11-July-2011
Kurtis F. Johnson
Dept. of Physics
Florida State University

Final Report on the FSU CMS Tier 3 System

SUMMARY:
In August 2010, the Florida State University HEP group was allocated K$68.5 in ARRA funds to build a CMS Tier 3 computer system. We have constructed a powerful and very cost-effective GRID-aware system comprising a head node, storage element (SE), compute element (CE), SQUID server, 3 RAID-6 storage units totaling 108 TB and 47 "compute nodes". In total, the system has ~400 cores controlled by a Sun Grid Engine (SGE) job manager. The GRID interface is provided by the OSG 1.2.19 package. Jobs may be submitted locally or via the CMS CRAB server. Our Tier 3 is in heavy use by the CMS Higgs and SUSY analysis groups, by our own HEP theory members and is available (and used by) other CMS members when CPU cycles are not otherwise in use. The OSG Site Metrics reports more than 2500 jobs/week.

Tier 3 centers are not allocated system administration resources, but must provide their own. Because of the heavy system maintenance load typically associated with computing systems, we have emphasized automating system tasks. For example, transfer of "official" CMS data is via an automated PhEdEx server to one of our RAID 6 storage units. To automatically update CMSSW software, an autonomous daemon polls the CERN download site and installs the latest CMSSW version within 2 hours of posting at CERN.

A unique aspect of the FSU site is that the hardware is assembled at FSU. Over the last few years, the FSU HEP group has developed in-house expertise in computer system assembly so as to maximize the performance per dollar. We realize a factor of about 2.5 increase in performance compared with purchasing similar equipment from commercial vendors.

Tier 3 Requirements:

The purpose of a Tier 3 site is to enable a group to carry out their analyses without burdening Tier 2 sites. The FSU Tier 3 site is designed to have both GRID and local interfaces so that jobs may be submitted using the CMS CRAB system or locally, directly via SGE; we also desired low maintenance (thus the automated analysis software updating), and to support our analysis activities at least 50TB of RAID-6 storage and substantial computational power.

Software:

To ease maintenance, all machines use the same OS, Scientific Linux 5, regardless of role.
The GRID interface is provided by OSG-1.2.19 and VOMS.
Automated transfer of CMS data is by the CMS PhEdEx package, controlled by our SE running BestMan-gateway. Again, to keep sysadmin workload low, we adopted FNAL central PhEdEx.
Automatic CMSSW updates use Doug Johnson's (U.C. Boulder) scripts.

Security:

With 50 high performance machines and substantial storage, T3 is a tempting target for crackers. Fortunately, all 47 of the compute nodes are connected in a non-routing 10.1.1.x sub-net for which the NAT is done by the t3 headnode; they are therefore not visible to the Internet. The headnode is itself visible only to our local HEP group subnet and thus somewhat protected. There still remain the GRID machines SE and CE, and the RAID units which must remain visible to the internet in order to fulfill their functions. These we try to protect by disabling all unnecessary services and using surveillance software (OSSEC) which scans log files for unusual activity and reports abnormalities to the site admin by email.

System Cost and Hardware:

Since we assemble our computers in-house, the cost of the system, aside from the constructor's time, is completely dominated by hardware costs, which we address below.

We use the same basic machine structure for all machines, based on the Intel Core i7 processor. The Core i7 is a high performance, unlocked, multi-core CPU. It is optimized for execution, not serving; we have benchmarked it on our typical jobmix against other CPU's and found it consistently better in performance/price. (See figure 1.)

We mount the motherboard/CPU onto an open plexiglass hanger which has the form of an inverted "T"; for the compute nodes one on each side (see figure 2). Two machines share one power supply; a "T" can be assembled in less than 40 minutes by unskilled labor. The advantages of this design - besides the low cost - are ease of maintenance and excellent cooling. Our compute nodes are clocked at 3.6GHz using the stock Intel cooler, and do not run hot, resulting in a 20% performance gain for our jobmix. We have checked this by observing core temp while loading the CPU with the mprime test program.

