

## **LARGE HADRON COLLIDER COMMITTEE**

### **LCG PROJECT COMPREHENSIVE REVIEW**

November 2007

#### **1. EXECUTIVE SUMMARY**

The fifth annual LHCC Comprehensive Review of the LHC Computing Grid (LCG) Project took place on 19-20 November 2007. The LHCC referees addressed the following areas: Management, Resources and Collaboration; Mass Storage and Networking; Distributed Fabric; Middleware Development and Deployment; Applications Area and Distributed Databases; and the issues of Services and Experiment Readiness. The LHCC acknowledges the considerable amount of work that has gone into the preparation of the LCG Project Comprehensive Review.

The LCG Project was created by the CERN Council in September 2001 with the aim of prototyping and deploying the computing environment for the LHC experiments. The formal launch of the project was at a workshop held in March 2002. Since that time, the LCG has demonstrated progress towards the realisation of the computing requirements of the experiments in time for LHC operation in 2008.

The LCG Project is a collaboration of the LHC experiments, the Regional Computing Centres, CERN and the physics institutes with the aim of preparing and deploying the computing environment that will be used by the LHC experiments to analyse the LHC data. The project includes support for applications and the development and operation of a computing service.

The LCG Project is divided into two phases. Phase I (2002-2005) had the objective of building a service prototype, based on existing Grid middleware, of running a production Grid service and of producing the Technical Design Report for the final system. Phase II (2006-2008) is building and commissioning the initial LHC computing environment. The LCG is not a Grid development project and it relies on other Grid projects for the middleware development and support.

The LHCC considers that the LCG Project has shown significant progress since the last Comprehensive Review in both the production and analysis phases and that the World-

wide distributed LCG (WLCG) is becoming a reality. In particular, considerable progress has been achieved in the stability, usage and interoperability of the Grid infrastructure and in the use of the Grid by the experiments for analysis; certain of the Tier centres have installed and run successfully the necessary hardware, while the Storage Resource Manager SRM v2.2-compliant services are being deployed, albeit after lengthy delays; the middleware services are in place and focus has shifted to ensuring stability of the installed features. The service level has constantly improved and the nominal data transfer through-put rate for 2008 has been achieved. A number of useful products have also been delivered by the Applications Areas and significant progress has been reported on the monitoring and reporting performance for both generic and experiment-specific issues.

However, the Committee did note some concerns. The Grid infrastructure has been only partially exercised and the analysis models are not yet fully defined. Although site stability and reliability have improved, they are not yet at the desired level and the required support model, especially for 24x7 operations, is not yet fully defined. The deployment schedule for the mass storage management and operation remains one of the most critical issues for the WLCG and its schedule is extremely tight for the upcoming Combined Computing Readiness Challenge CCRC08, which is an important milestone as all the services should be tested in their complete capacity at the same time for all the experiments prior to the start of LHC operations. Finally, it is also important that all pledged resources for 2008 are available for the CCRC08 in order to exercise the full system. The long-term guarantee of resources, both manpower and hardware, for the long-term also remains a concern.

The conclusions and concerns of the LHCC are given below. They will help the Committee to follow up outstanding issues and to monitor future progress of this project in forthcoming sessions of the LHCC.

## 2. OVERVIEW

- Considerable progress has been achieved in the stability, usage, and interoperability of the Grid infrastructure. The LHC experiments routinely use the Grid for production activities, with approximately 20% of the production done at CERN, 40% at the Tier-1 centres, and 40% at Tier-2 centres. Great progress has also been reported in using the Grid for analysis, although the system has only been partially exercised, and the analysis models are not fully defined.
- As part of the system commissioning, the CCRC08 has been planned for 2008, with two windows, in February and in May. The CCRC08 is an important milestone, when all the services should be tested at full capacity *at the same time* for all the experiments.
- A certain number of the Tier centres have proven to be able to install and run the necessary hardware. Still, the support model, especially for 24x7 operations, is

not well defined. Measures have been developed to monitor the site reliability. The improvement is constant, but the site reliability is not yet at the required level. The measure derived from both basic tests and experiment-specific tests should be published to have a better understanding of the site situation. The CERN Tier-0 centre seems to be well equipped to cope with the expected data rates. Operational issues are being addressed, but a full plan is not yet in place. Networking is in general adequate.

