CMS: Storage Life and Times

Lisa Giacchetti
CMS Computing Facility Department
Fermi National Accelerator Lab

Supported in part by the Department of Energy
DE-AC02-07CH11359

What's coming

- USCMS T1/LPC storage requirements
- That was then
- Evaluating storage futures
- This is (almost) now

Requirements

- Tier 1
  - Current: 11PB disk, 24PB tape
  - 2015 Pledge: 12PB disk, 32PB tape
  - Access via xrootd, srm, phedex
- USCMS LPC
  - Home area: 2GB default; POSIX compliant
  - Online disk: ~1Tb per individual user; 2+TB for physics groups (overall ~2.5PB)
  - Tape access available

Our Storage Challenges

- Heavy random access from hundreds of jobs running across farms
  - Scale up does not work
  - Scale out does but at what price
- Combining small home area with larger data storage area on one storage instance works until there are performance issues
- Need flexibility and expandability
- Need performance, reliability, stability manageability with low costs

Moving from....

HNas

- HDS HNas (formerly Bluearc)
  - Titan cluster with two 3200 heads
  - 300TB of Hitachi disk (two AMS 500)
  - Holds:
    - home areas (2GB/user)
    - per user quota'd data area (100GB – 1TB)
    - per physics group quota'd data area (1TB+)
    - unquota'd scratch which is automatically cleaned of old files
  - History of poor performance under load from the start.
    - Implemented system to find and stop jobs with too many reads/writes
    - Feb 2013: file systems set RO on workers

dCache

- ~13.5 PB solution comprised of
  - Admin servers: 15
  - Data servers: 300
    - Nexsans of various vintage used for storage backend; newest are E60 180TB units utilizing 3TB disks
  - V1.9
- Files scheduled for migration to Enstore tape backend immediately.
- Deletions from disk occur if space is needed
- Methods available to ensure files stay on disk
  - Pinning: Mainly used by production
  - Resilient dCache: multiple copies of files on separate data servers

Lustre

- Implemented in summer ’10 to handle the production unmerged area.
  - Had been located in dCache but was causing too much load
  - Config: MGS, MDS, 5 OSS’s backed by Nexsan disk, ~150TB disk
  - Added two more OSS’s later to handle network load from merge writes
- Solution worked for unmerged area but management was a significant load especially if something went wrong

EOS testbed

- EOS = CERN's Exploration of Storage file system
  - Used in production at CERN for all LHC experiments
  - Components are implemented as plugins for xrootd storage
  - https://eos.cern.ch/
- EOS Test bed built ~June 2012
  - 1 Main server
  - EOS Data servers: started with 3 and added on as storage became available
  - Access via fuse mount, xrootd
  - Provides scale out architecture like dCache but with POSIX compliant Linux command line access via Fuse.
- Invited users to kick the tires while we worked through its development/ growing pains. It filled a need when Bluearc was not performing.
- By the end of May 2013 had more than 600TB used and it was still not officially in production

Evaluation

Storage Evaluation

- Two projects in parallel:
  - Home/data area update/replacement
  - Project to separate dCache disk from tape
- Overall goals:
  - Reduce the number of file systems we had to manage
  - Provide better performance and accessibility
  - Reduce maintenance costs
  - Implement CMS required disk/tape separation
- Want to end up with 3 storage areas:
  - dCache like area for Tier1 production
  - POSIX compliant online data area for LPC analysis
  - Super reliable, POSIX compliant home areas
- Process: Evaluate for features, get pricing, test if possible
  - Testing done against a 300 node test farm

What Was Considered (1)

