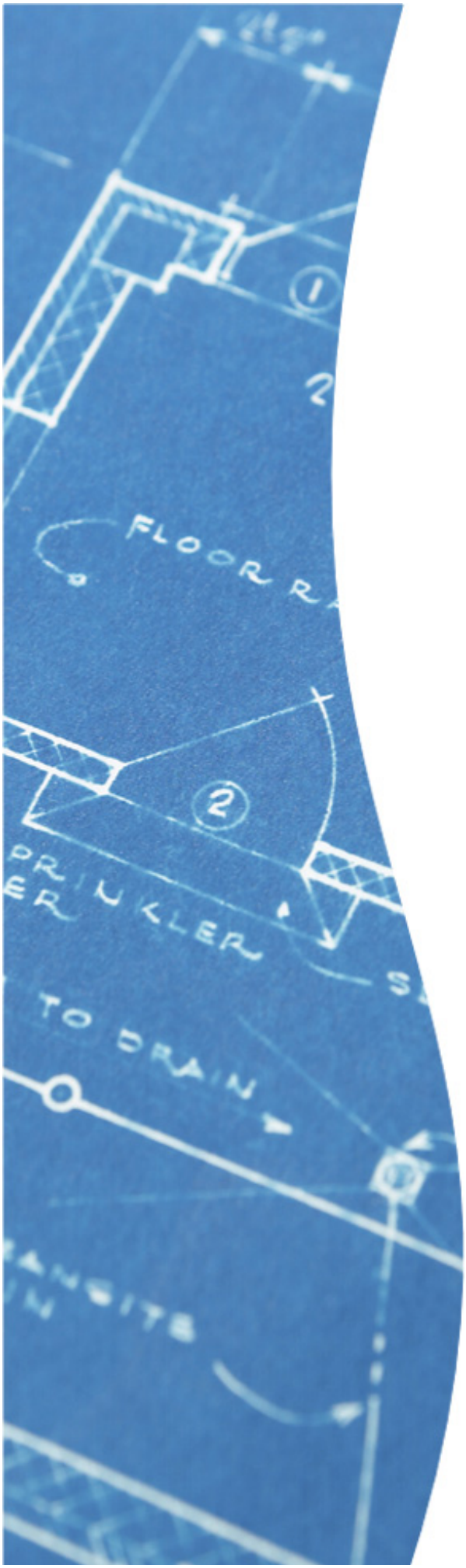

SUN'S HIGH PERFORMANCE **COMPUTING REFERENCE ARCHITECTURE**

Torben Kling-Petersen, Senior Technical Specialist
Börje Lindh, Senior Systems Engineer
Ola Tørudbakken, Distinguished Engineer

Sun BluePrints™ Online

Part No 820-7583-10
Revision 1.1, 6/1/09

Table of Contents

Introduction

High-performance computing (HPC) is rapidly becoming mainstream, and is now used in all areas of industry — from financial services, retail, manufacturing and aerospace concerns, to traditional areas such as research and education. Familiarity with the components that make up an HPC system is becoming vital not only for IT managers, but also for CIOs, CFOs, and CEOs. Insight into the strategic and competitive advantages offered by HPC infrastructure can and will change how companies do business and address their core missions.

While high-performance computing infrastructure is often seen as an essential part of academia or scientific research institutions, it is increasingly commonplace in almost all types of industries. Though the largest terascale and petascale HPC clusters generate headlines, smaller and more moderately-sized clusters now deliver applications across wide range of organizations. In fact, the majority of compute clusters built today are dedicated to areas as diverse as retail, insurance, traffic control, manufacturing, and other lines of business.

HPC infrastructure solutions also need not represent large complex systems. For example, a single highly-optimized system dedicated to a critical workload and shared by many users can be considered HPC infrastructure. Likewise, a compute cluster consisting of only a few systems might serve vital HPC applications. Smaller clusters can form the nucleus of a larger cluster as demand increases, allowing organizations to start small and grow as needed. Furthermore, compute clusters need not consist of a large number of identical systems. Historically, clusters were often constructed from what ever equipment was available, including back-end servers, dedicated compute resources, and even desktop systems. When not in demand for other purposes, these diverse computational resources could help solve critical problems. Today, compute clusters are often built from a selection of high-performance servers, and may run different releases of the Solaris™ Operating System (OS), Linux, and Microsoft Windows. Virtualization technology is also key to modern clusters.

Unfortunately, constructing HPC infrastructure has remained complex for many, and if not done properly, configurations can be difficult to scale as demand for compute power, storage, and interconnect bandwidth grows. Sun's HPC Reference Architecture is designed to address these challenges by providing an integrated end-to-end approach that invokes Sun's considerable experience deploying some of the largest supercomputing clusters in existence. Sun's HPC Reference Architecture addresses a broad range of individual HPC solutions, from single-rack clusters though terascale and petascale supercomputing clusters. Beyond server hardware, Sun's approach provides integrated and tested systems, innovative storage and file systems, high-speed

networking and interconnects, and HPC software. The reference architecture simplifies design and rapid deployment of clusters of virtually any size, without incurring arbitrary limitations on resulting HPC infrastructure.

As HPC technology has become more mainstream, IT managers and other professionals need to understand the components and strategies for building effective HPC infrastructure. This Sun BluePrints article is designed to provide IT managers with a grounding in the basic products and technologies that comprise Suns' HPC Reference Architecture. Key components are highlighted along with typical uses, issues, and architectures of HPC systems.

- "Sun's HPC Reference Architecture" on page 3 gives brief background on HPC, clustering, and introduces Sun's HPC Reference Architecture
- "Compute Systems" on page 11 provides an overview of compute node for clusters and highlights the diverse range of compute nodes available from Sun.
- "Interconnects and High-Performance Networks" on page 15 provides background on Ethernet networks and InfiniBand interconnects, and introduces Sun Datacenter Switches for DDR and QDR InfiniBand fabrics
- "File Systems and Archive Solutions for HPC" on page 20 outlines the methods for feeding data to the cluster as well serving home directories and providing solutions for archiving data
- "HPC Software, Resource Management, and System Management" on page 28 describes Sun HPC Software along with resource and system management software

Chapter 1

Sun's HPC Reference Architecture

In it's most basic form, HPC infrastructure simply lets computations run faster, or provides larger numbers of results, so that answers are generated faster and conclusions can be drawn more quickly. For this reason, HPC infrastructure has long been essential for simulating complex systems as with weather prediction or modeling of nuclear reactions. Experience gained in addressing these demanding applications is now benefiting smaller and mid-size HPC deployments.

Applications from diverse industries such as manufacturing, life sciences, and financial services are now requiring more and more computational resources. For many organizations, larger monolithic symmetric multiprocessing systems (SMPs) are often not cost-effective for the most computationally-demanding applications. Clusters of many smaller, less-expensive systems are providing compelling and economical solutions. In fact, clustering technology has provided the ubiquity and raw scalability required to approach the largest problems, even as it serves small to moderate computational needs. Sun's HPC Reference Architecture is designed to apply HPC technology to build clusters of virtually any size.

Improved efficiency through HPC

High-performance computing has traditionally been the viewed in the context of research and scientific pursuit, but this perspective is changing. HPC technology is now increasingly used for smaller and mid-range computational tasks — many related to making businesses run better or more efficiently. In fact, the computations performed by modern HPC infrastructure can benefit a range of diverse applications, including more timely risk analysis in financial services, airflow or thermal simulation in manufacturing, reservoir simulations in the oil and gas industry, or molecular screenings to help find new drugs. The automotive industry in particular uses HPC to simulate the movement of fluids and gases (such as the flow of air around a car), as well as simulation of deformations (as when simulating a car crash or forming of a piece of sheet metal).

Top500
The Top500 list is a list of the fastest super computers in the world, publicly announced by an independent consortium and updated twice a year at:
`http://www.top500.org`

The wide applicability of clustering technology in particular has made this rapid transition possible. The very same technologies that help enable the world's largest supercomputing sites are now facilitating rapid change in a wide range of commercial and other industries. Increasingly, effective HPC infrastructure is making complex computational tasks simple and accessible. Users may not even know that a given job is executed on a remote system — they only see the faster result. Distributed resource management (DRM) software takes care of locating, scheduling, and executing the job on the appropriate computing resource.

Ubiquitous clustering technology

Since its origins in the late 1950s and early 1960s, HPC technology has taken many 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 are powerful tools for solving problems that are not generally partitionable into multiple pieces. 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, none of them developed into an industry-wide standard approach.

In the 1990s, clusters and grids of commercial off-the-shelf x86 and x64 servers began to replace "conventional" supercomputers and their more exotic architectures. With increasingly standard and open source based distributed resource management technologies — and open software development projects such as Open MP[1] and OpenMPI[2] — these systems have made HPC technology available for more applications and larger numbers of industries and people. In fact, clusters and grids now easily dominate the Top500 List of SuperComputer sites[3].

Compute cluster
The definition of a compute cluster is a rather loose one. It can consist of a small or large number of homogenous or heterogeneous systems interconnected with some form of network. Compute clusters are sometimes referred to as "Grids" or "Grid computing".

Compute clusters are now common among scientific research, educational and commercial institutions alike, mainly due to their lower cost of acquisition, greater 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 usefulness of clusters is due in large part to new technologies based on open systems and standards.

