PROCEEDINGS OF SCIENCE

Progress in Computing

Ian Bird[1]
CERN
CH-1211, Geneva 23, Switzerland
E-mail: Ian.Bird@cern.ch

In the light of the experiences of the first few months with LHC data, the progress in building the grid infrastructure is reviewed, describing the programme of testing that was essential for creating a working grid. Some of the key achievements of the experiments are noted, particularly the remarkable success of producing physics results within weeks, as highlighted in this conference. Building on some of the lessons learned in the preparation phase as well as early experience with data, the outlook for evolution of the infrastructure in the future is described.

35th International Conference of High Energy Physics (ICHEP2010)
Paris,France
July 22-28, 2010

[1] Speaker

PoS(ICHEP 2010)530

1. Introduction

After just a few months of experience with LHC data taking, the four major experiments together with the Worldwide LHC Computing grid (WLCG) have shown that large scale distributed computing can be successful for High Energy Physics. Physics results have been shown at this conference based on data taken only days before – a remarkable achievement based on many years of preparation and testing.

1.1 The Worldwide LHC Computing Grid

The LHC Computing Grid (LCG) project [1] was approved in 2001 to prototype, develop, and deploy the computing environment for the LHC experiments. The distributed computing system was based on the MONARC model [2], which introduced the concept of Tiers of computing centres with data being processed and refined as it flowed outwards from CERN. The implementation of this model has been based on grid [3] technology, and has taken several years to develop sufficiently to provide the capabilities and robustness required by the experiments. In the second phase of the project, a Memorandum of Understanding [4] was adopted to provide a foundation for a collaboration – the Worldwide LHC Computing Grid (WLCG) – as the mechanism for the long term support and management of this computing environment. To date, 49 funding agencies have signed this MoU, representing 11 Tier 1 sites, and some 120 Tier 2 sites in 34 countries. The truly global extent of this collaboration can be seen in Figure 1. The original scale of resources expected for the first nominal years of LHC data taking were some 200, 000 CPU and 45 PB of disk, whereas in place in mid-2010 were in excess of 250,000 CPU cores, and close to 100 PB of disk.

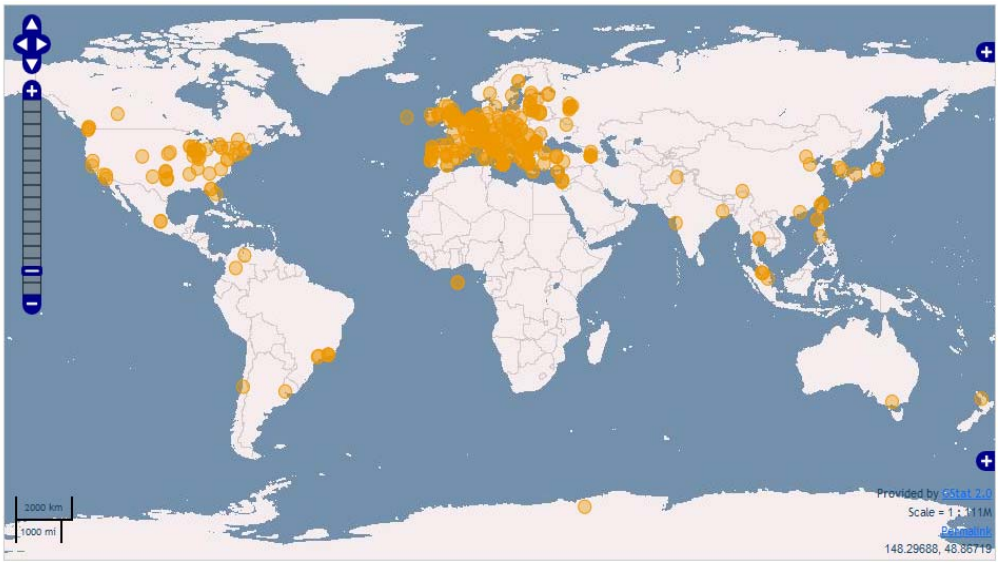

Figure 1: The global reach of the WLCG

2.Progress and Results

The WLCG computing environment has actually been is use as the main production environment of the experiments since 2004, supporting the production, reconstruction and analysis of simulated data. During this time a significant programme of testing and data challenges has been implemented. It is through these efforts together with more recent experience using cosmic ray data, that the experiments have been so well prepared to rapidly process and analyse real LHC data in the past few months.