The parts cost of a single machine is ~K$1. Following is a parts and price list, it should be understood that prices are approximate and vary with time:

CPU	$300
Mobo	230
RAM	270
HD(2TB)	100
1/2PS	70
1/2"T"	15
Total	$985

Hardware list of the FSU Tier 3:

t3 (cluster headnode for job submission)
SE (GRID aware storage element)
CE (GRID aware compute element)
SQUID (CMSSW constants cache; this is an older, Q6600 machine)
storage5 (20TB RAID6)
storage6 (44TB RAID6)
storage7 (44TB RAID6)
47 compute nodes (Most compute nodes have 12GB of RAM and clock at 3.6GHz.)
48-port gigabit switch

The RAID units deserve a special explanation as they are, at ~$100/TB, especially cost effective. Each comprises a single machine on its "T", a 3WARE RAID card (type SE9650-24, ~$1100) and 12 or 24 2TB hard drives (less than $100 ea.), with relevant drivers and monitoring systems installed. Our 44TB units cost less than K$4.5.

Ethernet connection graph:

All machines have standard 1Gb ethernet ports. The CE, SE, SQUID, headnode and storage units are connected to the FSUHEP subnet. The compute nodes are connected via the 48-port switch to the headnode which therefore also is the gateway for the compute nodes. The storage units are NFS mounted. If/when our jobmix becomes more heavily weighted towards CMSSW jobs (which tend to be transaction rich), the well known (but poorly quantified) inadequacy of NFS at high transaction rates will lead to rising rates of job failure. At that point we plan to move to the HADOOP storage system, for which we have already
purchased most of the disk drives.

System cost:

Besides the items already discussed, we purchased LAN cables and some LAN and video cards, additional hard drives for the HADOOP upgrade, UPS units, a monitor, DVD reader and other items that were useful during the construction phase. In round numbers, the system cost was:

CE, SE, headnode	-	K$ 3
47 compute nodes	-	47
storage5	-	2
storage 6 and 7	-	9
switch	-	1.2
other	-	6.3
Total		K$68.5

Conclusions:

FSU HEP has built a powerful, GRID aware tool for physics analysis which is already in heavy use by our theory group, our CMS members locally and is available for use via the GRID by other CMS members (see figure 3).

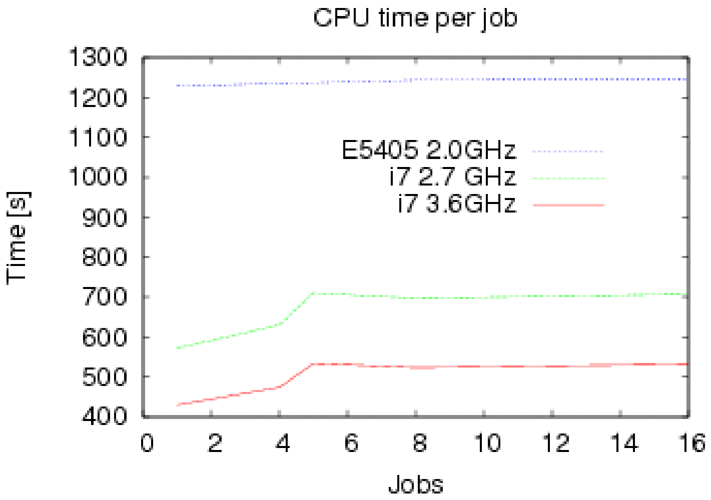


Figure 1: Example of a benchmark comparing the Core i7 execution times to another machine, with the number of concurrent jobs as a variable. Lower times are better.

