Grid Computing for UK Particle Physics

Jeremy Coles
GridPP Production Manager, Science & Technology Facilities Council, RAL, Chilton, Oxfordshire OX11 0QX and Senior Research Associate, University of Cambridge, Cavendish Laboratory, JJ Thomson Avenue, Cambridge CB3 0HE, UK
E-mail: j.coles@rl.ac.uk

Abstract. Over the last few years, UK research centres have provided significant computing resources for many high-energy physics collaborations under the guidance of the GridPP project. This paper reviews recent progress in the Grid deployment and operations area including findings from recent experiment and infrastructure service challenges. These results are discussed in the context of how GridPP is dealing with the important areas of networking, data storage and user education and support. Throughout, the paper offers feedback on observed successes and problems with the gLite middleware, grid usage and the experiment specific software which is required to make the Grid usable. The paper ends with a discussion of the decisions being taken to ensure a stable service and project through LHC startup until GridPP funding ends in 2011, including a brief discussion on monitoring and meeting the strict WLCG availability targets.

1. Background

1.1. The GridPP Project

The UK Particle Physics Grid (GridPP) [1] project is a collaboration of particle physicists and computer scientists from the UK and CERN. The collaboration is building a distributed computing Grid across the UK for particle physics. The project has recently been awarded funding (~£25M) to continue for a third stage (GridPP3 until April 2011). GridPP itself consists of a collaboration of 17 partner UK institutions and CERN. The organisation of these institutes into regionally managed Tier-2s is as shown in Figure 1. While the focus of the project is on LHC experiments a number of other projects are enabled on the infrastructure: NA48, Planck, Magic, CDF, H1, Babar, ILC, MINOS, D0, CEDAR, ZEUS, Supernemo and SNO. In addition, the GridPP project map1 shows the collaboration is active in many more areas than just deployment. The project has contributed to Ganga (a job submission framework), Phenogrid2 (which is developing and integrating experiment tools such as HERWIG and JETRAD to interpret events) and the EGEE Real Time Monitor among others.

1 http://www.gridpp.ac.uk/pmb/
2 http://www.phenogrid.dur.ac.uk/
1.2. Resources deployed

Although GridPP resources are for HEP-wide use, as can be seen from the Tier-1 fairshare plots in Figures 2 and 3, the resource allocations are predominately for LHC experiment usage, though the usage is somewhat less structured and use has been made by EGEE VOs like biomedical.

Overall deployment of hardware resources across GridPP Tier-2s can be seen in Figures 4 and 5. They show current (August 2007) deployments and give a comparison with the pledges made by GridPP (on a site-by-site basis) to WLCG. As can be seen the CPU situation is relatively good with most sites meeting or exceeding their pledges. There has not yet been a period when all available job slots have been used and much of 2007 has seen low utilisation (30-60%). Most CPU usage has been by LHCb (33%) with ATLAS (22%) and CMS (15%) the next biggest users.
Figure 3. Actual usage of the Tier-1 CPU from September 2006 to July 2007. Usage has been erratic but the farm has been well utilised.

Disk, while not in a poor state, is certainly not exceeding the pledges. There are a number of factors for this such as a reluctance to buy it until it is really needed (the third bar on the storage graphs shows current usage) due to the falling costs per TB over time but also uncertainty on expectations. Usage of storage elements (SEs) at sites tends to be dominated by one experiment. The SEs are expected to become better used with the upcoming CMS (CSA07) and ATLAS (M5) cosmic runs.

2. Recent advances

2.1. Challenges, tests and reliability

Over the last 12 months the LHC experiments have attempted to use the WLCG infrastructure in ways that simulate what is to be expected under normal LHC running conditions. Unfortunately this has not been straightforward and a new set of tests in 2008 (the Common Computing Readiness Challenges) are being proposed to have all the experiments testing the infrastructure simultaneously. There are a number of reasons this has been difficult so far, but principal among them is the instability of disk resources.

Figure 4. Pledged versus actually deployed CPU resource across GridPP Tier-2 sites. For each site, the blue stack represents the pledged CPU in KSI2K and the red stack the amount available at the time of writing.
Figure 5. Pledged versus actually deployed storage resources (TB). For each site the first stack (blue) shows the amount pledged. The second (red) the amount deployed and the third stack (cream) the amount used at the time of writing.

