

ACCELERATING PERFORMANCE AND REDUCING COSTS FOR DATA-INTENSIVE HPC WORKFLOWS

Sun Storage Solutions
White Paper
April 2009

Abstract

This paper provides an overview of the current trends and challenges in delivering the necessary I/O throughput for high-performance computing (HPC) applications. It presents Sun's offerings for addressing today's HPC challenges in the form of flash technology, network attached storage, parallel storage, and archive solutions. Readers will also learn about why open storage solutions are important to HPC and how Sun is utilizing open storage technologies to benefit HPC customers. Sun provides a wide range of cost-effective options for HPC storage performance, helping customers improve throughput for HPC applications while reducing costs.

Table of Contents

Executive Summary.....	3
Data-intensive HPC Applications are Driving New Requirements.....	3
Better Value through Open Storage.....	3
Eco-efficiency	4
Sun HPC Storage Solutions	4
Challenges and Requirements for Data-intensive HPC Applications.....	6
Requirements for HPC Storage	8
The Role of Open Storage in HPC.....	10
What is Open Storage?.....	10
Key Advantages of Open Storage.....	10
Sun Open Storage	11
Leveraging Open Storage to Benefit HPC Customers.....	11
Further Reading on Open Storage	12
Sun HPC Storage Solutions	13
Flash Technology.....	14
Taking advantage of flash technology with the ZFS file system.....	14
Network Storage.....	15
Sun Storage 7000 Unified Storage systems.....	15
Real-time visibility and simplified management	16
Parallel Storage	17
Lustre™ file system	17
Sun™ Lustre™ Storage System	18
Components of the Sun Lustre Storage System	19
Sun Fire™ X4540 Storage Server	19
Sun™ Storage J4400 array	19
Archive Storage.....	20
Sun StorageTek™ Storage Archive Manager	20
Sun Storage and Archive Solution for HPC.....	21
Conclusion	22
Try and Buy Offer	22
For More Information	22
Recommended white papers and Sun BluePrints.....	23

Chapter 1

Executive Summary

Data-intensive HPC Applications are Driving New Requirements

High-performance computing (HPC) has reached a turning point where server compute power is no longer the singular concern. Indeed, users of complex numerical applications from DNA research to crash testing and financial simulations are now concerned about storage system performance and data management as constraints to application throughput and user productivity.

With today's fast multicore servers and high-capacity storage systems, HPC and technical computing customers are generating greater amounts of data through sophisticated models and analysis programs. Today's HPC environments run more frequent and more complex simulations that absorb and produce increasing amounts of data while sharing the data across global teams of researchers, engineers, or analysts.

The impact of these trends on system architectures is the requirement to move greater amounts of data throughout the HPC workflow. Today's HPC systems thus require both higher storage capacity and higher I/O throughput rates. Traditional storage solutions and special-purpose HPC storage appliances are becoming too costly for today's exploding data volumes. Organizations are finding it challenging to contain storage costs while delivering the high I/O bandwidth, low latency, shared access, and data protection required for today's HPC and technical computing solutions.

Another challenge facing HPC users is that storage requirements vary significantly in different stages of the HPC data lifecycle. Storing and managing user home directories or saving pre-processed HPC input data does not require the same level of I/O performance as is needed for the HPC compute clusters where the HPC applications execute. Furthermore, many sites have a need to preserve their data over a period of multiple years. In some cases, multiple terabytes of data are generated every month, making it costly to preserve data without automated archival solutions that make efficient use of lower cost storage media such as tape.

In large HPC sites, different storage solutions are often deployed for these different categories of usage to help optimize performance and cost. Both large and small HPC and technical computing customers are looking for ways to improve storage performance while reducing the cost of storing and managing HPC data.

Better Value through Open Storage

Just as open-source software, industry-standard servers, and horizontal Linux clustering have radically changed the HPC landscape, similar trends are changing today's storage solutions. Sun is a leader in the open storage revolution and has a growing portfolio of storage products that leverage open storage trends to offer greater customer

value and excellent price/performance. Unlike other modern disk arrays and NAS devices that are constructed from expensive proprietary designs and specialized software, Sun Open Storage solutions offer significant savings by taking advantage of high-volume, industry-standard components and open-source software.

Eco-efficiency

With exploding data volumes, space, power, and cooling costs have become much more important for today's HPC solutions. Sun can help HPC customers improve eco-efficiency by providing energy-efficient, space-saving archive solutions and by utilizing flash technology in Sun servers and storage solutions. Recognizing the importance of eco-efficiency, Sun has made reduced space, power, and cooling a key design goal for all of its offerings. Sun's offerings can help customers improve application performance without adding more disk drives or servers to the HPC environment, thus reducing datacenter resource requirements.

Sun HPC Storage Solutions

A broad spectrum of storage requirements must be satisfied for organizations to be successful in using HPC solutions to meet their goals. Requirements can vary by industry, installation size, and even by the different stages of the HPC workflow. Recognizing this and investing accordingly, Sun is in a unique position to meet all of today's HPC storage requirements. Sun offers a full range of high-performance hardware and software offerings that address the entire spectrum of HPC storage requirements.