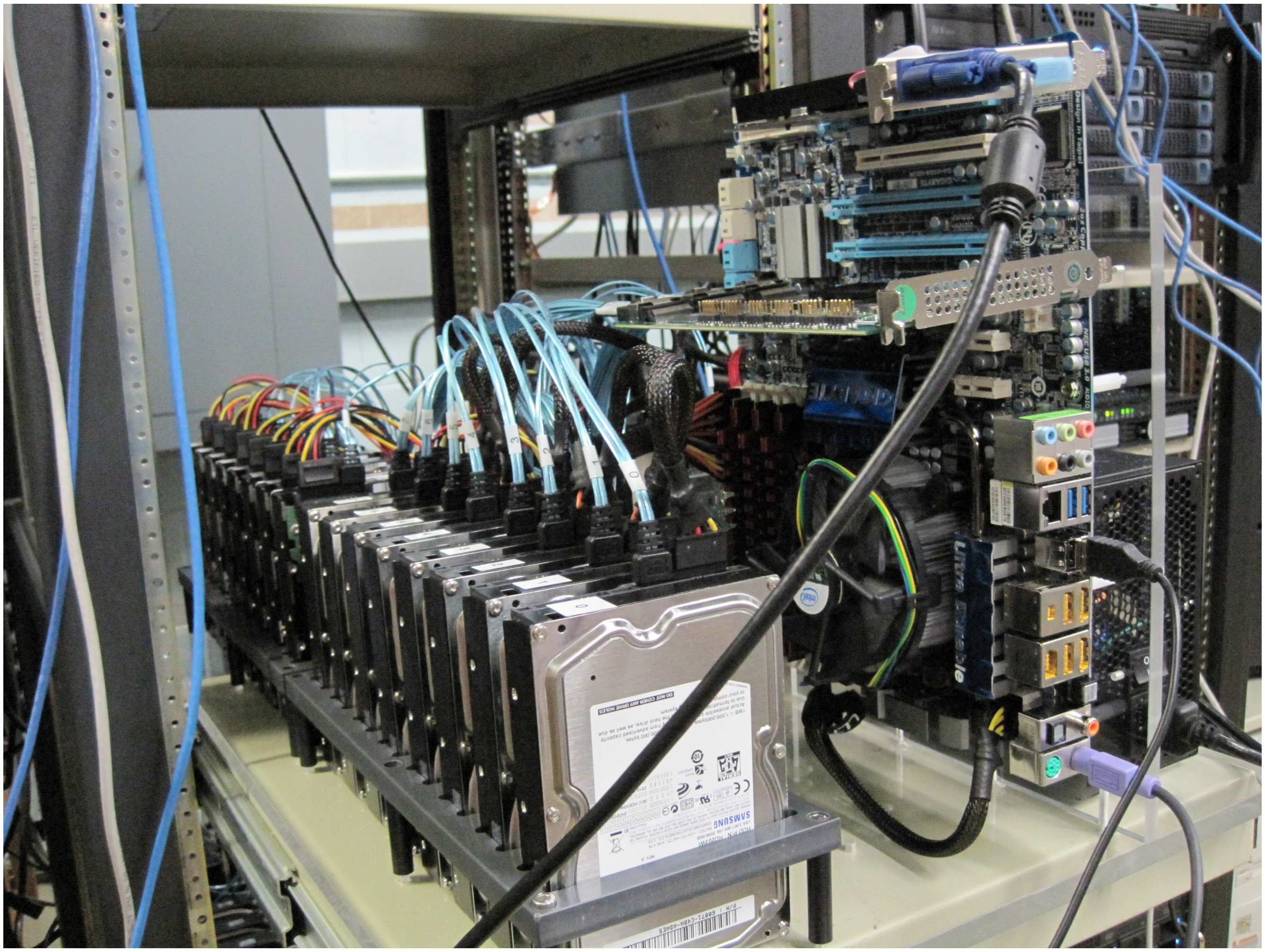
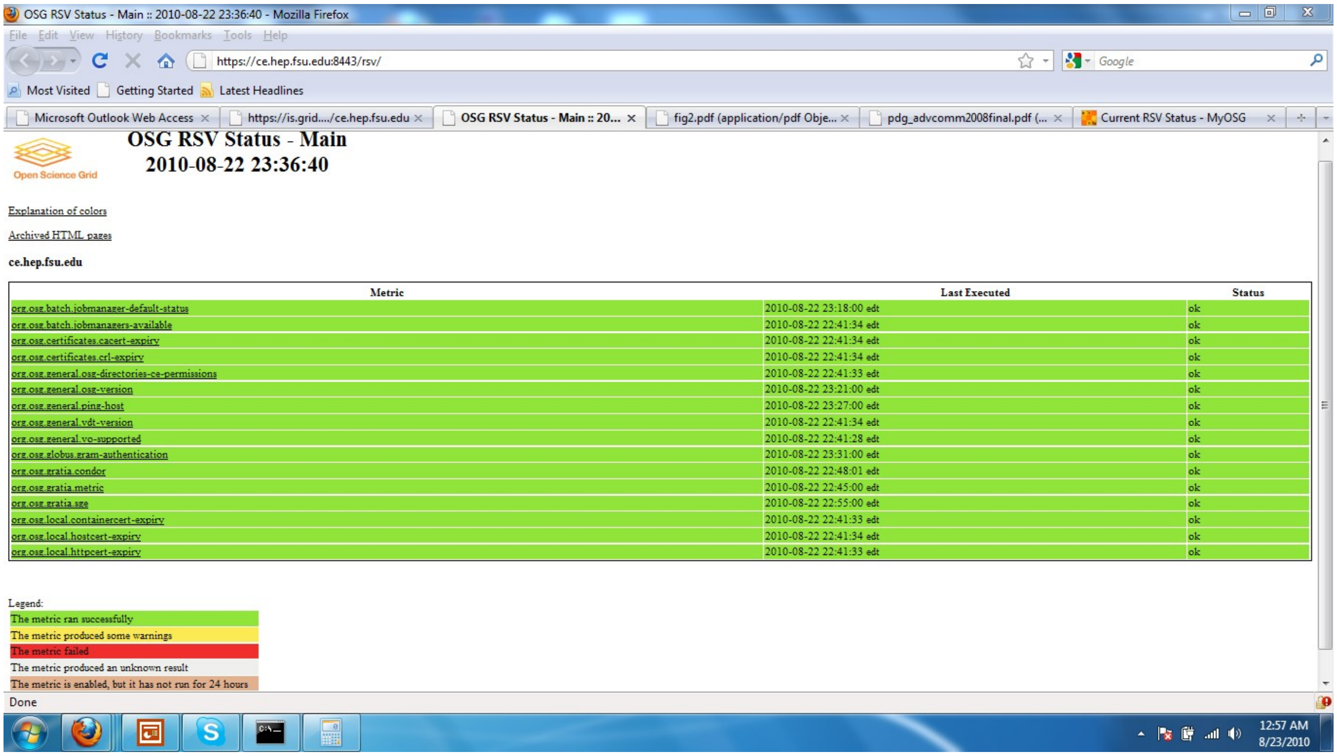


