A Survey of the Landscape:

Technology for Shared Storage

A Guide to SAN Management Software for Mac OS X

Updated August 2005

***Abstract**: This paper gathers and discusses known facts regarding different software products. Authored by the CommandSoft® people, who make the FibreJet® product line, it was written in the "work together" spirit and is encouraged background reading for all customers researching technology for shared storage. Customers are encouraged to perform their own due diligence and ask questions. This paper will be updated from time to time to include new relevant information. We would like to hear from you so send your questions, comments or suggestions to feedback@commandsoft.com.*

Version 1.0 Released January 2004

Version 1.0.1 Released March 2004

Per readers comments, updated FibreJet section to describe more details about its ability to "claim" and also "unclaim" storage. Also updated this section to describe some database failure situations. This should clarify how FibreJet behaves in these situations.

Version 1.0.2 Released August 2005

Added Xsan section. Update FibreJet, ImageSAN, SANmp and Charismac sections to reflect latest findings.

SAN Background

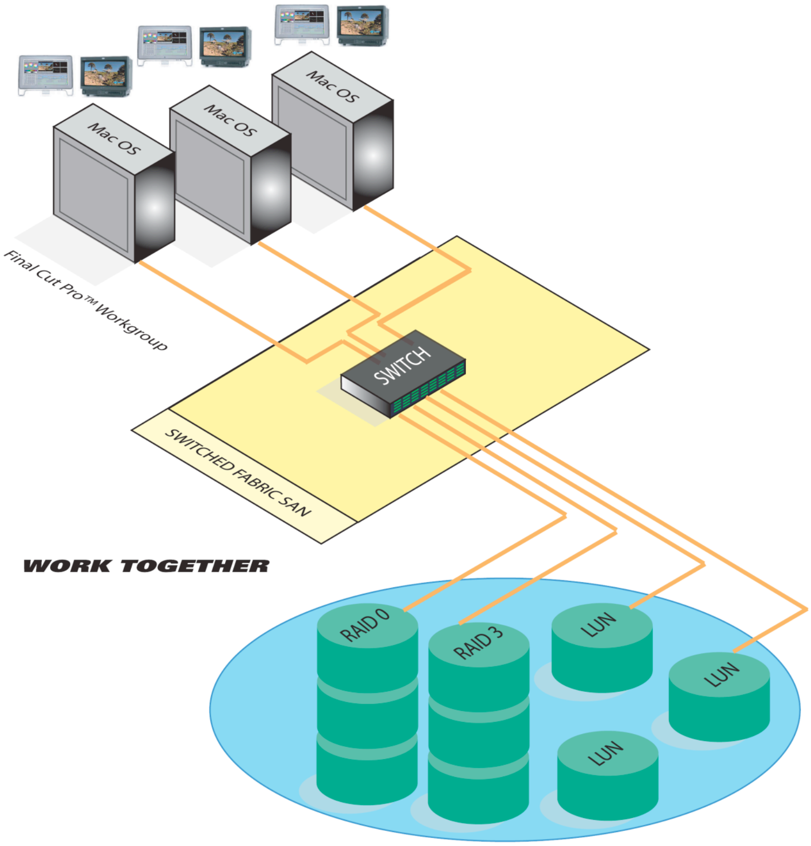
A **Storage Area Network** (**SAN**) is a network that allows computers to be connected to storage devices in a fashion that allows many computers to connect to many storage devices. Before SANs, storage was normally connected to only one computer at any given time, with few exceptions.

ONE-TO-MANY — STORAGE CONNECTED TO ONE COMPUTER

As more general purpose, and longer distance connection technologies such as fiber optics and the Fibre Channel protocol were developed, these SAN options became available to general computer users.

MANY-TO-MANY — A SAN, STORAGE CONNECTED TO MANY COMPUTERS

Although there are different hardware and protocols you can use to build a SAN, one popular method uses fiber optic cabling, and the **ANSI Fibre Channel** (**FC**) protocol standard. A network is built using FC switches that connect the computers and storage into what is known as a FC fabric. Connection to the computers is often done with a PCI based Host Bus Adapter (HBA) that connects via the fiber optic cabling to the FC fabric switches.



Final Cut Pro is a trademark of Apple Computer

A FIBRE CHANNEL FABRIC WITH CONNECTIONS TO COMPUTERS AND STORAGE

The SAN Problem: Operating Systems grab all disks

The problem with many computers to many storage networks is that the computers and their **Operation Systems** (**OS**) are used to assuming that all the storage the computer can access is fair game for the OS. This creates a multiple-writer situation in which a computer writing data to the storage is not aware of any other computer also writing data to the same storage, at the same time and at the same locations. This quickly results in corrupted data to the file systems.

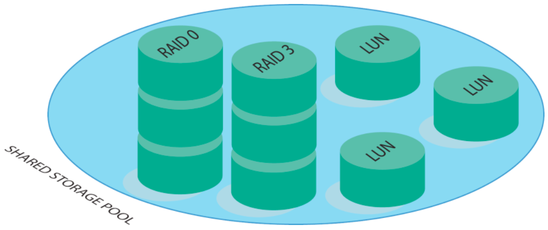
For SANs to work, something is needed to manage the file system traffic so that data is safe and doesn't get clobbered when everyone tries to write to the storage. That something is SAN management software that controls how the SAN uses shared storage connected computers.

How SANs benefit customers

For many environments, SANs allow significant gains in efficiency and cost savings as compared to storage directly attached to a single computer.

Utilization of resources — Storage and Server Consolidation

As computers utilize storage, SAN storage pools are ideal for changing demands for storage capacity by the various applications. It is much easier to allocate a piece of SAN storage to a computer through software verses the old method that would require upgrading the computer with more direct attached storage.



POOLS OF STORAGE ALLOW GREATER UTILIZATION

A SAN also addresses the problem in which some computers have unused storage that goes wasted as it is not being used. Because all the storage is in the SAN pool, utilization at all times is maximized. By consolidating storage resources into a dynamically reconfigurable, high-speed shared pool, the storage investment is optimized.

SAN Implementation Types

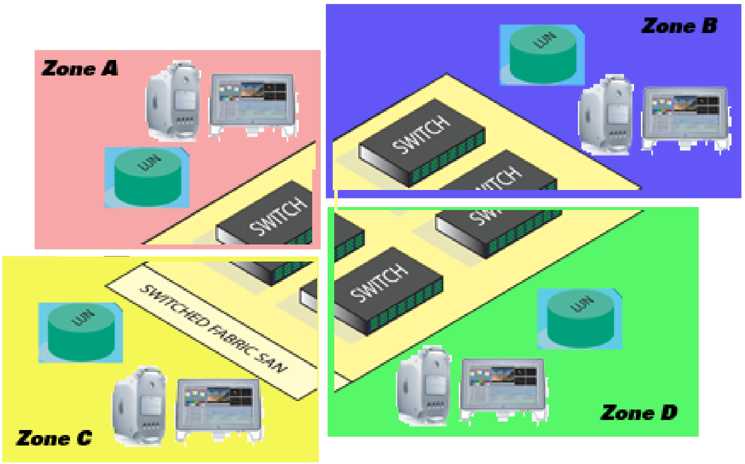
Shared Nothing - Storage Islands

Some systems approach the SAN problem by implementing a method called "shared-nothing". In this approach, the physical network connects everything, but access is restricted either by software, or hardware such as "switch" zoning, such that each computer would be allowed to see only that portion of storage it was granted access. The following summarizes disadvantages to this approach and why this approach is not appropriate in some situations:

- It defeats the ability to actually share the storage, at the same time, with others. If sharing data at the same time is one of the goals, then the remaining reasons are moot.
- Changes are not dynamic, and often require restarting one or more computers.
- It can be a management nightmare, requiring trained, full-time administration, to make changes.
- The granularity of control typically doesn't match real world use of storage, at the individual file system level, but is limited to an entire storage device, or storage device logical unit (LU) per access permission. This is not practical in today's world where a single LUN can easily be 1 **tera-byte** (**TB**) or larger.

So the main issue with this approach is that it defeats the main benefit of a workgroup SAN, which is the ability to actually share the storage, at the same time, with others. Another equally important issue with this approach is management, and flexibility. With this approach, in almost all circumstances today, changing the configuration, access permissions, or zones actually requires the editing station to do something before the changes will be seen. This inevitably involves restarting the computer to reflect the changes! Even worse, while the changes are being made, some computers involved might need to shutdown and thus stop working! Also, simply making the changes is complex and requires a trained administrator to manage the storage system, its allocations, and its permissions full-time. Lastly, these approaches greatly suffer because the granularity of control is often an entire storage device, and at best a single logical unit (LU) of the storage device, as opposed to an individual file system.

Nonetheless, some SANs that are designed for consolidation purposes utilize the storage pool and restrict access to pieces of storage to specific computers. Often done with switch zoning, this method prevents other computers from accessing any other storage that is outside its zone membership. This prevents the multiple-writer data corruption problem. In this environment the storage pool resources are available for allocation to many computers but nothing is actually shared at any one time.

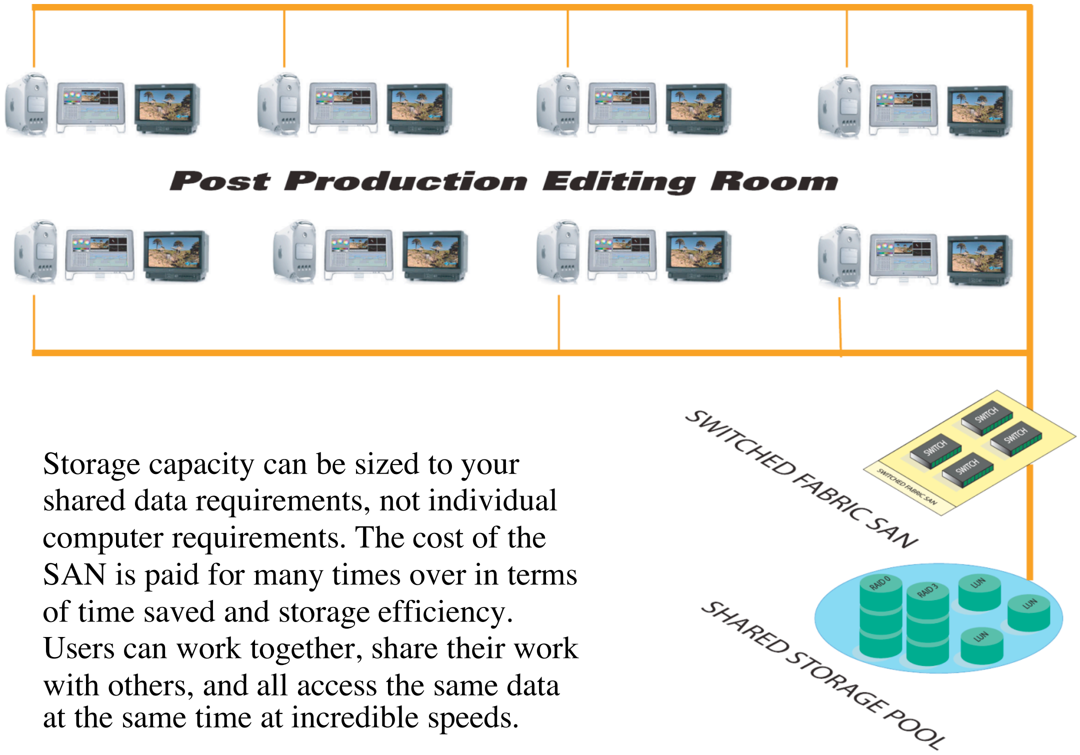


SHARED NOTHING — STORAGE ISLANDS COURTESY OF SWITCHED ZONING

As mentioned earlier, most switch based zoning available today is limited in granularity to an entire storage device (see the section later on called "Granularity of control — LUN structure background" for examples). This fact severely limits utilization choices as an entire storage device may be many TB in size, which is not always practical to assign to a single computer. Some expensive higher end switches allow zoning at the sub-device level known as **Logical Unit** (**LU**) zoning, but a LU is still a very large chunk for a single computer. The ANSI Small Computer System Interface (SCSI) storage protocol that storage devices utilize, allow for a sub-device addressing unit called **Logical Unit Number** (**LUN**). LUN addressing is mostly utilized by expensive storage devices known as **RAID** (**Redundant Array of Independent Disks**). A user configures the storage device and creates a number of LUs. This process is similar to partitioning a storage device using the OS, which is required before a file system can be placed on the device. Once a LU is created, it still needs to be partitioned in the OS, and is often broken down into multiple smaller pieces, each of which contain a separate file system. This is because a LU is usually limited to a minimum size, which is usually very large.

Shared Everything - Shared Storage

Another approach to the SAN problem and the idea behind Shared Storage is to allow a means for everything to be shared. What is needed is a method to transparently and dynamically reconfigure the storage networks at will without restarting any computers. Also the method would prevent multiple writers to a single file system at any given time while still allowing multiple computers to read the same data at the same time. Another desirable attribute would allow unrestricted multiple writers to the same file system at the same time without corrupting any storage.



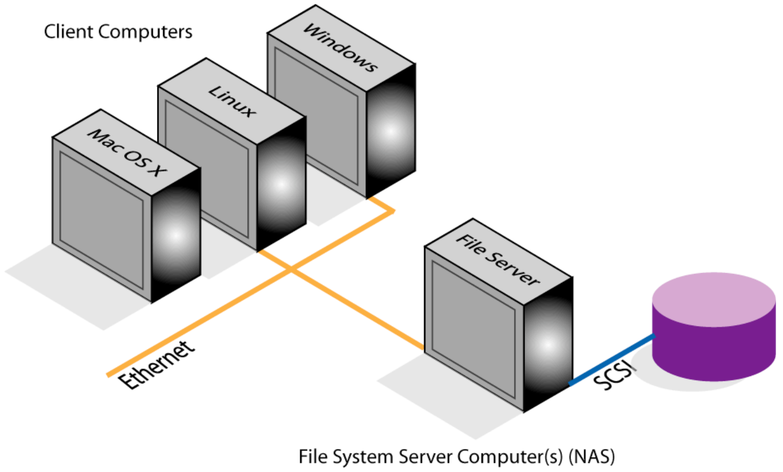
Storage capacity can be sized to your shared data requirements, not individual computer requirements. The cost of the SAN is paid for many times over in terms of time saved and storage efficiency. Users can work together, share their work with others, and all access the same data at the same time at incredible speeds.

