickly, and get more life out of deployed technology.

*Table 1. Building out a 3,456-node HPC supercomputing cluster using the Sun DS 3456 provides drastic reductions in both complexity and cost over traditional InfiniBand switching technology.*

| Category | Sun DS 3456 | Traditional InfiniBand infrastructure | Consolidation Ratio |
|---|---|---|---|
| Number of required leaf switches | 0 | 288 | — |
| Number of required core switches | 1 | 12 | 12:1 |
| Total 4x switch ports needed to connect 3,456 nodes | 3,456 (1,152 physical 12x connections) | 10,368 (4x) | 3:1 |
| Number of leaf-to-core trunking cables | 1152 (12x) | 3,456 (4x) | 3:1 |
| Number of host channel adapter (HCA) to fabric cables | 0 | 3,456 | — |
| Total InfiniBand cables | 1,152 | 6,912 | 6:1 |

## The Sun Constellation System

With an overall datacenter focus, Sun is free to innovate at all levels of the system — from switching fabric through to core system and storage design. In 2006, this approach allowed Sun to help the Tokyo Institute of Technology rapidly build the largest supercomputer in Asia and one of the largest outside the United States. The TSUBAME grid included hundreds of systems incorporating thousands of processor cores, terabytes of memory, and 1.1 petabytes of storage. Best of all, this supercomputing grid was built in just 35 days using advanced multiprocessor, multicore Sun Fire™ X4600 servers as compute nodes, and innovative Sun Fire X4500 servers for data storage.

Even larger supercomputing grids are easily made possible with the Sun Constellation System. Sun's systemic focus provides a highly-integrated InfiniBand interconnect together with computational and storage solutions that functions as a singular system. Best of all, organizations can obtain all of these tightly-integrated building blocks from a single vendor with a unified management approach.

Components of the Sun Constellation System include:

- The Sun DS 3456 core switch for an ultra-dense switching solution and support of 3,456 nodes per switch
- The Sun Blade 6048 modular system, providing an ultra-dense blade platform with support for 48 Sun Blade 6000 server modules in a custom rack-sized chassis
- Sun Fire X4500 storage clusters, with support for up to 480 terabytes in a single rack

In particular, the Sun Blade 6048 modular system was designed to work seamlessly with the Sun DS 3456, providing extreme density and performance and helping to enable construction of the largest supercomputing clusters in existence. For example, the Texas Advanced Computing Center at the University of Texas at Austin is working with Sun to build one such supercomputer that will employ the Sun DS 3456 and the Sun Blade 6048 modular system. This 3,936-node cluster is expected to provide a peak rating of over 500 teraflops of performance when complete[1].

The Sun Blade 6048 modular system offers extreme density and performance through a general-purpose chassis that leverages a choice of standard UltraSPARC® T1 and T2, Intel® Xeon®, and AMD Opteron™ processors. Using Sun's modular design, these systems combine compelling performance and large memory support with power and cooling efficiency and large I/O capacity — all tightly integrated with the Sun DS 3456 core switch. Key aspects of the Sun Blade 6048 modular system are listed below.

- The Sun Blade 6048 modular system is provided in a single rack chassis with four integral shelves.
- Each modular shelf provides space for up to 12 Sun Blade 6000 server modules along with dedicated I/O, power, cooling, and management.
- The custom-designed Sun Blade 6048 InfiniBand Switched Network Expansion Module (NEM) provides a 24-port InfiniBand leaf switch that can connect 12 server blades to the Sun DS 3456 core switch using only four 12x cables.

Additional information on the Sun Blade 6048 modular system can be found in Chapter 4 and in the Sun Blade 6000 and 6048 Modular System technical white paper, and additional information on the Sun Fire x4500 server can be found in the Sun Fire X4500 Server technical white paper. The remainder of this document describes the architecture and environment of the Sun DS 3456.

1.The TACC Ranger supercomputing cluster currently uses components that have not yet been announced by Sun.

Chapter 2

# Sun Datacenter Switch 3456 System Architecture

The architecture of the Sun DS 3456 is unique. Setting out to economically build the largest InfiniBand switch possible, Sun engineers derived an elegant and efficient architecture based on currently available off–the-shelf technology. Beyond the switch chassis and its components, Sun also applied considerable innovation to cable connectors, cable management, and integration with compute nodes.

## Constraints for Large InfiniBand Switches

Fabrics for supercomputers have traditionally been based upon either Torus topologies or Clos topologies. Some of the most notable supercomputers based upon Torus topologies include IBM's BlueGene and Cray's XT3/XT4 supercomputers. Torus topologies consist of $k^d$ nodes, where d is the number of dimensions and k is the number of nodes per axis, yielding $dk/2 - 1$ stages (hops), with a bisection width of $2k^{(d-1)}$. Torus fabrics have had the advantage that they have generally been easier to build than Clos topologies. Unfortunately, torus topologies represent a blocking fabric. In addition, torus topologies provide variable latency due to variable hop count, implying that application deployment for torus fabrics must carefully consider node locality.

First described by Charles Clos in 1953, Clos networks have long formed the basis for practical multi-stage telephone switching systems. Clos networks utilize a "Fat Tree" topology, allow complex switching networks to be built using many fewer crosspoints than if the entire system were implemented as a single large crossbar switch. Clos switches are typically comprised of multiple tiers and stages (hops), with each tier comprised of a number of crossbar switches. Connectivity only exists between switch chips on adjacent tiers. Clos fabrics have the advantage of being a non-blocking fabric in which each attached node has a constant bandwidth. In addition, an equal number of stages between nodes provides for uniform latency. The historical disadvantage of Clos networks is that they have been more difficult to build.

Clos fabrics consists of crossbar switches organized in tiers. The number of switches is determined by $(2T - 1)(n/2)^{(T-1)}$ where $n$ is the number of ports in a crossbar switch element, and $T$ represents the number of tiers. The maximum number of hops (or switching stages) is represented by $2T - 1$, and $2(n/2)^T$ end nodes are supported. To achieve a higher number of nodes, it is preferable to have a larger number of ports per crossbar switch element ($n$) than a large number of tiers ($T$) — since increasing $n$ will increase node count with fewer tiers/stages, and provides lower latency. The Mellanox InfiniScale III switch chip used in the Sun DS 3456 provides 24 ports. With 24-port switches it is possible to generate Clos networks supporting 24, 288, 3456, and 41472 ports for 1, 2, 3, and 4 tiers respectively. It is expected that next

generation InfiniBand switch silicon implemented in 65mm and 45mm chip production processes will support larger port counts, thus allowing both larger as well as more compact topologies. A summary of Clos fabric parameters and 3D torus parameters can be found in Table 2.

*Table 2. A comparison of Clos and Torus parameters and properties.*

| Properties | Clos (Parameters T=3, n=24) | | Torus (parameters d=3, k=15) | |
|---|---|---|---|---|
| Stages (S) | 2T-1 | 5 | dk/2-1 | 44 |
| Nodes (N) | $2(n/2)^T$ | 3,456 | $k^d$ | 3,375 |
| Switch elements (S) | $(2T-1)(n/2)^{T-1}$ | 720 | $k^d$ | 3,375 |
| Bi-section width (W) | $2(n/2)^T$ | 3,456 | $2k^{d-1}$ | 450 |
| Connections (C) | $2T(n/2)^T$ | 10,368 | $dk^d$ | 10,125 |
| Average distance (A) | 2T | 6 | dk/4 | 12 |

A maximal 3-tier, 5-stage Clos fabric is depicted schematically in Figure 2, with each vertical row of switch elements defining a stage.

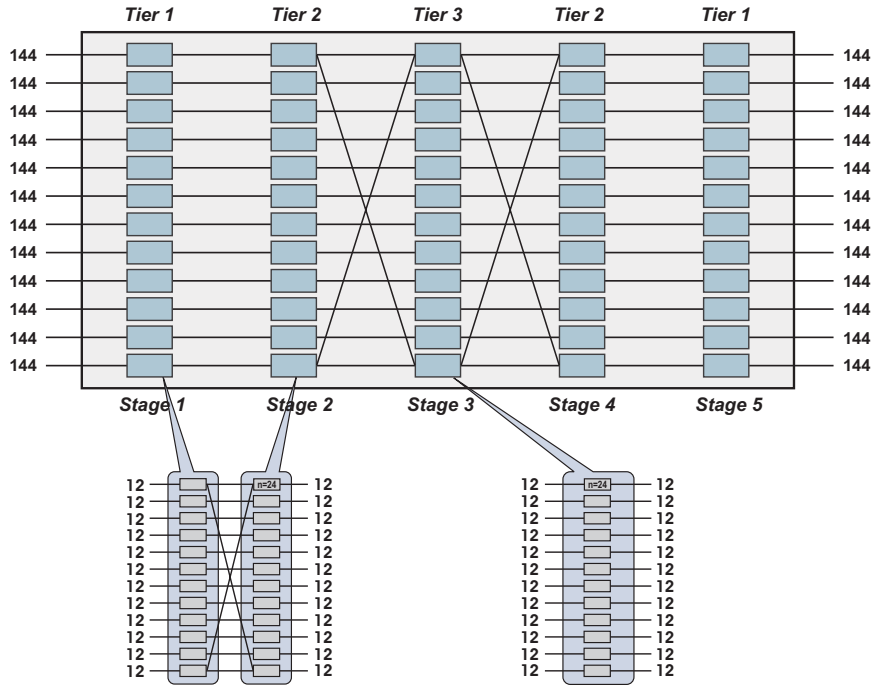



*Figure 2. A 3-tier, 5-stage Clos fabric can achieve single-system connectivity of 3,456 ports using 144 24-port switch chips per stage and 720 chips total.*

Building very large InfiniBand Clos switches is governed by a number of practical constraints, including the number of ports available in individual switch elements, maximum printed circuit board size, and maximum connector density. The Sun DS 3456 employs considerable innovation in all of these areas and implements a maximal three-tier, five-stage Clos fabric as depicted above. Each 144-port output in the figure represents a single line card deployed in the switch chassis.

## System Overview

Each Sun DS 3456 core switch is provided in a 24-slot double-wide rack chassis. Extensive integrable cable management is also provided (described later in this section). Labeled perspectives of the front and back of the chassis are provided in Figure 3 and Figure 4.