1. Open Multiprocessing is an application programming interface (API) that supports multiplatform shared-memory multiprocessing programming in C/C++ and FORTRAN on many architectures, including UNIX and Microsoft Windows platforms. For more information, please see `http://www.openmp.org/`
2. An open source implementation of the Message Passing Interface standard for communicating between HPC nodes. For more information, please see `http://www.open-mpi.org`
3. `http://www.top500org/stats/list/29/archtype`

Key technologies are now helping clusters scale while making the most of the resources involved:

- Distributed Resource Management technology such as Sun™ Grid Engine software helps assign computational resources appropriately, allowing the cluster to be shared effectively between multiple users and projects.
- Virtualization technology is increasingly important across systems, storage, and networking, allowing fine-grained assignment of resources where they are needed most.

Sun's end-to-end architecture for HPC

Effective and consistent end-to-end HPC infrastructure has been elusive for many organizations. Some approaches have been disappointing in their inability to start small and scale gracefully. Other approaches worked well as large complex deployments, but failed to be practical in smaller configurations. By developing a complete end-to-end HPC architecture that is designed to grow from a single rack to terascale and petascale deployments, Sun can easily deploy small and mid-range HPC clusters that can expand as required.

Grids

While synonymous with most implementations of HPC clusters, the word "grid" encompasses solutions not necessarily involving HPC resources. The term "grid" is increasingly being phased out as a definition for HPC.

One of Sun's stated design goals with regard to individual compute systems is a balanced architecture, where no single component acts as a bottleneck for the overall system. Sun's approach to HPC infrastructure is similarly focused on creating balanced and scalable cluster architecture that uses higher-level components — systems, networks, and storage — to their best potential. To this end, Sun's HPC Reference Architecture provides a full end-to-end solution for HPC infrastructure. While not all components are necessary in all deployments, the architecture provides the means to design and scale a solution in a tested and supported manner, delivering infrastructure that is ready to be deployed in the datacenter.

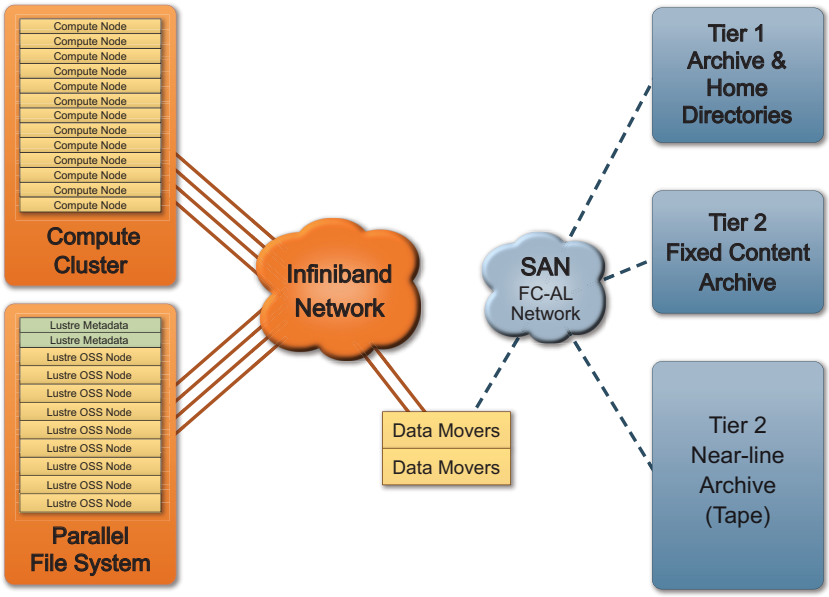

Figure 1. Schematic high-level perspective of Sun's High Performance Computing Reference Architecture

Sun's Reference Architecture for High Performance Computing (Figure 1) is designed to address the needs of the largest supercomputing clusters, but its components are also designed to address smaller clusters effectively. For larger configurations, an InfiniBand interconnect is key, offering the high throughput and low latency required for moving large amounts of data, and connecting compute clusters, and storage systems. Smaller clusters can use either InfiniBand or 10 Gb Ethernet networks. Compute clusters are comprised of powerful and dense servers in both rackmount and modular blade server configurations. A range of storage systems are available, employing either the Lustre™ parallel file system for fast simultaneous access to large amounts of data, or the scalable 128-bit Solaris ZFS™ file system for home directories and tier-1 archival.

Effective and flexible storage has never been more important. Today's powerful compute clusters can rapidly generate very large amounts of data. For larger clusters, Sun's Reference Architecture for HPC provides specialized Data Movers to help transition data from the cluster scratch space (storage cache) where it is created and processed, to a Storage Area Network where appropriate storage systems are matched to the needs of specific data. General-purpose storage systems serve both Tier-1 archives and home directories. Specialized fixed-content storage systems serve data that doesn't change often. Near-line archiving to tape is provided to make the most of available disk storage while keeping archived data close at hand in case it is needed.

Proven reference designs for HPC

Based on customer-driven projects, Sun has developed a number of proven reference designs for HPC. These architectures[1] range from departmental solutions to supercomputing clusters in the 500+ TeraFLOPS range such as the Ranger supercomputing cluster at the Texas Advanced Computing Center (TACC)[2]. In addition, Sun's commitment to eco-efficient IT also combines the most power-efficient solutions and cooling technologies, helping to enable sustainable system design. Important to Sun's HPC Reference Architecture, many of the same components are used for both small and large clusters, providing consistency, investment protection, and innovation across the range of cluster scenarios. For more information of Sun HPC solutions, reference designs, and customer references please visit `http://www.sun.com/hpc`.

Using dense and modular compute elements powered by the latest x64 and SPARC® CPU architectures, these clusters employ InfiniBand and Ethernet switching technology, scalable and parallel file systems, and powerful storage server technologies. Combined with integrated and supported open source Sun HPC Software, Sun's reference designs for HPC provides a simple and fast methodology for building large HPC solutions. Sun provides a broad range of solutions in the context of the larger reference architecture, described in the following sections.

Solutions for larger clusters

For larger clustered HPC environments, the *Sun Constellation System* provides integrated system building blocks that allow for rapid cluster deployment.

- The *Sun Blade™ 6048 Modular System* provides a dense and flexible blade platform with support for up to 48 server modules and up to 96 compute nodes n a single rack. A choice of Intel® Xeon®, AMD Opteron™, and Sun SPARC processors are provided, and operating system support includes the Solaris OS, Linux, and Microsoft Windows.
- Along with the Sun Blade 6048 Modular System, *Sun Datacenter Switches (DS)* provide compact and scalable support for both dual data rate (DDR) and quad data rate (QDR) InfiniBand Interconnects.
- For clustered storage, the *Sun Fire™ X4540 storage server* offers support for two quad-core x64 processors and up to 48 terabytes of tightly-coupled storage in only four rack units (4U). The Sun Fire X4540 server is ideal for deploying the Lustre parallel file system to serve as scratch space for the compute cluster.

---

1. For an indepth description of petascale solutions, see: `http://www.sun.com/servers/hpc/docs/pathways_to_petascale.pdf`
2. `http://www.tacc.utexas.edu/ta/ta_display.php?ta_id=100379`

Solutions for small-to-medium clusters

For small to medium clusters, Sun provides a series of configurable Sun Compute Cluster configurations that scale from a single rack upwards. Different Sun Compute Clusters are offered for specific application spaces as described below, and are constructed from Sun Blade 6000 Modular Systems, Sun Fire x64 rackmount systems based on the latest Intel Xeon and AMD Opteron processors, and Sun SPARC Enterprise® servers based on the chip multithreaded (CMT) UltraSPARC® T2 and T2 Plus processors.

- The *Sun Compute Cluster* is a streamlined HPC solution specifically designed to help solve the most computationally-intensive business or scientific problems. Consisting of either rackmount servers or blade servers along with networking, interconnects, and software, the Sun Compute cluster represents an easy-to-deploy building block for building small to medium clusters.
- The *Sun Compute Cluster for Research* arrives ready to switch on for more productivity and faster time to results. The Sun Compute Cluster for Research starts with 16 rackmount servers per rack or 48 blade servers per rack, and optionally adds Ethernet or InfiniBand interconnects, as well as storage components for the Lustre file system and management nodes.
- The *Sun Compute Cluster for Mechanical Computer Aided Engineering (MCAE)* provides an integrated solution for deploying mechanical analysis applications. The Sun Compute Cluster for MCAE starts with Sun Blade X6250 server modules hosting quad-core Intel Xeon processors, and adds management nodes, 10 Gb Ethernet networking, and storage capacity in the form of the Sun Storage 7000 Unified Storage System.
- The *Sun Compute Cluster for High-Performance Trading* lets trading firms innovate to increase the transaction speeds of their trading platforms, all consolidated into a single rack. Utilizing the Sun Blade 6000 Modular System chassis, the cluster integrates Sun Blade X6250 server modules equipped with Sun Blade 6000 disk modules. Open network storage is provided in the form of the Sun Storage 7000 Unified Storage System along with Gigabit Ethernet networking.

Figure 2 illustrates a typical smaller-scale cluster based on the Sun Blade 6000 Modular System. In this example, a single rack provides up to 30 Sun Blade 6000 server modules and up to 120 processor sockets for x64 processors. Rackmount servers provide file services, distributed resource management, and administrative services such as server and application provisioning. Separate networks address the need for administration and provisioning, as well as file services and job submission. Optional high-speed, low-latency interconnects based on Ethernet or InfiniBand provide for interprocess communication for multiprocess applications.

