White Paper

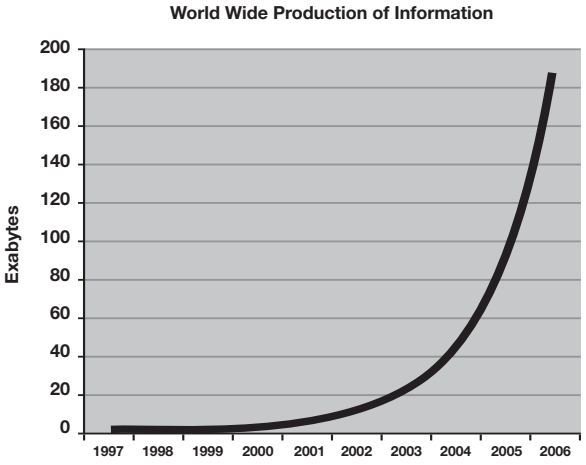
Storing, Accessing and Managing Multi-Petabyte Data Sets

Geoffrey Wehrman

Table of Contents

1.0 Introduction

The data explosion shows no sign of abating. Organizations engaged in high performance computing (HPC) across a wide variety of industries are doubling the amount of data they are storing every 6 to 18 months. Assuming a doubling time of 6 months, a company with 1TB of data today—a modest amount of storage by current standards—will need a petabyte of storage in 5 years.



Source: Gartner

Fig. 1. Description to go here. Description to go here. Description to go here.

If these numbers sound frightening, the reality can, in some circumstances, be even worse, with rates of data production increasing even faster during times of heightened innovation within many industries. For example, car manufacturers currently perform a limited number of computer simulations which mimic physical crash tests. Each test produces hundreds of gigabytes of data. Tomorrow's virtual crash tests—enabled by continuing improvements in HPC—will use stochastic methods to try hundreds or thousands of variations to identify more subtle design issues, resulting in hundreds of terabytes of data.

The data demand may be just as pressing on the consumer side. Today's 5Mpixel digital cameras deliver lower resolution than film. Increasing resolution beyond the level of 35mm film will push data sizes towards 100MB per photo, compared to the 2MB per photo that is typical today. Providing this resolution in high fidelity digital video means that filming that all important school play will require some 5.4GB/s, or 19TB for a single hour.

It should be obvious that today's mainstream storage systems don't scale to provide these levels of capacity and bandwidth. However, the largest HPC organizations are already learning to cope with data sets reaching into the petabytes and the resultant challenges which include:

- **Scaling Capacity.** Even with the cost of storage dropping at 40% per year and storage capacity growing at 80% to 130% per year, economically storing petabytes of information is no simple matter.
- **Increasing Bandwidth.** Current mainstream networking and storage networking technologies offer usable bandwidths in terms of hundreds of megabytes per second. (Gigabit Ethernet: 100MB/sec, 2Gbit/sec fibre channel: 200MB/sec; 4Gbit/sec fibre channel: 400MB/sec). This is adequate for moving multi-gigabyte files, but moving a terabyte of data over a single unloaded connection takes over half an hour. Even with 10 Gigabit Ethernet becoming available (1GB/sec), greater bandwidth is necessary to effectively access large data sets. The fact that disk capacities are growing faster than disk bandwidth poses another serious challenge. If this trend continues, a disk farm with adequate capacity may be far short of the necessary bandwidth.
- **Data Sharing.** As individual files and data sets get ever larger, maintaining multiple copies becomes ever more wasteful of storage, network bandwidth and user time, making efficient, high-speed data sharing a necessity.
- **Ensuring Data Protection.** With disk farms containing hundreds thousands of disks, instances of individual disk failure are a certainty, as is user error. Ensuring adequate protection for critical data is a necessity, but traditional methods are too slow and disruptive to ongoing work.
- **Developing Efficient Management Strategies.** Storing petabytes of data on disk is undesirable in terms of purchase cost, floor space, electricity consumption, and heat production but archiving data to tape or other media, managing a mixed media environment, and accessing archived data when necessary becomes more and more difficult as data set size increases.

Creating a practical, economical multi-petabyte storage infrastructure today requires significant effort. Storage technologies that integrate data lifecycle management and data protection while accommodating the heterogeneity characteristic of modern computing environments are becoming essential.