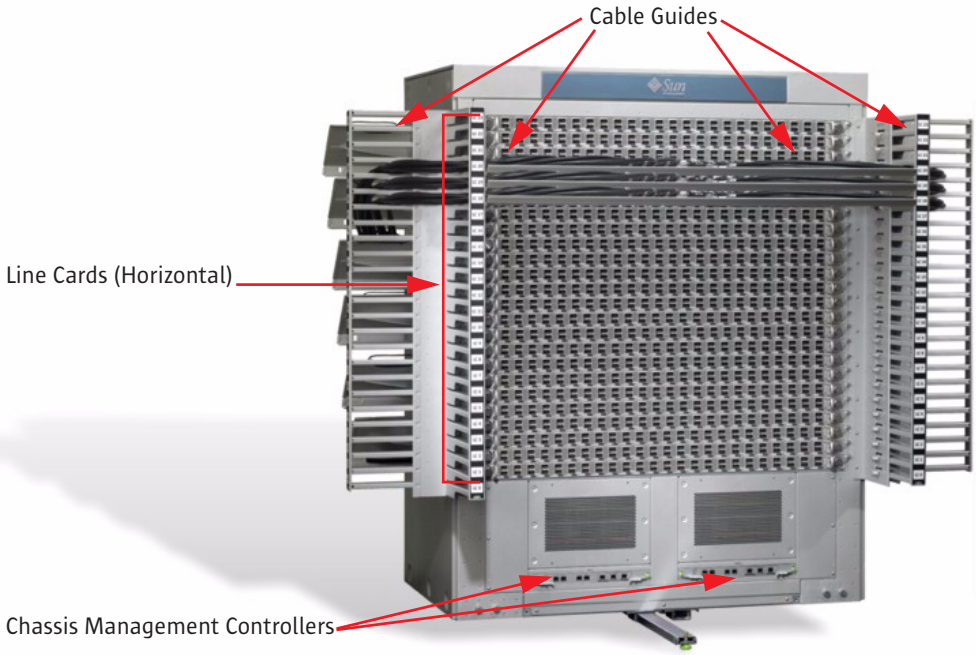



*Figure 3. Chassis front perspective*

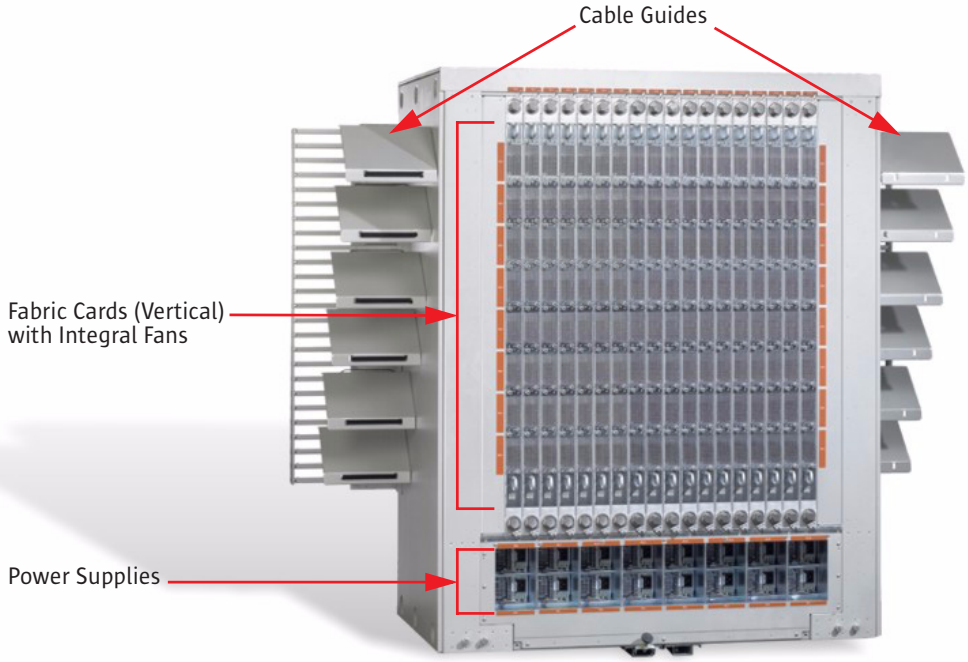



*Figure 4. Chassis rear perspective*

High-level system components include the following:

- Up to 24 horizontally-installed line cards insert from the front of the chassis with each providing 48 12x connectors delivering 144 DDR 4x InfiniBand ports. Each line card connects to pass-through connectors in a passive orthogonal midplane.
- 18 vertically-installed fabric cards insert from the rear, and are directly connected to the line cards through the orthogonal midplane. Each fabric card also features eight modular high-performance fans that provide front-to-back cooling for the chassis.
- Two fully-redundant chassis management controller cards (CMCs) insert from the front of the system. These modules monitor all critical chassis functions including power, cooling, line cards, fabric cards, and fan modules.
- Six to 16 (N+1) power supply units (PSUs) insert from the rear of the chassis with number depending on the line card population in the chassis. The PSUs are divided into two banks of eight, with each bank providing power to half the line cards and half the fabric cards.

The fabric card, line card, and chassis management controller are shown in Figure 5.

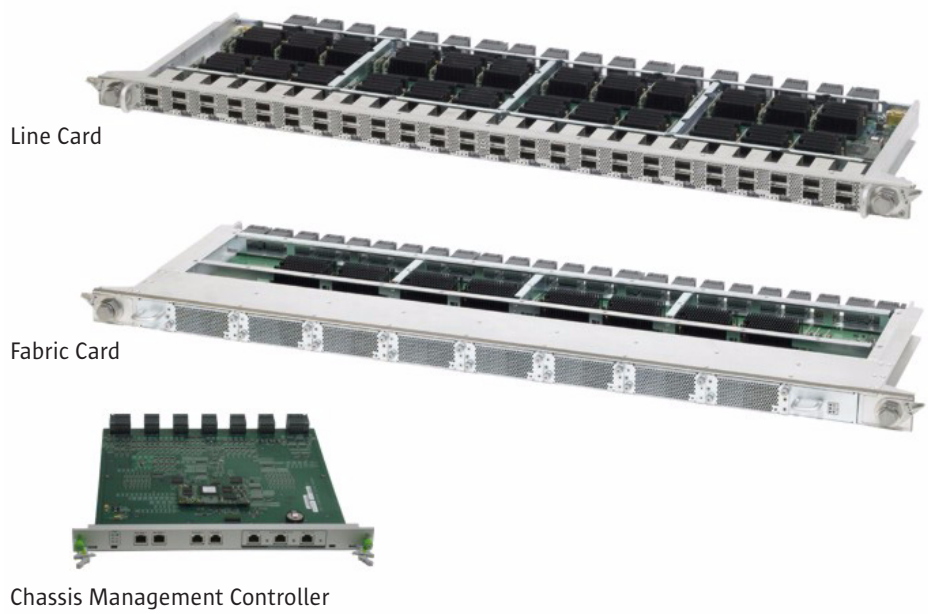



*Figure 5. Line card, fabric card, and chassis management controller (CMC) for the Sun DS 3456*

Table 3 lists the physical chassis specifications both with and without cable management.

*Table 3. Sun DS 3456 physical characteristics*

| Physical Characteristics | Without Cable Management | With Cable Management |
|---|---|---|
| Height | 62 inches | 62 inches |
| Width | 48 inches | 71.6 inches |
| Depth | 28.5 inches | 44.4 inches |
| Weight | 2,561.27 pounds | 2,908.8 pounds |
| Power requirements (maximum for a fully-populated chassis) | 36 kilowatts | 36 kilowatts |

## Innovative Interconnect System

With large InfiniBand deployments, cable management can present a significant challenge. Cable length limitations restrict the number of systems that can physically be placed close enough to switches without increasing latency. Literally thousands of cables are typically required, both to connect to servers as well as providing interconnectivity between core and leaf switches.

Beyond drastically reducing the number of cables required by serving up to 3,456 servers from a single switch, Sun has addressed InfiniBand cabling technology directly. Sun engineers designed innovative passive 12x cables and custom-tooled miniature 12x connectors that support three times the the density of conventional CX4 cables. Each 12x cable carries three discrete 4x InfiniBand connections. The direct result is a 3:1 consolidation of cabling, with 3,456 connections requiring only 1,152 12x connectors and cables. In addition, active circuity in both the line card and the Sun Blade 6048 InfiniBand Switched NEM effectively doubles the supported cable length for DDR to 15 meters.

The new 12x connector is mechanically superior to other solutions. Several 12x cables with their iPASS Plus connectors are shown inserted into line cards in Figure 6. Two 12x cables plug in to each set of dual (double-stacked) iPASS sockets on each line card. The double-stacked connector provides additional density and helps conserve valuable space on the edge of line cards. An integral gasket in the back shell assembly of the connector provide an EMI seal to cage interference and a flexible pull tab provides ease of unmating. A hybrid passive 12x splitter cable is also available to support connectivity to three conventional CX4 end-nodes, with active circuity embedded in the CX4 connector.


*Figure 6. The iPass Plus double-stacked connectors accept up to two 12x iPASS Plus connectors.*

## Integrated Cable Management

Each Sun DS 3456 chassis includes integrated cable management designed to route a maximum of 1,152 12x cables to either floor or ceiling datacenter cable management infrastructure. Shelves at the front of each line card guide cables around to the sides where a series of supportive armatures coax the cables 180 degrees toward the back of the chassis (Figure 7).


*Figure 7. Cable trays and cable support brackets guide 12x cables easily toward the rear of the chassis.*

Strategically mounted elbow cable hangers on the side of the chassis then direct cables toward the floor or ceiling for distribution.

