Early Experiences with Storage Area Networks and CXFS

John Lynch
Aerojet
6304 Spine Road
Boulder CO 80516

Abstract

This paper looks at the design, integration and application issues involved in deploying an early access, very large, and highly available storage area network. Covered are topics from filesystem failover, issues regarding numbers of nodes in a cluster, and using leading edge solutions to solve complex issues in a real-time data processing network.

1 Introduction

Aerojet designed and installed a highly available, large scale Storage Area Network over spring of 2000. This system due to it size and diversity is known to be one of a kind and is currently not offered by SGI, but would serve as a prototype system.

The project's goal was to evaluate Fibre Channel and SAN technology for its benefits and applicability in a second-generation, real-time data processing network. SAN technology seemed to be the technology of the future to replace the traditional SCSI solution.

The approach was to conduct and evaluation of SAN technology as a viable replacement for SCSI, consolidate redundant data in the system and overcome limitations of traditional file sharing mechanisms. Paramount in the design criteria was to carefully balance the risk associated with deploying a large-scale SAN prototype system in a mission critical production environment.

SAN technology can be categorized in two distinct approaches. Both approaches use the storage area network to provide access to multiple storage devices at the same time by one or multiple hosts. The difference is how the storage devices are accessed.

The most common approach allows the hosts to access the storage devices across the storage area network but filesystems are not shared. This allows either a single host to stripe data across a greater number of storage controllers, or to share storage controllers among several systems. This essentially breaks up a large storage system into smaller distinct pieces, but allows for the cost-sharing of the most expensive component, the storage controller.

The second approach allows for all storage devices to be accessed by all hosts on the storage area network and also filesystem sharing across multiple hosts. SGI designed a shared filesytem, Clustered XFS, (CXFS), to allow all of the systems in a cluster to have access to the same filesystem. However, they also

have access to it a full fibre channel performance as if it was direct-atttached.

To date, the industry trends tend to be on the share-hardware only approach. SGI is clearly the industry leader in the sharing of data with cluster filesystems, in addition to the sharing of hardware.

The system design uses elements from both approaches but really takes advantage of the CXFS approach offered by SGI.

2 Hardware Design

2.1 Requirements

During the design phase of the system some general requirements were levied on the SAN technology.

In general, the access speed from server to disk had to meet the performance of direct-attached SCSI. This was not viewed as difficult since 1Gbit Fibre Channel is over twice times the transfer rate as Ultra SCSI (40 MB/s).

No failure in any one server or workstation can affect the operation of another.

A single failure in the SAN fabric cannot create a system outage.

The clustered filesystem private network had to work with the existing network infrastructure.

The SAN system had to be economically justifiable to warrant the additional cost of employing the system.

The SAN system needed to work with existing tape storage devices. The program has compatibility requirements with other programs that have legacy Exabyte 8mm SCSI tape drives.

2.2 Design Methodology

The system design consists of four different Storage Area Networks. Due to the unique mission of each SAN, and the current port limitations on switches, design decisions were made on the makeup of the SAN architecture to limit the risk of failure in the implementation phase.

All of the SAN fabrics are based on the same building block principles that are outlined here.

In the original system, the Fast-Wide SCSI solution was replaced with the direct-attach fibre channel RAID.

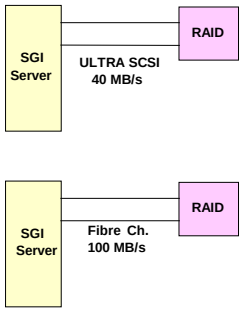

Figure 1: SCSI Replacement with Fibre Channel

A single Brocade Silkworm 2800 fibre channel switch was built into the system, connecting the server and the storage.

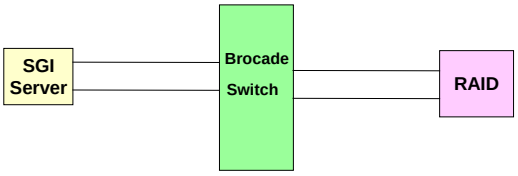

Figure 2: Addition of Brocade Silkworm Switch

After the functionality was verified, the addition of a SCSI Exabyte 8mm tape drives to the system through a Crossroads fibre channel to SCSI router.