As a leading innovator in computing, visualization, and storage, SGI is helping today's HPC leaders tackle these challenges. HPC sites worldwide have communicated a desire for larger single filesystems, more files per directory and increased support for heterogeneous operating platforms. Immediate requirements from large sites include:

- Single filesystems capable of accommodating 1 petabyte of storage
- 100 million files per directory
- Cluster filesystems capable of supporting 64-nodes
- Filesystems with wide area support to eliminate the need for data copying

SGI is well on the way to meeting these needs. In the longer term, customers are requesting:

- Multi-petabyte single filesystems
- Billions of files per directory
- Unlimited clusters

This paper explores the storage technologies currently in available to meet these demands and the strengths and weaknesses of each approach. The best practices for managing today's largest and most demanding data sets are illustrated through selected customer examples from the SGI user base. Since today's cutting edge is often tomorrow's mainstream, these examples should provide insight into what mainstream storage infrastructures might look like in the future.

2.0 Technologies for Creating Multi-Petabyte Storage Infrastructures

SGI is already helping customers create storage systems capable of accommodating a petabyte or more of data. These technologies are not mutually exclusive. In fact most multi-petabyte sites depend on flexible storage infrastructures that use most or all of the technologies discussed in the later subsections. Section 3 will describe how various technologies have been mixed and matched in the real world to solve today's storage problems and prepare for the future.

2.1 Evaluation Criteria

To compare and contrast the various storage technologies in use today and understand where and how they are best applied, it is first necessary to establish criteria on which these technologies can be evaluated. As discussed in section one, for multi-petabyte data sets these criteria include:

- Performance
- Connectivity
- Capacity
- Management

Different applications may require different characteristics in each area so a user should evaluate his expected or desired data flow against these criteria to determine which solution or set of solutions will best match current and expected future needs.

Performance. When considering a storage technology, it is important to consider all factors which may limit performance including disk bandwidth, network bandwidth, internal system bandwidth and CPU performance. Beyond that, how efficiently will the solution scale? Will increasing capacity increase performance, or are there hard limits? Is scaling the solution relatively cheap or expensive?

Connectivity. When considering the ability to scale the connectivity of a solution, one must consider how many hosts a solution can support and whether or not heterogeneous system platforms can be accommodated. Is data coherency guaranteed to all sharers? What does it cost to scale connectivity? Finally, is connectivity limited to the local area or can it be extended to cover geographically remote sites?

Capacity. When scaling capacity it is necessary to consider whether the solution has any hard limits and whether performance will scale along with capacity. Will scaling up capacity offer adequate reliability and tolerate failure of individual components? Does the solution offer or integrate with nearline or offline storage? What does it cost to add capacity?

Management. When considering the manageability of a solution, it is important to consider the ability of the solution to consolidate data logically and physically, whether the solution supports data lifecycle management (DLM), and whether centralized backups are possible.

The following table summarizes the evaluation of the various technologies discussed in the following subsections based on the above criteria:

	Traditional Filesystems	Copy-based File Sharing	Network File Sharing	Data Lifecycle Management	Shared Filesystems
Performance					
Limiting Factors	Disk bandwidth, software design	Network bandwidth	Network bandwidth, file server bottlenecks	Bandwidth to secondary storage, software design	SAN Bandwidth, software design
Scaling Efficiency	High	Low	Medium	Medium	Medium
Cost to Scale	Low	High	Low	Medium	High
Connectivity					
Number of Hosts	1	Any	High	NA	<100
Heterogeneous Hosts	NA	Any	Any	Yes	Yes
Data Coherency?	Yes	No	Not guaranteed	Yes	Implementation dependent
Cost to Scale	NA	NA	Low	NA	Medium
Wide Area Connections	No	Yes	No	No	Possible
Capacity					
Hard Limits?	Implementation Dependent	No	No	No	Implementation Dependent
Effect on Performance?	Scales with capacity	No	May not scale with capacity Decreases	Scaling capacity may not increase perf.	May not scale with capacity
Effect on Reliability?	Decreases	NA	Possible	Minimal	Decreases
Integration with DLM?	Possible	NA	Medium	NA	Possible
Cost to Scale?	Low	NA		Lowest	Medium
Management					
Data Consolidation?	No	No	Yes	Yes	Yes
Centralized Backups?	No	No	Yes	Yes	Yes

Table 5. Characteristics of group integration in Non-VAN and VAN enabled organizations.

2.2 Traditional Filesystems

In the early days of computing as different types and brands of storage became available, it became impractical for each software application to manage underlying storage devices directly. Filesystems were therefore developed to provide a uniform interface as an abstraction to hide low-level storage details from application software. Today, traditional filesystems use direct-attached storage or SAN storage, although SAN resources must be allocated for the exclusive use of each filesystem.