- Mass storage management and operation remains one of the most critical issues for the LCG. After many delays, the SRM v2.2-compliant services are being deployed. The experiments have done only limited tests on the new services. The deployment schedule is extremely tight for the CCRC08. The CASTOR2 system has at the end of October 2007 released a new version fixing many issues, but it remains to be seen whether this new version has the required stability and performance. Deployment at remote sites is essential in order to reduce service difficulties. dCache v1.8 deployment has begun well, but its overall schedule remains very tight.
- Basic middleware services are in place and the focus of the past year has been more on stability than on new features. The gLite middleware failed to deliver the computing element software. Development of this has been stopped and replaced with Web-based services (CREAM). The LHCC is concerned by the announced manpower cuts of European Union funds in the area of middleware development. The real test of the system will only happen with the first data, and it is essential that key experts are retained until then. A continued focus on stability and production quality software is essential for the coming year.
- The Application Area is in general in very good shape. The long-awaited migration from SEAL (Shared Environment for Applications at LHC) to the ROOT interactive tool for analysis is delayed because of unforeseen although not fundamental difficulties. The manpower level is matched to the requirements, but turnover can become a problem if the overlap is not sufficient to ensure proper knowledge transfer. Once again, in the view of the Committee, retention of key experts until the system has been fully exercised with real data would be very prudent. The 3D distributed database project is well on track.
- The service level has constantly improved and the nominal data transfer throughput rate for 2008 has been achieved, although only for short periods. Sustained, stable operation still needs to be achieved. Planning for the CCRC08 requires a strong coordination among all sites, a robust data transfer and management service, and a clear definition of the criticality of the various services as well as a failure recovery mechanism.
- Resources for the WLCG are provided through a Memorandum of Understanding for which, however, some signatures are still missing. Not all the pledged resources have been deployed in time and in particular the scarce disk space has

caused significant problems to the production activities. In addition, a large fraction of the Tier-1 and Tier-2 sites have not confirmed the 2008 pledges. It is important that the pledged resources for 2008 are available for the CCRC08 in order to exercise the full system prior to LHC data taking. Up to 2010, the pledged resources match reasonably well the experiments' requirements, except for ALICE, where a 50% short-fall is still observed. The pledged resources for 2011-2012 are significantly below the experiments' requirements.

- Communication with remote sites has improved significantly, and a system to plan and track the progress of each site (especially the Tier-1 centres) has been set up. Many level-1 milestones are not being met, including the 24x7 support, the VO boxes support, and the site reliability. Stronger coordination is required to ensure that the sites are ready for CCRC08 and for data taking.

### **3. MANAGEMENT, RESOURCES, AND COLLABORATION**

The organizational structure of World-wide LCG (WLCG) is unchanged with respect to the previous Comprehensive Review. The ramp-up of Tier-1 resources during last year has proceeded largely according to plan, although the actual installations tend to exhibit a slower increase than the pledges which foresee a steep jump in July 2008. As a result of this, parts of the installations have been lagging behind the requirements. Resources for the WLCG are provided under a Memorandum of Understanding (MoU) for which some signatures are still missing and about half of the remote sites have not yet confirmed their pledges for 2008. It is important that the pledged resources for 2008 are available for the CCRC08 in order to exercise the full system prior to LHC data taking. As of October 2007, the vast majority (80%) of the CPU used within WLCG has been provided from the Tier-1 and Tier-2 centers outside of CERN. The experiments have recently revised and updated their resource projections through 2008-2012, and these resources have been compared to the present status of confirmed pledges by the external centers. For ALICE, ATLAS and LHCb the resources are roughly balanced between CERN and the outside centres. In particular for 2008, the ALICE experiment still sees a significant overall short-fall of resources of up to 50%. The long-term perspective displays a slackening of the ramp-up, which might lead to significant overall resource deficits by 2012, which is however still a preliminary observation since not all pledges have been updated. This long-term development will have to be closely monitored. The recently formed Resource Scrutiny Group (RSG) is expected to play an important role in this process. Collection of the MoU signatures from the Tier-1 federations is almost complete.

