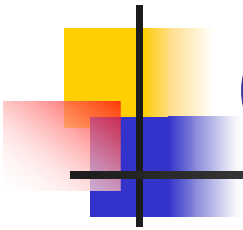
CD Storage and Data Movement for Run II

D. Petravick

Director's Review of Run II Computing

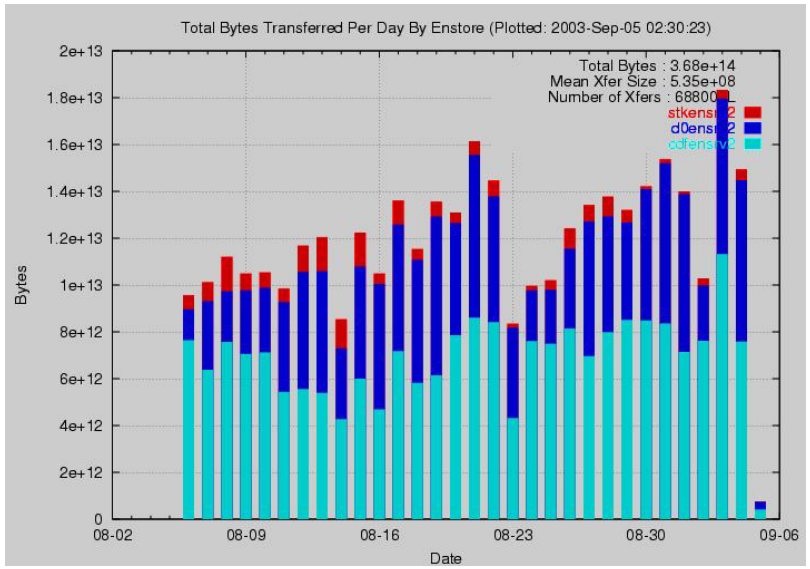


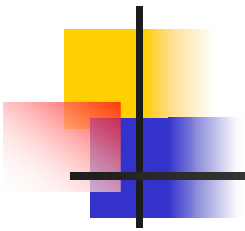
Overview

- Primary Emphasis is on data movement.
- Disks (dCache) fronting tape (Enstore).
- One copy permanent data archive as well.
- Part of a common software infrastructure that include the CMS Tier-1 Facility, LQCD, and other Scientific activities.
- Development.
  - Substantial FNAL Components.
  - DESY, LBL(SRM), GGF(GridFTP), UWisc(NeST).

Enstore Status

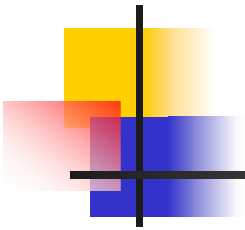
- > 1040 TB total data.
- 10 TB< 1 day < 20 Moved.
- ~50 TB/month ingest.
- Volumes.
  - 9940A > 10000.
  - 9940B > 2400.
  - LTO-1 > 1100.
- 3 instantiations on partitioned hardware.





Enstore Projects

- Federation
- Dispersal
- Migration
- Security
- Lifetime/Maintenance
- FYI – Non RII work
  - Ingest SDSS, KTeV data



Federation

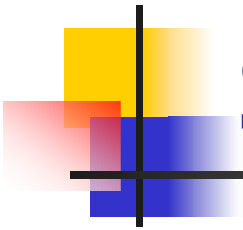
- Currently, 3 instance of Enstore software.
  - A bit balkanized, historical reasons
- Goal of Federation:
  - Increased administrative efficiency, given the evident evolution of these systems.
  - Retain and enhance scalability.
  - Not reducing performance in transition.
- Technical details are being defined:
    - Scalable file, volume clerks, common log files, configuration and accounting, sharing tape drives w/ fair share.

Dispersal

- Goal: Reduce the risk of total data loss due to the data being in one physical location.
- Proposal:
  - Proposal: Volumes in a file family alternate between facilities, subject to facility availability.
  - Disperse, not Duplicate the data.
  - Mitigate induced increased unreliability induced by distributed physical plant.
  - Requires Federation.
- Budget analysis for wide band/current data scenario, work to do.

