PATHWAYS TO PETASCALE COMPUTING

The Sun™ Constellation System — Designed for Performance

White Paper
February 2008

"Make everything as simple as possible, but not simpler"

— Albert Einstein

Table of Contents

Executive Summary

From weather prediction and global climate modeling, to minute sub-atomic analysis and other grand-challenge problems, modern supercomputers often provide the key technology for unlocking some of the most critical challenges in science and engineering. These critical scientific, economic, and environmental issues are complex and daunting — and many require answers that can only come from the fastest available supercomputing technology. In the wake of the industry-wide migration to terascale computing systems, a predictable path to petascale supercomputing environments has become essential.

Unfortunately, the design, deployment, and management of very large terascale and petascale clusters and grids has remained elusive and complex. In fact, most existing terascale architectures are unlikely to reach petascale for fundamental reasons — not because of inherent limitations, but due to practicalities of attempting to scale those architectures to their full potential. Seemingly simple concerns — heat, power, cooling, cabling, and weight — are rapidly overloading the vast majority of even the most modern datacenters. Sun understands that the key to building petascale supercomputers lies in a balanced and systemic design approach to infrastructure, along with careful application of the latest technology advancements.

Derived from Sun's innovative design approach and experience with very large supercomputing deployments, the Sun Constellation System provides the world's most scalable HPC computing architecture — built entirely with open and standard hardware and software technologies. Cluster architects can use the Sun Constellation System to design and rapidly deploy tightly-integrated, efficient, and cost-effective supercomputing grids and clusters that scale predictably from a few teraflops to over a petaflop. With a totally modular approach, processors, memory, interconnect fabric, and storage can all be scaled independently depending on individual needs.

Best of all, the Sun Constellation System is an enterprise-class Sun-supported offering that leverages general-purpose computing, interconnects, and storage components that can be rapidly deployed[1]. In fact, existing supercomputing grids and clusters are already being built using the system. For instance, the Texas Advanced Computing Center (TACC) at the University of Texas at Austin is partnering with Sun to install the Sun Constellation system as their Ranger supercomputing cluster — expected to have a peak performance rating of over 500 teraflops when complete[2]. This document describes the key challenges and constraints involved in the build out of petascale supercomputing architectures, including network fabrics, multicore modular compute systems, storage, and general-purpose I/O.

1.In 2006, Sun helped the Tokyo Institute of Technology deploy the TSUBAME grid in only 35 days, initially featuring 47.38 teraflops and 1.1 petabytes of storage.
2.Some of the components of the TACC installation are based on products which Sun has not yet announced.

Chapter 1

Pathways to Petascale Computing

Most practitioners in today's high performance computing (HPC) marketplace would readily agree that the industry is well into the age of terascale systems. Supercomputing systems capable of processing multiple teraflops are becoming commonplace. These systems are readily being built using mostly commercial off-the-shelf (COTS) components with the ability to address terabytes of storage, and more recently, terabytes of system memory (generally as distributed shared memory and storage pools, or even as a single system image at the high end).

Only a few years ago, general-purpose terascale computing clusters constructed of COTS components were hard to imagine. Though they were on several industry road-maps, such systems were widely regarded as impractical due to limitations in the scalability of the interconnects and fabrics that tie disparate systems together. Through competitive innovation and the race to be the fastest, the industry has been driven into the realm of practical and commercially-viable terascale systems — and now to the edge of pondering what similar limitations, if any, lie ahead in the design of petascale systems.

For many that have been in the industry since the turn of the previous millennium, just contemplating the throughput implications of a petaflop of compute capacity can be simply overwhelming. The existence of a single file larger than a petabyte and the amount of refinement needed to deliver a petabit's worth of I/O in a single second, truly challenge the boundaries of what was thought possible fewer than 10 years ago. For those who have come into the industry as it turns its attention away from terascale and more towards petascale design challenges, such boundaries seem foolish. After all, terabyte disk drives are now relatively inexpensive consumer-grade items. These observations and economies of scale serve multiple purposes — to remind the industry not to accept arbitrary limitations, and to inspire the design of systems of even greater capacity and scalability.

The Unstoppable Rise of Grids and Clusters

In the last five years, technologies used to build the world's fastest supercomputers have evolved rapidly. In fact, grids and clusters of smaller interconnected rackmount and blade systems now represent a majority of the supercomputers on the Top500 list of supercomputing sites[1] — steadily replacing vector supercomputers and other large

1.www.top500.org

systems that dominated previously. Figure 1 shows the relative shares of various supercomputing architectures comprising the Top500 list from 1993 through 2007, establishing clear acceptance of clusters as leading supercomputing technology.

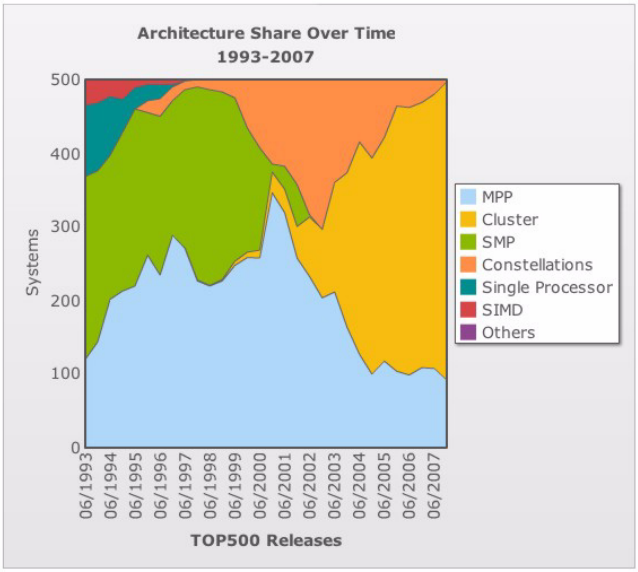

Figure 1. In the last five years, clusters have increasingly dominated the Top500 list architecture share (image courtesy www.top500.org)

Not only have clusters and grids provided access to supercomputing resources for larger and larger groups of researchers and scientists, but the largest supercomputers in the world are now built using cluster architectures. This trend has been assisted by an explosion in performance, bandwidth, and capacity for key technologies, including:

- Faster processors, multicore processors, and multisocket rackmount and blade systems
- Inexpensive memory and system support for larger memory capacity
- Faster standard interconnects such as InfiniBand
- Higher aggregated storage capacity from inexpensive commodity disk drives

Unfortunately, significant challenges remain that have stifled the growth of true petascale-class supercomputing clusters and grids. Time-to-deployment constraints have resulted from the complexity of deploying and managing large numbers of compute nodes, switches, cables, and storage systems. The programability of extremely large clusters remains an issue. Environmental factors too are paramount since deployments must often take place in existing datacenter space with strict constraints on physical footprint as well as power and cooling.

In addition to these challenges, most petascale computational users will also have unique requirements for clustered environments beyond those of less demanding HPC customers, including:

- *Scalability at the socket and core level* — Some have espoused large grids of relatively low-performance systems, but lower performance only increase the number of nodes that are required to solve very large computational problems.
- *Density in all things* — Density is not just a requirement for compute nodes, but for interconnect fabrics and storage solutions as well.
- *A scalable programming and execution model* — Programmers need to be able to apply their programmatic challenges to massively scalable computational resources without special architecture-specific coding requirements.
- *A lightweight grid model* — Demanding applications need to be able to start thousands of jobs quickly, distributing workloads across the available computational resources through highly-efficient distributed resource management systems.
- *Open and standards-based solutions* — Programmatic solutions must not cause extensive porting efforts, or be dedicated to particular proprietary architectures or environments, and datacenters must remain free to purchase the latest high-performance computational gear without being locked into proprietary or dead-end architectures.

The Importance of a Balanced and Rigorous Design Methodology

As anyone who has witnessed prior generations of supercomputing and high performance computing architectures can attest, scaling gracefully is not simply a matter of accelerating systems that already perform well. Bigger versions of existing technologies are not always better. Regrettably, the pathways to teraflops are littered with the products and technologies of dozens of companies that simply failed to adapt along the way.

Many technologies have failed because the fundamental principles that worked in small clusters simply could not scale effectively when re-cast in a run-time environment thousands of times larger or faster than their initial implementations. For example, ten gigabit Ethernet — though a significant accomplishment — is known in the supercomputing realm to be fraught with sufficiently variable latency as to make it impractical for situations where low guaranteed latency and throughput dominate performance. Ultimately building petascale-capable systems is about being willing to fundamentally rethink design, using the latest available components that are capable of meeting or exceeding specified data rates and capacities.