- dCache 2.2.7:
  - For: online/nearline data
  - Why: Handles large amounts of data, POSIX interface, performance, good support and long term dev plans
  - Tested: Yes; Set up test stand
- Hitachi HNas:
  - For: home, online data
  - Why: Stable product, a known entity, potential for using CCD* system for homes
  - Tested: Yes ;Baseline tests of CMS T1 installation
- Lustre 1.8.6:
  - For: online/nearline data
  - POSIX interface; Management experience; performance
  - Tested: Yes; against our installation
- EOS 0.2.29:
  - For: online/nearline data
  - Why: POSIX interface, xrootd, easy deployment
  - Tested: Yes; Used our test stand to evaluate
- Isilon:
  - For: home and online data
  - Why: Scale out solution, competitive pricing
  - Tested: No; Talked to other sites that had it installed
- Hadoop 2.0:
  - For: online/nearline data
  - Why: OSG support, additional tools available, POSIX
  - Tested: Yes; set up small test instance

*CCD = Core Computing Division

What Was Considered (2)

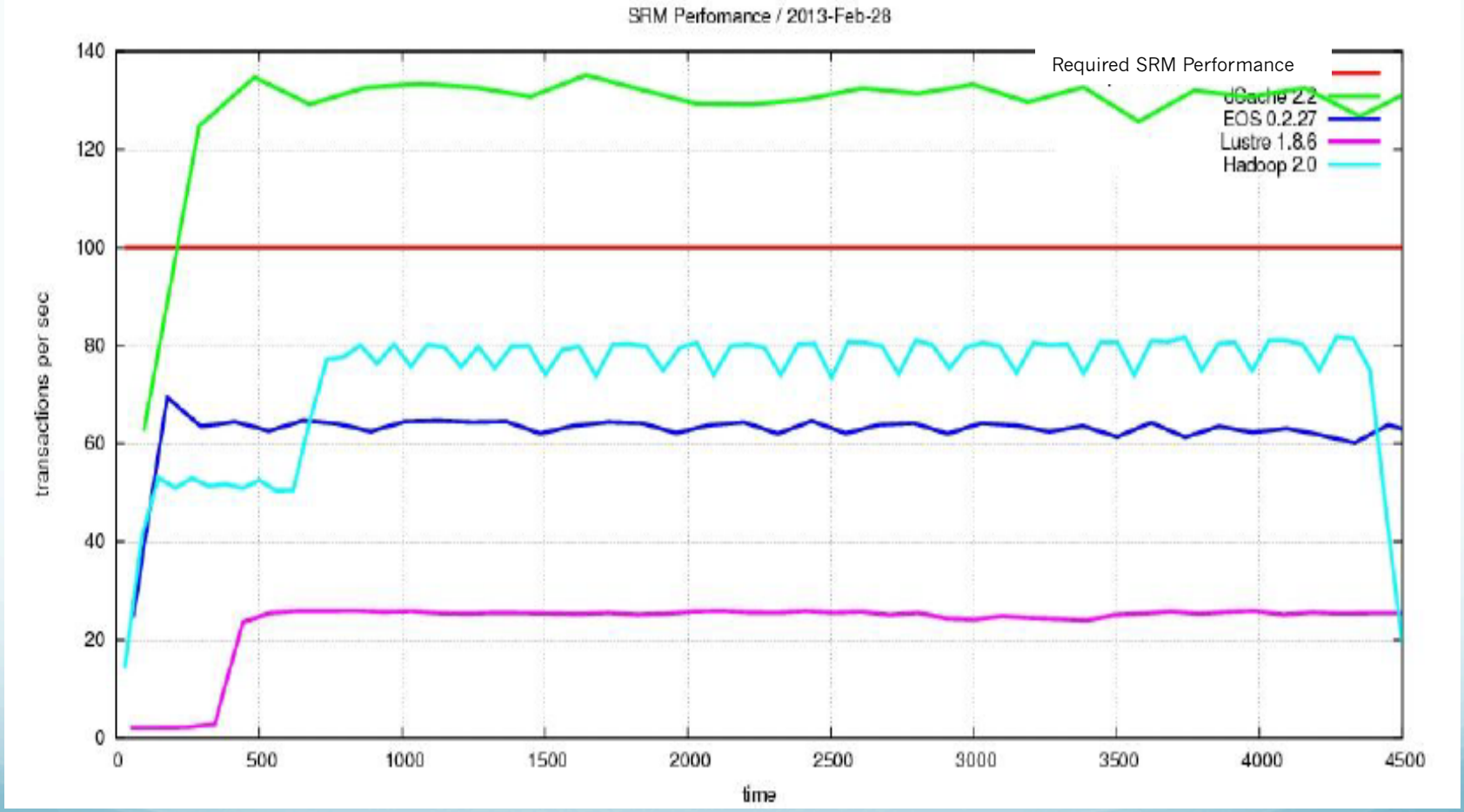
- Nexsan 5000:
  - For: home
  - Why: Good pricing, vendor relationship
  - Tested: Yes, Eval unit tested
- Overland SnapScale:
  - For: home, online data
  - Why: Scale out solution; pricing
  - Tested: No, company did take our scripts and were able to test with them
- Netapp:
  - For: home, online data
  - Why: Big player in this area, new options from them
  - Tested: No
- GPFS:
  - For: home, online data
  - Why: several other HEP sites using it successfully
  - Tested: No