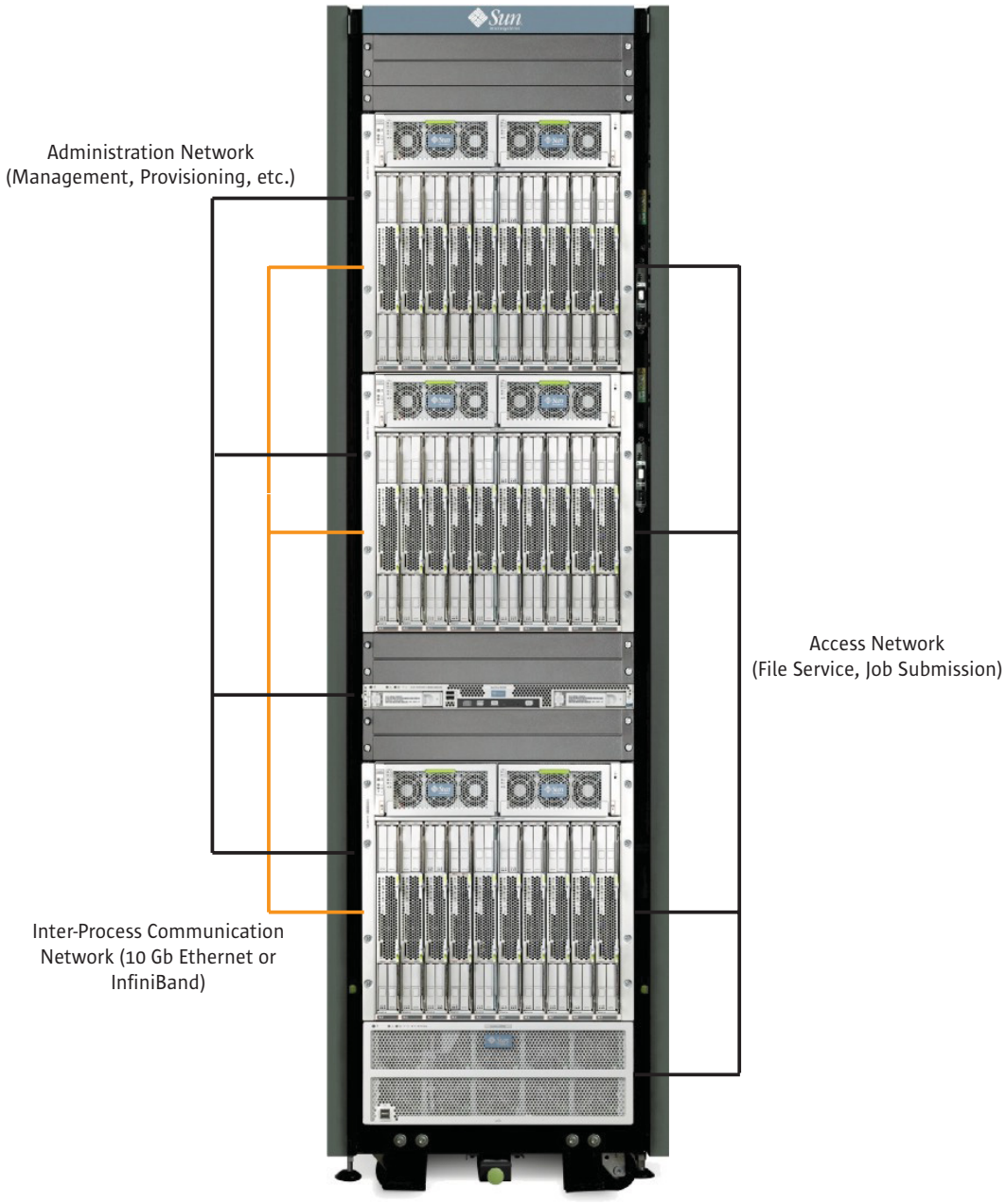

Figure 2. Smaller-scale clusters can be built using the Sun Blade 6000 Modular System and a range of rackmount servers

Storage and storage clustering solutions

Storage is especially important for HPC environments, and Sun provides specific modular solutions that address the diverse needs of HPC storage infrastructure.

- *The Sun Storage Cluster* employs Sun Open Storage products and the Lustre file system to eliminate I/O bottlenecks and provide very high rates of I/O bandwidth, scaling, and capacity. Sun Fire X4540 storage servers are deployed to support components of the Lustre file system with Sun Storage J4400 arrays provided for additional JBOD (just a bunch of disks) storage capacity.
- *Sun Unified Storage Solutions for HPC* can help accelerate workflows by providing shared access to all stored data, such as user directories and tier-1 archive data. By providing an innovative use of enterprise flash technology in Solaris ZFS hybrid storage pools, the Sun Storage 7000 Unified Storage System can dramatically reduce latency and I/O performance bottlenecks for applications.
- Beyond getting ever-increasing amounts of data out of the cluster, organizations need to be able to store and archive large amounts of data in a manner that is both simple to manage, and economical to run. By utilizing open standards, the *Sun Storage and Archive Solution for HPC* provides a modular and scalable approach utilizing Sun StorageTek™ Storage Archive Manager, Sun StorageTek QFS software, and the Solaris 10 OS.

Chapter 2
Compute Systems

The choice of computational architecture should always be based on the intended applications. The set of applications determines the operating system and will generally also help guide the hardware decisions. For example, if Microsoft Windows applications are a requirement, x86/x64 based hardware will be needed. Other applications may only be available for for the Solaris OS or Linux. Multithreaded workloads with larger shared-memory needs may scale best on chip multithreaded (CMT) technology such as UltraSPARC T2 and T2 Plus processors. In simple cases solutions can be built with a single OS and hardware platform, but multiple operating systems will typically be required for all but the most simple clusters.

General requirements

Correctly determining the requirements for the compute node is paramount. Many legacy HPC solutions were comprised of only a single compute node, usually a specialized and expensive vector processor. Today most HPC solutions consist of a cluster of smaller nodes, connected by some kind of network or interconnect. In addition to the OS, the applications that will run on the cluster also help dictate the hardware requirements for the nodes.

Highly-parallel applications may be able to take advantage of individual systems with many processor sockets and cores. Less parallel applications may be best served by smaller numbers of faster processors. Application memory requirements are also important. Reading from memory is 10,000 times faster than reading from disk, so sufficient memory is critical for all systems, particularly in HPC solutions. If an HPC cluster is built for multiple applications or multiple problem sizes, a range of compute nodes is often desired — some with more processors and memory and some with less.

x86/x64 servers for HPC

x86 and x64
x86 is the common name for Intel's most successful instruction set used in most desktop and servers over the last 10 years. x64, while adhering to the x86 instruction set, is AMD's name for 64 bit capable CPUs.

Most HPC solutions today are built from fairly small, two-socket to four-socket x86/x64 rackmount servers or blade servers. Off-the-shelf systems are often used, though the present trend is to use systems that are specifically designed for HPC. Configurations vary from a single large machine, to a handful of very large servers, to thousands of small servers or blade servers. In some cases spare resources on existing desktop or other systems are made available for those who need more resources for a job.

The trend today is away from specialized hardware, and as a result, more and more HPC solutions are based on general-purpose processors. Besides the instruction set, the performance characteristics of different CPUs are important. In the x86/x64 space, there are several processor families available from both Intel and AMD. These

processors have some architectural differences but the main execution pipeline is virtually identical. Actual application performance will vary with each application, selected processors, and system architecture. With a broad range of Sun x86/x64 based rackmount and blade servers based on the latest Intel Xeon and AMD Opteron processors, organizations can populate their clusters with the compute nodes that best serve their application needs.

Sun servers offer from one to four sockets for Intel Xeon processors, including support for the latest Intel Xeon Processor 5500 Series CPUs (formerly Nehalem). A sampling of Sun rackmount and blade servers based on Intel Xeon processors is shown in Table 1.

Table 1. A range of Sun servers suitable for HPC infrastructure based on Intel Xeon processors

Server	Processors	Memory Capacity
Sun Blade X6275 server module[a]	Two nodes per server module, each with two Intel Xeon Processor 5500 Series CPUs	Up to 96 GB per compute node
Sun Blade X6270 server module[b]	Two Intel Xeon Processor 5500 Series CPUs	Up to 144 GB
Sun Blade X6250 server module[b]	Up to two dual-core or quad-core Intel Xeon processors	Up to 64 GB
Sun Blade X6450 server module[b]	Two or four dual-core, quad-core, or six-core Intel Xeon processors	Up to 192 GB
Sun Fire X2270 server (1U)	One or two Intel Xeon Processor 5500 Series CPUs	Up to 48 GB
Sun Fire X2250 server (1U)	Up to two dual-core or quad-core Intel Xeon processors	Up to 32 GB
Sun Fire X4170 server (1U)	One or two Intel Xeon Processor 5500 Series CPUs	Up to 144 GB
Sun Fire X4270 server (2U)	One or two Intel Xeon Processor 5500 Series CPUs	Up to 144 GB
Sun Fire X4275 server (2U)	One or two Intel Xeon Processor 5500 Series CPUs	Up to 144 GB
Sun Fire X4150 server (1U)	One or two dual-core or quad-core Intel Xeon processors	Up to 64 GB
Sun Fire X4250 server (2U)	Two dual-core or quad-core Intel Xeon processors	Up to 64 GB
Sun Fire X4450 server (2U)	Up to four dual-core, quad-core, or six-core Intel Xeon processors	Up to 128GB

a.For configuration in the Sun Blade 6048 Modular System chassis
b.For configuration in either the Sun Blade 6000 or 6048 Modular System chassis

Sun x64 servers based on AMD Opteron processors offer from one to eight sockets and support the latest Third-Generation Quad-Core AMD Opteron processors. Sun rackmount and blade servers based on AMD Opteron processors are shown in Table 2.