Shared Everything — Providing the important option to work together

Client / Server Architectures — Multiple-writers to file system at a time

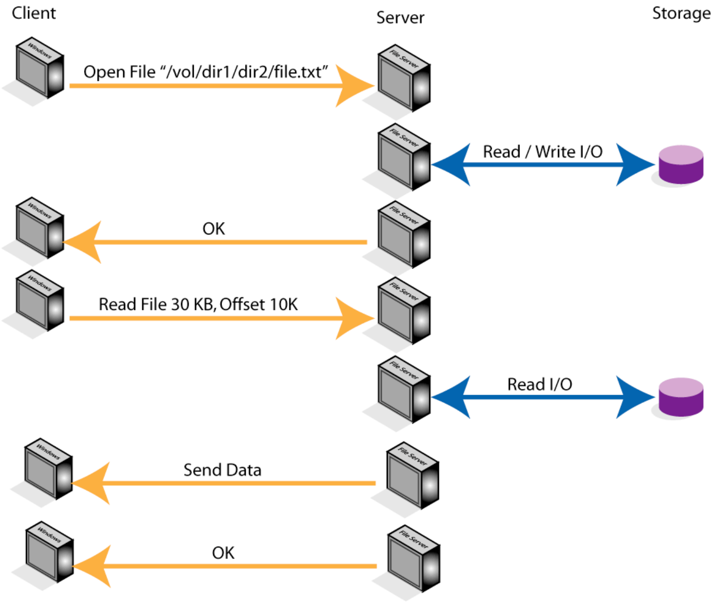
In a variation on traditional client / server file sharing, some SAN architectures utilize third party transfer which enables the client itself to directly transfer file system data to and from the SAN storage.

One must have a basic understanding of traditional **Network Attached Storage** (**NAS**) and how it fits into the client / server model before the SAN architecture modifications can be understood.



Traditional Client / Server network for file system sharing via NAS

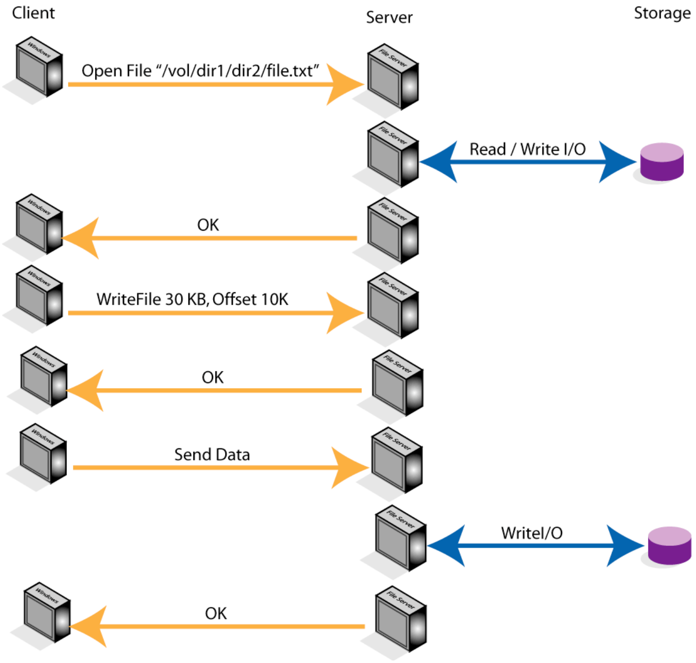
To understand the overhead involved with this type of network, and thus the SAN optimizations that can be done with it, we give some illustrated examples. Keep in mind that many clients can burden the Ethernet network and bottleneck the file server as requests are processed.


Client
Server
Storage
Open File "/vol/dir1/dir2/file.txt"
Read / Write I/O
OK
Read File 30 KB, Offset 10K
Read I/O
Send Data
OK

TRADITIONAL CLIENT / SERVER FILE OPEN AND READ OPERATION

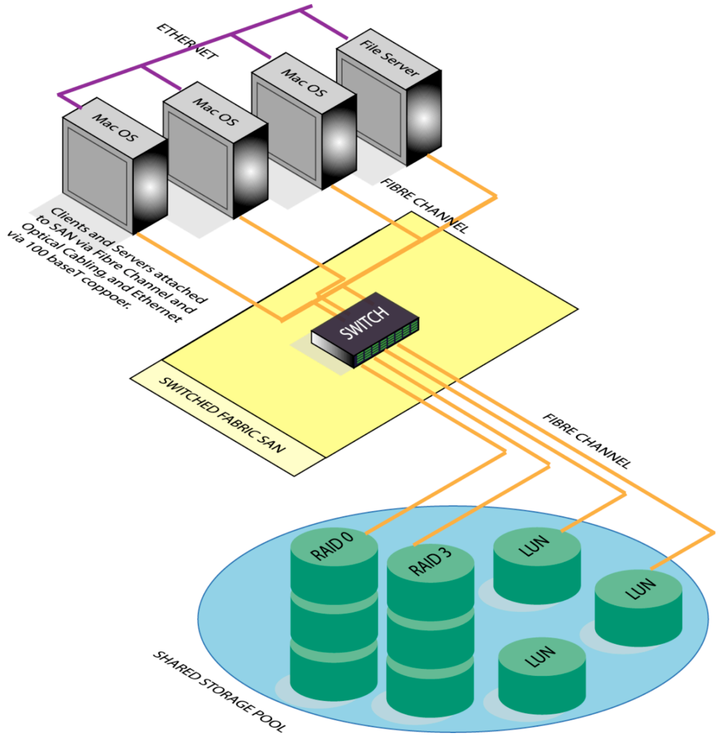
Actual operations involve even more traffic at the detail level, so these illustrations are a simplified generalization. They also do not reflect certain optimizations, such as caching or prefetching among others that might be applied to the problem.

These simplified illustrations make the point — to get or set any information about a file system, or read or write any data to the file system, the transaction must involve the client and the server. This overhead quickly becomes a big bottleneck for the server. Any operation such as obtaining a directory listing, obtaining information about the date and time a file was modified, opening, closing, reading and writing to a file etc…are examples of such transactions and take place over the Ethernet network.


Client
Server
Storage
Open File "/vol/dir1/dir2/file.txt"
Read / Write I/O
OK
WriteFile 30 KB, Offset 10K
OK
Send Data
WriteI/O
OK

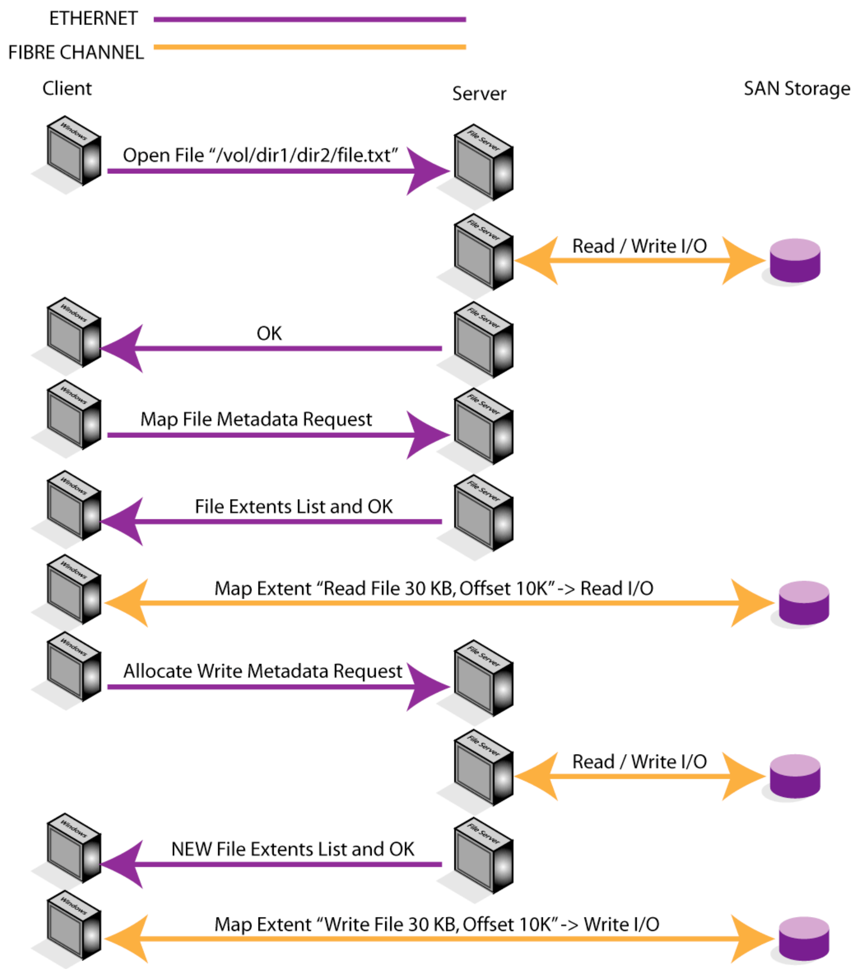
TRADITIONAL CLIENT / SERVER FILE OPEN AND WRITE OPERATION

Time-wise, a large burden in traditional client / server file system sharing is the multiple memory copies, or store and forward operations the server must perform in order to transfer data to and from the client. On the client to server side data must be packed or unpacked and prepared for sending over the Ethernet network. On the storage side memory must be prepared for a storage I/O operation. All in all it takes a lot of CPU overhead on the server side to process data associated with an I/O request. This area is where a common SAN optimization is applied since the client in a SAN can also directly access the storage.



A CLIENT / SERVER SAN ILLUSTRATING CONNECTIVITY TO STORAGE BY ALL COMPUTERS

These SAN architectures work by having a workstation serve metadata requests to clients over a standard LAN such as ethernet. In addition to all the standard file system LAN traffic, there is metadata traffic that includes the file system extents related to the requested I/O. An extent describes a contiguous portion of data in a file that can be used to map to a storage I/O request to the physical storage. For example, a read request would return metadata containing the sectors on the SAN for which the read request maps. The client would then issue the read I/O's directly on the SAN to actually transfer the data. Other metadata requests include open, close, get/set date, get/set attributes, delete, etc… operations. All these operate in the standard client/server file model, such as **NFS** (**Network File System**) or **CIFS/SMB** (**Common Internet File System / Server Message Block**) protocols, with the addition of the extent information communicated for reads and writes. For more information, refer to the United States Patent #6,044,367 — Distributed I/O Store.


ETHERNET
FIBRE CHANNEL
Client
Server
SAN Storage
Open File "/vol/dir1/dir2/file.txt"
Read / Write I/O
OK
Map File Metadata Request
File Extents List and OK
Map Extent "Read File 30 KB, Offset 10K" -> Read I/O
Allocate Write Metadata Request
Read / Write I/O
NEW File Extents List and OK
Map Extent "Write File 30 KB, Offset 10K" -> Write I/O

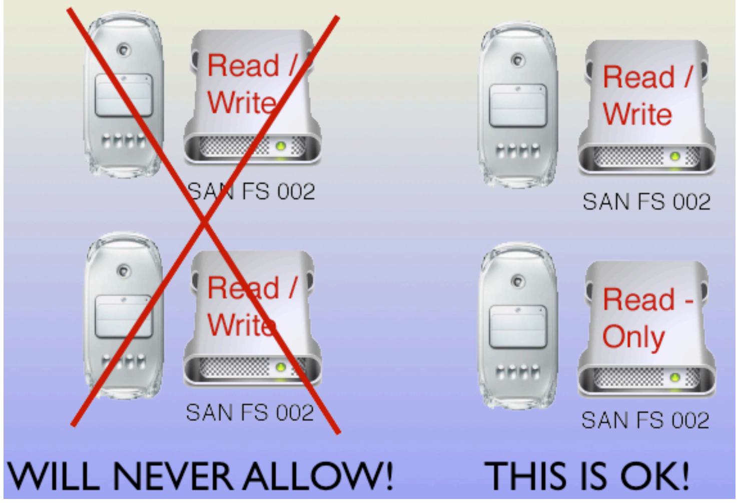
SAN OPTIMIZED CLIENT / SERVER FILE OPEN, READ AND WRITE OPERATION

Bottom line, these SAN architectures suffer from several high availability and scalability issues. Because they require a "metadata server" to serve out the metadata for a file system, it becomes a potential single point of failure, and a point for bottlenecking. For the failure aspect, if anything happens to the metadata server then all access to that file system is lost. For the bottlenecking issue, since a single coherent metadata server is responsible for all I/O and other metadata requests for all clients of the file system(s) that it is responsible for serving, and this data travels over a slower LAN network, the metadata server suffers from scalability issues similar to traditional client/server architectures for file servers.

In practice, with NFS metadata traffic, we have found this overhead of metadata over the LAN to quickly reach a scalability limit on small networks for clients involved in high-speed high-bandwidth I/O's (video/film) or low-bandwidth I/O but many requests (many audio tracks) such as used in Post Production environments. We have also found, in particular, that the NFS and CIFS architectures, which some of these SAN architectures use as a basis, have inherent I/O limitations related to maximum blocks of data per I/O that have serious implications to Post Production environments.

Serverless Architectures — Single-writer to file system (or LUN in some cases) at a time

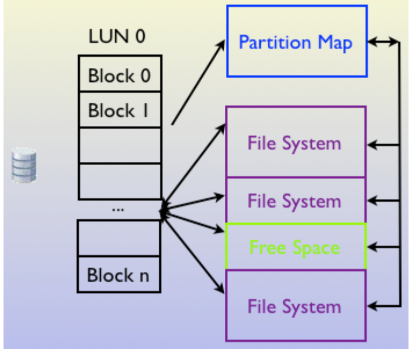
For environments that can be designed with workflow practices allowing new content to be created in a single-writer to a file system (or LUN) at a time fashion this architecture has several advantages for users. First, there is no single point of failure. Second there are no scalability or bottlenecking issues. The reason for this is because this architecture does not involve any server whatsoever.