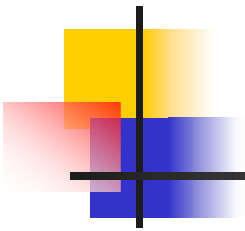
More Automated Migration

- Goal:
  - Conserve Library Slots,Data Center Foot Print.
  - Retire old Technology.
  - Reuse media where possible.
  - Preserve data layout.
- > 10,000 volumes of 9940A media.
  - FY03 ~2000 volumes in FY2003. (400 TB new).
  - Remaining ~7000 volumes (1.4 PB).
- A Project is warranted to increase efficiency.
  - Current practice: Manual QA and nannying



Security

- Write-Protect tabbing.
  - One copy permanent store.
  - Threats: Software error, of malicious employee.
  - Being proceduralized.
    - QA for tab checking underway.
- Security risk analysis.
  - Based on NIST model.
  - At request of FNAL Computer Security Staff.
- Dispersal – already discussed.



Lifetime/Maintenance

- Goal: Continuing improvements to sustain system performance in the face of change.
- Specifics:
  - Maintaining Flexible Media posture – LTO-2.
  - Accounting databases.
  - Libdb -> POSTGRES.
  - Operational sensing and monitoring.
  - Scaling operations – ~70 drives in service.

Administration

Scan three distinct systems (no integrated view – Balkan heritage).

Investigate tape faults – > 10/week.

Deal with other alarms.

Increasingly.

diverse physical plant. (70 Drives).

complex server plant 7servers/system.

Expand to Dcache.

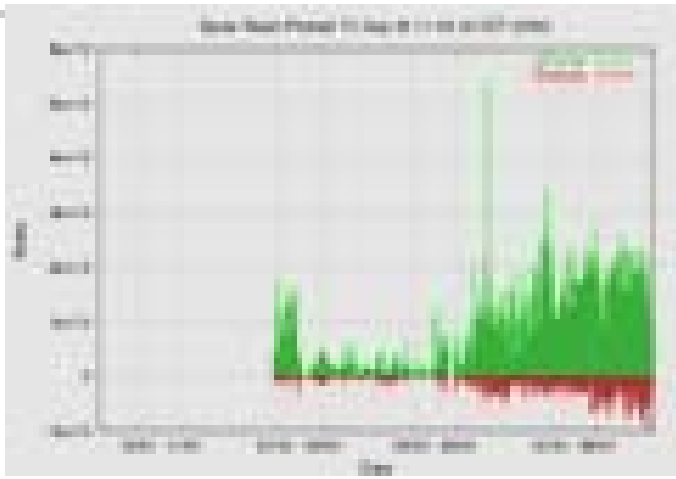
- Figures of Merit:
  - ~3000 mounts/day
  - Mainline Drive Utilization ~100%
  - Run II – 4 Silos, 1 3–QT AML/2

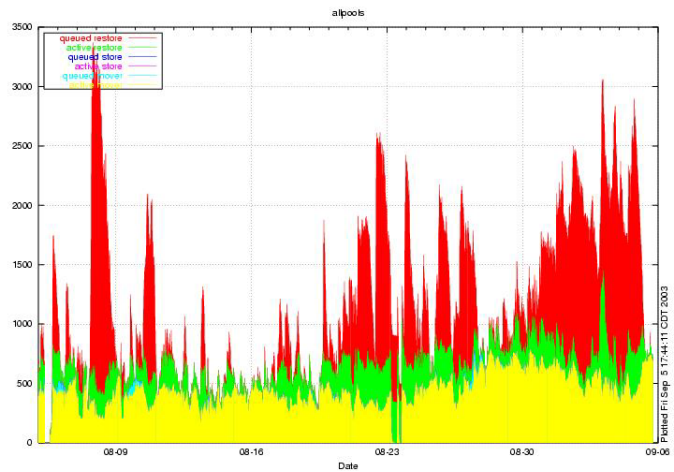
dCache – What Is It?

