White Paper

Delivering Guaranteed-rate I/O to SANs in Video and Film Postproduction Facilities

Eric Eppe

Table of Contents

1.0 Introduction

This SGI white paper discusses the use of shared filesystems and storage area network (SAN) technologies in video and film postproduction. It examines how standard SAN technologies can allow the storage infrastructure to become overburdened and as a result introduce unacceptable delays to critical media data flows. It then explains how the new SGI Guaranteed-Rate I/O version 2 (GRIO V.2) product, along with the SGI® InfiniteStorage Shared Filesystem CXFS™, can greatly help facilities get the best return on investment from a SAN.

2.0 Common Use of Digital Assets in Postproduction Facilities

Most of today's films, episodic television programs or advertisements are postproduced as a collaborative effort. Multiple skilled creative operators use specialized computer-based applications for effects, nonlinear editing, color correction, and so on to handle different aspects of the creative process.

Traditionally, these different applications were done sequentially, and content was copied from one machine to another – as media (film, audio or video) and more recently as data.

Increasingly, companies are looking to streamline the production cycle and enable their creative talent to be more productive by finding ways to eliminate file copying yet still maintain reliable delivery of content to the artists when they need it.

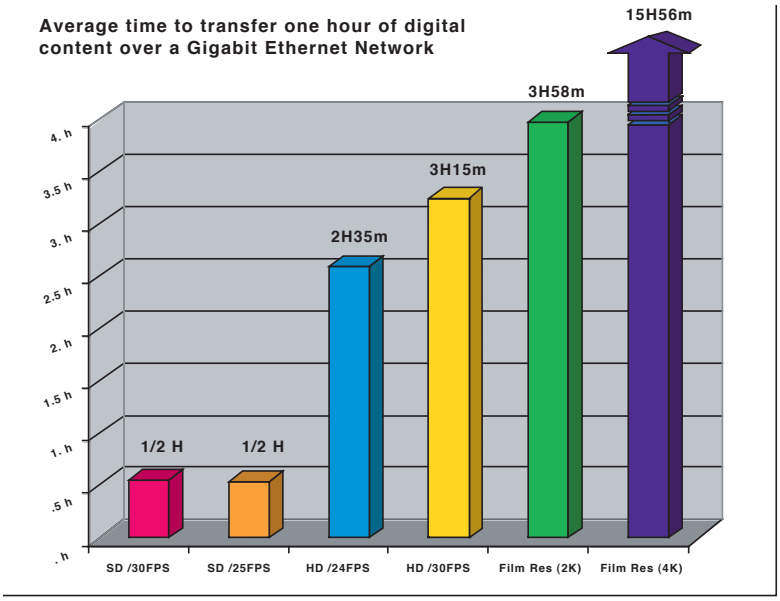
These applications require that data be delivered to or from the filesystem at a fixed rate with little scope for interruption or variation. This requirement is clearly incompatible with the direct use of a network filesystem such as NFS or SMB because it is almost impossible and not cost effective to guarantee that a TCP/IP network will deliver content at an acceptable and reliable rate. In any event, standard networks only deliver a fraction of the total bandwidth of a Fibre Channel network.

Thus, digital assets are commonly used on local dedicated Direct Attach Storage (DAS) by one computer workstation (such as color correction), then copied to the next workstation in the workflow (such as effects) via NFS, CIFS, or FTP.