SERVERLESS SAN — SINGLE WRITER TO A FILE SYSTEM AT A TIME

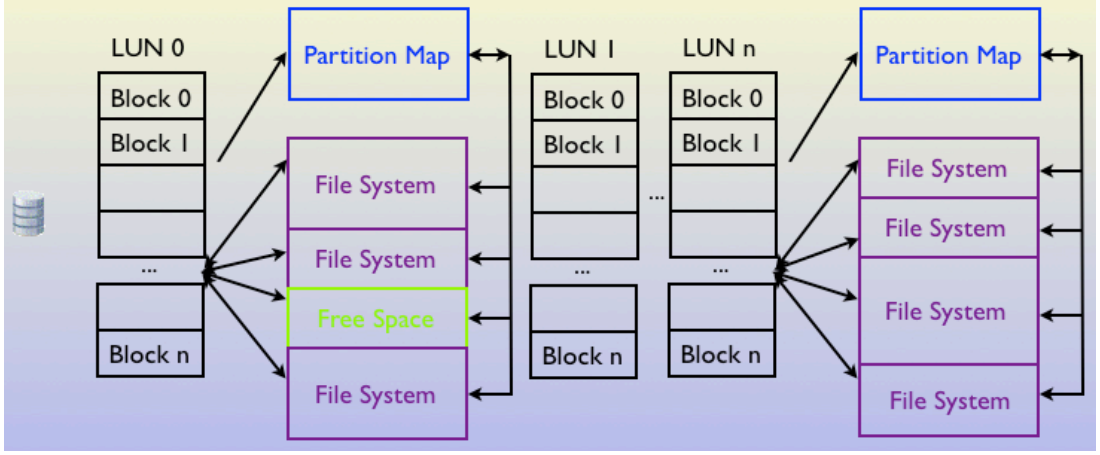
Granularity of control – LUN structure background

Beware however that not all of these serverless SAN architectures are equal. In fact some have a serious limitation in the area of granularity of control. Some only allow control at a whole LUN level, while others allow control at a file system level (potentially sub-LUN, or partitions of a LUN, or spanning part or whole multiple LUNs, as would be with a RAID-0 set).



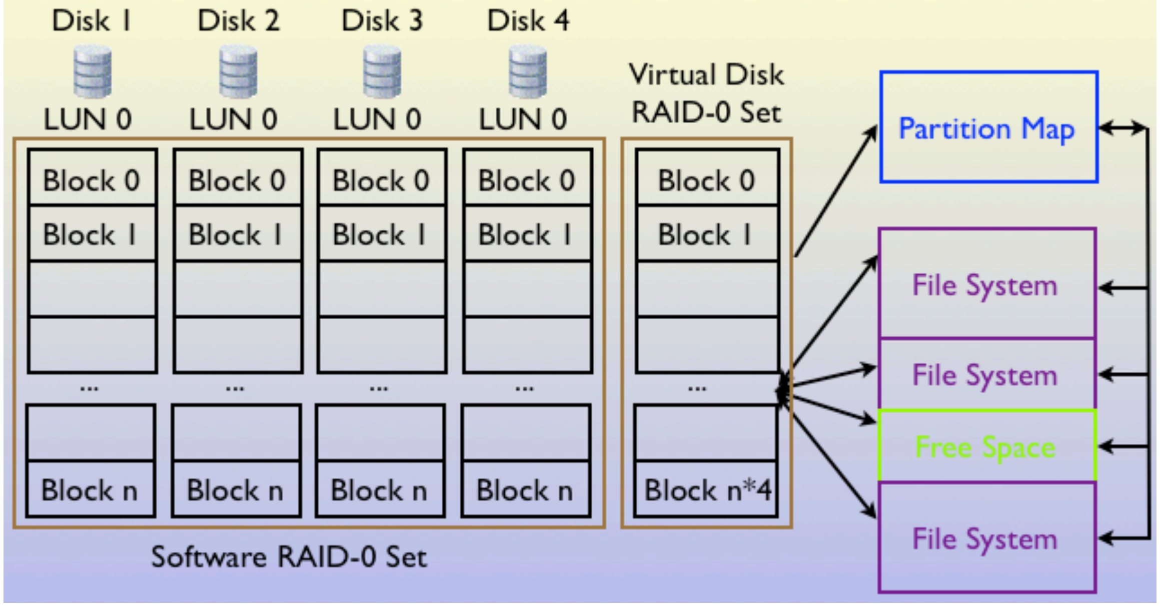
A STORAGE DEVICE WITH SINGLE LUN AND ITS PARTITION MAP AND FILE SYSTEMS

A single disk with a single LUN will have all storage addressable by block numbers 0 through n. At the beginning of the disk the OS records a partition map that describes the contents of the entire LUN, including its partition map (which contains entries for all areas of the LUN), file systems, and unused space.



A STORAGE DEVICE WITH MULTIPLE LUNS, EACH WITH ITS PARTITION MAP AND FILE SYSTEMS

A single disk with multiples LUNs may have each LUN contain a different number of total blocks. Each LUN itself however contains the same expected information describing its content.



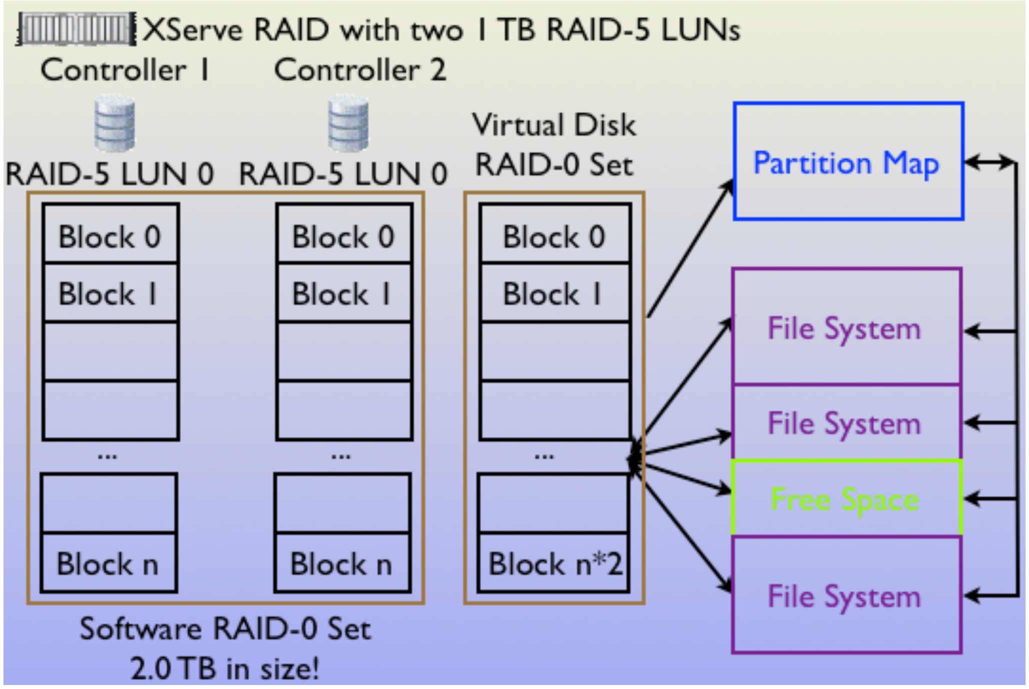
MULTIPLE STORAGE DEVICES STRIPED ACROSS THEIR LUNS TO CREATE A RAID-0 SET

RAID-0 is an example that uses multiple LUNs striped together to, in essence, create a larger virtual LUN. In the end, however, it also needs the information to describe its content such as partition maps, file systems, free space, in addition to information that describes the RAID-0 set and its structure itself. In this case it is interesting to note that a file system in this case actually has pieces of it spread across multiple physical disks verses the single LUN case in which it only occupies a portion of one LUN. In RAID-0 there is a striped segment size that dictates the maximum size of a contiguous chunk of data that can be written to a single LUN before moving to the next LUN to allocate the remaining data. If the chunk size was 256 KB, and 1024 KB was written to the RAID-0 set on a chunk size boundary, then 4 I/O's one each to the 4 LUNs would be issued in parallel to accomplish the write request. RAID-0 is known for high-performance because of the parallel nature utilizing all disk spindles simultaneously to accomplish I/O. Beyond the scope of this paper, there exist different types of sets as well including generalized virtual LUNs, spanning virtual LUNs, and other RAIDs that all use multiple LUNs in some fashion to map a file system.

Granularity of control – Why file system is better than LUN

Now that we have covered some background around LUN usage and how file systems can exist on single LUNs, parts of LUNs or across many LUNs, we can move on to the point about granularity of control and SAN software.

Because users interface with storage at the file system level, this is the preferred granularity of control. Many problems arise with granularity at the entire LUN level due to the fact that it is unclear how this maps to actual file systems a user wishes to access, as the prior examples illustrate the virtually limitless options.



WHY FILE SYSTEM GRANULARITY IS SUPERIOR TO LUN GRANULARITY

Consider a 2 TeraByte (TB) XRAID. The minimum LUN you can create in this configuration is 100 GB. This device has two controllers, each controlling half of the disks. A user might create 1 RAID-5 LUN on each controller, each 1 TB in size. The user might then software stripe (RAID-0) between the two controllers, and then partition the storage into 100 file systems. This both maximizes and load balances all available bandwidth with I/O's that are issued to the XRAID while maximizing the usability by allowing many file systems for use by many different users on the SAN. Given this scenario, granularity of control at the LUN level would be impractical because a user requiring write access to the storage would need to acquire exclusive write access to both 1 TB LUNs and thus all 100 file systems, even if that user only need to write a file to 1 of the 100 file systems on the storage! Given this scenario with file system level granularity, 100 different users could be writing to the storage at the same time with each having write access to 1 of the 100 file systems! File system granularity is clearly better.

A Word About Unix File System Permissions

Before SANs were available, users wishing to share large amounts of data on disks used a method commonly referred to as Sneaker-Net. This is because removable disk drives were hauled from one room to another for use on different computers. Computers, including Unix computers like Mac OS X have no problem with this method of data sharing.

A SAN can be thought of essentially as Sneaker-Net save for the physical act of hauling a removable disk from one room to another being replaced by a SAN software command to instantly reassign storage to another computer.

Mac OS X allows multiple users per computer. Each of these users has access to their own files. Each file is associated with an owner, group, and other permission settings that dictate who and how the files may be accessed (read, write or none). Users and groups are standard, and numbered the same on a machine. For example, when you install the OS, the first user created is always numbered 501. This 501 number is what is recorded on the disk. Therefore, if another computer accesses this storage (thinking it is directly attached), the first user on the computer (501) would appear to be the owner. Also when the OS is installed some standard users and groups are also created, such as system, wheel, root, etc. These also have the same numbers across computers.

In a shared storage SAN users typically want to share storage. If they don't, the administrator can configure the SAN so that only particular computers access particular storage.

Because a file system's numbers for users and groups will be translated to the local meaning on a computer, a SAN user and administrator simply needs to be aware of this fact. Under Mac OS X, the user has the option whether or not permissions are to be obeyed by the OS on a per disk basis.

If an Administrator wants to still use permissions in a shared-storage SAN, and is not comfortable with the local translations, then he can set up each SAN computer to have the same users, and groups, created in the same order, so that they have the same numbers, to enforce a SAN-global mapping of permissions settings.

Normally this is not at all necessary as everyone in a shared-storage SAN that has been given access to the shared-storage by the Administrator will typically want to share their creations. If not, they have the option of password protecting and encrypting their data, or implementing a SAN-global mapping of permissions. Normally the user wants others to read and write to the files they create, so the permissions for “others” are set to read/write. A simple SAN-global group is an easy method for ensuring read/write access as well. By default however, the first user, 501, will appear as the owner on each local machine, so as long as the owner has read/write access everything will be able to be shared with ease.

Mac OS X SAN Product Comparison

The following information was gathered from a variety of sources including, in some cases, direct experience with testing the various products, information from talking to users, and information obtained from the websites of various products.

CommandSoft®'s FibreJet®

FibreJet

Architecture

FibreJet is based on shared storage technology utilizing a serverless SAN architecture. FibreJet is available cross-platform with Windows XP or later and Mac OS 8.1 or later and all versions of Mac OS X. Persistent network state is recorded in a SAN located database that all SAN attached computers should have access. Because the database only records the persistent state, if it is corrupted or lost, users will continue to operate with the file systems in their last state until the SAN database is either repaired, replaced, or restored. More detail on this fact below.

Installation and Setup

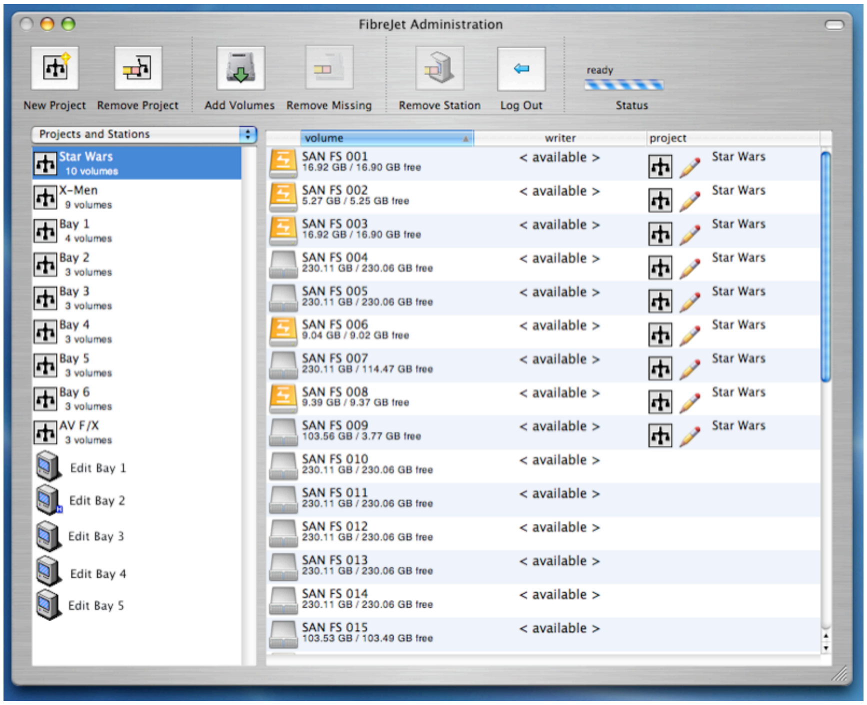
Installation is from CD using standard Apple installer technology and a single installer file. Each SAN attached computer should have the software installed. The software is protected with a USB hardware key that must be attached to each computer using the software.

