

Experience in Running LCG as a Production Grid Service

Ian G. Bird

CERN/ LCG and EGEE

Abstract

By summer 2004 the LCG service will have been in real production use for some 6 months during which period the LHC experiment data challenges will have been running. This paper reports on the experiences and lessons learned. We describe what services have been set up and what policies put in place to enable this, how the production has been supported, and finally we discuss how the service is expected to evolve over the next year. During this period, the LCG infrastructure has formed the basis of the EGEE production service. The paper will report on that experience and how the service has been expanded to include other application domains and what might be needed as further development in this direction.

1. Introduction

The LHC Computing Grid (LCG) project was set up to prototype the computing environment for the LHC experiments. The project is in two phases; the first from 2002-2005 devoted to prototyping the computing environment, developing common applications, culminating in a Technical Design Report for the computing environment based on the experience gained in that phase. The second phase of the project from 2006 to 2008 then building the computing environment and ramping up to be ready for data taking commencing in 2007.

The project is organized in four areas: Applications – building common application software infrastructures; Fabric – building the Tier 0 infrastructure; Middleware – providing re-engineered middleware; and Deployment – deploying and operating the grid infrastructure. The latter two areas are now areas joint with the EGEE project.

An important goal of the Deployment activity in phase 1 of the project was to put in place a basic production environment to support the experiments' Data Challenges during 2004. In this paper we report on some of the early experiences in doing that.

1.1 LCG Deployment Organisation

The organisation of the Deployment Area of the project is illustrated in Figure 1. The Deployment Area manager has a team based at CERN providing three essential functions: Certification and testing of the middleware, coordination and support of the deployment and operations, and support for integrating the experiments' software with the grid middleware. The manager is advised by the Grid Deployment Board, under aegis of which

collaborative projects and activities are set up. The activities of the Deployment Area respond to functional requirements set by the applications, but also to operational, security, and management requirements set by the regional computer centres. Of course, often these two sets of requirements are mutually exclusive!

In addition to coordination across all the computer centres involved in LCG, the Deployment Area also works closely with other grid development and deployment projects to seek common solutions, and to address issues of interoperability. Other organisations and forums are used to help coordinate and bring expertise into the discussion. These include Global Grid Forum, and other activities such as the HEPiX meetings where High Energy Physics computer centre managers discuss issues of common interest.

1.2 Certification and Testing

The certification and testing activity has been one of the most important mechanisms for preparing a consistent set of robust middleware for use in the 2004 data challenges.

The process is depicted in Figure 2. Middleware components received from the developers are subject to a series of functional tests, including a matrix of performance testing to be used as regression tests to verify that subsequent updates and releases do not degrade the quality and performance of the existing release. As part of this process site certification suites have been also developed. These are still quite basic but are successfully used to demonstrate correct installation of the middleware distribution at remote centres. In the past year, a significant activity during certification has been integrating components coming from different sources – particularly

