Fermilab Computing Division Systems for Scientific Data Storage and Movement

D. Petravick

Fermi National Accelerator Laboratory, Batavia, IL, U.S.A.

Abstract

In the past two years, the Fermilab Computing Division has constructed software and facilities which are used for the D0 experiment in run II and other elements of the scientific program. ⋆

The most recent activity has been centered around open source and commodity technologies: Intel based Linux computers for data movement, commodity networks, commodity tape drives, and HEP specific storage software. These systems are being commissioned for Run II and advocated for broad use throughout the Fermilab scientific program.

I provide an overview of all facilities and software constructed and supported by the Fermilab Integrated Systems Development Department, describe the general technical directions pursued, and describe progress to date.

Keywords: tape,storage,data network

1 Introduction

As preparations for Run II complete, the Fermilab Computing Division has constructed and begun to operate storage systems built around commodity components and open software. Commodity local area networks, “White Box” Pentium based computers running the LINUX operating system for servers and data movement from tape, and the likely use of commodity tape drives for Run II, with LINUX EIDE disk caches for general users are in operation or planned.

The supporting software, which is mostly open source, has been built by taking components and design ideas from the DESY data management group, and has included a jointly constructed Disk Cache software system.

This paper briefly describes these elements of our system.

2 Data Movers and Servers

Perhaps the most innovative element of the system is its extensive use of low cost-Intel-based servers for every task in the storage system. The Linux operating system was chosen for its support of low cost personal computers. Additionally, Linux allows us to exploit the experience gleaned from other projects the department has undertaken [1] , and allows us to exploit the Linux-related body of knowledge at HEP labs.

In detail, we have deployed “white box” computers. The primary components are: Intel Lancewood L440-GX+ Mainboards and Intel EtherExpress Pr/100 100 mbps Ethernet adaptors. The Intel Mainboards are notable for their BMC (baseboard management controller) which allows

⋆ Work supported by the U.S. Department of Energy under contract No. DE-AC02-76CH03000.

[1] see `http://www-isd.fnal.gov/`

production monitoring of critical hardware operating parameters, for example case temperatures and CPU fan speeds.

Monitoring for the whole ensemble of computers is via RS-232 serial lines. The server computer used for the monitoring is equipped with a Cyclades Cyclom-YeP serial line multiplexer. Each production computer is monitored using two serial lines. One serial line transmits the console logs and runs a getty, allowing for login should the network fail. The other provides access to the server BIOS at boot time, and then allows access to the EMP (Emergency Management Port) features of the Intel Motherboards.

The system has active, Ethernet controlled power management to assist in general administration, especially recovery form power failure.

The low hardware cost obtained by this approach allows us to build test and development stands very inexpensively. Administrators and developers can run storage systems on their desktop work stations. Redundant servers can be purchased at reasonable cost. Performant test systems can be constructed apart from the production systems.

2.1 Servers

Server computers are built from the commodity Linux systems mentioned above. Server computers manage data movement, but do not transmit data. Because of their low cost, a somewhat luxurious number of servers can be provided and utilized, because the underlying software system, Enstore, is distributed. For example, a separate server computer is employed solely as a WWW server to make sure that other servers perform even though there is intense web activity.

Apart from high availability, the demands on a server computer are not great. In part, high availability is supported by using the watchdog feature provided by the BMC on the Intel Lancewood motherboard. The watchdog feature reboots a computer if is it not accessed via software regularly. The most unusual aspect of the server operating system configuration is the use of Linux disk mirroring in support of the various databases deployed on the servers,

2.1.1 Data Movers

When configured as data movers, the Intel Lancewood systems are extended with 512MB of S-DRAM. The provisional Run II tape drives, Exabyte MAMMOTH-1, are interfaced using Adaptec 2944 differential SCSI adapters. We have 13 of the provisional drives in production, and have SONY AIT-1 and Quantum DLT7000 drives in less intense service as well. The demonstration of flexibility is important, as the final hardware for Run II has not been chosen.

One candidate Run II tape drive will move data (uncompressed) at 12MB/second. This drive has been used a design exemplar for the system. One lancewood computer can support two such tape drives. The drives are intrinsically robust again start/stop, and the 512 MB of memory on the mover computer provides a substantial smoothing buffer, given that the envisioned Run II file size is 1 GB.

The D0 run II system is specified to sustain 150 MB/sec from tape continuously. For D0 it is a system design principle that many of the computer systems must accept data transfers at or close to tape speed. As a consequence, we imagine deploying very little storage system side disk buffering. The thinking is that this would merely be a cache which would provide no real benefit, waste a significant amount of resources, and introduce additional unreliable mechanical elements into the system. Rather, a large memory buffer, available at commodity memory prices and a modest tape drive cost make it economical to network attach the tapes without a storage system side disk buffer.

