Deploying the LHC Computing Grid – The LCG Service Challenges

Ian Bird
CERN, Geneva, Switzerland
Ian.Bird@cern.ch

Les Robertson
CERN, Geneva, Switzerland
Les.Robertson@cern.ch

Jamie Shiers
CERN, Geneva, Switzerland
Jamie.Shiers@cern.ch

Abstract

The world's largest scientific instrument, the Large Hadron Collider (LHC) is currently being assembled near Geneva, Switzerland. When operational, several petabytes of data will be generated every year for a period of at least ten years. These data will be acquired at rates up to nearly 2GB/s and will be analysed by thousands of physicists worldwide. In order to exploit the full discovery potential of the LHC, a worldwide Grid is currently being deployed. As part of the commissioning of this Grid, a series of service challenges is currently being conducted, ramping up the service progressively. These challenges address not only the need to distribute data reliably between many sites around the world – not in burst mode but 24x7 for essentially all of the production lifetime of the machine, but also and much more importantly meet the needs of the experiments for all of their offline data processing.

1. Introduction

Four large experiments – each being a collaboration of thousands of people from hundreds of institutes in tens of countries – will take data at the LHC using massive detectors up to 35m in length and 25m in height.

The raw data rates from the detectors varies from around 300MB/s during proton-proton running for ATLAS and CMS, to 1.25GB/s during heavy ion running for ALICE.

For a variety of reasons, assembling sufficient processing power and storage capacity to enable the analysis of the data that will be produced in a single place is excluded. Instead, an approach based on Grid technologies has been adopted, exploiting the resources of institutes worldwide. The LHC Computing Grid project – or LCG – is providing the distributed production environment for physics data processing for the LHC experiments. This is being performed in conjunction with other national or regional grid projects, including EGEE and the OSG to name but two.

As part of the ramp-up of the LCG services to full production level, a series of "Service Challenges" is in progress. The first two of these challenges have already taken place and focused on attaining reliable file transfer services between some of the main sites involved in the LCG for extended periods of time in a production-like environment. The remaining two challenges – scheduled for the 2nd half of 2005 and mid-2006 respectively – need to build on this infrastructure work to provide services satisfying the full requirements of the LHC experiments for processing and analyzing their data at progressively higher data rates involving more and more sites. The setup that is used for the final service challenge becomes the initial production service in Q3 2006 – an extremely aggressive timescale given the overall complexity of the problem.

2. Overview of the LCG

Broadly speaking, the LCG adopts a hierarchical model involving Tier0 – Tier*n* sites with functionality that differs slightly by experiment. The Tier0 is defined as the host laboratory at which the data is acquired and at which a full copy of the raw data is stored. First-pass processing ("reconstruction") of the data is also performed at the host laboratory, with copies of the raw data and the output of the reconstruction being distributed across some 6 – 12 Tier1 centres per experiment. During proton-proton running, this distribution of data needs to keep up with the arrival of new data from the detectors, whilst for the heavy ion running it is foreseen to spread this out over the four month winter shutdown of the accelerator. The fraction of the data that is sent to a given Tier1 site will likely depend on the available capacity at that institute, the number of active physicists served on the corresponding experiment, by physics channels as well as by regional interests. As a very first approximation we first assume an equal split across all sites. This gives a data rate per

Tier1 between some 50MB/s and 200MB/s, although in reality a figure of some 150MB/s per site with peaks of some 200-250MB/s are more likely.

Of the four LHC experiments – ALICE, ATLAS, CMS and LHCb – most maintain a single copy of the raw and reconstructed data spread across all Tier1s, with a 2nd full copy being maintained at CERN. ATLAS intend to store two copies of the reconstructed at the Tier1s with an additional full copy at Brookhaven National laboratory (BNL) in the US. As well as storing these data, the Tier1 sites are responsible for reprocessing of these data, as improved calibrations and algorithms are developed, serving the end-user analysis needs of physicists as well as providing additional services to the Tier2 sites. The Tier2 sites – which typically number some 15 – 30 per experiment – are largely devoted to the production and processing of simulated data, improved calibrations and also in most cases end-user analysis. They do not provide long-term archival storage – this being one of the main services provided to them by the Tier1 sites (another important service being the delivery of analysis data as needed by the local user community). This model is clearly simplistic, but serves to provide an outline of the responsibilities of each site and the services that they require and / or offer.