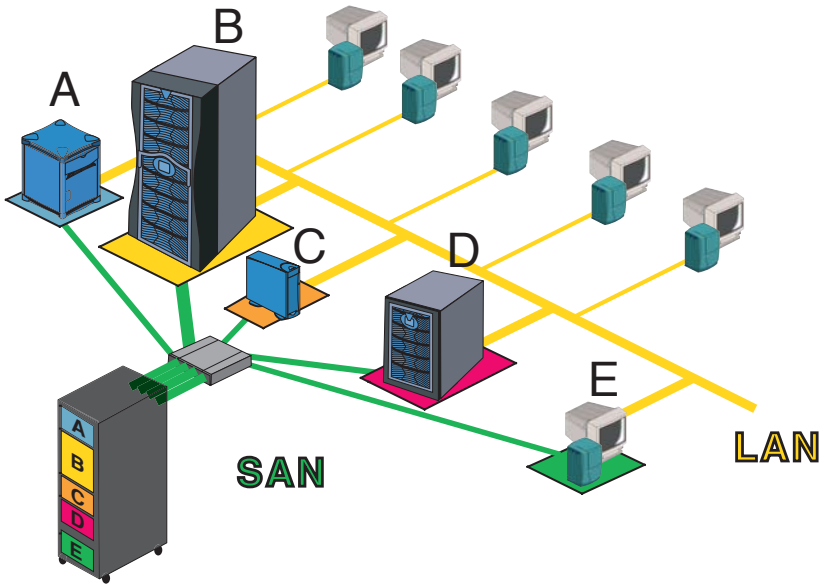
In addition, the need for digital assets with increased quality (such as High Definition, 2K or 4K and greater resolution formats), and therefore increased data size, has introduced data-management issues.

As an example, transferring a 1-hour 2K-film-resolution sequence (~1.1 TeraByte) over a standard Gigabit Ethernet network typically takes many hours, and in the case of busy networks, can take days. It also requires redundant, equivalent disk space on the next workstation in the workflow. This is an inefficient use of the resources, plus a loss of time and money.

Several new technologies have arisen to allow a better postproduction workflow, but until now none of these has allowed applications to share digital assets and be guaranteed that I/O operations will occur without dropping frames. GRIO V.2 addresses this requirement.



Storage Consolidation SAN



3.0 Storage Area Network (SAN)

A SAN is a dedicated, specialized network that transports data at high speeds between a set of disks (targets) and a number of computers. This concept has grown with the availability of the Fibre Channel Arbitrated Loop (FC-AL) standard. It is now commonly used in its switched fabric mode (FC-SW), meaning that many computers can access multiple disk arrays without having dedicated connections to each of them. A specific Fibre Channel switch (comparable to an Ethernet switch) performs real-time switching between the resources and all connected computers.

Because Fibre Channel is dedicated to data transport (whereas Ethernet was designed with terminal services in mind), it is much more suitable for data serving than is Ethernet. As a comparison, Ethernet has a standard and fixed payload (size of the vehicle that will carry real data) of 1500 bytes, whereas Fibre Channel can dynamically adjust its payload size from 512 bytes to 64KB, enabling a more efficient use of the available bandwidth.

While a SAN offers a comprehensive way to consolidate storage resources that previously used directly attached storage (DAS) into a single array, it does not provide data sharing mechanisms between storage volumes in the SAN.

It is only since 1999 that a true high-performance SAN shared filesystem has existed: SGI's InfiniteStorage shared filesystem, CXFS.

4.0 Shared Filesystems

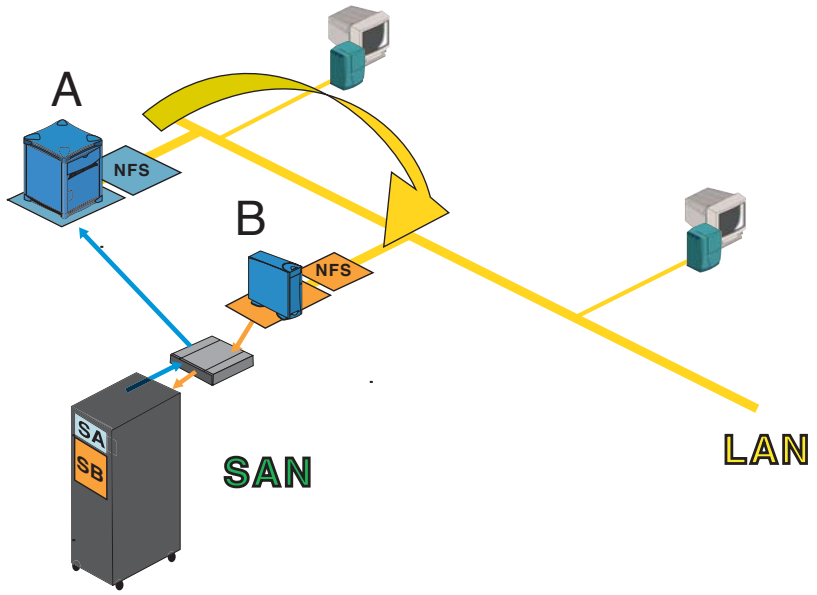
SANs have allowed more flexible use of storage resources and a more effective way for organizations to invest in storage equipment. However, a SAN by itself does not permit the computers connected to it to share a volume, which is a set of disks organized to look like a single disk. Shared filesystems allow many computers to mount the same filesystem and treat it as if it was local, thus enabling direct access to files using a Fibre Channel SAN link.

