Monitoring the Grid at local, national, and global levels

P D Gronbech [1]
1 GridPP Project Manager, University of Oxford

Abstract

The World-wide LHC Computing Grid is a global infrastructure set up to process the experimental data from the experiments at the Large Hadron Collider located at CERN. The UK component is provided by the GridPP project across 19 sites at the universities and Rutherford Lab. To ensure that these large computational resources are available and reliable requires many different monitoring systems, ranging from local site monitoring of individual components, through UK-wide monitoring of Grid functionality, to the worldwide monitoring of resource provision and usage. In this paper we describe the monitoring systems used for the many different aspects of the system, and how some of them are being integrated together.

1. Local Site Monitoring

Local GridPP [1] site monitoring covers cluster load, batch system utilization, network bandwidth monitoring and fault condition monitoring. Ganglia [2] is the most common software used to monitor a cluster (Figure-1). It is easily installed on clients and allows data to be collected on a master node and displayed via a web server.

Figure 1- Ganglia Monitoring showing a group of Worker Nodes

Ganglia monitors and collects many parameters, such as load, memory used, network activity etc. The top part of the page shown in Figure-1 shows an aggregate picture across the nodes in a particular logical group, such as worker nodes where the main priority is to ensuring they are busy. The web interface allows other aggregate parameters to be selected or the full details for any individual node can be displayed. The data are recorded, allowing time periods from an hour up to a year to be displayed. This can be extremely useful for spotting trends over time. Other parameters can be added to the Ganglia monitoring setup, but many sites are satisfied by the default setup.

Monitoring specific to the batch system used at a site is also typically used. Many GridPP sites use the torque batch system [3] (developed from PBS) monitored with pbswebmon [4], which provides a graphical way to monitor the occupancy of the cluster, and the job shares and efficiencies for each user (Figure-2). This package clearly shows which users jobs are running, something the ganglia programme does not do without modification. The efficiency of the job is also reported which can guide a systems administrator to investigate jobs which may be suffering through lack of I/O bandwidth to the disks or the network.

