Purdue Tier-2 Site Report

US CMS Tier-2 Workshop 2010

Preston Smith

Purdue University

Outline

- Site Overview
  - Dedicated Capacity
  - Shared Capacity
- Current Resources
  - Computing
  - Storage
  - Networking
  - Hardware status
- Acquisitions 2010
- Storage plans
- Analysis Model/User support
- Development Activities

Computation

- **Dedicated**: **Today**, CMS has access to 2012 computational cores
  - 1240 2.3 GHz 64-bit Xeon cores, 16 GB memory (May 2008)
    - 155 dual-processor, quad-core Dell 1950 systems
    - 16 GB DDR2-667 memory, 2 1 TB disks
    - 10318.35 HSA06
  - 560 2.3 GHz 64-bit Opteron 2376 (**Refreshed Spring 2009**)
    - 70 dual-socket, quad-core Sun Fire X2200 nodes
    - 8 GB DDR2-667 memory, 2 Seagate Barracuda 750GB disks
    - 4728.5 HSA06
  - 212 2.3 GHz 64-bit Xeon cores, 16 GB memory (May 2008)
    - 106 dual processor Dell 1950 systems (Steele)
    - 7056.42 HSA06
  - All running RHEL 5.4
- Total: ~20,103.27 HSA06 (dedicated nodes)

All HSA06 numbers from "the table"

Shared Capacity

- **~15,000 possible opportunistic batch slots**
  - In community clusters
  - BoilerGrid campus grid

- **125,018.46 HSA06 of shared capacity potentially available to CMS at Purdue**

Network Infrastructure

- All nodes have PUBLIC IP addresses
- WAN connections:
  - 10 Gb/s network to TeraGrid
  - 1 Gb/s network to Internet2, via I-Light
  - 10 Gb/s network to FNAL via StarLight
    - Provides access to NLR and major research networks via CIC OmniPOP
- LAN connections:
  - 20 Gb/sec Core (Cisco 6509)
    - 1 Gb/sec connections to Force10 C300

Networking infrastructure **NOT** purchased with project funds

Storage Overview

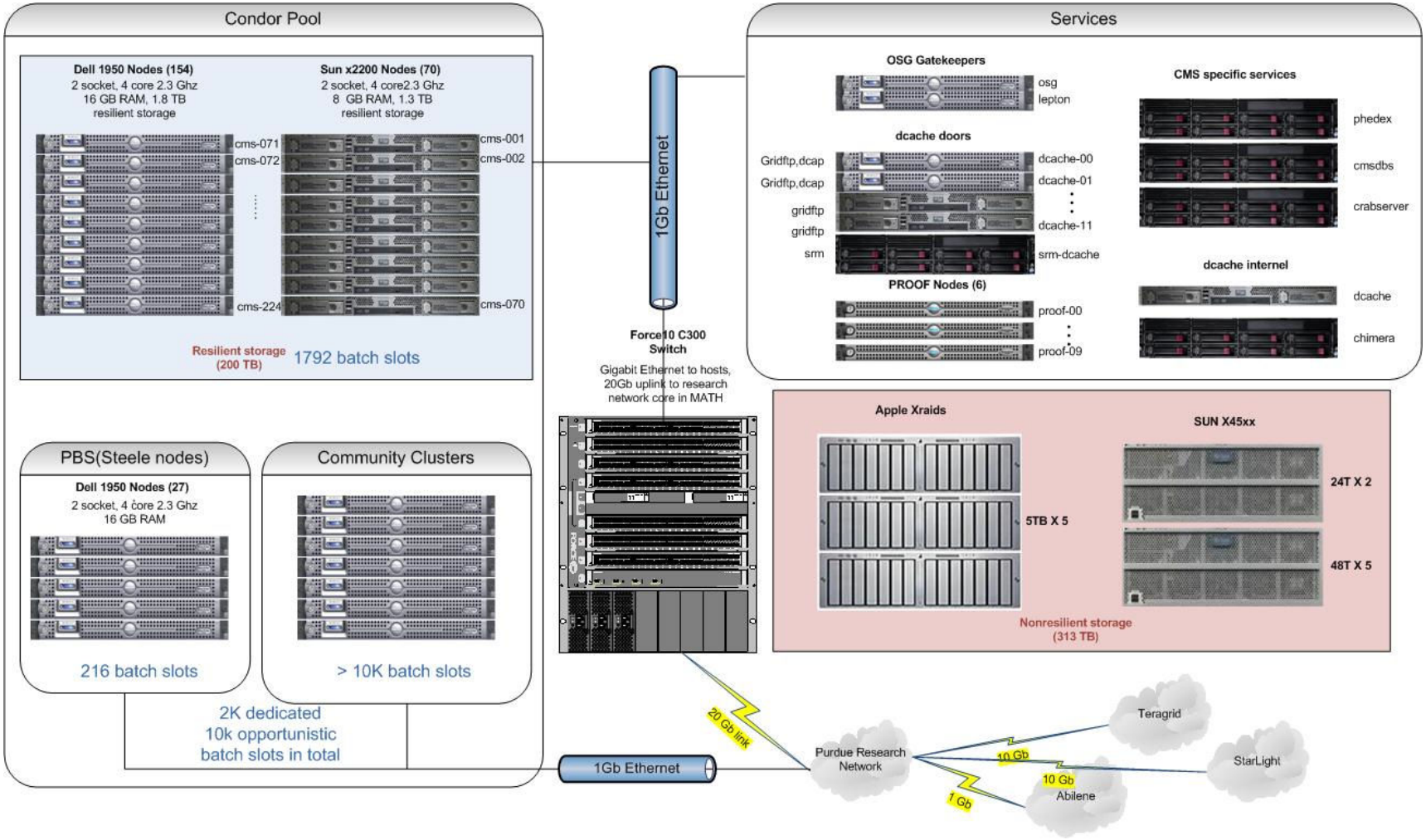
- Home directories:
  - All homes in RCAC served by 60TB BlueArc Titan NAS
    - Local CMS users and users from OSG all get BlueArc space
- General-purpose scratch:
  - NFS - not parallel filesystems
    - Second 120TB BlueArc Titan NAS provides enterprise-wide scratch
    - Shared application space
- dCache:
  - non-resilient dCache, using Apple RAIDs and Sun x4500/x4540
  - Plus resilient pools in worker nodes

