Fermilab Computing Division Systems for Scientific Data Storage and Movement

D.L. Petravick

Fermilab

CHEP2000, February 2000

Abstract

- Fermilab CD has constructed software and facilities for The D0 Experiment
- The facilities use commodity technologies;
  - INTEL server computers
  - Commodity networks
  - Likely commodity tape drives
- The software has been built in collaboration with the DESY scientific data group.
  - Enstore tape staging system which uses the DESY PNFS namespace.
  - Fermilab-DESY disk cache.
  - a large body of software to assist in operating the system.

High Level goals

- Meet requirements of experiments
- Be generally applicable
- Apply most cost effective hardware,
  - commodity networks
  - commodity computers
  - commodity tape drives
- Collaborate to accumulate a body of open software

Linux Computers

- Linux supports low cost personal computers.
- Access to source code helps to build performant systems
- Most salient deployed components are:
  - Intel Lancewood L440GX+ mainboards
  - Intel EtherExpress Pro/100 100 mbps Ethernet adaptors
  - Monitoring via two serial lines per computer
    - console logging and getty
    - BIOS and Emergency Management Port
  - Adaptec SCSI adaptors.
  - Detailed configuration management necessary

Server Computers

- Luxuriant (4!) number of servers because of low cost.
  - PNFS server for PNFS name space, and data bases.
  - Configuration server, alarms, logging, web, inquisitior.
  - Console server and metadata backups.
  - Library Manager and Media changers.
- Linux disk mirroring.
- BMC watchdog and IPMI monitoring.

Data Mover Computers

- Data movers:: $700 per tape drive: includes 256MB smoothing buffer.
- 13 provisional Run II (MAMMOTH-1) drives in production.
- 2 DLT-7000 drives in production.
- 2 AIT-1 drives (dormant)
- Have moved 1.2 TB/day. (not a stress test)
- D0 will have about 20 such mover computers.
- Implementing movers for 9 STK 9840 tape drives as well.
- 11 Misc drives in test stand.

Networks

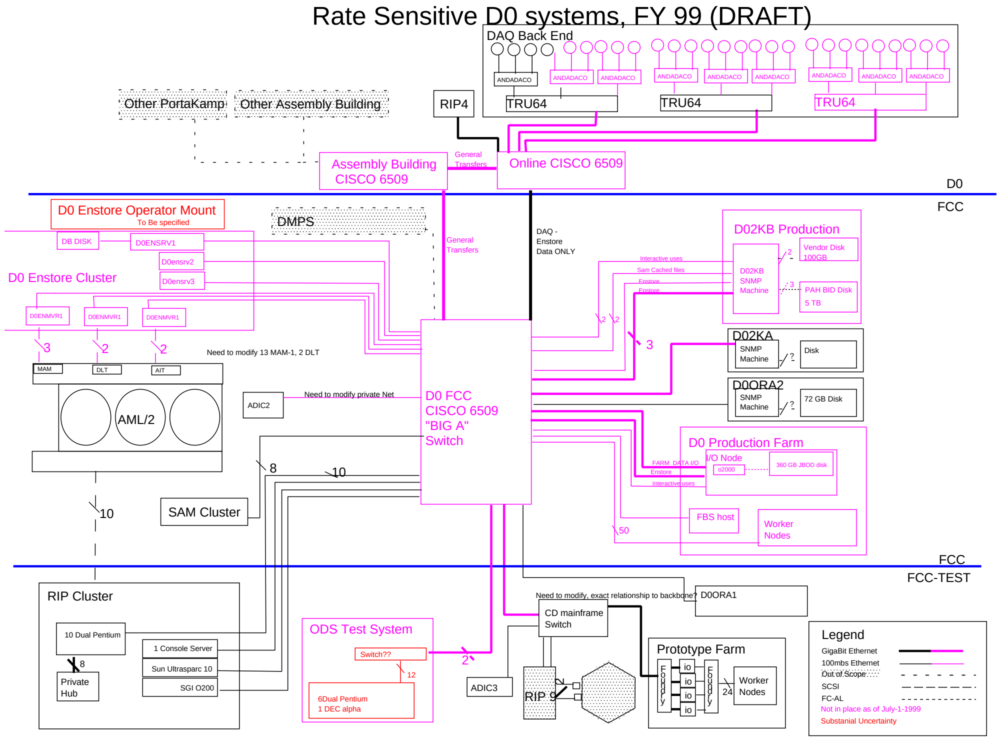
- For D0 tape drives are attached via commodity Ethernet networks.
- Enstore side: 100 mbps Ethernets, in a carefully constructed environment.
- We achieve full wire rate in tests. (11.7 MB/sec, Standard MTU)
- Receiving computer, software shall be constructed to catch full rate.
  - There are software considerations, not just “fast disks”.
  - SGI server has reserved CPUs to catch the data and ensure performance
  - Real-time scheduler for receiving process for Linux...

ADIC AML/2 Tape Library

- DIstinctly Non Commodity tape library.
- Implements Run II “late choice” strategy.
- 3 quadro towers
- 5000 8mm media/tower
- 250 drives/library possible

Run II deployment for D0

- Commodity Computers (Intel Lancewoods)
- Commodity Tape (Mammoth-1) as interim
- Commodity Network (100 mbps Ethernet)
- ADIC AML/2 library.)



Rate Sensitive D0 systems, FY 99 (DRAFT)

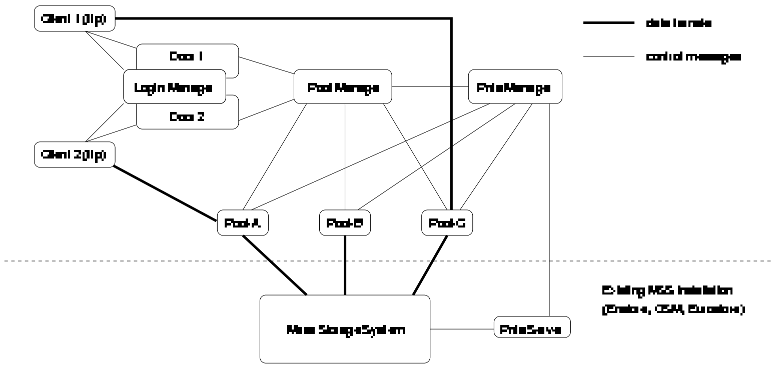
Enstore

- Tape staging system.
- Independent data movers – scales to very large data rates.
- Flexible tape drive support to implement flexible media strategy.
- OPEN source philosophy.
- Resues PNFS name spce from DESY.
- Implemented as discussed at last CHEP.
- http://isd.fnal.gov/enstore/

FNAL-DESY Collaborative Work

- Distributed, scalable disk cache
- Prototyped in Python, coded in Java
- Deploying this Spring, on small farm of lancewood EIDE disk servers.

FNAL/DESY Disk Cache for Mass Storage Systems
V1.0 Implemented Fall '99, C.G. Waldman, P. Fuhrmann



Future Work

- Deploy STK equipment for general use at Fermilab.
- Complete Run II deployment when final tape drive is selected.
- Deploy the FNAL-DESY disk cache for general use.
- Find collaborative partners for future needs.
- Test with OOFS and other interfaces to high level object systems.
- Use the systems level expertise in other areas.

Summary

- FNAL has a leading edge application of commodity technologies.
- The system successfully meets Run II needs.
- Forms a basis for collaboration with other labs