Put simply, getting to petascale requires balance and massive scalability in all dimensions, including scalable tools and frameworks, processors, systems, interconnects, and storage. Key challenges include:

- Keeping floating-point operations (FLOPs) to memory bandwidth ratios balanced to minimize the effects of memory latency (with each FLOP representing at least two loads and one store)
- Allowing for the practical scaling of the interconnect fabric to allow the connection of tens of thousands of nodes
- Exploiting the considerable investment, momentum, and cost savings of commodity multi-core x64 processors and tools
- Architecting to account for the opportunity to take advantage of external floating point, vector, and/or general purpose processing on graphics processing unit (GPGPU) solutions within a cluster framework
- Designing the highest levels of density into compute nodes, interconnect fabrics, and storage solutions in order to facilitate large and compact clusters
- Building systems with efficient power and cooling to accommodate the broadest range of datacenter facilities and to help ensure the highest levels of reliability.

These challenges serve as reminders that the value of genuine innovation in the marketplace must never be underestimated — even as design cycle times shrink and the pressures of time to market grow with the demand for faster, cheaper and standards based solutions.

The Sun Constellation System

Since its inception, Sun has been focused on building balance and even elegance into its system designs. The Sun Constellation System represents a tangible application of this approach on a grander scale — in the form of a systematic approach to building petascale supercomputing clusters. Specifically, the Sun Constellation System delivers an open architecture that is designed to allow organizations to build clusters that scale seamlessly from teraflops to petaflops of performance.

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 elements. As a systems company, Sun looks beyond existing technologies toward solutions that minimize the simultaneous equations of cost, space, practicality, and complexity. In the form of the Sun Constellation System, this systemic focus combines a massively-scalable InfiniBand interconnect with very dense computational and storage solutions — in a single architecture that functions as a cohesive system. Organizations can obtain all of these tightly-integrated building blocks from a single vendor, and benefit from a unified management approach.

Components of the Sun Constellation System include:

- The Sun Datacenter Switch 3456 (Sun DS 3456), currently the world's largest standards-based InfiniBand core switch with support for up to 3,456 nodes per switch, and up to 13,824 nodes with multiple core switches
- The Sun Blade™ 6048 modular system, providing an ultra-dense InfiniBand-connected blade platform with support for 48 multiprocessor, multicore Sun Blade 6000 server modules in a custom rack-sized chassis
- Sun Fire™ X4500 storage clusters, serving as an economical InfiniBand-connected parallel file system building block, with support for up to 48 terabytes in only four rack units and up to 480 terabytes in a single rack,
- A comprehensive and open HPC software stack, encompassing integrated developer tools, Sun Grid Engine infrastructure, provisioning, monitoring, patching, and simplified inventory management.

The Sun Constellation System provides an open systems supercomputer architecture designed for petascale computing — in the form of an integrated and Sun-supported product. This holistic approach offers key advantages to those designing and constructing the largest supercomputing clusters:

- Massive scalability in terms of optimized compute, storage, interconnect, and software technologies and services
- A dramatic reduction in complexity through integrated connectivity and management to reduce start-up, development, and operational connectivity
- Breakthrough economics from technical innovation that results in fewer more reliable components and high-efficiency systems in a tightly-integrated solution

Along with key technologies and the experience of helping design and deploy some of the world's largest supercomputing clusters, these strengths make Sun an ideal partner for delivering high-end terascale and petascale architecture.

Chapter 2

The World's First Non-Blocking InfiniBand Petascale Switch

Building the largest supercomputing grids provide significant challenges, with fabric technology paramount among them. Sun set out to design the Sun DS 3456 for maximum fabric scalability, and to drastically reduce the cost and complexity of delivering large-scale HPC solutions. Achieving these goals required a delicate balancing act — one that weighed the speed and number of nodes along with a sufficiently fast interconnect to provide minimal and predictable levels of latency.

The Fabric Challenge

For many applications, the interconnect fabric is already the element that limits performance. One unavoidable driver is that faster processors require a faster interconnect. Beyond merely employing a fast technology, the fabric must scale effectively with both the speed and number of systems and processors. Interconnect fabrics for large terascale and petascale deployments require:

- Low latency
- High bandwidth
- The ability to handle fabric congestion
- High reliability to avoid interruptions
- Open standards such as OpenFabrics and the OpenMPI SW stack

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 remote direct memory access (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 of service are also well served, since they each receive their own dedicated bandwidth.

Even with these advantages, building the largest InfiniBand clusters and grids has remained complex and expensive — primarily because of the need to interconnect very large numbers of computational nodes. 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. It is clear that density, consolidation, and management efficiencies are important not just 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. Interconnect topologies play a vital role in determining the properties that clustered systems exhibit. Torus (or toroidal) and Clos topologies are popular choices for interconnected supercomputing clusters and grids.

Torus Topologies

In torus topologies, each node connects to its neighbors in the x, y, and z dimensions, with six connecting ports per node. Some of the most notable supercomputers based upon torus topologies include IBM's BlueGene and Cray's XT3/XT4 supercomputers. Torus fabrics have had the advantage that they have generally been easier to build than Clos topologies. Unfortunately, torus topologies represent a blocking fabric, where interconnect bandwidth can vary between nodes. Torus fabrics also provide variable latency due to variable hop count, and application deployment for torus fabrics must carefully consider node locality as a result.

Clos Fat Tree Topologies

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, allowing 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 exists only between switch chips on adjacent tiers.

Clos fabrics have the advantage of being non-blocking, in that each attached node has a constant bandwidth. In addition, an equal number of stages between nodes provides for uniform latency. Historically, the disadvantage of large Clos networks is that they have been more difficult to build.

The World's Largest InfiniBand Clos Fat Tree Switch

Constructing very large InfiniBand Clos switches is governed by a number of practical constraints, including the number of ports available in individual switch elements, maximum achievable 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 inside the switch as depicted schematically in Figure 2. In the illustration, each vertical row of multiple switch elements defines a stage. The core switch element in the Sun DS 3456 is the Mellanox Infiniscale III 24-port InfiniBand switch chip.

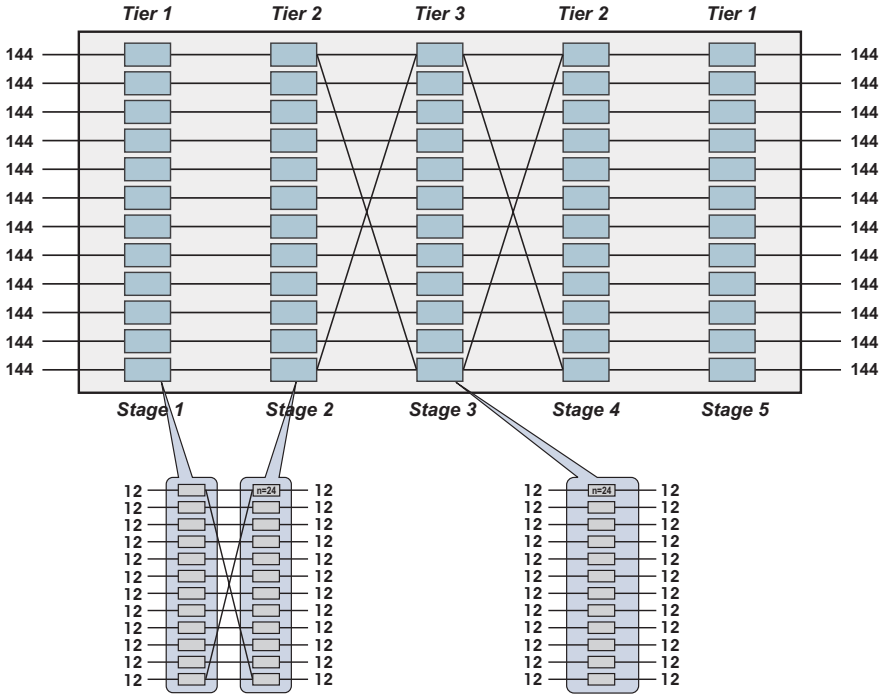

Figure 2. A 3-tier, 5-stage Clos fabric can achieve single-system connectivity of 3,456 ports using 144 24-port switch chips per stage and 720 chips total.