In a standard SAN, the computers are only linked to their own storage volumes. Each computer can only see its zone, which may include one or more volumes. It is not possible for one computer to use the same volume as another computer concurrently, without immediately corrupting data.

In today's world, most businesses must implement a data workflow of some sort, meaning some data generated on one computer must be transferred to another computer. To implement this workflow, most organizations have implemented network file serving.

In the example below, computer A serves its data using NFS. Computer B must read data from A, thus computer A uses an NFS server to read data from its storage volume SA. Data is transported through the Ethernet network, then computer B writes this data on its storage volume, SB. SB happens to be on the same storage array as SA.

File Movement within a SAN using the LAN



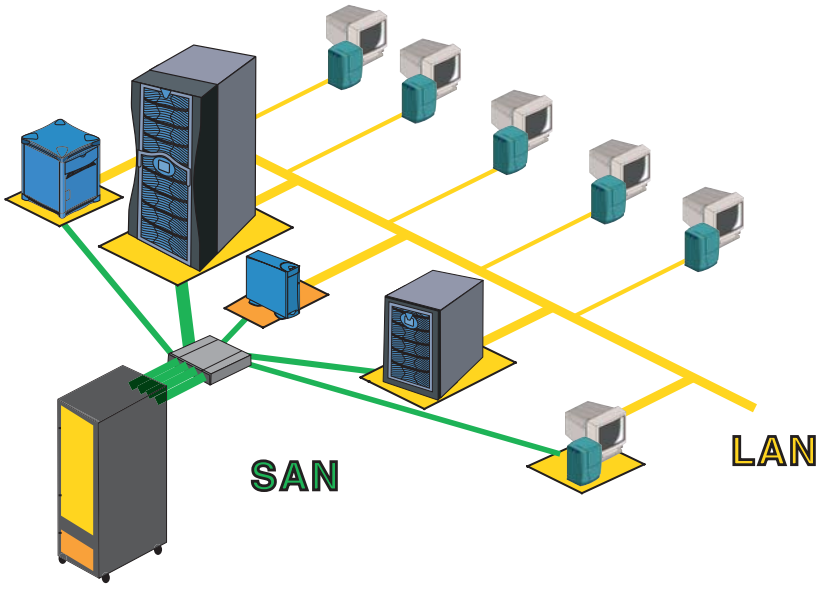
Because computers cannot natively share the same SAN array, data must flow from one part of the storage array (SA, owned by A) to another part of the same storage array (SB, owned by B) through the network. This transfer of data wastes disk space, network resources, and user time.

Over the last 25 years, many computer companies have tried to build a clustered filesystem. Examples include Digital Equipment Corp with the VaxCluster File System and Cray Inc. with the Cray® Shared File System. Although these filesystems were effective for their time, they were bound to very specific storage architectures, homogeneous computer architecture, and the same operating system across the board.

In the Media marketplace, some vendors such as AVID have recently delivered such systems as the Avid Unity™. They could be seen as Shared/SAN filesystems although they are in fact dedicating overbuilt disk space to a very restricted number of homogeneous (AVID) workstations. Hence, these solutions are not coping with the heterogeneous nature of the equipment found in most postproduction facilities.