Equipped as described, the data mover computer cost is about $700/tape drive, and compares favorably to the envisioned $5000 cost of the tape drive. The D0 system will have about 20 such mover computers.

We are implementing a similar system for an STK Powderhorn Silo, which is equipped with 9840 tape drives for general purpose use at Fermilab, and as a redundant system for D0 Data Acquisition, should the provided AML/2 robot fail. The hardware configuration of the mover computers for these tape drives is identical to the Run II data movers deployed for Run II.

3 Networks

We attach tape storage to computer systems using commodity, scalable and proven networks. Commercial, reasonable cost Ethernet switches have immense bandwidth compared to the experimetns required throughput. For example, the CISCO 6509 switch which is the backbone switch for the D0 offline systems has a throughput of 32 gbps, or approximately 4 gigabytes per second at very reasonable cost per port, especially if 100 mbps is the preferred wire speed.

We are able to obtain 94mbps (11.75 MB/sec) end to end throughput on the Ethernet, and have studied other issues which lead to rate degradation in practical systems, such as the need for proper support in the target operating systems scheduler.

Currently we prefer 100 mbps Ethernets for their low cost, $200/ port, and a direct connection into the salient backbone network. As we have argued, Ethernets are economical to implement on the storage system side, but, given the accepted constraint of a 1500 byte Maximum Transfer Unit (MTU) are CPU intensive on an expensive SMP computer. However, the tradeoff of using what some consider to be expensive CPUs to import data into an SMP computer has been accepted.

4 Tapes and Tape Libraries

We are operating two tape libraries in this type of storage system.

For Run II, we have an ADIC AML/2 tape library with three quadro-towers. This library supports almost any type of tape drive and tape media in a modular fashion. Currently the library holds AIT, DLT, and most significantly 13 Exabyte MAMMOTH tape drives, which are the interim tape drives selected for Run II development work.

The library has two robotic arms, and will support over 200 tape drives. It holds approximately 15,000 8mm tapes, and is potentially useful for DVD type technology.

For general purpose use, we have an STK Powderhorn library, which holds approximately 5000 tape volumes. It has 5 9480 tape drives and will be put into production this spring.

5 Software

Our main direction in software a system called Enstore for tape staging and a disk cache, jointly developed by Fermilab and DESY. In addition we have software to support the operational deployment of these systems.

5.1 Enstore

Enstore is a distributed tape staging system. It supports network and locally-attached tapes. It uses PNFS a general purpose distributed name space, which is especially appropriate for storage systems and was developed at DESY. The system is freely available, and uses no licensed software. It supports STK, Exabyte, SONY-AIT and DLT tape drives. It supports STK and ADIC AML/2 robots. It is extensible, and suports “lights out” operations.

5.2 Fermilab DESY Disk cache

The disk cache software is portable, and available, using no licensed software components, and is written in Java. The system is scalable and supports FTP access to a tape system. While we believe that high rate, large volume data transfers to commodity equipment should not use a buffer in the storage system, comprehensive solution for a laboratory as diverse as Fermilab provision should have provisions fo r other data flows.

We envisage the software's function is to:

- Accomadate smaller, less well-engineered flows, by providing the capability to implement some degree of buffering in the storage system.
- Provide a gateway to the wide area network, which has different transfer characteristics than the Fermilab Data lan.
- Be a software component experiments can implement on their own computers, which provides a usable buffering and caching system.

5.3 Operational and Monitoring software

The operational aspects of these systems are important. Part of the system development is to expose information to administrators, who can take pro-active steps to maintain and enhance these storage systems, which have ambitious throughputs and availabilities. Extensive monitoring of networks, computers, and tape drives is necessary if the system is to have good performance. Detailed status from inside the storage software, as well as from its supporting hardware has been integrated into WWW displays. Further work continues in this area.

6 Conclusions

Appropriate and usable storage systems have been constructed for Run II and other uses at Fermilab. The types of software and hardware used coincide with trends in the field towards open source software and commodity hardware. The effort to construct and deploy these systems builds on and enhances this technical direction, and constitutes some pioneering work at Fermilab.

References

1 `http://www-isd.fnal.gov/enstore/`

2 `http://d0ensrv2.fnal.gov/enstore/`

3 `http://d0db.fnal.gov/sam/`

4 `http://mufasa.desy.de/`

5 `http://ods.fnal.gov/ols/doc/pcfarms/`