Sun's offerings include:

- **Flash-based storage**—The latest Sun storage solutions, including Sun™ Storage 7000 Unified Storage Systems, take advantage of flash technology and the Solaris ZFS™ file system to deliver high performance at up to 75% lower cost than traditional storage solutions. Sun has also fully integrated flash technology into its new Sun Blade™, Sun x64, and Sun CoolThreads™ servers to help customers boost application performance without adding more servers or more high-speed disk drives to their HPC environment. Flash technology can be used with HPC applications to accelerate hot data files, alleviate the need to overbuy disks for performance, and reduce the need for installing large memory pools in compute nodes.
- **Network storage** — While traditional NFS solutions allow data to be centralized for easier distribution and sharing, they are not generally built to handle high performance, nor a high volume of I/O throughput. Sun Storage 7000 Unified Storage Systems provide extreme ease of use and high performance data sharing at dramatically lower prices than traditional NFS server and NAS technologies.
- **Parallel storage** — Sun™ Lustre™ Storage System addresses the I/O challenges of HPC environments by providing an incredibly powerful, scalable, and simple-to-deploy storage solution based on the Lustre™ file system. Sun servers and Sun Open

Storage products such as the Sun Fire™ X4540 Storage Server and cost-effective Sun™ Storage J4000 Arrays can be assembled into a cluster configuration in the Sun Lustre Storage System, providing a range of options for different customer needs.

- **Archive storage** — To address the escalating cost of storing large volumes of HPC data, Sun has a complete line of tape and archive products and software solutions for archival management. Sun StorageTek™ tape drives and libraries provide power efficient and cost-effective media for storing large volumes of data. Built around Sun StorageTek™ Storage Archive Manager software, the Sun Storage and Archive Solution for HPC provides automated policy-based archiving and on-demand, transparent file retrieval. The cost of building and managing large data repositories can thus be reduced through low power consumption, low cost media, and reduced effort for data management and operations.

Sun storage solutions can help HPC and technical computing customers achieve higher performance and reduced risk at dramatically lower prices.

Chapter 2

Challenges and Requirements for Data-intensive HPC Applications

Traditionally, the emphasis with high-performance computing (HPC) applications has been on CPU performance. HPC systems were measured in terms of floating point operations per second and the world’s largest systems have been tracked by similar performance metrics for more than 15 years on the Top 500 list (top500.org).

Today, however, the emphasis is shifting. Figure 1 shows the results of an IDC survey of medium to high-end HPC User Forum members. Users were asked if they had applications that were constrained by various I/O factors. Not only were more than 90% of respondents concerned about I/O in general, but they also expected further I/O constraints in the next three years¹.

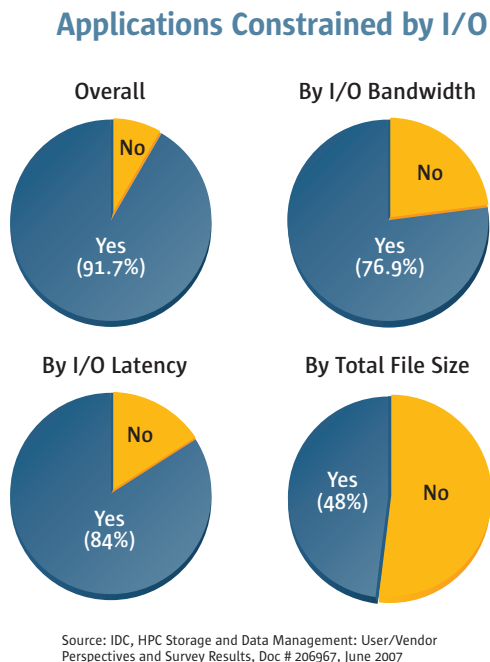


Figure 1. More than 90% of HPC users surveyed had applications constrained by I/O.

According to IDC, demand for HPC storage and data management is being driven primarily by the rapid increase in data that has been made possible by enormous performance gains on the computational side. Server processing power has been doubling every 18 months for many years, enabling HPC applications to create, analyze, and utilize increasingly large volumes of data.

1. “HPC Storage and Data Management: User/Vendor Perspectives and Survey Results,” IDC, Doc # 206967, June 2007.

As processing power and storage capacity have become more affordable, more thorough analysis is possible and practical. Applications can thus generate larger and more frequent results. In addition to storing and managing more data, there is a need for sharing these large data files across growing numbers of users often spread across global teams that can span multiple organizations.

These trends have created an even stronger need for faster storage I/O. Yet this is precisely the area where technology advances have been slower to manifest. In the past 15 years, disk drive capacities have increased dramatically, growing from 2GB to 1TB drives. However, disk drive rotational speeds have merely tripled in this time. So, traditional storage architectures are having difficulty matching the I/O capacity that today's applications need.

Storage performance has thus become a common bottleneck for HPC and technical computing applications and is driving up the cost of HPC solutions. Customers are facing the following key challenges with their HPC storage environments:

I/O performance bottlenecks

Storage I/O performance bottlenecks used to be something that only large HPC research institutions would encounter. Today, even small clusters often suffer CPU idle time while waiting for storage systems to complete I/O. Keeping high volumes of I/O flowing to and from the CPU is key to fast computations and requires low latency access to storage. However, fast access for single reads and writes is not enough by itself. With multiple processors working in parallel, overall throughput of data movement is also a key performance factor. Even more challenging is the fact that performance and cost are intimately connected. What customers are really looking for are solutions that provide the most I/O throughput per dollar.

Explosive growth of data

New application types and higher fidelity analysis are driving creation of more data per site, making scalable I/O processing an important requirement for today's HPC solutions. In addition to processing greater volumes of data, today's HPC systems often must maintain data for longer periods of time and the data must be shared throughout a global community of users. With explosive growth of data, it has become impractical for many HPC sites to store all of their data on disk drives where users can immediately access it. Management complexity also increases dramatically as the volume of data and number of users continue to grow, making storage management tools a key requirement for today's HPC customers.

