Fermilab Mass Storage System

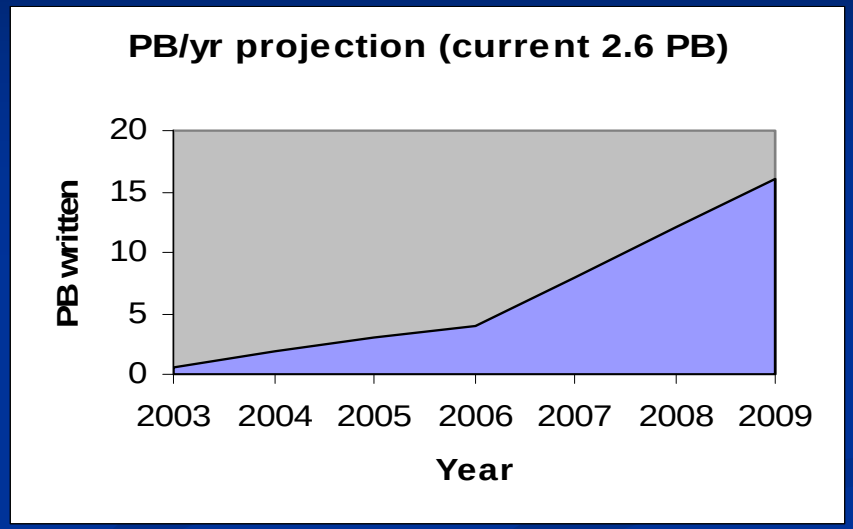
Gene Oleynik Integrated Administration, Fermilab

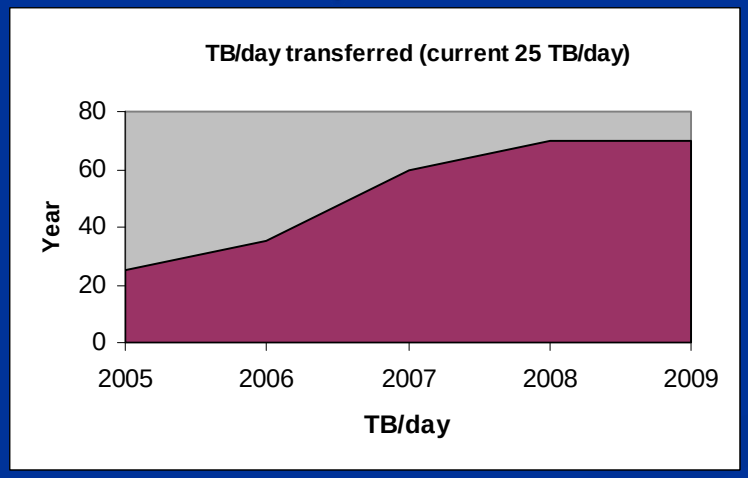
Introduction

- Fermi National Accelerator Laboratory (FNAL)
  - The premier, highest energy, particle accelerator in the world (for about two more years)
  - Accelerate protons and anti-protons in a 4 mile circle and crash them together in collision halls to study elementary particle physics
  - Digitized data from the experiments from experiments written into the FNAL Mass Storage System
- Overview of the FNAL Mass Storage System
- Users and use pattern of the system
- Details on MSS with emphasis on features that enable scalability and performance

Introduction

- Multi-Petabyte tertiary tape store for world-wide HEP and other scientific endeavors. THE site store
- High Availability (24x7)
- On and off-site (off-continent) access