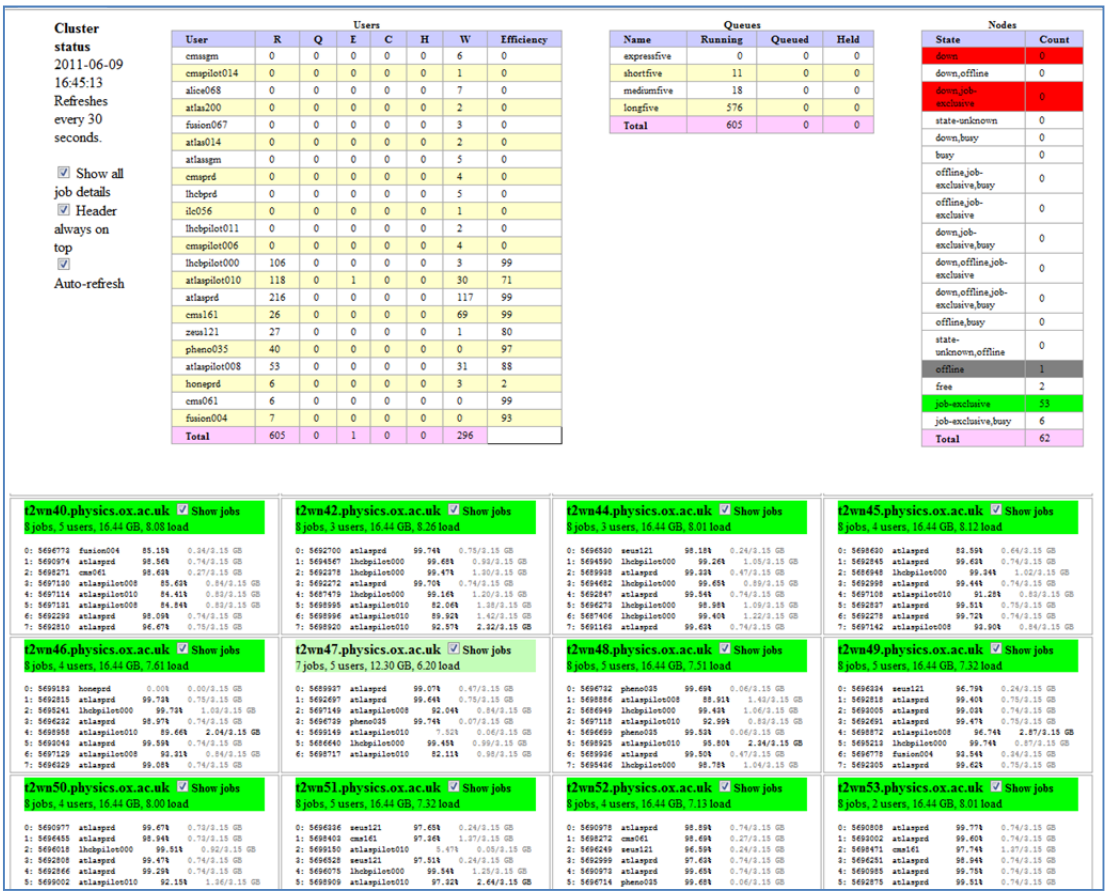

Figure 2 - pbswebmon

For security reasons it is important for sites to keep their systems package rpms up to date. Switching on automatic “yum” updates can help this, but for some critical packages this could break something so it is often not done. Similarly Kernel updates require a reboot and this is often delayed to avoid

downtime. A useful way of seeing how many outstanding updates are required on a particular system is to run Pakiti [5]. This package, originally known as yumit, was developed by GridPP staff at the Rutherford Appleton Lab. On each client it effectively runs a "yum check-update" command to list any packages that have new updates available. The results are stored on a database on the Pakiti server and displayed on a web page (Figure 3). The graphical display allows systems administrators to spot systems that have failed to pick up expected updates, or are behaving differently to that expected.