Table 2. A range of Sun servers suitable for HPC infrastructure based on AMD Opteron processors

Server (Enclosure)	Processors	Memory Capacity
Sun Blade X6240 server module[a]	Two Enhanced Quad-Core AMD Opteron Series 2300 processors	Up to 64 GB
Sun Blade X6440 server module[a])	Four Quad-Core or Enhanced Quad-Core AMD Opteron Series 8300 processors	Up to 256 GB
Sun Fire X2200 M2 server (1U)	One Quad-Core or Enhanced Quad-Core AMD Opteron Series 2000 processor	Up to 64 GB
Sun Fire X4140 server (1U)	One or two Quad-Core, or Enhanced Quad-Core AMD Opteron Series 2000 processors	Up to 128 GB
Sun Fire X4240 server (2U)	One or two Quad-Core, or Enhanced Quad-Core AMD Opteron Series 2000 processors	Up to 128 GB
Sun Fire X4440 server (2U))	Two or four Dual-Core, Quad-Core, or Enhanced Quad-Core AMD Opteron Series 8000 processors	Up to 256 GB
Sun Fire X4540 server (4U)	Two Quad-Core AMD Opteron Series 2000 processors	Up to 64 GB
Sun Fire X4600 server (4U)	Two, four, six, or eight Quad-Core or Enhanced Quad-Core AMD Opteron Series 8000 processors	Up to 512 GB

a.For configuration in the Sun Blade 6000 or 6048 Modular System

Chip multithreaded (CMT) SPARC® systems for HPC

Sockets

As more and more CPU architectures have multiple cores (presently up to eight) CPU sockets are becoming the basic measurement of how many compute elements a system can have.

Though not traditionally considered for HPC workloads, Sun's chip multithreaded (CMT) systems can offer advantages for some applications and environments. Systems based on the UltraSPARC T2 and T2 Plus processors in particular are proving ideal for applications such as string matching. In searching for patterns, string-matching algorithms represent an interesting challenge because they must traverse a large data structure in memory, and must effectively share the data structure among multiple threads. With the ability to support a large number of threads and rapidly switch between contexts, these systems effectively hide memory latency, and present a convenient general-purpose programming model.

Each UltraSPARC T2 and T2 Plus processor offers considerable computational resources. Up to eight cores are provided in each processor, with each core able to rapidly switch between up to eight threads as they block for memory access. In addition, each core provides two integer execution units, and a floating-point unit, so that a single

UltraSPARC T2 or T2 Plus processor core is capable of executing up to two threads at a time. Applicable Sun servers based on UltraSPARC T2 and T2 Plus processors are shown in Table 3.

Table 3. A range of Sun servers suitable for HPC infrastructure based on UltraSPARC T2 and T2 Plus processors

Server (Enclosure)	Processors	Memory Capacity
Sun Blade T6320 server module[a]	One quad-core, six-core, or eight-core UltraSPARC T2 processor	Up to 128 GB
Sun Blade T6340 server module[a]	Two six-core or eight-core UltraSPARC T2 Plus processors	Up to 256 GB
Sun SPARC Enterprise T5120 server (1U)	One quad-core, six-core, or eight-core UltraSPARC T2 processor	Up to 128 GB
Sun SPARC Enterprise T5220 server (2U)	One quad-core, six-core, or eight-core UltraSPARC T2 processor	Up to 128 GB
Sun SPARC Enterprise T5140 server (1U)	Two quad-core, six-core, or eight-core UltraSPARC T2 Plus processors	Up to 128 GB
Sun SPARC Enterprise T5240 server (2U)	Two quad-core, six-core, or eight-core UltraSPARC T2 Plus processors	Up to 256 GB
Sun SPARC Enterprise T5440 server (2U)	One to four eight-core UltraSPARC T2 Plus processors	Up to 512 GB

a.For configuration in the Sun Blade 6000 or 6048 Modular System

Chapter 3

Interconnects and High-Performance Networks

Once the specific compute node architecture is defined, clustered systems must be connected together into a network. In fact, interconnects and high-performance networks between the compute nodes to some degree form the basis of many HPC systems. The specific requirements for the interconnect depend on the characteristics of the applications, and the size of the cluster. In some cases a standard 100 Mbit Fast Ethernet is sufficient while in other cases a higher-performance interconnect is required. Some systems may even have multiple networks. For example, a Gigabit Ethernet network might be used for job submission and user files, with a low-latency InfiniBand (IB) network used for interprocess communication and for scratch file transfer. Sun provides both Ethernet and InfiniBand network technology to match a wide range of HPC clustering needs — including both adaptors and switches.

Ethernet

Since its invention at Xerox PARC in 1973, Ethernet local area networking has been the dominant local area networking (LAN) technology for both wired and wire-less networking. Ethernet is now making inroads into new application areas ranging from clustering and storage networks to wide area networking. The initial speed of Ethernet was 10 Mb/sec. Fast Ethernet (100 Mb/sec) was adopted as an IEEE standard in 1995. With the explosion of the Internet and with the need to move larger and larger amounts of data, Gigabit Ethernet (GbE) became an IEEE standard (802.3z) that was adopted in 1998. A key benefit of GbE was the usage of the same Category-5 copper cables deployed for Fast Ethernet, thus preserving infrastructure investment. The need for more bandwidth become more apparent in the following years, and in 2002, the IEEE ratified the 10 Gigabit Ethernet standard (802.3ae). This standard defines a version of Ethernet with a nominal data rate of 10 Gb/sec. Both Gigabit Ethernet and and 10 Gb Ethernet are switched network technologies.

Generally speaking, Ethernet is typically used for applications with minor communication demand, often referred to as embarrassingly parallel applications. In contrast, InfiniBand is typically used for applications requiring low latency (1 us), high-throughput (20 - 120 Gb/sec), low CPU overhead (less than 10%), and high scalability.

InfiniBand

InfiniBand was first standardized in October of 2000. Since that time, InfiniBand technology has emerged as an attractive fabric for building large supercomputing clusters and storage systems — where high bandwidth and low latency are key requirements. As an open standard, InfiniBand presents a compelling choice over

proprietary interconnect technologies that depend on the success and innovation of a single vendor. Similar to gigabit Ethernet and 10 Gb Ethernet, InfiniBand is a serial point-to-point full-duplex interconnect.

InfiniBand also presents a number of significant technical advantages, including:

1. Loss-less network fabric with flow-control to avoid packet loss due to buffer overflow — negating the need for re-transmission and improving general performance
2. Congestion Management to improve fabric performance
3. Service differentiation through Virtual Lanes (VL) to help enable quality of service (QoS) — allowing for fair allocation of bandwidth, and making it possible to separate latency-sensitive traffic from other traffic through high and low priorities
4. Multipath routing to help enable load balancing and redundancy
5. Virtual cut-through routing for lowest latency packet forwarding
6. Host channel adaptors with Remote Direct Memory Access (RDMA), supporting the off-loading of communications processing from the operating system, leaving more CPU resources available for computation

Not only does InfiniBand perform well today, but the InfiniBand Trade Association (IBTA) has established a performance road-map to accommodate future demands for bandwidth. InfiniBand links consist of groups of lanes (1x, 4x, 8x, and 12x) that support a number of different data rates, with 4x being the most common. Supported data rates and link widths are given in Table 4.

Table 4. InfiniBand data rates

Data Rates	**Per Lane and Per Link Bandwidth**			
	1x	**4x**	**8x**	**12x**
Single Data Rate (SDR) 2.5 Gbps	2.5 Gbps	10 Gbps	20 Gbps	30 Gbps
Dual Data Rate (DDR) 5 Gbps	5 Gbps	20 Gbps	40 Gbps	60 Gbps
Quad Data Rate (QDR) 10 Gbps	10 Gbps	40 Gbps	80 Gbps	120 Gbps
Eight x Data Rate (EDR) 20 Gbps	20 Gbps	80 Gbps	160 Gbps	240 Gbps
Hexadecimal Data Rate (HDR) 40 Gbps	40 Gbps	160 Gbps	320 Gbps	480 Gbps

Interconnect fabrics for clusters have traditionally been based upon either mesh, torus, or Clos topologies.

Torus topologies

Some of the most notable supercomputers based upon Torus topologies include IBM's BlueGene and Cray's XT3/XT4 supercomputers. Torus topologies are referred to as direct networks, in which each node contains an embedded switch and connects to its neighbors. For example, in a 3D-Torus topology, each node connects to its neighbors in the the X, Y, and Z direction. Torus fabrics are generally considered to be easy to build, in that the cabling and cable distance is generally less than that of Clos topologies.

The disadvantage to Torus topologies is a lack of bandwidth and a variable hop count that leads to differences in latency depending on the location of the server in the Torus fabric. The variable hop count implies that node locality needs to be carefully considered during application deployment. However, for some specific applications that express a nearest-neighbor type of communication pattern, a torus topology is a good fit. Computational fluid dynamics (CFD) is one such application.