With the advent of SAN, engineering a heterogeneous, kernel-based, high-performance and scalable shared filesystem has become the Holy Grail for many companies and universities. Because of the complexity, only a handful of companies have come close to achieving this goal. In 1999, SGI has been the first to achieve it by delivering CXFS.

SAN with CXFS Shared Filesystem



In the SAN with CXFS Shared Filesyetem example, each computer uses the same large CXFS shared filesystem (in yellow). There is no need to transfer data from one computer to another using the Ethernet network because data resides on the same equally accessible filesystem. This simplifies the data workflow and optimizes the available disk space by removing the need to duplicate files.

CXFS is a truly heterogeneous, high-performance, no-compromise shared filesystem. It supplies shared filesystem services to all major operating systems including SGI® IRIX®, Linux®, Microsoft® Windows®, Apple® Mac OS® X, Sun Microsystems Solaris™, and IBM® AIX®.

5.0 Benefits for the Film Mastering and Postproduction Business

Generally speaking, there are many benefits to using a SAN with a shared filesystem. The following are particularly important for a film studio, mastering and postproduction facility:

- Overall improvements in data access times and an increase in available bandwidth by avoiding NFS and LAN bottlenecks. While LAN throughput can scale to gigabits/sec, SAN bandwidth can effectively scale to gigabytes/sec (GB/s) per computer. For example, CXFS has demonstrated up to a 12-GB/s bandwidth on a single production computer. This is especially important because new film resolutions are requiring more bandwidth than before, such as 1.2 GB/s for a 4K stream. Using CXFS results in a dramatic reduction in the overall time needed to edit and finish a sequence because it avoids unnecessary movement of data between computers in the workflow. Operators do not need to know about digital assets location; the assets are visible to everyone as if they were local to each computer.

- An increase in the overall available space. In most facilities, each computer has a local storage subsystem that must be big enough to handle the work on a specific dataset size. For example, 15 minutes of 2K film resolution with all temporary files can easily consume one Terabyte of storage disks. Assuming a 80% utilization, there are 200 GB unused per workstation. In a facility, these unused spaces can easily sum to several Terabytes. With a shared filesystem, it is possible to reduce the overall amount of space required and to use the resulting free space for other computers or to ease dramatically the storage administrator's job by canceling daily space-freeing activities. By reducing the number of copies of a file in the workflow, a shared filesystem further reduces the aggregate demand for disk storage.

- A decrease in backup time and costs by eliminating backup complexity and bottlenecks. Backup traffic is removed from the LAN and a single shared volume means one centralized backup.

- Optimization of the facility's biggest financial investments: the film Scanners and Telecinés. CXFS allows faster and concurrent access to scanned frames and permits color correction to be done directly from disks instead of tying up the Teleciné to a single color-correction system.

- Creative talent, the facility's most important asset, gets to spend their time being more productive, instead of wasting time moving or waiting for files.

- Unlimited scalability. CXFS can scale to 18 million Terabytes, allowing any facility to store, theoretically, millions of full-length feature movies. It is no longer necessary to move video clips to tape when expanding storage capacity. Optionally it is possible to manage the data life cycle by using SGI® InfiniteStorage DMF which will automatically move files to the appropriate and most cost effective medium, such as slower disks and/or to tape libraries. No user intervention will be needed but all files will remain visible to the SAN users.

- No single point of failure because CXFS provides redundancy of all components.

6.0 Unprotected SAN

An *unprotected SAN* is a one that is vulnerable to oversubscription from attached computers, resulting in unpredictable performance degradations which in turn can result in application failures, such as dropped frames. This can result in a real loss of time and money when using expensive assets like a film scanner, a Teleciné, or when a digital intermediate session needs to be re-done due to loss of data.

The factors involved in dropping frames are numerous, but the most frequent is delayed disk operations, a syndrome that occurs when a scheduled set of I/O – such as writing 24 DPX frames a second to disk – is subject to uncontrolled concurrent I/O activity from another application that is using the same paths to the target storage system. Application vendors usually minimize these risks by providing architectural workarounds, but with no real guarantee on frame rate accuracy.