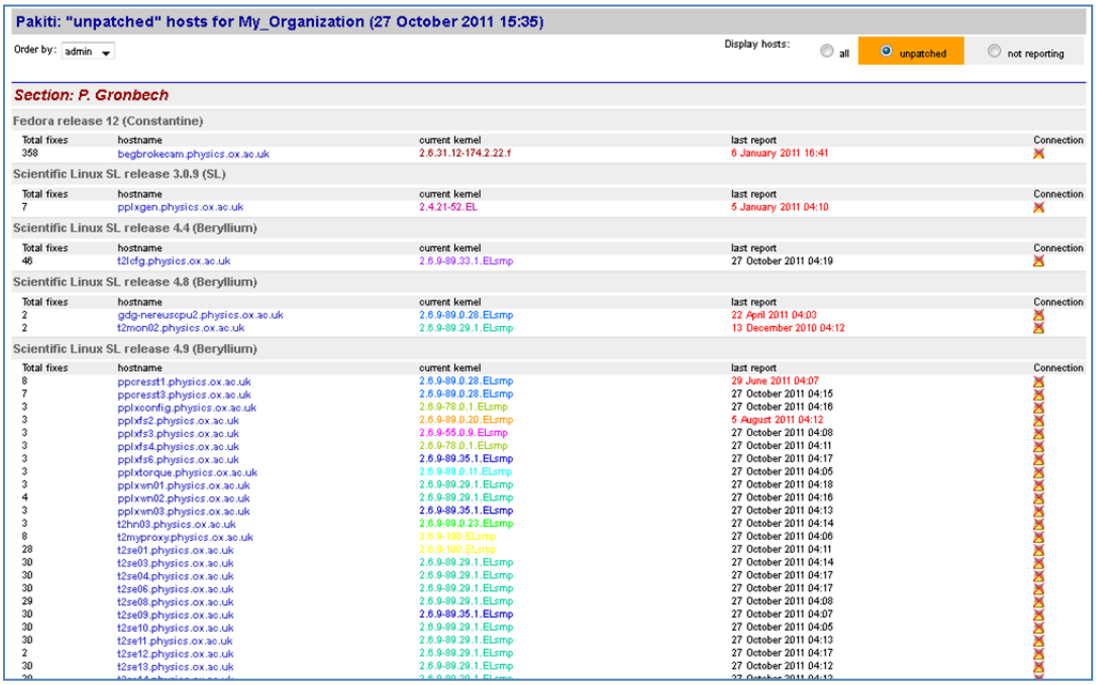

Figure 3 - Pakiti

The above three packages display results via a web interface that must be actively monitored. However, systems administrators want to be notified by an email or SMS message when systems start to fail for either hardware or software reasons. Nagios [6], which provides a very powerful framework, can be used to monitor the status of systems by running tests at specified intervals and then performing actions depending on the results. This could be emailing a warning message or running an event handler that takes remedial action to solve a problem. The advantage of Nagios is that it removes the need to monitor web pages when all is well but can provide notification via email, or SMS (as well as the web) when there is a problem. There are repositories of checking scripts that can be used to build up a customised array of tests for each system. If a specific check is not already available it is relatively simple to create it: For example to check the health of a particular hardware RAID array controller on a storage server or to check a particular piece of Grid middleware is running properly. Each time a new problem occurs the systems administrator can enhance the local Nagios monitoring, to watch for this new failure mode. Simple examples may be local disk partitions getting dangerously full, or a more sophisticated batch systems "black hole spotter". A black hole, in this

context, describes a batch worker (CPU node), which develops a fault such that it accepts new jobs from the queue, then immediately fails them and then takes another job. In this situation all the jobs on the queue can very quickly be drained via this node and no further jobs will get run by the cluster, as that worker node is always apparently free.

Network health, usage and bandwidth should also be monitored at sites. Many sites use Cacti [7] and/ or Network Weathermap [8] to monitor the switches in and around their clusters. The switch connecting a cluster to the outside world is of particular relevance to Grid sites as almost all the work comes from other WLCG [9] sites around the world, and only a minority from the sites own local or Tier 3 facilities. Monitoring the usage of this link helps the site to know if they have enough bandwidth to feed their cluster. If it is continuously saturated for many days then thought should be given to increasing the site connection to JANET [10]. (UK sites are mostly connected at 1Gbps with a few sites at 2-5Gbps and a couple at 10Gbps. Over the next couple of years most sites are expected to move to approximately 2-5Gbps). Cacti can also be setup to monitor other systems that support the Simple Network Management Protocol (SNMP), such as the Power Distribution Units in each rack. (Figure 4)

Figure 4 - Cacti being used to monitor power and network switch traffic

2. National Testing

Although Cacti monitors the traffic actually flowing, it does not provide a measure of the available bandwidth between sites. The Gridmon [11] project was setup to address this problem and a dedicated 'Gridmon' test box is installed at each of the 19 GridPP sites to perform a matrix of iperf, udpmon and other network throughput tests. Each site performs tests against the other sites in its distributed Tier 2 and the Tier 1. The results are stored on a central database with a web frontend (Figure-5). This provides a useful history of the available bandwidth and can clearly show when the

network gets degraded allowing problems to be diagnosed. A reduction in bandwidth can be easily seen which could be caused by a rate-cap that has been applied, or a faulty piece of networking equipment.

