DISK STORAGE UPGRADE IN TRIUMF'S CENTRAL CONTROL SYSTEM

M. Mouat, E. Klassen, K. Lee, J. Pon, P. Yogendran TRIUMF, Vancouver, Canada

Abstract

The TRIUMF 500 MeV cyclotron's central control system (CCS) was upgraded a few years ago. Since then, evolving requirements have prompted changes to various sub-systems of the control system. Recently, the disk storage has been upgraded. The disk sub-systems are a major component in the control system and in part determine allowable configurations and software/ hardware maintenance issues. The old and new configurations are described along with the new requirements and considerations. Functionality, performance, expandability, and other characteristics of the new system are also covered. In particular, the multi-porting of the disks is reviewed.

1 INTRODUCTION

The TRIUMF 500 MeV cyclotron's Central Control System upgrade was completed in 1997 [1]. This upgrade took more than 3 years to complete and since then the CCS has continued evolving. By the end of 1997, the application and infrastructural software, the computers, the network, the disk systems, the magnetic tape backup, and the console displays, had been changed. Since then, as a part of the ongoing evolution, computers, network, and tape backup have been changed.

One area that had not evolved sufficiently was the disk storage. Additional, faster, larger disks had been added along the way but the performance was lagging and several other features were unavailable. A product from Compaq/DEC called DSSI was the primary disk facility. One of the most powerful features of the DSSI was the ability to multi-port the disks. This is a configuration that allows more than one computer to simultaneously mount and access any of the disks. The operating system must support this concept. Typically, a clustered computer environment with a distributed lock manager is needed, such as in OpenVMS. Unlike a setup such as an NFS mount where the disk is attached to just one computer that serves other computers, multi-porting allows each computer to independently and, to all appearances, simultaneously access each disk. One major advantage is that any computer can be shutdown, for maintenance or other reasons, without affecting the access of other computers to the disks. Multi-porting also allows a computer to do its disk IO without adding a burden to some other computer and the network between them.

In addition to the performance requirements, the need for increased reliability of components, redundancy, electrical decoupling of the computers, greater flexibility in racking the disk systems, product availability for larger disks, and a variety of other factors, precipitated a change in the disk storage.

2 TECHNOLOGY CHOICES

For the purpose of this paper, the requirements for new disk storage can be viewed from the categories of essential, desirable, and 'nice but not important'. In the essential group are features such as support for OpenVMS, multi-porting, significant performance increases, RAID (Redundant Array of Inexpensive Disks) functionalities, easy expandability, and price. Desirable features include such items as hot swap, redundancy up to 'no single point of failure' capabilities, error logging and enunciation, remote management, multiple operating system support (Solaris, Windows, Linux), optical isolation between the computer and the disk storage, use of an established standardized technology with a future of growth and evolution, and multi-supplier equipment sourcing. In the 'nice but not important' category are features such as a graphical management tools, security components (secure management connections), performance reporting, direct tape backup, and online expansion.

A review of the possible storage technologies finds that there are currently three basic choices; Direct Attached Storage (DAS), Network Attached Storage (NAS), and Storage Area Networks (SAN) [2][3]. A combination of these could also be considered.

DAS is an older style, but is still the most common configuration. In this case, the disks are directly attached to one computer. That computer can do block level accesses, which is efficient for its applications. The computer can serve to other computers, usually via ethernet, but if the computer goes down the disk storage is unavailable. In addition, this computer can be fully occupied just file serving, effectively becoming a NAS configuration.

NAS is a setup where a network, normally ethernet, is used to connect disk storage equipment to one or more computers. This configuration typically does only file level serving. Commonly the same network connection that is used for standard network activity is also used for disk accesses. Ethernet is a flexible medium but it was not designed for congestion or guaranteed point-to-point connections so both the disk IO and the other needs can suffer in this arrangement.

SANs allows multiple computers to connect to the disk storage on a separate, optimized/specialized, network. The hardware is known as Fibre Channel, and is standardized under the ANSI X3T11 Fibre Channel Standards. Block level access is provided. The SAN's configuration often looks similar to a common ethernet network with hosts

and switches. Point-to-point, loop, and switched topologies are supported, as are copper and fibre media.

Only Fibre Channel configurations met our essential requirements so the decision was straightforward.

