DATA STORAGE FOR THE SLS

L. Sekolec, D. Vermeulen
Paul Scherrer Institut, 5232 Villigen PSI, Switzerland

Abstract

When it comes into operation in August 2001, the Swiss Light Source will require a large (> 1 Terra Byte) amount of data storage capacity, both for controls and data acquisition. Already during the initial stages of testing and commissioning, the parameters of the control elements will need to be archived. Over time, the data storage requirements will certainly grow as new beam-lines and new detector technologies are used, so the storage system must be easily expandable. Due to high availability requirements, the hardware implementation must include fault tolerance, which implies hardware RAID with detectable alarm for disc or port failure. Expansion of storage capacity on the fly, without disrupting services, must be possible. High bandwidth is needed, as transfer rates of up to 80 MByte/second are expected. Additionally, multi-platform access over long distances, using standard network protocols, must be available. A modern concept for storage must be implemented which gives us the ability not only to store increasing amounts of data, but also to archive them. Backups, including databases, must be made on-line, as we cannot afford to bring down the system during this activity. A prototype system is being built to meet these requirements, based on commercial systems using Fibre Channel technology.

1 INTRODUCTION

The Swiss Light Source (SLS) will require storage capacity in excess of 1 TByte, both for controls and data acquisition when it comes into operation in August 2001. During the testing and commissioning stage, the on-line storage capacity should not exceed 200 GByte, but the requirement will grow considerably as the experiments start to deliver data. Only time will show how much storage capacity will finally be required, so we need a storage system that is, besides being reliable, also easily expandable.

Various stages of storage must be considered. On-line storage for regularly accessed data should be on fast media. Near-line storage for infrequently accessed data can be on slower media e.g. tapes or slow, high capacity discs. Off-line storage would be on media in a store and requires operator intervention. This data is automatically deleted after a predefined period, i.e. the media is made available for new data. Archive storage is the same as off-line, except that the data are kept for 10 years or removed on owner request. Hierarchical Storage Manager software handles the automatic transfer of data between the different stages of storage.

Any hardware implemented must be fault tolerant. This implies hardware RAID with detectable alarm for disc or port failure. Expansion of storage capacity on the fly, without disrupting services, must be possible. High bandwidth is needed; transfer rates of at least 80 MByte/second are expected. Additionally, multi-platform access over long distances, using standard network protocols, must be available.

A modern concept for storage must be implemented which gives us the ability not only to store increasing amounts of data, but also to archive them. Backups, including databases, must be made on-line, as we cannot afford to bring down the system during this activity. A technology that gives us these possibilities seems, at the moment, to be Fibre Channel connections to Network Attached Storage (NAS) with possible expansion to Storage Area Networks (SAN).

2 FIBRE CHANNEL

2.1 Fibre Channel Technology

The current SCSI technology is limited both in transfer rates, 40 MByte/sec, and in the length of the bus. A maximum of 3 m between devices and a total length of 12.5 m per bus, 25 m for a differential SCSI bus, is allowed. Additionally only 15 devices per bus can be attached.

With Fibre Channel technology, initiated in 1993 by SUN, HP and IBM, transfer rates of 100 MByte/sec are reachable, even 200 MByte/sec in duplex mode. Rates of 300 MByte/sec are expected in the near future. The distance between devices can be 30 m with copper media and up to 500 m with Fiber Optics in multi mode. With Fiber Optics and mono mode, up to 10 Km are possible. Up to 126 devices or nodes can be connected to a Fibre Channel bus. A node can be a RAID system, Tape Library or a single device. Normal SCSI devices can be connected by use of SCSI to Fibre adapters.

Several protocols are supported by Fibre Channel, the most popular of these being SCSI. This means that applications that currently use SCSI devices can without problems implement Fibre Channel hardware. Other protocols include Internet Protocol (IP), ATM and IEEE 802.2 LAN.

Two basic topologies are possible with Fibre Channel, Point to Point Topology and Arbitrated Loop Topology.

2.2 Point to Point Topology

Point to Point topology is the simplest implementation of Fibre Channel where disc systems are attached to a server node. This can also be considered as simple Network Attached Storage (NAS) and appears as a "black box", with dedicated software providing NFS or CIFS volumes on the network.

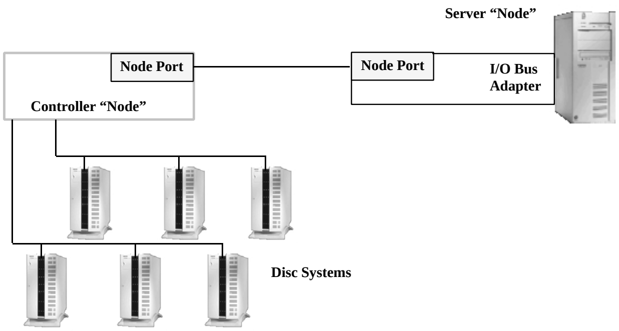

Figure 1: Point to Point Topology

2.3 Arbitrated Loop Topology

Arbitrated Loop topology is an extension of the Point to Point topology. A server is connected to the I/O devices by means of a Fibre Channel loop or even a dual loop for dual ported devices. This ensures that if one device falls out, the others are still reachable. Such a topology also enables transfer of data directly between devices, without having to communicate through the server. It is hoped that some time in the near future, intelligent I/O devices will be available to enable such communication.