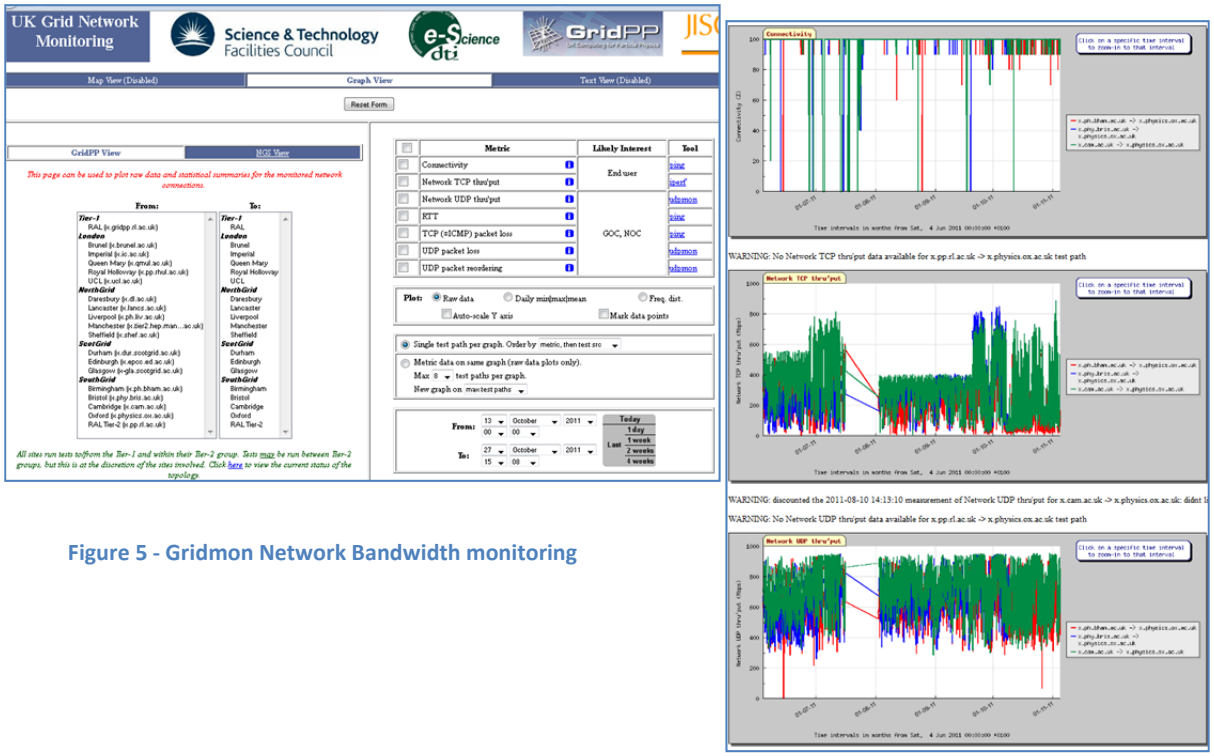

Figure 5 - Gridmon Network Bandwidth monitoring

Other UK wide testing includes a GridPP-developed summation of relevant WLCG tests coupled with dedicated UK tests [12] developed by Prof. S. Lloyd at QMUL (Figure 6). This started out by just summarizing the Global Service Availability Metrics for the UK sites, but has expanded to include experimental tests and customized tests required by the UK.