Today, most computer-based postproduction applications that require a high QoS are using a combination of directly attached storage (DAS) of different kinds (SCSI JBODs, Fibre Channel

JBODs, or RAID), and sometimes a specific computer and application architecture that minimizes the side effects of oversubscription. This dedicated unit is controlled and owned by the attached computer. The volume management and the filesystem can be specific to the platform and are also controlled by either the operating system or the user application itself.

This architecture gives good performance and guarantees but is not meant to give equal access to the content for all workstations in the facility. In fact, in most "real time" architectures, the filesystem or its equivalent is locked by the application and even the operating system has no control over it.

Accommodating this architecture in a SAN is a risky solution because other computers will attempt to use the same storage subsystem. The contention for disk access, using the same switch or the same set of RAID controllers, can cause a storage system to be overburdened and make reliable delivery of media data streams impossible.

Because the SAN architecture offers several key benefits, such as a centralized management and a consolidated space allocation, several vendors have adopted this architecture and try to mitigate the risks by over-building the storage subsystem. However, over-building is costly. The following example shows how much an overbuilt SAN would cost to a typical postproduction facility.

XYZ Productions is a (fictitious) postproduction facility that operates in the film business and also does some television advertisements. Its SAN needs are as follows:

For XYZ Productions, an overbuilt SAN of the required size can be calculated using the following formula, where 0.65 is the average sustained bandwidth utilization for a RAID (average sustained bandwidth / peak theoretical bandwidth), and 1.3 is the overbuilding factor that will secure a seamless operation to all workstations:

Target system sustained bandwidth requirement
= 1.3 x 3201 / 0.65

In this case, the bandwidth that XYZ Productions would have to pay for is equal to 6400 MB/s (~2 times the price of its real bandwidth requirements). For the same available disk space, this increased bandwidth infrastructure is likely to cost an additional $350K to XYZ Productions.

It is unlikely that all the workstations will need the sustained bandwidth at the same time because operations are not synchronous by nature. However, the decision to overbuild a SAN infrastructure is always a trade-off between the investment to be made and the time you can potentially lose on a specific production if you do not get the data in a timely manner.

Rather than building an overbuilt SAN that is not cost effective and does not fully ensure real bandwidth protection, XYZ Productions does have another choice. A software-based alternative does exist that enables applications to be given guaranteed bandwidth, resulting in real money and time savings. That alternative is SGI InfiniteStorage Guaranteed-rate I/O Version 2 (GRIO V.2).

Definition/ bandwidth requirements	Bandwidth per node	Number of workstations	Total bandwidth needed
Standard Definition (SD) at 30 Fps in RGB @10 bits def	42 MB/s	10	420 MB/s
High Definition (HD) 24P RGB @ 10bits def	199 MB/s	4	796 MB/s
2K film resolution	306 MB/s	2	612 MB/s
4K film resolution	1223 MB/s	1 scanner and 1 station running on local storage	1223 MB/s
Unqualified applications doing I/O	Less than 10 MB/s Average 5MB/s	30	~ 150MB/s total
		Bandwidth Grand Total	3201 MB/s

7.0 Building Quality of Service for All SAN Users

In the overbuilt SAN approach, the attention is focused at the device level; this method tries to correctly size every component in the data path to avoid any contention at the workstation level. This bottom-up approach is lengthy, costing time and money – and although it offers some level of comfort, it does not offer a guarantee that no frame loss or jitter will occur. Any change in the configuration can significantly alter the results.

A different approach is to make no guess on any individual component performance behavior in the presence of contention, but rather, to provide a top-down design that will prevent any oversubscription of the available bandwidth. SGI GRIO V.2 follows this path.

The idea is to qualify the globally available bandwidth on the SAN (the qualified bandwidth) that can be consistently delivered with the necessary responsiveness and then provide mechanisms for applications and/or computers in the SAN to be granted access to dedicated portions of this bandwidth.

This ensures that no other application or computer connected to the SAN will disturb this bandwidth guarantee.

