A highly distributed, petascale migration from dCache to HDFS

Dan Bradley, Sridhara Dasu, Will Maier, Ajit Mohapatra
{dan,dasu,wcmaier,ajit}@hep.wisc.edu

University of Wisconsin - High Energy Physics

HEPiX Fall 2011 - Vancouver, BC
October 24-28 2011

New storage landscape for US CMS Tier2s

Since 2004, seven Tier2 centers have provided analysis, simulation and storage to the US CMS community.

- In 2010–2012, all of the centers will have completed a storage migration
  - Except Vanderbilt, which joined in 2011
- How do we continue to meet our commitments to the CMS community while making big changes?
  - What does it mean that we're making these changes at all?

Wisconsin CMS Tier2

Wisconsin combines storage and compute resources in a hybrid model built on commodity hardware.

- OSG middleware, SL5.7
- 2.5k dedicated slots, up to 2k opportunistic slots (GLOW campus grid)
- 1.1 PB of writable storage
- On a good day:
  - 60k hours (105% of dedicated cores)
  - 20k jobs (burst to 80-90k with fun workflows)
  - 5-30 TB of data transfers

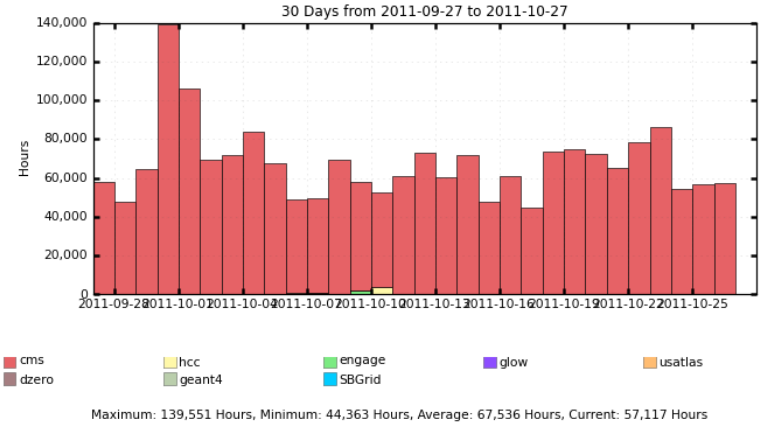

Figure: Hours spent on jobs by VO at Wisconsin, October 2011 (OSG Gratia)

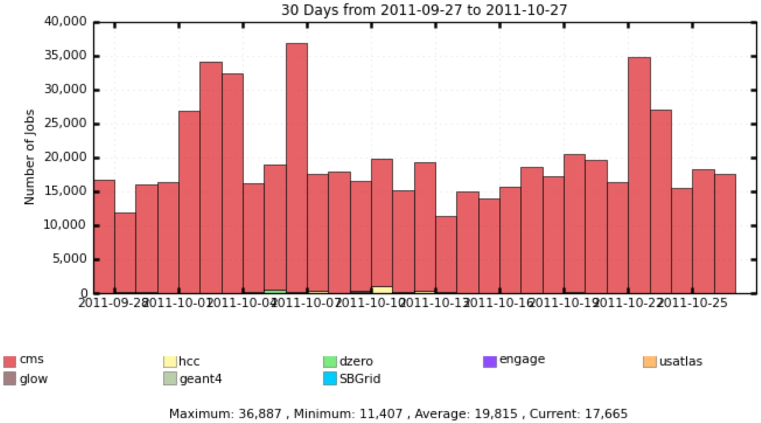

Figure: Jobs completed by VO at Wisconsin, October 2011 (OSG Gratia)

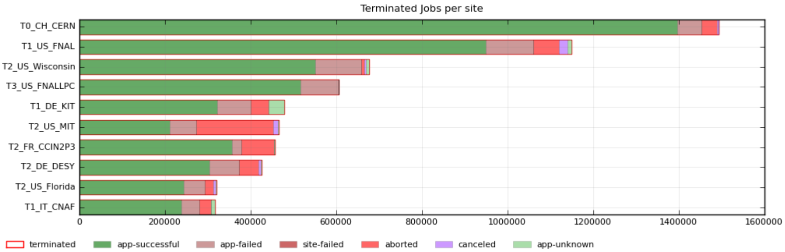

Figure: Top ten sites by job termination status, October 2011 (CMS Dashboard)

2004–2011: dCache at Wisconsin

dCache is feature-rich, highly configurable, stable and performant.

- Dedicated nodes for central services: PNFS, SRM, admin
  - 3 PostgreSQL DBs (PNFS, SRM, companion)
  - Gridftp, dcap, SRM doors
- $\sim$250 pools running on Condor worker nodes
  - Mix of 1U Opteron/Xeon with 1-10 TB of directly attached SATA disks
  - Some dedicated 4U pools with up to 70 TB of SATA RAID
- All disk cache, no tape backend
  - Source and restage official data from FNAL as necessary
  - All data replicated twice locally for performance/availability
- Stable software, strong user community and expert support from FNAL/DESY

2009–2010: Enter HDFS

HDFS is highly fault tolerant and designed to be run on commodity hardware.