FibreJet is available as a fully functional application demonstration for a nominal fee that is refundable upon purchase. The nominal fee covers the expense of the hardware key that is programmed for demonstration mode to expire after a given number of times FibreJet is run. The number of activations can be reset given a code that is typed in the software. Upon purchase, a code can be entered from FibreJet to reprogram the demonstration hardware key into a normal hardware key. Software updates are available via the Internet and users can be entered into a mailing list informing when updates are available.

There are no special storage or storage software requirements as FibreJet utilizes standard utilities like Apple's Disk Utility or ATTO's ExpressStripe software for partitioning the storage. Therefore, no proprietary disk utility software is required when installing a FibreJet SAN.

The SAN database will be created and located on a file system that the user names FibreJet. When the software starts and doesn't find a SAN database, it will next look for a file system named FibreJet. If it finds one it will allow the user the option to create the SAN database using this file system partition. Once the user agrees to this, the partition is erased and converted from a file system into a cross-platform SAN database partition and the initial setup is complete.

From this point the user may enter administration mode, set passwords, create projects, and restrict access as appropriate.



Star Wars
10 volumes
X-Men
9 volumes
Bay 1
4 volumes
Bay 2
3 volumes
Bay 3
3 volumes
Bay 4
3 volumes
Bay 5
3 volumes
Bay 6
3 volumes
AV F/X
3 volumes
Edit Bay 1
Edit Bay 2
Edit Bay 3
Edit Bay 4
Edit Bay 5

volume	writer	project
SAN FS 001 16.92 GB / 16.90 GB free	< available >	Star Wars
SAN FS 002 5.27 GB / 5.25 GB free	< available >	Star Wars
SAN FS 003 16.92 GB / 16.90 GB free	< available >	Star Wars
SAN FS 004 230.11 GB / 230.06 GB free	< available >	Star Wars
SAN FS 005 230.11 GB / 230.06 GB free	< available >	Star Wars
SAN FS 006 9.04 GB / 9.02 GB free	< available >	Star Wars
SAN FS 007 230.11 GB / 114.47 GB free	< available >	Star Wars
SAN FS 008 9.39 GB / 9.37 GB free	< available >	Star Wars
SAN FS 009 103.56 GB / 3.77 GB free	< available >	Star Wars
SAN FS 010 230.11 GB / 230.06 GB free	< available >	
SAN FS 011 230.11 GB / 230.06 GB free	< available >	
SAN FS 012 230.11 GB / 230.06 GB free	< available >	
SAN FS 013 230.11 GB / 230.06 GB free	< available >	
SAN FS 014 230.11 GB / 230.06 GB free	< available >	
SAN FS 015 103.53 GB / 103.49 GB free	< available >	

ADMINISTRATION MODE OF FIBREJET SHOWING PROJECTS, STATIONS AND FILE SYSTEMS

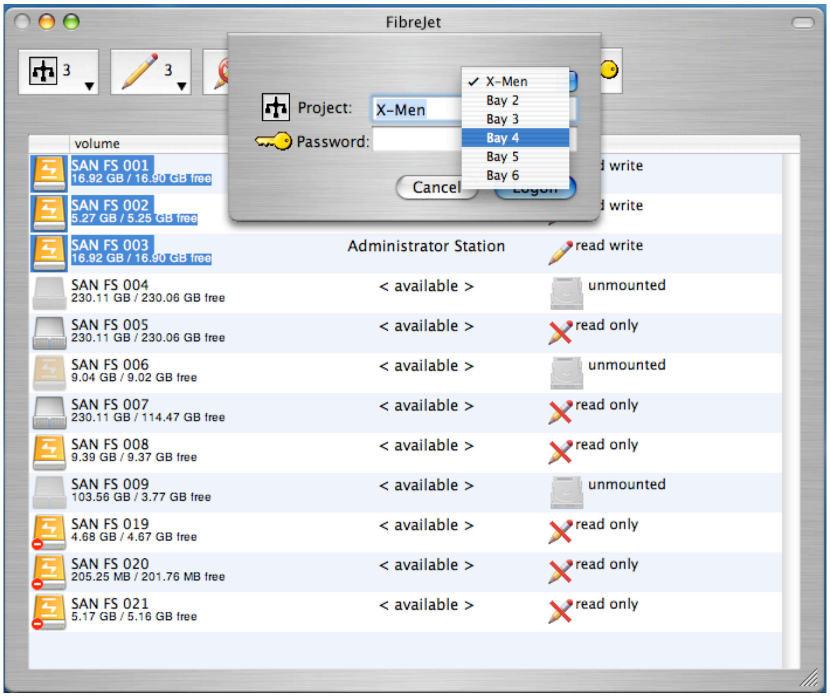
From administration mode the user can perform Disk maintenance, arrange projects (which are named groups of file systems), set access modes and configure passwords using a simple interface. The interface supports drag and drop functionality, as well as extensive contextual menus, so there are always multiple choices for the user to accomplish the same thing.

Usability model

FibreJet inherits the user account security of the user that has logged into the OS, thus leveraging the security model that is setup and enforced by Mac OS and the IT administrator. The computer itself is recognized and recorded automatically in the SAN database before operations are allowed.

Once the user logs into the computer with their account and password, they then run FibreJet (which can be set to auto-launch in the users startup items). From here what happens depends on how the Administrator configured the system.

If the Administrator configured FibreJet for open access, then the user might automatically see all the disks they have access to (or last had access to) and it might also mount them in the last state so the user can get right to work and completely ignore the fact that FibreJet even exists on the system.



Project: X-Men

Password:

volume		
SAN FS 001 16.92 GB / 16.90 GB free		write
SAN FS 002 5.27 GB / 5.25 GB free		write
SAN FS 003 16.92 GB / 16.90 GB free	Administrator Station	read write
SAN FS 004 230.11 GB / 230.06 GB free	< available >	unmounted
SAN FS 005 230.11 GB / 230.06 GB free	< available >	read only
SAN FS 006 9.04 GB / 9.02 GB free	< available >	unmounted
SAN FS 007 230.11 GB / 114.47 GB free	< available >	read only
SAN FS 008 9.39 GB / 9.37 GB free	< available >	read only
SAN FS 009 103.56 GB / 3.77 GB free	< available >	unmounted
SAN FS 019 4.68 GB / 4.67 GB free	< available >	read only
SAN FS 020 205.25 MB / 201.76 MB free	< available >	read only
SAN FS 021 5.17 GB / 5.16 GB free	< available >	read only

STRAIGHT FORWARD PROJECT BASED SECURITY USER ACCESS

On the other hand, if the Administrator protected the SAN with various projects and / or passwords then the user will be required to log into the various projects they have been given the passwords to in order to begin their work and see the file systems they are allowed access.

FibreJet allows the user to mount disks in read/write or read-only mode depending on access privileges set by the Administrator. FibreJet allows the user to unmount volumes. FibreJet has a read-only project feature that is useful for libraries and other collections that should not be modified, but are good to share among users.

FibreJet supports dynamic reconfiguration of the SAN storage pool, including requesting write access from another user. These write requests can be accompanied by a text message describing the need. The user receiving a request for access may grant or deny the request, and type a response message giving any explanation.

Interesting Aspects

FibreJet is the only software that supports media sharing with AVID® and ProTools® users. This is done with a special AVID® read-only mode that only FibreJet has that is used to avoid the media going offline. Media going offline can easily happen otherwise, when accessing new content. Without this feature, the user at worst needs to restart their computer, and at best needs to restart the AVID® software, before the new media can be recognized. With this feature the user can access the new content without quitting their AVID® application!

FibreJet's overall look and feel is Apple® Finder® like (icon based drag and drop, contextual menus, key board and menu shortcuts, dock menus, etc.) so the user does not have to learn a new way to work with file systems other than what they already know.

FibreJet's write request / response mechanism includes a request and response message that users exchange giving additional explanation.

At startup, FibreJet detects and warns the user if multiple SAN databases have been discovered.

The FibreJet SAN database has been designed so that the size can be configured giving the ability to have virtually unlimited scalability in number of volumes, projects, and computers.

FibreJet includes a command to Backup and Restore the SAN configuration as a means to protect against the SAN database being lost.

FibreJet supports dynamic storage, including the ability to dynamically recognize when new storage comes online.

FibreJet allows the user to display or hide missing file systems it expected to find based on prior usage. If storage dynamically appears to the OS containing any of these file systems, then they will become available for normal use dynamically.

FibreJet prevents rogue hosts (those without any SAN software installed) from mounting the standard volumes. This prevents inadvertent and unauthorized access to storage if a computer is attached to the SAN that does not contain the FibreJet software.

FibreJet contains an event history logging feature that can be used to trace what and when operations were done in the application.

FibreJet display's the OS reported icon for the file system, including its size and available free space.

FibreJet allows flexible sorting of file system information by name, size, status, available free space and owner.

FibreJet by default doesn't allow force unmounting. Force unmounting is the ability to unmount a volume despite it being used by some piece of software.

The Administrator may enable the option which would allow users to choose to force unmount.

FibreJet allows configuration of individual timers (including enabling, disabling, firing, and setting how many seconds between running) for checking the SAN database for pending requests, and for auto-updating file systems (which detects when a file system has actually been changed by the user with write access so it will update the changes for the read-only users to see).

FibreJet includes an option to automatically suspend FibreJet when it is not the foreground application. This limits all impact to the CPU in terms of any I/O to the SAN database for example, which otherwise might possibly to interfere when certain real-time storage applications that press the fringe of system performance.

FibreJet is not limited to only HFS+ volumes.

FibreJet does not require workstations to be connected to a LAN.

FibreJet filters SCSI START-STOP commands from the SAN preventing the Fibre Channel disks from being spun down. This feature is available only when using the CommandSoft QLogic Fibre Channel HBA driver.

FibreJet has the ability to "unclaim" disks it has claimed, if needed.

FibreJet has the ability to force Mac OS X to rediscover all storage partitions which is crucial for 10.3.x environments with over 100 partitions.

FibreJet supports Dock menu functionality so that the user doesn't need to switch into the FibreJet application to perform many common operations.

FibreJet has a safe-mode that it goes into when adverse database conditions are detected so that work can continue uninterrupted.

General Support List

FibreJet has been tested with HBAs from QLogic, ATTO, Apple and Astera.

FibreJet has been tested with switches from Brocade, QLogic and Emulex.

FibreJet supports 1 and 2 GB/s Fibre Channel

The FibreJet product includes a driver written by CommandSoft for QLogic cards.

FibreJet is also available for Mac OS 8.1 through 9.22 and is cross compatible with Mac OS X version.

FibreJet has been tested with Mac OS 10.2.2 through 10.3.3

FibreJet works with just about any SAN storage

FibreJet is available for Firewire based SANs

More about claiming, unclaiming, and adverse database conditions

FibreJet does not loose any SAN file systems, even file systems located on the same disk/LUN as the database, as a result of a corrupted or lost FibreJet database. Nor does it lose any data as a result of its file system claiming process, because at worse, this process can be simply reversed at any time.

Depending on the environment, FibreJet may "claim" the file system partitions. The purpose of this "claiming" process is to prevent rouge host issues in which another SAN-attached machine, without any SAN software, might otherwise recognize and mount file systems with write access, thus causing multi-writer corruption. With the file systems claimed, a normal computer doesn't recognize them as something it is able to mount. If there is ever a need, FibreJet allows an Administrator to also do the reverse, which is to "unclaim" a file system or all the file systems, as needed. This puts them back in their original condition that would allow any SAN-attached machine to mount them. One worst case scenario with respect to "database corruption" may make the ability to "unclaim" a file system very important, as described later.

With FibreJet, if communication to the FibreJet database is lost (the ultimate corruption in a sense), FibreJet continues to operate in a safe mode in which the user continues to have access to the file systems they were using. Additionally, FibreJet allows a user other safe operations to known file systems such as mounting them read-only (if they were authorized of course). When FibreJet detects this condition, it will continue to warn the user of the loss of communication and its operation in safe-mode, as each operation is attempted.

If, instead of complete loss of communication to the database, the database gets corrupted, FibreJet is built to handle database errors, and will continue to operate as best it can, and should at worse, go into its safe mode.

There are many options for remedying the situation when one of the above (loss of communication or partial or complete corruption) occurs. First, know that any mission-critical systems that are in the SAN operating and relying on file system communication will continue to operate in safe-mode so applications will continue to fully function. Users just doing their work will continue to do so, however limited to safe-mode. An Administrator, from any machine, can address the issue of getting the database back online fully.

If the corrected database is going to be in the same exact storage location, there are several commands in Administration mode for quickly getting a fully functioning database. First is a Zero Database command that will recreate an empty database in its present storage location. The Administrator also has a Save and Restore configuration command that allows for the backup and restore of important database information like projects and volume membership. Administrators are encouraged to regularly use these commands to backup their database.

Most SANs are small, and even if a user didn't use the Save configuration command to backup the database, it is a simple matter of recreating the various projects in a modest sized SAN (the hosts and volumes automatically get recreated in the database upon discovery). For background purposes it should be noted that the main purpose of the database is ensuring that two users do not obtain write access to the same volume, and also enforcement of storage security for authorized access. The database is also used to record persistent network state and exchange messages and requests between users.

Because Mac OS X largely learns about partition and storage device layout when it starts up, it will still remember where the database is located since it hasn't moved in this case. Each SAN machine will simply have to Suspend FibreJet, quit FibreJet and relaunch FibreJet once the Administrator has done the above. Then the users can continue to use FibreJet in its normal mode, rather than safe-mode.

If the problem were the more severe case, in which the physical device where the database is located is no longer available (it was destroyed for instance), will mean that the Administrator needs to create a new database in a different location. Then, once this is accomplished, the users will need to take the additional step of either restarting their computer so that Mac OS X rediscovers the storage layout and finds the new location of the database, or use FibreJet's "Rediscover SAN Network" command which forces Mac OS X to rediscover all the storage without restarting the computer. Then quitting and running FibreJet will find the newly created database allowing the user to continue in its normal mode.