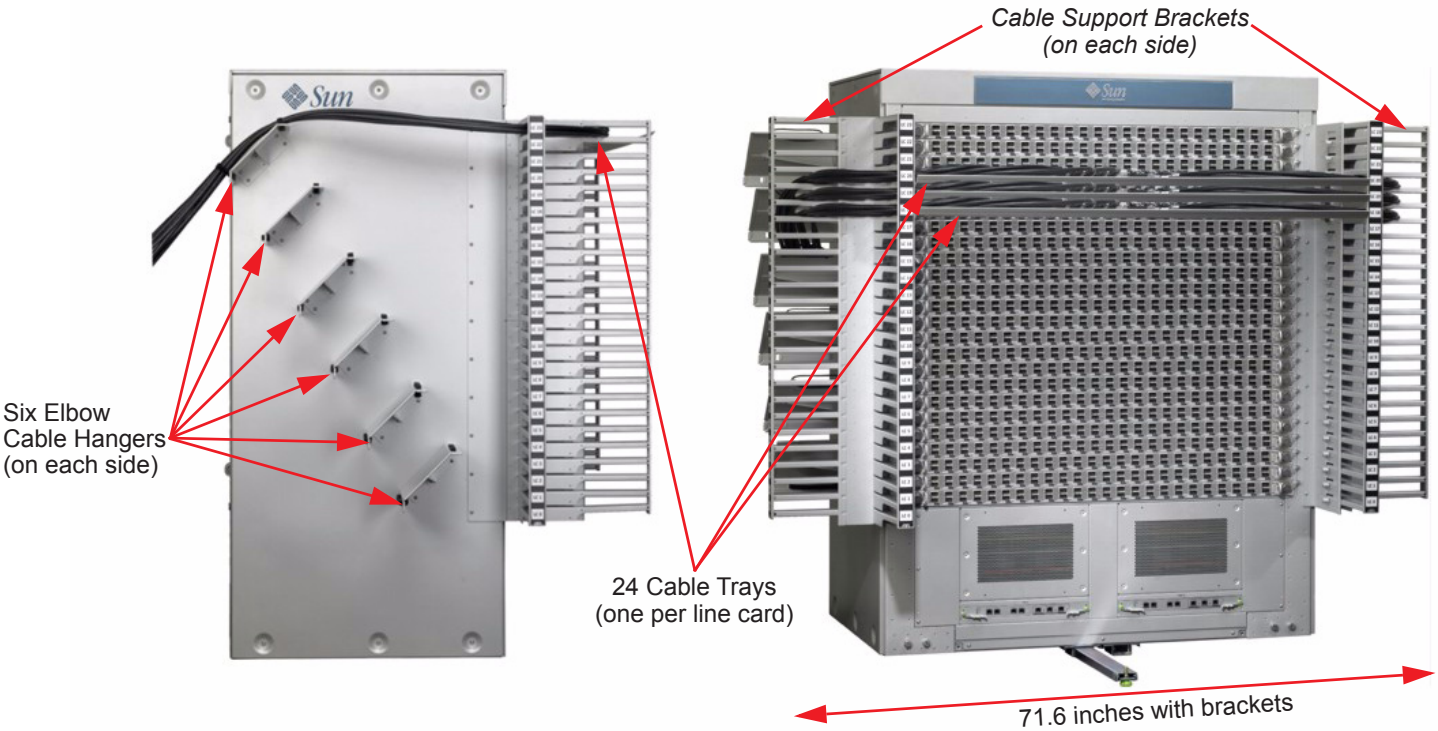



*Figure 8. Integral cable management directs cables around the chassis and either to the floor or ceiling for distribution to servers.*

## Routing

The primary challenge in an interconnect fabric is achieving high utilization of the topology in terms of throughput and latency. Different applications have different communication needs leading to a variety of communication patterns — all of which should be efficiently supported by the communication fabric. Using multiple paths is widely regarded as an efficient method for increasing network performance and providing some flexibility with regard to different traffic patterns.

Using a multiple path approach is especially useful for statistically non-blocking Clos networks. In addition, the ability to utilize multiple paths through the network yields a degree of fault tolerance. Assuming that the different paths are completely disjoint, any network fault may disconnect at most one of the paths. The Sun DS 3456 uses a multipath routing algorithm that provides multiple paths between communicating endpoints. Sun engineers have been able to show that they can consume and sustain in excess of 90 percent of the total fabric bandwidth of the core switch when fully populated.

## Congestion Management

Clos fabrics provide a statistically non-blocking fabric, which will maintain high utilization. At the same time, congestion is a known and expected issue in large interconnect fabrics, such as those provided by the Sun DS 3456. Usually a function of traffic patterns, fabric utilization, and sub-optimal topologies, congestion spreading in Clos fabrics is typically managed using limited buffers at each stage, which are quickly consumed during the onset of fabric congestion. Credit-based flow-control per virtual lane (VL) ensures no packet loss, but propagates congestion (tree saturation congestion spreading) such that other traffic on the same VL cannot pass.

The Sun DS 3456 implements Forward Error Correction with Notification, and Backward Error Control with Notification (FECN/BECN) for congestion control. This fairly straightforward communication mechanism works through communication between source and destination HCAs. The switch sets the FECN bit when detecting congestion. The destination HCA sends BECN back to the source HCA, and the source HCA then throttles the data rate. Traffic monitoring in the switch provides the switch buffer setting. Using FECN/BECN it has been demonstrated that high fabric utilization can be sustained during hot-spot traffic conditions.

## Power and Cooling Design

Most datacenter operators are now acutely aware of power consumption, both in terms of the systems they employ, and the heat they must handle to keep them operating efficiently. The Sun DS 3456 features front-to-back airflow, in keeping with popular hot/cold aisle configurations in many modern datacenters. Eight field-replaceable fan modules are attached to each of the 18 fabric cards for a total of 144 fan modules. Airflow moves from the line cards across the fabric cards, with a total airflow of 3,600 cubic feet per minute.

The Sun DS 3456 power system is comprised of up to 16 power supplies in two groups of eight. All of the power supplies furnish 3.3 volts of stand-by power to the system loads. The lower power domain (power supplies PS0 through PS7) and the upper power domain (power supplies PS8 through PS15) each provide 12 volt power to nine of the fabric card locations and twelve of the line card locations in the system. Each power supply has its own power cord and plugs into a power distribution board. The 3.3 volt power is distributed through the system midplane, bus bar structures are used to distribute the 12 volt power.

Each of the two power domains may include one or more redundant power supplies. For example, a maximized system may contain eight power supplies in each power domain while only seven operational power supplies are needed to power each domain (N+1 redundancy). The power domains themselves are not redundant and both power

domains must be operational for the system to operate. The number of installed power supplies depends on the number of configured line cards as shown in Table 4. All configurations require 18 fabric cards to implement the switch fabric.

*Table 4. Installed line cards and required number of power supplies*

| Power Supplies per Power Domain | Total Number of Power Supplies | Line Cards Supported |
|---|---|---|
| 3 | 6 | 2 |
| 4 | 8 | Up to 6 |
| 5 | 10 | Up to 10 |
| 6 | 12 | Up to 16 |
| 7 | 14 | Up to 20 |
| 8 | 16 | Up to 24 |

**Note –** Not all of the configurations described in Table 4 represent defined product offerings

## Component Architecture

The Sun DS 3456 is designed to efficiently provide the most InfiniBand connectivity in the most compact and efficient manner. This approach can eliminate multiple core switches, and literally thousands of leaf switches, cables, and connections found with other InfiniBand deployment approaches. The result is greatly reduced latency with significantly improved reliability and resiliency. All of this is accomplished with standard off-the-shelf components and an innovative architecture that puts them to their best use.

### Mellanox InfiniScale III DDR 24-port InfiniBand Switch Chip

The Sun DS 3456 employs the InfiniScale III 24-port Mellanox InfiniBand Switch chip on its fabric cards and line cards. Representing Mellanox's third generation InfiniBand silicon device, the device supports 24 4x ports or eight 12x ports, or a combination of both port types. The double data rate (DDR) version of the InfiniScale III supports 20 Gbps per 4x port and 60 Gbps per 12x port, delivering up to 960 Gbps of aggregate bandwidth.

The InfiniScale III switch chip architecture features an intelligent non-blocking packet switch design with an advanced scheduling engine that provides quality of service with switching latencies of less than 200 nanoseconds. The InfiniScale III architecture also scales efficiently to allow designers to create large core switches such as the Sun DS 3456 which can scale to support up 13,824 nodes in a single cluster. The combination of

unmatched bandwidth, inherent low latency design, and the reliability of InfiniBand fabrics makes the InfiniScale III an ideal building block . A block-level diagram of the InfiniScale III architecture is shown in Figure 9.

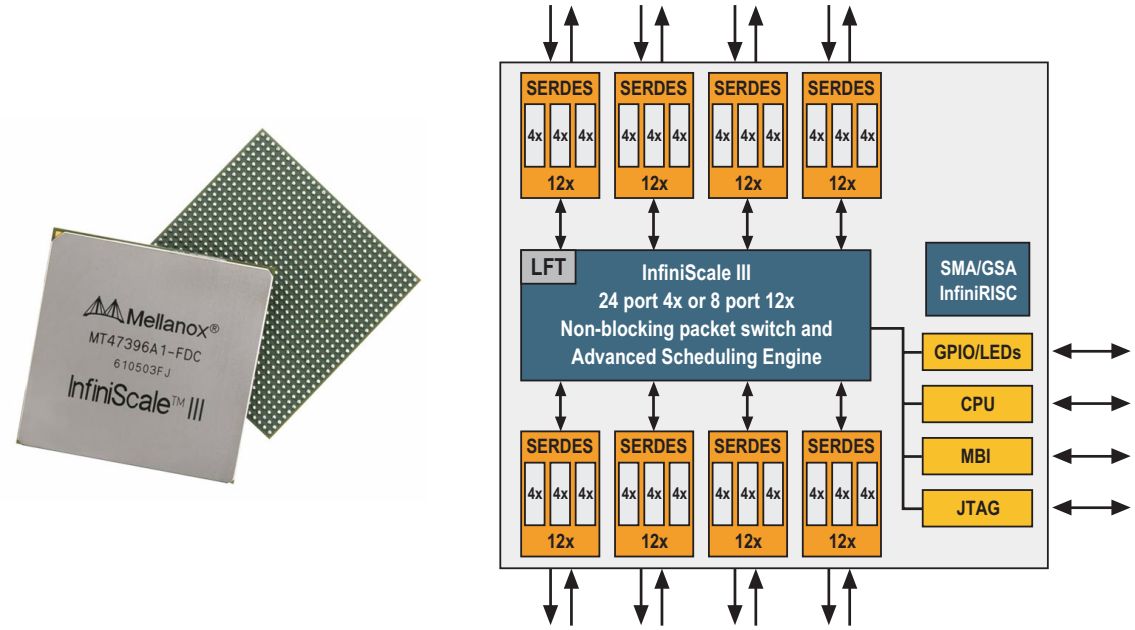



*Figure 9. Block-level diagram of the Mellanox InfiniScale III 24-port switch chip*

The InfiniScale III switch chip integrates ninety-six 5 Gbps SerDes interfaces in a single 961-ball package — reducing power, system cost, and PCB size and complexity. This density is particularly important for integration into the Sun DS 3456. Power efficiency is also very important, and each InfiniScale III typically consumes only 34 watts.