2.1 WLCG Testing Programme

The overall programme of testing and challenges can be seen sketched in Figure 2.

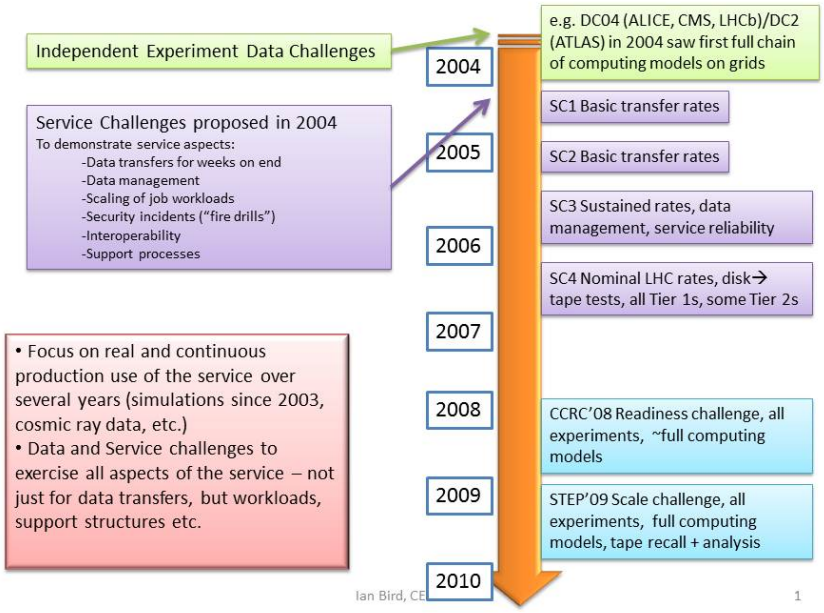

Figure 2: The Programme of Testing and Data/Service Challenges

Even prior to 2004 there were significant efforts to ensure that data could be streamed from the online systems to tape at CERN at the required rates. From 2004 to 2006 the focus was on ensuring the ability to transfer data from CERN to the Tier 1 centres at the required speeds, including copying the data to tape in the Tier 1s. In parallel various aspects of the overall WLCG service were tested during dedicated service challenges, to ensure that the service was manageable and adequate. In 2008 and 2009 there were extended periods of readiness challenges, exercising the full computing models of the experiments, including full tape recall at the Tier 1s, and analysis testing at the Tier 2s. In 2009 the experiments also acquired cosmic ray data and passed this through the full processing chain, permitting full alignment and calibration of the detectors prior to LHC data taking. This exercise was also beneficial in testing aspects of the computing environment. During this entire period, extensive

production of Monte-Carlo was also carried out, so that the service was really working in production with full operational support and management processes being continually refined.

2.2 Evolution of Computing Models

During the 10 years from the proposal of the MONARC model, the experiment computing models all evolved – although all are variations on the same theme. The ATLAS model is closest to the original ideas, with Tier 2s being associated to a specific Tier 1, and obtaining data only from that Tier 1, and sending simulated data to the Tier 1 for archiving. Physics analysis work is sent to the site hosting the relevant data sets. The CMS model is less hierarchical in its implementation, allowing data to move between any Tier 1 and Tier 2 sites. CMS thus had to validate many more end-end links than ATLAS, but data distribution can be changed to adapt to different policies as priorities change. Tier 2s are also formally attached to a given Tier 1 for support and for archiving MC productions, but can request data from any Tier 1 or other Tier 2.