In the UK case, the RAL Tier-1 has had a number of obstacles to overcome in the last year. Firstly a large amount of purchased disk failed to meet its acceptance criteria – disks in raid 6 experienced multiple dropouts. It took many months to trace the problem to a firmware (missing return statement) issue whereby after a drive head stayed in one place too long it repositioned to allow lubricant to flow back and did not return afterwards. Secondly a decision was taken during 2006 to move away from the RAL ATLAS data store interface to tape (a site bespoke solution) and adopt CASTOR. CASTOR was developed at CERN and initially had many dependencies on the CERN framework – for example use of the Lemon monitoring system. Like any new software package it was only when it became more widely used (in this case outside of CERN) that significant problems showed up (for example in network tuning, reserve space pile up, pool setup and tape access speeds) and these have kept the development and deployment teams very busy for the last year. At the time of writing, single CASTOR 2.1.3 instances for each experiment appears to have allowed a period of relative stability for ATLAS and CMS transfers.

Throughout the last 12 months a significant amount of production simulation has taken place (>100M KSI2K hrs). From this it is important to note that across WLCG over 40% of CPU time has come from Tier-2 sites; in the UK this is close to 85%. The continued testing has also seen the network links between sites exercised. CMS in CSA06 (Nov 2006) saw up to about 550 MB/s, though this was not sustained for long periods. PhEDEx transfers have kept rates above this level since about April 2007 which shows that steady progress in addressing stability issues is being made. As mentioned earlier, concurrent running has not been possible so far but is being planned for early 2008. Throughout 2006 and into 2007 GridPP deployment team members have pursued the systematic testing of site-to-site transfers to test both the site wide area networking capabilities and the Storage Element (SE) set up. This work also enabled the first heavy loads to be placed on the Tier-1 networking and revealed unexpected throughput limits in the site firewall and the subsequent implementation of caps until the underlying issue is resolved by the firewall vendor. Throughout, the constant testing of networking links by a dedicated monitoring framework (GridMon\(^3\)) has proven invaluable in determining the time points at which problems later impacting transfers began.

\(^3\) [http://gridmon3.dl.ac.uk/gridmon/graph.html](http://gridmon3.dl.ac.uk/gridmon/graph.html)
The WLCG management board has for some time now monitored the site availability figures for the Tier-1 sites using the ops Virtual Organisation (VO) Site Availability Monitoring (SAM) test results⁴. These have shown that availability at over 95% for sustained periods is possible. In GridPP the Tier-2 sites have also been monitored and each month their performance is compared to the average result for all the GridPP sites⁵. This comparative approach (see Figure 6) is allowing GridPP to ramp up performance targets in line with what is seen to be possible.

2.2. Feedback from sites

In the second quarter of 2007, the GridPP Project Management Board (PMB) set about reviewing the readiness of GridPP sites for LHC start up. Each site was sent a comprehensive questionnaire covering 6 core areas: Management and context; Hardware and network; Middleware and operating systems; Service levels and staff; Finance and funding; and Users⁶. Each Tier-2 was then visited by a review team to discuss the responses and issues that arose. As well as revealing some of the findings mentioned earlier in this paper, the reviews concluded that the Tier-2 sites were working well together with good technical cooperation taking place. As a result the approaches to installation are quite uniform (use of kickstart and local scripts, YAIM and tarballs for the worker nodes – some sites were moving on to more advanced configuration with Cfengine). There was widespread concern about the lack of full testing of some components of middleware prior to their use in production. The rest of this section highlights some of the technical concerns and areas in which future progress would help sites.

One long standing that came up and has been partially resolved during 2007 is the porting of the gLite middleware to run on Scientific Linux 4. Workarounds had been found to run the SL3 compiled WN versions on SL4. Even with the emergence of a production version there have been minor issues in identifying OS dependencies outside of the middleware but needed by the experiment software (the EGEE packaging group (SA3) had steadily removed support for external dependencies – the

---

⁴ The site results can be accessed via https://lcg-sam.cern.ch:8443/sam/sam.py
⁵ For the table of results see http://www.gridpp.ac.uk/wiki/SAM_availability:_Monthly_summary_table
⁶ Some of the material can be accessed here: http://www.gridpp.ac.uk/tier2/
recommendation being to use repositories like DAG to meet them). The missing RPMs were identified
(and inserted into the VO identity card on the operations portal) but with many only being found once
the experiments ran on updated production sites, the need for a re-evaluation of the role of the Pre-
Production Service of EGEE in experiment testing of new components was noticed. As for the move
to native SL4 built worker nodes, some GridPP sites that had new hardware requiring SL4 moved
relatively quickly. Other sites are currently in the process of upgrading or will wait until they
commissioned new clusters. Experiences here need to be remembered for future OS moves.