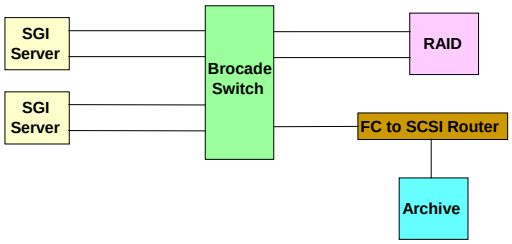

Figure 3: Tape Drive Added

A fibre channel to SCSI router is a device that converts fibre channel protocol to SCSI protocol and vice versa. It is used to integrate legacy SCCI devices into a new generation fibre channel architecture.

Host bus adapters (HBAs) were then added to the system to eliminate single point of failures in accessing the storage devices..

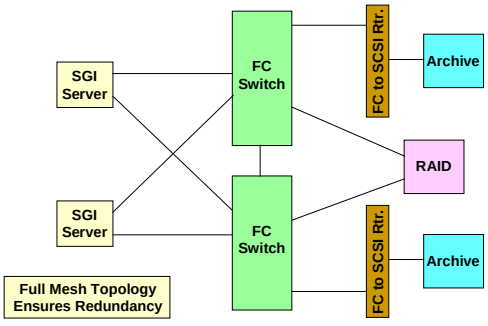

Figure 4: Full Redundant Architecture

The technique of gradual insertion of technology enabled the smooth integration of fibre channel technology into an existing architecture. All four fabrics are based on are based on these building blocks.

2.3 Final SAN Fabric Design

Fabric One is used by the servers to hold live satellite downlink data, common data that is used by all servers such as satellite ephemeris, common maps and imagery, a single copy of custom ground station software.

This fabric has two 32 processor, five 16 processor, and one 4 processor Origin 2000 servers accessing data on eight different RAID devices each containing two storage processors. Total storage capacity is 2.3TB.

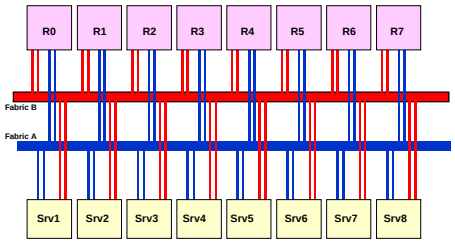

Figure 5: Fabric One Layout

Fabric Two is used by the Sybase servers. The operational databases that are used to store satellite trending data as well as ground station configuration databases are stored here.

This fabric has two eight processor Origin 2000 servers accessing data on four different RAID devices each

containing two storage processors. Total storage capacity is 512MB.

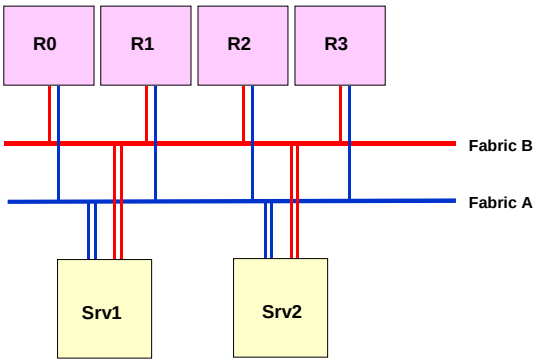

Figure 6: Fabric Two Layout

Fabric Three consists of all the workstations and storage that contains common data. The common data used by all workstations contains satellite ephemeris, common maps, our custom ground station software, and IRIX auditing data.

This fabric has thirty-seven 4 processor Onyx2 workstations accessing data on three RAID devices each containing two storage processors. Total storage capacity is 384MB.