The Sun Datacenter Switch 3456

Providing the world's largest InfiniBand core switch 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. 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 virtual data lanes
- 1 virtual management lane
- 4096 byte maximum transmission unit (MTU)
- Ability to scale to 13,824 end nodes[1] when deployed with the Sun Blade 6048 InfiniBand Switched Network Express Module (NEM)

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

With its extensive cable management facilities, the Sun DS 3456 occupies slightly more than the space required for two standard racks. Labeled perspectives of the front and rear of the chassis are provided in Figure 3 and Figure 4.

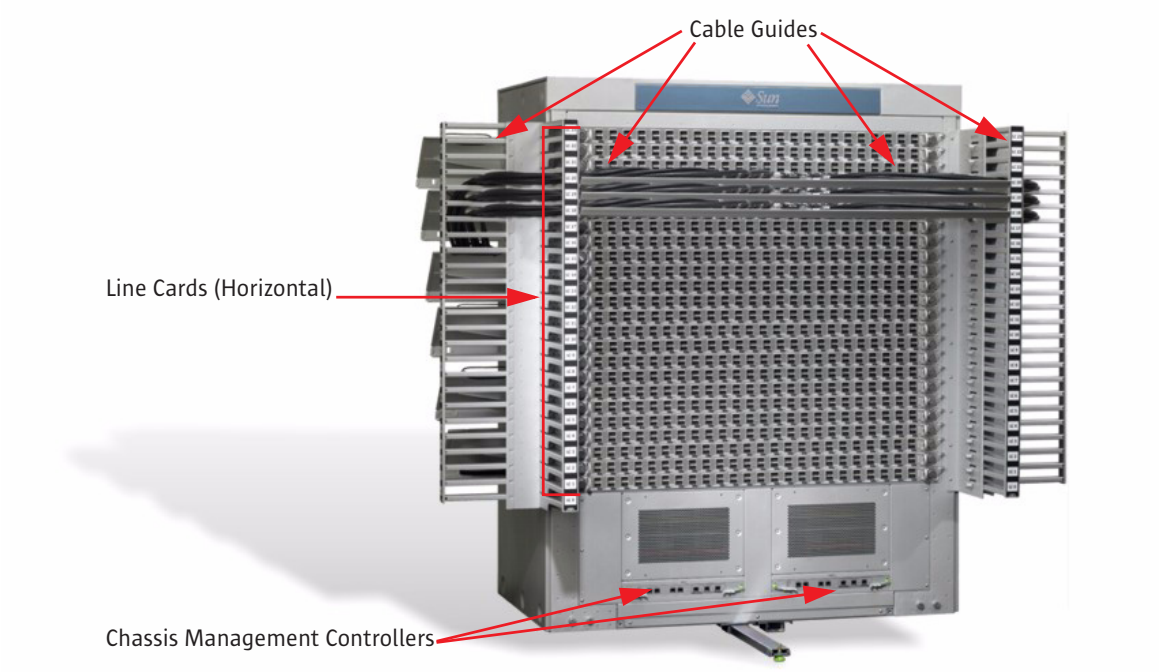

Figure 3. Sun DS 3456 front perspective

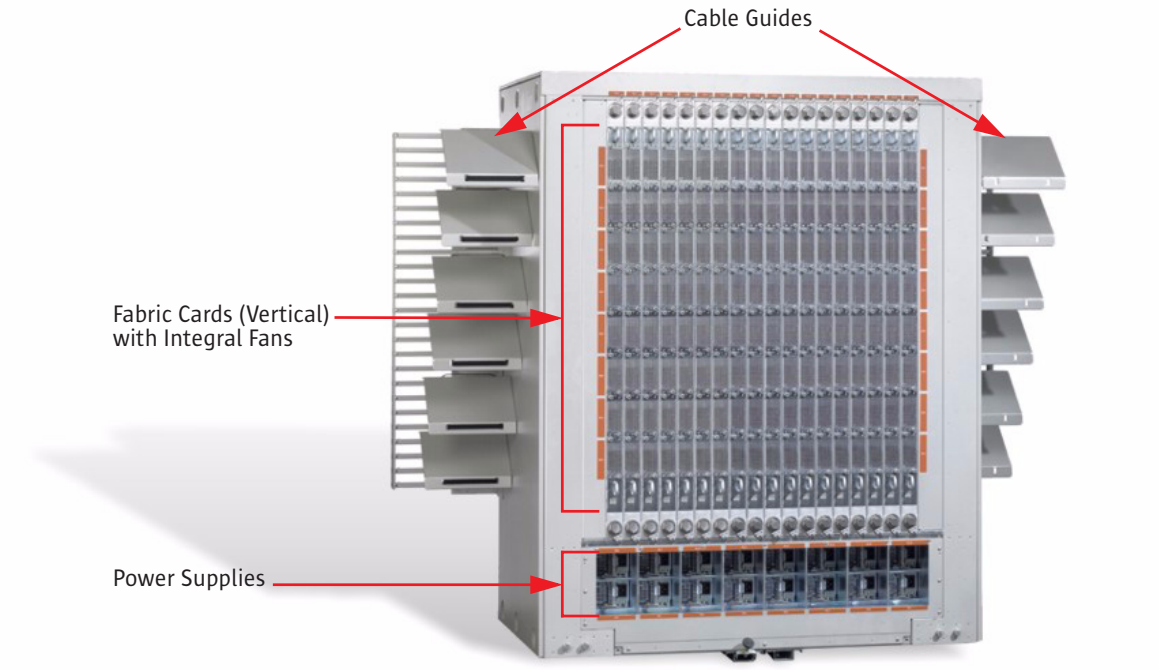

Figure 4. Sun DS 3456 rear perspective

High-level system components of the Sun DS 3456 include the following:

- Up to 24 horizontally-installed line cards are inserted 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 are inserted 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) are inserted from the rear of the chassis with the total 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.

Key Technical Innovations

Building the largest supercomputing and HPC configurations demands a new approach, one that rapidly brings new technology to bear on the most important problems and questions. Going beyond available InfiniBand switching, cabling, and host adapters, Sun engineers used their considerable networking and datacenter experience to view InfiniBand technology from a systemic perspective. The Sun DS 3456 represents a complete system that is based on multiple technical innovations:

- The Sun DS 3456 chassis implements a three-tier five-stage Clos fabric with up to 720 24-port Mellanox InfiniScale III switching elements, integrated into a single mechanical modular enclosure (Figure 5 illustrates a three-dimensional representation of the switch topology).
- 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. An optional splitter cable is available to convert one 12x connection to three 4x connections, for connectivity to legacy InfiniBand equipment.
- Integrated cable signal conditioner circuitry is provided that optimizes 4x signal integrity to support long cable runs. This approach effectively doubles the supported cable length for dual data rate (DDR) to 15 meters, and facilitates the use of copper cables that are thinner, less expensive, and more flexible than alternatives.

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

- 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 Sun Fabric Management software, and can assist the system administrator during construction and fault isolation.
- A custom-designed double-height PCI Express 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, where any system is only seven hops distant from any other.
- Fabric topology for forwarding InfiniBand traffic is established by a redundant host-based Subnet Manager running Sun Fabric Management software. A host-based solution allows the Subnet Manager to take advantage of the full resources of a general-purpose multicore server.