Unsustainable costs

Data-intensive applications have impelled many HPC sites to deploy high-cost proprietary storage solutions that are designed specifically for HPC environments. A new more cost-effective approach to storage is needed where general-purpose systems and open-source software can be used to deliver the necessary performance while keeping storage costs within budget. The need to manage growing volumes of

data over time is also driving customers toward multi-tier storage solutions. As data becomes less active, it should be archived to a lower cost media such as tape where it consumes less power and cooling resources. To manage the archival process efficiently requires an automated approach and tools that simplify retrieval of archived data. Data protection and recovery are also key issues, making archival functionality an important capability for today's HPC environments.

Pressure for fast deployment

It's becoming increasingly important for HPC and technical computing projects to get up and running quickly in today's business environment where immediate results are critical to success. Customers need solutions that can be deployed quickly and easily so they can begin running applications and storing data without a lengthy setup and integration project.

The trends of faster and faster processors and increasing needs for storage capacity are expected to continue. The need to address storage performance will not go away. Customers must find new and more cost-effective ways to get additional performance out of their HPC storage environments while simplifying deployment and management of the HPC storage infrastructure.

Requirements for HPC Storage

The HPC workflow creates some unique requirements for storage systems not only because I/O requirements are demanding, but also because there can be multiple stages with differing requirements. The HPC data lifecycle typically includes three primary stages as listed below. For large environments, each of these stages is typically served by different storage and a separate file system. For smaller sites, a single storage architecture may serve all of these needs. The storage requirements for these stages include:

- **Staging/User Home Directory** — This directory area provides a private workspace for HPC users such as researchers and engineers. It includes the home directories for HPC users and stores working data such as test data sets and sometimes post-processing data. It also generally holds the HPC application executable files and staged input files or pre-processed data that will be used by HPC applications during batch execution.
- **Scratch/Computation** — High performance (high bandwidth and low latency) is required for temporary access to data being used by HPC applications during execution. This includes reading and writing input files and output files (e.g. analysis results files) as well as checkpoint/restart log files if they are generated during longer batch runs of HPC applications.
- **Archive** — This area is for long-term retention of HPC input and output data. It can also provide network access to these data files so they can be shared across a wide area network by global teams of users without necessarily moving data back to the staging area for shared access.

The data management, data access, and performance characteristics of the storage system(s) supporting these three categories of usage are dramatically different. The requirements are summarized in Table 1.

Table 1. Storage requirements for different categories of HPC usage.

Usage Category	Storage System Requirements
Staging/User Home Directory	<ul style="list-style-type: none"> • Support for industry-standard protocols and transports such as NFS and TCP/IP • Adequate performance for multiple or many HPC users • Cost-effective storage that is easy to backup and manage • Security controlled access based on user roles and the ability for authorized users to copy files to and from other storage environments
Scratch/Computation	<ul style="list-style-type: none"> • High performance (IOPS, bandwidth, response time) to keep HPC applications from waiting for data • Fine grained control over physical data placement on the media (striping, chunking, striding, etc.) in order to optimize utilization of the available media • Able to scale across all three key performance metrics: capacity, bandwidth, and IOPS • Support for industry-standard transports and protocols to allow for cost-effective connection by any grid client • Not necessarily backed up since storage is typically temporary
Archive	<ul style="list-style-type: none"> • Protect intellectual property against data loss (often for multiple years) • Scale to multiple petabytes if necessary • Extreme cost-effectiveness in terms of cost per GB as well as eco-efficiency (low power and cooling costs and minimal space requirements) • Dynamically migrate data to the most appropriate media type as data becomes older and/or less active • Support for industry-standard transports and protocols • Direct access by users with no human intervention required to load tapes, etc. • Reasonable performance (100's of MB/sec. bandwidth) for both ingest into the archive as well as read access by users • Store data in open, stable formats so that it can be read and understood by any application now and into the future

Sun is in a unique position to meet these requirements with a full range of hardware and software storage offerings addressing the entire data lifecycle within an HPC workflow.

Chapter 3

The Role of Open Storage in HPC

What is Open Storage?

Today's HPC systems often utilize open-source operating systems and open-source HPC software applications and tools to enable rapid innovation and help reduce costs. In recent years, the shift from custom-built, massive parallel processing (MPP) systems to horizontal clusters of low-cost servers, has dramatically reduced the cost structure of HPC deployments.

Now the same shift to open systems is taking place in HPC storage technology, with similar benefits and advantages for today's HPC applications. Traditional proprietary storage arrays simply cannot compete with the cost-effectiveness and flexibility offered by open storage solutions.

When purchasing traditional storage arrays or proprietary storage solutions, customers are generally forced to purchase the storage controller and all related software from the same storage vendor. They often pay for individual software features, usually with capacity-based software licenses. In some cases commodity disk drives used within proprietary storage solutions are marked up as much as five times the original cost. In the past, proprietary HPC storage solutions have had a significant performance advantage, so customers were willing to pay extra. Today's open storage solutions offer similar performance at a fraction of the cost of special-purpose HPC storage arrays.

Key Advantages of Open Storage

Open storage solutions enable a new approach to HPC storage — one that can better meet the tremendous demands of HPC applications while also reducing costs. Some of the key advantages of using open storage solutions for HPC include:

Better management tools

Because open storage solutions are based on general purpose computing platforms running a mature server operating system, they can leverage sophisticated management tools for optimizing performance and dramatically reduce the complexity of large pools of HPC storage. Administrative tools that have matured along with server technology over the last two decades, are suddenly available to be used within the storage environment. Open storage systems can also run the same operating system that has been evolving in the server environment for many years, thus leveraging the maturity and advancements in server technology.