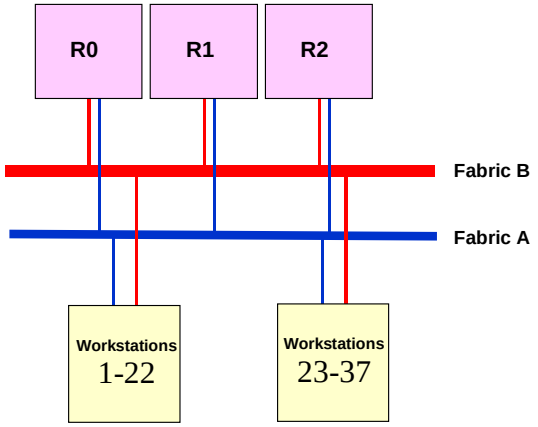

Figure 7: Fabric Three Layout

Fabric Four is used by intelligence operators. It contains all the large-scale imagery data used to support a real-time missile-warning mission.

This fabric has two 4 processor Onyx2 workstations accessing data on a single RAID device containing two storage processors. Total storage capacity is 1TB.

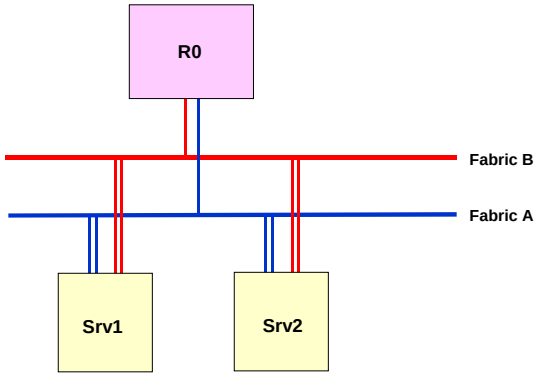

Figure 8: Fabric Four Layout

This SAN architecture was designed to enhance the existing system by reducing the number of software images that needed to be installed, consolidating the system audit records into a single location. This also reduces the dependency on NFS as a mechanism to share files among the servers.

3 Reliability

To ensure maximum reliability, each Origin 2000 server is equipped with dual Adaptec Fibre Channel Host Bus Adapters (HBAs). This will ensure that each server has a redundant path to the storage arrays. Each server is connected to a primary fibre channel fabric (fabric A) and to a completely separate secondary fibre channel fabric (Fabric B). Each fabric consists of Brocade

2800 fibre channel switches with redundant power supplies.

Media Interface Adapters (MIAs) that were installed on a copper RAID storage processor to provide an optical connection to the server were removed. The fibre channel switches were outfitted with both copper and optical interfaces. The optical interfaces were used in connecting all the Origin 2000's and Onyx2's to the switches. The copper interfaces connect the switches to the RAID storage processors. Removing the MIAs increased the reliability of the connection in eliminating a high failure rate item.

The two fabrics are completely isolated from each other. The RAID devices are SGI Fibre Channel RAID. Each contains two Thor storage processors, one connecting to the primary fabric and the other connecting to the secondary fabric. A typical fabric consists of four Brocade 2800 switches. These switches are connected to each other using three Inter-switch Links (ISL). The data most likely accessed by a particular server or workstation was configured on the same switch as that server or workstation, further reducing latency. It is uncommon to have ISL traffic in this configuration.

This configuration provides the most redundancy for path diversity from server to storage. Any failure in any HBA, switch, or interconnection, will cause the system to automatically find an alternate path through the redundant components. A failure in either of the two storage processors in a RAID unit will automatically trespass the LUNs from one storage processor to another. Data availability is far more important than I/O performance for this application. All LUNs within the storage system are configured as RAID level 5.

Normal access rate to a single RAID is 80MB/sec for each storage processor, with two storage processors capable of delivering 160MB/sec. In the event of a failure in one of the storage processors, the system is capable of maintaining access to the storage, but at a degraded state of all of the LUNs being supported by the surviving storage processor.

Initial implementation of the SGI's Cluster File System, also known as CXFS, was using IRIX 6.5.6f. It was in its infancy and cluster stability was an issue. The cluster database would become corrupt, get replicated across the other members in the cluster and the entire cluster would become unusable. Since upgrading to IRIX 6.5.10f with patch 4156, significant improvements have been made to cluster stability.