Figure 2: Photograph of the “storage7” 44TB RAID-6 storage unit which illustrates the open architecture of the FSU HEP group's computer cluster. On the right side the motherboard of the host computer is visible mounted on the inverted vertical strut of a plexiglass “T”; on the left side is a row of 2TB disks.


OSG RSV Status - Main :: 2010-08-22 23:36:40 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
https://ce.hep.fsu.edu:8443/rsv/
Google
Most Visited
Getting Started
Latest Headlines
Microsoft Outlook Web Access
https://is.grid..../ce.hep.fsu.edu
OSG RSV Status - Main :: 20...
fig2.pdf (application/pdf Obje...
pdg_advcomm2008final.pdf (...
Current RSV Status - MyOSG
Open Science Grid
Done
12:57 AM
8/23/2010

OSG RSV Status - Main
2010-08-22 23:36:40

Explanation of colors

Archived HTML pages

ce.hep.fsu.edu

Metric	Last Executed	Status
org.osg.batch.jobmanager-default-status	2010-08-22 23:18:00 edt	ok
org.osg.batch.jobmanagers-available	2010-08-22 22:41:34 edt	ok
org.osg.certificates.cacert-expiry	2010-08-22 22:41:34 edt	ok
org.osg.certificates.crl-expiry	2010-08-22 22:41:34 edt	ok
org.osg.general.osg-directories-ce-permissions	2010-08-22 22:41:33 edt	ok
org.osg.general.osg-version	2010-08-22 23:21:00 edt	ok
org.osg.general.ping-host	2010-08-22 23:27:00 edt	ok
org.osg.general.vdt-version	2010-08-22 22:41:34 edt	ok
org.osg.general.vo-supported	2010-08-22 22:41:28 edt	ok
org.osg.globus.gram-authentication	2010-08-22 23:31:00 edt	ok
org.osg.gratia.condor	2010-08-22 22:48:01 edt	ok
org.osg.gratia.metric	2010-08-22 22:45:00 edt	ok
org.osg.gratia.sge	2010-08-22 22:55:00 edt	ok
org.osg.local.containercert-expiry	2010-08-22 22:41:33 edt	ok
org.osg.local.hostcert-expiry	2010-08-22 22:41:34 edt	ok
org.osg.local.httpcert-expiry	2010-08-22 22:41:33 edt	ok

Legend:

- The metric ran successfully
- The metric produced some warnings
- The metric failed
- The metric produced an unknown result
- The metric is enabled, but it has not run for 24 hours

Figure 3: Screenshot of the all-green OSG status board when the FSU Tier 3 GRID interface first became completely operational.