It should be clear that creating a database in a new location is a simple matter. FibreJet simply requires you to name a normal file system "FibreJet", and if no database is found, it gives you to option to destroy that file system and use it as the FibreJet database. It is recommend that when initially partitioning the SAN storage the user should set aside a small partition just for this purpose to save someone the step of finding an appropriate place later or trying to create one later (no one has had to use it yet by the way). If a user didn't do this ahead of time, they would then have to find a piece of storage (about 10 MB in size minimum) that could be partitioned or repartitioned for this purpose.

Lastly, if there were some sort of emergency in which a database was lost or corrupted, no one was already in FibreJet for it to operate in "safe" mode, and you needed access to all the SAN storage so you could do work or begin to find a place for a new database, then FibreJet allows you to "disassemble" all SAN file systems, or just individual file systems, back into their original state. This is the all important "unclaiming" part we described above. At this time the user would be able to see and use all the storage from a single workstation safely and then recreate the database from a set aside partition, or if needed, repartition. This process is very quick and if a user found the SAN in this state (like coming in the morning with no one online, but the database storage was lost over night) then the user can get it all back up in as little as 5 minutes.

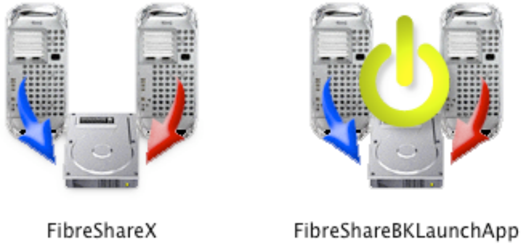
This should clarifies why FibreJet will not find itself in the same state as other products mentioned in this document and why with FibreJet, a user will never find themself not being able to preserve all thier file systems in operational state. For these reasons, FibreJet will not be the cause of why one would lose any of their file systems. Also, because FibreJet is “always-on” even when things go bad, users continue to get their work done.

Apple Computer's XSAN

Reference

Please download from http://www.commandsoft.com/content_files/comp/FibreJet_and_Xsan_2005.01.pdf for a complete comparison of XSAN to FibreJet.

Charismac Engineering, Inc.'s FibreShare



Architecture

FibreShare is based on shared storage technology utilizing a serverless SAN architecture. Persistent network state is recorded in a randomly located SAN database not of the users choosing (may be distributed/replicated across all SAN storage). All SAN attached computers should have access to this SAN database, which can become a challenge, since the user does not know on which SAN disk it is located. Because the database only records the persistent state of the file systems, if it is corrupted or lost, users will continue to operate with the file systems in their last state, however, repairing the problem can prove difficult for several reasons.

First, not knowing where the SAN database is located is a challenge. Secondly, if it is damaged or lost, the software just starts using another random SAN disk as the database. Third, at this point all or most passwords, users, groups, file system privileges and administrator settings are lost. From our tests it did not appear like the database was consistently distributed/replicated across all the SAN storage (as has been reported), as when we removed the storage we had identified with the database access, things quickly deteriorated with the 2.0.4 version we tried. Problems especially occurred when altering the storage configuration by adding and removing drives, and combining two FibreShare SANs.

FibreShare has a zero database command that can be used to start the SAN database from scratch when there is a problem. There is also an export and import administrator database command to preserve the administrators, users and groups. The problem here is that this feature does not preserve any privilege settings of the file systems so all will need to be reconfigured.

This can be a daunting task because each file system needs each user and/or group explicitly assigned, which amount of work is regulated by the following formula:

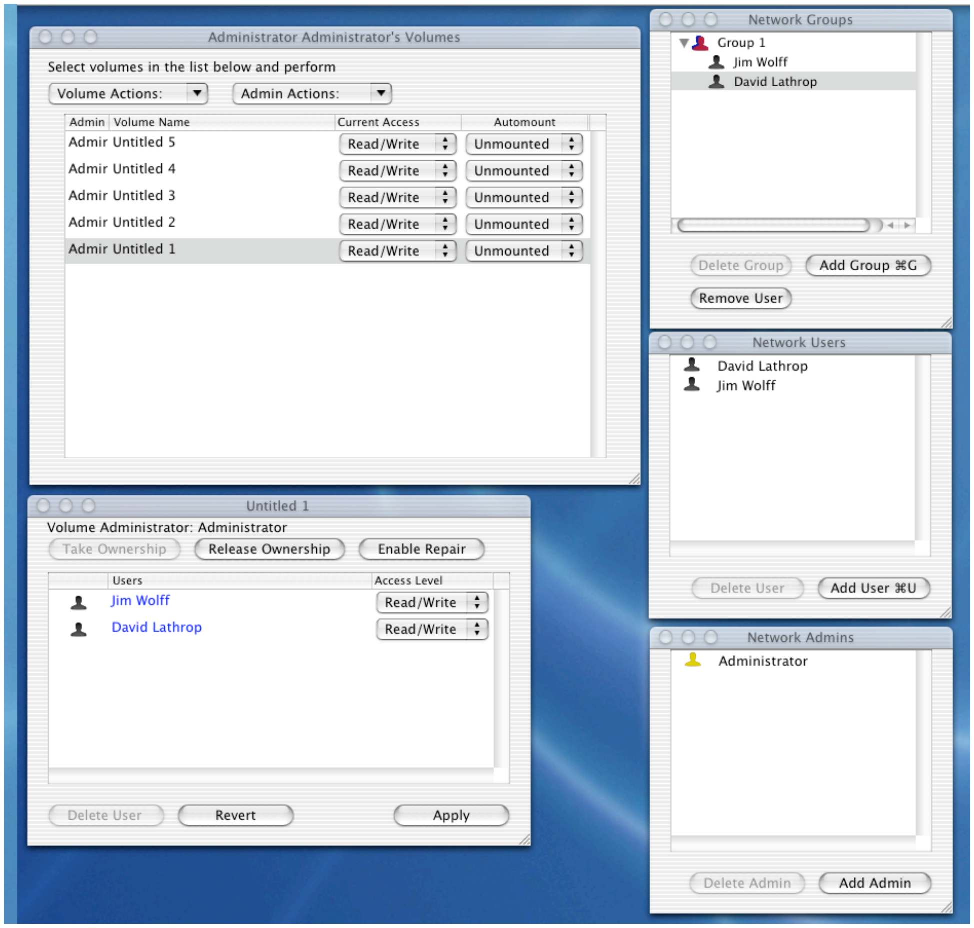
$a=(g+u)*f$

Here, a = total assignments in the GUI by the administrator, g = total number of groups of users, u = total users which are not already in appropriate groups, and f = total number of file systems. When calculated it is easy to see that quite a lot of work is required of the Administrator. Given 130 volumes and 13 users (not uncommon in a medium size SAN) this results in at best 130 individual GUI assignments (each in a separate window that must be brought up) that are required to reconfigure the file systems. This example is an easier case where all 13 users are in 1 group and not repeated for the assignments (g=1, u=0, f=130)! The same SAN in another arrangement could easily results in many hundreds of assignments to initially configure or later reconfigure the SAN.

Installation and Setup

Installation is from CD using standard Apple installer technology and a single installer file. Each SAN computer should have the software installed. The software is protected with a USB hardware key that must be attached to each computer using the software.

FibreShare requires all SAN storage to be set up using Charismac's proprietary Anubis partitioning utility. FibreShare will not work with standard disk partitioning utilities, thus requiring existing storage to be reconfigured. Only other CharisMac products will work in conjunction with FibreShare.



Select volumes in the list below and perform

Admin	Volume Name	Current Access	Automount
Admir	Untitled 5	Read/Write	Unmounted
Admir	Untitled 4	Read/Write	Unmounted
Admir	Untitled 3	Read/Write	Unmounted
Admir	Untitled 2	Read/Write	Unmounted
Admir	Untitled 1	Read/Write	Unmounted

Volume Administrator: Administrator

Users	Access Level
Jim Wolff	Read/Write
David Lathrop	Read/Write

Group 1
- Jim Wolff
- David Lathrop

David Lathrop
Jim Wolff

Administrator

FIBRESHARE ADMINISTRATION WINDOWS

The SAN database is created on a random location among the SAN disks. In actuality it seems that each SAN disk has the potential to store the SAN database as room for it is reserved for each disk that is setup using the Anibus utility. Because the SAN database could potentially be anywhere, you must be careful in changing anything about the SAN disks, including repartitioning or repurposing any disk as it might contain critical information for the operation of FibreShare, so it must first be "released" from the FibreShare system.

Usability model

FibreShare implements a separate security model for the SAN, including its own users, administrators, groups and privileges, thus requiring the user to log into the FibreShare system before any storage can be accessed. This security system is separate and apart from the users, administrators, groups and privileges that may have been configured on the Mac OS X network by the IT administrator. Be aware therefore that this security model is added work for the IT administrator as it creates another separate and unrelated system that needs to be managed.

Once the user logs into the FibreShare system, they will see the entire list of volumes that they can potentially access, according to how the users and groups are assigned privileges to the file systems at that time. This could be overwhelming in even a moderately sized SAN with 130 file systems for example, if that user was assigned the ability to access all file systems at that time. The only way to change this configuration is through work the Administrator must perform by reconfiguring file system privileges using the a=(g+u)*f formula discussed earlier. A single user in a ProTools environment for example, usually works with fewer than 30 file systems at a time.

FibreShare allows the user to mount disks in read/write or read-only mode depending on the access privileges set by the Administrator. FibreShare allows the user to unmount volumes. FibreShare can restrict users to only mounting volumes in a read-only state that is useful for libraries and other collections that should not be modified, but are good to share among users.

FibreShare supports dynamic reconfiguration of the SAN storage pool, including requesting write access from another user. These write requests can be granted or denied, without explanation.

Interesting Aspects

FibreShare requires storage to be setup with Charismac's proprietary Anubis formatting utility with the special option to create a FibreShare volume.

FibreShare will not work with existing data or storage formatted in any other way so it must be backed up before you can begin.

FibreShare's SAN Database information, including access states of the volumes, and any group, user and administration information is stored in a special partition set aside when the storage is partitioned with the Anubis utility.

FibreShare passwords for users and administrators are stored in clear-text for anyone to see who knows how to use any standard utility to look at the data on the disk. This completely renders the security model moot as anyone may access any of the SAN storage in any state by obtaining these passwords. On a positive note, this feature can come in handy if you forget your password since it is so easy to find.

FibreShare has SAN database problems for users with changing SANs. For SANs with active storage changes taking place (new storage, combining SANs, removing storage, migrating data etc…), it is easy to create a situation where FibreShare encounters multiple locations where administration information is stored about the users, administrators, passwords and groups. When this occurs, it simply picks the first one it finds and begins to use this, without any warning to the users. This can result in unexpected configurations as configured information may not be used anymore or may have gone away since it is stored in an ambiguous state and location in the SAN. This at best results in confused users that all of a sudden can't use the SAN in the way they expect. At worst it results in lost access to the SAN completely, for no apparent reason. This can be a difficult problem to diagnose and may require the entire SAN database to be zero’ed out and reconfiguration and setup to take place again.

The state of how storage was last mounted is not automatically preserved between boots or sessions. FibreShare however does contains an Auto-Mount setting for each volume that the user can choose as a default preference to mount the volume in a chosen state after the user logs into FibreShare.

The user must log in each time. There is no preference to automatically log in and just go to work.

There are no “Projects” paradigm, access is strictly controlled and managed by a system administrator. A system administrator has an administrator password and is allowed to administer a set of storage. The Administrator for storage can then create users and groups of users, and assign storage and allowed access privileges, to the storage. This basically results in a much larger burden for ongoing administration management to reassign storage privileges.

With FibreShare, a User logs into the system with a password and then can access the storage that has been assigned to him by the Administrator. If a User’s storage needs change, the Administrator must reconfigure the system to the users new needs.

FirbeShare’s administration model is to assign users or groups of users to each volume, as opposed to assigning volumes to users. This creates a lot of additional work for the administrator to manage. If you have 130 volumes with FibreShare, the administrator has to go to each of the 130 volumes and set which users can access the volumes before anyone can do anything.

FibreShare’s overall look and feel consists of text lists, pop-down and contextual menus. There are limited drag-and-drop targets, but in general the user must click or select a lot in the GUI to get anything done. During pending operations there is no visual feedback.

FibreShare does not have an event log function.

FibreShare has a configurable timer for auto-updating read-only volumes, however it attempts to update them every time the timer expires, even when no updating is needed. This results in many unnecessary unmounting and mounting of read-only volumes. This can be very annoying to the user.

FibreShare does not have any way to suspend its database and disk operations, which occur to each SAN disk, so the only way to stop FibreShare from not interfering by sending unwanted I/Os to the storage during critical times, such as digitizing, is to grab exclusive access to all the volumes on that piece of storage which can also be annoying to the user.

FibreShare prevents rogue hosts on the SAN at the cost of requiring all storage on the SAN be setup using the Charismac Anubis utility.

Rorke's ImageSAN™

Architecture

A company called Tiger Technology Ltd., out of Colorado, appears to have contracted in whole or in part with Rorke to create the ImageSAN® software. ImageSAN is based on shared storage technology utilizing a Client / Server architecture. ImageSAN suffers from the same high availability and scalability issues common to these SAN architectures. It requires metadata server(s) to traffic all I/O to the file systems.

From a OS perspective, ImageSAN is in the I/O path of the system and couples NFS file system traffic, special metadata file mapping requests, and direct to SAN storage I/O to accomplish reading and writing to files. At the cost of high availability, scalability, and performance it does allow multiple writers to the same file system at the same time. This limits is applicability to post production environments streaming audio and video data.

Under the ImageSAN architecture, a failure of a machine on the SAN, such as a metadata server, will result in the complete loss of access to the file system(s) it was responsible for serving. ImageSAN requires a separate Ethernet LAN network from the SAN Fibre Channel to be in place to communicate the NFS and metadata traffic.

Installation and Setup

Installation is from CD using a single installer file. The user must enter license keys in order to activate or deactivate the software. There are also license keys that allow ImageSAN to operate in demonstration mode for 30 days. Each SAN attached computer should have the software installed.

ImageSAN requires that the file systems are of the HFS+ type, and that the SAN network use ATTO Fibre Channel HBAs (according to their documentation).