The LHCb model is somewhat different in that the Tier 2 sites are only used for producing Monte Carlo, with reconstruction, stripping, and analysis being performed at the Tier 0 and Tier 1 sites. ALICE also has a distinct model, with no real difference between the Tier 1 and Tier 2 sites in terms of what work is performed at each. The Tier 1 sites do nevertheless retain responsibility for data archiving. ALICE allows data movement between any sites, and even permits remote data access when (part of) a data set is not available at a site.

In all cases a second copy of the raw data is distributed between the Tier 1 sites of the experiment, providing insurance against loss of a single copy.

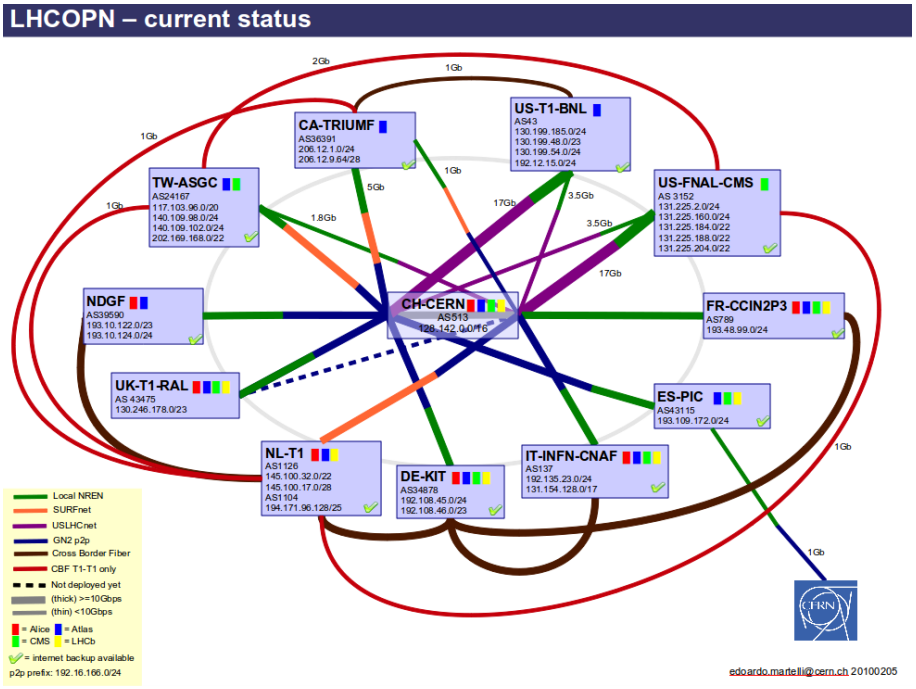

Figure 3: The LHC Optical Provate Network - OPN

It was not clear 10 years ago that wide-area network capacity would be sufficient for the needs, or that the network would be reliable enough. This resulted in the current models, with many services replicated due to the worry that network connections would not be reliable. In fact during the time between the original ideas and the start of the LHC, the network capabilities have exceeded all expectations both in terms of capacity, and in reliability. Today we have also secondary connections between all Tier 1s and the Tier 0. This is shown in Figure 3.

2.3 Progress in Data Transfer and Distribution

One of the important sucesses of WLCG is the ability to manage and sustain data transfers globally at significant data rates, in excess of those planned for and anticipated. Figures 4 and 5 illustrate this – first showing the success during the STEP'09 challenge, and then data export rates from the Tier 0 to Tier 1s with real LHC data. The original targets were to be able to support rates from CERN of 1.3 GB/s. It can be seen that in reality this is often exceeded, and at times rates of 70 Gbits/s have been observed over the OPN, corresponding to the start of ATLAS reprocessing campaigns. While these rates are significant, it is important to realise that they caused no problems and the OPN worked as designed. The secondary routes are vital in providing reliability, as physical links have been cut several times.