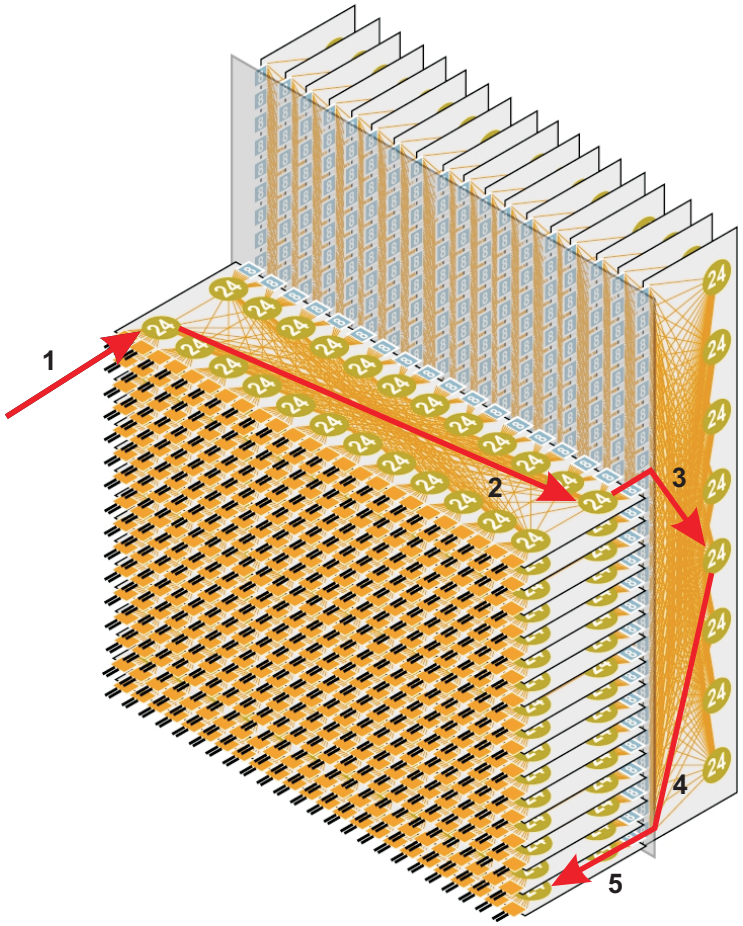

Figure 5. An InfiniBand transaction takes five hops through the five stages of Sun Datacenter 3456 Switch (the Sun Blade 6048 InfiniBand Switched NEM adds two hops)

Massive Switch and Cable Consolidation

Cost and complexity figure importantly, given the scale involved with building supercomputing clusters and grids. 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 — dramatically lowering acquisition costs. With its ultra-dense connectivity, a single Sun DS 3456 cabinet can replace as many as 300 smaller switches, providing up to a 6:1 reduction in rack space and weight. In addition, replacing physical switches and cabling with switch chips on printed circuit boards drastically improves reliability. Sun's new 12x InfiniBand cables and connectors coupled with the Sun Blade 6048 InfiniBand Switched NEM eliminate thousands of additional cables, providing additional cost, complexity, and reliability improvements. Overall, the Sun DS 3456 provides radical simplification of InfiniBand infrastructure.

Table 1 illustrates Sun estimates for deploying a 3,456-node HPC supercomputing cluster using the Sun DS 3456 and a typical 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. Through simplification, organizations can also build large supercomputing clusters more rapidly, 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

Chapter 3

Deploying Dense and Scalable Modular Compute Nodes

Implementing petascale supercomputing clusters depends heavily on having access to large numbers of high-performance systems with large memory support and high memory bandwidth. As a part of the Sun Constellation System, Sun's approach is to combine the considerable and constant performance gains in the standard processor marketplace with the advantages of modular architecture. This approach results in some of the fastest and most dense systems possible — all tightly integrated with the Sun DS 3456 core switch.

Compute Node Requirements

While systems such as IBM's Blue Gene employ very large numbers of slower proprietary nodes, this approach will likely not translate easily to petascale. The programmatic implications alone of handling literally millions of nodes are not particularly appealing — much less the physical realities of managing and housing such systems. Instead, building large terascale and petascale systems depends on key capabilities for compute nodes, including:

- ***High Performance***
  Compute nodes must provide very high peak levels of floating-point performance. Likewise, because floating-point performance is dependent on multiple memory operations, equally high levels of memory bandwidth must be provided. I/O bandwidth is also crucial, yielding high-speed access to storage and other compute nodes.

- ***Density, Power, and Cooling***
  The physical requirements of today's ever more expensive datacenter real estate dictate that any viable solutions take the best advantage of datacenter floor space and environmental realities. Solutions must be as energy efficient as possible, and must provide effective cooling that fits well with the latest energy-efficient datacenter practices.

- ***Superior Reliability and Serviceability***
  Due to their large numbers, computational systems must be as reliable and servicable as possible. Not only must systems provide redundant hot-swap processing, I/O, power, and cooling modules, but serviceability must be a key component of their design and management. Interconnect schemes must provide for redundant connections, and systems should only need to be cabled once and reconfigured at will as required.

Blade technology has offered considerable promise in these areas for some time, but has often been constrained by legacy blade platforms that locked adopters into an extensive proprietary infrastructure. Power and cooling limitations often meant that processors were limited to less powerful versions. Limited processing power, memory capacity, and I/O bandwidth often severely restricted the applications that could be deployed. Proprietary tie-ins and other constraints in chassis design dictated networking and interconnect topologies, and I/O expansion options were limited to a small number of expensive and proprietary modules.

The Sun Blade™ 6048 Modular System

To address the shortcomings of earlier blade computing platforms, Sun started with a design point focused on the needs of the datacenter and highly scalable deployments, rather than with preconceptions of chassis design. With this innovative and truly modular approach, the Sun Blade 6048 modular system offers an ultra-dense high-performance solution for large HPC clusters and grids. Organizations gain the promised benefits of blades, and can deploy thousands of nodes within the cabling, power, and cooling constraints of existing datacenters. Fully compatible with the Sun Blade 6000 modular system, the Sun Blade 6048 modular system provides distinct advantages over other approaches to modular architecture.

- ***Innovative Chassis Design for Industry-Leading Density and Environmentals***
  The Sun Blade 6048 modular system features a standard rack-size chassis that facilitates the deployment of high-density computational environments. By eliminating all of the hardware typically used to rack-mount individual blade chassis, the Sun Blade 6048 modular system provides 20 percent more usable space in the same physical footprint. Up to 48 Sun Blade 6000 server modules can be deployed in a single Sun Blade 6048 modular system. Innovative chassis features are carried forward from the Sun Blade 6000 modular system.

- ***A Choice of Processors and Operating Systems***
  Each Sun Blade 6048 modular system chassis supports up to 48 full performance and full featured Sun Blade 6000 series server modules, including:

  - The *Sun Blade X6250 server module* provides two sockets for Dual-Core Intel® Xeon® Processor 5100 series or two Quad-Core Intel Xeon Processor 5300 series CPUs with up to 64 GB of memory per server module.
  - The *Sun Blade X6220 server module* provides support for two Next Generation AMD Opteron™ 2000 Series processors and support for up to 64 GB of memory.
  - *The Sun Blade T6320 server module* offers support for the massively-threaded UltraSPARC® T2 processor with either four, six, or eight cores, up to 64 threads, and support for 64 GB of memory.
  - The *Sun Blade T6300 server module* provides a single socket for an UltraSPARC T1 processor, featuring either six or eight cores, up to 32 threads, and support for up to 32 GB of memory.

Each server module provides significant I/O capacity as well, with up to 32 lanes of PCI Express bandwidth delivered from each server module to the multiple available I/O expansion modules (a total of up to 142 Gb/s per supported per server module). To enhance availability, server modules don't have separate power supplies or fans, and each server module features up to four hot-swap disks with hardware RAID built in. Organizations can deploy server modules based on the processors and operating systems that best serve their applications or environment. Different server modules can be mixed and matched in a single chassis, and deployed and redeployed as needs dictate. The Solaris™ Operating System (OS), Linux, and Microsoft Windows are all supported.

- ***Complete Separation Between CPU and I/O Modules***
  Sun Blade 6048 modular system design avoids compromises because it provides a complete separation between CPU and I/O modules. Two types of I/O modules are supported.
  - Up to two industry-standard PCI Express ExpressModule (EMs) slots are dedicated to each server module.
  - Up to two PCI Express Network Express Modules (NEMs) provide bulk IO for all of the server modules installed in each shelf (four shelves per chassis).

  Through this flexible approach, server modules can be configured with different I/O options depending on the applications they host. All I/O modules are hot-plug capable, and customers can choose from Sun-branded or third-party adapters for networking, storage, clustering, and other I/O functions.

- ***Sun Blade Transparent Management***
  Many blade vendors provide management solutions that lock organizations into proprietary management tools. With the Sun Blade 6048 modular system, customers have the choice of using their existing management tools or Sun Blade Transparent Management. Sun Blade Transparent Management is a standards-based cross-platform tool that provides direct management over individual server modules and direct management of chassis-level modules using Sun Integrated Lights out Management (ILOM).

  Within the Sun Blade 6048 modular system, a chassis monitoring module (CMM) works in conjunction with the service processor on each server module to form a complete and transparent management solution. Individual server modules provide support for IPMI, SNMP, CLI (through serial console or SSH), and HTTP(S) management methods. In addition, Sun xVM Ops Center (formerly Sun Connection and Sun N1™ System Manager) provides discovery, aggregated management, and bulk deployment for multiple systems.

