PURDUE CMS

PURDUE USCMS Tier-2

Compact Muon Solenoid Experiment

http://www.physics.purdue.edu/Tier2/

# Purdue Tier-2 Site Report

US CMS Tier-2 Workshop
LIGO Livingston
March 3, 2009

Norbert Neumeister, Tom Hacker, Preston Smith,
Fengping HuHaiying Xu, David Braun

Purdue University

Presented by Preston Smith

# Outline

- Community Clusters
- Site Overview
  - Dedicated Capacity
  - Shared Capacity
- Resources
  - Networking
  - Storage
- Acquisitions to Date
- 2009 Plan
- User Information
- Development Activities

# Community Clusters

- Clusters in RCAC are arranged in larger “Community Clusters”
  - One cluster, one configuration, many owners
  - Leverages Rosen Center'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
- New cluster “Coates” coming online in spring 2009
- And “Abell” in 2010…. and so on…

# Computation

- **Dedicated**: **Today**, CMS has access to 1750 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
    - 3963k SI2k
  - 288 2.2 GHz / 1 MB cache 64-bit Opteron 2214 (Jan 2007)
    - 70 dual-processor, dual-core Sun Fire X2200 nodes
    - 4 GB DDR2-667 memory, 2 Seagate Barracuda 750GB disks
    - 448k SI2k
  - 212 2.3 GHz 64-bit Xeon cores, 16 GB memory (May 2008)
    - 106 dual processor Dell 1950 systems (Steele)
    - 678k SI2k
  - All running RHEL 4.7
- Total: ~5089k SI2k (dedicated nodes)

# Shared Capacity

- **~8000 possible opportunistic batch slots**
  - In community clusters
  - BoilerGrid campus grid

- **25.57 M SI2k 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)
  - CMS dedicated equipment in CMS machine room (MANN)
    - 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 “Thumpers”
  - Plus resilient pools in worker nodes

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

# Facilities

- Still unused capacity in CMS machine room for upcoming acquisitions
- New data center spaces on the drawing board for 2010 and beyond

New spaces large enough for two clusters even larger than Steele

# dCache

- dCache system today:
  - Running dCache version 1.8p15
  - 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 528 TB

# 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)**<br>**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) |

# And This Year?

- At target capacity now, with minimal investment of project funds.
  - FY08 funds are not yet spent, FY09 funds will not be spent in 2009
- A dollar spent yesterday will buy less compute power than it will a year from now

- **Spending project funds as late as possible maximizes CMS’s investment in hardware**

# 2009 Specifics

- Some 2008 funds will be used in hardware refresh
  - CE Node hardware upgrade, replace older servers
    - For example, PNFS server, phedex system is early 2005-vintage Dell 1850
    - Apple Xserve RAID systems date to early 2005
      - Replace with additional resilient capacity in fall

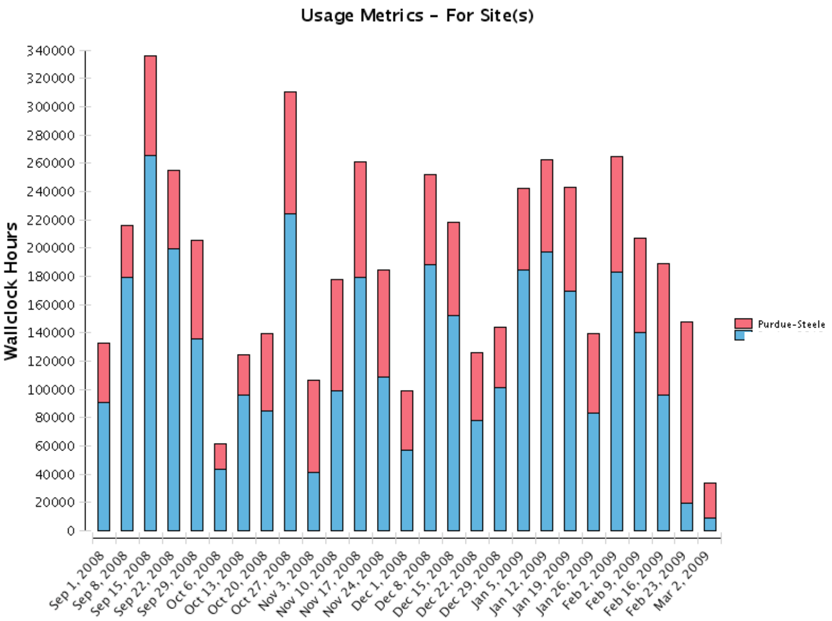
  - Remainder of FY 2008 funds will buy compute capacity
    - Upgrade dual-core Sun nodes to Shanghai?
    - Buy 50-100 new dual-socket multi-core nodes? (Fall)