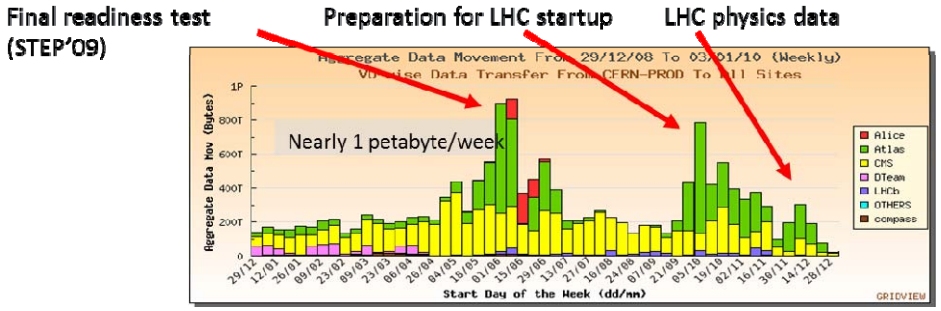

Figure 4: Data transfers during STEP'09 and later

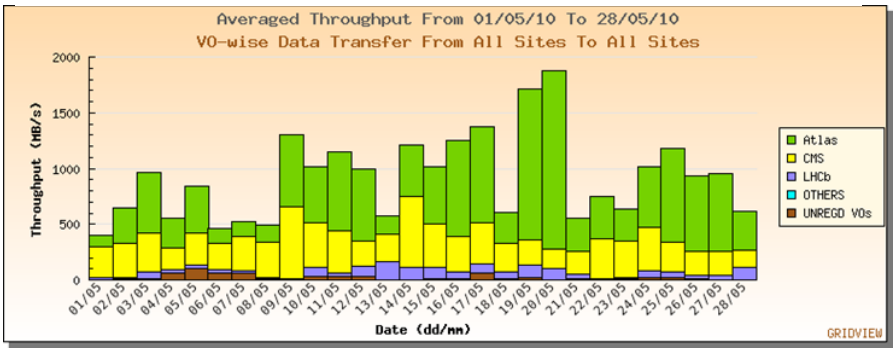

Figure 5: Data export during the first weeks of LHC running

While Tier 0 to Tier 1 transfers rely on the LHCOPN, those between other sites make use of the academic and research networks. Again significant transfer rates globally have been demonstrated. ATLAS for example, report being able to reach peaks of 10 GB/s global transfer

rates. In these early weeks of LHC data taking all experiments have been able to deliver data to physicists for analysis within hours of data taking. This is a significant achievement.

2.4 Workflow and Job Management

At the time of writing the Technical Design Reports for LHC computing, the workloads anticipated by the experiments in nominal data taking years was of order 1 million jobs per day. As can be seen in Figure 6, this has been achieved already, following a steady and coninuous ramp up over the past few years.

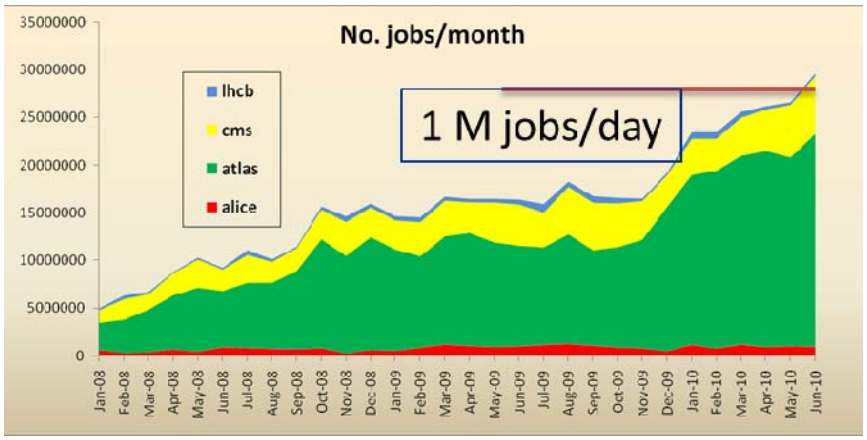

Figure 6: No. jobs/month across WLCG

CPU time delivered is also in excess of 100,000 cores continuous use (with much higher peaks) – shown in Figure 7. This is also a key figure as the anticipated need for CPU was of order 100,000 processors.