System Overview

By utilizing the lateral space that would otherwise be required 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 6.

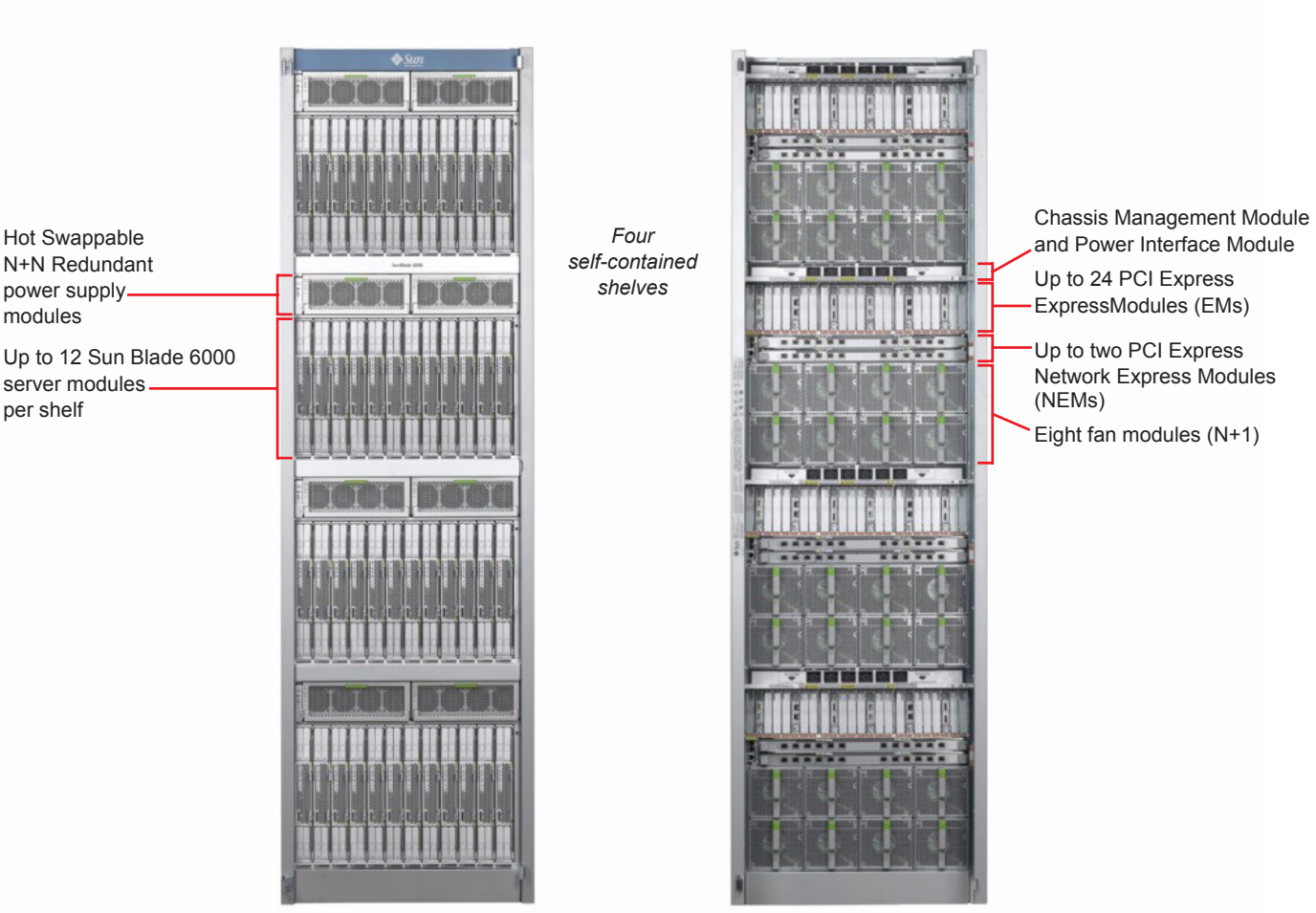

Figure 6. Front and rear perspectives of the Sun Blade 6048 modular system

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 of the chassis, with 12 modules supported by each shelf.
- A total of eight hot-swap power supply modules insert from the front of the chassis, with two 8,400 Watt 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 available 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.

Standard I/O Through a Passive Midplane

In essence, the passive midplane in the Sun Blade 6048 modular system is a collection of wires and connectors between different modules in the chassis. Since there are no active components, the reliability of this printed circuit board is extremely high — in the millions of hours. The passive midplane provides electrical connectivity between the server modules and the I/O modules.

All modules, front and rear connect directly to the passive midplane, with the exception of the power supplies and the fan modules. The power supplies connect to the midplane through a bus bar and to the AC inputs via a cable harness. The redundant fan modules plug individually to a set of three fan boards, where fan speed control and other chassis-level functions are implemented. The front fan modules that cool the PCI Express ExpressModules each connect to the chassis via self-aligning, blind-mate connections. The main functions of the midplane include:

- Providing a mechanical connection point for all of the server modules
- Providing 12 VDC from the power supplies to each customer-replaceable module
- Providing 3.3 VDC power used to power the System Management Bus devices on each module, and to power the CMM
- Providing a PCI Express interconnect between the PCI Express root complexes on each server module to the EMs and NEMs installed in the chassis
- Connecting the server modules, CMMs, and NEMs to the management network

Each server module is energized through the midplane from the redundant power grid. The midplane also provides connectivity to the I2C network in the chassis, letting each server module directly monitor the chassis environment, including fan and power supply status as well as various temperature sensors. A number of I/O links are also routed through the midplane for each server module (Figure 7), including:

- Two x8 PCI Express links connect from each server module to each of the dedicated EMs
- Two x8 PCI Express links connect from each server module, one to each of the NEMs
- Two gigabit Ethernet links are provided, each connecting to one of the NEMs
- Four x1 Serial Attached SCSI (SAS) links are also provided, with two connecting to each NEM (for future use)

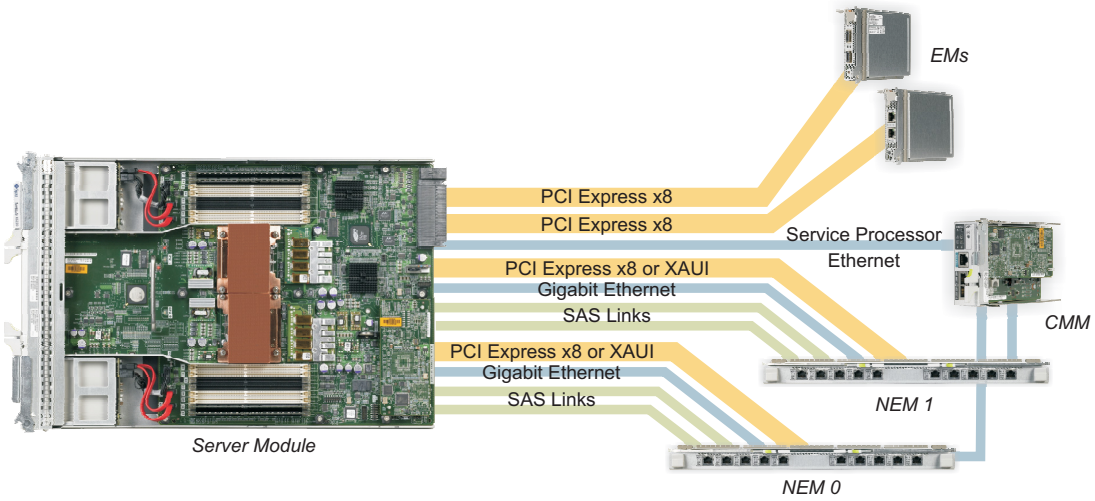

Figure 7. Distribution of communications links from each Sun Blade 6000 server module

Tight Integration with the Sun DS 3456

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 dual-height NEM for the Sun Blade 6048 chassis. Designed to work with the Sun DS 3456, the NEM uses common components, cables, connectors, and architecture. A block-level diagram of the NEM is provided in Figure 8, aligned with an image of the back panel.