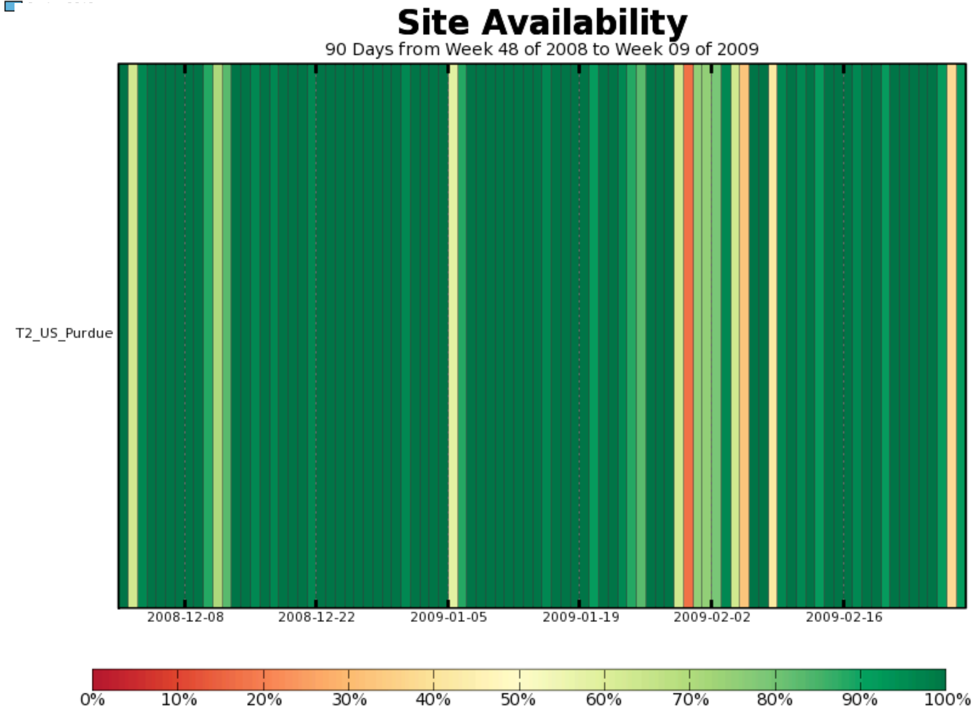
# The Theme for 2009

- A robust facility
  - Increase reliablity
  - Decrease complexity

- For example
  - Multiple CE nodes for redundancy and load balancing
  - SAZ in place now
  - Refresh aging hardware
    - Faster, greater density
  - Improve dCache architecture
    - Split srm from dcache admin host

# Metrics





Rosen Center for Advanced Computing

# Data Hosting

| **Group Name** | **Number Datasets** | **Total Size (TB)** | **Total Num Files** |
|---|---|---|---|
| DataOps | 140 | 71.07 | 22243 |
| Exotica | 44 | 9.433 | 3213 |
| JetMet | 29 | 17.60 | 5867 |
| Muon | 15 | 58.35 | 70548 |
| Purdue | 9 | 75.67 | 26962 |
| **Totals** | **237** | **232.1** | **128833** |

Special requests are accepted…

Hosted 50 TB extra for a two week span for JetMet on top of what is above

# Problems

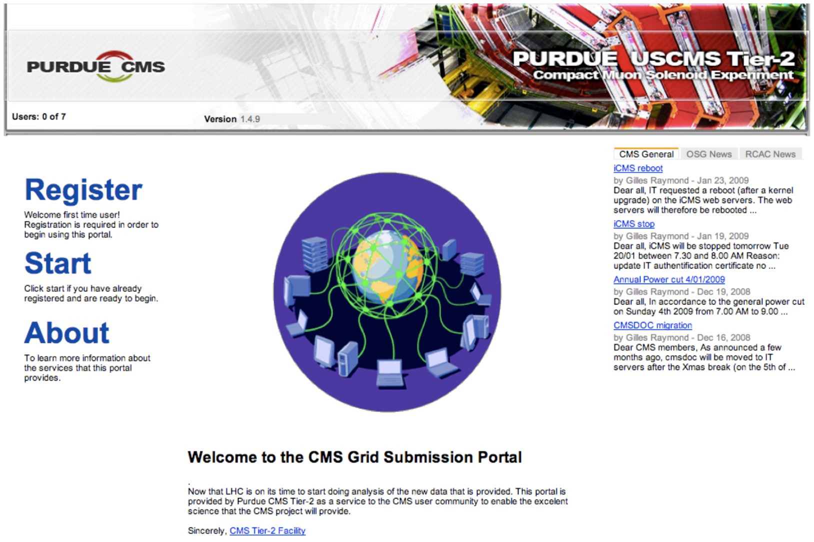
- No operation can go 100% trouble-free..
  - Nscd process spinning
  - Gatekeepers overloaded
  - Disk quotas exceeded with out-of-control output
    - Madgraph productions
  - Facility-related problems
    - 3 unexpected outages (power or chilled water) at CMS machine room in one year
  - Equipment failures
    - Loss of 4 hard drives in RAID pools in a little over 2 years
    - Node failures minimal – 4-5 hardware failures in 2006 and 2008 equipment

# Resources for Users

- Interactive login node
  - CRAB submissions, direct submission into batch queues
  - AFS access
  - Most recent CMSSW versions
- PROOF cluster
  - 8 nodes for PROOF
- Any user working in associated physics groups can potentially get an account
  - With an account, a user gets BlueArc access, dCache access
- Documentation and User support

# Development Activities

- CRAB Portal
  - Job submission to both a local crab and crab server.-
  - VOMS support for proxy generation.
  - File templates for crab.cfg and pset files.
  - Simple wizard for basic crab.cfg configuration.
  - File browsing and download.
  - Sharing user defined projects.
  - Project cloning.



Users: 0 of 7

Version 1.4.9

**Register**

Welcome first time user! Registration is required in order to begin using this portal.

**Start**

Click start if you have already registered and are ready to begin.

**About**

To learn more information about the services that this portal provides.

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 ...

CMSDOC migration
by Gilles Raymond - Dec 16, 2008
Dear CMS members, As announced a few months ago, cmsdoc will be moved to IT servers after the Xmas break (on the 5th of ...

**Welcome to the CMS Grid Submission Portal**

.
Now that LHC is on its time to 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

# OSG Activities

- Purdue team involved in OSG integration
- Recently completed work to standardize advertisement of MPI capability, and simplify execution of MPI jobs through Globus

# Questions?