Table 1. Tier1 Centres for the LCG

Location	ALICE	ATLAS	CMS	LHCb
Amsterdam (NIKHEF/SARA)	Yes	Yes		Yes
Barcelona (PIC)		Yes	Yes	Yes
Batavia (FNAL, IL)			Yes	
Bologna (CNAF)	Yes	Yes	Yes	Yes
Brookhaven, (BNL, NY)		Yes		
Distributed (Nordic)	Yes	Yes		
Didcot (RAL)	Yes	Yes	Yes	Yes
Karlsruhe (GridKa)	Yes	Yes	Yes	Yes
Lyon (CCIN2P3)	Yes	Yes	Yes	Yes
Taipei (ASCC)		Yes	Yes	
Vancouver (TRIUMF)		Yes		

The Tier1 sites that have currently been identified are shown in the table above. Whilst many of the European sites support all or several of the experiments, this is not currently true for the sites elsewhere, although the size of communities that the latter serve needs also to be taken into account.

3. Roll out of the LCG Service

At the time of writing, the LHC accelerator and the detectors are in the process of being assembled. First collisions are expected in mid-2007, although data generated by the interactions of cosmic rays in the detectors will begin in the summer of 2006. A significant amount of work on developing the offline computing models and software of the experiments has been underway for many years now and for these purposes it is necessary to ramp up the services earlier still. As is shown in the figure below, the initial service needs to be in place as early as September 2006.

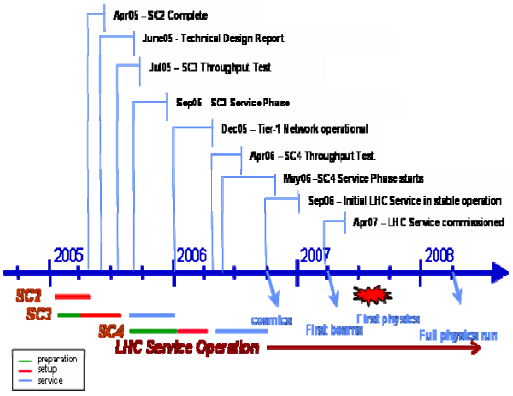

Figure 1. Roll out schedule for the LCG

4. Summary of Service Challenge 1

This initial service challenge had rather modest goals. It did not attempt to transfer real files, nor was tape storage involved. The primary motivation was to understand the issues involved in running high throughput transfers for extended periods of time in full production mode. Although more sites took part in this challenge than originally foreseen – and significant valuable experience was gained – the main target of running for a prolonged period was not achieved. This served to underline the fact that whilst much can be achieved by a super-human effort for short periods, this is not a realistic solution for services than need to run 24x7 for many months at a time for a total period in excess of a decade.

5. Summary of Service Challenge 2

Service challenge 2 was intended to build on the previous challenge, adding additional sites – each of which should sustain transfers of 100MB/s disk to disk – with an aggregate throughput out of CERN of 500MB/s. It was intended that this rate be maintained for 10 days, after which transfers to individual sites would be pushed to the maximum. As can be seen from the figure below, the throughput goals were exceeded, with a total of 500TB being transferred at average rates of some 600MB/s.

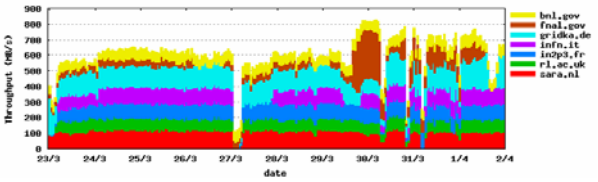

Figure 2. Throughput to Multiple Sites

A rate of 800MB/s was achieved between CERN and Fermilab alone, as shown in the plot below. This plot also shows that the link itself was severed by a trawler in the Atlantic, as well as the traffic failing over to a backup link. Unfortunately, the effort required to perform these transfers was still very high and much work remains to be done to move to full production service mode.

The sites that took part in this challenge were BNL, CNAF, FNAL, FZK, IN2P3, NIKHEF and RAL, together with CERN.