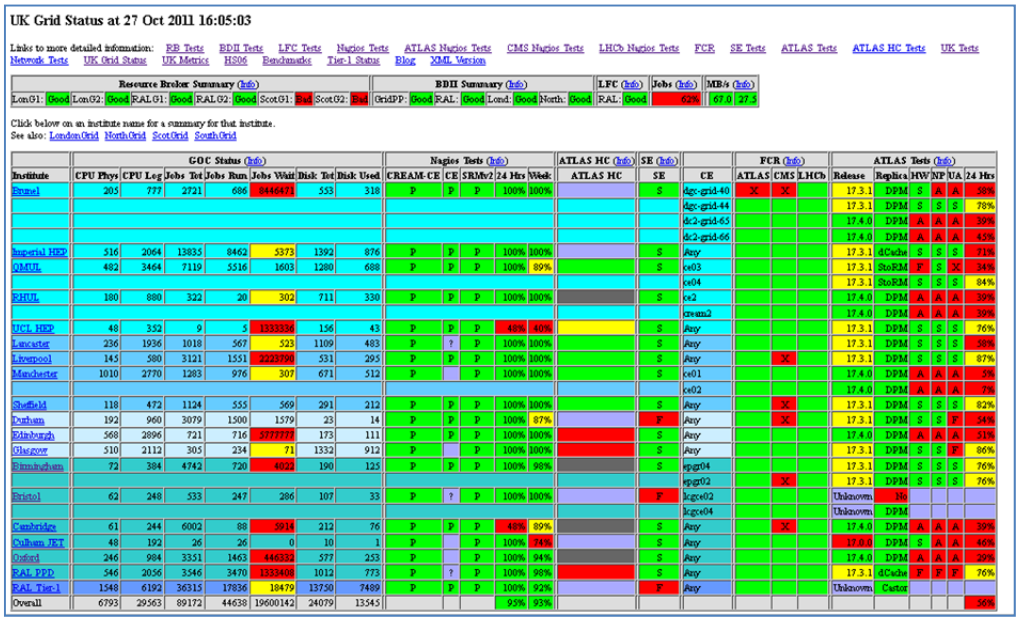

Figure 6 - UK Grid Status

The UK regional Nagios based Service Availability Monitoring (SAM) [13] is run by Oxford University. This service queries a central database (GOCDB) and Grid information services to create a list of sites and systems to be tested. The services offered are tested and the results of the tests are sent via an active MQ message bus to the EGI Regional Operations Dashboard [14] (Figure 8). Each region has an operator on duty that can raise alarm tickets against sites that have failed critical tests. All EGI [15] and WLCG sites have to meet agreed SLA's, to respond to such tickets within a defined time depending on their tier status. The nagios web interface (Figure 7) can be used by Systems Administrators to investigate the error message resulting from the failed test. An alternative interface known as MyEGI is also available on the same server [16] that provides historical plots of availability.