This approach is deterministic in the fact that it will not allow applications to use more bandwidth than available globally. By comparison, the overbuilt SAN will not stop hungry applications from consuming more bandwidth than available, potentially resulting in frame losses and/or jitter.

For the customers requiring Quality of Service on their shared SAN, SGI is now delivering GRIO V.2 as an optional feature to the XFS® filesystem or to CXFS SAN.

GRIO V.2 has the following basic components:

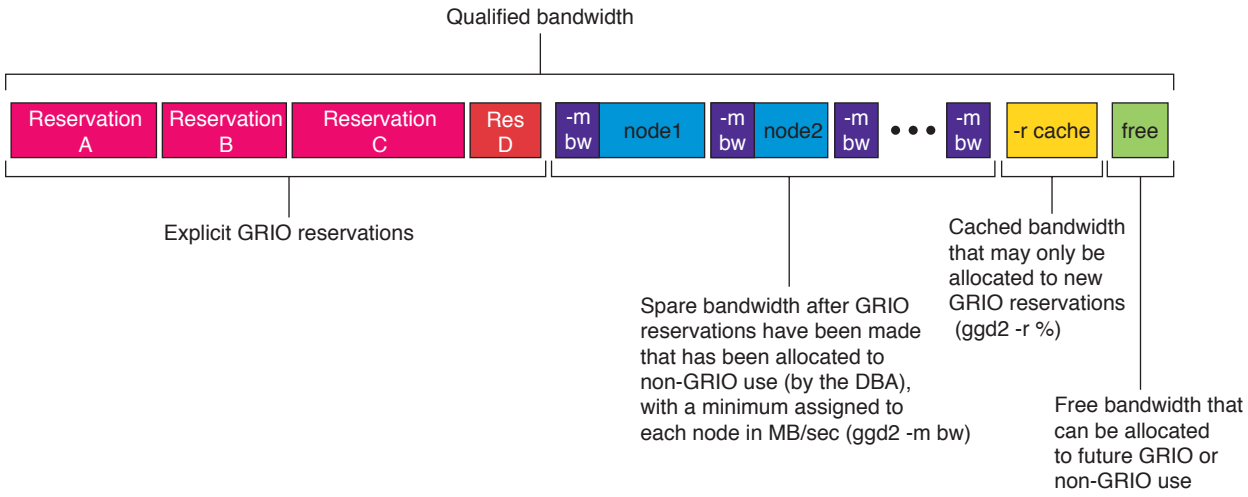
- It is available either as a standalone component to guarantee bandwidth to a single node with directly attached storage (DAS) or as a component in a SGI CXFS SAN, providing a SAN-wide bandwidth guarantee to all nodes participating in it. (This paper only discusses the SAN mode.)

- It provides several mechanisms for an application to request a bandwidth guarantee, known as stream creation.

- API-based dynamic bandwidth reservation. In this mode, user applications are GRIO-aware in the fact that they are modified and linked to the GRIO libraries. This mode is the one that offers the most flexibility in terms of stream creation, stream change, and so on. This API offers the platform for ISVs to take full benefits of the shared SAN without having to take care of bandwidth issues in the SAN. This API is freely available to any vendor or customer.

- Per-node static bandwidth allocation. As most applications are not GRIO-aware today, it is important that a mechanism exists to provide them with guaranteed bandwidth. In this mode, the bandwidth is guaranteed on a per-computer basis, for all applications that run on it. The stream creation is requested by the user, with a command-line program, for a specified number of bytes per second.

- Anonymous stream creation. In a postproduction workflow, some computer workstations must get a guarantee for their I/O, in particular the editing workstations, the effects workstations, and the scanner; other applications might not need any guarantee, just access to the same filesystems.



However, any computer accessing the SAN is performing I/O and hence has the potential to disrupt other allocated bandwidth reservations made by higher-priority computers. That is why all computers participating in the CXFS cluster must be GRIO aware. To deal with this issue, the computers that have not explicitly made a reservation (with the API or the command line) are considered lower priority.