Therefore, customers not currently using ATTO HBAs must purchase them for ImageSAN use. Rorke has told us however that it is possible to use ImageSAN with other HBAs, although at the time, we were not able to get this working.

ImageSAN terminology names the machine that serves metadata for a file system the Master. There are three states for a file system: Public, Private and Not Available. Configuration is susceptible to human error as many settings must be manually coordinated across the entire SAN, including servers and clients, without computer assistance.

The Administrator sets a file system to Public Master on the machine that will server the metadata, and Public Client on all the other machines in the SAN. A Private file system is one for which ImageSAN does not provide SAN data protection. The Administrator must be careful when using a Private file system and make sure that it is set as Not Available on all the other ImageSAN computers otherwise data corruption is likely. If any mistakes are made, ImageSAN is unforgiving.

As you can see, in an active SAN with any kind of storage or workload changes, an ImageSAN network requires tedious configuration changes resulting in a heavy administrative burden. The price for human error is data corruption.

Additionally, administration is not centralized. This means that the Administrator physically has to manage each SAN workstation to maintain the correct configuration, as opposed to just sitting at one computer and configuring the entire SAN network. This is equivalent to another completely separate network system that will burden the IT administrator.

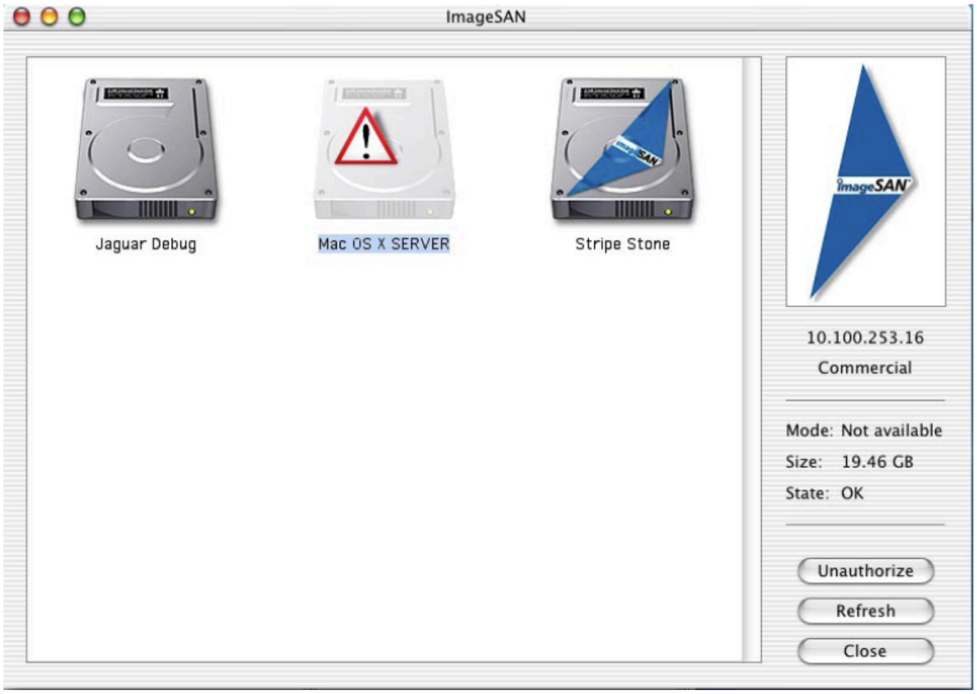
The old saying about the more moving parts that there are in an object, the more likely there are to be problems, applies to ImageSAN. Because of its complex nature, modifications to NFS, working with SAN and LAN traffic, network port configurations (8100, 111, 2049), firewall concerns, different I/O path alterations and complex metadata transactions, it is susceptible to breaking easier. This will also result in more software maintenance as new OS releases take place.

The setup requires setting the correct network ports for communications, and that the NFS startup items are configured correctly. Then, once every machine individually has all the right settings for Public Master, Public Client, Private, and Not Available for the file systems, the setup is complete and the machines must be restarted.

Usability model

ImageSAN launches automatically when the user logs into the computer, and volumes are mounted as appropriate. The application displays all file systems on the SAN. Icons for the file systems indicate its present state. Icons with ImageSAN stamped are either Public Master or Public Client file systems.

Standard volume icons are Private file systems. Standard volume icons that are transparent are Not Available file systems. Icons that are broken indicate a file system with an error (for example the file system Master could not be contacted). Icons with an attention sign indicate a change that has not been applied yet (for example from one state to another).



10.100.253.16
Commercial

Mode: Not available
Size: 19.46 GB
State: OK

IMAGESAN OSX WINDOW SHOWING MODES OF ALL SAN FILE SYSTEMS

ImageSAN inherits the user account security of the user that has logged into the OS, thus leveraging the security model that is setup and enforced by Mac OS and the IT administrator. However, ImageSAN does not protect against rogue hosts, which are computers without ImageSAN installed and attached to the SAN. If this happens you will corrupt data.

Information about a file system can be viewed, one at a time, by selecting it in the main window and looking to the right to see the file systems mode, size and state (but not free space). This information can be refreshed and changed. Information about the file system is changed in another window, for example to Public Master.

Permissions for a file system are configured on each machine, for each Public Client file system, for each user of the machine in question (a lot of configuration indeed!). These are the list of local users of the machine, and you can set an option such as to allow read and write access for that user. NOTE: This must be done on each machine individually.

ImageSAN has a feature that allows a mapping of a central server to client security mapping of users from the local list of users on the clients workstations. It also has a file and folder permissions option, on a master machine, that applies to the master machine's local users, that allows configuration of a domain (like a group). Then you can have the owner have read/write access to a folder and groups for example only have read access.

ImageSAN includes a file system performance test which results can be used to configure the ImageSAN drivers settings, including maximum client cache size, write expand size, direct read minimum size, among other settings. This must be done on each client.

Write cache is a dangerous per-client setting that must be used with care. If there is a problem with the computer between when it cached the data and actually wrote the data out to disk, then that data will simply be lost. This can become a big problem with applications; such as databases that utilize log structured, check-pointed, or transactional semantics. This is because they falsely believe that data has been written out when in fact it has not, leaving the system in an inconsistent and often unrecoverable state. In the case of plain file system level caching, corruption of data is also likely.

Another per-client setting, Write expand size, was necessary because of the performance limitations of this type of architecture due to the metadata server traffic. Many applications write performance were terrible due to the fact that the metadata server is the single cache-coherent point at which file system disk allocations occur. This option allows allocating large chunks of data at a time (from 100 MB to 500 MB), even if not needed at the moment, so that locally, data can be written faster. The excess allocated data is later released. Even with this, write performance is often not fast enough for various types of real-time sensitive digitizing, such as for audio (e.g. ProTools) and video capture (Digital High Definition), because the act of allocation is global to the server and happens over a potentially congested LAN network.

For this reason, the best hopes of digitizing are always on the machine that is the Master for the file system in question, and should only be attempted when clients are not accessing the file system. Also for this usability reason, the famed multiple-writers to the same file system at the same time feature of this architecture becomes moot because in real applications, this performance is not acceptable, so you cannot use this capability.

In our tests, ImageSAN performed at a fraction of the SAN speed as compared with other products. Many timelines couldn't play at all because of the slow performance.

The multi-writer to a file system at the same time ability is mostly why customers are interested in this architecture. At first it sounds natural to their local disk experience, or network experience. It is not until they get into how these architectures are actually setup and used do they discover the limitations that negate having this ability leaving them back with figuring out how in the heck they will get enough performance to scale their workload.

Interesting Aspects

ImageSAN requires volumes to be formatted as HFS+

ImageSAN only works with ATTO Fibre Channel HBAs (according to their documentation, although it is supposed to work with other HBAs)

ImageSAN does not protect against rogue SAN hosts.

ImageSAN requires the workstations to all be connected to the same segment of a LAN, have a unique IP, and be able to ping each other (DHCP is not really an option).

ImageSAN involves a complicated installation, setup, and administration model that introduces many changes to the workstations, requires at least one metadata server, and introduces many things that can go wrong. Please refer to the ImageSAN documentation, or demo, to illustrate this complexity.

ImageSAN requires extensive ongoing Administration on each machine in the SAN, especially if any of the permissions features are used at all.

ImageSAN's configuration is highly susceptible to human errors that can easily result in data corruption. Imagine a file system with two Public Masters!

ImageSAN's write cache feature can be very bad for postproduction environments. Similar to the store-and-forward operation of servers, it results in multiple memory moves as the data is first copied from an application buffer to the ImageSAN cache, and then is eventually written out to disk with another operation. This especially can interfere with other PCI card performance.

ImageSAN is sensitive to IP address changes, especially on the Master machine(s). If this changes, the clients must be restarted.

The reasons customers are interested in "file-level" control are actually completely negated when one understands how this system performs and how it must be configured and maintained.

At the time of this writing ImageSAN still did not work with Mac OS 10.3.x.

Studio Network Solutions SANmp™ (OEM XSHARE)

SANmp™

Multi-Platform Management Software

Architecture

SANmp is based on shared storage technology utilizing a serverless SAN architecture. Persistent network state is recorded in a SAN database that is distributed/replicated across all SAN storage. Because the database only records the persistent state, if it is corrupted or lost, user will continue to operate with the file systems in their last state until the problem can be fixed. A reset database command will clear the database on a selected disk, at which time users need to be recopied to that database and permissions reset.

Installation

Installation is from CD using standard Apple installer technology. Each SAN attached computer should have the software installed. The software is protected with a USB hardware key that must be attached to each computer using the software.

SANmp has a minimum file system size requirement of 1 GB.

SANmp requires the Administrator to selectively decide which disks are in the SAN from a list of all disks known to the OS, and then convert those disks to SANmp disks, causing all data on those disks to be erased, before it can be used. According to SANmp documentation “Converting a disk will destroy all information on all of the volumes of the disk. Are you sure you wish to convert the disk?” You can convert the disk back afterwards, but if you ever convert in back again a SANmp disk, you will again loose all data on the disk.

SANmp requires all SAN computers sleep standby mode.

Because each disk must be converted to a SANmp disk, space for a SAN Database for that disk is then created. If you add a user to the system, it is then added to all disks SAN Database. This appears to have serious implications in a changing SAN, or a SAN that is being combined with other SANmp volumes that may have some, overlapping, or different users on those disks. This will result in ambiguous behavior.

Once a SANmp disk is converted, the user may not simply use standard disk utility programs, such as defragmenters, or repair utilities, according to SANmp documentation. In fact, we were unable to even repartition disks without disabling the SANmp software first.

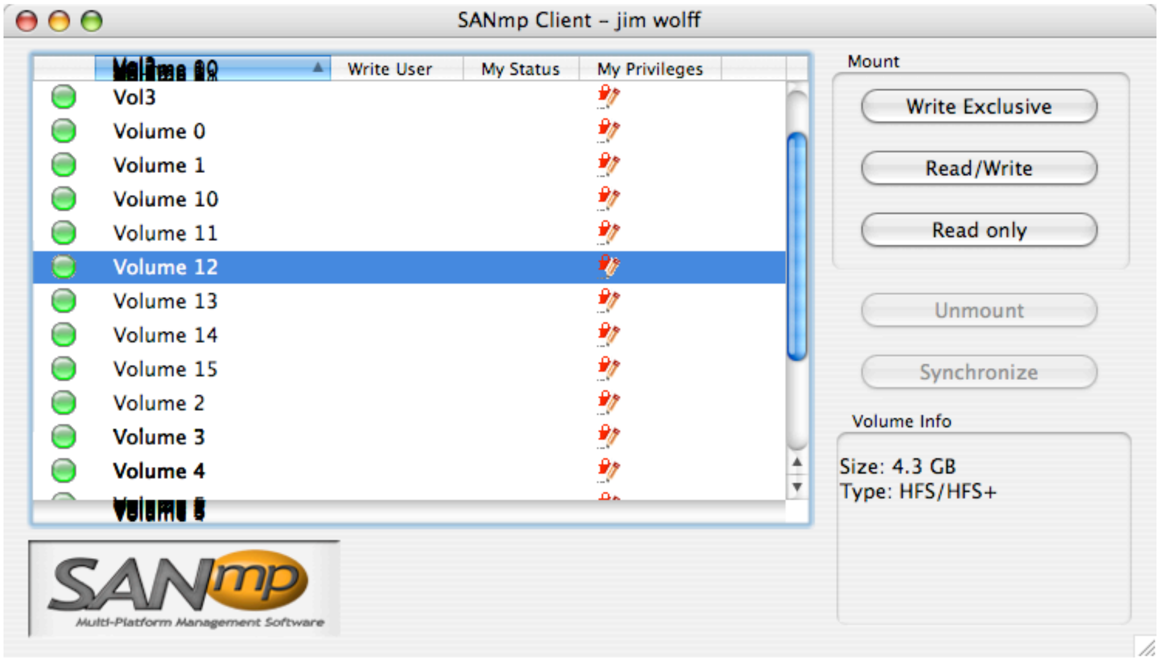
SANmp says not to use disks that are journaled. This is bad because journaled disks are the most reliable. It is even more important since SANmp doesn't appear to simply work with standard repair utilities without disabling their software.

Once disks are converted and erased, users are created and placed on each disk. Then privileges must be set before the user can mount any file systems.

Interference to post productions applications, such as capture or playback is unclear as SANmp interactions to the disk are not based on configurable timers. For instance, reference to automatic recovery of a write access request for a user that has crashed with write access says to wait 15 seconds for this to automatically recover. This feature implies a heartbeat operation that must be taking place on the writer users disk that conducts write I/O to the disk. This additional I/O can cause problems, especially in loaded systems. There is no way to disable or suspend this additional activity.

Usability model

Once the user logs into SANmp with a name and password, they are allowed to see the file systems that they have privileges for (read/write, exclusive, or read) as indicated by the first column in the GUI.



	Volume	Write User	My Status	My Privileges
●	Vol3			✎
●	Volume 0			✎
●	Volume 1			✎
●	Volume 10			✎
●	Volume 11			✎
●	Volume 12			✎
●	Volume 13			✎
●	Volume 14			✎
●	Volume 15			✎
●	Volume 2			✎
●	Volume 3			✎
●	Volume 4			✎
●	Volume 5			✎

Size: 4.3 GB
Type: HFS/HFS+

The volume column in the GUI displays the name of the file system.

