

# NFS Fibre Channel Implementation

Greg Shaw  
Unix Operations Group  
Storage Technology Information Services  
Revision: June 1, 1999

## Summary

## Introduction

Storage Technology, through server consolidation, has centralized NFS services to a single pair of Sun 6000 servers in a High Availability (HA) configuration. When originally designed in 1996/1997, the NFS server pair were connected to StorageTek Nordique 9137 arrays with an approximate capacity of 1200 gigabytes (GB) of information. Since that time, the use of the NFS servers has increased, effectively filling all available space on the NFS servers.

A band-aid in the form of another reallocated 9137 allowed expansion of the NFS servers, but failed to address the real need to radically expand the capacity of the NFS servers. The final solution entailed installation of a Fibre Channel Storage Area Network (SAN) with StorageTek 9145 disk arrays. This document attempts to explain the transition, configuration and installation of the Information Services Storage Area Network.

## Terms and Acronyms

In any technical document, many terms will be reduced to acronyms. See the following list for definitions of the acronyms used in this document.

| <u>Acronym</u> | <u>Term</u>                                                                                                                              |
|----------------|------------------------------------------------------------------------------------------------------------------------------------------|
| HA             | High-Availability (redundant)                                                                                                            |
| NFS            | Network File System (network based file services)                                                                                        |
| SAN            | Storage Area Network                                                                                                                     |
| SP             | Storage Processor (within a 91xx Array)                                                                                                  |
| RAID           | Redundant Array of Inexpensive Disks                                                                                                     |
| LUN            | Logical Unit Number (a logical disk within a 91xx Array)                                                                                 |
| GB             | Gigabyte                                                                                                                                 |
| JBOD           | Just a Bunch Of Disks (non-RAID attached storage)                                                                                        |
| Bound, Binding | The logical configuration of a set of RAID drives into a single logical unit (LUN)                                                       |
| 91XX           | Conglomeration of both 9131, 9137, 9140 and 9145 arrays, representing any and all arrays.                                                |
| Online Primary | The default state in a two-server configuration, where both servers are working normally; no failures have been found.                   |
| Dual Services  | When a single server has failed in a two-server pair, the resulting server, doing double work, is referred to as being in dual services. |

## Scope

The current NFS HA configuration represents only the capabilities of Fibre Channel as of early 1999. Since the Fibre Channel capabilities remain an evolving standard, future installations may vary greatly from this configuration.

The reader should have an understanding and some practice with the following subjects:

- High-Availability with regard to the Veritas FirstWatch software
- Basic SCSI and Fibre Channel fundamentals (media types, basic functionality and capabilities).
- Veritas Volume Manager fundamentals, in particular Volume Manager Disk Groups, Volumes, and on-line data movement capabilities.

## Assumptions

- Performance remains a secondary goal to high availability.

- The NFS server solution must be a transitionable configuration in a production environment. At no time may the NFS servers be unavailable for an extended period of time for data movement.

## NFS Server SCSI Configuration

The original NFS server configuration, designed in 1996 and expanded in late 1998 consisted of the following components:

Sun 6000 Servers (each):

### Hardware Configuration

- 6-250mhz processors
- 2.2gb memory
- 3 CPU Boards
- 6 I/O Boards
- 24 Ultra-Fast-Wide Differential SCSI Controllers
- Sun 622Mb/s ATM adapter

### Software Configuration

- Solaris 2.5.1 with the recommended patch set
- Veritas Volume Manager 2.5.2
- Veritas FileSystem 3.2.1
- Veritas FirstWatch 2.2.4
- Application Transparent Failover (ATF) version 1.2.6
- Arrayguide array monitoring software
- BMC Patrol monitoring software (system availability)

### 9137 SCSI attach RAID Arrays

- 30-9gb disks
- 2 Tuscon Storage Processors with 64mb of cache memory per processor
- Internal disk configuration consisted of 5 RAID 5 LUNs of 5 disks each, 1 RAID 5 LUN with 4 disks, and a single disk hot spare.
- Flare code 9.55, PROM code 1.73
- Cache Parameters:
  - 8MB Read Cache
  - 52MB Write Cache

See Figure 1 for the following discussion.

The NFS servers have been cabled through every Storage Processor (SP), resulting in a SCSI bus for each Storage Processor terminated on each end by a server. By using this configuration, each server can access all 9137 attached disk. One problem remained; SCSI maintenance wasn't possible, as removal of a SCSI cable from either host resulted in a non-terminated SCSI cable, with undefined results (generally, any I/O on that SCSI connection failed, causing server problems).