Performance. Traditional filesystems yield great performance. Performance is typically limited only by available disk bandwidth or the bandwidth of connection between disk and host. Adding hardware generally results in linear performance scaling at low incremental cost. Tuning may be required to ensure storage layout, I/O request sizes, cache sizes, etc. are optimally configured. Inefficient algorithms used in some filesystem designs may limit performance. A filesystem must be multi-threaded to achieve the highest possible performance on multi-processor systems.

Connectivity. Only single host access is possible.

Capacity. Many filesystems have relatively low limits on the overall capacity that can be managed by a single filesystem. Likewise, some filesystems may limit individual file sizes and number of files stored which can be onerous given today's data growth. The cost of scaling capacity is relatively cheap since minimal storage hardware is required, but since resources must be allocated to a single system, scaling capacity for a traditional filesystem limits flexibility.

Management. Traditional filesystems do not lend themselves to data consolidation since each host must have its own storage. Even in a SAN, individual disks or LUNs must be more or less permanently allocated to individual hosts. Backups must be performed for each individual host, which can be difficult in environments with many systems.

The best-of-breed traditional filesystem is the SGI InfiniteStorage Filesystem XFS. XFS is a journaled, full 64-bit filesystem carefully designed to scale to meet the needs of the most demanding HPC environments by providing the utmost performance and capacity without imposing artificial limits.

Underlying XFS is the SGI InfiniteStorage Volume Manager XVM. XVM allows highly efficient performance scaling across extremely large numbers of disks or RAID LUNs. Still, XFS and other existing filesystems are not designed to continue operation when underlying storage components are not operational, so they are only as reliable as the least reliable piece of underlying hardware. This may impose a practical limit on the size of XFS files systems.

SGI released an open source version of the XFS filesystem in 2000 to extend the benefits of XFS to a broader user community.

2.3 Copy-based File Sharing

Because of the single-system connectivity limitation of traditional filesystems, it quickly became necessary to develop ways to share data between disparate computer systems. File transfer protocol (ftp) and similar programs developed as one of the first methods of sharing files between systems. The method is still in wide use, however, because networks—particularly wide area networks—typically lack the bandwidth to make other forms of file sharing feasible.

Performance. Available network bandwidth is typically the performance limiting factor. It often takes hours for large files to transfer.

Connectivity. Because ftp is almost ubiquitous, it is almost always possible to establish a connection between two systems for file transfer, so connectivity is never a problem, and the solution works over wide area networks although bandwidth may be an issue. No mechanism is provided to ensure coherency between copies.

Capacity. As individual files and data sets become ever larger, copy-based file sharing becomes increasingly undesirable since every copy wastes valuable storage space.

Management. Copying data between systems results in duplication, defeats data consolidation, and can also create problems with data integrity and security. Data management in an environment that relies heavily on copy-based file sharing can be a serious headache.

2.4 Network Filesystems

Because of the problems inherent in copy-based filesystems, the next step in the evolution of storage virtualization was the development of network filesystems that allowed data to be stored on a central file server and shared over networks to client systems. The most common network filesystem, NFS, was first released in 1984 and quickly became available for virtually every computing platform. More recently, the Common Internet File System (CIFS, also referred to as SMB or SAMBA after a popular implementation for Unix and Linux platforms) has become popular for sharing files with Microsoft Windows systems.

To simplify the deployment of network storage, file server vendors pioneered a new type of storage beginning in the 1990s: Network Attached Storage (NAS). NAS systems are generally single-purpose, dedicated file servers that provide a large amount of RAID storage and broad network connectivity in a simple and easy-to-deploy package.

While network file sharing has had dramatic benefits for computing in general, it has always been problematic for HPC problems that demand high I/O bandwidth. For this reason, many HPC users continue to use FTP to transfer files to local storage before processing. To date, no network filesystem in wide use has exhibited good performance over wide area connections.

Performance. The greatest limiter of network filesystem performance is typically the available network bandwidth. The TCP/IP protocol adds a significant amount of overhead to network transactions and limits performance. When network bandwidth is adequate, individual file servers or NAS devices often become a bottleneck, especially when accessed simultaneously by multiple network clients. To achieve the greatest possible performance it may be necessary for client and server side file sharing software to be multi-threaded so they can take full advantage of multiple processors. Because these software technologies have their roots in a time when uniprocessor systems were more common and connectivity was the priority rather than performance, this is often not the case.

Connectivity. Because of the ubiquitous nature of network file sharing protocols, connectivity is a particular strength. Client and server software exists for virtually every operating system available. In general, connectivity is limited to the local area for all but the most casual levels of access. Data coherency may not be guaranteed by all solutions. For instance, NFS provides an optional locking mechanism to coordinate application access on different platforms, but its use is not required.