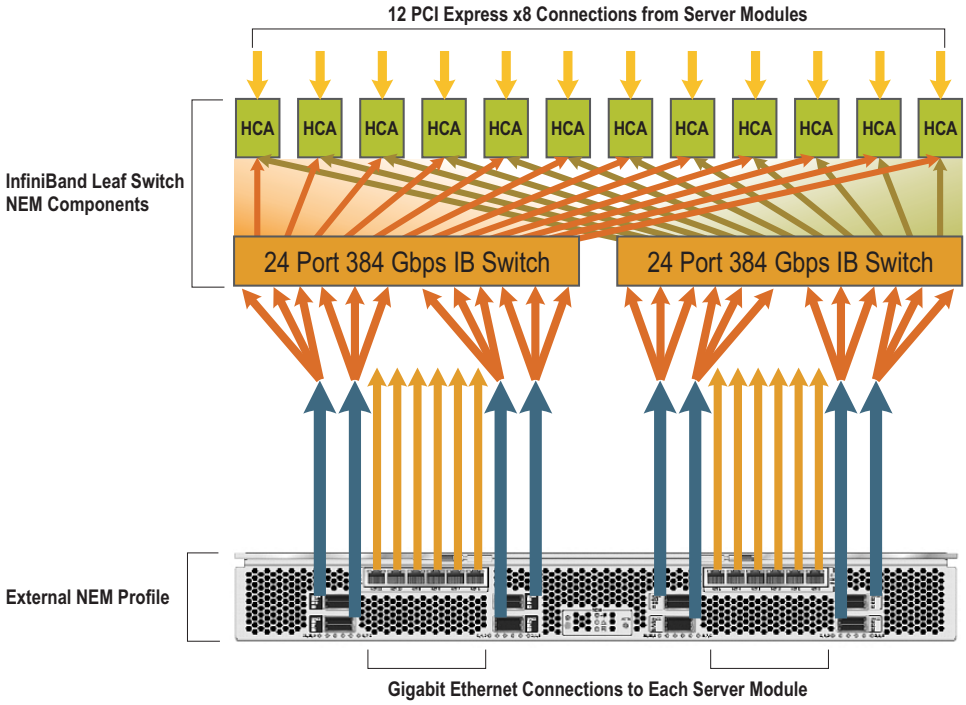

Figure 8. The Sun Blade 6048 InfiniBand Switched NEM provides eight switched 12x InfiniBand connections to the two on-board 24-port switches, and twelve pass-through gigabit Ethernet ports

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 13,824 Compute Nodes

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
- Four core switches can serve up to 13,824 server modules

To maintain a non-blocking fabric, each Sun Blade 6048 InfiniBand Switched 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 InfiniBand Switched NEM connects to every core switch with either one or two 12x cables. Table 2 lists the connections, maximum supported servers, and Sun Blade 6048 modular systems supportable with one, two, and four Sun DS 3456 core switches.

Table 2. 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)

Chapter 4
Scalable and Manageable Storage

Large-scale supercomputing clusters place significant demands on storage systems. The enormous computational performance gains that have been realized through supercomputing clusters are capable of generating ever-larger quantities of data at very high rates. Effective HPC storage solutions must provide cost-effective capacity, and throughput must be able to scale along with the performance of cluster compute nodes. Users and systems alike need fast access to data and storage, and longer-term retention and archival are increasingly important in HPC and supercomputing environments. These demands require a robust range of integrated storage offerings.

Storage for Grids and Clusters

Along with the general growth in storage capacity requirements and the shear number of files stored, large HPC environments are seeing significant growth in the numbers of users needing convenient access to their files. All users want to access their essential data quickly and easily without having to perform extraneous steps. Organizations also want to get the best utilization possible from their computational systems. Unfortunately, storage speeds have seriously lagged behind computational performance for years, and HPC users are increasingly concerned about storage benchmarks, the increasingly complexity of the I/O path, and the range of solutions required to provide complete storage solutions.

Of particular importance, large HPC environments need to be able to effectively manage the flow of high volumes of data through their storage infrastructure, requiring:

- Storage that acts as a *resilient compute engine data cache* to match the streaming rates of applications running on the compute cluster
- Storage that provides longer-term *retention and archive* to store massive quantities of essential data to tiered disk or tape hierarchies
- A range of scalable and parallel file systems and integrated data management software to help project file system data from near-term cache to longer-term retention and archiving and back on demand

Even as the capacities of individual disk drives have risen, and prices have fallen, high-volume parallel storage systems have remained expensive and complex. With experience deploying petabytes of storage into large supercomputing clusters, Sun understands key issues needed to deliver high-capacity, high-throughput storage in a cost-effective and manageable fashion. As an example, the Tokyo Institute of Technology (TiTech) TSUBAME supercomputing cluster was initially deployed with 1.1 petabytes of storage provided by clustered Sun Fire X4500 servers and the Lustre parallel file system.

Clustered Sun Fire™ X4500 Servers as Data Cache

Ideal for building storage clusters to serve as a data cache, the Sun Fire X4500 server defines a new category of system in the form of a general-purpose enterprise-class x64 server that is closely coupled with high-density storage — all in a very compact form factor. Supporting up to 48 terabytes in only four rack units, the Sun Fire X4500 server also provides considerable compute power with dual sockets for Second Generation AMD Opteron processors. The server can also be configured for high-throughput InfiniBand networking — allowing it to be connected directly to the Sun DS 3456. With support for up to 48 internal 250 GB, 500 GB, 750 GB, or 1 TB disk drives, the Sun Fire X4500 server is ideal for large cluster deployments with Solaris ZFS or the Lustre parallel file system. As with other Sun x64 servers, support is provided for the Solaris OS as well as Linux and Microsoft Windows.

Figure 9. The Sun Fire X4500 server provides up to 48 terabytes of compact storage in only four rack units

The Sun Fire X4500 server represents an innovative design that provides throughput and high-speed access to the 48 directly-attached, hot-plug Serial ATA (SATA) disk drives. Designed for datacenter deployment, the efficient system is cooled from from front to back across the components and disk drives. Each Sun Fire X4500 server provides:

- Minimal cost per gigabyte utilizing SATA II storage and software RAID 6 with six SATA II storage controllers connecting to 48 high-performance SATA disk drives
- High performance from an industry-standard x64 server based on two Second Generation AMD Opteron processors
- Maximum memory and bandwidth scaling from embedded single-channel DDR memory controllers on each processor, delivering up to 16 GB of capacity and 12.8 GB/second of aggregated bandwidth with two processors and eight 2 GB DIMMs
- High-performance I/O from two PCI-X slots to delivers over 8.5 gigabits per second of plug-in I/O bandwidth, including support for InfiniBand HCAs
- Easy maintenance and overall system reliability and availability from redundant hot-pluggable disks, power supply units, fans, and I/O

Tight Integration of the Solaris™ ZFS Scalable File System

The Sun Fire X4500 server is tightly integrated with the Solaris ZFS scalable file stem. Open-source Solaris ZFS offers a dramatic advance in data management with an innovative approach to data integrity, tremendous performance improvements, and a welcome integration of both file system and volume management capabilities. A true 128-bit file system, Solaris ZFS removes all practical limitations for scalable storage, and introduces pivotal new concepts such as virtual storage pools that de-couple the file system from physical storage.

This radical new architecture optimizes and simplifies code paths from the application to the hardware, producing sustained throughput at near platter speeds. New block allocation algorithms accelerate write operations, consolidating what would traditionally be many small random writes into a single, more efficient write sequence. The Solaris 10 OS with Solaris ZFS is also the only current general-purpose OS designed to provide end-to-end checksumming for all data, reducing the risk of silent data corruption.

Parallel File System Support

The Sun Fire X4500 server is now a standard component of many large supercomputing cluster deployments around the world. Large grids and clusters need high-performance heterogeneous access to data, and the Sun Fire X4500 server provides both high throughput as well as essential scalability that allow parallel file systems to perform at their best. In addition to multiple operating system support on the Sun Fire X4500 server, additional parallel file systems may be run.

- ***The Lustre Parallel File System***

  Many large HPC systems have deployed the Lustre parallel file system to manage data in their production Linux environments, including some of the world's largest supercomputers. Lustre's state-of-the-art object-based storage architecture provides ground-breaking I/O and metadata throughput, with considerable reliability, scalability, and performance advantages.

  The Lustre file system currently scales to thousands of nodes and hundreds of terabytes of storage — with active deployment to support tens of thousands of nodes and petabytes of data. Together with the Lustre parallel file system and the Linux OS, the Sun Fire X4500 server also serves as the key component for the sun Customer Ready Scalable Storage cluster (*Chapter 5*).