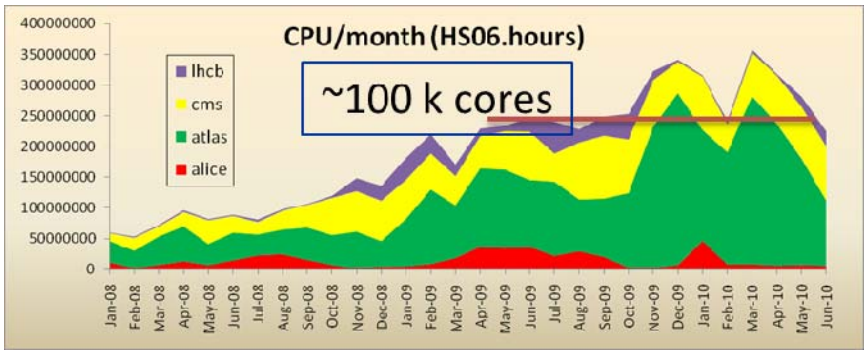

Figure 7: CPU time delivered/month across WLCG

What is equally important to illustrate is that the provision of this CPU is really distributed across the full infrastructure, with the Tier 2s delivering well in excess of 50% of the total. Figure 8a) shows the distribution of CPU time in May 2010 between CERN, the Tier 1s, and the Tier 2s, while Figure 8b) shows the partition of the Tier 2 CPU time by country. This distribution is close to the capacities pledged by the various countries. It is important to understand that the Tier 2s are being used, and that it is not necessary to do analysis at CERN as

some had feared. It is also clear that the WLCG allows the experiments to really make use of all resources provided to them, no matter where in the world those resources are located.

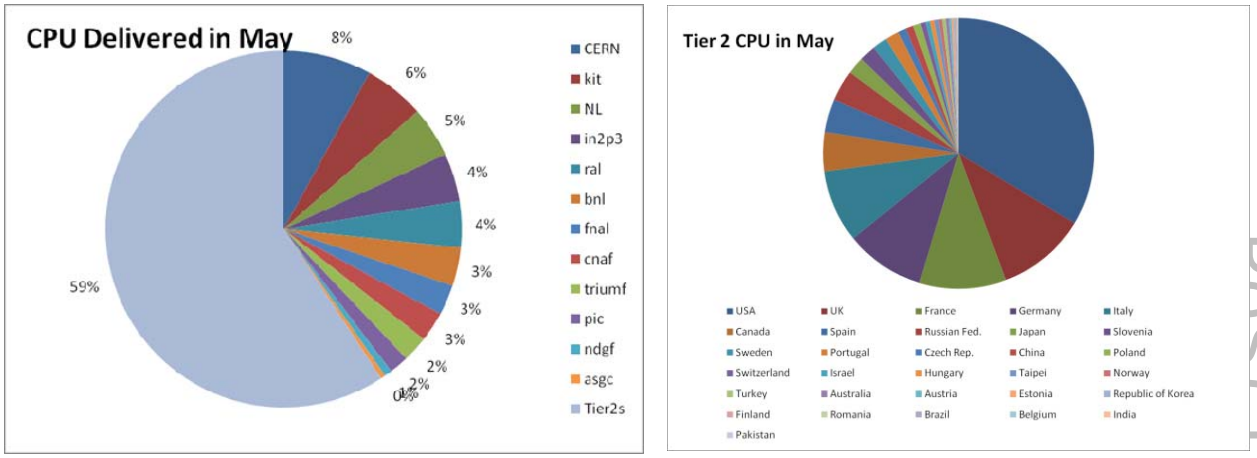

Figure 8: CPU delivered May 2010: a) Tier0, Tier1s, Tier 2; b) between Tier 2 countries