Typical use for the systems is an *Online Primary* state, with both systems acting as NFS servers. In the event of a failure, the heartbeat links between the systems will indicate, to the available system, that the other system has failed. At that time, the available system will mount those disks formerly available to the other system, and make them available to the clients. This state is described as *Dual Services*, representing a single server doing both its own work, and the work of the failed server..

With this in mind, the servers have the following restrictions:

All available disks (other than the system disk) must be available to both servers at any time.

A channel failure can impact both systems. Thus, any channel failure must be repaired as quickly as possible, or, in the event of system maintenance, both servers must be taken off-line.

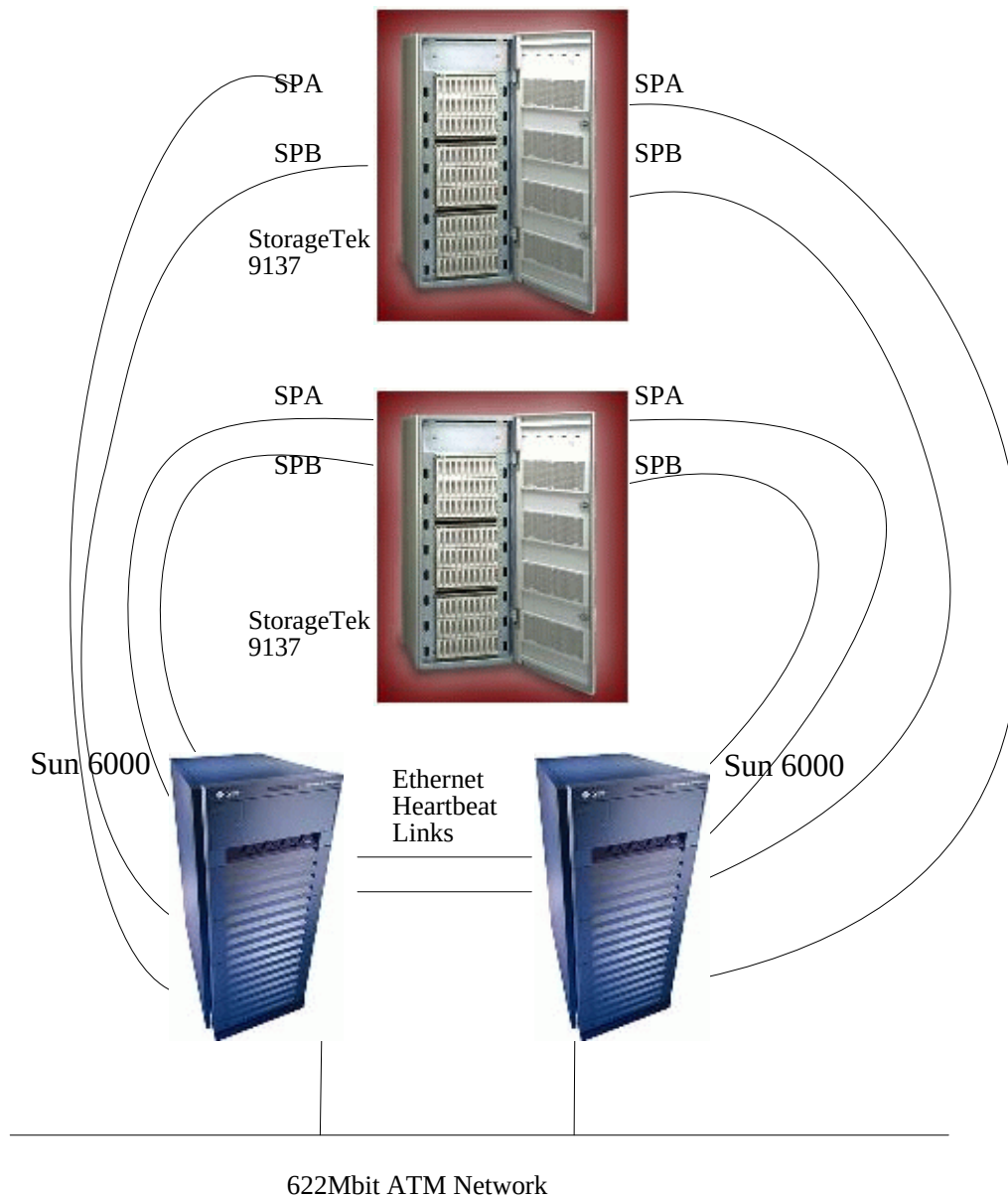
Maintenance of the servers can be accomplished singly by running in dual services, with the restriction that maintenance that could impact the running server be minimized.

## Outstanding Problems

The following problems were found on the SCSI NFS server configuration:

A single SCSI cable failure could cause both servers to fail, as the cable failure would cause cable termination failure. At that point, the server doing I/O to that channel would slowly fail. When the failing server was then stopped, to allow the other server to take over, the channel would remain failed, causing that server to fail also.