4 Integration with SGI

From the very early stages in the design process, Aerojet and SGI were working hand in hand to come up with a solution. Program requirements for the Storage Area Network exceeded SGI's testing and supported configurations. Aerojet and SGI collaborated to design and develop a solution that successfully balanced the requirements for the system against the risk of using an unsupportable configuration.

The CXFS engineering team was very supportive of this effort. Early on SGI was eager to develop any IRIX patches to aid on problem resolution. It was through close cooperation that SGI and

Aerojet was able to make great strides in larger SAN architectures.

5 Performance

5.1 Processor Utilization

Program requirements dictate that a 100 percent processing margin must be maintained on the systems during normal operations. Aerojet was concerned about the additional system overhead incurred by the introduction of CXFS and SAN technology to the system. To mitigate the concern Aerojet measured the impact of running CXFS on the system to ensure the processing margin requirement was not compromised.

Performance Co-Pilot was used to measure the processor utilization of each of the CXFS processes. The processes measured were the cluster configuration daemon (clconfd), cluster reset daemon (crsd), cluster administration daemon (cad), cluster monitor daemon (cmond), and the cluster database (cdb). The objective was to monitor the processor utilization two states of the cluster, the idle and all nodes in the cluster writing data simultaneously.

Initial measurements were taken with the cluster in an idle condition. Idle condition is defined as the cluster is up and all nodes are reporting green status. In this condition no nodes are accessing the clustered filesystem.

Performance Co-pilot was recording all data during the testing for later analysis. Figure 9 illustrates the processor utilization of the CXFS daemons during the idle condition.

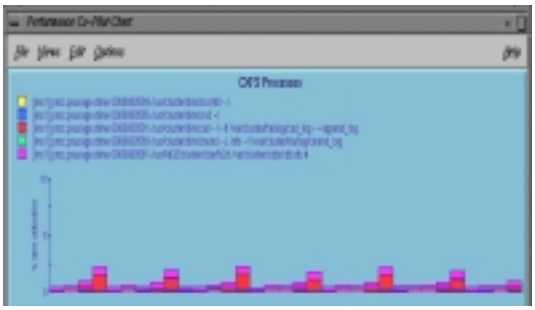

Figure 9: CPU Utilization (Cluster Idle)

Observations during the test indicated that the most active process was the Cluster Administration daemon (cad). The cluster admin daemon shown in red used 2.5 percent of the total number CPU's. The test system was an Origin2000 with 32 R12K 250Mhz processors. Translating to CPU usage, in the idle condition the cad uses .8 of a single CPU (32 * .025).

Also observing the above graph reveals the heartbeat message the cad sends out to all other nodes at 8-second intervals. The heartbeat message is used to query other node to test the state of the CXFS services, whether they are active, down, or unknown.

Comparisons were made against the idle condition to test if increases in cluster activity increased the cpu utilization.

The test cluster was the server cluster with eight nodes. This cluster was chosen because the amount of cluster data is the greatest. According to SGI the most stressing scenario on the cluster is when multiple nodes are writing data to the cluster. The more data and nodes that write data the more metadata has to be generated and sent to all other nodes.

Figure 10 shows the test cluster with all eight nodes writing system audit records on to the clustered filesystem simultaneously. To further increase the stress on the system, all nodes wrote to the same filesystem on the cluster.

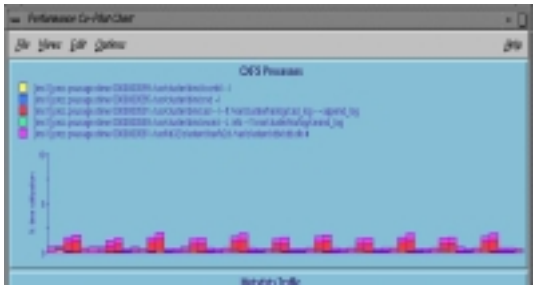

Figure 10: CPU Utilization (Cluster Write)

Results were surprising in that the cpu utilization was unaffected by cluster activity.

It seems from our testing that the processes required for CXFS are static in their use of system resources. Additional testing will need to be undertaken to validate this observation on different cluster sizes and different node types.