Lower cost platforms

With today's multicore processors and chip multithreading (CMT), open storage systems can be assembled with more processing power than traditional modular storage controllers, and at significantly lower cost. Customers can leverage an industry-standard server in place of an expensive, proprietary disk controller. This platform is then combined with commodity disk drives and open-source software to provide the operating system, management functionality, and data services.

Increased innovation

Software and application innovation is no longer limited by a single vendor's business goals or research and development budget. Instead a burgeoning community of open-source developers can continually add new features or functions at a pace that cannot be matched by a single vendor's development team.

No proprietary lock-in

Open-source storage software liberates customers from proprietary vendor lock-in and expensive storage software license costs. By avoiding proprietary solutions, customers avoid proprietary premiums and have more freedom to expand their storage solutions in cost-effective ways, including taking advantage of all of the benefits listed above.

Sun Open Storage

Sun Open Storage solutions leverage Sun's more than 25 years of innovation in server technology as well as the influence of a community of passionate open-source developers. The solutions are based on general-purpose computing servers and utilize open-source software such as the OpenSolaris™ operating system (OS), the Solaris ZFS file system, and the Lustre file system.

Leveraging Open Storage to Benefit HPC Customers

Sun is building open storage into its products, enabling a new approach that meets the needs of today's HPC storage environments and offers the following business benefits:

Massive scalability and performance

To avoid performance bottlenecks in HPC applications, storage solutions must be able to scale to support high I/O throughput while also offering a large capacity file system. Sun Open Storage solutions offer virtually unlimited file system capacity and can scale throughput well beyond that of traditional storage solutions — all at a fraction of the cost.

Better storage economics

Sun Open Storage solutions leverage open-source software and low-cost commodity disks and require no extra software licensing fees. Customers can thus avoid expensive, proprietary HPC components and eliminate the need for additional software license fees based on storage capacity. When used with the Solaris ZFS file system, Sun Open Storage solutions can leverage flash technology to deliver much greater price/performance than is possible with traditional solutions based on expensive high-speed disk drives or proprietary NAS systems.

Improved service levels

Storage diagnostics and tuning analysis helps administrators quickly address and resolve performance issues in production systems.

Faster deployment

Because Sun Open Storage solutions are packaged as appliances and based on general-purpose computing resources and an open architecture, they can be deployed quickly without a lengthy integration effort. It is also easy to repurpose open storage components as HPC deployments evolve over time.

Further Reading on Open Storage

Additional information about open storage trends and Sun Open Storage offerings is available in the white paper entitled, “Open Storage Adoption,” which can be found on the Web at sun.com/offers/details/OpenStorage_Adoption.html.

Chapter 4

Sun HPC Storage Solutions

Sun HPC storage solutions span the entire HPC data lifecycle and provide industry-leading performance and scalability at much lower cost than special-purpose HPC storage offerings. The primary categories of Sun's storage offerings for HPC are highlighted in Figure 2. Figure 2 shows a sample HPC scenario that includes multiple HPC user environments connected via network fabrics to three storage file systems that address each of the major stages of the HPC workflow as described in Chapter 2.

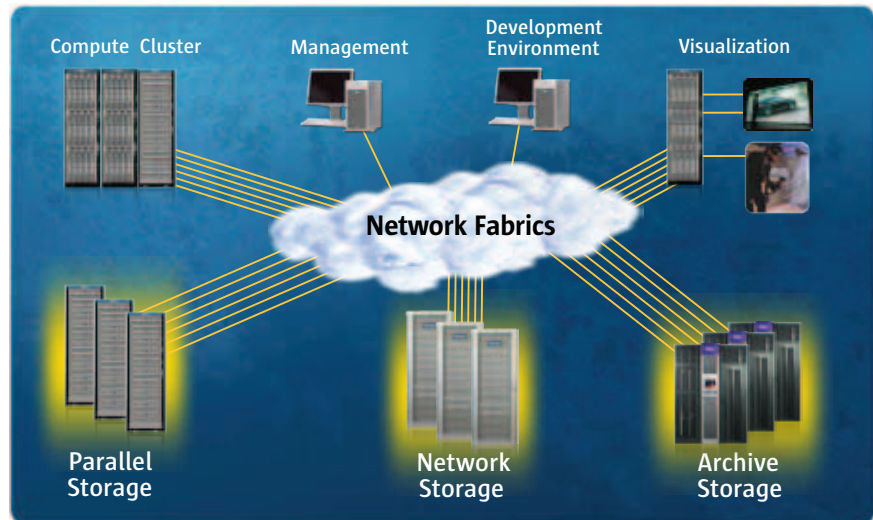


Figure 2. Sun HPC storage solutions address the entire HPC workflow.

Sun's storage portfolio is also augmented by recent investments in Sun products that take advantage of flash technology to boost I/O performance and reduce the cost of storage. Sun has integrated flash technology into many of its server and storage solutions and is developing a complete portfolio of flash technology offerings.

Sun HPC storage offerings can be divided into four categories as illustrated in Figure 3 and further described in the sections that follow.