The write user colume displays the name of the user than currently has write access to the volume, if any.

The my status column displays how the volume is currently mounted.

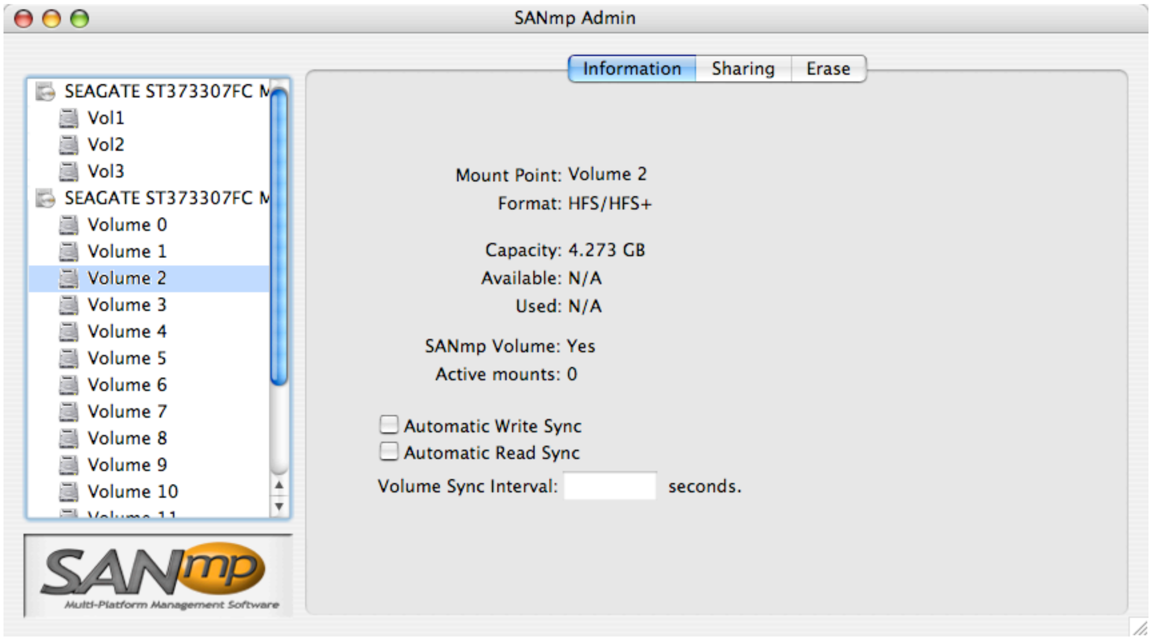
The my privileges column indicates what level of permissions have been granted to the user for this volume.

The last column indicates availability of a volume. When unlocked, the user can mount the volume. When locked, the user cannot mount the file system (e.g. the Administrator has disabled the file system).

SANmp allows the user to mount disks in read/write, exclusive, or read-only mode depending on the access privileges set by the Administrator. SANmp allows the user to unmount volumes. SANmp can restrict users to only mounting volumes in a read-only state that is useful for libraries and other collections that should not be modified, but are good to share among users.

There is no drag-and-drop or contextual menu functionality in the SANmp GUI. You must select the file system to mount and either click a button, use a hotkey, go to a menu command, or use the CLI. This is the case for unmounting as well.

A separate administration application is used to administer the SAN. This is where disks are converted and unconverted. This is also where timers can be enabled to synchronize writing and reading among users. Additionally the number of current users that have the volume mounted is displayed.



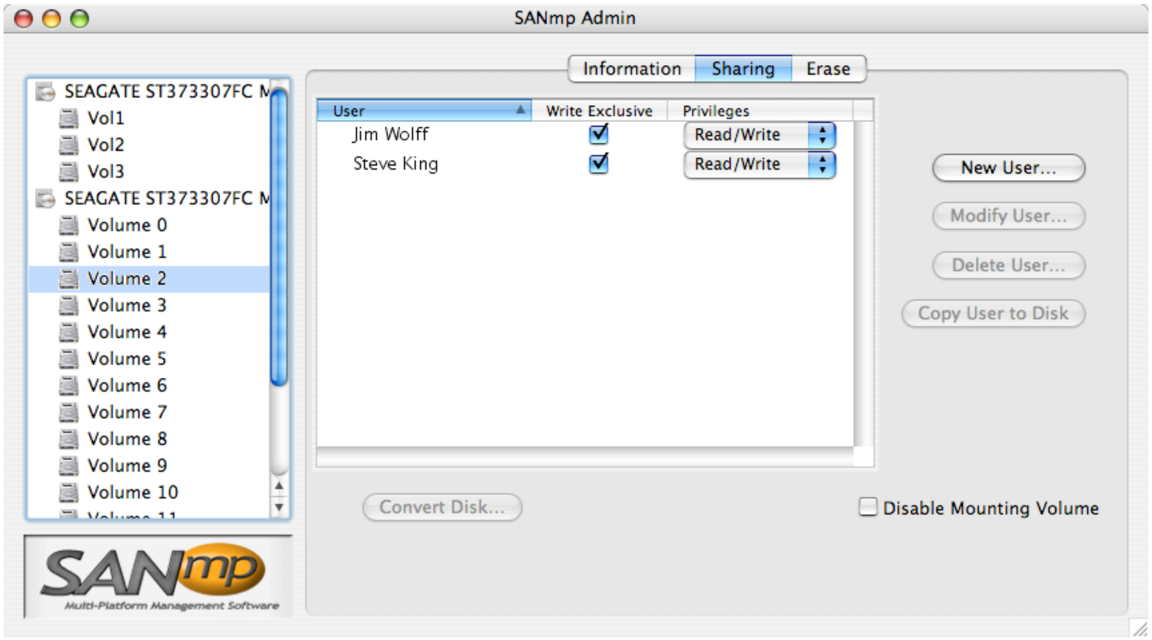
Mount Point: Volume 2
Format: HFS/HFS+

Capacity: 4.273 GB
Available: N/A
Used: N/A

SANmp Volume: Yes
Active mounts: 0

☐ Automatic Write Sync
☐ Automatic Read Sync
Volume Sync Interval: seconds.

The sharing tab in the administration application is where users and privileges are managed.



User	Write Exclusive	Privileges
Jim Wolff	☑	Read/Write
Steve King	☑	Read/Write

Interesting Aspects

FibreJet and SANmp (August 2005)

The following is a summary of some important information to know when comparing FibreJet to SANmp. All the following has been confirmed against SANmp version 1.5.0 build 16 tested on Mac OS X 10.4.2 in August of 2005 (the latest version available then). The reader is encouraged to investigate if there is a later version that has addressed any potential issues, or if the behavior is different for an older or newer Mac OS version than that which was used herein, as manufacturers are always updating software features.

Support for software RAID-0 under Mac OS 8-9 and OS X

SANmp: Big problem. There is no cross platform software RAID-0 support between Mac OS 8-9 and OS X. SANmp offers software RAID-0 on Mac OS 8-9 through its compatibility with Intech Software's Hard Disk Speed Tools. There is no software RAID-0 support on Mac OS X, although the documentation claims otherwise (so they may fix this). When we tested striping using Apple's Disk Utility SANmp didn't recognize it. For straight non-striped partitions (using any utility), cross platform is supported.

FibreJet: No problem. FibreJet supports standard non-striped volume partition format as well as striped (RAID-0) volumes from ATTO technology (e.g. StudioNet FC, AVID, ExpressStripe, OS 9 & X). In fact, it does this cross-platform so you don't have to modify any existing data on any existing disk, no conversion, no data migration required. By the way, FibreJet supports SANmp formatted volumes so converting up to FibreJet is painless. In addition to software RAID-0 striping using ATTO's utility, straight non-striped partitions (using any utility) are also supported cross platform.

Support for standard OS X partition utilities

SANmp: Under Mac OS X, any standard utilities that use straight partitioning (e.g. non-striped) are supported. There is no support for any software RAID-0 (striped) sets created by any Mac OS X utility. NOTE: If you have a Mac OS 9 computer on the SAN, you must use the OS 9 utility that SANmp provides (Hard Disk SpeedTools) to partition, otherwise the OS 9 machine will not be able to access the file systems.

FibreJet: Under Mac OS X, any standard utilities that use straight partitioning are supported. In addition, software RAID-0 is supported when using CommandSoft's StorDirector, Apple's Disk Utility, or ATTO's striping utility. Cross platform software RAID-0 (between Mac OS 8-9 and OS X) is supported if ATTO's striping utility is used. To see disks between OS 8-9 and OS X, any straight partitioning utility can be used on either platform to create the file systems. To use software RAID-0 between OS 8-9 and OS X, you should use ATTO's utility to partition.

File System Limitations and standard disk utilities

SANmp: SANmp has a minimum file system size requirement of 1 GB. SANmp says not to use disks that are journaled. This is bad because journaled disks are the most reliable. It is even more important since SAMmp doesn't appear to work with standard repair utilities (they mention once converted you cannot simply use standard disk utility programs, such as defragmenters).

FibreJet: No limitations on file system size. You can, and are encouraged to use journaled disks where performance allows (because they never get file system corruption on computer crashes). You can use standard disk utility and repair utilities on FibreJet disks using our "Volume Maintenance Mode", although you probably won't ever need to if you use the disks journaled in the first place!

Data Loss at setup time and later

SANmp:SANmp requires the Administrator to selectively decide which disks are in the SAN from a list of all disks known to the OS, and then convert those disks to SANmp disks, causing all data on those disks to be erased. SANmp requires that the Administrator convert a disk before it can be used on the SAN. This will convert the disk to a SANmp disk, and is a permanent step. According to SANmp documentation "Converting a disk will destroy all information on all of the volumes of the disk. Are you sure you wish to convert the disk?" You can convert the disk back afterwards, but if you ever need to convert in back to a SANmp disk again, you will again loose all data on the disk. This is problem because it means all your data must be completely backed up before converting to a SANmp disk (so that it can be restored afterwards). With the version we tested, we were never able to conver/unconvert/and reconvert successfully. The reconvert always failed, requiring disabling the SANmp software, completely reformatting the drive, and reenabling the SANmp software, before the reconvert would succeed. Any convert/reconvert always erased all existing data on the drive. There is no option to convert or unconvert a single file system on a disk with multiple file systems.

Non-SANmp disks (those that have not been converted) are unprotected in the SAN and will result in multiple writers if more than one computer is present in a SAN with unconverted storage.

FibreJet: Uses data in-place and does not erase it or require repartitioning or data migration. You can use your file systems with FibreJet without losing any data. You can later not use them with FibreJet and not loose data. If you want to use them again with FibreJet later, there is again no data loss. Also, FibreJet works cross platform with OS 8-9 and OS X, and allows you to keep using your existing striped and non-striped volumes in place.

When FibreJet is installed on a SAN attached computer, new storage put into the SAN is always protected from multiple-writers automatically (unlike SANmp). Additionally, an optional feature called Rogue host protection, can be employed which will change the signatures on available storage in the SAN so that even computers without any SAN software installed will be prevented from accidentally mounting the file systems and creating a multiple-writer situation (like a SANmp converted disk).

SAN Database issues and Disk changes

SANmp:

Because each disk must be converted to a SANmp disk, space for a SAN Database for that disk is then created. If you add a user to the system, it is then added to all SAN Databases online at that time. When changing the SAN without all storage online, or a SAN that is being combined with other SANmp volumes that may have some overlapping, or different users on those disks, a password synchronization problems can arise that must be corrected.

A Copy Users To Disks command must be used for new disks to manually replicate the SAN Database information. After the SAN Database is replicated to the new disks, users must be manually configured for privileges for the additional file systems. There is no project or group concept for file systems, users, or permissions.

Once a user is added, the privileges must be set before the user can mount any file systems.

The administrator password is stored locally, so if multiple machines are going to be used for administration, passwords much be changed at each machine to keep them in sync.

Interference to post productions applications, such as capture or playback is unclear as SANmp interactions to the disk are not based on configurable timers. For instance, reference to automatic recovery of a write access request for a user that has crashed with write access says to wait 45 seconds for this to automatically recover. This feature implies a heartbeat operation that must be taking place on the writer users disk that conducts write I/O to the disk. This additional I/O can cause problems, especially in loaded systems. There is no way to disable or suspend this additional activity.

Logging in multiple times as the same user from multiple machines is possible and has the following effects to be aware: the user is prevented from mounting the same file system more than once; the user may only access the file systems from a single physical disk LUN at a time (meaning the second login cannot use any of those file systems).

When adding new storage SANmp recommends that all machines be powered to avoid the possibility of a rouge host. There is no support for dynamic storage.

SANmp implements a separate security model for the SAN network, including its own users and privileges, thus requiring the user to log into the SANmp system before any storage can be accessed. This security system is separate and apart from the users and privileges that may have been configured on the Mac OS X network by the IT administrator. The administrator can manage users using the SANmp admin GUI or the CLI.

FibreJet:

FibreJet dynamically detects new storage in the SAN without any reboot or restart required (as long as the underlying HBA supports the dynamic drive model) and allows users to use the new storage automatically as provided for by the Administrator. If the storage was not previously in any project created by the Administrator, the Administrator simply drags the new storage into any existing or new projects to allow access.

FibreJet allows all timers that touch the SAN database and SAN disks to be highly configurable and to be disabled as an option. FibreJet leverages the Mac OS X security model that IT administrators already manage so there is no additional need for separate users. FibreJet manages storage access by grouping into Projects that can be given passwords, made invisible to others, restricted access to the file systems and many other options allowing a completely open or a completely closed environment depending on the needs of the customer.

FibreJet also allows a mode that allows "users" to be created for customers that are more comfortable with a user-based paradigm instead of Projects. The administrator manages projects.

User Interface

SANmp:

There is no drag-and-drop, contextual or dock menu functionality in the SANmp GUI. You must select the file system(s) to mount and either click a button, use a hot-key, or go to a menu command. This is the case for unmounting as well.

The version we tested had a minor scrolling problem that garbled the volume list in the main window when scrolling up or down.

There is no indication of the free space available in a file system through the SANmp client GUI, so the user cannot tell at a glace using the client GUI where to place new material.

There is also no immediate indication of the size of the file system. The users can view this information for an individual file system by selecting it in the main window.

