Preparation for WLCG Production from a Tier-1 Viewpoint

M J Bly and R A Sansum
STFC Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, DIDCOT, Oxfordshire, OX11 0QX, United Kingdom
Email: m.j.bly@rl.ac.uk, r.a.sansum@rl.ac.uk

Abstract. The GRIDPP Tier-1 Centre at RAL is one of 10 Tier-1 centres worldwide preparing for the start of LHC data taking in late 2007. The RAL Tier-1 is expected to provide a reliable grid-based computing service running thousands of simultaneous batch jobs with access to a multi-petabyte CASTOR-managed disk storage pool and tape silo, and will support the ATLAS, CMS and LHCb experiments as well as many other experiments already taking or analysing data. The RAL Tier-1 is already well advanced towards readiness for LHC data-taking. We describe some of the reliability and performance issues encountered with various generations of storage hardware in use at RAL and how the problems were addressed. We also describe the networking challenges for shipping large volumes of data into and out of the Tier-1 storage systems, and system to system within the Tier-1, and the changes made to accommodate the expected data volumes. We describe the scalability and reliability issues encountered with the grid-services and the various strategies used to minimise the impact of problems, including multiplying the number of service hosts, splitting services across a number of hosts, and upgrading services to more resilient hardware.

1. The RAL Tier-1

The GRIDPP Tier-1 at RAL is the UK Tier-1 for the World LHC Computing Grid, supporting ATLAS, CMS, and LHCb, as well as existing HEP experiments. The facility currently has around 800TB of usable disk storage space, approximately 1200 CPU cores running in the batch system, and a tape silo with a 10,000-tape capacity. The latter has approximately 18 T10000 tape drives and 10 9940B tape drives. There are approximately 100 systems running grid services of various sorts, as well as the usual monitoring nodes and system services such as mail, authentication etc. The computing hardware is generally rack-mounted to save space, and has network-accessible power controllers.

The hardware has been procured in stages with major purchases each year adding to the batch and disk storage capacity, and to the tape drive and tape inventory for the tape silo. The planned lifetime of the CPU and disk hardware is 5 years, of which 3 years is under full warranty and maintenance, the fourth year carries now maintenance with an expected mortality of 10% and the fifth year no maintenance and progressive decommissioning around the end of the fifth year. Processing and disk storage capacity is dominated by the most recent procurements which totalled 516TB useable disk space and 550,000 SPECint_base2000 [1] processing power. A typical batch worker is a 1U rack-mounted unit and has two CPU chips with single or dual CPU cores, 1GB RAM per CPU core, a system disk with 50GB capacity plus 50GB per CPU core, and dual 1Gb/s Ethernet ports. The storage servers provide between 1.6TB and 9.5TB useable space per system.
2. Hardware reliability
The Tier-1 operates the majority of its services on commodity hardware. The services nodes have
tended to be re-tasked batch workers which do not have any of the usual additions to harden the
services against failure: dual hot-swap capable disks and power supplies. The storage hardware has
been mainly commodity solutions with some additional hardware for reliability: hot-swap fans, and
power supplies, and in most cases hardware or software raid systems disks. The data storage arrays
themselves are SCSI-attached IDE and SATA disk arrays with RAID5 and more recently direct
attached SATA arrays with RAID5 or RAID6 capability.

2.1. Storage reliability
Annual disk failure rates have averaged between 2% and 3%, depending on the generation of storage
hardware, with no generation significantly different from another. Commissioning issues aside, all the
technologies have proved reasonably reliable but have highlighted two big issues.

2.1.1. Interconnects. With the SCSI-attached generations, we have observed that thermal cycling of
the hardware due both to daily temperature changes and to air conditioning events can cause problems
with the interconnect cables. It appears that regular thermal cycling in a machine room where the
ceiling is directly under the flat roof of the building is sufficient to work cables loose enough to cause
signal issues on the interconnect, and cause the SCSI transport layer in the software to observe various
errors. This usually causes the array to be dropped offline, but has on several occasions caused data
corruption when the system persisted in trying to write data to the array. Most cases are solved by
completely detaching the cables, cleaning the contacts reattaching them. Where the array has been
dropped offline, it has usually been possible to recover the data with a simple file system check (fsck),
but in some cases, files have been discovered to be corrupted. In one or two cases, significant data
corruption has occurred, necessitating recovery from tape where possible.