5.2 Metadata Network Traffic

Concerns were also raised with the capacity of the metadata network. SGI initially recommended a private dedicated for the metadata traffic. The system installed and utilized a 100 Mb/s switched Ethernet system. The metadata was isolated on a different subnet from the rest of our mission data. System requirements dictated a 150 percent margin on all networks. This meant that the metadata network could not use more than 40 Mb/sec of bandwidth.

Using Performance Co-pilot, the amount of metadata traffic was measured in our two test conditions.

Figure 11 shows the network traffic in bytes/sec on the metadata network.

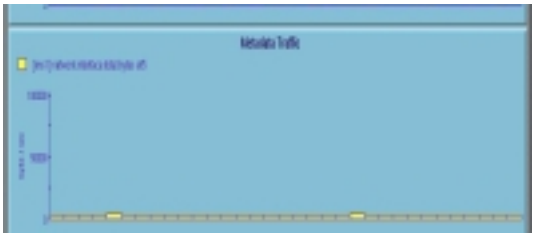

Figure 11: Metadata Network (Cluster Idle)

Measurements indicated virtual no metadata traffic. This was expected due to the cluster being in an idle mode.

Figure 12 shows the network traffic in bytes/sec on the metadata network during the max write mode.

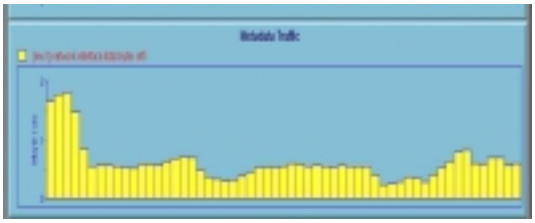

Figure 12: Metadata Network (Cluster Write)

In the max write mode, a lot of metadata is generated due to the new files being written. During this test system audit records were being written to the same filesystem on the cluster. Measurements indicate that the metadata traffic consumed a peak of approximately 2 MB/sec and average bandwidth of 1MB/sec.

Initial results are extremely promising. Additional testing on larger clusters is required to reveal if more nodes increase the bandwidth required in a linear fashion, i.e., double the number on nodes and bandwidth required doubles.

5.3 Filesystem Failover

The time it takes to recover from a failure in the storage area network is a great concern. On the original system, filesystem failover was not an option. Since SCSI disk were used, a failure in the path caused the entire server to crash.

Using the SAN, the ability to sustain a failure in the path to disk and recover was very useful. Filesystem failover works when and I/O error is detected in trying to access the storage device. The server builds a table of possible paths to a target. There is two was of controlling this list and failover priority. One is to allow the operating system to determine the paths possible and upon failure try alternate paths to disk. The other option is to use the failover.conf file. This file sets up the path definition to the targets for the node. It also controls the priority in which the alternate paths are selected.

During the initial configuration of the system, the operating system was chosen as the method to control the failover. This was chosen over the failover.conf file due to its simplicity, and each node had no more than two paths to each target.

Initially, failover time was measured at 4 minutes. This is the default time that is preconfigured with CXFS. The system required a better failover time. The IRIX kernel was tuned to lower the time to failover from 4 minutes to 15 seconds. 15 seconds was chosen as the time for the system to prevent possible filesystem failover bounce. This is defined in an intermittent path to disk that caused constant failover of the filesystems.

Recovering from the failover is a task in a large system. If the path fails from the fibre channel switch to the target causes all nodes in the cluster to failover to an alternate path. A large cluster with a lot of nodes, each node has to be failed over by hand by issuing a scsifo –t <current path>.

A workaround solution was found on the SGI RAID controllers. Using the IRIX Fibre Channel RAID software GUI, it is possible to enter a special engineering mode. This mode is not documented and only appears to work with this brand of RAID.

While in the engineering mode, the GUI will allow the administrator to drag and drop the LUNs onto the correct storage processor. Once this is performed, the storage processors perform a LUN trespass and cause all node to failover to the new path. Without this ability, it normally took about 30 minutes to failover all the nodes in the cluster one at a time.