Clos topologies

First described by Charles Clos in 1953, Clos networks have long formed the basis for practical circuit-connected multi-stage telephone switching systems. These topologies have also evolved to include packet-based switching. Clos networks utilize a "Fat Tree" topology, allowing complex switching networks to be built using many fewer cross points 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 built from 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 — in terms of the number of ports available in individual switch elements, maximum printed circuit board size, and maximum connector density. Sun Datacenter Switches address these issues directly.

Sun Datacenter Switches for InfiniBand infrastructure

As a part of the Sun Constellation System, Sun Datacenter Switches provide dense, compact, and scalable InfiniBand fabrics. Tightly integrated with the Sun Blade 6048 Modular System, these switches employ innovative architecture, consolidated industry-standard 12x cabling, and switched PCI Express Network Express Modules (NEMs). From a high level, Sun Datacenter Switches are comprised of the following key components:

- Ultra-scalable DDR or QDR InfiniBand switching fabrics
- High-performance host adapters (NEMs)
- Industry-standard high-density cabling systems
- An open management platform

Sun Datacenter switches are available to implement both DDR and QDR InfiniBand fabrics. Various members of the Sun Datacenter Switch family are described in the sections that follow.

- The Sun "Magnum M9" switch provides DDR/QDR connections for up to 648 individual compute nodes, when combined with the Sun Blade 6048 Modular System and the Sun Blade 6048 InfiniBand QDR Switched Network Express Module (NEM). Together with industry-standard 12x cables and connectors, multiple Magnum M9 switches can connect up to 5,184 compute nodes.
- Deployed in a single datacenter cabinet, the Sun DS 3456 connects up to 3,456 nodes, and implements a maximal three-tier, five-stage DDR Clos fabric as depicted in Figure 3. Each 144-port output in the figure represents a single line card deployed in the switch chassis. Multiple Sun DS 3456 can be combined to connect up to 13,824 nodes.

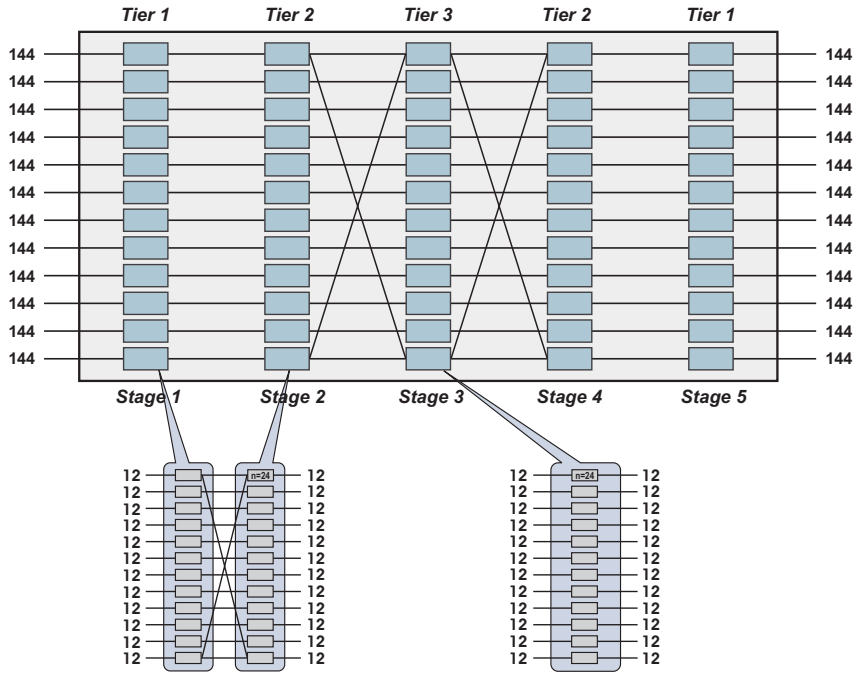

Figure 3. A 3-tier, 5-stage Clos fabric can achieve single-system connectivity of 3,456 ports using 144 24-port switch chips per stage and 720 chips total.

- Building on the considerable strengths of the Sun DS 3456, the Sun DS 3x24 helps to reduce the cost and complexity of delivering small to moderate-scale HPC clusters. The Sun DS3x24 offers 72 InfiniBand 4x DDR ports in a 1 rack unit (1U) 19-inch chassis. Figure 4 provides a high-level perspective of a 288-node three-stage Clos fabric implemented using the Sun DS 3x24. Each orange box depicts a single Sun DS 3x24 with its three switch chips. Each Sun Blade 6048 InfiniBand Switched NEM provides an additional Mellanox InfiniScale III switch chip that connects to the 12 Sun Blade 6000 server modules in each of the four shelves in a Sun Blade 6048 modular system chassis.

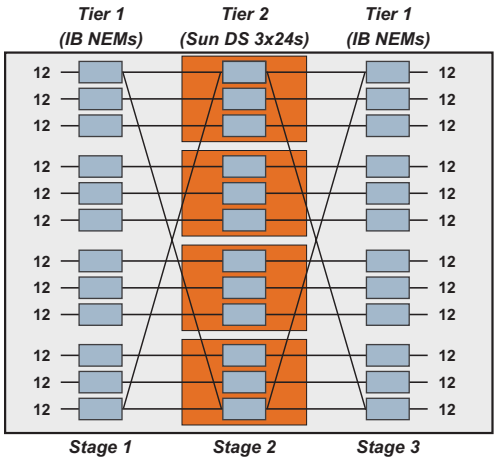

Figure 4. Together with Sun Blade 6048 InfiniBand Switched NEMs, four Sun DS 3x24 can implement a 288-node three-tier, three-stage Clos fabric

While these switches are principally designed for the Sun Blade 6048 Modular System, Sun offers adapter cables to integrate rackmount servers into InfiniBand clusters. Sun also provides third-party InfiniBand switches that are appropriate for smaller clusters based on rackmount servers.

More information on the Sun Blade 6048 modular system, Sun InfiniBand Switched NEMs, and multiple switch deployments is provided in the *Sun Datacenter Switch 3456 and 3x24 Architecture white paper*. Additional information on InfiniBand technology is provided in the Sun white paper entitled: *Making Clusters Easier with Sun InfiniBand Technology*.

Chapter 4

File Systems and Archive Solutions for HPC

Data management in HPC environments can present significant challenges. With growing clusters and more powerful servers, getting data in and out of the cluster is more challenging than ever. HPC applications need fast access to high-capacity data storage for writing and retrieving data. Many HPC clusters also require scalability, both in the amount of data, and the numbers of nodes that require access. In addition, data needs to be stored, managed, and archived in ways that are both straightforward and economical for the organization.

Sun has extensive experience developing file systems, from the Network File System (NFS) to the Lustre parallel file system, the scalable 128-bit ZFS file system, and Sun StorageTek QFS. This chapter describes parallel and scalable file system technology appropriate for implementing cluster storage, and also describes the Sun Storage and Archive Solution for HPC.

Network File System (NFS)

In most smaller cluster environments, the cost and complexity of implementing a scalable file system may not justify any potential performance gain. As a result, traditional file systems such as the Network File System (NFS) can be used for small clusters rather than the Lustre file system or other parallel file systems. NFS can typically serve up to 128 nodes, although the actual number depends on the throughput requirements of the applications being run on the cluster. One exception to this general rule is small clusters with requirements for high I/O performance per node. Clusters such as these may need a parallel file system to meet performance goals.

NFS is a network file system protocol originally developed by Sun Microsystems in 1984, allowing a user on a client computer to access files over a network as easily as if the network devices were attached to local disks. In compute clusters, this ability lets all servers access a common pool of resources (such as applications, libraries, and data) using standard Ethernet connectivity. Depending on the applications and their data requirements, NFS over standard Gigabit Ethernet is often good enough to provide a core fabric of communication for a smaller compute cluster.

Lustre™ parallel file system

The Lustre parallel file system[1] is an open source, shared file system designed to address the I/O needs of compute clusters containing up to thousands of nodes. The Lustre file system is intended for environments where traditional shared file systems,

1. For and in-depth description of the Lustre parallel file system, see `http://www.sun.com/software/products/lustre/docs/lustrefilesystem_wp.pdf`

(such as NFS) do not scale to the required aggregate throughput, or to support a sufficiently large number of nodes. Larger clusters using low-latency, high-speed interconnects will generally require a scalable parallel file system like the Lustre file system or Sun StorageTek QFS.

The Lustre file system is a software-only architecture that supports a number of different hardware implementations. The main components of a Lustre architecture are the Lustre file system clients (Lustre clients), Metadata Severs (MDS), and Object Storage Servers (OSS). Lustre clients are typically compute nodes in HPC clusters. These nodes run Lustre client software, and access the Lustre file system via InfiniBand (or Ethernet) connections. Metadata Servers and Object Storage Servers implement the file system and communicate with the Lustre clients (Figure 5).

