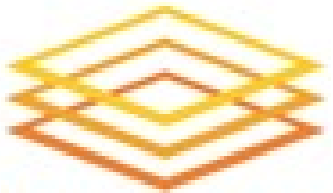
Tier 3: CMS Planning

Rob Snihur
University of Nebraska, Lincoln

OSG Storage Forum
Fermilab



USCMS Tier 3 Overview

- About 20 US Tier 3 sites exist
  - Various hardware/software configurations & support levels
  - Expect many more in the next year
- Goals:
  - Easy startup & monitoring
  - Minimize admin while operating (0.25 FTE)
  - Efficient data analysis (& MC production)

T3 Use Cases

- **Analysis**
  - **Full CMS framework: cmsRun exe, submit via CMS Remote Analysis Builder (CRAB)**
  - **ROOT/PAT ntuple analyses**
- **Monte Carlo production**
  - **opportunistic**
- **Derive alignment, calibration constants**
  - **Short intense projects**

US T3: Resources

GRID-enabled = 91% *(e.g., Minnesota is not)*
Only local users = 36% *(e.g. Minnesota allows collaborators)*

Allow CMS users = 64%
Allow non-CMS users = 55% *(Maryland allows only CMS VO)*
Priority to local users = 91%

Priority policy ranges from strict enforcement to lax
- *FIT local:USCMS:other = 100:20:10*
- *Maryland "All CMS users map to single account so local batch users win"*

Some sites do not (yet) have resources to allow non-local users (i.e., no SE or even CE).

Several sites plan to open up resources to the CMS VO.

US T3: Support

Each T3 site is supported by up to a few individuals

- grad students, faculty, USCMS software engineers, campus computing staff
- they usually have other responsibilities as well
- they install and maintain non-CMSSW software

Bockjoo Kim (Florida) installs CMSSW on any T3 if wanted.

USCMS Tier 3 coordinator: Bob Clare (UC Riverside).
USCMS dedicated T3 support person:
Rob Snihur (@FNAL) & Doug Johnson (¼ FTE @Colorado)

Additional support from staff at FNAL, OSG, and at T2s.

- dedicated hyper news forum for osg-tier3
- community-support meetings every other week

Survey - Hardware

- **Head nodes**
  - **Most sites have a single node**
  - **A couple have multiple**
- **Storage**
  - **~50% have a single storage element (SE)**
  - **0.1 – 100 TB**
  - **Raid boxes: RAID5, RAID6 ==> O(10) TB**
  - **nfs mounted**
  - **No tape storage**
- **Worker nodes**
  - **From 2 to 1400 (Vanderbilt) cores ; generally 10's to 100's**
  - **Many sites planning to expand**
- **Clusters**
  - **Most sites have a single cluster**
  - **Princeton, Texas Tech have more**

A proposed $100k Tier 3

Assumptions:

- 6 physicists, (1.4 + 1) TB each
- Process sample in 24 hrs
  → 16 nodes w/ 8 cores each
- Flush & update sample in 12 hrs
  → 600 Mb/s networking

- Upgradeable RAID chassis: **$33k**
- 16 worker nodes: **$41k**
- 24-port Gigabit switch: **$12k**
- 3 server nodes: **$9k**
- Racks & infrastructure: **$5k**

Install of a Tier 3

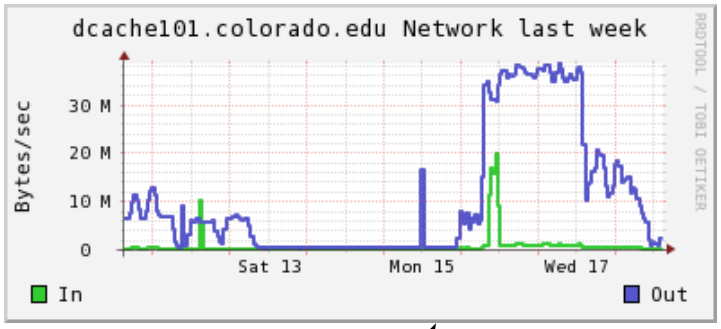
- **OSG: CE & SE, Worker Nodes**
- **BeSTMan**
- **PhEDEx – transfers data files**
- **CMS SoftWare (CMSSW)**
- **Squid (i.e., dbFrontier)**
- **Certificates, passwords, register via web pages, keys, copies of config files for different functions, ...**
- **Many steps, see Malina's guide**