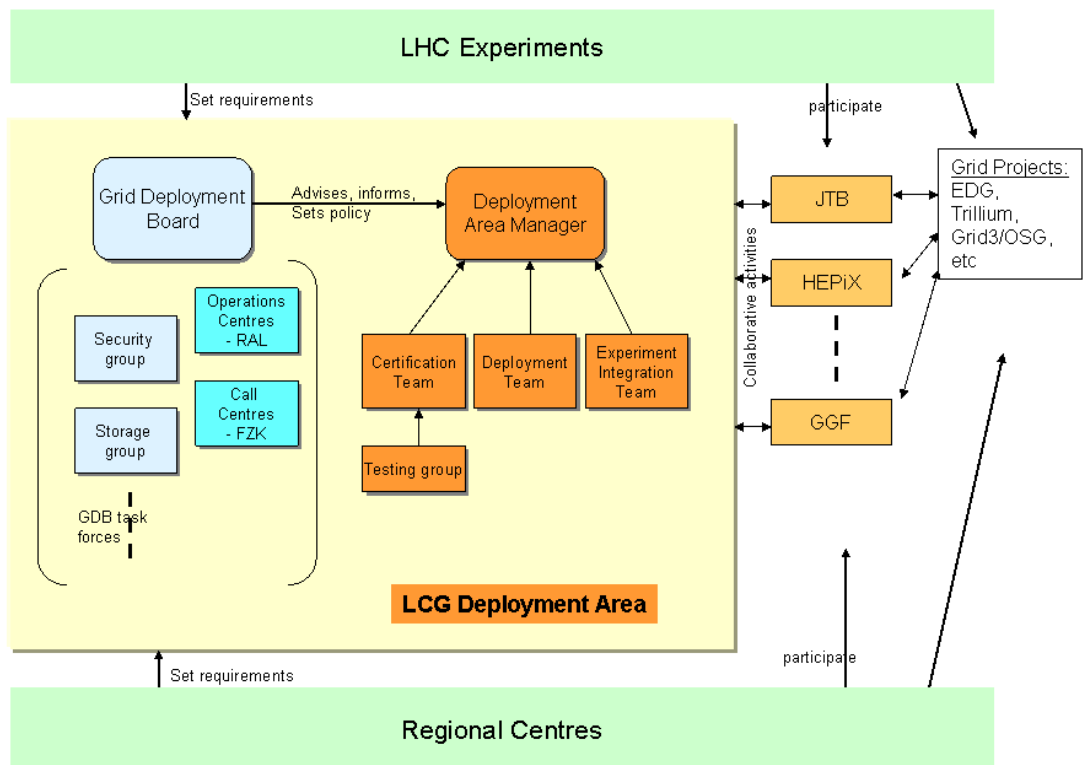


Figure 1: Deployment Organisation

those provided by VDT, and those from the European DataGrid projects. The resources used in the certification testbed are quite significant (~60 machines) as most of the deployment scenarios – different batch systems, site configurations, etc. are tested.

Once the candidate release has been certified it is passed to the applications integration groups and the deployment groups for application testing and deployment preparation. A small testbed is operated for the applications groups to test basic integration of their software with the middleware and to provide feedback before the release is widely deployed. Initially this work was done by the applications on the certification testbed but this was found to be unworkable and was split out.

The preparation for deployment includes verifying the installation process, providing the installation instructions for both automated and manual installation, and providing the release notes and documentation to accompany the release.

At each stage the problems may need resolving by developers and the process is iterated until the targets set for the release are

achieved. At this point the release can be deployed to the collaborating computer centres.

1.3 Deployment

The deployment team at CERN is responsible for coordinating and supporting the installation of the middleware release. The support model includes direct support for the Tier 1 sites which are typically large computer centres. These in turn provide support for the Tier 2 centres in their geographical region. This model has worked well in regions with large numbers of Tier 2 sites: UK, Italy, Spain, where the Tier 2 installations have been done entirely by the Tier 1 sites with little CERN involvement. Other sites however, have required direct support from the CERN team, and with more than 60 sites now involved in the LCG-2 service this is reaching the limit of manageability.

Operational support problems are intended to be addressed primarily by the Grid Operations Centres, with the CERN team providing second-level support. Lack of available staff in the prototype GOCs at RAL and Taipei has meant that the CERN team has had to deal directly with a high load of operational problems.