Figure 1:



## Test Environment Configuration

Prior to implementation in the production environment, all fundamental changes to the NFS server environment must be tested in a similar configuration. To satisfy this need, the following server configuration was used to emulate the NFS server configuration:

Sun 1000 Servers (each):

Hardware Configuration

- 2-50/60mhz processors
- 256mb memory
- 2 CPU and I/O Boards
- 2 Ultra-Fast-Wide Differential SCSI Controllers
- Sun 155Mb/s ATM adapter

Software Configuration

- Solaris 2.6 with the recommended patch set
- Veritas Volume Manager 2.5.2
- Veritas FileSystem 3.2.6
- Veritas FirstWatch 2.2.4
- Application Transparent Failover (ATF) version 1.2.6
- Arrayguide array monitoring software

9131 SCSI attach RAID Arrays

- 20-4gb disks
- 2 Phoenix Storage Processors with 64mb of cache memory per processor
- Internal disk configuration consisted of 3 RAID 5 LUNs of 5 disks each, 1 RAID 5 LUN with 4 disks, and a single disk hot spare.
- Flare code 9.55, PROM code 1.73
- Cache Parameters:
  - 8MB Read Cache
  - 52MB Write Cache

In addition to the above configuration, the following hardware/software was added to for Fibre Channel connectivity:

Hardware

Jaycor 1063 32-bit Fibre Channel SBUS cards (2 per server)

- Copper attachment with Media Interface Adapters (MIAs) to convert to multimode optical fibre
- Initial microcode level 11.0.9

2 Storage Technology Access Hubs

- 3 Multimode optical fibre interface cards with 8 ports each (24 ports total)
- Microcode revision 1.0.3

2 Storage Technology 9140/9145 Disk Subsystems

- 8 9140 JBOD Enclosures:
  - 10 18gb disks
- 1 9145 Disk/Storage System
  - 10 18gb disks per tray
- 2 Thor Storage Processors with 256MB of read and write cache

Software

- Fibre Channel drivers (Jaycor driver v1.22)
- Navisphere client array monitoring software

## **Initial Testing Configuration**

The initial testing configuration for the Sun/9145 configuration was meant to test basic connectivity, availability and performance.

Both Sun 1000's were taken off-line to test the basic capabilities of Fibre Channel and access of 9145 units. The following steps were taken:

- Install the Fibre Channel cards
- Install the Media Interface Adapters to convert from copper (a 9 pin plug) to multimode optical fibre.
- Install optical fibre from Sun 1000's to 9145's (direct attach)
- Bring up the systems
- Install the Fibre Channel drivers and related software
- Configure the sd.conf to access the additional LUNs within the 9145
- Reboot the systems with a reconfigure option to see the new arrays

Once the 9145 units were visible to the system as hard disks, basic testing was possible:

- Labelled the LUNs that the 9145 came with (LUN 0 and 4 assigned to SPA and SPB, respectively)
- Created a filesystem on the default LUNs
- Did performance testing to verify that the link was solid (standard I/O tests -- create files, delete files, etc.)

After some basic configuration problems (e.g. bad hardware, incorrect settings, etc.) were ironed out, the system accessed the 9145s as standard (RAID) disks.

## 9145 Internal Configuration

Once basic connectivity had been attained, it was now possible to do more useful testing. In particular, the 9145 comes with only two 5-disk RAID 5 LUNs within the unit, one assigned to SPA, one assigned to SPB. Any other disks configured within the unit haven't been configured. To configure the remaining storage, the Navisphere agent must be loaded and configured on the server, and the Navisphere Supervisor, a GUI front end running on a Windows 95 or NT system used to access and configure the 9145. After configuring the agent software and loading the Supervisor software on a PC, configuration of the remaining disks was possible.

The internal configuration of the 9145 unit consisted of the following hardware configuration:

- 1-9145 Unit with 10 internal 18gb drives
- 8-9140 Units with 10 internal 18gb drives

As described above, the default shipping configuration has the 10 drives within the 9145 bound as two 5 disk RAID-5 LUNs to allow basic system connectivity to the box. The Navisphere software requires some level of disk connectivity to the 9145 subsystem to communicate with the 9145; hence the shipment of the system with two LUNs bound, one to SPA and one to SPB.