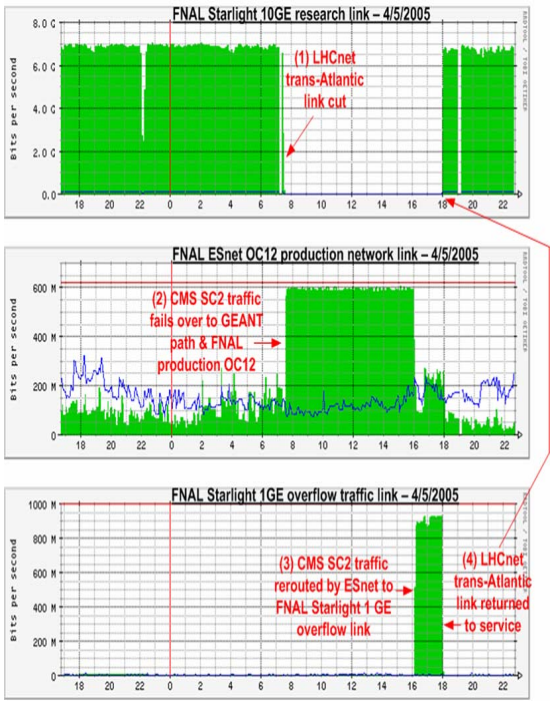

Figure 3. Transfers to Fermilab

6. Milestones for Service Challenge 3

This challenge consists of a setup phase, which includes a throughput test, followed by a much more important service phase. It is intended that the throughput phase include both disk – disk transfers at 150MB/s per Tier1, with a total aggregate bandwidth out of CERN of 1GB/s sustained over a period of some 10 days, followed by a more modest 60MB/s from disk to tape at the Tier1s. The main goal of this phase is to demonstrate that the infrastructure is ready for exercising more realistic transfers and use cases as well as providing a baseline against to which other tests can compare.

Following the throughput phase, the challenge will ramp up in complexity, simulating the file sizes and access patterns of the experiments before exercising the offline software of the experiments directly to generate and process the data. This will involve numerous additional software components and services and it is felt essential to proceed incrementally, allowing time to debug and resolve problems before a full and extended production phase of some 4 months can be started.

The service phase should exercise in production mode all primary offline use cases of the experiments except for analysis, the latter being a primary goal of Service Challenge 4.

This challenge includes also a limited number of Tier2 sites, as shown below.

Table 2. Initial Tier2 Sites for SC3

Site	*Tier1*	*Experiment*
Bari, Italy	CNAF, Italy	CMS
Turin, Italy	CNAF, Italy	Alice
DESY, Germany	FZK, Germany	ATLAS, CMS
Lancaster, UK	RAL, UK	ATLAS
London, UK	RAL, UK	CMS
ScotGrid, UK	RAL, UK	LHCb
US Tier2s	BNL, FNAL	ATLAS / CMS

As the role of the Tier2 sites in this challenge will be mainly the production and processing of simulated data, and as significant compute resources are required for this purpose, rather modest data rates between the Tier2 and Tier1 sites are expected. Instead, the main focus will be on functionality and reliability with the aim of demonstrating reliable file transfer rather than high throughput. The following table shows the resources required – in kSI2K seconds – to reconstruct and simulate proton proton and heavy ion events for the different experiments (LHCb has no heavy ion programme). Given that Tier2 sites currently provide between 10 and 1000 kSI2K years and that the event size involved is around 2.5MB (ATLAS raw + reconstructed simulated event) it is simply not possible to generate a high data rate from Tier2 sites to the Tier2 sites – where the output of the simulation is archived – even if all compute resources at the Tier2s are devoted to this task.

Table 3. Resources Required for Simulation

	ALICE		ATLAS	CMS	LHC
	pp	Pb-Pb	pp	pp	b
reconstruction	5.4	675	15	25	2.4
simulation	35	15000	100	45	50

7. Overview of Service Challenge 4

Service challenge 4 needs to demonstrate that all of the offline data processing requirements expressed in the experiments' Computing Models, from raw data taking through to analysis, can be handled by the Grid at the full nominal data rate of the LHC. All Tier1 sites need to be involved, together with the majority of the Tier2s. The challenge needs to successfully complete at least 6 months prior to data taking. The service that results from this challenge becomes the production service for the LHC and is made available to the experiments for final testing, commissioning and processing of cosmic ray data. The analysis involved is assumed to be batch-style analysis, rather than interactive analysis, the latter expected to be performed primarily "off the Grid". The setup phase ends with a throughput demonstration sustaining for three weeks the target data rates at each site as defined in the following table. The throughput is measured network-tape at each Tier-1, and disk-network at CERN. The target date for completing the throughput test is end April 2006.

The service phase of Service Challenge 4 will include the basic software components required for the initial LHC data processing service, as defined in the LCG Technical Design Report. The service must be able to support the full computing model of each experiment, including simulation and end-user batch analysis at Tier-2 centres. The service phase is scheduled to operate for four months from May to September 2005.

In parallel, the various centres need to ramp up their capacity to twice the nominal data rates expected from the production phase of the LHC, to cater for backlogs, peaks and so forth. The analysis involved is assumed to be batch-style analysis, rather than interactive analysis, the latter expected to be performed primarily "off the Grid". The total aggregate data rate out of CERN that needs to be supported is double that of Service Challenge 3, namely 2GB/s.