Capacity. Scaling storage capacity is typically simple and relatively inexpensive with network filesystems and NAS. However, their may be some hidden costs associated with scaling capacity. As a NAS device reaches its capacity or performance limit, it must either be replaced or an additional device must be added. This can create significant spikes in scaling costs.

Management. Individual NAS systems are easy to manage, but proliferation of NAS systems can unnecessarily complicate data management and limit flexibility. Creating a network file sharing infrastructure with the performance necessary for HPC may require balancing the data and workload across multiple NAS devices, further increasing complexity.

To simplify the deployment of networked storage for the most demanding environments, SGI has developed the SGI InfiniteStorage NAS 2000 and NAS 3000 storage systems. These highly scalable NAS platforms are capable of supporting over 100TB of storage in a single system while delivering outstanding data throughput that can range from megabytes per second to multiple gigabytes per second

Conventional NAS systems have significant scalability limits in three key areas: capacity, connectivity, and performance. Whenever, a limit in any area is reached an additional system is required, leading to excessive proliferation of systems that needlessly increases storage complexity. Each SGI NAS system is designed to replace a large number of conventional NAS systems in terms of both capacity and performance. Both systems are fully upgradeable to incorporate support for the SGI InfiniteStorage Filesystem CXFS for complete integration between NAS and SAN.

2.5 Data Lifecycle Management

Because of the cost and complexity of provisioning large amounts of disk storage, computer vendors during the mainframe era developed hierarchical storage management (HSM) systems capable of migrating data transparently between disk and lower cost storage media such as tape. Although such technologies have always had a market in the HPC arena, they never really caught on for general purpose computing because of the decreasing cost and increasing capacity of disk storage.

Now with data exploding everywhere, plus increased regulatory and corporate governance requirements, there is a renewed interest in data migration technologies that can match the storage medium to the usage and importance of each piece of data to reduce storage costs and provide data lifecycle management (DLM). DLM allows data to be appropriately and economically stored based on automated processes rather than time-consuming and error-prone manual archiving. DLM systems are not a replacement for the other technologies described in this section, but serve as an adjunct to them.

In typical DLM systems, whenever free primary storage drops below a preset threshold it triggers a migration process. Data on primary disk is evaluated against user-defined criteria such as time of last access, file size, owner, group, etc. and migrated to appropriate secondary storage such as less expensive SATA disk storage or tape libraries. Multi-tier hierarchies that use, for example, primary disk as the first tier, SATA disk as the second tier, and tape as the third tier, are possible. Files area automatically promoted or demoted within the hierarchy based on usage patterns. Any attempt to access a file results in it being moved back to disk storage.

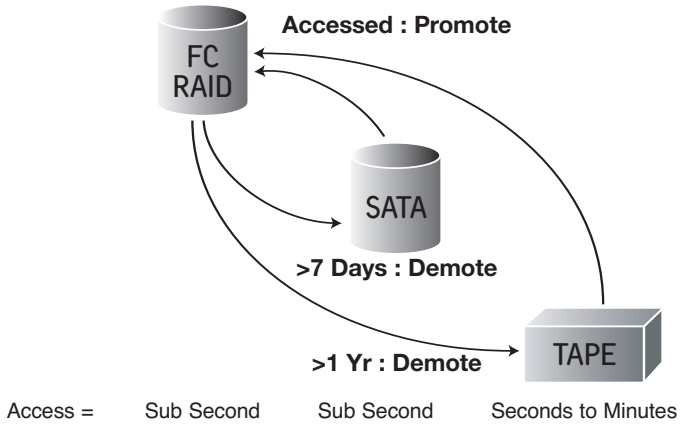


Fig. 2. Example 3-tier DMF deployment with high-performance primary disk, lower cost nearline disk (SATA), and tape. Data moves automatically and transparently up or down the storage hierarchy based on user-defined criteria such as time of access, size, owner and group.

Performance. Performance of DLM systems is typically characterized by how much total data can be migrated on a daily basis and how quickly a migrated file can be accessed. In practice, HPC users have found that they can create DLM systems that provide good performance while greatly reducing overall storage costs. Enough primary disk storage is provisioned to store the active working data set, while inactive data is stored on less expensive media where it is still available for immediate recall. Creating a system with adequate performance requires careful assessment of needs to ensure that the primary storage pool is adequate and that secondary and tertiary storage pools have adequate bandwidth.