With the default configuration, 10 drives of 90 have been allocated. Within the Navisphere Supervisor, the physical location of the remaining drives has little to do with the use of the remaining disks. In particular, the remaining 80 drives can be thought of as a column of drives 8 high, with 10 drives per row:

Figure 2:

|      |     |     |     |     |     |
|------|-----|-----|-----|-----|-----|
| 9140 | 18g | 18g | 18g | 18g | 18g |
| 9140 | 18g | 18g | 18g | 18g | 18g |
| 9140 | 18g | 18g | 18g | 18g | 18g |
| 9140 | 18g | 18g | 18g | 18g | 18g |
| 9140 | 18g | 18g | 18g | 18g | 18g |
| 9140 | 18g | 18g | 18g | 18g | 18g |
| 9140 | 18g | 18g | 18g | 18g | 18g |
| 9140 | 18g | 18g | 18g | 18g | 18g |
| 9140 | 18g | 18g | 18g | 18g | 18g |
| 9145 | 18g | 18g | 18g | 18g | 18g |

In the above figure, the 9145 unit consists of the actual 9145 unit, along with 8-9140 Just a Bunch Of Disks (JBOD) devices. In the default unit, the bottom row is split into two 5-disk RAID-5 LUNs. The rest of the array is undefined.

Through use of the Navisphere Supervisor agent, the remaining disks can be configured with any of the following RAID systems:

| <u>RAID Value</u> | <u>Short Description</u>                                        |
|-------------------|-----------------------------------------------------------------|
| RAID-0            | Concatenation                                                   |
| RAID-1            | Mirroring                                                       |
| RAID-3            | Write Performance Optimized with redundancy (similar to RAID-5) |
| RAID-5            | Striped Redundancy                                              |
| RAID 0/1          | Concatenation with Mirroring                                    |
| RAID 1/0          | Mirroring with Concatenation                                    |

The discussion of the attributes, costs, and related configuration of the different RAID values are beyond the scope of this document.

For cost efficiency, RAID-5 was chosen as the underlying hardware redundancy system.

The 9145 unit can configure RAID-5 in any configuration of 5 through 16 disks. The manual recommends two configurations, a 5-drive RAID LUN for high-transaction environments, and a 9-drive RAID LUN for larger I/O patterns.

To test the different RAID-5 drive configurations, a number of RAID-5 LUNs were configured within the array:

| <u>Number of Disks</u> | <u>Disk Stripe</u>        | <u>RAID Stripe Size (in sectors)</u> |
|------------------------|---------------------------|--------------------------------------|
| 5                      | Across 5-9140 units       | 128                                  |
| 5                      | Within a single 9140 unit | 128                                  |
| 6                      | Within a single 9140 unit | 128                                  |
| 7                      | Within a single 9140 unit | 128                                  |
| 8                      | Within a single 9140 unit | 128                                  |
| 9                      | Within a single 9140 unit | 128                                  |
| 10                     | Within a single 9140 unit | 128                                  |

In addition, the internal 9145 caching parameters were set to the following values:

256MB Read  
256MB Write

It has been shown in other testing that read caching can be less than effective in open systems. In particular, open systems use internal memory for disk caching, and generally will yield the best throughput without the need for hardware read caching.

In this situation, however, read caching could help in the following two situations:

- Veritas Filesystem uses an extent based allocation method with a transaction log. This results in a small region (~3mb) of disk being used at the beginning of the disk for all I/O.
- Filesystem inode updates are, by necessity, a synchronous operation. Any file update must go to disk in some fashion for the I/O to complete.

To be truly effective, locking of the above information into the read cache would be ideal. However, since there doesn't seem to be a way to lock that information into the read cache (only ongoing I/O will determine the contents of the cache), the read cache will be a hit-or-miss proposition.

Performance testing brought out the following results:

- Striping across multiple 9140 units and within a single 9140 unit were identical
- Sun Sparc 1000s are not fast enough to drive the 9145 effectively, with little measured difference between 5 through 10 disk LUNs.

Since the above tests showed that the Sun 1000s were incapable of exercising the 9145 effectively, further performance testing was restricted to failure-related testing. In particular, performance testing was used to validate that the systems could handle high I/O rates for extended periods of time.

Since the Sun 1000s were unable to drive the 9145s to show different performance characteristics, the layout decision fell to the following performance parameters:

- The more disks found in a RAID-5 grouping, the less data space (and resulting cost) lost for RAID. In a 5 Disk RAID-5 group, a single disk's capacity is used for parity storage (no single disk contains the parity), resulting in a 20% storage loss. In a 10 disk -5RAID group, a single disk's capacity used for parity storage results in a 10% storage loss.
- The NFS servers have Sun ATM 622Mbit/sec interfaces, with a maximum theoretical throughput of 40MB/s.