2.1.2. Multi-disk failures. The second issue is the recovery or rebuild time for the hardware RAID
arrays. For RAID5 arrays, a single disk failure will trigger a rebuild onto a hot spare disk. While this is
happening the array is in a degraded state and is vulnerable to any problem with the remaining disks or
the hot spare, but the array will continue to operate and read or write data. Rebuild times vary
depending on the system tuning and I/O load – when the system is under high I/O load the controller
cannot progress the array rebuild quite as fast as it can when under no load. This increases the
vulnerability window. The size of the array also effects rebuild times – bigger arrays with more and
larger disks take longer to rebuild, increasing the vulnerable period, and the probability of a second
disk failure. If a second disk does experience an error before the array is rebuilt the controller may be
able to recover depending on the issue, but it may not, causing potential data loss. RAID6 helps with
this issue in that there are two sets of redundancy information in the array. This means that if one disk
fails, although a rebuild may not immediately start if there is no hot spare, the array is still operating
with the a complete parity set intact (essentially in RAID 5 mode) and can suffer a second disk failure
without loss of data. The Tier-1 now procures storage systems requiring RAID6 or multiple
redundancy capability to reduce the likelihood of a dual disk failure leading to data loss. The added
cost is not significant and the saving in staff effort recovering damaged data systems is significant.

2.2. Service hardware
In general, the services hardware has been reliable, with the expected complement of disk and RAM
failures depending on age of the hardware. However running grid services which require maximum
uptime and the ability to survive system issues on re-tasked batch workers has met with mixed
success. In most cases the hardware copes with the loads, but cannot cope with disk, ram and PSU
failures.
Various strategies can be employed to guard against service failures, ranging from backups, fast reinstallations or re-instancing, multiple disk software or hardware system disks, and redundant power supplies right up to expensive fully redundant hardware and more recently virtualization.

To give some added robustness to the services, we have elected fit selected sets of services nodes with additional disks and employ software raid configurations and backups strategies where needed to enable systems to survive and maintain availability. This has worked very well with several hosts maintaining service availability after a disk failure during silent hours, due to their RAID1 system disk pairs. Service interventions for these systems can then be planned and announced in advance. An added advantage in some cases in that of increased I/O speed, particularly for reading data, increasing the performance of the service.

More recently services such as the 3D database project have made use of systems with not only built-in hardware redundancy but also Oracle Real Applications Cluster (RAC) technology, enabling both the hardware and the database itself to be more robust against failures.

In future, we recognise that the reliability and availability requirements of the hardware for the Grid services will require not just hardware with redundant features but also much more powerful hardware to run the increasingly demanding Grid software.

3. Grid Services Performance
The RAL Tier-1 runs a wide range of Grid services ranging from Compute Elements (CEs) and Resources Brokers (RBS) to Proxy servers, Local File Catalogues, a File Transfer Services (FTS) and the R-GMA central service host. In the early days of operation, a single reasonably powered host system was sufficient to run each service. In some cases, a single host was able to share two or more services, for example, the local BDII service at RAL was co-hosted with the CE. As the processing and data transfer loads on the WLCG service have risen, and the complexity of the services has increased, single instances of many types of services node have proved inadequate to provide a reliable service, and shared hosting of services on a single host has become very difficult.

3.1. The RAL Compute Element
As mentioned previously the CE host at RAL originally hosted the CE and the local BDII. As the number of grid jobs arriving at the Tier-1 increased, the performance monitoring of the host began to show that the host itself (a dual-chip Xeon system) was overloaded and unable to keep up with the service requirements, and would eventually grind itself to a standstill, with little or no response to services requests and no response on its console.

The first move was to implement the local BDII service on a new host to remove the load it represented, and to safeguard that part of the local grid service against issues on the CE host. This had a marginal effect at best and it soon became clear that the CE was still underpowered. The main effect of the move was to increase the site reliability in response to information system requests to the local BDII.