SANmp offers no way to dynamically request write access from another user.

SANmp does not support saving settings so that a users file systems will automatically be mounted in its last state next time they go to work. However, the user can utilize the CLI to write a script that lists the commands to mount certain file system in a certain state. This can then be run the mount the file systems as defined by the script.

SANmp's default admin password, adminpw111, can be used by anyone installing the software who has an admin HASP key to bypass all security in the system, as long as that user has access to the installation CD. Therefore, they suggest keeping that CD in a safe place to protect the SAN.

SANmp volume access is all or nothing and highly administrator intensive. There is no grouping of storage concept; the administrator is responsible for and making individual file system privilege assignments to specific users.

SANmp offers a command-line interface, which is an alternative way to control some functions of the application, and can be combined with scripting to programmatically control it.

FibreJet:

FibreJet provides many ways to accomplish the same thing. FibreJet GUI supports all Mac-style interfaces including extensive use of drag-and-drop, contextual menu functionality, dock-menu functionality, and customizable tool bar with drop targets and pop-up lists.

FibreJet lists the file system size and file system free space for all accessible volumes right in the main window, and also allows you to sort the list by different attributes, including space.

FibreJet supports dynamic passing of write access between users. Additionally a messaging facility allows a message from the requester and the responder so they can communicate.

FibreJet provides a means to save a users last settings so next time they work, the system will automatically come up with everything as they left it. This can also be disabled if the customer wishes to be more protective of the environment (e.g. requiring them to log into projects).

If the administrator forgets their password, we can provide a one-use code that will work to reset the password. If they forget their password again in the future the same code will not work and they will have to obtain another once-use code.

FibreJet's Project paradigm is superior for Post production environments. Users log into Projects, which are just groups of storage with different access characteristics. For instance, a "Star Wars" projects with all the storage they need for that project. Or, "Effects Library" for the storage that has common stuff, but is set so that users cannot modify or destroy its contents. Users can log into multiple projects. Storage can belong to multiple projects; in that case, the most permissive permissions of the aggregate projects are used for the in common storage.

FibreJet supports AppleScript, Apple's preferred method for programmatically controlling an application. This would be used for instance as part of an automated a SAN backup process.

Astera Timba

Architecture

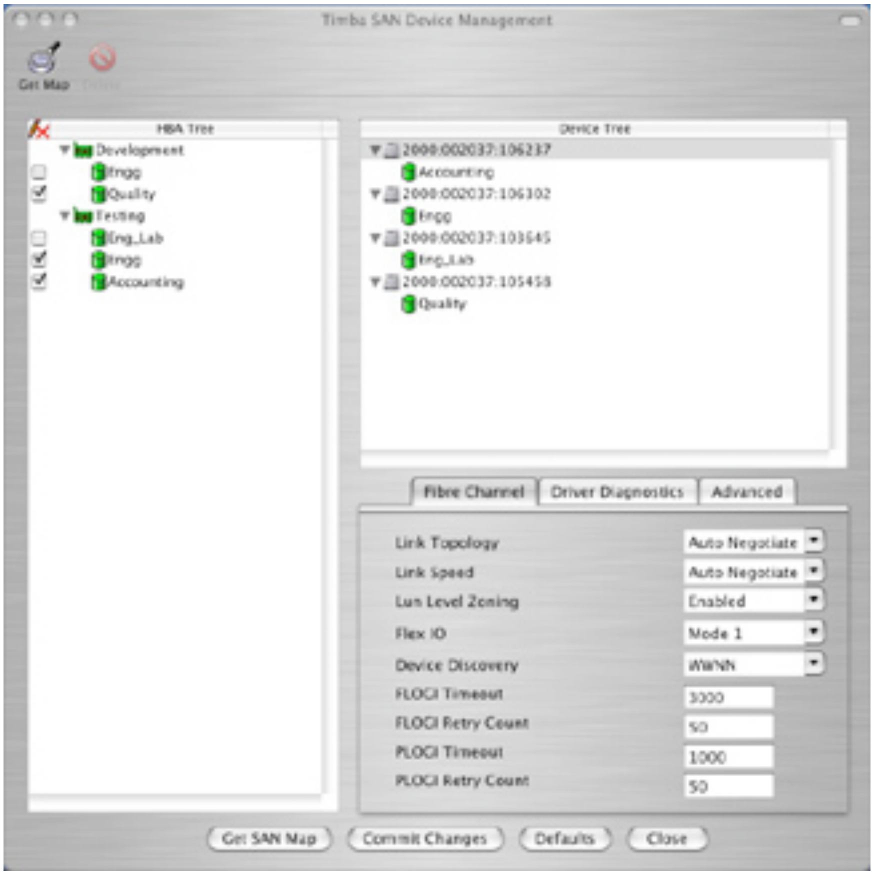
Timba is based on shared storage technology utilizing a serverless SAN architecture. It is believed the persistent network state is available in each SAN workstation and changes are negotiated over the Ethernet LAN network that each SAN workstation must be attached.

Timba is a LUN level SAN product. Its granularity is at the entire LUN levels making it somewhat obsolete in all but the most limited SAN situations. This is because an entire LUN, which potentially might have 10’s or 100’s of file systems on it normally, must be acquired for write access or read-only access as a whole unit. This is not how users typically work in a SAN and would require strict workflow design to be used at all.

Installation

Timba requires Astera HBAs to operate in each SAN machine. The Timba software was built on the Jumanji HBA control software, and must be installed on each machine.

Usability model



HBA Tree

- Development
  - □ Engg
  - ☑ Quality
- Testing
  - □ Eng_Lab
  - ☑ Engg
  - ☑ Accounting

Device Tree

- 2000:002037:106237
  - Accounting
- 2000:002037:106302
  - Engg
- 2000:002037:103645
  - Eng_Lab
- 2000:002037:105458
  - Quality

Link Topology	Auto Negotiate
Link Speed	Auto Negotiate
Lun Level Zoning	Enabled
Flex IO	Mode 1
Device Discovery	WWNN
FLOGI Timeout	3000
FLOGI Retry Count	50
PLOGI Timeout	1000
PLOGI Retry Count	50

Interesting Aspects

None.

SGI CXFS InfiniteStorage Share Filesystem

Architecture

CXFS for Mac OS X was not available at the time of this writing.

CXFS is based on shared storage technology utilizing a Client / Server architecture. Therefore, CXFS, just as ImageSAN, suffers from some of the same issues common to these SAN architectures. It requires metadata server(s) to traffic all I/O to the file systems.

SGI has optimized CXFS however using proprietary metadata transactions rather than piggybacking on NFS or CIFS traffic. This should result is less metadata burden on the metadata server.

The underlying file systems are based on the SGI XFS format. As a result, all the metadata servers are required to be SGI IRIX OS machines, which are very expensive.

This means only SGI tools can be used to maintain, repair etc…these types of volumes because they are not HFS+ based.

SGI CXFS is supposed to be able to fail-over metadata servers if they fail. If this happens, the user will still loose data in the files they were transacting to if they were in the metadata servers cache at the time. Therefore, although the file system will not be corrupt, the file data itself can be corrupt. Additionally, although the SAN network will still be up and running, and machines accessing the metadata server at the time may still crash or have their application that was writing crash because of the lost cached data.

Miscellaneous Information

iSCSI, Gigabit and 10 Gigabit Ethernet

A new storage traffic transport protocol called iSCSI encapsulates SCSI traffic over standard TCP/IP networks. As Network Interface Cards (NICs) are developed to offload the packetization steps for TCP/IP data from the CPU, protocols like this for transporting raw SCSI disk blocks over longer distances and standard infrastructure become more practical. At the same time, these could still be considered SANs and require similar or the same management software depending on how the disks are presented to the OS layers.

HD Video Streams

Digital Video standards, including the so-called high-definition (e.g. 10-bit 1080i 59.94 fps), pack a lot of data. These streams are about 160 MB/s with an I/O size of about 1 MB.

SD Video Streams

Standard Definition video streams at 10-bit resolution are about 30 MB/s, and have a smaller than 1 MB I/O size typically.

DV Video Streams

Digital Video streams are about 4 MB/s, and have yet smaller I/O sizes.

Professional Audio

Environments specializing in sound, such as AVID ProTools shops, are characterized by many streams (e.g. 128) of audio. Although a single audio stream is less MB/s than video, it is seek-time intensive, especially when working with over 100 streams at the same time. That is why it is possible for storage systems to actually be taxed more in an audio shop than a video shop.

Final Cut Pro

Apple’s Final Cut Pro® product (version 4.1) is a find tool video editing in small to medium sized timelines. Bug fixes and updates are available online via the built in software-update feature of Mac OS X.

AVID Media/Film Composer

AVID’s Media/Film Composer and other related products are well known in many environments and have been the standard for many years. AVID is well suited for large complex projects with long timelines.

AVID ProTools

AVID DigiDesign ProTools is a professional high-end audio editing solution, for the most demanding audio editing environments.

Media 100

Media 100 is a mid-level non-linear editing solution, that combines easy of use with many high-end functionality.

AJA Kona and I/O Box

AJA is the maker of fine video capture boards both for SD and HD. The I/O box is a firewire breakout box that gives you many input and output options attaching to the computer through firewire.

Blackmagic

Blackmagic is a company that in addition to making drivers for the AJA Kona cards, also makes it own SD and HD cards as well as other products.

Pinnacle Cinewave

Pinnacle makes the Cinewave line of capture boards that support SD and HD. These boards also feature real-time effects that can save the user from the time require to render while making changes to the timeline.

Aurora Ignighter

Aurora makes the Ignighter line of capture boards.

Brocade

Brocade makes 2 GB/s Fibre Channel switches.

QLogic

QLogic makes 2 GB/s Fibre Channel switches and its own line of HBAs.

Emulex

Emulex makes Vixel 2 GB/s Fibre Channel switches and its own line of HBAs

McData

McData makes 2 GB/s Fibre Channel switches

Apple HBA

Apple OEMs the LSI Logic's Fibre Channel HBA

ATTO

ATTO Technology makes its own Fibre Channel HBA which is based on the QLogic chipset.

Astera

Astera makes its own Fibre Channel HBAs, as well as the Timba SAN Management software.

Zoning

Zoning is a function of Fibre Channel switches that isolates the network ports into functional groups. In some cases, it is necessary to configure zoning so that SAN host computers do not interfere with each other. A feature of switches, called switch change notifications, can sometimes interfere with playback or capture in Post Production environments. By having a separate zone for each host, and in each of those zones including all the SAN storage, the network is isolated from events cause by hosts coming on or off the SAN network.

High Availability

High Availability refers to the ability of a system to continue to operate in different failure conditions. It also refers to the ability of the system to access a resource via multiple means, or paths.

Scalability

Scalability refers to the ability of a system to grow in size and performance. It also refers to the ability of a system to grow in size and performance with regard to access to a single logical resource.

JBOD

Just a Bunch Of Disks (JBOD) refers to a storage arrangement of individual disks, often placed in an enclosure, and individually accessible and addressable. Offering no group failure recovery, JBODs are often software striped (RAID-0) to create a single large virtual piece of storage that is very fast as it stripes data that is written or read across all available disks in the RAID set.

RAID

Redundant Array of Independent Disks (RAID) is a standard for applying different algorithms to multiple physical disks. RAID-0 is striping, RAID-1 is mirroring, RAID-3 is striping with parity, RAID-5 is striping with alternating parity, and there are other combinations and RAID levels. RAID-5 is the most common used in hardware devices (so-called hardware RAID). Levels with parity, or error correcting code, allow certain amounts of failures to the physical disks while allowing continued access to the data without loss. Problem can be fixed online, including replacing bad disks, which can be reconstructed all while I/O continues to happen in the system. When a failure occurs, the system operates with degraded performance until the problem is addressed.

Apple XRAID

Xserve RAID is a hardware RAID developed by Apple. The RAID controller is developed by Accusys, a Chinese company. This low-cost RAID is very popular for its cool looks and average performance. XRAID is on the low-end of RAID devices and does not support dual-redundant controllers so there are still single points of failure with this device.

Chaparral RIO line

Chaparral makes a mid to high end RAID device that performs at the high-end of functionality and performance. A single device has been tested with up to 4 streams of HD.

Apple Stripe (Disk Utility)

Apple's Disk Utility is a low-end disk utility for Mac OS X. It provides basic disk services including partitioning, repairing, and disk imaging. It supports RAID-0 and RAID-1. Apple's RAID-0 does not allow the user to further partition the device into smaller chunks, which limits its use in SAN environments.

ATTO Stripe

ATTO's ExpressRAID and ExpressStripe utilities for Mac OS X provide flexible partitioning and performance testing options. They allow standard, RAID-0, and RAID-1 partitioning on the same LUN. Also allowed are the ability to partition the RAID sets so that the user can create many individual file systems for use on the SAN.

Legal Notices

CommandSoft® and FibreJet® are trademarks of CommandSoft, Inc.
Apple®, Xserve, Xserve RAID, Final Cut Pro are trademarks of Apple Computer, Inc.
Charismac, FibreShare, Anubis are trademarks of Charismac
Rorke, ImageSAN are trademarks of Rorke Data Systems, Inc.
Studio Network Solutions, SANmp are trademarks of Studio Network Solutions
XSHARE is a trademark of RAID, Inc.
Astera, Jumanj, Timba are trademarks of Astera Technologies
SGI, CXFS, XFS, InfiniteStorage Shared Filesystem are trademarks of Silicon Graphics Incorporated
AVID, Media Composer, Film Composer, DigiDesign, ProTools are trademarks of AVID Technology, Inc.
Media100 is a trademark of Media100, Inc.
AJA, Kona, I/O Box are trademarks of AJA, Inc.
Blackmagic is a trademark of Blackmagic Design, Inc.
Pinnacle, Cinewave are trademarks of Pinnacle Systems, Inc.
Aurora, Ignightger are trademarks of Aurora Systems
Brocade is a trademark of Brocade Systems
QLogic is a trademark of QLogic Corporation
Emulex, Vixel is a trademark of Emulex Corporation
McData is a trademark of McData Corporation
ATTO, ExpressStripe, ExpressRAID are trademarks of ATTO Technology, Inc.
All other trademarks, are trademarks of their respective companies.

Index