Storage Technology Futures

Brian Bockelman

OSG AHM 2009

- We enjoyed sunshine, crawfish, and a mini storage revolt in the CMS T2 session.
    - Nebraska (inadvertently) laid down the gauntlet with HDFS.
    - About 2 weeks later, Nebraska, UCSD, Caltech, and Wisconsin held a pow-wow in San Diego.

2009: The Year of Storage

- UCSD and Caltech ended up switching to HDFS.
- Wisconsin thought about HDFS, but stuck with dCache (too late to change?).
- Purdue appears (very?) happy with dCache.
- MIT - still on dCache, maybe not happy.
- Florida is heading toward Lustre.

Entering 2010

- T2s:
  - Hadoop: Caltech, Nebraska, UCSD.
  - Lustre: Florida.
  - dCache: Purdue, MIT, Wisconsin.
- T3s:
  - Hadoop: UCD, UColorado.
  - Xrootd: Cornell - most likely others.
  - Others? I don’t know; lots of NFS probably.

2010 State of Storage

- Last year, we went through many upheavals. More than 50% of sites made major changes to the SE.
- For 2010, we're concentrating on "nailing things down".
- I believe, despite changes and experimentation, the "state of storage" is stronger than before. We have multiple choices available to each site.
  - We are now a diverse collection of technologies; failure of any one wouldn't be fatal to the program.

HDFS, Lustre

- Status for HDFS and Lustre were given by Mike Thomas and Yujun Wu, respectively.
- These SEs are significant in that the LHC has little-to-no control over the direction the software takes - we're pure users, not stakeholders.

dCache 2009

- dCache has had a pretty big 2009:
  - Chimera is maturing and deployed at many sites. No big disasters in conversions.
  - New pool metadata provider.
  - New info provider.
  - NFSv4.1 support is headed toward reality (I've run at least one job on it!).
  - Fairly quiet on the SRM front - a good thing compared to previous years
- 1.9.5 is "Golden release" - long term supported release, for LHC 2010 run.

dCache

- Sites running dCache in 2010:
    - T1s: BNL, FNAL
    - USATLAS T2: MWT2_IU, MWT2_UC, AGLT2,
    - USCMS T2: Purdue, MIT, Wisconsin
    - T3s/other: Illinois, UConn

OSG and dCache

- OSG maintains its own configuration and packaging of dCache.
  - Current release version is 2.3.4 (based on dCache 1.9.5)
  - Value-add includes storage probes, transfer probes, and integration with OSG GIP.

dCache 2010

- OSG Storage will continue to support current dCache release during 2010.
  - Chimera support is planned - get on Chimera ASAP.
- No other release expected during OSG's currently funded lifetime (updates, critical fixes only).

dCache Summary

- dCache still has a healthy ecosystem of developers and large users.
- I can’t imagine FNAL using anything else!
- It’s still a complex distributed system - several databases, many cells.
- It has controls (such as queueing mechanisms for movers) that provide protections nothing else has.
- Well in-tune with the needs of the LHC community - esp. the T1s.

Xrootd

- In 2009, Xrootd did lots of maturing:
  - Client received better support in CMSSW.
  - Initial OSG support and packaging.
  - Release process, bug tracking, and versioning became appropriate for a collaboration.
- Still no stable/unstable branch, versioning is awkward for sysadmins (latest version number is 20091028).