8. Initial LHC Service

The initial LHC service is scheduled to enter operation by end September 2006, capable of handling the full nominal data rate (see Table 2). The service will be used for extended testing of the computing systems of the four experiments, for simulation and for processing of cosmic data. During the following six months each site will build up to the full throughput needed for LHC operation, twice the nominal data rate.

9. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [1]. Multiple citations [1, 2] should be in a single set of square brackets.

10. Summary

The service challenges are a key element of the strategy for building up the LCG services to the level required to fully exploit the physics potential of the LHC machine and the detectors. Starting with the basic infrastructure, the challenges will be used to identify and iron out problems in the various services in a full

production environment. They represent a continuous on-going activity, increasing step-wise in complexity and scale. The final goal is to deliver a production system capable of meeting the full requirements of the LHC experiments at least 6 months prior to first data taking. Whilst much work remains to be done, a number of parallel activities have been started addressing variously the Tier1/2 issues, networking requirements and the specific needs of the experiments. Whilst it is clear that strong support from all partners is required to ensure success, the experience from the initial service challenges suggest that the importance of the challenges is well understood and that future challenges will be handled with appropriate priority.

11. Acknowledgements

The work described above involves many individuals, institutes and projects. In particular, the Tier1 sites listed in table 1 have played a significant role in the initial Service Challenges. The contribution of all of these partners is gratefully acknowledged.

Please direct any questions to Jamie Shiers – Jamie.Shiers@cern.ch

CERN, the European Organization for Nuclear Research, has its headquarters in Geneva. At present, its Member States are Austria, Belgium, Bulgaria, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Italy, Netherlands, Norway, Poland, Portugal, Slovakia, Spain, Sweden, Switzerland and the United Kingdom. India, Israel, Japan, the Russian Federation, the United States of America, Turkey, the European Commission and UNESCO have Observer status.

Brookhaven National Laboratory conducts research in the physical, biomedical, and environmental sciences, as well as in energy technologies and national security, and builds and operates major scientific facilities available to university, industry and government researchers. BNL is operated and managed for the U.S. Department of Energy's Office of Science by Brookhaven Science Associates, a limited-liability company founded by Stony Brook University, the largest academic user of Laboratory facilities, and Battelle, a nonprofit applied science and technology organization.

CCIN2P3, the Computing Center of the National Institute of nuclear physics and particle physics is located in Lyon, France. Its main mission is to provide computing resources and storage of experimental data to the physicists of the Institute involved in the major experiments of the discipline and particularly in international collaborations. In the field of grid computing, CCIN2P3 is one of the leaders of the French grid effort and is deeply involved in the main European grid projects for science.

Fermi National Accelerator Laboratory is located in Batavia, Illinois, USA. Fermilab is operated by Universities Research Association, Inc., a consortium of 90 research universities, for the United States Department of Energy's Office of Science.

Forschungszentrum Karlsruhe, a member of the Helmholtz Gemeinschaft Deutscher Forschungs-zentren (HGF), constructs and operates the GridKa computing center for the German particle physics community and is the designated German Tier 1 for the LHC.

INFN-CNAF is the National Center for Research and Development in Technology, Computer Science and Data Transmission of INFN (Istituto Nazionale di Fisica Nucleare), and is the major computing facility of the INFN grid infrastructure. INFN, Italy's national nuclear physics institute, supports, coordinates and carries out scientific research in sub-nuclear, nuclear and astroparticle physics and is involved in developing relevant technologies and a significant outreach program.

SARA is the National Center for Computing and Networking Services and NIKHEF is the National Institute for Nuclear Physics and High Energy Physics in the Netherlands. The two institutes have joined forces to become an important LHC data storage and analysis center. The Advanced Internet Research Group of the University of Amsterdam has strongly contributed to this Service Challenge in manpower and equipment.

The UK Council for the Central Laboratory of the Research Councils (CCLRC) works with the other UK research councils to set future priorities that meet UK science needs. It also operates three world class research centres: the Rutherford Appleton Laboratory in Oxfordshire, the Daresbury Laboratory in Cheshire and the Chilbolton

Observatory in Hampshire. These world-class institutions support the research community by providing access to advanced facilities and an extensive scientific and technical expertise. RAL is one of the collaborating institutions in the GridPP project, the UK's contribution to the LHC Computing Grid project.

Web sites:

- LHC Computing Grid (LCG) project: http://www.cern.ch/lcg/
- Enabling Grids for E-SciencE (EGEE): http://public.eu-egee.org/
- Grid3: http://www.ivdgl.org/grid3/
- GridPP: http://www.gridpp.ac.uk/
- INFNGrid: http://grid.infn.it/
- Open Science Grid (OSG): http://www.opensciencegrid.org/