2.5 Other Metrics

Even in the early weeks of data taking as reported here, the number of individuals within each experiment using the system to do analysis are significant. ATLAS and CMS both report around 500 different users performing analysis in a period of 1 month, with LHCb and ALICE reporting somewhat less (~250 each), but consistent with the relative sizes of their collaborations. This is far more than reported even during the analysis phases of the readiness challenges, and demonstrates both that the experiments have made considerable investments in hiding the complex details of the grid from the users, and that it really is feasible to do analysis successfully in this distributed environment.

From the operational point of view, the effort required to maintain the service in reliable operation is still fairly significant, with a certain level of manual intervention and coordination required. However, this is now at a level that is sustainable, whereas even only 1 year earlier, during the STEP'09 challenge it had not been clear that this would be the case. This is clearly an area for development in the future. Some of the problems arise from a level of unreliability of the services and infrastructure at a site (e.g. power or cooling failures), and seems unavoidable at some level. In the future the computing models must not assume the permanent availability of a given site.

3.Anticipated Evolution

The experience in the past 6 years or so, together with the first months of real LHC data have provided some lessons to be learned, and some directions for the future. There are several different aspects of the overall WLCG service and its technical implementation where work for the future is already going on.

Today we have demonstrated that we have an infrastructure capable of supporting LHC data processing and analysis, and there are significant operational structures behind it. Not least of these is a world-wide authentication/authorization mechanism which has been the enabler of these large collaborations being able to run anywhere in the world, coordinated security policies and operational response, and the set of operational procedures, monitoring tools, alarms and reporting activities that the infrastructure relies on. However, it is essential that the technical implementation of the grid (i.e the middleware) can be evolved without losing these key aspects. For the future it is clear that we must adapt to changing technologies, which will require a re-think of storage and data access methods, enabling the use of multi-core CPU and other novel processor types, and making use of virtualisation where appropriate. Behind this is the need to move towards a sustainable structure built on mainstream technologies, but with the added values described above. It is now clear that the network is a very reliable service and we must invest in this area and ensure we can make full use of the distributed system that we have.

Work on refining data management has already begun, with the realisation that the strictness of the original MONARC hierarchy is no longer necessary or appropriate. Other points of discussion include the simplification of the use of tape – becoming really a true archive rather than an active medium as now, with the disk resouces used as a more dynamic cache with data distribution for analysis driven by need rather than pre-defined placement. The models should now recognise that not all data has to be local to a site – remote access to data should be permitted. It is notable to realise that it is often much faster to fetch data remotely than to access local tape. The goal of these activities is to build a more reliable and robust data access system, recognising the limitations and possibilities inherent in a distributed system.

4. Summary and Conclusions

The experiments have demonstrated truly distributed computing models, although today there is a lot of interaction and support needed from sites. The network traffic is far in excess of what was anticipated but is supportable. Planning for the future of the network is under way together with the networking providers. This system has enabled physics output in a remarkably short time, with a large number of people doing analysis at Tier 2 centres worldwide.

Experience with real LHC data and real users highlights several areas for improvement, and the technical implementation of the WLCG system will evolve within the strong collaborative and management framework.

References

[1] *Proposal for Building the LHC Computing Environment at CERN*, CERN/2369/Rev., 2001, http://cern.ch/LCG/PEB/Documents/c-e-2379Rev.final.doc
[2] M. Aderholz et. al., *Models of Networked Analysis at Regional Centres*; MONARC Phase 2 Report, CERN-LCB-2000-001, March 2000
[3] I. Foster & C. Kesselman, *The Grid: Blueprint for a New Computing Infrastructure*; Elsevier 2004.
[4] *Memorandum of Understanding*, CERN-C-RRB2005-01, http://lcg.web.cern.ch/LCG/mou.htm.