3 CCS SAN CONFIGURATION / DESCRIPTION

The CCS has two clusters of computers, one for production running and one for development. Each of the groups of computers had its own DSSI disk system and now has its own Fibre Channel disk system. Hardware installation and reorganizing the disk structure is complete, and moving the files to the new system is ongoing.

The new configuration has 5 basic hardware components. The first is the host bus adapter (HBA), which is a card that normally sits on the host computer's PCI bus. The second is the connection (fibre or copper; fibre in our system) from HBA to the Fibre Channel switch. The third is the switch, a component that routes packets between hosts and controllers. The fourth is the connection (fibre or copper; fibre in our system) from switch to storage system controller. The fifth is the controller and the shelves of disks (SCSI or Fibre Channel; SCSI in our system). Our configuration (see figure 1) is modest with a low level of redundancy. These components can be installed in parallel providing a storage system with no single point of failure and automatic failover. Presently, the fibre optic cabling can be up to 100 km, providing wide geographic separation.

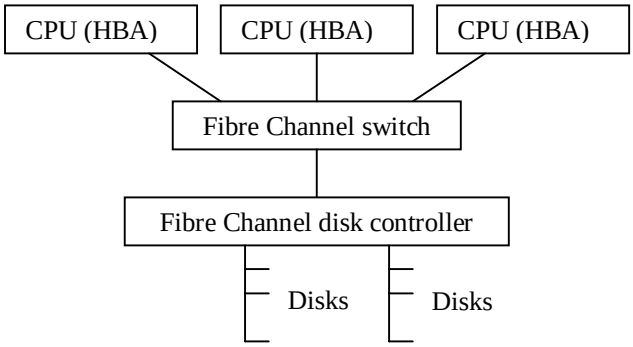

Figure 1. Configuration for Development and Production Clusters

The disk controllers and shelves support features such as RAID structures (mirroring, striping, parity, etc.), JBOD (Just a Bunch Of Disks) organization, partitioning, local and remote management, support for multiple operating systems, automatic sparing of failed disks, hot swap of disks, and dynamic expansion of raid sets.

The user interface to the disk system management has 3 levels; local connection to the disk controller, connection via a host computer (a computer attached to the Fibre Channel system) running a software 'agent' and a local application program, and connection via a remote TCP/IP client (software running on a PC/Windows computer) that talks to either a local or remote 'agent'. Both the local-to-controller and the local-to-agent management programs are command line interfaces. The Windows client application has a graphical user interface.

Although we are presently using the disk controller to connect only to OpenVMS systems, the controller can simultaneously serve a variety of other operating such as Windows, Solaris, HP-UX, Linux, etc. This does not mean that the controller will let two operating systems talk to the same file, it means that the disk storage can be partitioned to support various separate computer systems at the same time. Centralized disk management permits the efficient allocation of new and unused disk resources.

One of the major new features that this Fibre Channel system provides is the protection given by RAID. Previously, when a disk failure occurred there was a high likelihood that cyclotron applications would stop and beam delivery would be terminated. This can lead to significant downtime. With the new system configured for mirroring or striping with parity, a disk failure does not stop access to the disk files although performance may be slightly degraded for a while. The system provides the ability to designate spare disks and when a disk failure occurs, the failed disk is automatically moved (a software reconfiguration) from the 'raidset' to a 'failedset', a spare disk from the 'spareset' is allocated to the 'raidset'. The spare disk is rebuilt online and there is no loss of access, only some loss of performance while the rebuild occurs.

When more storage is needed, this can be accomplished without shutting down the system or any loss of access to the disk volumes. A new disk can be inserted 'hot' and then allocated to a storage set. The disk volume simply appears bigger when the procedure is finished.

This Fibre Channel hardware is rated at 1.0625 Gb/s full duplex from the HBA through the switch to the disk controller. The disk controller has Ultra3 SCSI (160 MB/s) buses and disks. Our controller will support two disk shelves of 14 disk drives each or slightly more than 2 terabytes if 72 GB drives are used. The majority of our disks are 36 GB, 10K RPM drives. The controller has an expandable cache buffer that currently contains 256 MB. Our Fibre Channel switch can run at 100 MB/s per port in each direction (full duplex) and has only a small latency (<2 μs). In general, the actual performance that is attained will depend on the manufacturers and models of equipment.