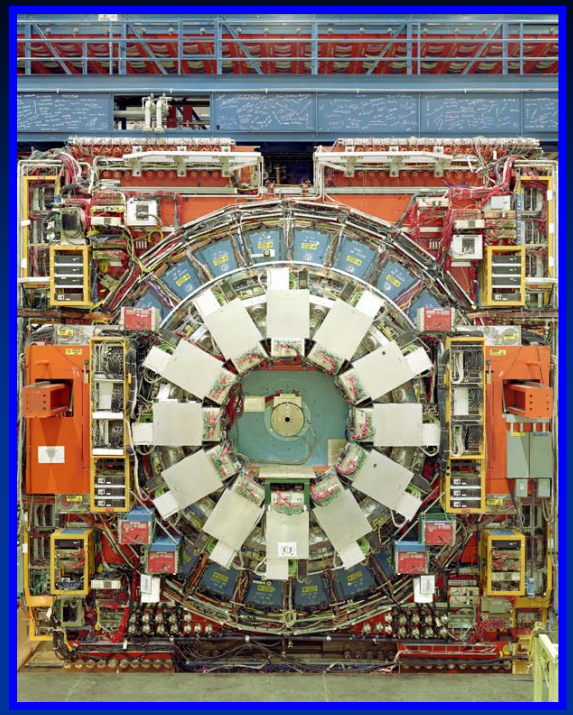


- Scalable Hardware and Software Architecture
- Front-end disk caching
- Evolves to meet evolving requirements

Users

- Local HEP experiments: CDF, D0, minos, mini-boone (> 1 PB/yr, 25 TB/day peak transfers). \
- Remote HEP experiments: Tier 1 site for CMS via Grid from CERN Switzerland (3.5 PB/yr 2007+)
- Other remote scientific endeavors: Sloan Digital Sky Survey (ship dlt tapes written at Apache Point Observatory, New Mexico), auger (Argentina)
- Quite a few others: Lattice QCD theory for example