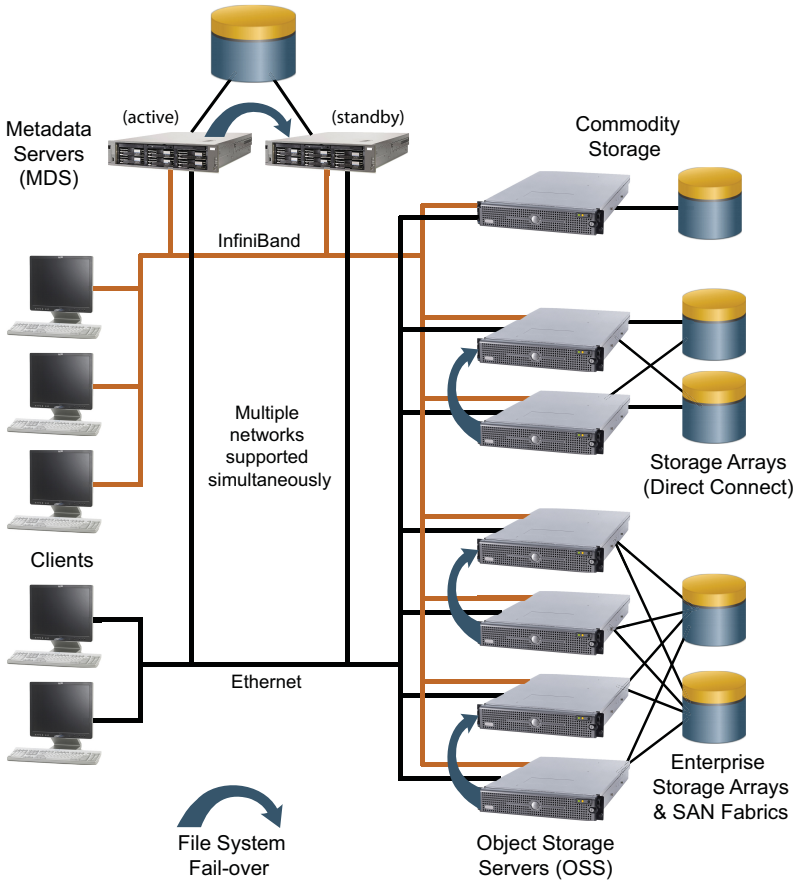

Figure 5. High-level overview of the Lustre file system architecture

The Lustre file system employs an object-based storage model, and provides several abstractions designed to improve both performance and scalability. At the file-system level, Lustre technology treats files as objects that are located through Metadata Servers. Metadata Servers support all file system name space operations, such as file lookups, file creation, and file and directory attribute manipulation. File data is stored

in objects on the OSSs. The MDS directs actual file I/O requests from Lustre clients to OSSs, that in turn manage the storage that is physically located on underlying storage devices. Once the MDS identifies the storage location of a file, all subsequent file I/O is performed between the client and the OSSs.

This design divides file system updates into two distinct types of operations:

- File system metadata updates on the Metadata Servers, and
- Actual file data updates on the Object Storage Servers.

Separating file system metadata operations from actual file data operations not only improves immediate performance, but it also improves long-term aspects of the file system such as recoverability and availability. Lustre technology can support a variety of configuration options including a choice of interconnects, single or dual MDS, and different storage attachment methods for the Object Storage Servers.

Sun Storage Cluster and the Lustre parallel file system

The Sun Storage Cluster[1] is intended for environments that need a scalable high-performance file system (such as the Lustre file system) for a large cluster of compute nodes. In general, these environments contain more than 256 nodes, require large data throughput, and typically use a low-latency high-speed interconnect based on InfiniBand. The Sun Storage Cluster supports Lustre clients using InfiniBand interconnects, dual Metadata Servers, and commodity storage servers for the OSS.

The Sun Storage Cluster is designed to deliver key functionality for HPC clusters, including:

- *Highly-scalable throughput* — The solution delivers high data throughput and near linear scalability of up to 95 % in real world deployments. Configurations can scale by adding Sun Fire X4540 storage servers for incremental throughput.
- *Increased data density* — Sun storage solutions such as the Sun Fire X4540 storage server offer dense storage with up to 48 TB in only four rack units.
- *Quick deployment* — The design and architecture are based on Sun's experience gained building large-scale HPC storage solutions. Deploying factory-integrated HPC storage solutions based on configurations field-tested by Sun is far quicker than custom-built solutions. A fully-integrated hardware and software stack — configured and tested before delivery — help reduce both risk and time to deployment.
- *Scalable, modular design* — The modular design of this approach lets organizations start small and expand as their storage needs increase.

1. http://www.sun.com/hpc/storagecluster/

Designed as “fast scratch space”, solutions such as the Sun Storage Cluster provides the data storage capacity and throughput for large clusters. The Sun Storage Cluster architecture is depicted from a high level in Figure 6.

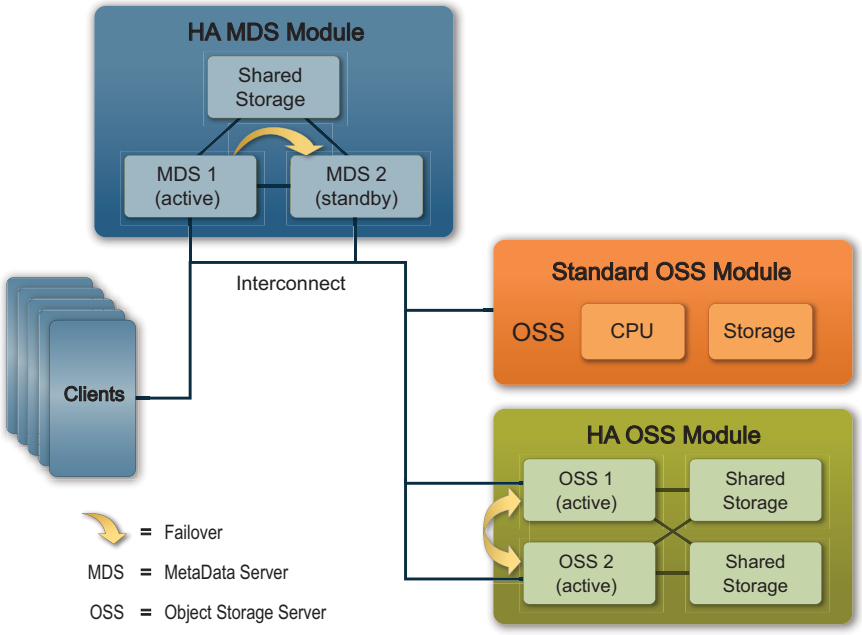

Figure 6. The Sun Storage Cluster implements the Lustre parallel file system

Key components of this integrated data storage solution include:

- High performance Sun Fire X4250 servers as Lustre MDS modules
- High-performance Sun Fire X4540 storage servers as Lustre OSS modules and High Availability OSS modules
- Sun Storage J4400 arrays offering inexpensive JBOD (just a bunch of disks) technology
- Linux operating systems, Lustre file system software, and the Open Fabrics Enterprise Distribution (OFED), along with monitoring and management software

Solaris™ ZFS™ and Sun Storage 7000 Unified Storage Systems

Beyond the data caching file requirements of compute clusters, HPC environments need storage infrastructure for user’s home directories and for tier-1 archival. Reliability, scalability, and robustness are paramount for these file systems as they can represent an organization’s most important data — including specifications, completed designs, and data from successful simulation runs. Coupling the scalable Solaris ZFS file system with optimized servers and storage components, the Sun Storage 7000 Unified Storage System represents an ideal solution.

In addition to the scalability provided by a 128-bit file system, Solaris ZFS provides extra reliability and robustness, even protecting valuable data from silent data corruption. Sun Storage 7000 Unified Storage Systems incorporate an open-source operating

system, commodity hardware, and industry-standard technologies. These systems represent low-cost, fully-functional network attached storage (NAS) storage devices designed around the following core technologies:

- General-purpose x64-based servers (that function as the NAS head), and Sun Storage products — proven high-performance commodity hardware solutions with compelling price-performance points
- The general-purpose OpenSolaris™ Operating System
- The Solaris ZFS file system, the world's first 128-bit file system with unprecedented availability and reliability features
- The high-performance Solaris networking stack using IPv4 or IPv6
- Solaris DTrace Analytics, that provide dynamic instrumentation for real-time performance analysis and debugging
- Sun Fault Management Architecture (FMA) for built-in fault detection, diagnosis, and self-healing for common hardware problems
- A large and adaptive two-tiered caching model, based on DRAM and enterprise-class solid state devices (SSDs)

Sun Storage 7000 Unified Storage Systems rely heavily on Solaris ZFS for key functionality such as Hybrid Storage Pools — combining the strengths of system memory, enterprise flash technology, and commodity disk storage. By automatically allocating space from pooled storage when needed, Solaris ZFS simplifies storage management and gives organizations the flexibility to optimize data for performance. Sun Storage 7000 Unified Storage Systems utilize ZFS Hybrid Storage Pools to automatically provide data placement, data protection, and data services such as RAID, error correction, and system management. By placing data on the most appropriate storage media, Hybrid Storage Pools help to optimize performance and contain costs.

Key capabilities of Solaris ZFS related to hybrid storage pools include:

- *Virtual storage pools* — Unlike traditional file systems that require a separate volume manager, Solaris ZFS introduces the integration of volume management functions.
- *Data integrity* — Solaris ZFS uses several techniques to keep on-disk data self consistent and eliminate silent data corruption, such as copy-on-write and end-to-end checksumming.
- *High performance* — Solaris ZFS simplifies the code paths from the application to the hardware, delivering sustained throughput at near platter speeds.
- *Simplified administration* — Solaris ZFS automates many administrative tasks to speed performance and eliminate common errors.

Sun Storage 7000 Unified Storage Systems feature a common, easy-to-use management interface, along with a comprehensive DTrace analytics environment to help isolate and resolve issues. The systems support NFS, CIFS, and iSCSI data access protocols, mirrored and parity-based data protection, local point-in-time (PIT) copy, remote replication, data checksum, data compression, and data reconstruction.