## Passive Midplane, Fabric Cards, and Line Cards

Focusing on density, the 36-inch x 40-inch Sun DS 3456 passive midplane provides 432 8 x 8 Molex iTRAC connectors arrayed in an 18 x 24 grid (Figure 10). These orthogonal connectors feature 64 signal pairs in a small 32.2 millimeter square connector.

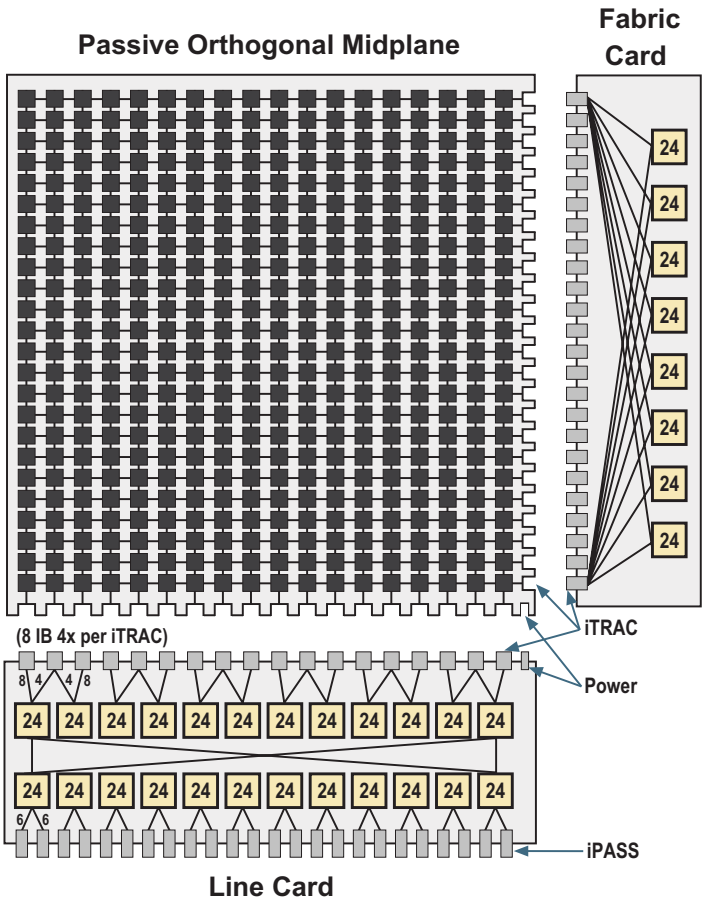



*Figure 10. ITRAC connectors join the fabric and line cards through the passive midplane while 12x iPASS connectors provide external connections along the edge of each line card*

Fabric cards install vertically and connect to the midplane from the rear of the chassis, and all 18 fabric cards must be installed for non-blocking behavior. Each fabric card connects eight InfiniScale III switch chips to the passive midplane. Up to 24 line cards install horizontally from the front of the chassis, and may be installed in subsets of the maximum. Each line card provides 24 InfiniScale III switch chips, 12 interfacing to the passive midplane, and 12 interfacing to the 12x iPass connectors at the front of the line card. A total of 144 4x InfiniBand ports is provided by each line card, expressed as 48 physical 12x iPASS connectors. Both line cards and fabric cards also include an intelligent platform management interface (IPMI) controller for environmental chassis management and monitoring. Line cards are hot-swappable, and provide indication of logical link status via visible LEDs.

A three-dimensional perspective of the fabric provided by the core switch is shown in Figure 11 with an example route overlaid. With this dense switch configuration, InfiniBand packets traverse only five hops from ingress to egress of the switch, keeping

latency very low. The Sun Blade 6048 InfiniBand Switched NEM adds only two hops for a total of seven. All InfiniBand routing is managed using a redundant host-based subnet manager as described in Chapter 3.

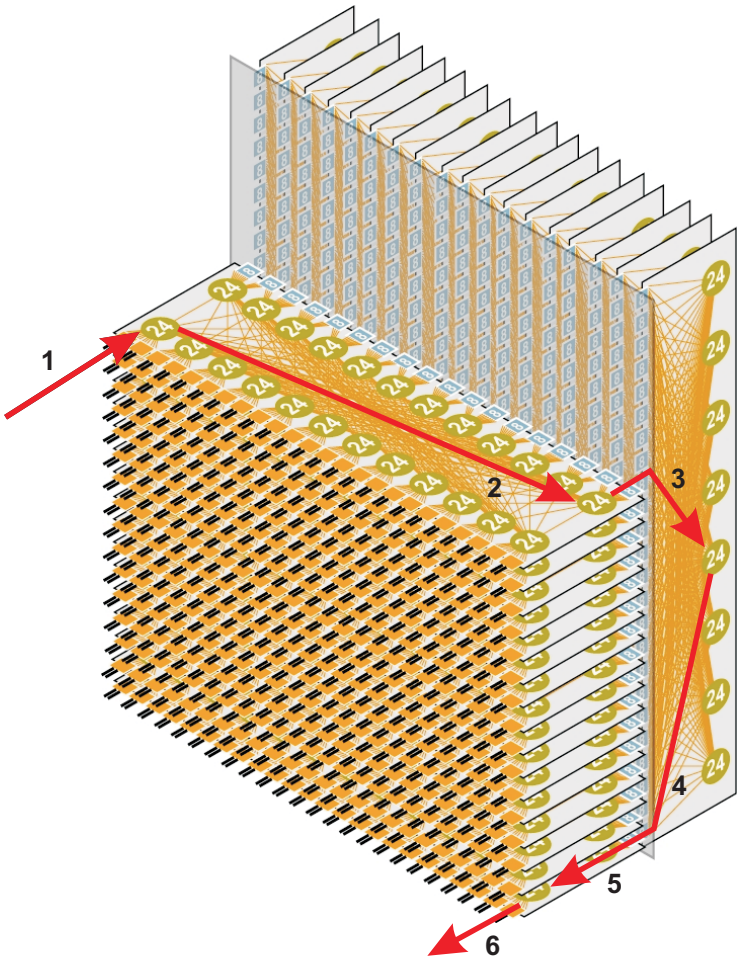



*Figure 11. A path through a Sun DS 3456 core switch connects two nodes across horizontal line cards, a vertical fabric card, and the passive orthogonal midplane.*

Chapter 3

# Integrated Fabric Control and Management

Managing the largest datacenters requires coordinated and comprehensive management technologies. With so many systems involved, compute nodes, interconnects, operating systems, and applications must all be managed effectively to help ensure that valuable resources aren't wasted or under-utilized. To this end, Sun provides comprehensive management tools in key areas, including:

- *Cluster Management* — Sun Grid Engine (formerly Sun N1™ Grid Engine) provides provisioning and scheduling of workloads onto cluster resources, communicating with per-node agents via the data network.
- *Hardware and System Management* — Sun xVM Ops Center (formerly Sun Connection and Sun N1 System Manager) provides "bare-metal" management of servers by interfacing with server service processors via the management network.
- *InfiniBand Switch and Fabric Management* — Sun Fabric Management software provides control and management for the Sun DS 3456 and integrated InfiniBand interfaces.

This chapter discusses InfiniBand management while cluster and system management tools are discussed in Chapter 5.

## InfiniBand Management System

Given the scale of the Sun DS 3456, and its pivotal role in the datacenter, a robust and fully-integrated InfiniBand control and management system is essential. In addition, scaling InfiniBand subnet management present significant technical challenges that must be met. Sun's management model has been to view the cluster as a single logical fabric infrastructure — one that provides a single compute resource.

### System Overview

The Sun DS 3456 provides a fully-featured management stack as illustrated in Figure 12. As a completely integrated system, coordinated management elements are provided at multiple layers of the system. Redundant management servers running the Solaris™ Operating System (Solaris OS) and Sun Fabric Management software are connected to the switch via both direct InfiniBand connections as well as Ethernet connections to the chassis management controller (CMC). Software elements running on the CMCs communicate with agents that reside within the switch itself.

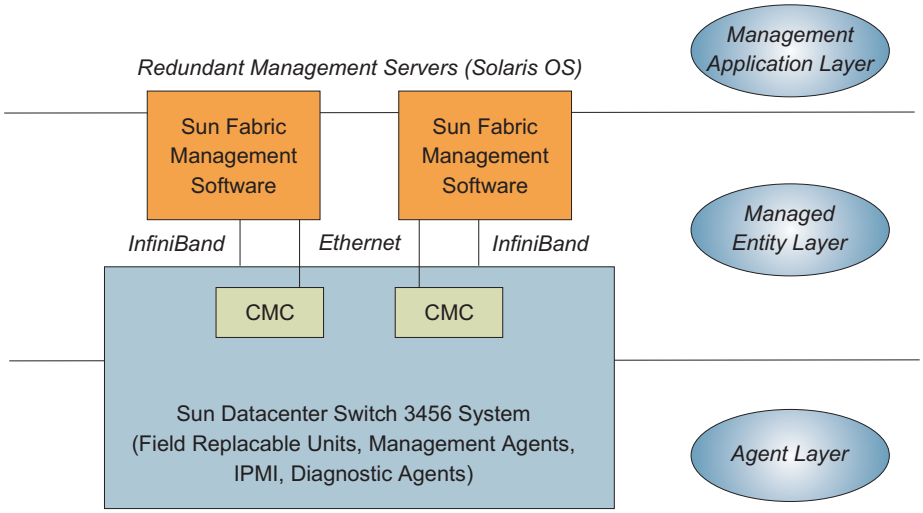



*Figure 12. Sun DS 3456 management stack*

## Scaling Requirements for InfiniBand Subnet Management

Subnet management and administration of InfiniBand clusters on the scale of those achievable with the Sun DS 3456 can present significant challenges. Given the sheer number of attached nodes, maintaining accurate and up-to-date information is difficult, and equally essential. Scalability challenges take a number of forms, including:

- *Effective routing* — The routing of the subnet must allow maximized throughput and minimized contention for arbitrary workloads and traffic patterns within the constraints imposed by the fabric topology.
- *State change handling* — Fabric initialization must converge in bounded time even with unstable links in the fabric, and state-change event processing and reporting during run-time must not prevent other operations from making reasonable progress.
- *Availability* — The Subnet Management and Subnet Administration functions must be highly available within the subnet. Clients must be able to fail-over communication to alternative paths/routes in case of failure, and path fail-over should not require Subnet Manager intervention.
- *Topology and configuration control* — Correct cabling and connectivity within the system must be automatically checked.
- *Real-time performance and error monitoring* — All relevant performance and error counters for all ports in the system must be monitored and aggregated in a timely manner without severely impacting basic Subnet Management/Administrator activity, and without impacting any client host operation.
- *Efficient link diagnosis* — It must be possible to observe state information from each side of a suspect link, and also to isolate communication problems to individual links (or nodes) in a timely manner.