The DZero Experiment



The CDF Experiment

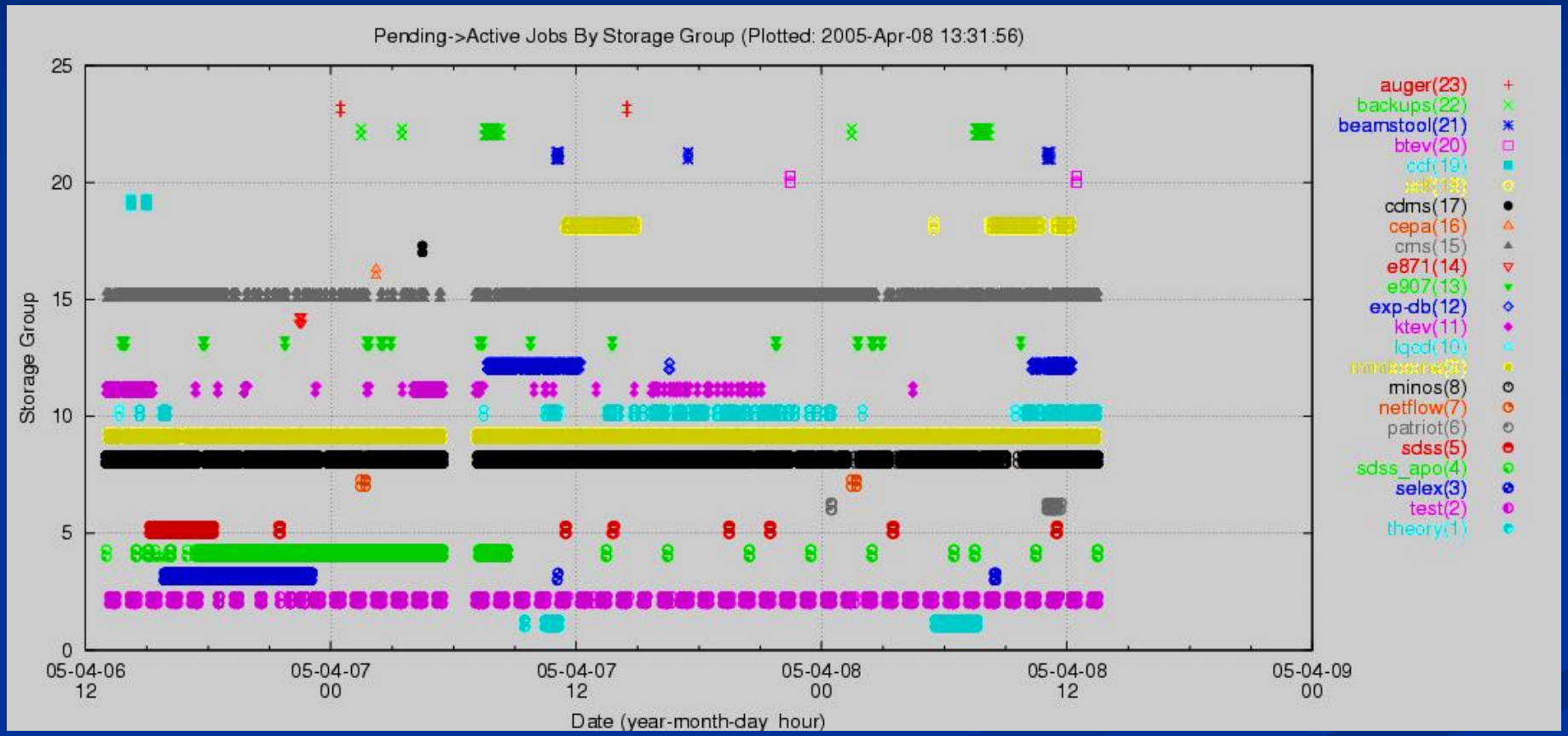
Sloan Digital Sky Survey

The CMS Experiment

And Many Others
KTev, Minos,
Mini-boone, …

- Users write RAW data to the MSS, analyze/reanalyze it in real-time on PC "farms", then write results into the MSS
- ~3 bytes read for every byte written
- Distribution of cartridge mounts skewed to large numbers of mounts
- Lifetime of data on the order 5-10 years or more

Plot of User Activity



Software

- enstore
  - In-house, manages files, tape volumes, tape libraries
  - End-user direct interface to files on tape
- dCache
  - Joint DESY and Fermilab, disk caching front-end
  - End user interface to read cached files, write files to enstore indirectly via dCache
- pnfs
  - DESY file namespace server (nfs interface) used by both enstore and dCache

Current Facilities

- 6 Storage Tek 9310 libraries (9940B, 9940A)
- 1 3-quadratower AML/2 library (LTO-1, LTO-2, DLT, 2 quads implemented)
- > 100 tape drives and associated "mover" computers
- 25 enstore server computers
- ~100 dCache nodes with ~225 Terabytes of raid disk
- Allocated amongst 3 instantiations of enstore:
  - cdfen (CDF experiment) – 2 9310 + 40 drives + plenty dCache
  - D0en (D0 experiment) – 2 9310 + 40 drive, 1 AML/2 quad + 20 drives
  - Stken (General purpose + CMS) – 2 9310 + 23 drives 1 AML/2 quad (dlts) + 8 drives
- Over 25000 tapes with 2.6 P of data

Architecture

- File based, user presented with nfs namespace. ls, rm, etc. but files are stubs
- Copy program to move files in and out of mass storage, encp
- Copy program to move files in and out of mass storage via dCache, dccp, ftp, gridftp, SRM
- Details of the storage all hidden behind the file system interface and copy interfaces

Architecture

- Based on commodity PCs running Linux. Typically dual Xeon 2.4GHz
- Mover nodes for data transfer
- Server nodes for managing metadata – library, volume, file, pnfs name-space, logs, media changer. External RAID Arrays. postgreSQL
- Administrative, monitor, door, and pool nodes for dCache

Architecture

- Technology independent
  - Media changers isolate library technology
  - Movers isolate tape drive technology
- Flexible policies for managing requests and accommodating rates
  - Fair share for owners of drives
  - Request Priority
  - Family width

Architecture

- Number of nodes of any type unrestricted
- Number of active users is unrestricted
- Scale
  - storage by building out libraries
  - transfer rates by building out movers, tape drives
  - Request load by building out servers
- Client's needs are somewhat unpredictable, typically must scale on demand as needed

Managing File & Metadata Integrity

- File protection procedural policies
  - File removed by user are not deleted
  - Separation of roles – owner in the loop on recycles
  - Physical protection against modification of filled tapes
  - "cloning" of problem or tapes with large mount counts
- File protection software policies
  - Extensive CRC checking throughout all transactions
  - Automated periodic CRC checking of randomly selected files on tapes
  - Flexible read-after-write policy
  - Automated CRC checking of dCache files
  - Automated disabling of access to tapes or tape drives that exhibit read or write problems.