<p>Network Storage Breakthrough economics through open source, standard components, and integrated flash technology</p>	<p>Flash Technology Leveraging the newest media for faster results</p>
<p>Parallel Storage Lustre file system coupled with open storage for high performance and compelling value</p>	<p>Archive Storage World-leading StorageTek tape libraries and multi-tier archival solutions</p>

Figure 3. Four primary categories of Sun HPC storage offerings.

Flash Technology

Recent advances in the production of flash technology have made solid-state drives (SSDs) and flash array products much more cost-effective, enabling a new approach to storage architectures. While SSDs are more expensive than mechanical drives, they are non-volatile and significantly cheaper than DRAM. They also offer much higher performance and greater power efficiency than hard disk drives, making them a strong alternative that provides a means to rebalance system and storage I/O performance.

Reliability characteristics of flash and SSDs have also improved with MTBF exceeding that of hard disk drives. Like hard disk drives (HDDs), enterprise SSDs also support bad block management, wear leveling, and error correction codes (ECC) to foster the highest level of data integrity and reduce service downtime. The solid-state nature of flash also allows enterprise SSDs to withstand significantly higher shock and vibrations than HDDs. They also require less power and cooling and can operate in a wider range of environmental conditions.

Sun Storage 7000 Unified Storage Systems take advantage of flash technology to deliver high performance and up to 75% lower cost than traditional storage solutions. Sun has also fully integrated solid-state drives (SSDs) based on flash technology into its new Sun Blade, Sun x64, and Sun CoolThreads servers, enabling up to 65x better response time than traditional hard disk drive solutions.

Taking advantage of flash technology with the ZFS file system

Taking advantage of the performance and cost characteristics of flash technology requires a file system that recognizes different types of storage media and can transparently optimize data placement to drive better application and file system performance.

The ZFS file system can take advantage of SSDs and flash arrays today by caching data on the high-performance media, allowing customers to achieve immediate performance gains. Unlike less sophisticated file systems, ZFS recognizes different media types and will optimize how it handles each type to maximize system throughput. For example, ZFS can take advantage of the performance characteristics of high-speed SAS drives when they are present. Now ZFS can also leverage SSDs, where available, enabling even more significant performance gains.

As shown in Figure 4, ZFS manages different types of storage media using Hybrid Storage Pools to combine high-speed SSDs with high capacity HDDs, boosting the performance of the overall storage solution while also reducing costs.

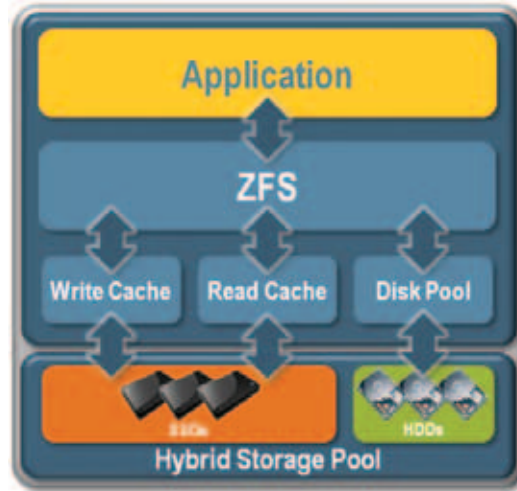


Figure 4. Hybrid Storage Pools isolate applications from the storage media, enabling ZFS to optimize performance across multiple media types.

Using ZFS and Hybrid Storage Pools, HPC applications can be completely isolated from storage media while taking advantage of the performance characteristics of different media types. This technology has the potential to unlock new levels of performance for HPC applications and deliver significant cost savings without the need to manually migrate data to or from different storage media. More information about Hybrid Storage Pools and ZFS can be found in the white paper titled, “Deploying Hybrid Storage Pools With Flash Technology and the Solaris ZFS File System.” This paper can be found on the Web at wikis.sun.com/display/BluePrints/Deploying+Hybrid+Storage+Pools+With+Flash+Technology+and+the+Solaris+ZFS+File+System.

Network Storage

For organizations using network-attached storage (NAS) to support their computational environments, the Sun™ Storage 7000 systems can serve as a NAS appliance that radically simplifies storage provisioning and data management while changing the economics of storage by providing performance that approaches that of storage area network (SAN) technologies, but at dramatically lower prices.

Based on the OpenSolaris OS, these systems are an ideal fit for a range of uses in HPC environments, including data staging, user home directories, and the cluster working space where single file system throughput needs can be up to 2GB per second (1GB/sec sustained).

Sun Storage 7000 Unified Storage systems

Sun Storage 7000 systems (Figure 5) utilize a high-performance Hybrid Storage Pool architecture with solid-state drives (SSDs) to dramatically improve I/O performance and power efficiency. The high-end configuration, the Sun Storage 7410 system, offers up to 576TB² of raw capacity using a combination of solid-state drives (SSDs),

2. The initial release of the Sun Storage 7410 system offers 288 TB in usable capacity and a free software upgrade will be available in the near future to enable the system’s full 576 TB of capacity to be utilized.



Figure 5. Sun Storage 7000 family.

DRAM, and HDDs. SSDs are used to provide an optional read and/or write cache, enabling I/O throughput similar to special-purpose HPC appliances. Sun has analyzed and verified that the Storage 7000 system is nearly twice as fast as similarly configured NFS systems for delivering data to MCAE applications. In recent MSC Nastran benchmark trials performed by Sun, the Sun Storage 7210 server with Hybrid Storage Pools was compared to the Sun Fire X4540 server, which represented a nominal NFS server. Both systems served a twenty node Sun Blade 6000 server, which was running the MSC Nastran benchmark. The Sun Storage 7210 server with flash technology was able to process 1300 jobs/hour, providing 90% more job throughput than the Sun Fire X4540 server, which processed just under 800 jobs/hour.