Users on our computers see a noticeable improvement in performance with the new disk systems. Any activity that involves the disks seems to benefit. Image activation, file IO, tape backups, etc. are significantly faster. In areas where the old disks were a bottleneck, the change in performance is about a factor of 5.

Software utilities to diagnose problems, or monitor use and gather statistics are not currently in use. Host computers do support an SNMP 'agent', but we have not yet explored theses capabilities or found high-level applications that support this low-level functionality.

There is a facility to trigger email messages or a pager call on some types of system problems.

4 UPGRADE ISSUES

As the upgrade progresses, experience has been gained on hardware, software and system management issues. Initially when selecting a hardware provider, a third party supplier was chosen. This company's equipment had desirable features and price but they could not resolve their problems and had to remove their equipment. At that point, Fibre Channel hardware from Compaq, the suppliers of OpenVMS, was purchased and the installation proceeded smoothly. The hardware assembled easily, and the system management and local console commands are straightforward to understand and use. Until fairly recently, we used only the local command line interface for managing the storage system. Addition of the graphical user interface and the ability to remotely monitor and manage this equipment is clearly a step forward.

Installing and commissioning the hardware is not the largest part of upgrading the disk storage. Cleaning up the existing files and file organization, moving the files from the old system to the new, and related tasks are a much bigger job. The Controls Group decided that this was a good opportunity to reorganize the file structure, remove obsolete files, and improve the level of indirection in referencing files.

The new storage system presently has 2 storage sets. The first is a shared system disk (a mirror set), and the second is what appears to be a single large disk (a striped raidset with parity). On the raidset, a file organization was selected that initially has 6 primary directories; users, data, documentation (docs), applications (apps), backups, and scratch. With the exception of system files, which are on the system disk, all files are being placed in a directory structure under one of these top-level directories.

Over the years, many users have come and gone, leaving more than a few orphaned accounts and files. There are almost 200 accounts on the two computer clusters. Establishing which files are still required is not always a simple task. The cleanup has been quite aggressive and if the need arises deleted files can be restored from tape. Each user account has to be adapted for the new disk storage system (disk quota, default device, default directory), the user's directory structure has to be moved over, and references to disk names must be changed wherever they occur (mostly in script files and in application programs). In addition, many of the symbols and logical names (the OpenVMS method of named indirection) also have to be updated. The use of physical disk names or even logical names for the physical disks is avoided. For example, we use RUN SCAN$BIN: SCAN.EXE instead of RUN DSK1:[APPS.SCAN.BIN] SCAN.EXE. Virtually all user defined logical names are located in one file so maintenance is easy. Scripts were developed to help, where possible, to automate moving files to the fibre channel storage.

The time to complete full and incremental backups of this new, larger, disk system was a concern. Our previous configuration of disks and tape drive was limited by the speed of the tape drive. A faster disk system did not improve the backup time. New tape drives, about 8 times the speed and more than 5 times the capacity of the old ones, have been purchased. As expected, the time to do a backup is much shorter and takes fewer tapes.

5 FUTURE DIRECTIONS

There are currently Windows and Solaris systems that could take advantage of the new Fibre Channel facilities.

It is possible to attach tape drives to the Fibre Channel system itself and share the drives between various host computers. In a similar fashion to the multi-porting of disks, this allows computers to be shutdown without the loss of access to the tape drives by the other computers.

Fibre Channel is a standards based technology and like other standards such as SCSI and ethernet, future developments are constantly being planned and implemented. The specifications for speeds up to 4 Gb/sec are already formalized and 2 Gb/sec are commercially available. Upgrading the SAN will likely become as common as upgrading the LAN.

6 SUMMARY

The disk storage in TRIUMF's central control system has been enhanced with the addition of 2 Storage Area Networks. Installation and configuration of this Fibre Channel hardware is complete. The task of moving old files to the new disk and directory structure, and making the appropriate changes, is underway and proving to be a much bigger job than installing the hardware. Substantial performance improvements, capacity, reliability, and functionality have been realized.

7 REFERENCES

[1] M. Mouat et al, "Status Report on the TRIUMF Central Control System", International Conference on Accelerators and Large Experimental Physics Control Systems '97, Beijing China, November 1997

[2]http://www.storage.ibm.com/software/roadmap/softwareroadmap.pdf

[3]"The Compaq Enterprise Network Storage Architecture: An Overview", Doc #12L8-0500A-WWEN, Compaq Computer Corporation, May 2000