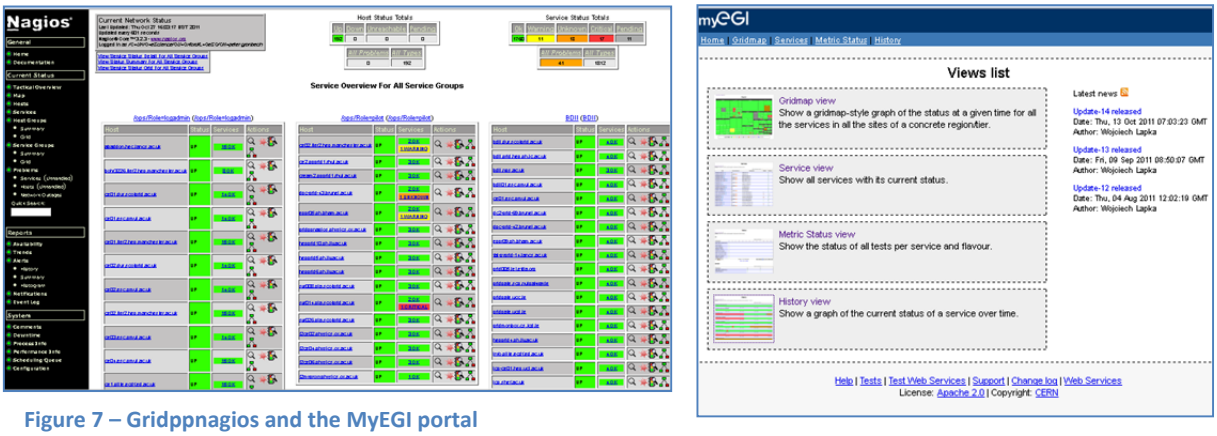

Figure 7 – Gridppnagios and the MyEGI portal

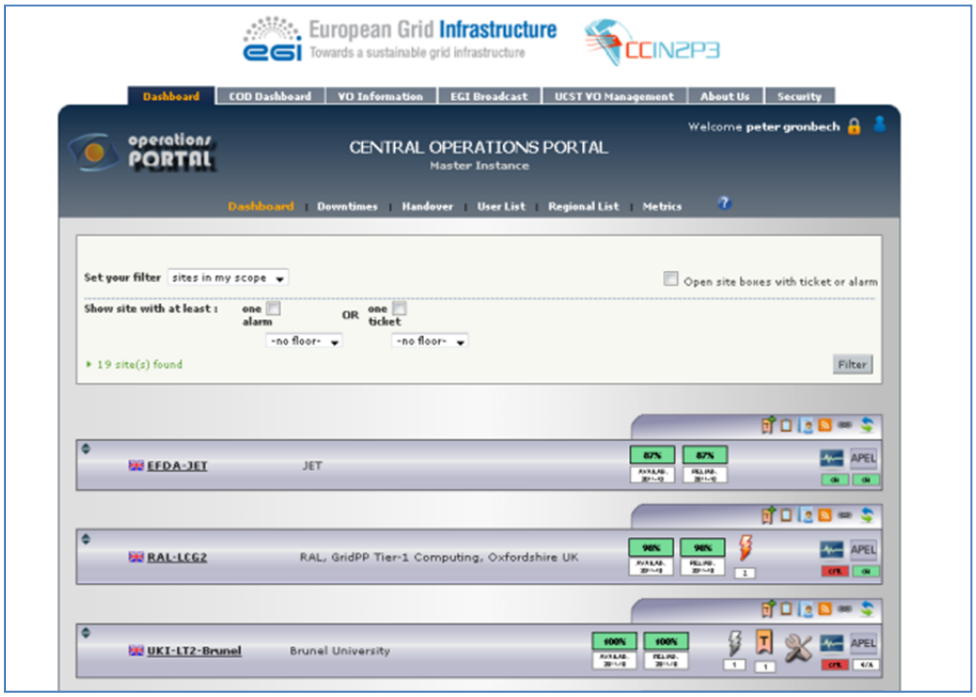

Figure 8 - Regional Operations Portal

3. Global Monitoring

There are many monitoring pages that look at all the European Grid Infrastructure (EGI) grid, or focus particularly on the WLCG grid.

The GOCDB (Grid Operations Central Database) [17] registers details of all sites, their production and certification status and the contacts and services hosted there. Each site publishes the same

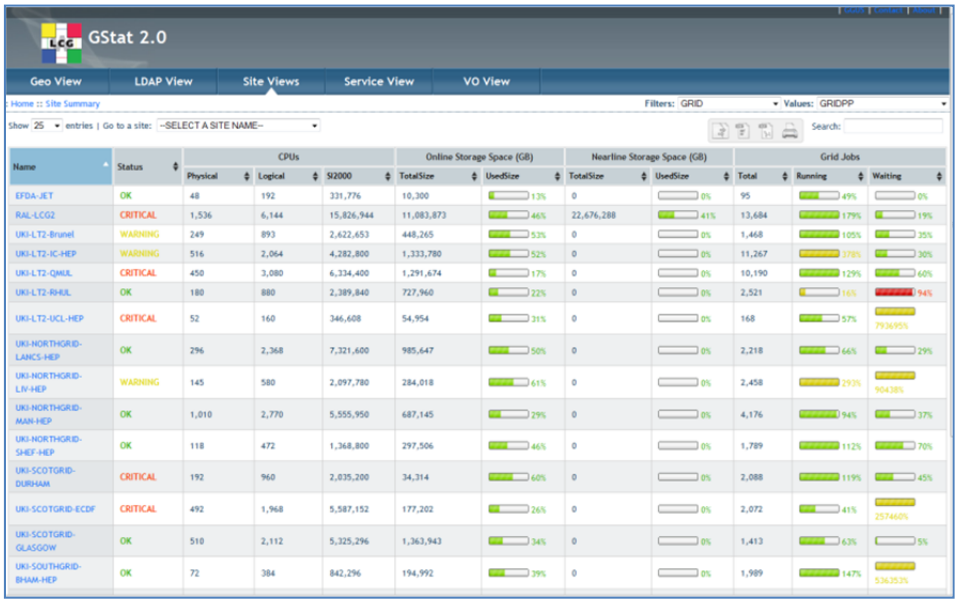

Figure 9 - GSTAT

information plus additional real time attributes via a site-BDII using LDAP. This can be seen via LDAP queries or by using the GSTAT [18] web pages (Figure-9). The GOCDB lists what is expected to be there and the GSTAT shows what is actually there, if there is a difference there may be a problem and this would be shown up by the Regional monitoring. An extension to GSTAT is the WLCG REBUS (Resource, Balance and Usage) [19] systems that makes use of published capacities and compares them against the pledged resources to the LHC [20] experiments, over time. This can highlight sites that have failed to provide what they promised.

The actual work done at a site is measured by the Accounting system APEL [21], each night batch system logs are processed and correlated with the grid middleware logs to publish the number of jobs, and CPU time they took for each Virtual Organisation. This is all collected together in a database and can be accessed by a very flexible web interface[22], which can show details of work carried out at individual sites, distributed Tier2s or by country, and can show the variations across time and users.

The main experiments at CERN submit jobs to the sites, with sophisticated job submission frameworks. The four LHC experiments dashboards now have a common frontend web page [23]. The dashboards track the jobs and report job submission, running, success or failure. Sites that have high failure rates are ticketed by the experiment team shifters, and can be automatically black listed from receiving new jobs.