Sun Storage 7000 systems enable customers to quickly scale in multiple dimensions. To adapt to the changing needs of HPC users, customers can add storage capacity, more computational power, or additional read/write cache to improve performance. The Sun Storage 7000 series family comes in three different configurations plus a two-node cluster configuration providing high availability.

Real-time visibility and simplified management

Storage Analytics software is a free software component in Sun Storage 7000 systems, and provides the industry's only comprehensive and intuitive analytics environment. It gives administrators all of the tools they need to quickly identify and diagnose system performance issues, perform capacity planning, and debug live storage and networking problems.

Administrators can drill down for in-depth analysis of key storage subsystems using built-in instrumentation that provides real-time visibility throughout the data path. Real-time statistical graphs can be used to quickly locate and isolate problems down to the user and file level and can help administrators optimize storage performance and capacity utilization—all while systems continue running in production.

Sun's Unified Storage systems also provide unmatched simplicity and ease-of-use through an intuitive and powerful management interface that takes the guesswork out of system installation, configuration and tuning.

Additional information about real-time visibility and other software features available with Sun Storage 7000 systems can be found on the Web at sun.com/storage/disk_systems/unified_storage/features.jsp.

Parallel Storage

While network storage appliances such as Sun Storage 7000 systems are quick to deploy, simple to manage, and provide powerful tools that limit the need for specialized administrative expertise, traditional shared file systems such as NFS were not architected to scale to the performance levels required for bandwidth-intensive clusters. As a result, some customers require a parallel file system that is designed to serve the bandwidth needs of cluster environments. Sun provides a very high performance parallel storage offering with the Sun Lustre Storage System based on the open-source Lustre file system.

“Lustre gives us double the storage, at three times less cost of competing solutions. If our render farm gets bigger, we can quickly and easily add more nodes to the Lustre file system.”

Daire Byrne
Senior Systems Integrator
Framestore

Lustre™ file system

The power behind the Sun Lustre Storage System is the groundbreaking Lustre file system. This open-source shared file system is designed to address the I/O needs of compute clusters containing up to thousands of nodes. A number of HPC sites use the Lustre file system as a site-wide global file system, servicing clusters on an unprecedented scale. It is best known for powering the largest HPC clusters in the world, with tens of thousands of client systems, petabytes of storage, and hundreds of gigabytes per second of I/O throughput. The Lustre file system is used by 42% of the Top 100 Supercomputers as ranked by top500.org in the November 2008 listing. Respondents to IDC’s survey of medium to high-end HPC User Forum members identified Lustre as the most frequently installed storage management software³.

The Sun Lustre Storage System leverages technologies developed for these large-scale sites, and makes these technologies easier to deploy and use for a variety of mainstream cluster environments. Figure 6 illustrates the difference in how performance scales between the Lustre file system and NFS file systems.

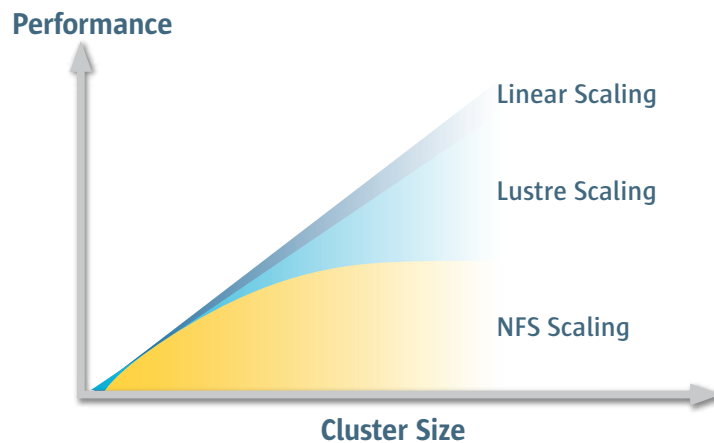


Figure 6. Lustre file system offers near linear performance scaling.

3. “HPC User Forum Survey, 2007 HPC Storage and Data Management: User/Vendor Perspectives and Survey Results.” IDC.

“We use a Sun Lustre Storage System to provide very high aggregate bandwidth for applications that run on Ranger. Applications have measured sustained throughput of 35GB/sec to our largest file system, increasing the productivity of researchers that generate and analyze large amounts of data.”

Tommy Minyard, Ph.D.
Associate Director
Advanced Computing Systems Group
Texas Advanced Computing Center

Sun™ Lustre™ Storage System

The Sun Lustre Storage System simplifies design, deployment, and scalability for the most demanding storage performance needs. Consistent with Sun’s open storage strategy, the Sun Lustre Storage System utilizes a combination of open-source software and Sun Open Storage platforms to deliver compelling capabilities and value.

Sun Lustre Storage System virtually eliminates I/O bandwidth and storage capacity constraints. Its I/O performance can scale from just two GB/sec to more than 100GB/sec of throughput while capacity can also be scaled to multiple petabytes.

A key to achieving these levels of performance is the horizontal scaling model enabled by the Lustre file system.

With traditional storage solutions, scaling is done “vertically” by adding more capacity to a system until it is filled. The maximum performance available is limited by the processing capability of the individual device. In contrast, the Lustre file system distributes files “horizontally” across a cluster of servers and storage. The performance and capacity of this environment can then be scaled in a near linear fashion by adding new storage devices to the cluster.