Connectivity. DLM software does not typically support large numbers of directly connected clients, and may have limited support for heterogeneous client types. However, a system or systems with DLM can work in conjunction with network or cluster filesystems to extend the benefits to greater numbers and types of clients. A large number of clients may make it difficult to anticipate peak loads an increase the risk of overwhelming DLM systems.

Capacity. DLM systems offer the most economical capacity scaling currently available since they substitute less expensive media for primary disk storage. DLM is normally used in conjunction with the other technologies to decrease storage costs.

Management. Properly implemented, DLM solutions can dramatically simplify ongoing data management by creating a large, scalable, virtual storage pool that adapts readily to changing usage patterns.

The SGI InfiniteStorage Data Migration Facility DMF is designed to provide complete data lifecycle management, with the scalability to tackle the challenges of today's most data-intensive environments. SGI customers are already using DMF to manage petabytes of storage at a fraction of the cost of disk-only solutions. Busy sites migrate as much as 3TB of data per day between primary and secondary storage with no administrator intervention and no loss of user or administrator productivity. A storage infrastructure using DMF saves SGI customers 65% on average versus the cost of an all disk solution.

SGI InfiniteStorage Data Lifecycle Management Server combines all the hardware and software elements needed for a complete DLM solution in an integrated and easy-to-deploy package that can be purchased for use in either NAS or SAN environments.

2.6 Shared Filesystems

The development of SAN technology capable of providing exceptional bandwidth between servers and storage systems created the need for a new type of filesystem. Initial SAN deployments simplified capacity planning and storage provisioning and provided physical connectivity between servers and storage systems, but each server had to be allocated dedicated storage resources and data sharing was impossible.

Over the past five years, shared filesystems have been introduced to fill this void, taking advantage of the bandwidth afforded by modern SANs to provide much higher performance shared data access than has been previously possible. Most often the SAN uses a Fibre Channel fabric, but interest in Ethernet networks with iSCSI is increasing.

Typical shared filesystems blend the characteristics of traditional filesystems and network filesystems. The choices that are made in the design process have a significant impact on the scalability of the final product.

Performance. In shared filesystems, performance may be limited by SAN bandwidth. However, SANs can be configured with multiple switches and multiple RAID devices to increase bandwidth with multiple non-blocking paths through the SAN. Fibre Channel offers near linear scaling. With a shared volume manager, multiple hosts can share volumes striped across multiple RAID devices. Proper sizing of storage layout, I/O request sizes, cache sizes, etc. improves scaling performance.

The scalability of iSCSI for shared filesystems is not yet definitive. However, interest in iSCSI is high because of the relative economy of Ethernet versus fibre channel. Fibre channel HBAs and switches may represent a significant investment.

There are several implementation issues that may affect the performance of shared filesystems. The first is whether or not the software is implemented in kernel or user space. Filesystems that are closely integrated with the system kernel offer much better performance and can be designed to be POSIX-compliant for wider compatibility and investment protection. As with other solutions, they should be multi-threaded to take best advantage of large numbers of system processors.

The second issue involves the way storage access is coordinated to prevent data corruption. Some solutions use the same scheme as network filesystems, in essence implementing a file server on the SAN. This file server becomes a bottleneck limiting the performance achievable. More advanced solutions provide some mechanism for distributing metadata and allow each member of the cluster to read and write file data direct from storage.

Connectivity. Shared filesystems offer intermediate degrees of connectivity, often supporting tens of systems in a cluster. However, the number of systems that can be supported in a single cluster is considered a limitation by many in the HPC community, and vendors are working to increase these numbers. Some solutions may sacrifice data coherency in order to achieve greater scalability.

Likewise, many solutions offer support for multiple system platforms, but there is still room for improvement in comparison to a ubiquitous network technology such as NFS. To the extent that SANs can span metropolitan and wide areas, shared filesystems may offer the potential to share data over larger geographical distances than has been previously possible.

Capacity. As with traditional filesystems, the capacity of a shared filesystem is limited primarily by choices made by the software designers. Many shared filesystems are implemented on top of existing traditional filesystems and thus inherit their underlying limits, strengths, and weaknesses. The cost of scaling capacity for a shared filesystem may be somewhat higher than direct-attached or NAS solutions because of the cost of SAN hardware including switches and HBAs. On the other hand, the greater connectivity of shared filesystems versus direct-attached storage and the greater performance versus NAS may reduce the total cost of ownership to justify the additional expense.

Management. Shared filesystems introduce some complexity due to the need to configure SANs and other infrastructure including private network connections between cluster members as required by some solutions. Once configured, ongoing management tasks are relatively minor. Because of the excellent consolidation offered by these solutions, the overall storage infrastructure may be simplified. These solutions also have the capability to offload backup tasks from busy servers.