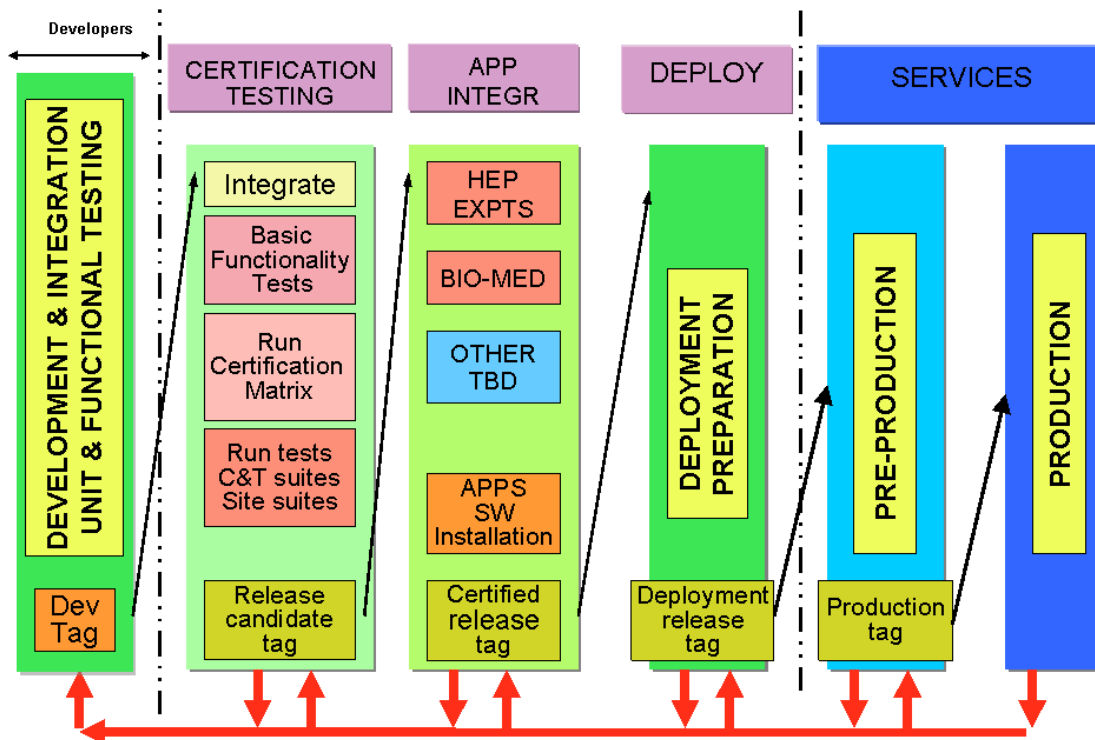


Figure 2: Certification and Deployment

Both of these issues will be helped with the migration of the support for the infrastructure itself to EGEE, which has a clear hierarchical model for operations and user support (see 3. below).

1.4 Grid Deployment Board

The Grid Deployment Board (GDB) was set up to address policy issues requiring agreement and negotiation between the regional computer centres. The members include representatives from countries providing computing resources, from the experiments, and the project management.

The GDB has set up a number of technical bodies to advise it, some being short lived, and some, like the security group being a standing group.

The GDB meets monthly to coordinate the activities in the regional centres. One of its first achievements was to get agreement from all countries on the requirement that there would be a single user registration for the entire LCG, avoiding the need for users to register individually at many sites. This included addressing the issues of gathering personal information which would have led to many legal problems. All sites agreed that collecting contact information, together with the means to contact the users' employers in case of misuse of the resources was sufficient.

1.5 Security Group and Operational Policies

The LCG security group has been very active and effective, and has drafted and overseen the implementation of the essential security and usage policies for LCG. These have been ratified by the GDB, and include usage rules, overall security policy, requirements and agreements on auditing, accountability, and incident response. The LCG usage rules have been proposed as general grid usage rules by several of the other grid projects associated with LCG.

From the beginning the security group has included members from other large sites outside of LCG since many of the LCG sites support user communities that also use those other resources. As the EGEE project has begun, the security group has become a joint security group for LCG and EGEE, with the U.S. Open Science Grid consortium also interested in participating. This will help in bringing the various grid infrastructures closer at least at the policy level.

1.6 Grid Operations Centre

The Rutherford Appleton Laboratory agreed in the GDB to take responsibility for building a first prototype Grid Operations Centre (GOC) for LCG. At the time of writing, a second

prototype based on the work done at RAL has been set up in Taipei.

The LCG GOCs are intended to provide grid infrastructure monitoring, troubleshooting and operational support. The current implementations include regular monitoring of essential services, verification that sites are up and responding, that basic test jobs can run. They are also responsible for ongoing site validation by running a set of validation tests. In addition, they check that the certificates associated with the essential services are not close to expiration.

1.7 Grid User Support

Forschungszentrum Karlsruhe (FZK) agreed in the GDB to take responsibility to set up and operate a user support call centre for LCG. This they have done, providing a portal where all LCG users are able to report all problems with using the infrastructure. In the GDB it was clearly expressed by the experiments that they wanted to provide the first-level triage of problems, before the operations and support teams in LCG should respond.

In practice the experiments and the LCG teams have used combinations of dedicated mailing lists and problem tracking systems to manage this process, while waiting for the FZK portal to be available. It is only now that migration to this more central service is possible and is just beginning. The intent is to ensure that all problems are assigned and addressed and that none “slip through the cracks”. FZK support group will provide the staff to follow up each problem report with the teams assigned to deal with them.

2. Data Challenge Experiences

The LHC experiment data challenges were scheduled to begin in January 2004. However, in all cases the schedules have slipped, and most were at least 1 month delayed. We now have some significant experience with ALICE, CMS, and LHCb, but ATLAS have only just started and as yet it is too early to report on that experience. However, even in the preparations for the challenges for all experiments it became clear that functional issues had to be addressed rapidly.