Reporting and monitoring within WLCG have considerably improved, and a common set of milestones for the Tier-1 centres has been established. Some high-level milestones, in particular in connection with implementing 24x7 support and service level agreements of VO boxes, are significantly late with some centres. Site availability and reliability metrics are routinely recorded, and downtimes followed up

in weekly operations meetings. Implementing experiment-specific job efficiency monitoring into the WLCG-level monitoring is still an ongoing project.

The previous Comprehensive Review had emphasized the need for a continuation of projects supporting the Grid infrastructure on the long term. While the current Open Science Grid (OSG) funding cycle extends until 2010, the Enabling Grids for e-Science in Europe (EGEE) funding ends in April 2008. An extension of EGEE until 2010 has been proposed. There are indications that this extension might only be granted with a significant cut in software support. Such a cut would come at a problematic time, since the final shake-down of the middleware through Full Dress Rehearsals and the CCRC08 will be well underway. It is important to ensure that key positions are not affected. A long-term perspective might open through the European Grid Initiative Design Study (EGI-DS), but it is yet unclear whether this will result in sufficient support for WLCG at the relevant time scale.

The review did not display much in terms of concepts for end-user analysis beyond the Tier-2 level. This area plays an important role for the final derivation of physics results, and deserves more attention in the future.

#### **4. MASS STORAGE AND NETWORKING**

The three storage systems, CASTOR2, dCache and Disk Pool Management (DPM), fill three different roles for storage. Between the three systems there are about 180 instances installed and in use. CASTOR2 is in use at the Tier-0 centre at CERN, at RAL and at CNAF. Stability and performance problems have been observed at all centres. The CASTOR2 system has released a new version at the end of October 2007 fixing many issues, but it remains to be seen whether this new version has the required stability and performance. Deployment to remote sites is essential to reduce service difficulties. The CASTOR2 performance at RAL limited the effective use of the UK Tier-2 centre disk for ATLAS. The source of these problems has been understood and mitigated, with improved procedures put in place for deployment at RAL. Further development on the tape layer is becoming essential.

The Site Availability Metrics (SAM) give a good handle on the performance of the systems and represent a significant advance in understanding the state of the storage deployments at the sites, and it is clear that this framework will be used for VO specific tests, giving additional and better indicators of a site's ability to deliver services to the user communities. Additionally, improved diagnostic tools have been requested by the sites. Requirements for these tools should be gathered and the development prioritized. In order to promote understanding of the systems, it is strongly urged that metrics are developed and published to enable users to evaluate the performance of the storage systems over a sustained time period.

Ensuring support for the necessary development and implementing a long-term viable support model deserves serious consideration. Some steps have been taken along

these lines, at an appropriate level for this phase of the project, with more planning necessary to achieve smooth operations and well-planned upgrades into the future.

Delivering SRM v2.2 implementations to provide experiment-requested functionality has been a focus of the storage projects, with substantial progress in the past year, although with significant delays. The CASTOR2 and DPM releases which support SRM v2.2 have been released although not yet fully deployed. The roll-out of dCache v1.8 which supports SRM v2.2 started in early November 2007, with two sites completed, and the rest scheduled before the end of 2007. While some problems have been noted with all of the implementations, they have largely been patched or do not impact substantially the deployment. With CERN running the interoperability test suite, it is largely felt that from the storage perspective the SRM v2.2 functionality has been well tested. Furthermore, backwards compatibility with SRM v1.1 enables sites to upgrade with minimal impact to the experiments. The experiments are beginning to test usage of SRM v2.2, with LHCb having moved along the furthest. Preliminary ATLAS testing enabled the identification and fixing of several problems. However, further work is required in this area for the experiments to reap the benefits. Planning for the deployment and the adoption by the experiments began at a workshop in November 2007 which included 60 participants and the planning will continue in December 2007 as it is highly desired for SRM v2.2 to be in use for the first phase of the CCRC08 in February 2008. The proposed roll-out and integration schedule for February 2008 is extremely tight and will require substantial planning, and significant co-ordination will be required to execute that plan, particularly with respect to the experiment framework updates. Development teams will have to be sensitive to the trade-offs of new functionality and stability should be considered with care, particularly since the inevitable shakedown and operational problems that will cause some amount of instability as the experiments increase their usage.