Improvements need to be made in allowing easier failover recovery from a single location. As clusters become bigger and bigger, recovery time will be just as important as failover time.

6 Trials and Tribulations

6.1 Network Connectivity

During the early integration phase some problems with CXFS were apparent. CXFS was designed to be very efficient in it's communication with other nodes. Originally CXFS only used multicast to distribute metadata between nodes in the cluster. Since the multicast network traffic does not adhere to RFC 1112 it was not routable between nodes on different subnets. Since our network has various subnets and need to be routed, another solution had to be implemented. After consultation with SGI, the solution was to make CXFS a little smarter in its communication process. A patch to IRIX resulted in that now if the nodes are on the same network then CXFS will use Multicast to communicate between nodes, and if the nodes are on different networks CXFS will establish a TCP connection. Once this was implemented all nodes then could begin communicating with each other.

On one of the two node clusters, the network infrastructure is two Enterasys Smartswitch 2200's with layer 3 services enabled. In this architecture, the two nodes could not communicate with each other. Each node is on the same network and CXFS uses Multicast to communicate with each other. But due to the fact that CXFS does not follow RFC 1112 for multicast traffic the network switches was not allowing the traffic to be propagated through the network. In this case, since Layer 3 services were not required for the system to operate, we turned off Layer 3 services and full communication was restored.

6.2 Cluster Size

Another problem is with clusters that contain only two nodes. Depending on how the nodes are weighted, set a tiebreaker node, or use the heartbeat/reset cable the cluster will behave differently.

During the design phase it was not recommended that the hardware/reset cable not be used. Both nodes in the two node cluster are weighted the same. Similar weighting was chosen to keep the quorum calculation simpler and does not really have an affect on a two-node cluster. A tiebreaker node was set to one of the nodes. Below is an explanation of the operation in a two-node cluster.

6.2.1 CXFS Cluster Membership Quorum

There can only be one active metadata server per filesystem. However, a problem can develop of the heartbeat/control network (which is used to transport metadata information between clients and the metadata server) has trouble. If the heartbeat/control network is somehow split in half (for example, due to a network failure), the network can become two smaller networks (segments). If this happens, it is necessary to ensure that only one metadata server is writing the metadata portion of the CXFS filesystem over the storage area network. To facilitate this, the concept of a CXFS cluster membership quorum of possible metadata servers is introduced.

6.2.2 Quorum Calculation and Node Weights

The quorum is calculated on the combined weight of nodes attempting to participate in the membership compared to the total weight of all nodes defined in the cluster.

Nodes have a default weight of 1. When all nodes have a weight of 1, an initial quorum calculation essentially becomes a majority of the number of nodes. In cases where a cluster consists of one or more CXFS metadata servers and multiple CXFS client-only nodes, one may want to configure the node weights, such that the quorum consists of only the possible metadata servers, regardless of the state of the clients. In this case, the metadata servers should have a node weight of 1, while the clients are weighted 0, the clients will be unable to form a membership by themselves. In this scenario, a membership will always require a quorum of metadata servers.

Note: At least one node must have a membership weight greater than 0. All potential metadata servers must have a membership weight greater than 0. Membership weight values other than 0 or 1 are not recommended.

For the initial CXFS cluster membership, a quorum requires more than 50% of the total defined weight for the cluster. For an existing membership, a quorum requires 50% of total weight of all nodes defined in the cluster.

Note: The tiebreaker node must have weight greater than 0 so that it can be a metadata server.

For a two-node cluster, if the hardware reset cable is not used a tiebreaker node should be set to avoid a network partition. However, if the tiebreaker node in a two-node cluster fails or if the administrator stops CXFS services, the other node will do a forced shutdown, which unmounts all CXFS filesystems.

When nodes that are potential metadata servers are given a weight of 1 and nodes that are only clients are given a weight of 0, a majority of nodes each with a configured weight of 1 must be accessible. If the network being used for CXFS heartbeat/control is divided in half, only the portion of the network that has the quorum of metadata servers and the tie-breaker node will remain in the cluster. Nodes on any portion of the heartbeat/control network that is not part of the quorum will shutdown from the cluster. Therefore, if the CXFS heartbeat/control network is cut in half, a metadata server will not be active on each half of the network trying to access the same CXFS metadata over the storage area network at the same time.