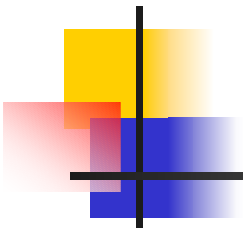
- Software: Scalable, Java based, "user code file system, DESY and FNAL, greater CMS-UF.
- Gives Network attached "file system" on a base of distributed disk.
  - Includes IDE disk on Linux computers.
- Optionally Parameterized with a backing storage system (e.g. enstore).
- LAN interface – dCap.
- WAN/Grid Interface – Ftp(s), SRM.

Dcache Status

- CDF.
  - 35 Nodes w 3 read pools each (68TB).
  - 13 Nodes being added.
    - Convert from static.
  - 5 Selection groups.
- D0.
  - Test system.
  - 2 pool nodes being comissioned.

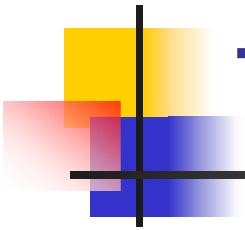






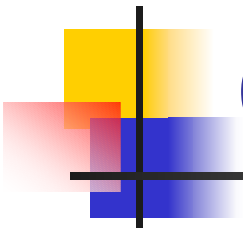
Dcache – Significant Projects

- Tapeless Data-path.
- Scratch Dcache Investigation.
- Grid/Wan interfaces.
  - SRM, FTP, VO.
  - Advanced networking.
- Scaling -- capacity and rate.
- Large File Support – finished, in testing.



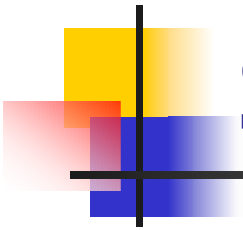
Tapeless Datapath

- Goal: Caching sufficiently capacious to reduce read access to tape.
- Technical Features.
  - Writing at run II scales.
  - Allow read hot spots to not interfere.
    - Replication of requested files to read pools.
  - Strongly authenticated writes.
  - Integrity - CRC checking as data flow.
  - Allow experiment to see when data recorded to tape.



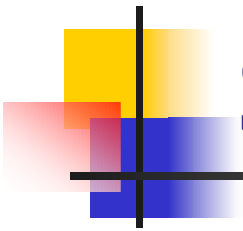
Grid Interfaces

- VO.
  - Allow for VO/SAZ to authenticate.
  - Eliminate our authorization file.
- GridFTP.
  - Substantial work on Grid FTP v1.1 protocol in GGF.
  - Implement V1.1 of protocol to address scaling.
  - Python Grid FTP client.
- SRM.
  - Promulgate management interface.
- Advanced Networking interfaces.



Scaling

- Goal: Investigation to identify scaling problems beyond today's largest production system.
- CDF scaling requests stand out for Run II computing.



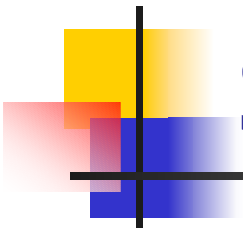
Scratch Cache

- Goal: Provide managed, reliable space on farm node disks.
- Suitable for an analysis environment.
  - Software must not assume sole control of the computer.
  - Software must have sufficiently small footprint.
- Data are NOT on tape.
  - Data are replicated to assure availability.
- CDF request is aligned w/ current CMS-UF hardware plan.

CCF Effort – Whole program

Task	
Administration	2.5
Physical Systems	1.0
Enstore/Tape software	1.8
Dcache software and depl	4.0*

* Includes > 1.0 FTE for CMS.



Summary

We see Enstore and dCache as building blocks useful to Run II for the foreseeable future.

RunII experiments are leveraged across the FNAL scientific program.

Cost: VERY significant non Run II work.

The Components will grow into LHC/ Open Science Grid compatible software, assuring this aspect of Run II computing is aligned with future trends in computing.