- Nebraska, UCSD, Caltech and Wisconsin began evaluation and testing in March 2009[1]
- Approved by USCMS September 2009[2]
- Nebraska, UCSD, Caltech deployed 0.19 in 2010
- Began 0.20 preparations, OSG integration
- Wisconsin waited. . .

[1] https://twiki.grid.iu.edu/bin/view/Storage/HdfsWorkshop

[2] http://indico.cern.ch/conferenceDisplay.py?confId=67969

Why migrate?

- Wisconsin had made a significant investment in dCache
    - Strong relationship with experts at FNAL
    - More than 10 years of combined local experience
- Early interest in HDFS
    - Strong fit for hybrid hardware model
    - Fast, in-memory namespace
    - Simple operational answers to node failure
    - Integration with external monitoring (Ganglia)
    - Large community (Yahoo, Facebook)
- Spring 2011: Chimera or HDFS?
    - Could we migrate to HDFS in less time than it would take to convert to Chimera?

Migration requirements: choose four

A migration from dCache to HDFS should be easy, safe, fast and undisruptive.

- Require little additional effort
- Preserve rollback capability as long as possible
- Minimize disruption to production service
- Provide real-world demonstration capacity before final cutover
- Resolve conflicts between dCache replicas
  - Rare but real; at Wisconsin, $\sim 1/10000$ replicas
  - Caused by odd corner cases (bad disks, power cuts, . . .)
- Exploit existing services
- Be distributed, parallel, idempotent, incremental, monitorable, scalable, throttlable

Options

- Build new cluster
  - Expensive, infeasible
- Wipe and retransfer
  - Consolidate user data on small subset of hardware
  - Convert rest of cluster to HDFS
  - Transfer user data, retransfer official data
  - Simple, slow
- Drain, proxy and fill
  - Migrate replica by replica
  - Use symlinks or HSM staging to proxy reads

Relevant features

Our migration strategy exploited several useful features of both dCache and HDFS.

- dCache stores files as files (not decomposed into blocks)
- dCache gracefully handles external renames and replacements of replicas
    - Happily follows symlinks
    - Serves read from open file handle, then closes and reopens on next read
- HDFS provides a mountable, POSIXish interface via FUSE

Preparation

- Deploy seed HDFS cluster
- Start HDFS daemons on dCache pools
    - Using the same data disks as dCache
- Mount HDFS on dCache pools via FUSE
- Map PNFSids to file names, checksums, metadata

Migration algorithm

- On each dCache pool, scan data directories
- Replace replicas with symlinks into HDFS FUSE mount
  - Could also use dCache HSM staging
- Repeat until all data is migrated
- When replica checksums disagree, choose the larger replica
- Use a simple tree of zero-sized files in HDFS as bookkeeping
  - Or an RDBMS

Migration: March - April, 2011

Drain-proxy-fill migration ran in background without disrupting regular analysis and production.

- Simple shell script[3] running in a loop
- Expanded from server to rack to row over two months
- Migration transfer rates scaled with number of pools
- Reads of migrated files up to 20% slower
    - Migrated files cost network, not disk
    - Large nodes showed worst performance (many more files, same available bandwidth)
- Writes to dCache were unaffected (since not proxied)
    - Picked up on next run of migration script

[3] http://hg.hep.wisc.edu/cmsops/hdfs/migration/file/tip/migrate-pool

Announcement

```
Date: Wed, 20 Apr 2011 17:27:40 -0500
From: Will Maier <wcmaier@hep.wisc.edu>
To: hn-cms-gridAnnounce@cern.ch
Subject: Finalization of HDFS migration 2011.04.21-22

Hi all-

On Thursday, 2011.04.21 at 0700 CDT, we will begin blocking
[write] access to our dCache storage service in order to
prepare for the final migration to HDFS.
```

Cutover: 2011-04-21 07:00

- Disabled writes, synced namespace and data
- Migrated SRM hostname/IP (HDFS services were already running)
- Most of migration time was spent updating PhEDEx scripts and verifying monitoring
- We left dCache running for a few weeks, just in case

All clear

```
Date: Thu, 21 Apr 2011 18:10:59 -0500
From: Will Maier <wcmaier@hep.wisc.edu>
To: cms-physics@physics.wisc.edu
Subject: Notes on the migration to HDFS

Hi all-

The HDFS migration is complete.
```