The SGI InfiniteStorage Shared Filesystem CXFS is a leading example of a shared filesystem that is in wide use at many of the most data-intensive HPC sites. CXFS has the largest number of installations and supports the greatest number of heterogeneous system platforms of any shared filesystem.

CXFS is built on the foundation of SGI's XFS filesystem, and benefits directly from the exceptional scalability and performance of XFS. CXFS is a fully multi-threaded, kernel-level implementation supported on all major operating systems (Linux, IRIX, AIX, HP-UX, Mac OS X, Solaris, and Windows) ensuring optimal performance. In practice, SGI customers have achieved data rates over 10 gigabytes per second with CXFS with even greater speeds possible using more and/or faster hardware. CXFS provides coherent file access to all hosts when buffered I/O is used. This may not be true for all shared filesystems. Direct I/O can be used for application-managed shared file access by multiple hosts.

CXFS provides synchronization by designating one system on the SAN as a metadata server, responsible for controlling file permissions and mediating shared access. Metadata transactions take place over a separate, dedicated network. Once the metadata server grants access, systems with CXFS read and write data directly over the SAN to and from disk.

Should a metadata server fail, a designated backup metadata server automatically takes over management of CXFS filesystems. This feature—in combination with an ability to failover metadata networks plus fully redundant SAN configurations and RAID storage—delivers extremely high availability along with exceptional performance.

CXFS is used with increasing frequency in wide area networks. A special version of CXFS for wide area networks, is available through SGI Professional Services. Wide Area CXFS utilizes networking technologies from either LightSand or YottaYotta to create a shared storage infrastructure. SGI and YottaYotta have demonstrated a CXFS cluster reading and writing to a shared file across thousands of miles at hundreds of megabytes per second.

CXFS works in concert with other SGI storage technologies including the SGI InfiniteStorage Data Migration Facility DMF for data lifecycle management. Many customers implement CXFS in conjunction with NFS to provide a single unified storage pool capable of delivering data at network speeds to desktop and workstation users and at extremely high speeds to compute servers for the most demanding computations.

3.0 Example Customer Deployments

The following case studies illustrate the deployment and use of the technologies discussed in the previous sections.

3.1 Geophysical Fluid Dynamics Laboratory

The Geophysical Fluid Dynamics Laboratory (GFDL) in Princeton, New Jersey, is at the heart of international scientific efforts to understand and predict the earth's climate and weather. GFDL is a Top 500 Computing Site and a recognized leader in applying advanced computing to simulate the behavior of the earth's atmosphere and oceans. Because of the very large data sets that result from advanced weather and climate simulations, storage and data management are as critical to GFDL's mission as computer systems.

GFDL needed a storage architecture capable of accommodating petabytes of data and very large individual files. Shared access with bandwidth of multiple gigabytes per second was critical to eliminate data duplication. A combination of CXFS and DMF software and state-of-the-art SAN hardware has been used to create a flexible storage architecture that can easily scale to meet GFDL's storage needs now and in the future.

CXFS provides a number of advantages for GFDL. It delivers shared bandwidth far in excess of what is possible with NFS and easily accommodates their largest files and filesystems. All cluster nodes have shared access to the same filesystems with bandwidth of from 2 to 4GB per second. Nodes are connected to shared storage—RAID and tape—by multiple brocade fibre channel switches in fully redundant configurations.

With DMF, GFDL manages several StorageTek® PowderHorn® silos with a combined tape capacity of 3.4 petabytes, a storage pool that is far in excess of what could be cost-effectively maintained with online storage alone. Archived data is always available and access to that data is transparent to GFDL researchers and will remain so as this environment grows to a projected 7 petabytes in 2006.

3.2 Weta Digital

Weta Digital is the digital effects arm of Weta Ltd., the Wellington, New Zealand production company that shot all three films in The Lord of the Rings trilogy. The scope and intensity required by work on the three films created significant data management challenges for the company.

SGI DMF was key to producing the entire trilogy. DMF delivers high-performance, reliable, and efficient data management with virtually unlimited storage capacity and dramatically lower total cost of ownership. On The Fellowship of the Ring, Weta Digital used DMF to manage 100TB of data and approximately 10 mil-

lion files, ranging from small to extremely large. A given file may consist of an element, a texture, one version of a shot, or a completely rendered image sequence. Adding the data from The Two Towers doubled Weta 's information storage to 20 million files and 230TB.