BlueArc Storage **NOT** purchased with project funds – provided by Rosen Center

dCache

- dCache system today:
  - Running dCache version 1.9.5-11
  - 6x 5.6 TB Apple Xserve RAID
  - 2x Sun Fire X4500 servers containing 14 TB storage each
  - 2x Sun Fire X4500 servers containing 48 TB storage each
  - 3x Sun Fire X4540 servers containing 48 TB storage each
  - 70 Sun x2200 nodes containing 105 TB
  - 155 Dell 1950 nodes containing 310 TB
  - Resilient capacity: 415 TB
  - Non-resilient capacity: 321 TB
- Total usable capacity of ~525 TB

Current Site Architecture



Equipment in Service

Resource	When Installed
Steele	Spring 2008
Sun Nodes	Acquired January 2007, CPU/RAM refresh 2009
Dell Nodes	Spring 2008
Gatekeepers	Summer 2009
Networking (Force10 C300)	Spring 2008
X4500 “Thumpers”	January 2007
X4540 “Thors”	Summer 2008, Spring 2009
Apple Xserve RAIDs	**2005** *(Servers refreshed 2008)*
dCache servers, chimera, phedex	Refreshed Spring 2009

Acquisition Summary

Early 2005	**Purdue contributes** 50 nodes (100 cpus) of ia32 cluster “Hamlet”
Mid 2005	**Purdue cost-share** purchases approx. 30TB of RAID storage
Mid 2005	CMS Tier-2 acquires 64 nodes (128 cores) of EM64T cluster “Lear” **(FY 2005 project funds)**
Mid 2006	**Purdue provides** 10Gbit connections to StarLight and TeraGrid WAN
Late 2006	**Purdue cost-share** adds 40TB of RAID storage (Sun X4500)
	CMS Tier-2 acquires 70 4-core Sun x2200 nodes **(FY 2006 project funds)**
Early 2007	**Purdue provides no-cost replacement** of CMS’s share of Hamlet with more Lear nodes
Mid 2007	**Purdue acquires** enterprise-class BlueArc Titan NAS systems for central storage, CMS file service migrated to BlueArc at **no cost to CMS**
April 2008	**Purdue cost-share** adds ~140 TB of RAID storage (Sun x4500)
May 2008	**Purdue provides no-cost replacement** of 212 cores of Lear with “Steele”, Xeon E5410
	CMS Tier-2 acquires 100 8-core E5410 Dell 1950 nodes **(FY 2007 project funds)** **Purdue cost-share** adds 55 nodes of the same configuration
	Purdue contributes Force10 C300 network switch for CMS
Feb 2009	**Purdue cost-share** adds ~140 TB of RAID storage (Sun x4540)
Spring 2009	CMS Tier-2 acquires CPU/RAM upgrades forSun x2200 nodes and dCahe Server refresh **(FY 2008 project funds)**

Facilities

- **Except for Steele –**
  - All equipment listed previously are located in new data center space used only by CMS Tier-2