Managing File & Metadata Integrity

- Metadata protection
  - Redundancy in database information (file, volume, pnfs)
  - RAID
  - Backups/journaling – 1-4 hr cyclic and archived to tape
  - Automated database Checks
  - Replicated databases (soon)

Migration & technology

- Migrate to denser media to free storage space
- Evolve to newer tape technologies
- Normal part of workflow – users never loose access to files.
- Built in tools to perform migration, but staff always in the loop
- Recent efforts:
  - 2004 CDF: 4457 60GB 9940A to 200GB 9940B cartridges. 3000 slots and tapes freed for reuse. Only 42 file problems encountered and fixed.
  - 2004-2005 stken: 1240 20GB Eagle tapes to 200GB 9940B tapes on general purpose stken system freeing 1000 slots.
  - 2005 Migration of 9940A to 9940B on stken about to start
  - And on it goes. Have LTO-1 to LT0-2 still to migrate

Administration & Maintenance

- The most difficult aspect to scale
- Real time 24 hours/day requires diligent monitoring
- 4 Administrators, 3 enstore, 3.5 dCache developers
- 24x7 vendor support on tape libraries and tape drives.

Administrators

- Rotate 24x7 on call primary and secondary
- Monitor and maintain 7 tape libraries with > 100 tape drives, 150 PCs, 100 file servers
- Recycle volumes
- Monitor capacity vs. use
- Clone overused tapes
- Troubleshoot problems
- Install new hardware and software

Administrative support

- Automatic generation of alarms by Enstore and dCache software. Generate tickets and page administrators.
- In-house on-hour helpdesk/off hour call centers generate pages and tickets.
- Extensive web based plots, logs, status, statistics and metrics, alarms, and system information.
- Animated graphical view of Enstore resources and data flows.

Administration Monitoring

- States of the Enstore servers
- Amount of Enstore resources such as tape quotas, number of working movers
- User request queues
- Plots of data movement, throughput, and tape mounts
- Volume information
- Generated alarms
- Completed transfers
- Computer uptime and accessibility
- Resource usage (memory, CPU, disk space)
- Status and space utilization on dCache systems

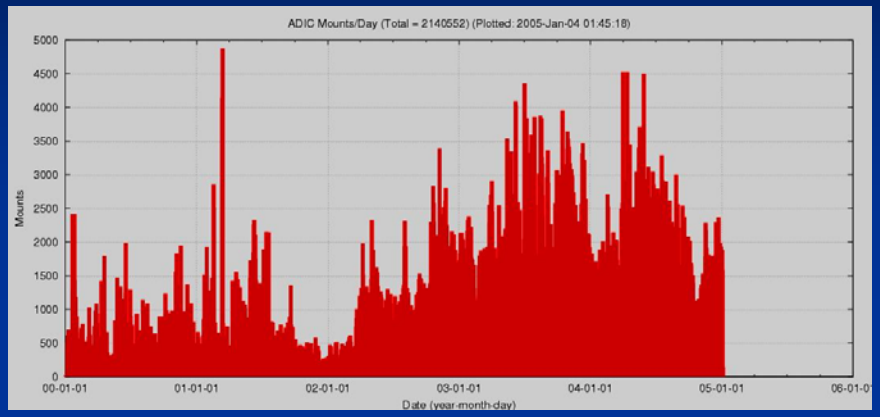
Admin metrics, random week

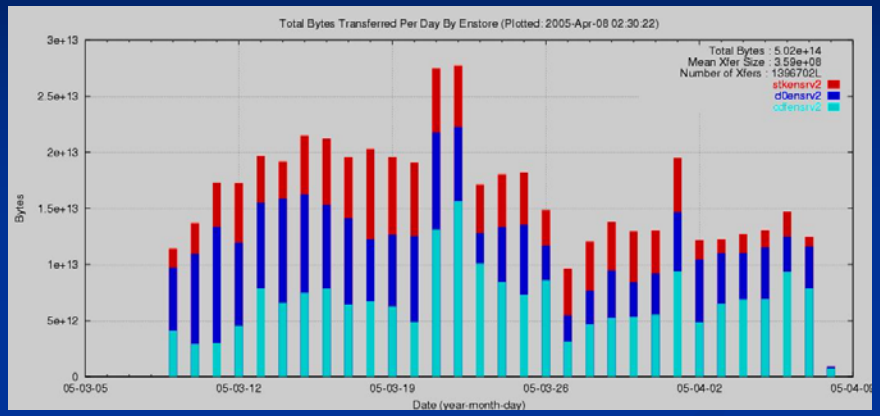
- 1 9940B, 1 9940A, 1 LTO2 replaced
- 3 mover interventions
- 4 server interventions
- 2 tape drive interventions
- 2 fileserver (dCache) interventions
- 3 tape interventions
- 4 file interventions
- 40 tapes labeled/entered
- 2 tapes cloned
- 3 Enstore service requests
- 1 data integrity issue