- GRIO V.2 allocates portions of the remaining bandwidth to these computers. The given bandwidth is then throttled up or down by the CXFS client itself to the given limit, this limit is reconsidered on a regular basis (every 2 seconds). This mode is considered lower priority by GRIO V.2.

- GRIO V.2 provides tools and mechanisms to monitor the qualified bandwidth and fine-tune the stream characteristics. The qualified bandwidth is monitored before the SAN becomes generally available to users by having the customer generating a realistic production workload. The number found (X number of MB/s) will serve as the maximum bandwidth available on the SAN, no bandwidth will be allocated beyond this number. Of course, if a change occurs in the SAN architecture (such as new disks, controllers, or computers), it will be necessary to re-qualify the available bandwidth.

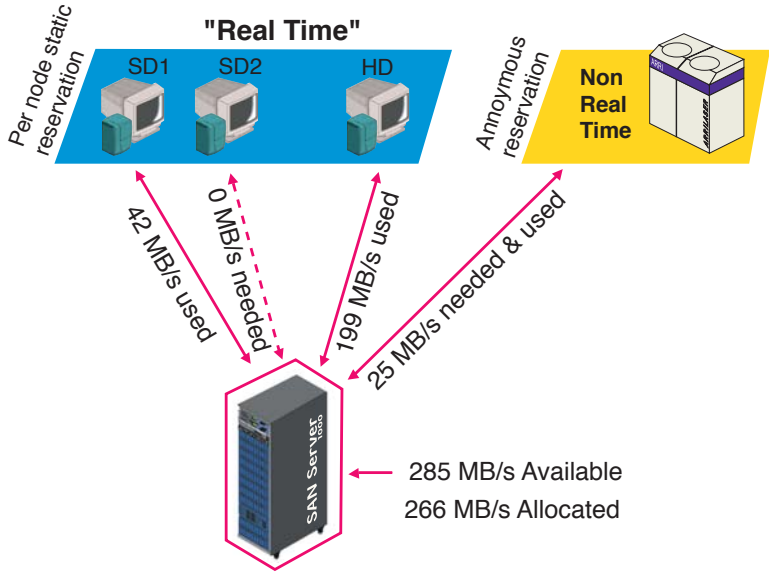
- Qualified bandwidth is managed centrally by a daemon generally located on a server capable node but has no relationship with the CXFS server. Client computers use fast RPCs to this server to manage bandwidth requirements. This service is fault tolerant, thus, in the event of a failure of the computer on which this daemon resides, another server on the cluster is dynamically elected and all further requests are directed to this new server.

- GRIO V.2 is available on computers running IRIX but will shortly be available on all CXFS platforms including Windows® XP and Windows® 2000, Linux 32-bit, and Mac OS X.

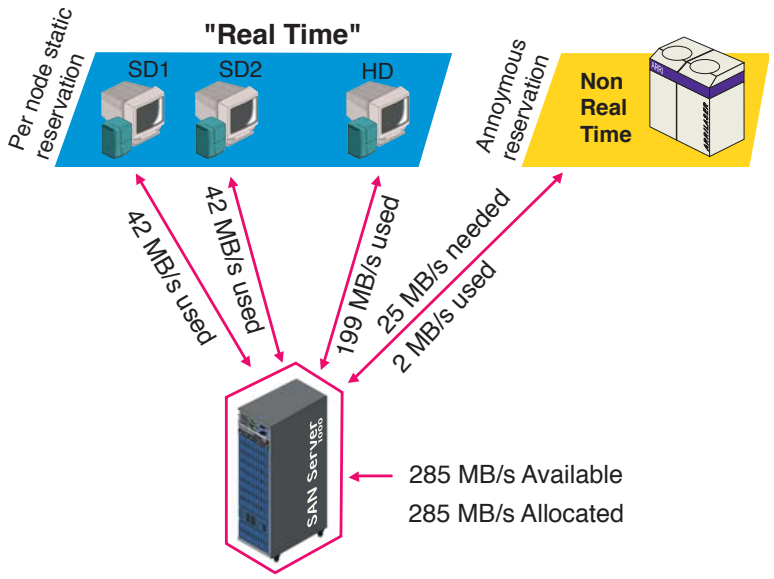
The following simple example illustrates some of GRIO features: This configuration has three editing workstations with high guaranteed-bandwidth needs (two at standard definition, SD1 and SD2, and one at high definition, HD), plus one film recorder that has no need for guaranteed bandwidth (maximum 25 MB/S).