With the above in mind, performance of the 9145 subsystem should be sufficient in almost every case to allow throughput of up to 40MB per second per server. Due to this, the remaining disks within the 9145 were configured with the following numbers:

- The original LUNs (0 and 4) remain, but LUN 4 was rebound as LUN 1, resulting in LUN 0 being SPA, and LUN 1 being SPB. This allows all even LUNs to be owned by SPA, and all odd LUNs to be owned by SPB.
- 7 10-disk RAID-5 groups
- 1 9-disk RAID-5 groups
- 1 Hot Spare

The hot spare and the 9 disk RAID group are co-resident within a single 9140 enclosure.

The resulting space available to the system (prior to filesystems being placed on the system):

| <u>Number of Disks</u> | <u>Size</u>      |
|------------------------|------------------|
| 5                      | 66gb (4x16.5gb)  |
| 9                      | 132gb (8x16.5gb) |
| 10                     | 148gb (9x16.5gb) |

Total space per cabinet: 1300gb (1.25tb)

## Access Hub Configuration

After basic testing with the 9145's directly attached to the Sun 1000 servers, a pair of Storage Technology Access Hubs were added to the configuration. Prior to the Access Hubs being available, the fibre channel configuration of the servers chained through the 9145 storage processors similarly to the previous SCSI configuration. The introduction of the hubs allows for more redundancy and flexibility.

The Storage Technology Access Hub operates similar to a dynamic ethernet hub. It allows configuration of any of the ports within the hub (which may consist of copper, multimode optical fiber or singlemode optical fiber) as participants in multiple Fibre Channel arbitrated loops. This has the following features:

- Direct attach has no redundancy for link (cable) failure. Should a cable fail, further I/O on the loop is unpredictable, similar to standard SCSI. The Access Hub has a feature that allows the logical loop to 'heal' itself in the event of a link failure. It will take approximately 3 seconds to heal the loop in the event of a failure, resulting in a small outage on the link (and a broadcast to the members on the loop that a unit has failed).
- The Access Hub uses a standard 10Mbit ethernet connection for configuration connectivity. The included Web-based configuration capabilities through the ethernet connection allow for incredible flexibility with regard to loop changes.
- The Access Hub, through using system-provided tools, can be reconfigured dynamically to change the loop structure. This would allow, for instance, a group of tape drives to be added to the loop to backup the system, removing after the backups had finished, while the system continued to run. Also, additional storage can be added to a loop to expand server storage capacity, requiring only a reboot to access.

With the above in mind, the Sun 1000s were cabled to two hubs, one for each Storage Processor. See figure 3 for a detailed drawing of the configuration.

Two Access Hubs (rather than one) were used for the following reasons:

- Two hubs would be more resistant to failure. Should a single hub fail, it should be possible to use the other hub to access all attached storage.
- The number of arrays ordered for Information Systems (IS) storage totaled 7 cabinets, with 20 storage processors total. The Sun systems to be cabled to the 9145 units would initially be taking 12 connections, minimum. This results in a total of 32 connections, completely filling a single hub, leaving no room for growth.

After initial configuration, the hubs were found to be running microcode version 1.0.1. After upgrading the hubs to microcode version 1.0.3, initial configuration testing could begin.

A loop within each hub was built to attach to all of the NFS 9145 units. The Sun 1000s were configured to access the arrays, and long-running I/O started to test the reliability of the loop.

The systems were configured with the following `/etc/system` changes:

```
set      maxphys=8388608
set      fcaw:force32=1
```

The `maxphys` configuration value can be found in the Jaycor readme (from their website, <http://www.jni.com>). The `force32` option refers to 64bit SBUS Jaycor cards, which were not being used at the time, but necessary for 64bit SBUS card operation in a 32bit SBUS server.

In addition to the above settings, the `sd.conf` file had to be configured to access the additional LUNs within the 9145 units. By default the system looks at only LUN 0, and must be configured to look for any additional LUNs.

A number of long term performance issues were found:

- The microcode on the card tested was version 1.0.9. After upgrading to 1.0.11, some problems went away.
- A bad Media Interface Adapter (MIA), was causing link on one of the loops to go up and down. After replacement, the loop operated properly. This was diagnosed from the console messages indicating loop up/down a number of seconds apart.
- The Jaycor card uses a daughterboard on the actual card. Some of these were found to be loose, causing intermittent problems. After reseating the daughterboard firmly, performance problems were no longer seen.

It should be noted that the Access Hubs had no evident problems. In fact, they were instrumental in diagnosing a microcode problem on the production NFS servers.

## Application Transparent Failover Configuration

As part of a high availability configuration, Storage Technology recommends the use of Application Transparent Failover (ATF) for Storage Processor redundancy. ATF is a software package that operates between the SCSI driver and the 91XX arrays to allow for a Storage Processor failure.