The CERN Tier-0 centre continues to demonstrate progress towards readiness in all key areas of supplying computing, disk and tape storage and networking and the facilities issues associated with running a large computing plant. Perturbations to the computing models have required additional hardware and re-allocation of funds. In addition to the immediate term operational issues and the expansion of the compute, disk and tape plants, long ranging planning is in place for a new facility and for making technology projections. One area of concern is whether the tape drive capacity is sufficient. This is coupled to the storage development in CASTOR2. Additionally, the DAQ to Tier-0 link has been demonstrated for ALICE, ATLAS, and LHCb, and with CMS scheduled for January 2008. The CCRC08 will be the first time when all experiments will exercise the Tier-0 centre performance simultaneously.

## **5. DISTRIBUTED FABRIC**

The LHCC heard reports from certain of the Tier-1 centres (FNAL, TRIUMF, GridKA) and Tier-2 centres in the Asia-Pacific region, the UK and the US.

Over the previous year, the Tier-1 and Tier-2 sites have demonstrated great improvements in stability of operations, as shown by the routine site availability monitoring. The Committee encourages publishing the experiment-specific metrics along with the basic functionality tests to allow for a better understanding of the site performance. An improved stability of the middleware during the last year, as well as the approach of actual data taking, have been key factors in reaching these results. Still, a good overall fraction of the sites do not perform at the required level and continued attention and improvement is required.

The sites above reviewed by the LHCC have the ability to deal with the required high number of CPUs, big storage systems, tapes, databases, and networks. However, the ability to implement the fast ramp-up of resources in a short period of time (end 2007-middle 2008) remains to be tested, as well as the actual impact of the SRM v2.2 migration foreseen in the coming months on Tier-1 and especially on Tier-2 sites. Coordination and collaboration within the Tier-1 and Tier-2 centres, which remains a key factor for the success of the operations and must be strongly encouraged in the future, has reached a good level in the UK and US but it must still be fully implemented in the Asia-Pacific region. Support for 24x7 operations has not been fully tested and operational models are not well defined. This is especially important for the Tier-1 sites. Software support from the middleware development and deployment team is essential for a prompt resolution of problems and must be made available at the Tier-1 and Tier-2 centres.

## **6. MIDDLEWARE DEVELOPMENT AND DEPLOYMENT**

The LHCC congratulates the WLCG for the impressive progress reported since the last Comprehensive Review in the development and deployment of the middleware. This progress followed an initial period of issues related to the proper functionality of the pre-production middleware. The LHC experiments continue to rely strongly on services provided by WLCG. The EGEE and OSG projects have now reached a well-established interoperability thanks to the existing levels of communication and coordination across the WLCG middleware activities. Both EGEE and OSG maintain a consistent activity on software development, validation and deployment of services across the Tier centers. The LHCC appreciates the fact that the WLCG now focuses on improvements to the stability and reliability of the existing services, in preparation for the CCRC08 milestone and towards the best possible performance at the LHC start-up in 2008. Very good progress has been reported on middleware tools for monitoring of the status of the Grid nodes. Progress has been reported on usage accounting and job priority assignment.

The Committee notes a change in strategy in the Computing Element (CE) software, where the gLite development has been stopped and replaced with Web-based services (CREAM). The Committee is concerned by the announced manpower cuts of European Union funds in the area of middleware development. The preliminary

indications are that there will be an approximately 50% cut in the middleware development, a 40% cut in applications support, and no cut in operations support. The real test of the system will only happen with the first data, and it is essential that key experts are retained until then. A continued focus on stability and production quality software is essential for the coming year.