The membership monitors itself with normal messaging and heartbeating. If a failure occurs, the offending nodes are removed from the membership or are reset to prevent further access to shared resources by these nodes. A node that discovers it has been eliminated from the membership (due to a communications failure) will forcibly terminate access to the shared disks and stop CXFS services, if it is not first reset.

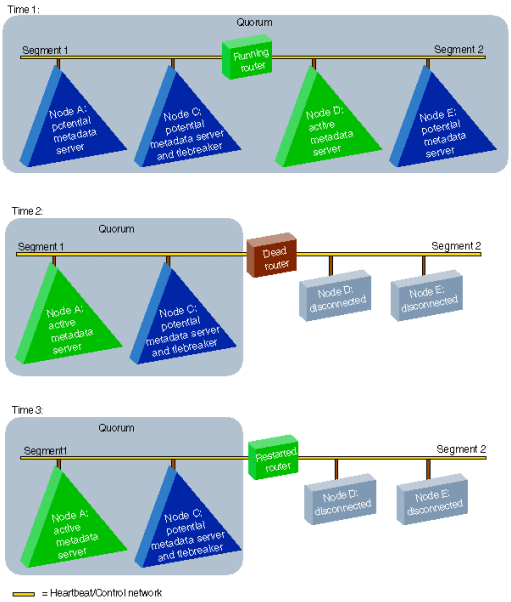

Figure 13: Changes in Quorum and Active Metadata Server due to Network Partitioning

Possible solutions to the two-node cluster that are being explored are to add a third machine to the cluster. This machine does not really do anything but raise the quorum so that one machine failing will not cause the entire cluster to die. Another possibility is to use two Origin200's as metadata servers, and use the remaining nodes as clients only. It is thought that because the Origin200's are only running IRIX and CXFS and not running user code they will be extremely stable and very reliable.

6.3 Hanging Nodes

Occasionally, nodes will lose contact with each other and will be reported as an unknown status. If attempts to stop and start CXFS services do not rectify the problem, resetting the node is a solution but not always possible. A work around procedure was developed and yields an 80% success rate.

Login to the node that is marked unknown. The CXFS services needs to be stopped by issuing a "/etc/init.d/cxfs stop" command. The second command "/usr/cluster/bin/cmgr –c 'admin cxfs_stop'" is required to halt CXFS in the kernel. This process is required to ensure all CXFS services are halted. To restart CXFS services, the kernel needs to be rearmed by issuing a "/usr/cluster/bin/cmgr –c 'admin cxfs_start'". Once CXFS services is running in the kernel, then issue a "/etc/init.d/cxfs start" command. In a few moments, the node should appear active and green on the CXFS GUI.

If, after a few moments, the node does not appear active or is still marked as unknown, the node will need to be reset.

6.4 Filesystem Layout

Extreme detailed planning of the filesystem layout is essential prior to system integration. Using XVM as the filesystem manager has some uniqueness that needs to be understood. XVM for instance owns all the disks. The disk is either labeled as a cluster disk or a local disk. A cluster disk is a disk that is owned by a particular cluster and can be accessed by any node that is a member of that cluster. A local disk on the other hand is owned only be a particular machine and cannot be access by any other machine on the fabric.

6.5 Tape Devices

In order to integrate legacy Exabyte Mammoth tape devices into the storage are network a scsi to fibre channel router was used.

During the integration phase interoperability with the tape drives was not successful. Initial testing with a single external Mammoth tape drive on the SCSI Router was successful. Tests writing data to the tape device to include path failover was successful.

The final system configuration called for an Exabyte X200 tape library to be accessible on the storage area network by all servers. The X200 consisted of a tape robot and four Mammoth tape drives in the library. Once connected initial results revealed that only the first device on the SCSI buss was accessible. Either the first tape drive was accessible or the robot was visible. Upon further investigation it was discovered the HBA driver was dropping all but the first LUN from visibility. That is the HBA's driver was only allowing the first LUN to be built in the hardware graph and the remaining LUNS were discarded.