The Subnet Manager and Subnet Administrator components of Sun Fabric Management software as well as the associated in-band and out-of-band fabric management components address key issues described above. These components are described in the sections that follow.

## Sun Fabric Management Software

Sun Fabric Management software provides a comprehensive management environment designed to provide connectivity and diagnosis throughout the Sun DS 3456. With Sun Fabric Management software, organizations can manage single or multiple switch chassis and thousands of interconnected nodes as a single system. Figure 13 provides an overview of the management infrastructure provided with Sun Fabric Management software.

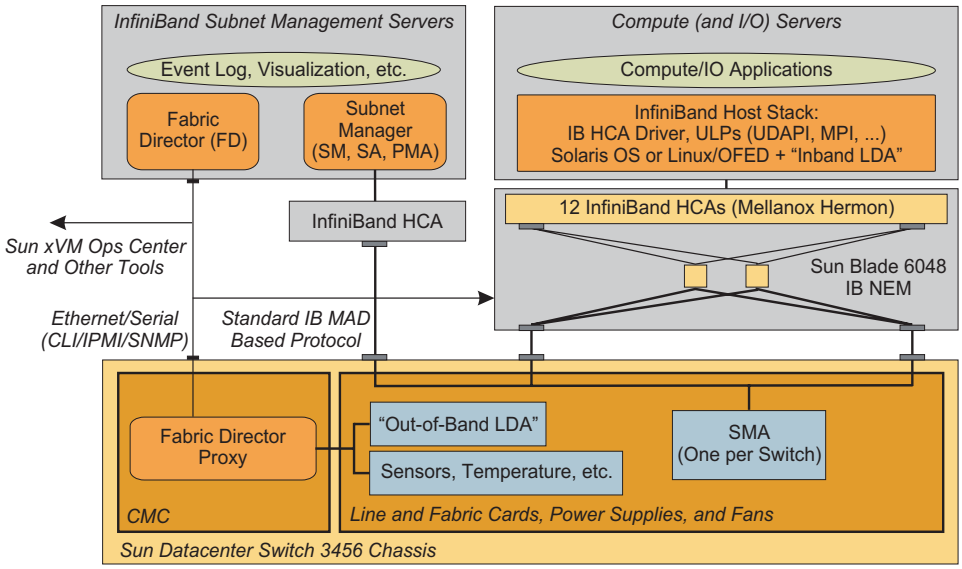



*Figure 13. System management overview showing component connections between Sun Fabric Management software, compute servers, and the Sun DS 3456*

Sun Fabric Management software is hosted on dedicated management servers running the Solaris OS. Both SPARC and x64 systems are supported as management servers. A single active master management server is defined with one or more hot standby systems that are available for failover purposes. Components of the Sun Fabric Management software running on the management server communicate with proxy components running on the chassis management controllers, and with agents associated with the various InfiniBand nodes (HCAs and switches) throughout the InfiniBand fabric. System requirements for the management servers, include:

- SPARC or x86/x64 CPU
- 8 GB of system RAM or higher
- Solaris 10 OS, release 11/06 or later

## Management Components

Sun Fabric Management software implements components that are defined by the InfiniBand specification, as well as Sun-specific fabric management components. These components are described in the sections that follow.

### Components Defined in the InfiniBand Specification

Sun Fabric Management software implements components defined in the InfiniBand specification. These components include:

- *Subnet Manager and Subnet Administrator*— The Subnet Manager (SM) and Subnet Administrator (SA) run on the active management server, operating on an InfiniBand HCA that is attached to the fabric as shown in Figure 13. The Subnet Manager is principally involved with configuring and routing the topology. The Subnet Administrator provides specific routes between nodes within the fabric.
- *In-band Discovery* — In-band discovery involves the discovery of arbitrary topologies using geographically addressed ("direct routed") Subnet Management Packets (SMPs). The Subnet Manager communicates with Subnet Manager Agents (SMAs) residing in individual HCAs and switches that provide basic error status and counters, before and after the associated links have become active.
- *Initialization* — To initialize the fabric, the Subnet Manager assigns LIDs (per-node Logical IDs) to end-nodes and performs routing (end-node-to-end-node routes via switches) using Subnet Manager specific policies and algorithms. Basic multi-cast routing is handled similarly. Routing is implemented by updating per-switch forwarding tables (LID to port mapping) using write requests to the switch Subnet Manager Agent (SMA).
- *Operational Subnet* — In order to establish communication between end-nodes, the relevant "path info" must be known. Normal data traffic is enabled when the Subnet Manager sets ports and links to an active state. Client nodes can learn about presence of other nodes and what "path" exists by sending (in-band) requests to the Subnet Administrator (SA).
- *Performance and Error Monitoring* — The per-node Subnet Manager Agent provides basic error status and counters before and after the associated links have become operational. The per-node Performance Management Agent (PMA) provides additional performance monitoring capabilities (per port packet counters, etc.)

### Sun Fabric Management Software Components

Sun Fabric Management software also provides additional components that help provide diagnostics and help to place the fabric in a state such that the Subnet Manager can start its discovery and initialization processes. These components include:

- *Fabric Director* — Co-located with the Subnet Manager in the management server, the Fabric Director (FD) communicates with a set distributed agents known as Fabric Director Proxies (FDPs). Communication can take place via the management network

or in-band on InfiniBand, keeping track of presence and state of fabric components and associated cabling and electrical links. The Fabric Director can check and verify that actual cabling of the fabric is in accordance with what has been intended for the system mechanical connectivity blueprint.

- *Fabric Director Proxy* — A Fabric Director Proxy (FDP) instance is associated with each switch chassis. In the Sun DS 3456, a specific Fabric Director Proxy operates on the Chassis Management Controller as an extension to the generic Shelf Manager software infrastructure. The FDP uses other chassis management functions to gather status, as well as embedded out-of-band connectivity to InfiniBand devices, links, cables and connectors.
- *In-Band Link Diagnostic Agent* — The In-Band Link Diagnostic Agent (LDA) can be considered a minimal Subnet Manager that uses Subnet Management Packets to track and test local link state and associated error counters. This functionality does not depend on an active Subnet Manager in the fabric. The LDA can also be used to monitor normal traffic using Performance Management Agent attributes in switches and HCAs (once the subnet has been made operational by a Subnet Manager). The LDA can test links by performing a sequence of Subnet Management Packets targeting either another LDA instance or a standard Subnet Management Agent. For instance, an LDA on a host node can operate against the SMA in the switch to which it is attached. An LDA instance can be associated with each host port in the system.
- *Out-of-Band Link Diagnostic Agent* — The Out-of-Band LDA is a generic term representing the complete embedded I2C and other out-of-band infrastructure for collecting and monitoring the presence and operational state of InfiniBand devices. These devices include both HCAs and switches, the state of associated ports and links, and the state IDs (serial numbers) of connected cables. In the Sun DS 3456, out-of-band infrastructure monitoring is implemented by the IPMI infrastructure within the Sun DS 3456 core switch, and using I2C infrastructure on the Sun Blade 6048 switched InfiniBand NEM.

## Chassis Management Controller Architecture

Each Sun DS 3456 chassis features two chassis management controllers that insert from the front of the chassis. The CMC runs the Fabric Director Proxy and provides access to IPMI sensors and out-of-band link diagnostic agents residing on the line cards and fabric cards. The CMC is based on the ShMM-1500 Shelf Management Mezanine card from Pigeon Point Systems, and features various system interfaces and external network and serial management interfaces.

Each CMC contains:

- ShMM-1500 Shelf Management Mezanine card from from Pigeon Point Systems deployed in a radial design
- Dual IPMI hubs/muxes, providing two sets of 42 connections for redundant IPMI communication with each of the 18 fabric modules and 24 line modules

- A dedicated CMC redundancy interface for communication with the other CMC for fail-over purposes
- Two 10/100Base-T RJ-45 and one serial RJ-45 management connections for external monitoring and communication with the Fabric director

In the Sun DS 3456, the two Chassis Management Controllers are configured in a redundant dual star architecture as shown in Figure 14.

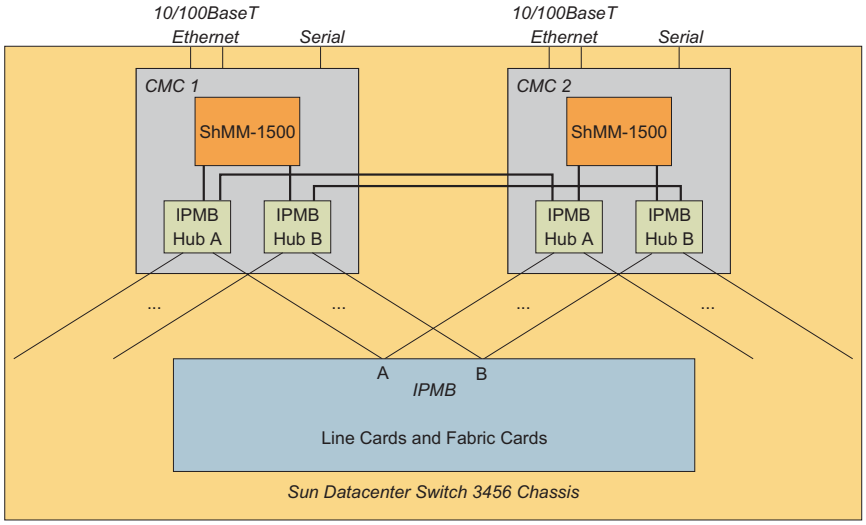



*Figure 14. The chassis management controllers in each Sun DS 3456 chassis are configured in a dual redundant star architecture, with both Intelligent Management Platform Bus (IPMB) interfaces on each CMC connected to each fabric card or line card.*

Chapter 4

# Deploying the Sun Datacenter Switch 3456

Deploying petascale supercomputing clusters can be a daunting challenge. Collecting and configuring thousands of high-performance systems into an InfiniBand network has historically resulted in considerable complexity. To combat this complexity, and to allow the largest supercomputing clusters to be deployed rapidly, Sun provides tight integration of the Sun Blade 6048 modular system with the Sun DS 3456. In fact, organizations such as the Texas Advanced Computing Center are already deploying these systems — building one of the largest supercomputing clusters in existence. This section describes the Sun Blade 6048 modular system and its InfiniBand connectivity, and provides some deployment examples.