4. Integrated Dashboards

Systems Administrators are often overwhelmed by the number of different web sites and monitoring systems they should track. Attempts to integrate output from several systems into a site dashboard have been made at the Tier 1 (Figure-10) and some of the larger sites (Figure-11).

The RAL Tier1 Dashboard provides an overview of the Tier1 status. It was originally conceived as a dashboard to provide information externally (to the users). In practice it has proven an invaluable tool for internal monitoring as well. The dashboard pulls together a number of different sources of information. Several key ganglia plots provide trend information for batch and storage. Summaries of both the current SAM tests status and entries in the GOC DB are obtained via php scripts that request the information in XML and parse the result. A list of disk servers in intervention is provided by extract from the local database used to manage these systems. In addition the dashboard provides an area where Tier1 staff can display messages which are also copied to Twitter.

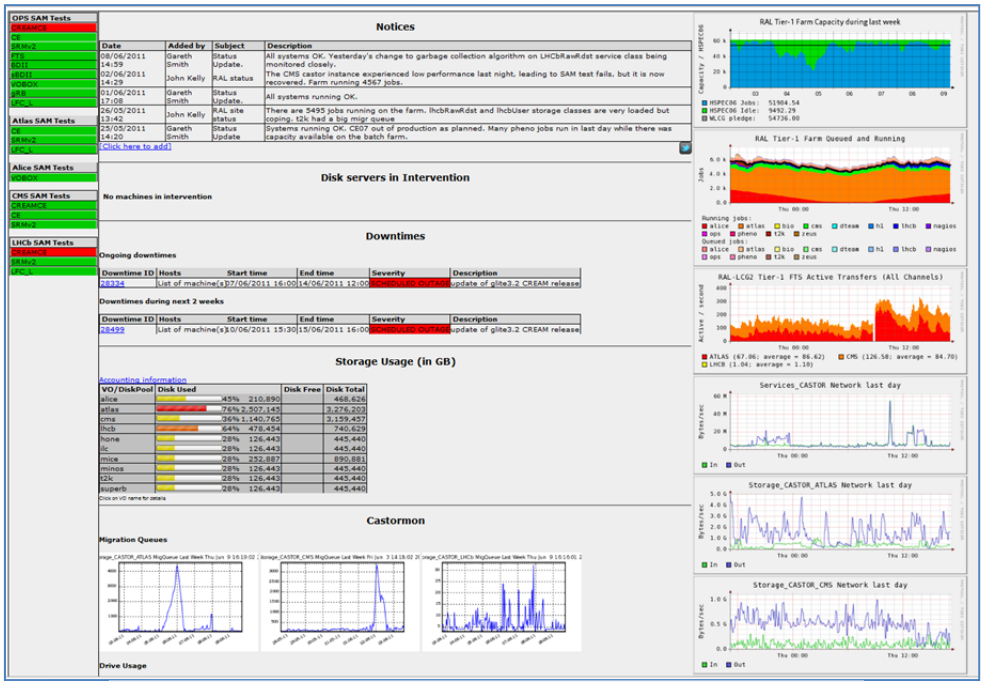

Figure 10 - RAL Tier 1 Dashboard

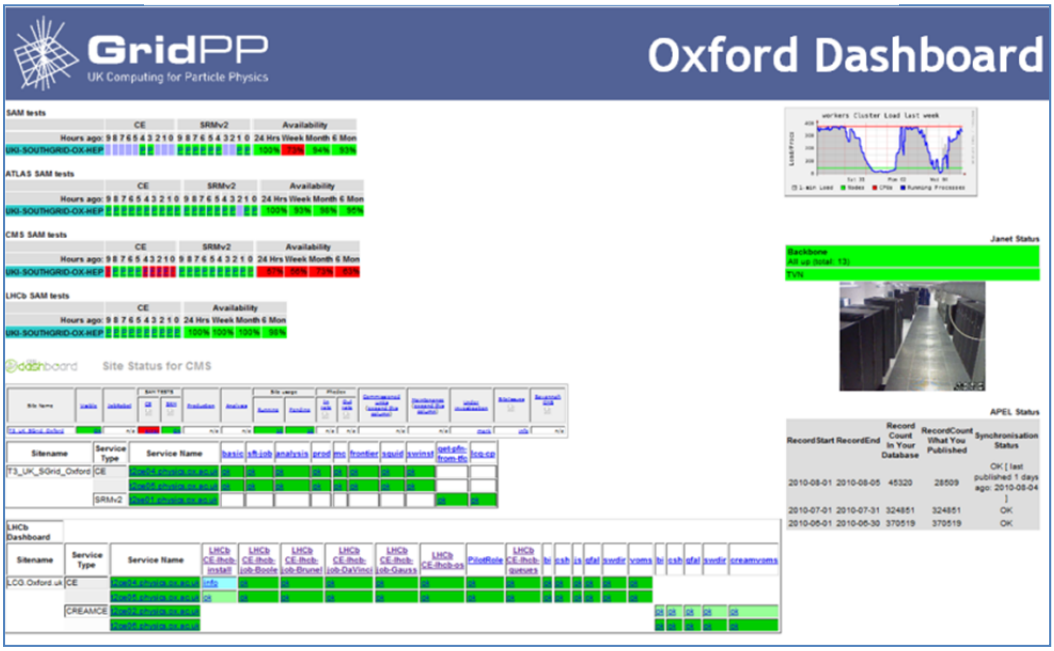

Figure 11 - Oxford Site Dashboard

The Oxford dashboard is based on php code originally developed at Glasgow. It web scrapes the tables from Steve Lloyds SAM summary pages, the ganglia WN load graph, and some experiment dashboards. In addition as the computer room is in a remote location we find having a live camera feed useful. By showing the most relevant information on a single page that can be on display continuously in the support staff room, it can draw attention to faults at a glance.

5. Conclusions