Community Clusters

- Clusters at Purdue are arranged in larger "Community Clusters"
  - One cluster, one configuration, many owners
  - Leverages ITaP's expertise for grid computing (TeraGrid, NW Indiana grid), systems engineering, user support, and networking
  - Today, CMS owns a share of one community cluster
    - Steele: 893 node Xeon E5410 (7144 core, 60+TF)

- Steele installed in 2008
- In 2009 – "Coates" ~8000 core Opteron 2380, 10GbE
  - Another similarly-sized community cluster coming later this spring

Acquisitions for 2010

- In 2009, all expenditures were simply hardware refreshes
  - Sun nodes -> quad core
  - dCache server updates

- **Much of FY09 unspent**
- 2010 Acquisition (April-May) will spend FY09 funds
  - Current plan:
    - Approximately 50% of funds in community cluster
    - Remainder on refresh of Apple RAID systems, add additional non-resilient storage arrays, 10Gb Ethernet for Sun thumpers

Deployment Targets

- Computation is already well in excess of FY10 target
  - Only a fraction of current computation purchased with project funds!
- Storage is ~25 TB short yet –

- ***This spring's acquisition will well exceed the targets***

Operational Issues

- Many lately are networking-related
  - Cable failures
  - Solaris network drivers
  - Packet corruption
- Phedex crashes
  - Seems to have improved recently with v. 3.3.0
  - This hurts our ‘readiness’
- Scratch issues
- Last year’s worth of equipment failures
  - 1 thumper controller replacement,
  - 1 failed system board on Sun node (out of warranty)
  - 2 failed hard disks
  - Dell 1950 SAS cards required reseatings as systems burnt

Storage Plans

- Shared disk
  - We've found the point to where BlueArc scales for scratch
  - CMS will continue to leverage BlueArc for homes and application space
- Storage Element
  - Currently no plans to switch from dCache
  - We have a great deal of operational experience with it – the enemy we know is better than the one we don't

dCache Observations

- Some things work really well
  - Chimera is great – fixes many problems that come along with pnfs
  - System is fast overall
  - We can implement powerful storage policies with combinations of replica manager/PFM and a little scripting
  - gPlazma is flexible for authentication/authorization

- ***That being said:***
  - Some things don't quite do what they promise
    - Secondary groups, ACLs
  - Other things need watched closely
    - dcap doors getting stuck – requires monitoring and automated restarts

Other Storage Efforts

- Staff developing expertise with Lustre storage
  - New community cluster will have new scratch subsystem – possibly a large (.5PB) Lustre
- Hadoop
  - My team using HDFS on other general purpose cluster, as well as providing MapReduce resource to campus
- New mass storage system coming to Purdue
  - Can CMS benefit from Purdue's new HPSS?

- Bases are covered – should USCMS mandate any SE changes, expertise is being developed

Interaction with Physics Groups

- We provide support and resources for the following physics groups:
  - Exotica, Muon, JetMet
- End of 2009: Swap JetMet and B-physics with MIT
- Very good interaction with Muon POG and Exotica PAG
- Each physics group should assign a link person to Tier-2s
  - Large requests should only come from link person or conveners
  - Better communication about priority users would be appreciated
- Allocated disk space is under-utilized by 2 out of 3 groups

User Support

- We support users from the following University groups
  - Carnegie-Mellon, Ohio State, Purdue, SUNY-Buffalo, Vanderbilt
  - /store/user space:
    - Purdue: 14 users
    - Carnegie-Mellon: 5 users
    - Vanderbilt: 4 users
    - SUNY-Buffalo: 2 users
    - Ohio state: 0 users
    - Others: 21 users
  - Local accounts: 28 users

New Grid Submission Portal

Home Register Start About

CMS Grid Submission Portal

Welcome to the CMS Grid Submission Portal

.
Now that the LHC is on its way, start doing analysis of the new data that is provided. This portal is provided by Purdue CMS Tier-2 as a service to the CMS user community to enable the excelent science that the CMS project will provide.