## Sun Blade™ 6048 Modular System Overview

The Sun Blade 6048 modular system builds on the considerable success of the Sun Blade 6000 modular system, and provides access to server modules based on the latest SPARC, Intel, and AMD processors

### Extreme Density, Performance, and I/O

In order to provide the high computational and I/O density required for top-end supercomputing clusters, the Sun Blade 6048 modular system takes a innovative new approach. Rather than mounting multiple chassis in a datacenter rack, the custom Sun Blade 6048 chassis takes the footprint of a standard datacenter rack, providing four modular shelves. Each of the four modular shelves is self contained and independent, providing space for server modules, standard PCI Express I/O modules, management modules, fan modules, and power supply modules. In keeping with its datacenter design, the Sun Blade 6048 modular system supports 1+1 power grid redundancy.

By utilizing the lateral space that would otherwise be used for chassis mounting hardware, the Sun Blade 6048 chassis provides space for up to 12 server modules in each of its four shelves — for up to 48 Sun Blade 6000 server modules in a single chassis. This design approach provides considerable density. Front and rear perspectives of the Sun Blade 6048 modular system are provided in Figure 15.

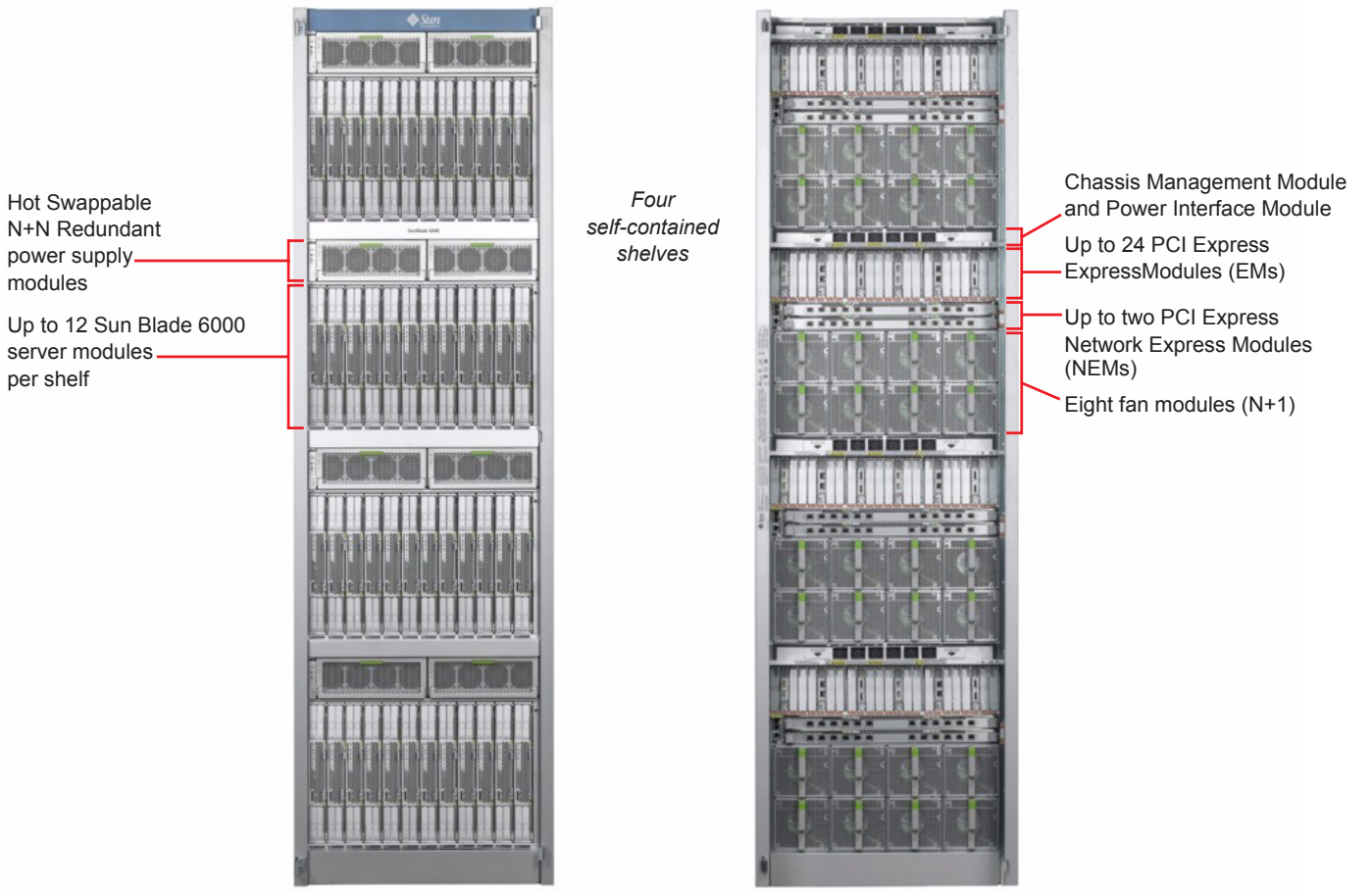



*Figure 15. Front and rear perspectives of the Sun Blade 6048 modular system*

With four self-contained shelves per chassis, the Sun Blade 6048 modular system houses a wide range of components.

- Up to 48 Sun Blade 6000 server modules insert from the front, including:
  - The Sun Blade T6320 server module based on the UltraSPARC T2 processor, the industry's first massively threaded system on a chip with four, six, or eight cores and support for up to 64 thread per processor and up to 64 GB of memory
  - The Sun Blade T6300 server module based on the UltraSPARC T1 processor, with support for six or eight cores and up to 32 threads per processor, and up to 32 GB of memory
  - The Sun Blade X6250 server module with two sockets for Dual-Core Intel Xeon Processor 5100 series or Quad-Core Intel Xeon Processor 5300 series CPUs, and up to 64 GB of memory
  - The Sun Blade X6220 server module with two sockets for Next Generation AMD Opteron 2000 Series processors, and up to 64 GB of memory

- A total of eight hot-swap power supply modules insert from the front of the chassis, with two 9000 W 12-volt power supplies (N+N) are provided for each shelf. Each power supply module contains a dedicated fan module.
- Up to 96 hot-plug PCI Express ExpressModules (EMs) insert from the rear of the chassis (24 per shelf), supporting industry-standard PCI Express interfaces with two EM slots configured for use by each server module.
- Up to eight PCI Express Network Express Modules (NEMs) can be inserted from the rear, with two NEM slots serving each shelf of the chassis. Four dual-height Sun Blade 6048 InfiniBand Switched NEMs can be installed in a single chassis (one per shelf).
- A chassis monitoring module (CMM) and power interface module are provided for each shelf. The CMM provides for transparent management access to individual server modules while the Power Interface Module provides six plugs for the power supply modules in each shelf.
- Redundant (N+1) fan modules are provided at the rear of the chassis for efficient front-to-back cooling.

## Sun Blade 6048 InfiniBand Switched NEM

Providing dense connectivity to servers while minimizing cables is one of the issues facing large HPC cluster deployments. The Sun Blade 6048 InfiniBand Switched NEM solves this challenge by integrating an InfiniBand leaf switch into a Network Express Module for the Sun Blade 6048 chassis. As a part of the Sun DS 3456, the NEM uses common components, cables, connectors, and architecture. A block-level diagram of the NEM is provided in Figure 16, aligned with an image of the back panel.

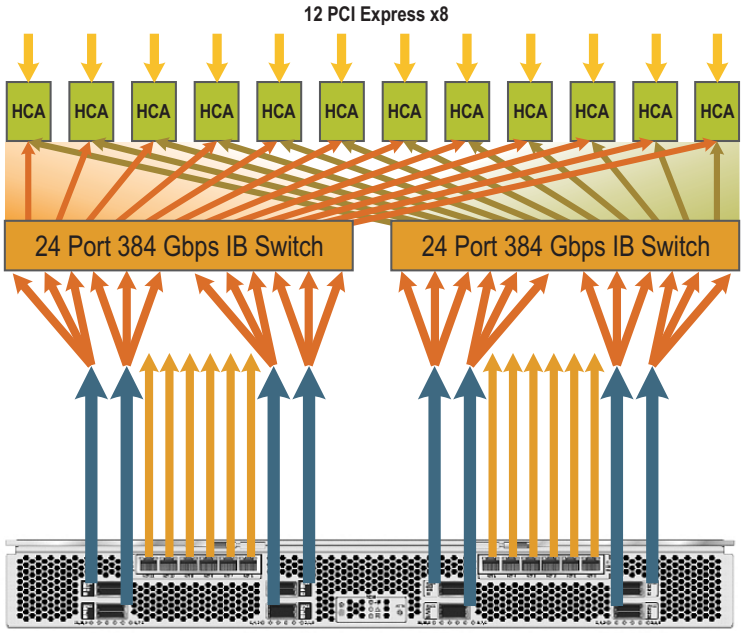



*Figure 16. The Sun Blade 6048 InfiniBand Switched NEM provides eight switched 12x InfiniBand connections to the two on-board 24-port switches, and twelve pass-through gigabit Ethernet ports, one to each Sun Blade 6000 server module*

Each Sun Blade 6048 InfiniBand Switched NEM employs two of the same Mellanox InfiniScale III 24-port switch chips used in Sun DS 3456 fabric and line cards. In this case, each switch chip provides 12 internal and 12 external connections. Redundant internal connections are provided from Mellanox ConnectX HCA chips to each of the switch chips, allowing the system to route around failed links. The same 12x iPass connectors are used on the back panel for direct connection to the Sun DS 3456. Additionally, 12 pass-through gigabit Ethernet connections are provided to access gigabit Interfaces provided on individual Sun Blade 6000 server modules mounted in the shelf.