The near linear scalability of the Lustre file system enables the Sun Lustre Storage System to achieve performance far exceeding that of traditional NAS or SAN storage devices. As a rule of thumb, the Sun Lustre Storage System is ideal for clusters requiring two GB/sec or greater sustained aggregate bandwidth.

One of the largest computing systems in the world for open science research, the Texas Advanced Computing Center (TACC), manages a 1.2 petabyte file system and demonstrates near-linear scalability with the Lustre file system. As shown in Figure 7, TACC has observed throughput rates of 46 GB/sec on a single Lustre file system deployed across 50 Sun Fire X4500 servers, each of which is capable of more than 900 MB/sec of throughput. In addition, TACC has experienced a remarkable achievement with throughput on a single application’s use of the Lustre file system yielding 35 GB/sec for that single application. More details about the TACC configuration and its performance are available at tacc.utexas.edu/resources/hpcsystems/#ranger.

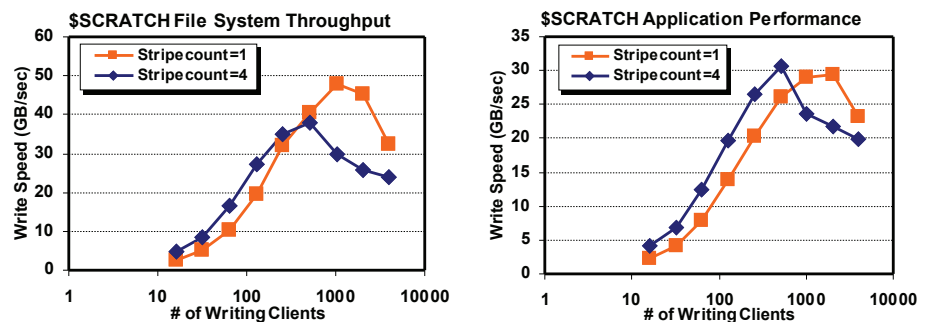


Figure 7. Proven scalability at Texas Advanced Computing Center.

Components of the Sun™ Lustre™ Storage System

The Sun Lustre Storage System consists of pre-defined modules that simplify system design and implementation, helping customers accelerate deployment and reduce risk. Sun has selected and tested the Sun Fire servers and the Sun Open Storage platforms that deliver the best performance at the most economical price point available for a Lustre deployment. These tested configurations offer reduced risk and deliver known performance.

The modules that comprise the Sun Lustre Storage System include:

- **High Availability Metadata Server (HA MDS) module** — The HA MDS module manages and stores metadata, such as filenames, directories, permissions, and file layout. The Lustre file system separates metadata from application data, enabling greater performance. The HA MDS Module consists of Sun Fire™ X4250 servers and Sun™ Storage J4200 arrays.
- **Object Storage Server Modules (High Availability OSS and Standard OSS)** — The OSS modules store application data and communicate directly to clients via LNET, a high performance and reliable data transfer protocol provided by the Lustre file system. Two or more OSS modules are usually deployed via the horizontal scaling model described earlier. Performance scales in a near linear fashion as OSS modules are added. Each Standard OSS module consists of a Sun Fire X4540 server and either InfiniBand or 10 Gigabit Ethernet connections. The HA-OSS modules consist of a pair of Sun Fire X4250 servers, four Sun Storage J4400 arrays, and utilize InfiniBand or 10 Gigabit Ethernet connections.
- **Lustre clients** — Lustre client code is used on each of the cluster nodes and enables the high-speed access to the file system via the performance-optimized LNET protocol. Client software is available for many different distributions of Linux.

Sun Fire™ X4540 Storage Server

The Sun Fire X4540 storage server is used as the building block for the standard OSS module in the Sun Lustre Storage System. Its integrated server and storage architecture enables a high density, high performance, and cost-effective platform for serving application data to a Lustre environment. The Sun Fire X4540 server is an open storage platform that incorporates industry-standard hardware components, including eight compute cores to run the OSS software and 48 1TB disk drives in a single, low cost enclosure. It is an ideal fit for HPC scratch space deployments. Compared to traditional servers and storage arrays, the innovative design of the Sun Fire X4540 server can save about half the cost and uses 50% less power.

Sun™ Storage J4400 array

Used as the storage component for the HA OSS Module in the Sun Lustre Storage System, Sun Storage J4400 arrays deliver breakthrough economics, excellent performance, and high availability. The HA OSS module combines Sun Storage J4400 arrays and Sun Fire X4250 Servers with open source software, including the Lustre software stack with the Linux operating system and RAID 6 data protection. Shifting the data

protection and data management functions to the compute cores on the Sun Fire X4250 server enables high performance and reliability with high volume industry-standard components. This results in a significant cost savings over specialized, custom-built, high-bandwidth storage solutions.

Archive Storage

By moving data that is not in current use for HPC projects to lower cost media such as tape, organizations are able to realize significant savings through lower cost equipment as well as reduced power and cooling costs. As the leading tape automation provider for the HPC market, Sun has a complete set of archive offerings that can be coupled with Sun's parallel storage offerings to address the complete lifecycle of HPC data.

Large HPC customers with multi-terabyte installations can take advantage of multi-tier archival solutions that combine disk and tape to simplify the archival process and provide additional protection. Smaller installations can select a cost-effective tape drive or tape library to simplify backup of their data on a regular basis. Sun offers a complete portfolio of tape drives and tape libraries to support right-sized solutions for protecting data.