Some of the many monitoring systems that are currently in use have been described. All areas; Fabric Monitoring, Grid Software; Experimental Job Submission and Accounting are continuously evolving. All providing views appropriate to each users particular interest. Sites running production resources have to take all of these into account. Systems administrators will use a combination of increasing the number of tests that local nagios monitoring systems check for and rectify automatically, coupled with use of site dashboards to reduce the time required to maintain an understanding of the health of their sites.

Acknowledgements

This work is carried out on behalf of and with the support of the GridPP project [24].

References

[1] GridPP: the UK grid for particle physics
D. Britton et al, UK e-Science All Hands Conference, Phil. Trans. R. Soc. A June 28, 2009 367:2447-2457; doi:10.1098/rsta.2009.0036
[2] Ganglia http://ganglia.sourceforge.net/
[3] torque http://www.adaptivecomputing.com/products/torque.php
[4] Pbswebmon http://sourceforge.net/apps/trac/pbswebmon/wiki
[5] Pakiti http://pakiti.sourceforge.net/
[6] Nagios http://www.nagios.org/
[7] Cacti http://www.cacti.net/
[8] Network Weathermap http://www.network-weathermap.com/
[9] WLCG http://lcg.web.cern.ch/lcg/
[10] JANET http://www.ja.net/
[11] Gridmon http://gridmon.dl.ac.uk/gridmon/graph.html
[12] UKGrid Monitoring S.Lloyd at QMUL http://pprc.qmul.ac.uk/~lloyd/gridpp/ukgrid.html
[13] Distributed Nagios based SAM testing https://www.gridpp.ac.uk/wiki/UKI_WLCG_Regional_Nagios
[14] Regional Operations Dashboard https://operations-portal.in2p3.fr/dashboard
[15] EGI http://www.egi.eu/
[16] gridppnagios https://gridppnagios.physics.ox.ac.uk/nagios/ and https://gridppnagios.physics.ox.ac.uk/myegi
[17] GOCDB http://goc.egi.eu/
[18] GSTAT http://gstat-prod.cern.ch/gstat/summary/EGEE_ROC/UK/I/
[19] WLCG REBUS http://gstat-wlcg.cern.ch/apps/capacities/sites/
[20] The Large Hadron Collider http://lhc.web.cern.ch/lhc/
[21] APEL Accounting https://wiki.egi.eu/wiki/APEL
[22] Accounting Web Portal http://www4.egee.cesga.es/accounting/egee_view.html
[23] CERN Experimental Dashboards http://dashboard.cern.ch/
[24] GridPP http://www.gridpp.ac.uk/