On the existing NFS cluster, the 9137 units experienced many Storage Processor failures. In almost all cases, this was due to memory problems within the 9137. The 9137s attached to the NFS servers were not installed with Error Correcting (ECC) Memory, which caused many failures when a memory problem was found. The ATF software, in the event of a Storage Processor failure, will transfer all I/O (and LUNs) from the failed SP to the (hopefully still) working SP, allowing the system to continue to operate without interruption.

However, a few problems with ATF on SCSI were found:

- Any failures were reported to the system via the syslog. However, since ATF doesn't use a standard controller, target, LUN, slice numbering, identifying the failed component was difficult.
- A number of mysterious problems were found with ATF. In particular, some supposedly safe commands (*atf\_restore\_all*, for example) caused the system to stop doing I/O at times, effectively causing a system outage rather than rebalancing the I/O.
- Any reconfiguration reboots with failed components caused ATF to be in an unbalanced configuration, resulting in potential unrecoverable failures. In particular, should an SP be unavailable, ATF will not acknowledge the SP during a reconfiguration reboot, resulting in an array with a single SP (and single point of failure). Should the remaining SP within that array fail, the system would be left with no way to access the LUNs.
- ATF represented a single-vendor HA solution. Configuring additional non-91XX disk on a system was difficult, as the system would see the disk, but ATF would not. After some number of *atf\_reconfigure* and reconfiguration reboots, the disks would then be available to the system. This complexity caused many problems with periodic maintenance.

The SCSI NFS servers were using version 1.2.6 of ATF. The 9145 product requires version 1.3.0, which supports both SCSI and Fibre Channel attach 91XX storage. This version was installed on the Sun 1000s for testing.

After installation and reconfiguration via *atf\_reconfigure* (see the ATF installation manual for a detailed installation process), the following tests were performed on the 9145 units, split through the Access Hubs.

| <u>Test</u>                                            | <u>Successful?</u> |
|--------------------------------------------------------|--------------------|
| Verification of all SPs seen by ATF.                   | Yes                |
| Basic I/O to the disk.                                 | Yes                |
| Test of a single SP failure.                           | Yes                |
| Test of a staggered, dual SP failure (5 minutes apart) | No                 |

On the final test, the test attempts to emulate a SP failure and reboot, and a resulting SP failure and reboot some time later. This emulates a SP failure during the night, or on a weekend, and a secondary SP failure at some later time. If the software can allow for this fault, a single SP failure will not be deemed a critical failure, and can be worked during business hours. On the previous version of ATF, version 1.2.6, a staggered SP failure would not cause an outage. ATF would mark the SP as failed, rerouting the I/O to the still-working SP. Once the original failed SP had rebooted and was working again (generally about 2 minutes later), the other SP could fail, and ATF would reroute the I/O back to the originally failed SP.

In testing version 1.3.0, ATF could not recover in the final test. When the second SP had failed (after waiting 5 minutes for the SP to become available), the system then started getting I/O errors to the disk, resulting in I/O corruption, and Veritas Volume Manager marking the disk as failed.

In addition to the above testing, a fundamental problem was found between ATF 1.3.0 and Veritas. In particular, when installing ATF, an error message about "Can't read symbolic link /dev/vx/dsk/rootvol" was noticed. When deinstalling ATF, it removed the rootvol link, resulting in a non-bootable system. The system was recoverable by booting from CDROM and changing the rootvol mount to a standard disk mount. However, this would not be an acceptable failure on a production system.

Due to the problems seen with the previous version of ATF, and the fundamental problem between ATF and Veritas, ATF was removed from the configuration. This results in a higher potential for failure, but due to the problems found with ATF, it was deemed a lower potential than the failures found through use of ATF.

## Integration Testing Configuration

The initial test configuration represents a smaller version of the current NFS SCSI configuration. The test systems split the capacity of the single 9131 array, with fail-over representing a movement of the LUNs within the array between systems. This configuration was tested prior to installing any Fibre Channel hardware or software to verify that the system worked similarly to the existing NFS configuration.

After initial testing of the SCSI configuration, the two systems were failed to a single system to represent a real-world test of upgrading one server while the other remained on-line.