How we tested

- Options that could be tested were limited.
  - Many vendors unwilling/unable to loan out units
  - One vendor was able to take our scripts and run tests at their site in a virtual environment
- Home/online data test:
  - Fermi Disk Test Suite: simulates IO of running jobs; 5 writes, 5 reads across multiple nodes and cores
  - In parallel watch time to write a 10MB file on the command line (interactive performance)
- Nearline storage test (run by Catalin Dumitrescu):
  - 1-1000+ testing threads/node (one file transfer per thread)
  - Pool of 100 files
  - Load increase every second
  - Test using srm, xrootd, dccp
- Advantages of these test procedures
  - Identify service saturation
  - Identify breaking point
  - Easy to find performance vs clients

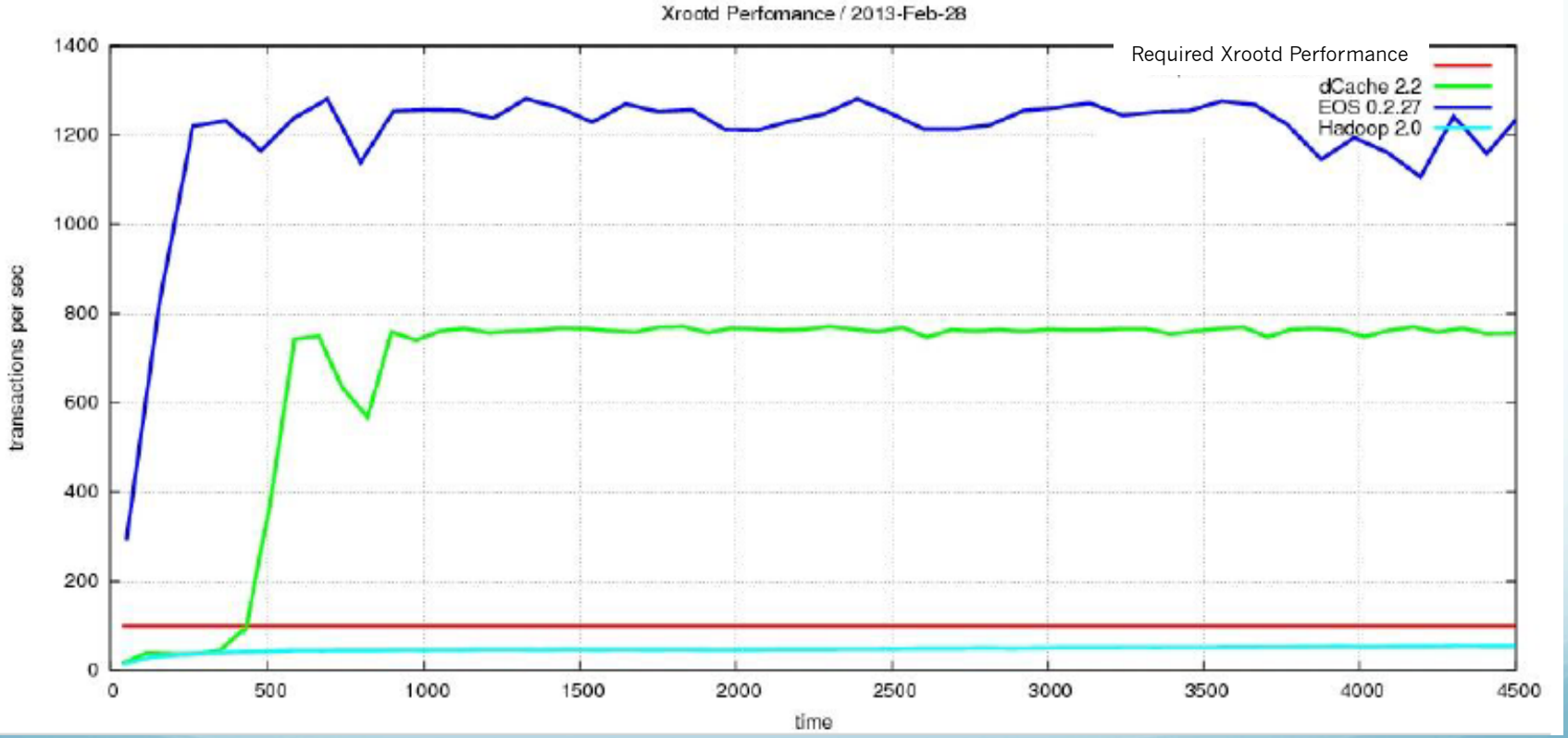
Results: SRM nearline tests

- 270 nodes; thousands of threads



Results: xrootd nearline tests

- Xrootd ops: 270 clients; thousands of threads



Results: home/online tests

- EOS performed very well in all tests we were able to run.
  - Limited by available disk on production system
- dCache POSIX interface not fully compliant; For example can not update files
- Decision to split out replacement of home area from the replacement of data areas (online)
  - Based on money available and EOS performance
- Isilon was a strong candidate for the home area replacement. Competitive pricing, 5 year warranty, others (Jlab, PNNL) using it in production

In the end…

- dCache 2.2 for nearline
  - Solid performance
  - Strong support and development plans
  - Enstore integration
- EOS for online
  - Excellent performance; The magic 10MB write number was always within the 5-10sec we need to see