To meet varied needs for capacity, reliability, performance, and price, the product family includes three different models — the Sun Storage 7110, 7210, and 7410 Unified Storage Systems. Configured with appropriate data processing and storage resources, these systems can support a wide range of requirements.

Sun Storage and Archive Solution for HPC

Most if not all active data used by the compute cluster is stored and accessed via the data cache system. However, given the large amounts of data created by typical compute clusters, longer-term storage and archiving is often required. As mentioned, storage for home directories is also needed. There are many different approaches to this problem and the technologies outlined below are only intended as a high level overview[1].

As with other components of the reference architecture, Sun 's HPC storage solutions are designed as part of an overall cluster solution. Storage solutions integrate with other complementary products, such as the Sun StorageTek SAM-QFS file system, that provide long-term data archive capabilities. For large clusters, high-performance servers act as Data Movers, efficiently moving data between the fast scratch storage of the Lustre parallel file system (or other file system) and long-term storage.

Hierarchical storage management

Hierarchical Storage Management (HSM) is a data storage technique that automatically moves data between high-cost and low-cost storage media. Sometimes referred to as tiered storage, HSM systems store the bulk of the organization's data on slower devices, such as SATA drives or tape. Data is copied to the faster cluster scratch space when needed.

In a typical HSM scenario, data files that are frequently used are stored on disk drives, but are eventually migrated to tape if they are not used for a certain period of time, typically a few months. If a user does reuse a file that has been migrated to tape, it is automatically moved back to disk storage. The advantage of this approach is that the total amount of stored data can be much larger than the capacity of the disk storage available. Since only rarely-used files are on kept on tape, most users will usually not notice any delay.

Sun Storage and Archive Solution architecture

The Sun Storage and Archive Solution for HPC performs a number of key functions for an HPC environment:

- Serving data to the compute cluster
- Storing and protecting data on various tiers

1. For details on HPC archiving please see:
   `http://www.sun.com/servers/hpc/docs/Sun_HPC_Ref_Arch.pdf`

- Providing a policy-based hierarchical storage manager to automate the movement of data across the storage tiers
- Implementing a continuous copy mechanism to protect critical data and alleviate the need for traditional backup applications

To provide flexibility and scaling, the solution is architected from a series of modules or building blocks as shown in Figure 7. These modules are connected to the various cluster networks as appropriate, and modules can be selected based on a particular set of needs. It is important to note that the modular approach allows individual modules to be swapped out based on the size and requirements of the cluster.

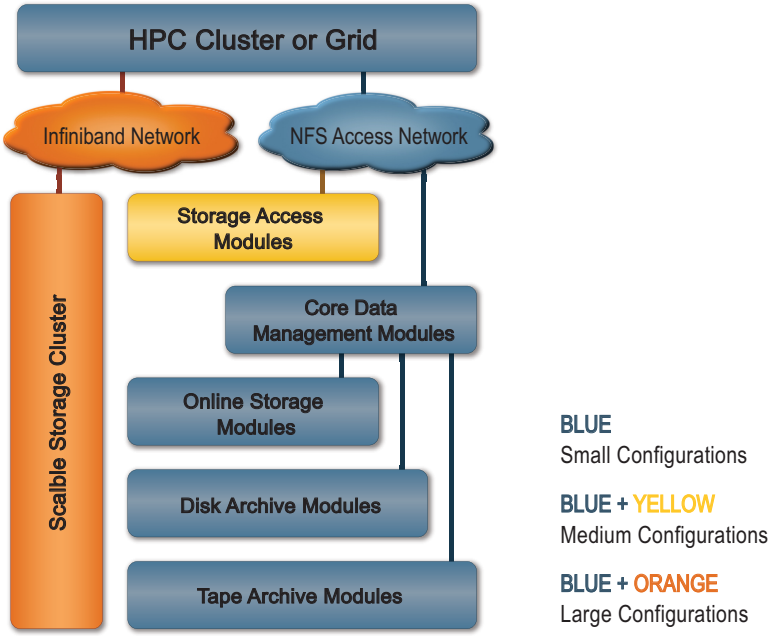

Figure 7. Sun Storage and Archive Solution architecture

Key modules of the Sun Storage and Archive Solution include:

- *Core Data Management Modules* — Consisting of Sun Fire servers running Sun StorageTek Storage Archive Manager and the Solaris 10 OS, the Core Data Management Modules control the overall operation of the solution. These modules manage the location and movement of data based on policies set by administrators, and manage metadata related to the environment.
- *Storage Access Modules* — Optional Storage Access Modules are comprised of Sun Fire servers running Sun StorageTek QFS software and the Solaris 10 OS, and can be deployed to deliver high levels of I/O performance to the HPC cluster. When deployed, these modules provide the compute cluster I/O services while the Core Data Management Modules continue to serve home directories and manage the overall environment. As needs grow, performance can be scaled by adding more modules. For clusters that require higher performance, the Sun Storage Cluster can

be substituted for Storage Access Modules. If the Sun Storage Cluster is used for cluster scratch space, the Online Disk Modules are sized only for user home directories.

- *Online Storage Modules* — Consisting of Sun modular storage, the Online Storage Modules provide reliable high-performance storage for user home directories, and metadata related to the environment. In small or medium sized configurations, Online Storage Modules can also provide cluster scratch space. Since the I/O profile of home directories, metadata structures, and cluster scratch space are typically very different, Online Storage Modules can be configured for different needs.
- *Disk Archive Modules* — Disk Archive Modules provide high-density, low-cost online capacity for storing large volumes of data that require regular access but are not under heavy usage by the cluster. Either via policy or by explicit command, the Core Data Management Modules move data from the Online Storage Modules to the Disk Archive Modules — freeing valuable cluster scratch space. Data management policies can also be set automatically based on file attributes, and data can be managed according to the storage and access requirements of each user on the system.
- *Tape Archive Modules* — Tape Archive Modules provide the same function as Disk Archive Modules, with the advantage of lowest possible cost and power consumption for stored data. Like the Disk Archive Modules, Tape Archive Modules are contained within the same name space as the Online Storage Modules, so data access is seamless, regardless of media type. With Sun StorageTek Storage Archive Manager software, files that are moved to tape archive are still visible and accessible to users, without requiring a backup operator to initiate a restore option from tape.

Chapter 5

HPC Software, Resource Management, and System Management

Ultimately, compute clusters and other HPC configurations are more than the systems, networks, and storage solutions that comprise them. Effective HPC infrastructure must also provide a seamless software environment that supports essential applications even as it provides for the efficient utilization of resources and management of the cluster. In the past, many of these components were left for organizations to collect, build, and customize on their own. Sun's HPC Reference Architecture provides and integrates these components, ranging from tested and optimized high-performance computing software, to resource and system management software.

Suns HPC Software for Linux and the Solaris OS

The software stack for any given HPC system depends on the applications themselves and their supported platforms. There are essentially three types of programs:

- Traditional single threaded programs occupy a single address space and utilize a single execution thread.
- Multithreaded programs occupy a single address space and provide multiple threads of execution. These applications need to be executed on the same machine, preferably in parallel on multiple processors.
- Multiprocess programs are comprised of multiple processes (single threaded or multithreaded) that communicate, with each addressing a subset of a given problem. Multiprocess applications may run on several machines throughout the cluster.

All of these applications require integrated and optimized HPC software in order to operate to their greatest advantage. In addition, organizations need to be able to deploy their clusters and HPC infrastructure quickly without excessive hand tuning and optimization. To this end, Sun provides Sun HPC software in OS-specific editions that bring pre-built and optimized versions of popular open source HPC software components together in a convenient and easy-to-deploy package.

- *Sun HPC Software, Linux Edition* is an integrated, open-source software solution for Sun HPC clusters. This software simplifies the deployment of HPC clusters by providing a ready-made framework of software components that lets organizations turn a bare-metal system into a running HPC cluster. Software is provided to provision, manage, and operate large-scale Linux HPC clusters — serving as a foundation for optional add-ons such as schedulers and other components not included with the solution.
- *Sun HPC Software, Solaris Developer's Edition* (in BETA release as of this writing) combines virtualization technology with a high performance computing platform. The software includes tools and technologies that lets organizations develop, test,

and deploy high performance computing applications. The software suite represents an integrated HPC product engineering environment that is certified on the robust Solaris 10 OS. The Developer's Edition release includes pre-configured Solaris 10 based virtual machines that include the HPC engineering environment for VMware and Sun xVM VirtualBox virtualization platforms. More than just a development environment, the software installs a mini three-node single-system cluster using virtual machines for testing applications before full cluster deployment.

No matter what the OS platform, Sun's HPC Software includes support for multiple operating systems, and provides HPC software components that are tested to work together.

- ***Parallelization subroutines***

  The parallelization subroutine layer has traditionally been provided by a software package called Parallel Virtual Machine (PVM) but today most parallel programs support the Message Passing Interface (MPI) and/or OpenMP. MPI is a standard, but there are many implementations, both open and proprietary. Some examples include MPICH, MVAPICH, Sun HPC ClusterTools™ (OpenMPI), IntelMPI and ScaliMPI.