One of the biggest concerns of UK system administrators is the lack of complete error messages in the
middleware and related components, which makes problem investigation and diagnosis difficult. SA3
have indicated that a checklist could be developed and new middleware will not be deployed if error
messaging is not satisfactory according to these prescribed checks. This is badly needed as often error
messages are unhelpful and impenetrable. A clearer log system and standard logging format is also
needed.

Another area that makes deployment difficult is the current lack of (and implementations to deal with)
VO quotas on shared disk pools – especially for DPM and dCache. There have been numerous cases
of one VO filling a shared pool and thus making the whole site unusable by any VO – in fact the ops
tests can fail as a result and the sites becomes blacklisted. VO changes at the site level are also in need
of further development to make it easier to add/remove/suspend a VO user. Actually controls of this
sort are needed in other areas too such as making it more straightforward to add, remove or drain a
queue; to close a site on the BDII and to close a storage area. To cover this it would be very welcome
if a common control interface to all grid services could be provided - these to provide (a minimum of)
remote stop, start and status.

The passing of parameters to the LRMS is needed to enable better use of resources – being able to
pass job information to the CE from the RB/WMS, such as required scratch space and memory
requirements would stop jobs failing when “known” limitations are hit. System administrators still call
for better failover in clients, tools and services – data management tools for example should be able to
use a backup BDII. It should be noted though that the provision of diagnostic scripts for each service
to allow checking of configuration and functionality would reduce the need for failover!
Documentation, such as release notes, should contain all information needed on required configuration
changes, detailed deployment steps, bug fixes and new functionality. They might also provide clear
definitions of maintenance and garbage collection/requirements for the main middleware nodes.
Project documentation has improved but still has room for further improvement.

In addition to the clear needs just mentioned there are other areas that would help but have so far been
worked around (often at the expense of many sysadmins redoing the same investigations). A list of
RPMs per service component and defined internal/external dependencies would help in early
configuration checking and diagnosis. Reducing the OS dependencies of the middleware would
increase the available resources and prevent the conflicts seen in several shared resource centres;
publishing of network cost information to prevent job submissions to sites where SE access is already
bottlenecked would reduce timeouts. Finally tools to deal with bulk SE operations would save a lot of
time – imagine being able to painlessly move data between SEs and having the catalogues updated
correctly – this would be of direct benefit now at RAL where the experiments are about to begin their
movement of data out of dCache and into CASTOR.

The reviews showed that sites generally understood the requirements which came with being part of
the WLCG project. On the whole management was found to be working and finance and funding were
well understood. Following the reviews each site (and Tier-2) received, where applicable, a feedback
report with recommendations for improvement. One area that remains difficult to address in some
cases is the lack of dedicated manpower to run the grid clusters. Other findings such as the need for better monitoring are being actively tackled.

2.3. Bridging the user-infrastructure divide

Since part of the purpose of this paper is to summarise GridPP progress and issues, it is useful to next look at current issues that need to be addressed with regards to users and VOs. Here there are concerns from the very start of the process even if this is no longer significant for the LHC VOs. Enabling and configuring new VOs is tricky for people new to VOMS and hard on the sites. Often the site required information is missing from the VO card and additionally site administrators have no direct way of checking the configuration once the work seems to be finished. This problem still arises in the context of LHC VO required changes since there is rarely any diagnosis software. This has been seen in the moves towards pool accounts for sgms. As an aside here, it is worth indicating that a mismatch between grid X509 based systems and mapping people down to the UID/GID pair in Linux has no simple translation requiring lots of fiddling with pool accounts, LCMaps and grid-mapfiles. This will get worse with VOMS groups and roles (the experiments have indicated a desire for many of each) mapping between security realms. Already there are problems in the non-unique ways in which information is extracted from the FQAN.

One of the biggest current problems with the grid from a user perspective is the instability of the SEs. There are several underlying reasons for failures including actual disk failure. In these instances it has proved difficult to inform all users with data on the affected disks. Actually the same communication problem is seen when the SEs get full and need clearing. GridPP’s storage group has members involved with SRM v2.2 testing and deployment planning. While there is still a gap in understanding how the experiments will use the new SRM v2.2 functionality, there is now a concerted effort to ensure the improved stability of v2.2 implementations gets translated into good user experience.