- ***Parallel NFS (pNFS)***

  The parallel network file system (pNFS) is a forthcoming standards-based extension to NFS v4 that allows clients to access storage devices directly, and in parallel — eliminating the scalability and performance issues often associated with NFS servers in existence today. pNFS resolves these issues by separating data

and metadata, moving the metadata server out of the data path. By bringing parallel file system I/O to the ubiquitous standard for network file systems, pNFS preserves investments while users gain increased performance and scalability.

Long-Term Retention and Archive

Staging, storing, and maintaining HPC data requires a massive repository of on-line and near-line storage to support HPC data retention and archival needs. High-speed data movement must be provided between computational and archival environments. The Sun Constellation System addresses this need by integrating with a wealth of sophisticated Sun StorageTek™ options, including:

- Sun StorageTek SL8500 and SL500 Modular Library Systems
- Sun StorageTek 6540 and 6140 Modular Arrays
- High-speed data movers
- Sun StorageTek 5800 system fixed-content archive

Sun's comprehensive StorageTek software offering is particularly key to facilitating seamless migration of data between cache and archival.

- ***Sun StorageTek™ QFS***
  Sun StorageTek QFS software provides high-performance, heterogeneous shared access to data over a storage area network (SAN). Users across the enterprise get shared access to the same large files or data sets simultaneously, speeding time to results. Up to 256 systems running StorageTek QFS technology can have shared access to the same data while maintaining file integrity. Data can be written and accessed at device-rated speeds, providing superior application IO rates. Sun StorageTek QFS software also provides heterogeneous file sharing using NFS, CIFS, Apple Filing Protocol, FTP, and Samba.

- ***Sun StorageTek Storage Archive Manager (SAM) Software***
  Large HPC installations must manage the considerable storage required by multiple projects running large-scale computational applications on very large datasets. Solutions must provide a seamless and transparent migration for essential archival data between disk and tape storage systems. Sun StorageTek Storage Archive Manager (SAM) addresses this need by providing data classification and policy-driven data placement across tiers of storage. Organizations can benefit from data protection as well as long-term retention and data recovery to match their specific needs.

Chapter 5 provides additional detail and a graphical depiction of how caching file systems such as the Lustre parallel file system combine with SAM-QFS in a real-world example to provide data management in one of the world's largest supercomputing installations.

Chapter 5

Deploying Supercomputing Clusters Rapidly with Less Risk

Sun has considerable experience helping organizations deploy supercomputing clusters specific to their computational, storage, and collaborative requirements. Complementing the compelling capabilities of the Sun Constellation System, Sun provides a range of services that are specifically focused at delivering results for HPC-focused organizations. Sun's partnership with the Texas Advanced Computing Center (TACC) at the University of Texas at Austin to deliver the Sun Constellation System in the 3,936-node Ranger supercomputing cluster is one such example.

Sun Datacenter Express Services

Sun's new Datacenter Express Services provide a comprehensive, all-in-one systems and services solution that takes the complexity and cost out of HPC infrastructure and procurement. Sun delivers a uniquely flexible approach to managing IT resources that allows organizations to maintain control of their environments. The offerings combine the cost savings and improved quality of the Sun Customer Ready Program, with the expertise of Sun Services to deliver:

- Improved availability and decreased cost
- Optimized system performance and reduced complexity
- Control of IT resources at all times

With Datacenter Express, Sun's Customer Ready Program builds and tests systems to exact customer specifications, while Sun Services provides complementary expertise based upon exact requirements. Having testing and component integration performed at Sun ISO-certified facilities helps reduce HPC system deployment time, installation issues, and minimizes unnecessary downtime. With Sun Datacenter Express, organizations leverage the expertise of Sun Services so that HPC solutions are easier to repair, maintain, and support — and configurations are easier to scale and modify.

Sun Customer Ready Architected Systems

The Sun Customer Ready program of factory-integrated systems provides powerful, cost-effective, and energy-efficient computing solutions. The program helps reduce the risk and time to deployment for clusters and grids through pre-configured solutions that are factory integrated and tested. As a part of this program, Sun Customer Ready architected systems combine Sun's powerful and cost-effective server and storage products with leading infrastructure or application software to provide cluster building blocks that are complete, easy-to-order, and fast-to-deploy.

- ***Sun Customer Ready HPC Cluster***
  The Sun Customer Ready HPC Cluster offers a high-performance computing cluster solution that can be customized to specific needs. With a choice of Sun Fire x64 rackmount servers and Sun Blade modular systems, Sun Customer Ready HPC cluster allows organizations to customize their cluster designs without compromise.
- ***Sun Customer Ready Storage Cluster***
  The Sun Customer Ready Scalable Storage Cluster provides a high-performance storage solution to support the demands of HPC clusters for fast access to working data sets. Built around the Sun Fire X4500 server and the Lustre scalable cluster file system, the Sun Customer Ready Scalable Storage Cluster supports large clusters of compute nodes where high data throughput rates and low-latency, high-speed interconnects are required.
- ***Sun Blade Scalable Units for HPC***
  Sun Blade Scalable Units provide cluster building blocks that arrive with integrated Sun Blade 6000 and 6048 modular systems, high-performance networking, and cabling already configured. Each Sun Blade Scalable Unit is built with server modules containing either AMD Opteron or Intel Xeon processors, with the Solaris 10 Operating System pre-installed, and support for InfiniBand fabric and networking.
- ***Sun Customer Ready Visualization System***
  The Sun Customer Ready Visualization System provides a complete solution that can delivery both immersive and remote visualization with a choice of Sun UltraSPARC and x64 systems and NVIDIA Quadro FX or Quadro Plex VCS graphics solutions. InfiniBand switches and channel adapters are configured along with Sun Scalable Visualization software and Sun Shared Visualization software.

A Massive Supercomputing Cluster at the Texas Advanced Computing Center

The Ranger supercomputing cluster now being deployed at TACC is testament to Sun's commitment to help design and build the world's largest grids and clusters. TACC is a leading research center for advanced computational science, engineering, and technology, supporting 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 peak performance, 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 Ranger Computation and Interconnect Architecture

TACC is deploying the Sun Constellation System to build the Ranger supercomputing cluster. When complete, the 3,936-node cluster is expected to provide a peak rating of over 500 teraflops of peak performance[1]. Figure 10 illustrates the initial deployment floor-plan, 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

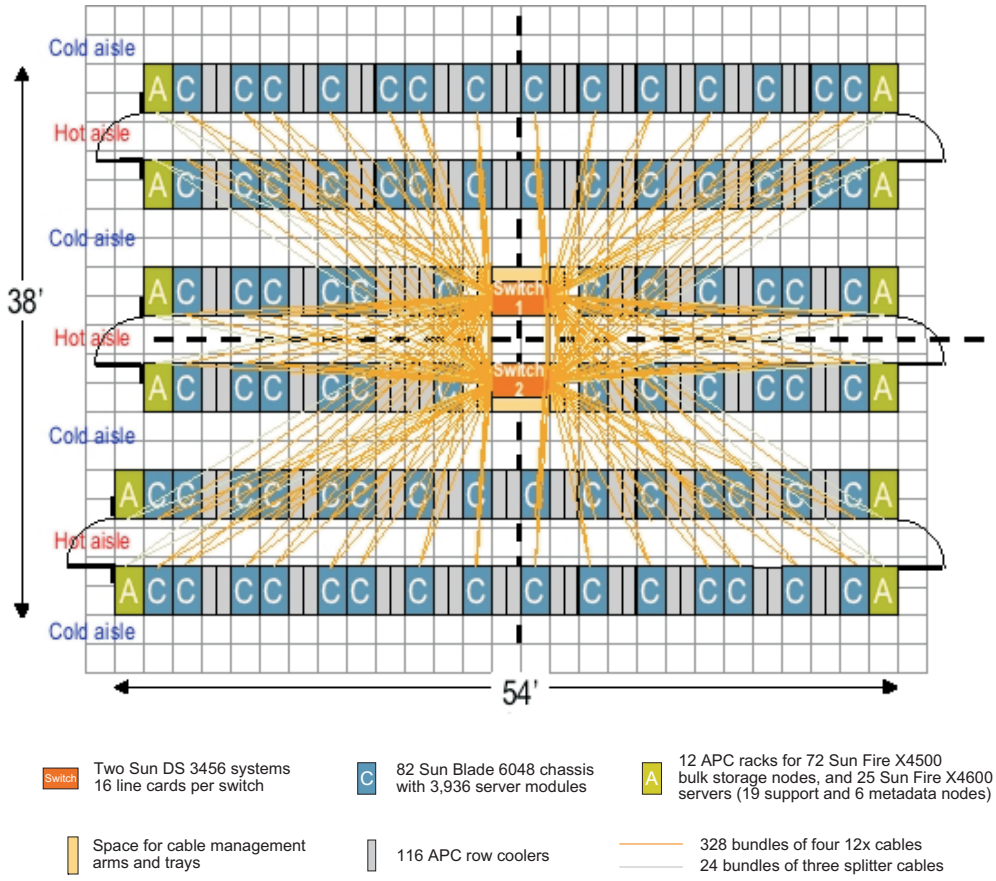

Figure 10. TACC initial deployment floor-plan featuring dual Sun DS 3456 core switches and 82 fully-populated Sun Blade 6048 modular systems