Sincerely, CMS Tier-2 Facility

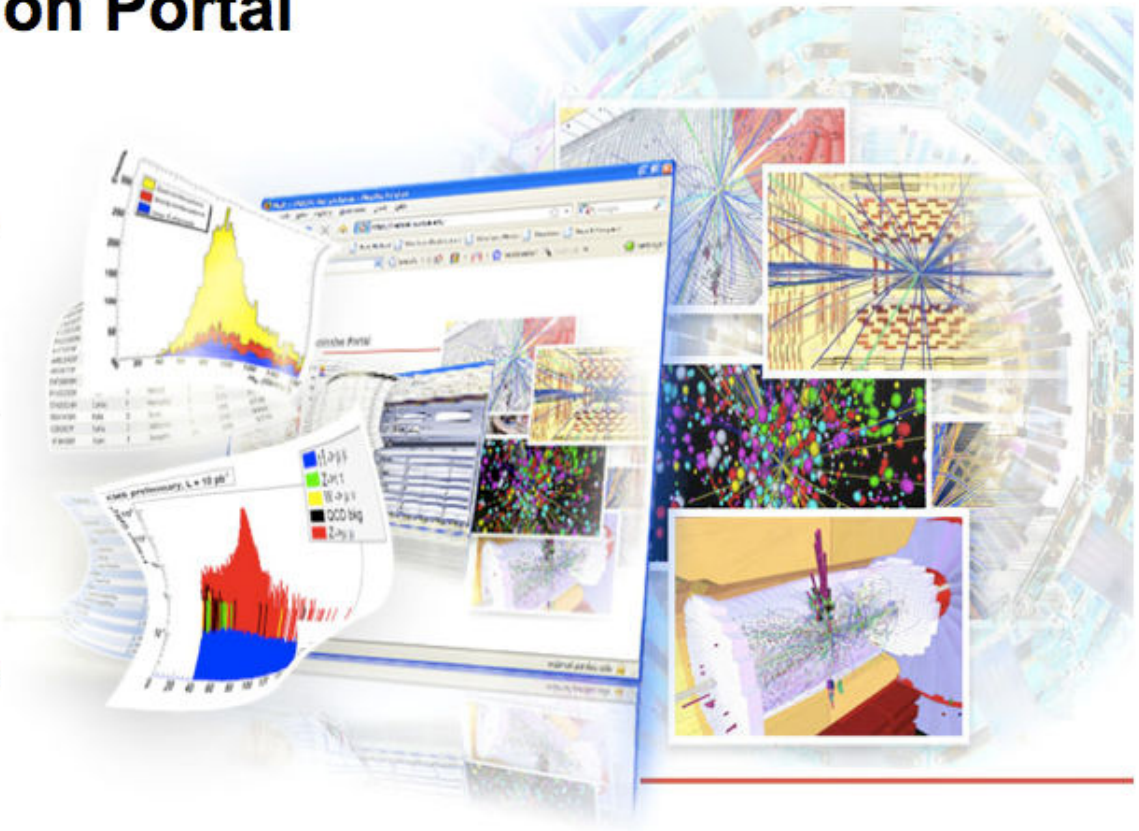
CMS General | OSG News | RCAC News

iCMS interruption 2009-03-19
by Gilles Raymond - Mar 18, 2009
Dear all, due to this ...

iCMS reboot
by Gilles Raymond - Jan 23, 2009
Dear all, IT requested a reboot (after a kernel upgrade) on the iCMS web servers. The web servers will therefore be rebooted ...

iCMS stop
by Gilles Raymond - Jan 19, 2009
Dear all, iCMS will be stopped tomorrow Tue 20/01 between 7.30 and 8.00 AM Reason: update IT authentification certificate no ...

Annual Power cut 4/01/2009
by Gilles Raymond - Dec 19, 2008
Dear all, In accordance to the general power cut on Sunday 4th 2009 from 7.00 AM to 9.00 ...



RCAC Rosen Center for Advanced Computing | ITaP Information Technology at Purdue | Purdue University

Submission Portal

- Woodstock JSF was deprecated
- Evaluation of IceFaces
  - Still requires JavaScript for complex Ajax interactions
  - Missing components limited the retro fit possibility
- Evaluation of GWT
  - GWT seemed much better for providing Ajax and components
  - GXT rich set of GWT based components
  - Using GWTDesigner to provide a graphical layout similar to JSF tools
  - Memory usage can be an issue with GWT applications
  - All method calls are asynchronous
- Work in progress for version 2 with GWT/GXT combination
  - Duplicate the existing functionality
  - Planning on evaluating http://www.cilogon.org/service to replace current authentication/proxy system

Reading Performance Benchmark

- Benchmarked typical analysis jobs
  - using CMSSW_3_4_1
  - reading from local disk, from scratch space, from non-resilient dCache (200 concurrent jobs) and from resilient dcache (100 concurrent jobs)
  - using cmsRun and FWLite executable

	local disk CMSSW	scratch space CMSSW	non-resilient dCache	resilient dCache	scratch space FWLite	local disk FWLite
Speed (evts/sec)	131.46	52.64	29.59	26.87	53.06	97.59

Side Remark

- The original funding period of the US-CMS Tier-2 program was 2005-2009.
- When we started the project at Purdue we negotiated a very attractive agreement with the University
  - Cost share, space in central machine room, free networking, central IT support, etc.
- In order to re-negotiate the deal, an official statement from the project management about the continuation of the US-CMS Tier-2 would be extremely useful.
  - Otherwise there is a risk to lose part of the potential leverage