## Scaling to Multiple Core Switches

Designers need the ability to scale supercomputing deployments without being constrained by arbitrary limitations. While clusters of up to 3,456 nodes may seem large, multiple Sun DS 3456 core switches can be combined to produce even larger configurations. As with single-switch configurations, a multiswitch system still functions and is managed as a single entity, greatly reducing management complexity.

- A single Sun DS 3456 can be deployed for configurations that require up to 3,456 server modules (nodes).
- Two core switches can be deployed to serve up to 6,912 servers, or to provide redundancy
- Four core switches can serve up to 13,824 server modules

To maintain a non-blocking fabric, each Sun Blade 6048 Switched InfiniBand NEM must connect via four 12x cables, independent of the number of switches. In a single-switch configuration, the four 12x cables connect to the one core switch. For configurations larger than 3,456 nodes, each Sun Blade 6048 Switched InfiniBand NEM connects to every core switch with either one or two 12x cables. Table 5 lists the connections, maximum supported servers, and Sun Blade 6048 modular systems supportable with one, two, and four Sun Datacenter 3456 core switches.

*Table 5. Maximum numbers of Sun Blade 6000 server modules and Sun Blade 6048 modular systems supported by various numbers of Sun DS 3456 core switches*

| Number of Core Switches | Maximum Nodes supported (Server Modules) | Maximum Sun Blade 6048 Modular Systems | 12x InfiniBand Cables per Shelf/Rack | Leaf to Core Trunking Connectivity (12x cables) |
|---|---|---|---|---|
| 1 | 3,456 | 72 | 4/16 | 4 to core switch (1,152 total) |
| 2 | 6,912 | 144 | 4/16 | 2 to each core switch (2,304 total) |
| 4 | 13,824 | 288 | 4/16 | 1 to each core switch (4,608 total) |

Configurations with multiple core switches also provide connection redundancy. For example, even if a core switch were to fail in a four-switch configuration, the remaining three core switches would provide full fabric connectivity, at a proportionally reduced throughput. Figure 17 illustrates a single Sun Blade 6048 modular system shelf connected to two core switches through the switched InfiniBand NEM, with a path traveling to an individual server module. The four ports to the second switch on the NEM remain available. As a result, a fully throughput-redundant configuration *could* be built with four core switches, albeit at the expense of connecting 144 additional Sun Blade 6048 racks.

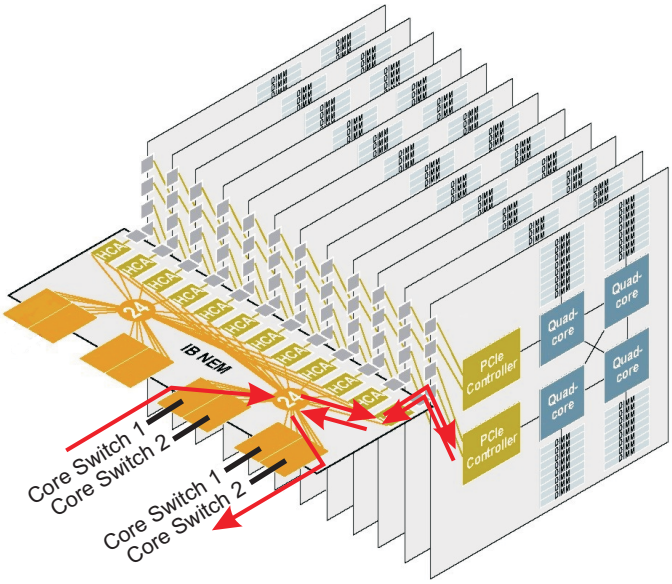



*Figure 17. The Sun Blade 6048 InfiniBand Switched NEM allows each Sun Blade 6000 server module to connect to multiple Sun DS 3456 core switches*

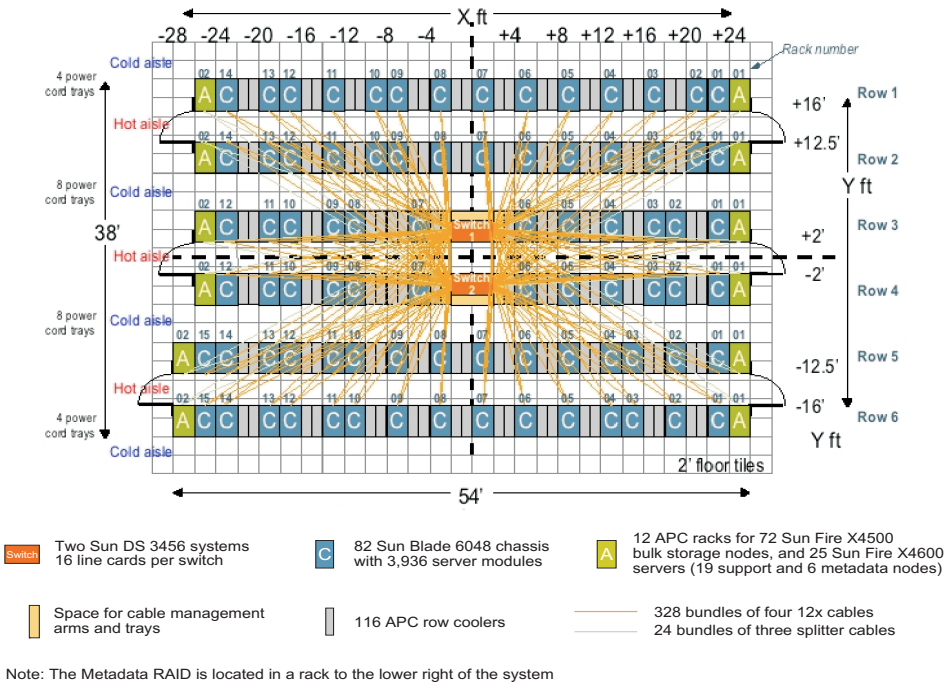
## A Massive Supercomputing Cluster at the Texas Advanced Computing Center

The Texas Advanced Computing Center (TACC) at The University of Texas at Austin is a leading research center for advanced computational science, engineering, and technology. TACC supports research and education programs by providing comprehensive advanced computational resources and support services to researchers in Texas, and across the nation. TACC is one of several major research centers participating in the TeraGrid, a program sponsored by the National Science Foundation (NSF) that makes high-performance computing, data management, and visualization resources available to the nation's research scientists and engineers.

As a part of the TeraGrid program, the NSF in mid-2005 issued a request for bids on a project to configure, implement, and operate a new supercomputer with peak performance in excess of 400 teraflops, making it one of the most powerful

supercomputer systems in the world. The new supercomputer will also provide over 100 terabytes of memory, and 1.7 petabytes of disk storage. TACC is partnering with Sun and using the Sun Constellation System to build the Ranger supercomputing cluster. Figure 18 illustrates the initial deployment, consisting of:

- Two Sun DS 3456 core switches with 16 line cards per switch
- 82 Sun Blade 6048 racks featuring 3,936 Sun Blade 6000 server modules
- 12 APC racks for 72 Sun Fire X4500 servers acting as bulk storage nodes, 19 Sun Fire X4600 servers acting as support nodes, and six Sun Fire X4600 metadata nodes
- 116 APC row coolers and doors to facilitate an efficient hot/cold isle configuration




Note: The Metadata RAID is located in a rack to the lower right of the system

*Figure 18. TACC initial deployment floor-plan featuring dual Sun DS 3456 core switches and 82 fully-populated Sun Blade 6048 modular systems*

Chapter 5

# Software for Standards-Based Supercomputers

Given the scope and cost of supercomputing clusters, adherence to standards is essential, both to help assure interoperability, and to get the most life out of computational, interconnect, and storage investments. In addition to the Sun DS 3456, Sun provides a wealth of standards-based tools designed to help datacenters manage both systems and computational resources. Sun also provides innovative visualization software to help scientists and engineers see and share the information behind their complex data.

## Sun Grid Engine

Sun Grid Engine (formerly Sun N1 Grid Engine) distributed resource management (DRM) software provides job scheduling and policy automation, unifying network-connected servers, workstations, and desktops while providing user access to the grid along with essential administrative and management interfaces. Computing tasks or jobs are distributed across the grid in accordance with resource requirements for the job, user requests for the job, and administrative and managerial policies. Usage accounting data is stored and made available so that it is possible to determine what resources were used in the execution of a job and for whom the job was run. Sun Grid Engine administrative overhead is quite low, letting overall resource utilization approach 100 percent (depending upon the optimization of resources to the workload).

Sun Grid Engine software can be used to create grids at both a department level and at a multi-department or enterprise level. Through the pooling of departmental resources into larger enterprise grids, multiple users, teams, and departments can share common resources while working on different projects with different goals and schedules, providing maximum resource availability to all users in a flexible, policy-based environment. Grid Engine software also offers heterogeneous support, enabling grids to be built from a wide variety of hardware systems and operating environments and spreading the benefits of grid computing across the entire IT infrastructure.

Distributed resource management (DRM) software and its batch queuing mechanism are the basic required element of a Grid Computing environment. As with traditional batch environments, in the normal operation of a compute grid, jobs are queued until required resources are available. DRM further enhances traditional batch queuing by monitoring host computers in the grid for proper balancing and load conditions.

Several different types of systems comprise a typical Sun Grid Engine implementation:

- ***Master host*** — A single host is selected to be the Sun N1 Grid Engine master host. This host handles all requests from users, makes job scheduling decisions, and dispatches jobs to execution hosts.
- ***Execution hosts*** — Systems in the grid that are available to execute jobs are called execution hosts.
- ***Submit hosts*** — Submit hosts are machines configured to submit, monitor, and administer jobs, and to manage the entire grid.
- ***Administration hosts*** — Managers use administration hosts to make changes to the grid configuration, such as changing DRM parameters, adding new compute nodes, or adding or changing users.
- ***Shadow master host*** — While there is only one master host, other machines in the cluster can be designated as shadow master hosts to provide greater availability. A shadow master host continually monitors the master host, and automatically and transparently assumes control in the event that the master host fails. Jobs already in the grid are not affected by a master host failure.

Jobs submitted to the master host are held in a spooling area until the scheduler determines that they are ready to run. Sun Grid Engine software then matches available resources to job requirements, such as available memory, CPU speed, and available software licenses. The requirements of different jobs may vary greatly, and only certain hosts may be able to provide the corresponding service. As soon as a resource becomes available for execution of a new job, Sun Grid Engine software dispatches the job with the highest priority and matching requirements.