2.1 Core Sites

Experience with LCG-1 in late 2003 showed a number of potential problems. The most serious and frustrating from the users’ point of view was the lack of stability of the services at many (smaller) sites, and the wrong configurations at

many sites. These problems were often manifest by a lack of sufficient support and response at those sites. This seems to be due to a lack of understanding of what amount of effort it really takes at the moment to run production services with unstable software. Many smaller sites were simply unused to running services at all for large applications.

For the data challenges, it was decided to focus effort on a set of Core Sites that provided significant resources to attract the experiments to commit effort to using them, and that could themselves commit sufficient effort to really support the service in the way that was needed. The LCG-2 service thus started with 8 core sites – which unsurprisingly were mostly the main large Tier 1 sites. It was agreed that a parallel information system would be set up to include all the other sites, and that as they were certified by the experiments as stable they could be moved into the production core sites.

The LCG-2 service has now been successfully deployed to more than 60 sites in 22 countries around the world (Figure 3), including USA, Canada, and Asia-Pacific.

The data challenges of the various experiments were all somewhat different and exercised different aspects of the system. Many issues have been exposed ranging from deficient functionality, to configuration problems, to wrong underlying architectural and model assumptions. This is all invaluable information upon which to base future short and long term developments.

2.2 Data Challenges and Experiences

The sections below summarize these experiences, and are presented in the chronological order of the challenges. Obviously, in all of these activities there were very many operational issues that occurred but are not reported here. In addition, the LCG-2 service has been very stable and reliable overall, especially in comparison with previous experience in use of Globus and EDG systems. However, a number of significant problems were revealed as described below.

The ALICE challenge

The ALICE data challenge is in three phases; the first which is where the only real experience has been so far started in February and finished in May, and consisted of event simulation. Simulated events were generated at remote sites and files sent back to CERN for storage. Subsequent phases will distribute the events out to remote sites for reconstruction,

and then perform user-level analysis on reconstructed events.

The most significant issues reported by ALICE were:

- Mis-match of sophisticated batch system capabilities with grid abilities to publish and make use of those capabilities. One particular issue was the amount of free resources advertised by a site does not simply indicate how much a particular VO can use. This can be addressed in part by providing a CE per VO.
- Lack of sufficient disk storage at remote sites. This reflects a general problem that many smaller sites have not planned sufficient resources to support the requirements.
- ALICE generated a large number of very small files. Whilst not a grid-specific problem this brings out an issue of the mismatch between convenience in processing time and file size and the (in-)ability of mass storage systems to handle many small files.

The CMS challenge

The CMS data challenge in April and May was focussed on the distribution of data from the Tier 0 centre at CERN to the CMS Tier 1 centres. This involved the use of the data management systems, including the Replica Location Service (RLS), Replica Management (RM) tools, but did not extensively test job submission. They plan an ongoing analysis activity that will use job submission during the rest of 2004.

The major problems found by CMS were:

- Bad performance of the RLS. Many issues were addressed and resolved during the data challenge, but the activity brought to light many different problems. Partly this was due to the way in which CMS used the service, but the fundamental problem was that the design of the RLS did not really respond to the basic requirements. It may be that it is only now that those requirements are clear, but there is much work to be done to produce an adequate file catalogue system.
- The lack of consistent grid-storage interfaces meant that CMS was required to build a data transfer system that understood the different interfaces. LCG is focussing on SRM as a common interface for storage systems, but this is certainly not deployed at all Tier 1 centres yet, and there is no existing general disk pool management system with an SRM interface that can be

widely deployed. This issue is being addressed by various development efforts.

- The data transport layer itself was not reliable. Again CMS developed a layer to provide that reliability. Again this is something that LCG is addressing in the remainder of this year.

The LHCb challenge

LHCb started their data challenge in May and it is still in progress at the time of writing. They have been using the job submission system extensively and have particularly stressed it by running much longer jobs (48-72 hours) than earlier experiences. The particular issues that have been found are:

- Lack of normalisation of published cpu power, batch queue lengths, etc. This is a real problem especially for long jobs that exceed many queue lengths if it is not clear what the normalisation factor is. This is being addressed by agreeing a normalisation scheme. In addition many sites did not provide long enough batch queues.
- The Globus gatekeeper model assumes that clusters are homogenous and that published capacities apply to the entire cluster. For many large sites this is not the case. A solution would be to have a CE for each set of different machines, but combined with the problem seen by ALICE where a CE for each VO is needed soon makes the problem of matching batch systems to the existing grid interfaces almost unmanageable.
- The job submission system does not provide bulk operations. This is clearly a problem in a production system aimed at batch productions.
- Many jobs apparently cancelled for unknown reasons. This illuminates a problem that the middleware as developed provides very few tools and mechanisms for troubleshooting and monitoring. This must be addressed rapidly.
- LHCb also ran into the problems of lack of sufficient hardware resources – particularly the mismatch of available disk space per cpu.