The qualified bandwidth available on the SAN is 285 MB/s, At instant T, SD2 is not active, hence the bandwidth available is largely enough for one standard-definition stream and one high-definition stream plus 25MB/s used by the film recorder.

Situation at instant T



Situation at instant T+1



As soon as SD2 becomes active (T+1), and 42MB/s are needed for this station, GRIO will throttle down the (non priority) file recorder station to the remaining available bandwidth (from 25MB/s to 2 MB/s) and then will allocate 42MB/s to SD2.

The benefits of this architecture are considerable:

- Because applications can get bandwidth guarantees, it is now possible for them to directly use files available on the shared SAN filesystem without wasting time copying files between workstations in the workflow.

- Better use of the resources by not overbuilding the SAN. In our XYZ Productions example, the protected SAN was overbuilt by a factor of 1.3, which could be directly saved using GRIO V.2.

- Noninvasive and open architecture. If some application vendors such as AVID are already selling SAN solutions with equivalent capacities, their use is restricted to the vendor's applications that have been customized to take advantage of the solution, and no other vendors can hook their applications to these SANS.

  The origin of an application makes no difference to GRIO V.2, whether it is a Discreet® smoke®, or Apple® Final Cut Pro®. CXFS with GRIO V.2 is the first general-purpose shared filesystem to provide support for these real-time media workflows.

- Interfacing flexibility. If a vendor chooses to take advantage of the full flexibility of GRIO V.2, the API is open. However, there is no obligation to use the API.

- Unlimited scalability. The SAN is not frozen in a specific configuration (as opposed to other vendors), but can easily be extended to accommodate growth or changes in operations. GRIO V.2 will manage the qualified bandwidth, no matter what its size.

- Seamless integration of 'silent' applications/computers that are often used in a post-production workflow for non-interactive jobs, such as film recorders.

- Simplified operations and greater flexibility. Because post-production facility workflows are dynamic by nature, bandwidth requirements for each workstation can easily change on a per-day or per-hour basis. GRIO V.2 allows site administrators and users to configure dynamically the bandwidth requirements and use only the needed resources, hence maximizing the use of the facility's equipment.

8.0 Summary

It is obvious that the standard network protocols such as Ethernet, NFS, or CIFS cannot cope with growing needs for high-definition file formats (HD, 2K, 4K); they are too slow for these file formats and even if the bandwidth was sufficient, they cannot guarantee that all frames will be reliably delivered in a timely fashion.

A SAN using Fibre Channel is a much better suited data-transport mechanism. It can scale almost linearly and, most importantly, with the SGI InfiniteStorage Shared Filesystem CXFS it offers a very efficient data-sharing mechanism, allowing a

seamless workflow of digital assets in postproduction facilities. With GRIO Version 2, SGI is now offering a unique quality-of-service mechanism that enables applications with high-bandwidth requirements to be guaranteed to share data, at full speed, while delivering the frame rates and quality of service needed to sustain high-performance media deployments.

GRIO V.2 will fit most postproduction facilities needs. CXFS offers support for most of the operating systems currently in use in this industry, including IRIX, Mac OS X, Linux 32-bit, and Windows. GRIO V.2 is also scalable to meet tomorrow's needs for guaranteed bandwidth.

Because there is no need to make expensive copies of the same digital assets, and due to the fact that GRIO enables highly demanding workstations to stop locally caching their digital assets, CXFS and GRIO can enable postproduction houses to save a tremendous amount of time and money.

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 +44 118.925.7500
Japan +81 3.5488.1811
Asia Pacific +1 650.933.3000

3647 [04.13.2004]

J14568