The Sun Grid Engine software scheduler takes into account the order the job was submitted, what machines are available, along with weighing both policy and the priority of the job. This tool gives administrators a great deal of flexibility, providing fine-grained control in managing an organization's resources. General software flow is illustrated in Figure 19.

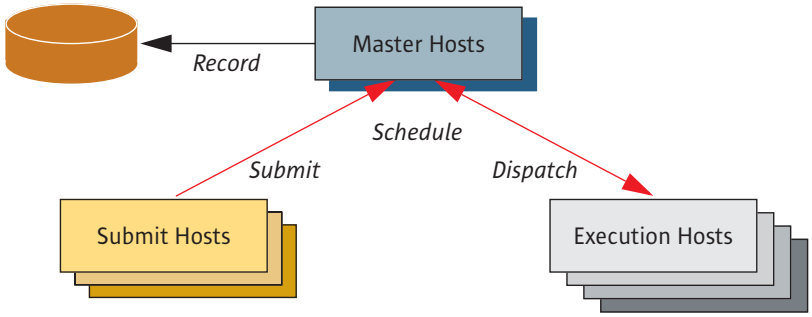



*Figure 19. Software flow in Sun Grid Engine distributed resource management software*

Jobs execute in Sun Grid Engine according to the following steps:

1. *Job submission* — When a user submits a job from a submit host, the job submission request is sent to the master host.
2. *Job scheduling* — The master host determines the execution host where the job will be assigned, assessing the load, checking for licenses, and evaluating any other job requirements.
3. *Job execution* — After obtaining scheduling information, the master host then sends the job to the selected execution host. The execution host saves the job in a job information database and initializes a shepherd process that starts the job and waits for completion.
4. *Accounting information* — When the job is complete, the shepherd process returns the job information, and the execution host then reports the job completion to the master host and removes the job from the job information database. The master host updates the job accounting database to reflect job completion.

To provide security and control access to the grid, the Sun Grid Engine master host maintains information about eligible submit and administration hosts. Only systems which have been explicitly listed as eligible submit hosts are able to submit jobs to the grid. Similarly, only systems which have been added to the list of eligible administration hosts can be used to modify the grid configuration. For example, to provide increased security for the grid, administration hosts can be limited to only those hosts in a physically-secured computer room.

Sun Grid Engine software can also be configured with one or more shadow master hosts, eliminating the master host as a single point of failure and providing increased availability to users. If the master becomes unavailable, the shadow master host automatically and transparently takes over as the master. Jobs already running on an execution host are not affected by a master-host failure, and users gain increased reliability and availability. This functionality is a fully-integrated part of Sun Grid Engine software, and the only prerequisite for its use is a highly-available file system on which to install the software and configuration files. In addition, Sun Cluster agents are available for Sun Grid Engine software, enabling highly-available configurations to be built from multiple clustered servers. Sun Cluster configurations provide automatic failover of Sun Grid Engine services.

## Sun xVM Ops Center

Beyond local and remote management capabilities, datacenter infrastructure needs to be agile and flexible, allowing not only fast deployment but streamlined redeployment of resources as required. Sun xVM Ops Center technology (formerly Sun N1 System Manager and Sun Connection) provides an IT infrastructure management platform for integrating and automating management of thousands of heterogeneous systems as a

single system. To improve life cycle management and change management, Sun xVM Ops Center supports management of applications and the servers on which they run, including the Sun Blade 6048 modular system.

Sun xVM Ops Center simplifies infrastructure life cycle management by letting administrators perform standardized actions across logical groups of systems. Sun xVM Ops Center can automatically discover and group bare-metal systems, performing actions on the entire group as easily as operating on a single system. Sun xVM Ops Center remotely installs and updates firmware and operating systems, including support for:

- Solaris 8, 9, and 10 on SPARC systems
- Solaris 10 on x86/x64 platforms
- Red Hat and SuSE distributions

In addition, the software provides considerable lights-out monitoring of both hardware and software, including fans, temperature, disk and voltage levels — as well as swap space, CPU utilization, memory capacity, and file systems. Role-based access control lets IT staff grant specific management permissions to specific users. A convenient hybrid user interface integrates both a command-line interface (CLI) and an easy-to-use graphical user interface (GUI), providing remote access to manage systems from virtually anywhere.

Sun xVM Ops Center provides advanced management and monitoring features to the Sun Blade 6048 modular system. The remote management interface discovers and presents the Sun Blade server modules in the enclosure as if they were individual servers. In this fashion, the server modules appear in exactly the same way as individual rackmount servers, making the same operations, detailed inventory, and status pages available to administrators. The server modules are discovered and organized into logical groups for easy identification of individual modules, and the system chassis and racks that contain them. Organizing servers into groups also allows features such as OS deployment across multiple server modules. At the same time, individual server modules can also be managed independently from the rest of the chassis. This flexibility allows for management of server modules that may have different requirements than the other modules deployed in the same chassis.

Some of the functions available through Sun xVM Ops Center software include operating system provisioning, firmware updates (for both the BIOS and ILOM service processor firmware), and health monitoring. In addition, Sun xVM Ops Center includes a framework allowing administrators to easily access inventory information, and simplify the task of running jobs on multiple servers thanks to the server grouping functionality.

## Sun Visualization System

Along with scalable performance from high performance computing (HPC) systems and clusters, visualization plays a valuable role in helping scientists, researchers and professionals from a wide range of disciplines understand their data in visual terms. Visualization technology is increasingly indispensable to the solution of the most important and complex problems. By simplifying and distilling information for human perception, visualization is aiding discovery and decision making in a host of scientific and engineering domains. Important new opportunities for visualization are also emerging in financial markets, national security, and public health.

The Sun Visualization System takes advantage of high-speed low-latency interconnects such as InfiniBand to allow effective clustering of visualization systems to address scalability. Innovative software now facilitates the scaling of standard OpenGL applications across multiple systems and graphics accelerators, and the sharing of those applications to multiple heterogeneous systems across the network. To this end, the Sun Visualization System includes both *Sun Scalable Visualization software* and *Sun Shared Visualization software*.

Visualization systems can be complex to design and build. The necessary integration of networks, high-speed interconnects, computing systems, graphics accelerators, and display technologies along with software selection and configuration can be daunting and time consuming. Unfortunately, as the industry has consolidated around commercial off-the-shelf (COTS) graphics and open source software, much of the burden of designing and building visualization systems has been pushed back onto the organizations that use them. Sun's approach addresses these issues by offering:

- Tightly integrated software components built to work together
- Highly configurable solutions to fit a wide range of needs
- Integration, installation, and configuration of systems, graphics, networks, and software into pre-cabled racks that are ready to deploy
- Design and integration with other key visualization equipment such as projectors, switchers, and haptic devices by partnering with certified display integrators, Sun CRS, and Sun's professional services group

Sun has considerable visualization experience and expertise, having designed and deployed both scalable and shared visualization systems for a variety of diverse purposes around the world. For instance, in addition to the Sun DS 3456, Sun has implemented a grid-enabled interactive visualization system at TACC. Now the Sun Visualization System brings the expertise gained from that experience and others to help take the time and complexity out of deploying visualization solutions.

The principal components of the Sun Visualization System include:

- *Sun Scalable Visualization software*, including tuned and supported versions of ParaView, Open Scene Graph, and Chromium for a wide variety of integrated multisystem visualization clusters constructed from multidisplay Sun x64 workstations and servers as well as single systems with multiple graphics cards
- *Sun Shared Visualization software*, providing tuned and supported versions of VirtualGL, Virtual Network Computing (TurboVNC), and extensive Sun Grid Engine software integration to aid with allocating and managing visualization resources on both SPARC and x64 systems
- *High-performance graphics accelerators*, such as NVIDIA graphics solutions for Sun x64 workstations and servers including the NVIDIA Quadro Plex Visual Computing System (VCS)
- *Powerful Sun SPARC and x64 based workstations and servers,* including those based on the latest SPARC and x64 processors.
- *High-speed, low-latency InfiniBand interconnects* including support for the Sun DS 3456 as well as supported InfiniBand switches from third parties such as Voltaire and Cisco

Chapter 6
# Conclusion

The Sun DS 3456 offers massive simplification and savings, and effectively removes many arbitrary limitations placed by complex conventional InfiniBand switching and cabling infrastructure. Implemented as a massive Clos-topology InfiniBand core switch, the Sun DS 3456 can provide up to 3,456 4x InfiniBand ports and up to 55 tbps of bi-sectional bandwidth. Together with the Sun Blade 6048 InfiniBand Switched NEM, up to four core switches can be combined to connect up to 13,824 nodes — allowing InfiniBand fabrics to scale simply and predictably.

By providing the world's largest non-blocking InfiniBand core switch, the Sun DS 3456 fundamentally improves large-cluster fabric deployments beyond traditional approaches. An innovative and robust new 12x connector and cable helps the system drastically reduce the number of switches required for large supercomputing installations, and provides a six-fold reduction in cabling. Organizations deploying the Sun DS 3456 can literally avoid hundreds of switches, and thousands of cables. Hot-swap power and cooling and a redundant external subnet manager let the Sun Datacenter Switch 3456 provide reliability and economies of scale for managing large InfiniBand deployments.

As an integral part of the Sun Constellation System, the Sun DS 3456 switch integrates tightly with the Sun Blade 6048 modular system and Sun Fire X4500 storage clusters to provide effective and scalable supercomputing infrastructure. Designed to work together, these systems redefine fabric, computational, and storage densities in compelling new ways. Systems built with the Sun Constellation System can scale from hundreds of teraflops to petaflops — proving that standards-based supercomputing on a massive scale is more feasible than ever.

## For More Information

To learn more about Sun products and the benefits of the Sun DS 3456, contact a Sun sales representative, or consult the related documents and Web sites listed in Table 6.

*Table 6. Related Websites*

| Web Site URL | Description |
|---|---|
| sun.com/ds3456 | Sun Datacenter Switch 3456 |
| sun.com/blades/6000 | Sun Blade 6000 and 6048 modular systems |
| sun.com/servers/x64/x4500 | Sun Fire X4500 server |
| sun.com/featured-articles/2007-0626/feature/index.jsp | Sun Constellation System |

**Sun Microsystems, Inc.** 4150 Network Circle, Santa Clara, CA 95054 USA **Phone** 1-650-960-1300 or 1-800-555-9SUN (9786) **Web** sun.com



SunWIN #: 515536 Printed in USA 11/07