Some of the load on the CE is generated by considerable local I/O to the disk, and the local disk space available was proving insufficient to meet the needs of the Grid job loading. The host type concerned had in the past exhibited limited performance to the (IDE) disk under load conditions so the CE was transplanted to a slightly more powerful machine (another dual-chip Xeon system) with a different main board chipset, more memory and a bigger disk. This made a small difference at first, by reducing swapping and speeding up I/O, but the system ultimately proved incapable of handling the load so a second move was made to a dual-chip system with dual-core (Opteron) CPUs and a faster (SATA) disk. This has proved adequate to the task so far but we now recognize that multiple CEs targeted at specific experiment communities is the way to reduce the load on individual systems and increase the overall facility reliability.
3.2. The UK BDII
The Tier-1 at RAL has run the UK-wide BDII service since the first releases of the WLCG software stack. The UK BDII provides the information service with data about which resources in the UK are available and was used by the WLCG Site Availability Monitor (SAM) tests and their successors to direct the availability testing jobs. Until mid-2006, the single service host proved adequate for the task but the increasing number of queries from more systems in the UK using it as their information service started to have a detrimental effect not only the host itself but also the availability of the whole WLCG service in the UK. If the UK BDII fails to respond, then the availability monitors do not know which resources to test for availability.

A two pronged solution was deployed to increase both the reliability and the response time of the BDII. Firstly, two new identical BDII hosts running on slightly faster hardware were prepared. Then the service was transferred to one of the new hosts using the existing host name as an alias. The old host was decommissioned after queries stopped being made against it, and the second new host was added to the first as a DNS round-robin pair with a short time-to-live set in the DNS. Thus queries would go to which ever of the two hosts was named in the DNS response.

The load of queries balanced out roughly evenly using this strategy but we noted that although the level of dropped or failed queries was now very low and the load on each host was not significant, a small number of queries were still being dropped. We therefore added a third (identical) host to the DNS round-robin set for the BDII, which reduced the query failure rate to zero.

The advantage gained from having three hosts providing the BDII service is that if one host fails, only one third of the queries will fail. While this is not ideal, it does allow some time to recover the failed host or provide a new host to replace the failed one – automatic deployment of BDII host configurations is quite quick. In the future, developments at RAL will allow the Tier-1 team to modify the DNS directly if such failures occur to take the failed host out of the DNS.

3.3. Resource Brokers.
The Resource Broker service was originally comprised of a single host, migrating from host to host the initial versions of the LCG software stack were released. It was evident over time that the RB was one of the more resource-hungry services, so when the LCG-2 stack was released, the RB service was installed on a dual-chip Xeon system, which performed well during initial production service running. However, the increased load of grid-jobs began to cause resource issues in the host, manifesting in very high load averages as the system tried to keep up with the transit of jobs. In addition, limits in the MySQL database structures meant that the RB databases needs to be regularly purged of old job records causing service down-time. VO groups were also raising reliability and availability issues.

To attempt to alleviate these problems, a second and later a third RB were added to the service. These were targeted at specific VOs and enable maintenance on the RB databases to take place without stopping all transactions.

3.4. Batch Scheduler
The batch system and local job scheduler used successfully at the Tier-1 before the grid was OpenPBS, firstly with its own FIFO scheduler, then later with tweaks for job priority reordering, and most recently with the Maui scheduler. The batch system software has now been migrated to Torque but still uses the Maui scheduler. Various hard limits within the Maui code have been patched to enable the larger number of jobs, queues and classes required for the Tier-1 setup. The batch and local job scheduling systems run together on the same host which has software RAID1 system disks to provide added protection against disk failure.

3.5. CASTOR
With the purchase of a new tape silo (an STK SL85000), a decision was taken to migrate the tape management system from the home-grown ADS system to CASTOR2 from CERN. The existing
system is to be run in parallel for older data but for the WLCG data storage, CASTOR2 would be used exclusively.

There have been a number of initial problems with the CASTOR software, notably in the area of reliability and stability. A number of issues were encountered for which fixes were sought from the CATSOR developers at CERN, but the rapid development and issue of newer versions meant that a production service on an older version was less likely to be fixed, the development team electing to provide fixes in later versions. This continued for some months until agreement was reached between CERN and RAL, and other interested parties establishing a framework under which controlled migration from production version to production version is supported at the same time providing for development versions to be tested in near-production instances prior to deployment. This has increased the confidence in more recent versions of CASTOR and stabilised the production instances at the Tier-1.

4. Networking Issues

The internal network within the Tier-1 was based on a small number of gigabit Ethernet switches providing fast access to data resources, and a larger number of 100Mb/s Ethernet switches providing general access to batch workers. The switches were all interconnected using gigabit Ethernet with a single gigabit Ethernet switch acting as a central ‘hub’, which also hosted several of the tape store servers and high access demand servers such as the home file system. The Tier-1 had a 1Gb/s link to the site backbone and from there relied on the site link to the UK wide area network (SuperJANET [2]) for data transfer to other sites, competing with other site traffic. In addition, there was (and is) considerable data transfer from the rest of the RAL site into the Tier-1 subnet to store data on the tape silo.