Use of CPU has generally become a success story, though the policy of resubmission that helps improve overall success rates for the experiments does not help sites to diagnose problems! Early in 2007 several periods of low efficiency were seen across the VOs and there is still variation within VOs. This is often blamed on I/O bottlenecks or a stage request to load new data getting stuck. GridPP has seen inefficient jobs as a problem and for the rest of 2007 is testing a policy of killing jobs that consume less than 10 minutes of CPU time in a 6 hour period (2.7% efficiency). Ironically one reason that jobs have been seen to fail is that some write large log files to /tmp and cripple the WNs (which then requires scripts to be written to clear up and recover the situation). There are site problems that also contribute to job failures including the “black hole” effect seen after worker node updates do not complete correctly and lead to jobs being allocated to nodes which are no longer running correctly. This particular problem will more easily be spotted with improvements in site monitoring.

Testing of new middleware by the experiments has already been mentioned in relation to the move to SL4. The move to 64-bit looks to be even more of a concern since memory requirements with present builds look too large. At a more basic level, improved experiment integration with pre-release and pre-production service testing would help to spot potential problems that should be avoided. Some recent examples include the loss of EDG_WL_SCRATCH functionality which was not spotted until it impacted production and the attempt to implement a mechanism to handle job priorities.

A user level issue which should be mentioned is that when tickets are created against a site there is sometimes a lack of sufficient information which leads to problems assigning such tickets and inevitably increased time in resolving them. Additionally, when a site has investigated and written a response some users are slow at acknowledging the solution or retesting (user feedback is needed to confirm jobs are running correctly) and in the interim the site receives regular reminders that they have
an open ticket. The opposite can also happen where too much contact is made with the site either through multiple tickets/queries or direct contact about the same problem. This can lead to confusion as the administrator thinks they have resolved a ticket but another is opened! Confusion can also come through rapid changes in direction requested by the experiment – for example ATLAS’s ACL change request (that was also accompanied by buggy scripts) and LHCb’s use of sgm accounts.

2.4. Working together

The last year of GridPP operations has seen a greater focus on working with the VO's/experiments to improve site availability as seen from a user perspective. One method introduced to do encourage this is the deployment of VO specific SAM tests which allow sites to monitor whether they are meeting additional requirements of the experiments. Since a frequent problem is the experiment installations at sites not being “clean” or up-to-date, there is some dependence on the experiment to solve certain problems. Sites can struggle to get the experiment support they need. To improve the situation for GridPP sites, weekly regional deployment team meetings are now attended regularly by members of the experiment teams. As well as keeping each other updated this offers an opportunity to follow up on things that may otherwise have been forgotten.

In addition to (and starting earlier than) the VO specific SAM tests, GridPP was running a set of ATLAS tests which worked through from basic job submission and testing the installed ATLAS software, to building a new Athena package and analyzing a file of AOD data. This was the first opportunity sites had to “directly” discover experiment perceived problems (the logs of each job being made available on a web page). As Figure 7 shows, there was a positive feedback loop which helped greatly improve the efficiency rates. A stable state of around 80% has been reached. This is a good outcome. Sites would not expect to always pass these tests because for example, if a site is very full and the test job times out while still queueing the site fails the test. This is the real user experience where job success and speed to get executed depends on the experiment queues and fairshares at each site. Additionally if a site is in maintenance the site will fail, and this is again a true measure of the site’s availability for a user.

![Figure 7](image)

**Figure 7.** The average GridPP site availability over time as indicated by the GridPP run ATLAS tests. Good progress was made once sites had this user view of their availability. Drops to zero are the result of the test mechanism failing and not the sites all having problems.

3. The Future

3.1. Challenges on the road ahead

There are clear challenges in the remaining period before LHC data taking starts. Not least for WLCG as a whole is the challenge of ramping up Tier-1 resources by a factor of 6 and Tier-2’s by a factor of
GridPP sites are already providing a sizeable fraction of the pledged resources (though it is evident that they are not able to contribute the pledged KSI2K hrs without a substantial increase in experiment activity) though there are still some sizeable shared central university computing resources to be grid enabled. For RAL Tier-1 a new machine room is currently being built (800m² for the machine room and offices with space for 200 racks and 2.3MW for power/cooling) and this is due for completion in the autumn of 2008. Moving staff and gradually transitioning hardware to this new building will be an additional burden on top of installing and maintaining new hardware.