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

TACC Ranger and Sun HPC Storage Solutions

For the TACC Ranger installation, it was essential to provide an effective data cache that could keep pace with the massive high-speed cluster while deploying tiered storage infrastructure for long-term retention and archival. Large capacities and throughput where essential throughout. Tight integration and ease of data migration were also required.

A Resilient Compute Engine Data Cache

HPC applications need fast access to high-capacity data storage for writing and retrieving data sets. TACC needed storage infrastructure that functions as an effective and resilient compute engine data cache — providing the maximum aggregate throughput at the lowest possible costs. This cache had to be capable of scaling to hundreds of petabytes, with very low latency for temporary storage, and accessible by all of the compute nodes in the cluster. Ideally, the data cache had to be easy to deploy, administer, and maintain.

Designed as “fast scratch space” for large clusters, the Sun Cluster Ready Scalable Storage Cluster provides the data storage capacity and throughput that these applications require. Key components of this data storage solution include high-performance Sun Fire X4500 servers, the Lustre scalable cluster file system, and high-speed InfiniBand interconnects, integrated into one system. As deployed at TACC, the system will scale to over:

- 72 GB/second sustained bandwidth
- 1.728 petabytes of raw capacity

The configuration includes a scalable storage cluster with 72 Sun Fire X4500 servers and over three thousand 500 GB disk drives — but only occupies eight physical racks.

Long-Term Retention and Archive

At the other end of the spectrum, TACC needed storage infrastructure for long-term retention and archival with a deep repository for massive amounts of data. Users needed to be able to access their data regardless of location. Integral management was required for the transparent movement of data from archival media (e.g. tape) in and out of the compute engine data cache.

The Sun Customer ready Scalable Storage Cluster that was deployed to implement the data cache was designed as part of an overall cluster solution (Figure 11). Sun Fire X4500 storage clusters were tightly integrated with other complementary products, including Sun StorageTek Storage Archive Manager and the QFS file system, that provide long-term data archive capabilities to other Sun StorageTek devices. High-performance Sun Fire servers act as Data Movers, efficiently moving data between the fast scratch storage of the Sun Customer Ready Storage Cluster and long-term storage.

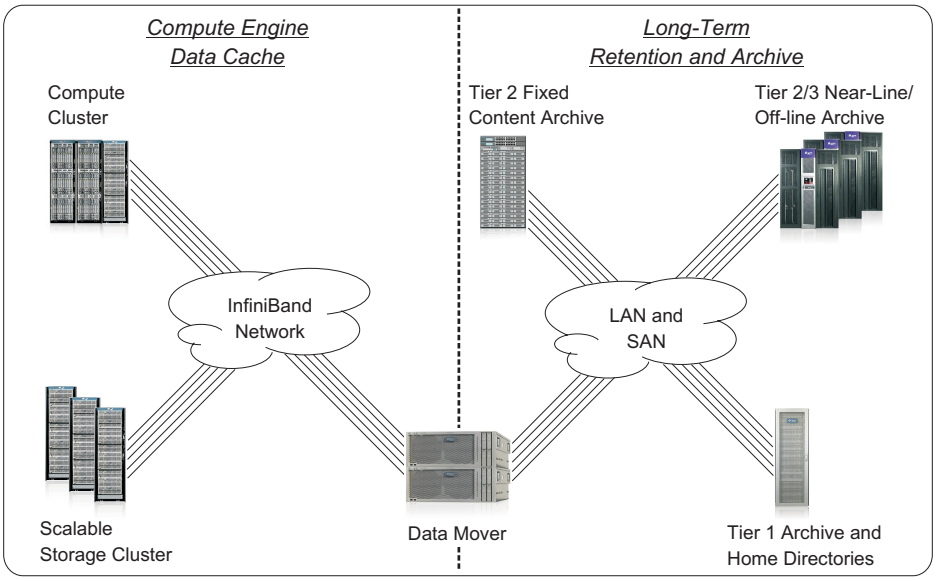

Figure 11. Data movers provide automated, policy-based data management and migration between storage tiers

As deployed at TACC the system will scale to over:

- 200 petabytes of near-line storage
- 3.1 petabytes of on-line storage

The configuration includes:

- Five Sun StorageTek SL8500 Modular Library Systems
- 48 Sun StorageTek T10000 tape drives
- 10 Sun StorageTek 6540 Arrays
- Six Sun Fire Metadata servers with SAM-QFS

Chapter 6
Conclusion

In spite of advancements in technology, delivering petascale clusters and grids has remained challenging. More than vast collections of nodes, Sun's approach to large terascale and petascale architecture provides a systematic and careful design of fabric, compute, and storage elements. The Sun Constellation System delivers an open and standard supercomputing architecture that provides massive scalability, a dramatic reduction in complexity, and breakthrough economics.

Implemented as a massive InfiniBand core switch, the Sun DS 3456 can support up to 3,456 4x InfiniBand ports and provides 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 in a single InfiniBand fabric. With an innovative and robust new 12x connector and cable, the system drastically reduces the number of switches required for large supercomputing installations, and provides a six-fold reduction in cabling — eliminating hundreds of switches, and thousands of cables.

The Sun Blade 6048 modular system is the first blade platform designed for extreme density and performance. With a choice of the latest SPARC®, Intel, and AMD processors, the Sun Blade 6048 modular system integrates tightly with the Sun DS 3456. Fully compatible with the Sun Blade 6000 modular system, server modules run standard open-source operating systems such as the Solaris OS and Linux, and can deploy general-purpose software that does not require custom compiles and tuning. A modular and efficient design realizes savings in both power and cooling.

Along with a breadth of Sun StorageTek storage offerings, the Sun Fire X4500 server provides one of the most economical and scalable parallel file system building blocks available. Supporting up to 48 TB in a single 4U chassis, the Sun Fire X4500 server effectively combines a powerful multiprocessor, multicore x64 server with large-scale storage, and direct InfiniBand connectivity. Scalable and parallel file systems such as Solaris ZFS and the Lustre parallel file system effectively utilize the resources of the Sun Fire X4500 server, providing direct access for cluster systems.

The Sun Constellation System also provides a comprehensive software stack that supports application developers and cluster users alike. Integrated tools such as Sun Studio 12 provide the fastest compilers available, tuned to get the most of Sun platforms. Sun HPC Cluster Tools enable the development of cluster-ready applications. Sun Grid Engine provides distributed resource management along with policy enforcement as it distributes jobs for execution. Sun xVM Ops Center provides monitoring, patching, and simplified inventory management for even the largest clusters and grids. Together these tools help ensure that development and management of large clusters remains manageable as scale towards petascale.

Acknowledgements

This work was inspired by and in part based on previous works from Andreas Bechtolsheim (International Symposium on SuperComputing 2007), Jim Waldo for his Sun Labs paper “On System Design” SMLI-PS-2006-6, and Ivan Sutherland’s paper “Technology and Courage” SMLI-PS-96-1.

For More Information

To learn more about Sun products and the benefits of the Sun Constellation System, contact a Sun sales representative, or consult the related documents and Web sites listed in Table 3.

Table 3. Related Websites

Web Site URL	Description
sun.com/sunconstellationsystem	Sun Constellation System
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/storagetek	Sun StorageTek storage products
sun.com/gridware	Sun Grid Engine software

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

Printed in USA 02/08