It was quickly clear from the published data transport requirements for the experiments that 100Mb/s connections for batch servers were not adequate, and that single gigabit Ethernet backbone links would not be adequate either. The published requirements for data import to the RAL Tier-1 also exceeded the available RAL site link to the WAN. Thus a program of improvements and upgrades to link capacities was started.

To cope with the expected data import rates of 150MB/s continuous and 400MB/s for short periods, a new link was developed, which became part of the LHC Optical Private Network (the OPN). This provided connectivity direct from CERN to RAL via two paired 1Gb/s links, running in the UK over the UKLight [3] development network circuits. Progressive upgrades took place over several months to bring the capacity to four 1Gb/s aggregated circuits, and then to a 10Gb/s circuit running over dedicated fibre on the Thames Valley Network, the local WAN.

The RAL Tier-1 is unique in that the four associated Tier-2s in the UK are in fact each a federation of sites. Each Tier-2 supports its own set of experiments but not each site within a federated Tier-2 supports the same experiments and they do not have identical storage resources to network connectivity. Since the data transfer requirements at the Tier1 for data to and from the Tier-2s vary from site to site, each site may be treated as a separate Tier-2. Thus the RAL Tier-1 can expect to be importing data from and exporting data to as many as 21 Tier-2 sites in the UK, and several outside the UK. Since Tier-1 to Tier-2 data traffic must pass over SuperJANET, the RAL site link with its firewall would be a considerable bottleneck. As the SuperJANET backbone network in the UK was being upgraded during this period, the RAL site connection was upgraded from a single 1Gb/s connected to the TVN to redundant 10Gb/s links connected directly to the new SuperJANET5 (SJ5) backbone.

To match the changes in the external links, the Tier1 internal network links were improved. Starting in late 2005, a major move began to stacks of commodity gigabit Ethernet switches with very high speed intra-stack backbone interconnects provided all capable systems with 1Gb/s connectivity. The inter-stack links were formed using four 1Gb/s link aggregated as trunks. As the OPN link speed increased from 2Gb/s to 4Gb/s, the link from the Tier-1 to the OPN end-point at RAL was also...
increased in bandwidth to match. To alleviate the bandwidth restriction on the link to the RAL backbone, the link was doubled to form a trunked pair at 2Gb/s.

The Tier-1 participated in the LCG Service Challenges during the period of these updates. The increasing complexity of the WAN and LAN interlinks highlighted issues with trunked links. It became apparent from network link monitoring that traffic on the WAN links as far as the access routers was balanced across the circuits. However the traffic to the WAN from the LAN and the traffic within the LAN was not being balanced across the aggregated links. The switch stacks were not balancing the data traffic, but were apportioning the various streams based on both source and destination addresses. Data transfer performance was compromised by the pre-existing non-random allocation of network addresses to the data servers, causing one or two links in a trunk set to be fully used and the others under utilized. It became clear that use of trunked links within the Tier-1 was not a long term option, and the existing medium-term plan to upgrade the internal backbone links to 10Gb/s was implemented by adding 10Gb/s capability to the individual stacks. The improvement was immediate, allowing other data server performance issues to be detected and analysed without the restrictions of bandwidth limits on the backbones. The initial ‘daisy-chain’ of 10Gb/s links has been upgraded to a star formation with a central hub built of commodity stackable switches, the same units used to provide the 1Gb/s switch stacks with 10Gb/s uplink capability.

We use the Cacti [4] tool to monitor the performance of the network, which provides long term traffic pattern data. So far data transfer rates within the Tier-1 and on the links to the WANs has been well within the capability of the backbones. However we are now considering whether the topology of the network may have to be revised to split the stacks handing the data servers into smaller stacks each with its own 10Gb/s uplink, or to double the uplinks to the bigger stacks, to ensure that data transfer traffic will not be limited.

5. Conclusion
The RAL Tier-1 has made good progress towards readiness for LCH data processing. Important lessons have been learnt about the types of hardware needed for running services, and the various methods of distributing services or service instances across multiple hosts. Elements of the software stack, particularly CASTOR, are now more stable and reliability in increasing. It remains for storage systems tuning work to be finished so that optimum performance can be attained for the various data transfer streams.

References
[1] http://www.spec.org/
[2] http://www.ja.net/
[3] http://www.uklight.ac.uk/
[4] http://www.cacti.net/