At that point, the following configuration steps were performed with I/O running to the systems at all times:

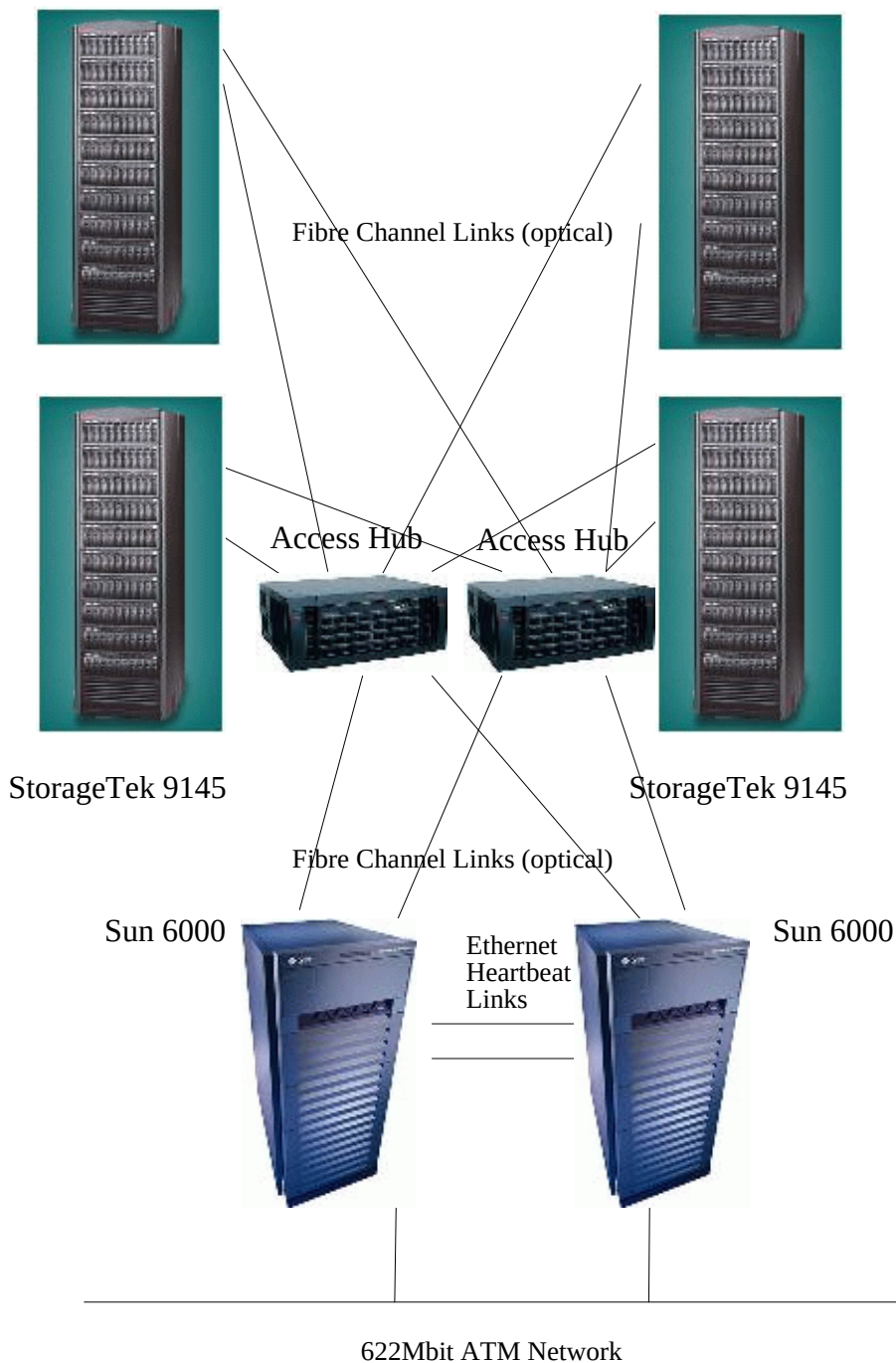
1. Installed Fibre Channel cards and software.
2. Configured the system to access the additional LUNs via `/kernel/drv/sd.conf`
3. Reconfiguration booted the system (boot -r)
4. Verified all 9145 LUNs were visible to the system.
5. Failed the other system over to the newly booted system.
6. Followed the above steps on the other system.
7. Moved the systems to on-line primary status. At this point, all of the LUNs (both SCSI and Fibre Channel) are available to both systems.
8. Added the 9145 LUNs to the existing disk groups.
9. Moved the data between 9131 (SCSI) and 9145 (Fibre Channel) disks live via Veritas Volume Manager.
10. Decabled the 9131 (SCSI) array and removed the SCSI cards after failing the systems alternatively.

It should be noted that Veritas Volume Manager (VxVM) has an excellent capability that eased the reconfiguration of the NFS servers. NFS clients access NFS volumes through a standard network via the major and minor numbers of the volumes. The NFS client uses those major and minor numbers to determine the validity of the remotely mounted disk. If the disk shared from the NFS server were to change those numbers by moving the shared disk between SCSI interfaces, the client then marks the disk as unavailable. The client shows this as a "Stale NFS File Handle", which is a relatively common problem in NFS environments, and generally requires a reboot of the client to fix.