Energy efficiency can be a big factor in reducing costs for archived data. Sun StorageTek Tape Libraries are up to 57 percent more energy efficient than competitive offerings. In fact, Sun has made energy efficiency a key design element and Sun's recent tape library offerings have reduced energy consumption by as much as 40 percent over previous generation libraries. In addition, the amount of data that can be stored on a tape cartridge has, in some cases, quadrupled in the past five years.

Sun StorageTek™ Storage Archive Manager

Sun StorageTek Storage Archive Manager (SAM) enables HPC sites to easily and efficiently exploit the cost savings between traditional disk storage and tape. Based on their retention and retrieval requirements, customers can actively manage and migrate data between the storage media types. When data is needed for an HPC job, it can easily be moved into the Sun Lustre Storage System to support high performance cluster applications. When the project is complete, it can easily be moved back to the archives where it will still remain accessible for recall by users.

SAM can manage multiple petabytes of data, easily keeping up with data growth by continuously, automatically, and transparently making copies or migrating new or changed files to inexpensive SATA disk, tape, or other secondary media. There is no need to stop access to the file system or files while Sun StorageTek™ SAM-FS software copies files. Therefore, no backup window is required. The actual physical file location — disk, tape, local, or remote — is completely transparent to the application thus requiring no special programming conventions to utilize StorageTek SAM-FS software.

The transparent migration capability can also be used to migrate data from outdated equipment to newer hardware. Open file formats help prevent vendor lock-in for archived data.

Sun Storage and Archive Solution for HPC

The Sun Storage and Archive Solution for HPC, combines Sun StorageTek Storage Archive Manager and Sun StorageTek tape libraries into a powerful and extremely cost-effective storage and archive management system for HPC data.

To simplify deployment while maximizing flexibility, the Sun Storage and Archive Solution for HPC consists of six different storage modules that act as scalable building blocks to provide comprehensive data protection and cost-effective, long-term retention. When deployed together with Sun offerings for parallel storage or network storage, the Sun Storage and Archive Solution for HPC helps customers manage the entire HPC data lifecycle.

Additional information on the Sun Storage and Archive Solution for HPC is available at wikis.sun.com/display/BluePrints/Sun+Storage+and+Archive+Solution+for+HPC.

“Sun is the only company that has a complete HPC portfolio, from hardware to software.”

Sik Lee, Ph.D.
Team Leader
Supercomputing Applications Team
KISTI

Chapter 5 Conclusion

Sun is leading an open storage revolution by combining open-source software with industry-standard system components to help customers reduce storage costs and avoid closed, proprietary, and expensive storage solutions.

As a leader in virtually every category of HPC storage offering, Sun is uniquely positioned to help HPC customers address the entire HPC data lifecycle with powerful and cost-saving storage options. Customers can eliminate I/O performance bottlenecks with Sun storage solutions that are a fraction of the cost of traditional HPC storage offerings.

Key advantages of Sun HPC storage solutions include:

- Industry-leading capabilities across all storage categories from network storage to parallel and archive storage offerings
- Dramatic cost reductions for high-performance storage through open solutions
- Innovative use of flash technology to deliver big performance gains at a fraction of the cost of traditional disk-only architectures
- Eco-efficiency with solutions that require less space, power and cooling than most competitive products

Try and Buy Offer

To gain firsthand knowledge of the benefits of Sun Open Storage solutions, consider Sun’s risk-free, try and buy offer for the Sun Storage 7000 system. Further details and an online application are available at sun.com/tryandbuy.

For More Information

For additional information on how Sun HPC Storage solutions can help reduce cost and complexity while optimizing the performance of HPC applications, visit the Web sites in Table 2 or contact a local representative.

Table 2. Web links for additional information.

Web Site URL	Description
sun.com/hpc	Sun High Performance Computing home page
sun.com/scalablestorage	Sun Lustre Storage System
sun.com/storage/openstorage/resources.jsp	Further reading and white papers about Sun Open Storage solutions
sun.com/storage/openstorage/products.jsp	Sun Open Storage product offerings
sun.com/unifiedstorage	Sun Storage 7000 systems
sun.com/j4000	Sun Storage J4000 Arrays
sun.com/x4540	Sun X4540 Storage Server
sun.com/lustre	Lustre file system
sun.com/flash	Sun Flash Storage
sun.com/qfs	Sun StorageTek QFS Software
sun.com/sam	Sun StorageTek Storage Archive Manager
opensolaris.org/os/community/storage	Storage Community within the OpenSolaris Community
opensolaris.com/200811/openstorage/storagemajor	Major OpenSolaris storage technologies

Recommended white papers and Sun BluePrints™

- *Revolutionizing the Storage Status Quo*
sun.com/offers/details/Revolutionizing_Storage_Status_Quo.html
- *Open Storage Adoption*
sun.com/offers/details/OpenStorage_Adoption.html
- *Deploying Hybrid Storage Pools With Sun Flash Technology and the Solaris ZFS File System*
Sun Blueprints™ Online Part No 820-5881-10
wikis.sun.com/display/BluePrints/Deploying+Hybrid+Storage+Pools+With+Flash+Technology+and+the+Solaris+ZFS+File+System
- *Rethink Storage with Sun Storage 7000 Systems*
sun.com/offers/details/Rethink_Storage.html
- *Sun Storage and Archive Solution for HPC white paper*
Sun Blueprints Online Part No 820-5304
wikis.sun.com/x/1Q7fAQ
- *Solving the HPC I/O Bottleneck: Sun™ Lustre™ Storage System*
sun.com/offers/details/820-7664.html