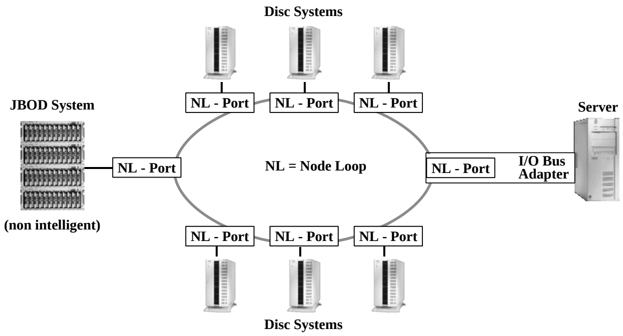

Figure 2: Arbitrated Loop Topology

2.4 Storage Area Network

A Storage Area Network (SAN) is a network, which can combine a variety of storage systems into one mass storage unit. This can be achieved by means of the Fibre Channel Fabric, a combination of switches and other connection interfaces, which are not as yet defined as standard. With appropriate Hierarchical Storage Manager software, this physical storage can be distributed as any number of virtual volumes. The SAN concept exists already with vendor specific solutions, but a general, vendor independent standard is not yet available. One needs to wait for another few years, before general hardware and software is available to implement a Storage Area Network of multi vendor storage systems.

3 FILE SERVER REQUIREMENTS

3.1 Storage Solution

Since the conditions for a general SAN at the Paul Scherrer Institut (PSI) do not as yet exist, we must decide on an intermediate storage solution. The simplest and immediately usable possibility is the NAS solution. A "black box", with dedicated software providing a mounted network volume, can be quickly installed and requires practically no management after the initial setup. In this case all data would be on-line, and any archiving would need to be handled separately. Archiving in the case of SLS is not very demanding, as the experimenters will take the data back to their home institutes. In principle only data needed for the control of the machine will need to be archived

3.2 Storage Hardware

The hardware implemented must be fault tolerant and expandable. This implies hardware RAID discs with detectable alarm for disc or port failure and addition, removal and replacement of discs on the fly.

Tape Libraries used for off-loading and archiving data should be capable of error recovery, error logging, tape loading on the fly and automatic drive cleaning. Libraries have at least two drives as opposed to an Autoloader or Stacker with only one. Several drives can read and write independently, and if one drive fails, the library is still functional.

3.3 Storage Capacity

Initially 200 GByte of on-line capacity, with easy upgrade possibilities to 1 TByte during the next 2 to 3 years are required. As more beam lines and experiments come into operation, and new detector technologies are implemented, the data storage requirements are expected to increase to 3 TByte. The data acquisition for the Protein Crystallography experiment itself is expected to gather 350 GByte per day.

Upgrades can be achieved by adding more discs or, when this is no longer possible due to space and/or system restrictions, the old discs can be replaced by new, higher capacity discs as they become available. As an additional solution, the NAS system can be expanded by clustering several units into a vendor specific SAN.

When the time is ripe, the SLS file server can be integrated into the general PSI SAN with all its possibilities of storage expansion and hierarchical management of data.

3.4 Multi-platform Access

The file server should support at least the following protocols: NFS, CIFS and http for the SLS web pages. Files on the file server must be visible from both UNIX and Windows NT in native form.

3.5 Data Integrity

Frequent backups need to be made. Not to effect the performance of the system, this should be in the form of snapshots. Snapshots can be considered as on-line backups and is a technology that is implemented by most mass storage system vendors. Incremental backups (every two hours) will be kept locally for instant retrieval, for the duration of the experiment - typically a few weeks, whereas full backups should be archived for a predefined period - 6 months.

Snapshots are intended to cater for errors such as, file deletion and data corruption. Hardware disc failure should be catered for by the RAID technology. Loss of data due to catastrophes such as fire or water can be avoided by mirroring the physical storage at different parts of the institute and/or archiving them to tape.

3.6 Data Export

Visiting scientists need to take experimental data back to their home institutes, therefore a variety of removable storage devices such as DLT tapes, CD and DVD drives must be provided. Devices such as CD drives are already capable of writing 1 MByte/sec, and DVD is about five times faster.

An "open shop" type of service should be provided at various places, particularly at the experimental beam lines, where the visiting scientist comes with his removable media and executes a menu driven copy procedure. File protections ensure that he can copy only his own data.

3.7 Access Speed

Where fast data transfers are demanded, 80 MByte/s for the Protein Crystallography experiment, one should consider Solid State Discs with access times measured in microseconds as opposed to milliseconds for disc drives. As long as only one experiment requires such high transfer rates, one can use Solid State Discs as a cache, directly connected with Fibre channel to the data acquisition system. When transfer rates of 80 MByte/sec and above become a more general requirement, such a cache would need to become part of the general storage concept. The connections between the cache and the on-line storage must of course also be Fibre Channel, to guarantee a continuous high transfer rate. Currently the price of Solid State Discs is still relatively high, $16 per MByte, but by the year 2002, the price is expected to fall to $4 per MByte.

4 CONCLUSION

PSI is working on a general SAN storage system but it will be some time yet before something can be implemented. Neither hardware nor software are ready for a vendor independent SAN. It is expected that one must wait for another two or three years before the necessary products are available on the market.

To ensure that the SLS project has the required data storage capacity available when it starts commissioning, we must decide on an intermediate solution. Such a solution must ensure that the storage requirements for the early stages of operation are also covered. Particularly, easy expansion of storage capacity must be guaranteed.

Since manpower is always a problem, we need a system with minimum management demands. The device should be installed and left to run. Important is also the reliability and support. If problems do arise with software or hardware, these should be solved very quickly; we cannot afford to wait for an expert to be flown in.

At the moment, a NAS storage system based in Fibre Channel technology seems to be the best solution. Such systems appear as “black boxes” that are installed, configured and left to run. Discs can be added as required and the files are accessible from UNIX and Windows NT in native form. Fibre Channel also satisfies the demand for high transfer rates.

When expansion by addition or replacement of discs is no longer possible, we still have the possibility of building a local SAN. By the time all growth possibilities are exhausted, the general PSI Storage Area Network will be available to provide additional storage capacity.