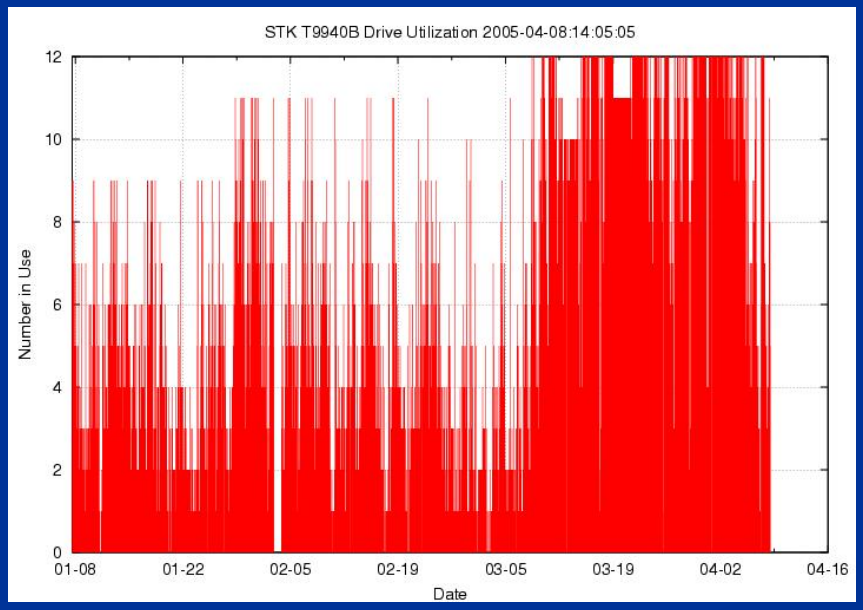
Administration Observations

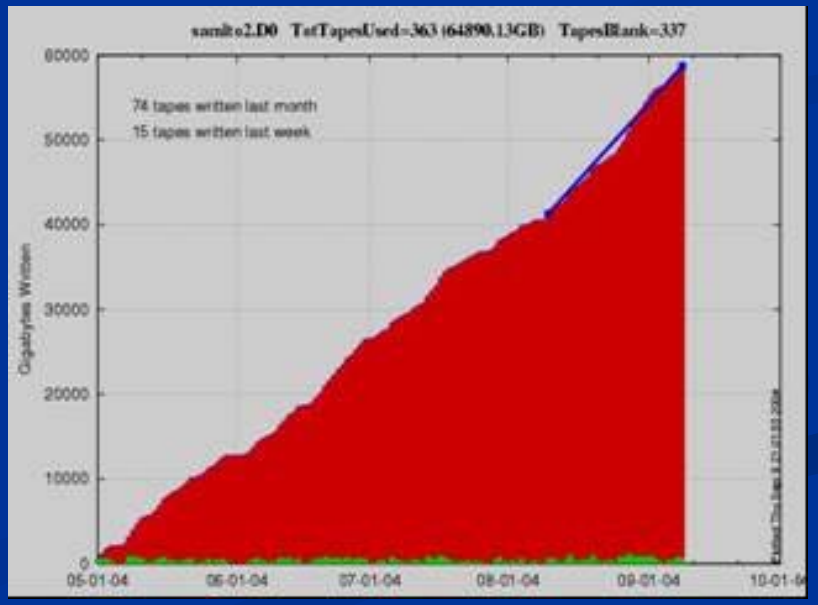
- Find for every drive we own, we replace 1/yr.
- With our usage patterns, we haven't discerned cartridge/drive reliability differences between LTO and 9940
- Lots of tape, drive interventions
- Large distributed system requires complex correlation of information to diagnose

Performance









Conclusion

- Constructed a Multi-Petabyte tertiary store that can easily scale to meet storage, data integrity, transaction, and transfer rate demands
- Can support different nd new underlying tape and library technologies
- Current store is 2.6PB @ 1PB/yr, 20TB/day expect to increase fourfold in 2007 (CMS tier 1 goes online)