- ***Networking protocols***

  Support for a wealth of network protocols is equally important. As InfiniBand evolves, the OpenFabric Alliance (OFA)[1], is developing the Open Fabrics Enterprise Distribution (OFED), a software stack that will support multiple protocols over InfiniBand. OFED includes OpenMP/OpenMPI as well as storage protocols such as the (SCSI RDMA Protocol (SRP), iSCSI over RDMA (iSER), Fibre Channel over IB (FCoIB), and network protocols such as IP over IB (IPoIB), Sockets Direct Protocol (SDP), and Ethernet over IB (EoIB).

- ***Optimized compilers***

  If the application is available in source form, choosing the right compiler and the optimal set of compiler flags can have a decisive impact on performance. By default, most compilers are optimized for the shortest possible compile time to optimize development work. Merely setting compiler optimization flags may speed execution up to two to five times at the cost of a longer compile time. Further gains can be made by looking at specific optimization flags.

  Sun provides performance tools that let the user look at the performance characteristics of the code and identify the subroutines and even specific lines in the subroutines that take the most time. These code sections can then be optimized (either rewritten or optimally pipelined in the processor). Sun Studio compilers support this approach, and it is available for both the Solaris OS and for Linux. Other compilers often used include gcc, Intel compilers, Portland Group, and others.

1. http://www.openfabrics.org

- ***Resource management***

  The resource management layer resides above the applications in the software stack. Resource management virtualizes the available computing resources and makes sure that jobs submitted by users are executed in an optimal fashion. Users should not have to keep track of the names and capabilities of computing resources, and whether they are available at the moment or not. Advanced resource management software such as Sun Grid Engine software can help ensure that critical projects get priority access to the resources that they need, when they need them. Sun Grid Engine software also helps make sure that shared resources are made available to different projects or departments according to predefined rules, while providing essential data for chargeback of consumed resources.

  Compute resources can be made available to users in many ways. In academic environments the users often submit their jobs manually, either through a submit command from a command shell or from a graphical user interface. In commercial environments it is more common that the computing resource is hidden from the users so that they are not even aware that the job is executed on a different system. This transparency requires some integration work by the system administrator or the vendor. It is also possible to use a Web portal to access the computing resource and both submit and check the status of jobs.

Resource management with Sun Grid Engine software

The resource manager is a key component of the HPC solution, as it handles the inventory of available compute resources, manages the queue(s) of submitted jobs, monitors execution (and in case of a failure, resubmits the job), and provides advanced reservation as well as accounting. The open source based Sun Grid Engine software is a very popular resource manager, but there are also other alternatives such as OpenPBS and Platform LSF.

Sun Grid Engine software provides policy-based workload management and dynamic provisioning of application workloads. A Sun Grid Engine master can manage a grid of up to ten thousand hosts, meeting the scalability needs of even the largest clusters. Sun Grid Engine software (or open source Grid Engine) runs on virtually all existing hardware from Sun and other vendors and supports all modern operating systems (including Microsoft Windows Server and MacOS X in addition to all major UNIX and Linux distributions). Sun Grid Engine software also contains accounting and reporting functions to provide statistics on usage and loads.

System management

Administering a small cluster with only several nodes is fairly simple, but when the number of compute nodes grows into the hundreds or even thousands, sophisticated tools are needed for hardware monitoring, upgrades, and OS installation. Sun

Microsystems has been active in developing system management tools for a long time. For traditional datacenter servers, Sun Management Center has provided a scriptable management interface.

Open source alternatives

Even those deploying commercial HPC solutions sometimes select non-commercial application stacks for system management. There are a large number of alternatives today providing easy management and software provisioning. Each of these solutions has different strengths and weaknesses, but all share the fact that they come from an open source community. The following list is by no means exhaustive and is provided as an example only:

- ROCKS[1] — ROCKS is mainly intended for Linux but Solaris OS support is now available.
- Ganglia[2] — Ganglia is mainly used for monitoring but it is deployed by a large number of companies.
- OSCAR[3] — OSCAR is generally used for clusters with MPI implementations.
- KickStart[4] — KickStart provides basic installation of Linux using server-based installation protocols and operating system images.

Many large institutions also have in-house developed solutions for different aspects of the required system management.

xVM Ops Center

For large compute clusters and systems without integrated management systems, other tools are necessary. Sun xVM Ops Center performs a number of critical tasks in a compute cluster, including:

- System discovery
- Hardware monitoring
- Operating system provisioning
- Comprehensive updates and patch management
- Firmware updates
- Hardware management from power up to production

Sun xVM Ops Center also provides support for cross-platform Linux and Solaris OS based x86/x64 and SPARC environments. As a result, Sun xVM Ops Center allows users to streamline their datacenter operations, minimizing costs as they more easily manage rapid growth, perform datacenter consolidation, and meet with stringent compliance requirements. xVM Ops Center is Web based for simplified usage but it also supports full scriptability for automated administration. In addition to systems from Sun, other

1. http://www.rocksclusters.org/
2. http://www.ganglia.info
3. http://svn.oscar.openclustergroup.org/trac/oscar
4. http://en.wikipedia.org/wiki/Kickstart_%28Linux%29

hardware platforms that support the Intelligent Platform Management Interface (IPMI) standard IPMI[1] can be managed through xVM Ops Center. IPMI helps facilitate “Lights Out Management and out-of-band management techniques.

1.http://www.intel.com/design/servers/ipmi/ipmi.htm

Chapter 6
Summary

Almost every organization can improve the efficiency of many internal processes by the effective use of HPC infrastructure. However, regardless of size, the design and acquisition of HPC infrastructure can benefit from a systemic and integrated design approach. Sun's HPC Reference Architecture is based on proven deployments of a range of HPC clusters — from single racks to some of the world's largest supercomputing clusters. The reference architecture provides systems, networks and interconnects, storage, software, and management components that are designed to work together.

The result of this approach is rapid deployment time and early results for even the largest clusters. At the same time, Sun's HPC Reference Architecture can provide proven components and building blocks that can directly benefit small to medium clusters. Only with a comprehensive end-to-end approach can organizations select the systems and components they need to accelerate their most important computing activities. Sun not only lends its IT knowledge but also deep insight into datacenter logistics involving power and cooling, software and hardware support, change management methodologies, and maintenance procedures. Sun's reference architecture even extends to software development — critical to creating the applications that can actually scale to the large multicore, multiplatform, data-intensive solutions required by business processes or research today. Only a few companies today can deliver a proven and tested high-performance solution today, and Sun Microsystems has the track record to be such a partner.

About the authors

The authors are listed in alphabetical order below.

- *Torben Kling-Petersen, Ph.D. — Senior Technical Specialist, HPC, Lustre Group*
  Torben Kling-Petersen has worked with high performance computing in one form or another since 1994 and is currently working as a Senior Technical Specialist for HPC in Sun's Global Systems Practice. Over the years, he has worked in a number of capacities such as lead architect for enterprise datacenter infrastructure, technical research lead and product specialist for high-end visualization to mention a few. In his present capacity, Torben works in a global role providing technical evangelism and solution architectures on petaflop-scale HPC projects.

- *Börje Lindh — Senior Systems Engineer*
  Borje Lindh has an M.Sc. in Chemical Engineering but has been working with Computers since 1987. He has been with Sun Microsystems since 1994, working in a number of different roles and has been involved with High Performance computing since 1997. He has written two Sun Blueprints on compute clusters and a some magazine articles mainly on processor architecture.
- *Ola Tørudbakken — Distinguished Engineer, Scalable Systems Group*
  Ola Tørudbakken has an M.Sc. degree from the University of Oslo, Department of Informatics in 1994, and has since then been working on High Performance Interconnects and Server Systems. He has been with Sun Microsystems since 2000, and his current responsibilities includes supervision and architectural definition of Fabric Products. He has published several papers in the field of interconnection networks and has participated in several standardization activities. He currently holds three US Patents, and has more than 30 US patents pending.

Acknowledgements

The authors would like to recognize Eric Liefeld, an independent technical specialist and writer for his assistance with this Sun BluePrints article. Eric is a former Sun Systems Engineer and a frequent contributor to Sun Microsystems technical documents.

References References to Sun BluePrints articles.

Wood, Chris and Flaskerud, Craig. “Sun Storage and Archive Solution for HPC, ” *Sun BluePrints Online*, May 2007”. To access this article online, go to `http://www.sun.com/servers/hpc/docs/Sun_HPC_Ref_Arch.pdf`

Top500 List of Supercomputing sites: `http://www.top500.org`

Open Multiprocessing: `http://www.openmp.org`

Open MPI: `http://www.open-mpi.org`

Texas Advanced Computing Center: `http://www.tacc.utexas.edu`

Lustre parallel file system: `http://www.lustre.org`

Sun Constellation System: `http://www.sun.com/sunconstellationsystem`

Open Graphics Language (OpenGL): `http://www.opengl.org`

ROCKS information: `http://www.rocksclusters.org`

Ganglia information: `http://www.ganglia.org`

OSCAR information:
`http://svn.oscar.openclustergroup.org/trac/oscar`

Ordering Sun documents

The SunDocs℠ program provides more than 250 manuals from Sun Microsystems, Inc. If you live in the United States, Canada, Europe, or Japan, you can purchase documentation sets or individual manuals through this program.

Accessing Sun documentation online

The `docs.sun.com` web site enables you to access Sun technical documentation online. You can browse the `docs.sun.com` archive or search for a specific book title or subject. The URL is
`http://docs.sun.com/`

To reference Sun BluePrints Online articles, visit the Sun BluePrints Online Web site at:
`http://www.sun.com/blueprints/online.html`

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 06/09