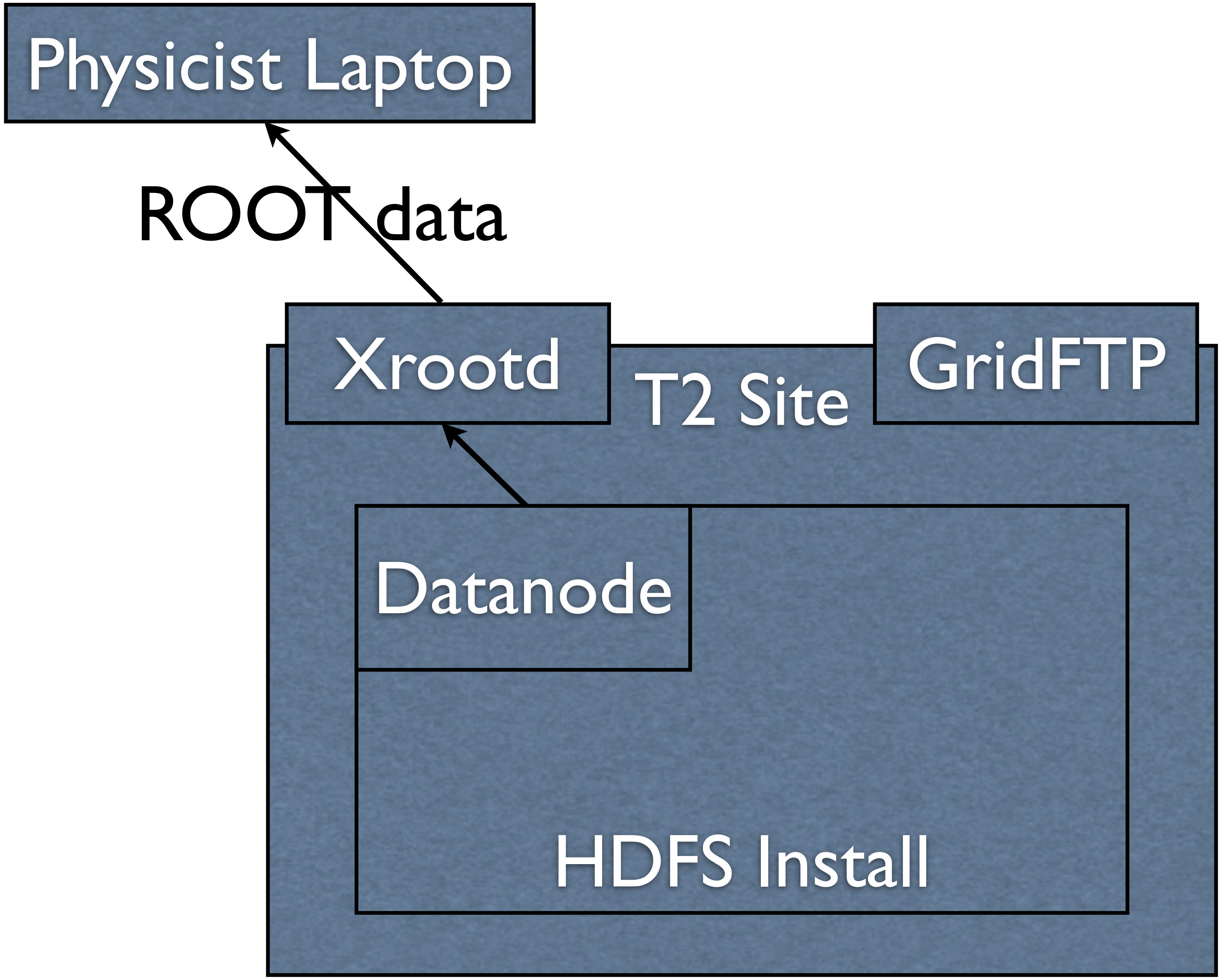
Xrootd in the US

- USATLAS T2s: SLAC/WT2, SWT2
  - MWT2 IU/UC experimented with it last year, but did not it.
- USATLAS T3s: Many (not familiar with the exhaustive list)
- USCMS T3s: Cornell (others?)

CMS and Xrootd

- No CMS T2 site is looking at Xrootd as its SE.
  - Nebraska and Caltech both run Xrootd servers to securely export their HDFS data.
  - Anyone w/ a cert in CMS can run against our site using xrootd.unl.edu

Xrootd at T2s



OSG and Xrootd

- OSG-Storage also provides packaging and support for Xrootd.
  - Popular with USATLAS T3s.
- We complement Xrootd with Gratia probes, BestMan, GridFTP, and do configuration with configure-osg/config.ini
- Caltech packages a separate version for HDFS integration.

Demo

- Fireworks from my laptop

CMSSW I/O

- We all know how crazy CMSSW analysis can get.
    - 4-5 reads per event.
    - 1 KB or less per read.
    - Everything is I/O bound - CPU efficiency around 50%.

CMSSW I/O

- We’ve been working on this!
  - With the current patches, # of reads per job falls by a factor of 10-100.
  - CPU efficiency 90-95%.
  - 1-2 reads per event; working on removing this.
  - Shooting for <1 read/evt.
- If you aren’t using these patches, go for it!
  - https://twiki.cern.ch/twiki/bin/view/Sandbox/CmsIOWork

CMSSW I/O

- 1-2MB/s *average* per batch slot still holds.
  - With our patches, you'll see 10MB/s of activity followed by ~60s of little-to-no IO.
- Latency (<30ms) doesn't matter.
- Tell your users, tell your friends. Get them to apply the patches.

Demo

- HadoopViz highlighting the CMSSW changes

Picking a SE

- With the improvements in CMSSW I/O, there is little to no analysis performance difference between the SEs.
  - And most any SE can support the necessary WAN traffic.
- We are left with factors that aren't easy to measure using Ganglia or Dashboard.

Things to Think About

- Some factors that might influence your decision:
  - System Stability
  - Project stability (how often are fixes needed, how many upgrades bomb)
  - Maintenance costs - one initial attraction to HDFS!
  - Surrounding community

SE Factors, Cont’d

- Existing hardware (Nebraska’s hardware would work poorly for Lustre; Florida’s hardware would work poorly for HDFS).
- Barriers to adoption - R&D needed, amount of effort required to change SEs.
- Specialization - if your site admin quit, how much training does the “new guy” need?

Where does this leave CMS?

- USCMS has always had a strong policy for site control.
  - We don't mandate Condor or PBS, so we shouldn't mandate dCache, HDFS, or Xrootd.
  - We do hold you responsible for the choices you make - you must hit MoU commitments.
- Sites should continuously review what SE their using. SE performance is more than IOPS or GB/s - how does yours measure up?