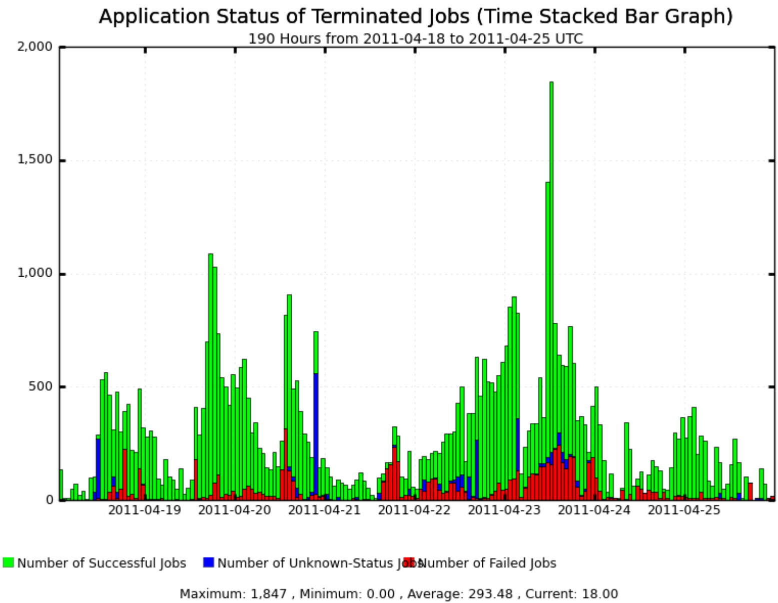

Figure: Application status of terminated jobs at Wisconsin, 2011-04-18–25 (CMS Dashboard)

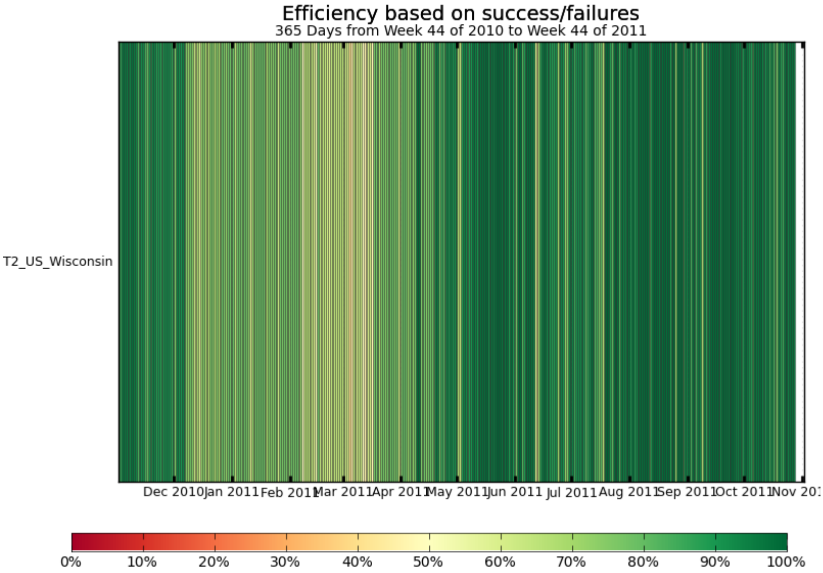

Figure: Job efficiency (successes/failures) at Wisconsin, 2010-10–2011-10 (CMS Dashboard)

Evaluation

In the event, the migration met most of our requirements.

- Minimal degradation of service
- Brief downtime
- Simple and manageable mechanism
  - Easier to use HMS staging?
- Proxied reads on dedicated storage nodes performed poorly
  - Avoid by partitioning placeholder symlinks evenly (or using HSM staging)
- Files that are re-created in dCache may confuse migration process
  - Files removed in dCache are not removed in HDFS without further intervention
  - If recreated files are smaller, simple algorithm loses data
- No other sites have used this technique (yet)
  - Only known migration since Wisconsin used wipe-and-retransfer

HDFS today

HDFS is a viable option for large, production sites on the grid and migration cost can be low.

- At least seven production deployments at CMS Tier2s in US, Estonia
  - More planned (US, Belgium(?))
  - Some Tier3 adoption
- Early adoption outside US
- 0.19 sites upgrading to 0.20 with community assistance
- 0.20 shipped as part of OSG storage stack
- CMS sites active on Apache and Cloudera lists

Closing thought: there's something happening here

HDFS is part of an increasing emphasis on the commonality of our problems across disciplines.

- Not just HDFS, but Chef/puppet, Lustre, native packaging (OSG), DVCS (git), virtualization
  - Leverage experience gained in web/cloud worlds when hiring
- Not just between sites, but between science and industry, HEP and the web
  - Collaborate through open source with innovative companies

Resources

- Wisconsin's HDFS configuration: `http://hg.hep.wisc.edu/cmsops/`
- OSG
  - Twiki: `https://twiki.grid.iu.edu/bin/view/Documentation/Release3/NavTechHadoop`
  - Announcement: `http://www.opensciencegrid.org/Hadoop_Announcement`
  - Packages: `http://vdt.cs.wisc.edu/components/hadoop.html`
- HEPiX 2009 performance comparison: `https://indico.cern.ch/contributionDisplay.py?contribId=16&sessionId=4&confId=61917`