SRM2.2 deployment is also going to provide a major challenge. This is not just in the deployment at sites but in the improved determination of experiment requirements in respect of it. Alongside this there will be new challenges coming from the hitherto unseen stressing of the infrastructure as users start to run analysis work over the grid. While GridPP members continue to test network and SE capabilities it is only when many (new) users get involved that serious debugging can start. For example, WN-SE links being unable to cope with certain access/usage patterns which can not be simulated. Related to this is a whole area that needs to be explored governing disaster (recovery) planning. While GridPP has started to look at scenarios this is undeveloped work at the present time and WLCG as a whole has yet to engage with it. Scenarios range from infrastructure service and software failures to security or even procurement problems. On the experiment side too planning is needed. What will happen if things do not scale as they should (from cataloging to transfer systems) or there is a significant loss of fabric or the computing models just do not work as envisaged?

Middleware stability (generally) has improved considerably over the last few years. There are still many concerns, including the recent abandonment of the gLite-CE in favour of the relatively unknown (to deployment testing) CREAM-CE. At some point a transition needs to be made to this new CE and also away from LCG-RBs to the now somewhat improved gLite-WMS. The timescales for moving to 64-bit computing are also unknown, and for the experiments at least in trying to contain their memory requirements, something for after the startup of the LHC. Against the quite significant changes yet to happen on the middleware side, there is also a growing problem of tracking and keeping up to date with a world of fast moving patches. This might improve with better monitoring and monitoring is sure to be a focus area over the coming 12 months. In GridPP there has been a steady increase in areas such as sharing Nagios scripts to detect problems early, this becoming ever more important as site resources grow beyond what the site administrators are used to running. The collaboration has also sought to improve the high-level service monitoring due to outages affecting many sites or becoming evident in regional testing. Hopefully the evolution of grid wide monitoring will not lead to a similar situation as happened in the last year whereby a decision to wrap all jobs with monitoring had a very large impact on wallclocktime of (especially) short jobs leading many to times out (a bug led to orphaned, in some cases thousands of gatekeeper, processes being left open).

For some months now, the topic of pilot jobs and use of job ID changing code (glexec being the main contender) on worker nodes has been hotly debated. GridPP has, through its deployment board, set down a series of pre-conditions prior to considering the use of such code across its constituent sites. Relevant policies need to be in place; the code needs to be audited; and above all a decision needs to be made about how to continue using shared resource centres where there is no direct access to the worker nodes. Virtualisation may be an option in the future but the issue needs to be resolved in the short-term given the increasing experiment dependence on this type of job.

3.2. The GridPP project

As mentioned at the start of this paper, GridPP has now received confirmation of funding until 2011, though not at the level the project requested. The award includes money for manpower and hardware.
The current 9.5 FTEs for Tier-2 hardware support increases to 13 in GridPP3. The Tier-1 sees an increase in funding for both hardware and manpower in line with the increased fabric pledged to WLCG.

Allocations of GridPP3 funding for Tier-2 hardware will be made to sites on the basis of their previous performance and in particular that during Q3 2007. The algorithm will be based on some combination of SAM test efficiency, CPU delivered and disk available.

4. Summary and outlook

GridPP serves many HEP communities but its primary focus is LHC computing. Tier-2s are providing over 8500 KSI2K of CPU and 700 TB of disk. These figures represent over 130% of the CPU and 84% of the disk agreed in a GridPP Memorandum of Understanding (MoU) signed in 2004. New MoUs are being drawn up for GridPP3.

While the sites are providing sufficient resources to WLCG they are not being sufficiently utilised. Steady progress is being seen in improving the availability of sites and services but some types of problem will only be uncovered as usage grows. Some sites struggle with lack of dedicated manpower and improved cross-site support is being explored within the Tier-2s to address this problem and to meet the defined response times in the WLCG MoU.

The GridPP Tier-1 has over the last year had to overcome several storage related problems. With CASTOR stability finally evident there begins the tricky process of migrating data from RAL dCache to RAL CASTOR.

The GridPP Site Readiness Reviews provided valuable interaction between sites and GridPP management, technical coordinators and experiment representatives. While some sites received more recommendations than others, overall the impression was very positive. There remain concerns with some aspects of the middleware but the underlying problems are manageable. Improved monitoring (and availability) remains a key target across all sites. A GridPP Tier-1 site review may take place soon to complement the Tier-2 reviews.

[1] Development of the UK computing Grid for particle physics The GridPP Collaboration, P.J.W. Faulkner et al. *Journal of Physics G: Nuclear and Particle Physics* 32 (2006) N1-N20

[2] Policy on Sites Stopping Stalled Jobs. GridPP internal document 113: [http://www.gridpp.ac.uk/pmb/docs/index.htm](http://www.gridpp.ac.uk/pmb/docs/index.htm)