## **7. APPLICATIONS AREA AND DISTRIBUTED DATABASES**

Both the Applications Area and the 3D distributed database project have made good progress and are currently in good shape for LHC operations. The Geant4 detector simulation of physics processes has significantly improved, albeit at a cost of about 50% increase in CPU usage. The event generator repository has been restructured and made easier for the user. A new FLUKA graphical user interface allows easy simulation of hadronic processes. Many new features have been added to ROOT in graphics and analysis. At the same time, the programme has become leaner in memory usage. A new bootstrap process is being developed to allow the user to only use those pieces of ROOT that are needed. PROOF allows user analysis which uses multi-processor/multi-core systems very effectively. The persistency framework, which is used in whole or in part by LHCb, ATLAS and CMS, is functional and undergoes steady improvement. The software process infrastructure has been restructured to use the Python object-oriented programming language and LCGCMT, a set of configuration files which contain instructions on how to use and build software provided by the LCG Applications Area, as a unified generating process and now include nightly builds. The 3D project is operational and has been tested successfully by the experiments. There is a good co-operation with ORACLE through the CERN *openlab*. Concerns in this area are the phasing out of SEAL, which is to be absorbed in ROOT, and of CORAL (Common Relational Abstraction Layer), which is going significantly slower than foreseen, and the decline and turnover of human resources, which may endanger retention of the necessary expertise in some areas.

## **8. SERVICES AND EXPERIMENT READINESS**

The baseline services include the storage element, file transfer and catalogue services, the synchronization of the databases, the compute element, the workload management system, information, monitoring and the management tools. Most of the services are in place, and the main goals of the last year have been to firstly deploy the residual services and to improve the reliability and secondly improve the performance and the capacity of the services. The residual services, such as the new version (SRM v2.2) for the Storage Resources Manager interface or the file transfer services (FTS), have been released and are in the deployment stage. These items had been pointed out as issues at the previous Comprehensive Review. The LHCC notes that an increased effort has been put in place for the monitoring, the accounting and the reporting of the services. Further efforts should focus on the ramping up of resource capacity, data transfer rates and stability. The corresponding scaling factors are now available for

the four experiments. They have also defined recently their needs for the most critical services. The rates in data transfer and data storage did not reach the nominal values – except during short periods - but the basic functionality and tools of WLCG are available and used extensively by the experiments.

The LHCC was impressed by the reports from the experiments describing that over the last year the four experiments deployed extensive Data Challenges in order to test the performance and reliability of the WLCG services. These Data Challenges include simulations, cosmic runs, reconstruction and analysis of the data. For all four experiments, a very large number of events have been successfully processed at the Tier-0, Tier-1 and Tier-2 sites. ALICE did saturate all available resources during the last four months. It is expected that this experiment will face a general resource short-fall in the future regarding both the proton and heavy ion runs. LHCb used successfully the new version of SRMD for the production of their data. An ongoing CMS Data Challenge (CSA07) should achieve soon 50% of their nominal computing requirements. ATLAS has used extensively the GRID for the transfer of cosmic data from the detector up to Tier-2 sites for analysis. It should be noted that all ten ATLAS Tier-1 sites are now in production operation. This should be considered as a major milestone for LCG services. The progress over the last year in services for the four experiments is, however, not uniform. The Committee notes that storage is late and harder than expected and that data movements continue to be a key area of concern.

The next important milestone for WLCG will be the CCRC08 data challenge planned for February and May 2008. The February tests will be essentially an integration challenge. It is clear that all services and resources will not be available for this period. The complete challenge is foreseen in May and should last four weeks. The CCRC08 will be a combined challenge by all four experiments and will be used to demonstrate the readiness of the services provided by LCG at a scale comparable to real LHC data taking. One of the main goals of CCRC08 will be to test the full transfer data matrix such as in particular *Tier-1-to-Tier-1*, *Tier-2-to-Tier-1* and the *Tier-2-to-Tier-2* transfers. The CCRC08 challenge should be done well in advance with respect to the real data taking period in order to identify and correct flaws and bottlenecks. During the coming six months, each experiment will continue or start new data challenges such as cosmic runs or Full Dress Rehearsals. It is clear that interferences between these challenges and the CCRC08 combined challenge have to be checked. The Committee notes that the overall schedule for the CCRC08 challenge appears to be very tight.