After numerous inquiries with SGI it was discovered that Exabyte is not a supported tape device on a storage area network. At that time the only tape device supported by SGI for SAN attach was the Storage Tek fibre tape drive.

Due to system requirements to maintain compatibility with tape devices and media type between sites, the X200 tape libraries were removed from the storage are network and attached directly to two of the Origin 2000's.

6.6 Raw Devices

A two-node cluster of Sybase servers was created. This time XVM was in cluster mode and the raw data partitions were again stripped across four storage processors. We created two equal raw data partitions all sharing the same four storage processors. Since only one of the Sybase servers is used at a time this solution is adequate.

In one particular instance we were setting up a database cluster that runs Sybase. Our Sybase implementation uses raw data partitions. On previous systems we experience significant delays in retrieving data from storage devices. In an attempt to solve the data access time, a raw data partition was create that was stripped across four storage processors and XVM was in local mode. This dramatically increased the data access time but the raw data volumes were only accessible by one of the servers.

Due to our particular Sybase implementation and requirements soft links had to be created to ensure that each Sybase server accessed only it's disk slices and no others. Our system required that the Sybase servers did not share the same filesystem. In this implementation we created a cluster that did not share the filesystem, but rather shared the fabric to strip the data over more storage processors to increase data access time.

This configuration has yielded performance improvements in system configuration time from five minutes to around 30 seconds or an 800 percent improvement.

6.7 OS Versions

IRIX versions and patch levels are extremely critical in the storage area network.

During an upgrade of the system from IRIX 6.5.8 to 6.5.9 it was discovered that if the IRIX versions or the CXFS patch level is not the same between all machines in a cluster the cluster will not come up or will be very unstable.

Upgrading IRIX or the patch level one node at a time with the cluster up and running proved not to be reliable. It was observed that once the node is upgraded and is rebooted, the CXFS database would become corrupt and unusable.

Trial and error has proven the best way to upgrade IRIX or the patch level is to stop CXFS services on all machines in the cluster. Once CXFS is stopped and the cluster is down, the new OS or patch can then be loaded. After the node is loaded and the new kernel is built, reboot all nodes and the cluster should come up and be stable.

One approach that should work just as well, but has not been tested, would be to load all the nodes with the OS or patch. With CXFS services running, rebuild the kernel. After the kernel is rebuilt, all nodes must be rebooted after all nodes have been upgraded.

7 Future Work

Future work includes creating even larger clusters of nodes. Extending the cluster size from its current maximum size of 24 nodes to a single SAN image that contains over 50 nodes.

We also intend to, at the earliest possible date, to being to create heterogeneous SAN architecture that contain not only IRIX platforms, but also Windows and other UNIX platforms as well.

8 Summary

The process of incorporating CXFS and SAN technology into the system has exceeded expectations. Early difficulties with CXFS were quickly overcome with close collaboration between SGI and Aerojet.

The system takes advantages of CXFS as a mechanism to enhance system usability, and consolidates duplication of data. With the exception of the tape drive support, the system clearly exceeds our requirements for performance and reliability.

CXFS and SAN technology is clearly the way ahead for the future in large-scale mission critical applications.

9 Acknowledgements

I'd like to express my deepest thanks to Fairy Knappe, Dale Oglesbee, Brian Gaffey, Neil Bannister, and Dan Bonnell from SGI for their collaboration and efforts on the SAN system.

10 References

Building Storage Networks, Marc Farley, Osborne/McGraw-Hill, New York, 2000

CXFS Software Installation and Administration Guide, SGI Document Number 007-4016-008, 2001

Request for Comments (RFC) 1112 Host Extensions for IP Multicasting, Steve Deering, Network Working Group, August 1989

Request for Comments (RFC) 1301 Multicast transport Protocol, Susan Armstrong, et al, Network Working Group, February 1992