Veritas Volume Manager abstracts the major and minor numbers with the system through two drivers, vxspec and vxio. This allows all of the disks on a NFS server to be reconfigured dynamically with no risk to the client of "Stale NFS File Handle".

Since VxVM allows the disks to move dynamically without client disk issues, moving data between the 9131 and the 9145 can be seamless, and may disregard the pathing associated with any disk connected to the system.

Figure 3:



## Production NFS Server Integration

The production NFS servers were almost out of disk storage. Beyond that, the 9137 SCSI raid arrays had already been allocated to other production machines which were out of disk storage. To say the least, the Fibre Channel storage was needed on the NFS servers as soon as possible.

On a Saturday, the NFS servers were taken off-line to install the fibre channel interface cards, and add the fibre channel storage. The test process was used, with a small difference in the case of the 7th array. The NFS servers had only a

single SBUS card slot available. Two card slots were required for SPA and SPB accessibility. The 7th array was converted from a SCSI channel per SP to a single SCSI channel for both SPs by chaining the two SPs together. The failure potential was raised by this chain, but for the short term, deemed acceptable.

After returning both servers to Online Primary, data copying was started via Veritas Volume Manager. All of the storage within the standard arrays was mirrored to the 9145 units sequentially.

Almost immediately, the NFS servers had problems. The I/O ran between the SCSI and Fibre Channel for approximately 10 hours before the fibre channel interface on one of the servers started having problems. The Fibre Channel card started showing the following errors:

```
unix: fca1: fca1: fca_cmd_timeout: EXPIRING cmd 0x645f8948 target:0 LUN:2
cmd_flags:0x20244 state:0x18 tgt_state:0x11
```

Some of these errors had been seen when hardware problems were found on the Sparc 1000s. After a week of troubleshooting, it was found that the cards received from manufacturing had contained cards with different microcode levels. The microcode used in test was level 11.0.9, while the microcode level in production (through dumb luck) was revision 10.0.

During this time, the test 1000s were used for ongoing performance testing. No problems were found.

Once the downlevel microcode was found, the systems were taken off-line to upgrade the microcode in the cards. The cards were upgraded to 11.0.11 microcode. However, the upgrade wasn't very successful. One of the cards failed to upgrade, and would no longer function. It appears that the microcode upgrade system, downloaded from Jaycor, has a problem with large servers. All further upgrades were done on a single processor Sparc 20, with excellent success.

When one of the cards failed to upgrade, the cards used in testing were removed from the Sparc 1000s and installed into the production machines.

From that point forward, no further microcode errors were seen. However, one performance issue arrived. During mirroring between the 9137 and the 9145 units, a number of mirrors were built at the same time (up to 8 at a time), resulting in a large amount of I/O to the 9145 units (approximately 4mb per second per mirror). It was found that multiple mirrors to a single LUN within the 9145 would saturate at a particular point, causing the system to see communication problems. It hasn't been determined whether the LUN or the microcode, patches, etc. were at fault. Mirroring was then scaled back to a single mirror per LUN. The total mirroring took approximately 10 days to complete. Note that the mirrors were initiated on an 'as available' time period. All new mirrors were initiated during regular business hours.

Once all of the data had been mirrored between the SCSI attach storage and the Fibre Channel storage, the SCSI mirror was then deleted, resulting in all data residing on the 9145s.

The servers were then taken down, similar to the production move, to remove the SCSI attached storage, and the unused SCSI hardware.

Since that time, the servers have had no Fibre Channel problems. They have been running 7x24 for a number of weeks.

## Future Direction

The NFS servers represent the first production SAN implemented within Storage Technology. Since this was the first implementation, a number of remaining items exist for future SAN connectivity:

- Jaycor 32bit SBUS cards are no longer being manufactured. Any future SAN connectivity will require the use of 64bit SBUS cards rather than 32bit.
- A true SAN should (theoretically) entail connectivity between any server with any Fibre Channel disk or tape product. This implementation doesn't attempt to use this capability, as a fundamental capability, Fabric Login, isn't universally available. Until Fabric Login arrives in all products (and they communicate with one another properly), true SAN connectivity will not be realized.

- Only SCSI protocols have been used in this system. SCSI and IP protocols would be advantageous for backing up the system across Fibre Channel. This would offload the large backups from the existing TCP/IP (ATM) network, allowing that network a longer useful life in its current form.
- This solution used only SBUS cards on Sun servers. Future connectivity should include PCI cards on both Sun and NT servers for more data transparency.

Beyond the above, one of the goals of the NFS Fibre Channel configuration involved the split of the NFS servers between physically distant buildings for disaster recovery purposes. In this configuration, a physical disaster impacting a single datacenter would not impact availability, as the data would be mirrored between the buildings, allowing the remaining server to continue in the absence of the other server.