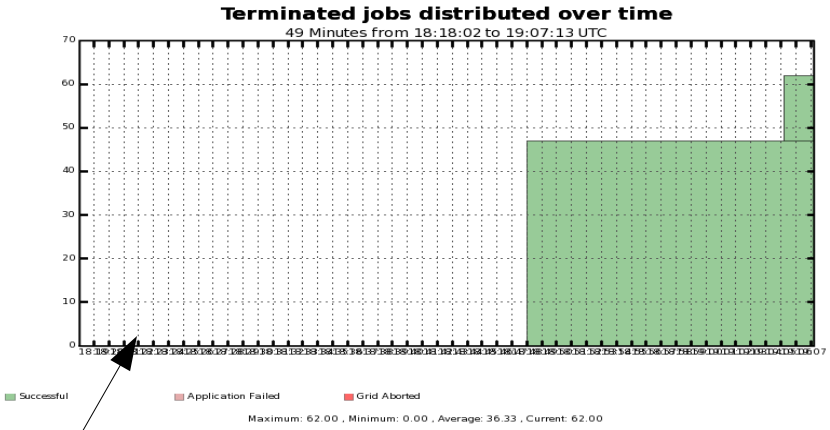
  http://hep-t3.physics.umd.edu/HowToForAdmins.html
- **Can we:**
  - **Simplify?**
  - **Standardize software stack & configuration?**
  - **Automate?**
  - **Not require root access?**

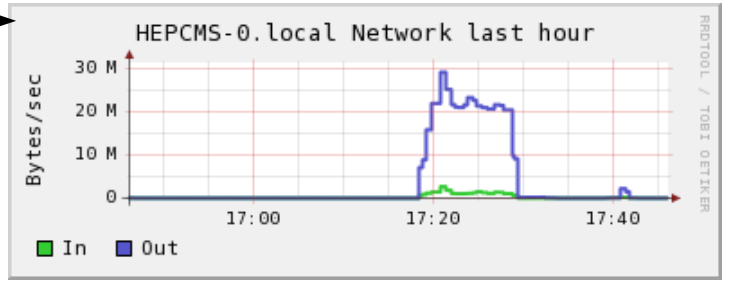
Types of Storage

- Data requirements/capabilities need more exploration
- BeSTMan
  - Simplest & most common on USCMS Tier 3s
  - VDT install
  - Full mode vs. gateway mode
- ReDDNet (see Kevin Buterbaugh's talk)
- Xrootd, LUSTRE, FUSE ?
- HADOOP? (see yesterday's talks)
- Learn about and test various solutions
  - T2 ==> large T3 ==> small T3
  - ATLAS

Monitoring







- Ganglia monitoring
  - Shows the cluster as a whole
- CMS dashboard
  - Shows limited info about your jobs
- Missing:
  - When can I expect my jobs to run & complete?
  - Are my jobs efficient or what are the bottlenecks (e.g., I/O)?
- Many metrics and statistics in job output produced by cmsRun

Questions

- When does nfs breakdown?
- File servers versus data disks on worker nodes
  - Which is cheaper?
  - Which has faster performance?
  - Which is easier to maintain & more reliable?
- How to compute on worker nodes with local data?
  - Wait for batch slots
  - HADOOP?
- Ntuple analysis: interactive vs. batch
  - PROOF
- How much storage is needed?
- How to simplify administration?

Virtualization

- Useful for test stands, improve/automate installs, reproduce problems at sites, security
- A tiny Tier 3 has been built at FNAL on a virtual machine
  - Chose xen technology
  - Flexible, grow into bigger site
- Can we package & distribute ~fully-installed worker and/or admin nodes?
- CERNVM?
  - Rpath handles hostnames
  - Hidden IPs?

Future

- ATLAS has 4 types of T3
- CMS: Start with two?
  - 1) $100k 2) Something larger?
  - open/closed to CMS VO?
  - Datasets: Must consider data formats (RECO vs AOD etc.)
- Federation of T3s
  - With your regional T2 for data
    - What do you want from T2?
  - Among peer T3s
    - Florida T3s experimenting with LUSTRE
- Virtualization
  - Virtual worker nodes and head nodes provide flexibility
  - Distribute fully-installed appliances?

Summary

- About 20 US Tier 3 sites exist
  - Various hardware & software configurations
  - Short term: get them up and running, simplify installs
  - Expect many more in the next year
- Storage:
  - BeSTMan primarily
  - File servers with RAIDs shared via nfs
  - As # cores in cluster grows, will need a scalable file system
  - HADOOP, ReDDnet, xrootd?