A key objective for WETA during production was to maintain as much free disk space for the artists as possible. DMF accomplished this by migrating data from online disk storage to tape storage provided on a StorageTek L700E robotic library with LTO multiple tape drives. This allows WETA to move data to nearline storage and retrieve it quickly as needed.

During peak operation, the 300 artists working on the films were moving over 1TB per day in and out of DMF-managed tape storage. A key advantage, besides reduced storage cost, is that artists never had to worry or care where data was stored. The quantity of storage required by effects-rich digital feature films, make data lifecycle management with DMF an extremely valuable solution. Active and completed films can be maintained by DMF so that artists can rapidly access whatever files they need at any time and focus on creativity and productivity rather than data management.

3.3 NASA Ames Research Center

NASA Ames Research Center provides supercomputing capabilities for many of NASA's most important projects. From modeling the aerodynamics of potential space shuttle replacements to predicting Earth's climate far into the future, these critical projects encompass some of the most challenging computational problems ever undertaken.

To meet storage needs for existing and future projects, NASA Ames has deployed a complete data lifecycle management solution using SGI DMF software to manage over 1,377TB of data with NASA researchers adding an additional 1 to 5TB per day.

NASA Ames has used DMF for over 6 years and has been pleased with the overall solution in terms of performance, reliability, and ease-of-use. NASA estimates that the cost of DMF with new tape drives and media delivers a cost savings of 5 to 10X versus the cost of new RAID storage[1].

The NASA Advanced Supercomputing (NAS) Division at NASA Ames recently worked with SGI and Intel to deploy a 10,240-processor system consisting of 20 SGI Altix nodes each with 512 processors and global shared memory. The supercomputer—named Columbia in honor of the crew lost in the 2003 shuttle accident—is currently the world's most powerful operational supercomputer. SGI CXFS and 440TB of RAID storage will be employed to provide a high-performance, shared data infrastructure for the system.

The combination of Columbia's CXFS infrastructure with the existing DMF data lifecycle management system will result in an integrated 1.4 petabyte storage system for seamless data access and management, arming NASA to pursue the many new endeavors targeted for coming years.

3.4 Meteo-France

Météo-France is the French national weather forecasting agency. Among its missions, Météo-France must collect detailed information about the ocean surface, the atmosphere and the Earth's snow cover. Based on this data, weather conditions are forecasted and made available throughout the country. Each day, more than 2000 weather maps are produced. In addition, Météo-France also acts as a repository for archived climatology data.

Applications such as weather modeling, climate evolution studies, forecasting pollution peaks and predicting ocean conditions require a huge amount of compute power. Météo-France uses supercomputers that sustain more than one Teraflop and process large volumes of data daily. According to Alain Beuraud, head of HPC and Storage Department at Météo-France, "The shorter the time required for the supercomputers to write their results, the more we save. Thus, we need high performance storage. The volume of data is so huge, that we could not store all data on expensive media". This is why French Met developed a migration policy for each of its fifteen applications, based upon data size and upon the probability of re-use of data files. "Our storage solution must be able to write on disk and tape, it must handle a multitude of small files, and it must transfer more than one Terabyte of data daily," adds Beuraud.

Météo-France relies upon SGI DMF software to meet these requirements. Data is stored in a four-level hierarchy which includes Fibre Channel and SATA disks and high performance and high capacity tape systems. A SAN provides continuous access to files at 2 GB/s. The solution also provides disaster recovery, which is automatically managed by DMF. The storage system currently contains more than 250 TB of data and data retained in the environment is growing by ~15% per year.

[1] See: "SGI Data Lifecycle Management at Nasa Ames" http://www.sgi.com/products/storage/success.html

The high performance of the storage system allows users to quickly browse through a large number of files (around 10 million). The reliability and robustness of the whole file service combined with the flexibility of DMF—which enables administrators to customize usage of the disk caches to decrease tape usage—has helped Météo-France to increase the amount of satellite-collected data, further refine its weather models, and meet other exploding application storage demands.

3.5 Pittsburgh Supercomputing Center

The Pittsburgh Supercomputing Center (PSC) is a joint effort of Carnegie Mellon University and the University of Pittsburgh together with the Westinghouse Electric Company. It was established in 1986 and is supported by several Federal agencies, the Commonwealth of Pennsylvania and private industry. PSC is the most powerful facility in the United States that is committed solely to public research areas such as earthquake preparedness, AIDS research, storm prediction and protein folding for biotechnology and pharmaceutical applications.