  - Easy to manage and update/expand
  - Continually refining and adding new features
  - POSIX interface
- CCD Hnas for home area
  - CCD in the business of providing home area storage
  - One less thing for CMS T1 to manage
  - Isilon was an attractive solution but CCD gave us an offer we could not refuse

Where we are now

Where we are now

- Move to dCache 2.2 is in progress
  - This involves the separation of our disk and tape subsystem while preserving all data
  - Will need to educate LPC users on the disk/tape separation changes
- EOS
  - Production ready now: added monitoring, maintenance processes, etc
  - More storage added Nexsan E60 260TB units using 4TB drives (removing out-of-warranty storage from it)
  - Need to merge BlueArc data and current dCache; users will migrate their own files; hopefully we will get them to clean up too
- CCD Hnas/Homes
  - CCD recently installed new cluster
  - Plan to migrate CMS homes to it by end of November
  - CMS T1 admins will have management access and will be first line support for our users
  - Need to migrate our scripts (user additions, quota changes, monitoring)

Some insights

- I was amazed at how many vendors were offering a Linux NFS solution that did not seem to understand Linux at all
- Testing commercial products is hard – very few vendors want to provide actual units to test

Questions?

Acknowledgements:

- Disk/Tape separation Project: Catalin Dumitrescu, Burt Holzman, Chih-Hao Huang, Krista Majewski Natalia Ratnikova
- Dimitry Litvintsev and Catalin Dumitrescu for nearline performance tests
- CMS T1 Team: Burt Holzman, Catalin Dumitescu, David Fagan, Krista Majewski, Rich Thompson, Tim Skirvin, Tony Tiradani, Merina Albert, Natalia Ratnikova, Chih-Hao Huang