PSC is also a member of the TeraGrid project the world's largest, fastest, distributed infrastructure for open scientific research. The TeraGrid includes 20 teraflops of computing power distributed at five sites and will also include high-resolution visualization environments, and toolkits for grid computing. These components are tightly integrated and connected through a network that operates at 40 gigabits per second—the fastest research grid on the planet.

Pittsburgh Supercomputing will use DMF software to manage all data generated by its Terascale Computing System, currently used by over 1,000 researchers worldwide as it expands its data store from 300TB to over 1PB. As a DMF user for over 10 years, PSC is extremely confident in both the scalability and reliability of DMF.

PSC chose to migrate data from its previous Cray environment to SGI and DMF for several reasons. First, it was not necessary to change the data format to accomplish the migration. All that was required was a database conversion with no change to the data itself. This seamless transition resulted in significant savings in time and money. In addition, the SGI® Origin™ system that replaced the retired Cray was compatible with the existing UNICOS® OS-based storage tapes. The Origin system interfaces with existing tape libraries, while DMF manages data placement and migration policies.

3.6 BBC Broadcast

BBC Broadcast is part of BBC Ventures Group, the BBC's new commercial media services business, offering a comprehensive range of services to play-out, publish, promote and provide media access for content across all media platforms. Key services include: new channel launches, play-out and channel management; channel branding and promotion; and subtitling and other media access services.

BBC Broadcast chose SGI® InfiniteStorage NAS 3000 and CXFS as the foundation of a tapeless environment for virtually all its processes. The system enables media assets to be available for a large number of requirements such as ingest, archive and playout. Every piece of content that touches BBC Broadcast will ultimately go through this system.

The combination of scalable SGI NAS architecture and CXFS shared filesystem provides BBC Broadcast with the ability to store and move broadcast assets from ingest through production to playout. CXFS provides a resilient high-performance core for activities such as transcoding and aspect ratio conversion, while NAS supplies the storage and bandwidth required by hundreds of concurrent 'desktop' browse video users.

4.0 Conclusion

Deploying multi-petabyte storage systems will become simpler over the next ten years as network speeds and storage densities continue to increase following Moore's Law (which predicts that data density and bandwidth will double approximately every 18 months). However, significant challenges will remain as HPC leaders continue to push the envelope with new methods that require ever faster storage access and greater capacity. There is no analogue to Moore's Law that applies to data management software. Creating the software solutions necessary to meet the challenges created by large, dynamic data sets can only be achieved through exceptional expertise, commitment, and strategy.

As the leading provider of data management software for HPC for over a decade, SGI continues to pursue the software breakthroughs that ease the burden of storage management so that scientists and engineers can focus on research and innovation without concern for where and how their data is stored. SGI is researching advances in a wide number of areas with important implications for future data management:

- Multi-petabyte storage systems make the traditional directory path and file name cataloging system used by filesystems both insufficient and impractical. With millions of files in a filesystem, managing the filesystem content based on filenames is a task that current tools cannot manage. A filename alone is rarely sufficient to describe the contents of a file. Tomorrow's filesystems must provide a built-in search capability that allows users to find files based on multiple attributes including those generated automatically based on file content and user-supplied information.

- High performance, high capacity filesystems necessarily span large numbers of hardware devices. If any hardware element is not available, access to the entire filesystem may be lost. Filesystems must become more reliable and serviceable so that a filesystem can remain online while underlying parts of the filesystem are offline due to maintenance, failure, etc. Tools are needed to allow maintenance and repair of partial filesystems without taking the entire filesystem offline.

- The computing environment of HPC customers is rarely restricted to a single geographic location. Filesystems of the future must be able to seamlessly span geographies without sacrificing performance, reliability or maintainability.

These and other innovations promise to not only simplify storage management, but to change the way people work and enhance the benefits of continued rapid improvements in HPC for the benefit of mankind.

Corporate Office
1500 Crittenden Lane
Mountain View, CA 94043
(650) 960-1980
www.sgi.com

North America +1 800.800.7441
Latin America +55 11.5509.1455
Europe +44 118.912.7500
Japan +81 3.5488.1811
Asia Pacific +1 650.933.3000

© 2005 Silicon Graphics, Inc. All rights reserved. Silicon Graphics, SGI, IRIX, Origin, Altix, XFS, and the SGI logo are registered trademarks and, CXFS, and FailSafe are trademarks of Silicon Graphics, Inc., in the United States and/or other countries worldwide. Mac OS is a registered trademark of Apple Computer, Inc. UNIX is a registered trademark of The Open Group in the U.S. and other countries. All other trademarks mentioned herein are the property of their respective owners.

3796 [04.2005]

J14888