The ATLAS challenge

At the time of writing the ATLAS data challenge is only just starting. However, in the preparation phase, much work was done at ATLAS' request to provide additional data management tools, and to replace some of the Java-based command-line tools with C-based tools to overcome the serious performance

limitations associated with the existing implementations. ATLAS had many of the same requirements on the RLS and RM tools as discussed by CMS.

2.3 Lessons Learned

It is clear that the expectations of the experiments towards grid technology have been set quite high and that in some cases solutions developed by the experiments themselves outperform grid solutions. However, in the long term the benefits of having a managed service providing the basic infrastructure will be of benefit. At the moment, the technology is still very immature and is at the level of research projects. The task we have is to take that middleware and to demonstrate that we can build a service around it, while re-engineering and developing the essential functionality so that it provides what the applications require. Additionally, it seems that it is only during the Data Challenges themselves, that some of the real requirements and usage patterns on the side of the experiments have become clearer.

For future middleware developments, there has to be a clear requirement from the infrastructure builders that the developers must focus more on a system-level view, and not simply on functionality. It seems also that earlier developments were full of good ideas that were not really in line with what was needed, and that solutions are often more complex than the real problem at hand.

On the other hand, although there has been quite some work on providing use-cases to the middleware projects, these were probably not detailed enough to guide and focus the development. It is very clear that the only path to success here is through an iterative development, deployment, testing cycle, until convergence is reached on requirements and implementations.

From the operational point of view, the stability of the infrastructure relies heavily at the moment on the level of commitment a site is prepared to put into the service. This at the moment means quite significant effort, until the infrastructure and middleware can be made more manageable and robust. This has been a

problem for many of the smaller sites who simply do not have this level of staff available.

It is important that the hierarchical model of support be extended – as is the intention with EGEE. The alternative, a truly distributed support infrastructure, while probably desirable in the long term, at the moment lacks the necessary tools such as interoperable trouble ticketing systems.

Finally, there are clearly many underlying architectural issues to be addressed. The second generation of infrastructure now being developed must provide a stable base platform on which to try different models of resource access, data management and other higher level services.

3. LCG to EGEE

As the infrastructure of LCG-2 will be operated by the EGEE operations group and expanded to include new European sites, and new applications communities brought in by EGEE, the operations model is evolving. This evolution is largely based on ideas and experience already prototyped during LCG.

The hierarchical operational support model of LCG is extended and formalized in EGEE, and is embodied in the Regional Operations Centres which take a front-line role in both operational and user support for geographical regions. The CERN team, coordinating the activity, remains as the second level of support. However, it is now the ROCs that support all of the smaller sites in their regions that were previously defaulting to the CERN team.

The ROCs also provide front-line user support, building and expanding on the central model used in LCG. However, this brings an additional layer of complexity as the need for coordination of support between the ROCs arises.

The Grid Operations Centre prototypes of LCG become encapsulated as part of the responsibilities of the EGEE Core Infrastructure Centres, which also have the responsibility of operating essential grid services which require expertise and infrastructure such as reliable database services and so on.



Figure 3: Sites with LCG-2 deployed

4. Future Work and Service Evolution

During the remainder of 2004, the LCG-2 service will be continuously operated to support the ongoing data challenges and regular batch production work of the experiments, and in the context of EGEE to support the work of other applications. Many of the issues raised during the data challenges will be addressed, at least as prototyping solutions to be improved upon in the EGEE developed middleware. The intention is to work with LCG-2 to validate as far as possible the basic computing model for the LCG Tier0 and Tier 1 interactions.

LCG is proposing a series of “service challenges” scheduled around the experiment and other production work. These service challenges are aimed at ensuring the basic underlying infrastructure can be made performant enough and stable. These include a basic reliable data transport service with performance at a significant fraction of the LHC data rate by the end of 2004. Other challenges include continuously filling the system with jobs to test the limitations of the system, and to work on security response challenges before a real incident really challenges us!

Finally, it is anticipated that the current generation of middleware is replaced by that currently being hardened and re-engineered by

the EGEE and other collaborative activities. This is expected about a year from now, but it is absolutely clear that from an operational point of view we must be extremely conservative and critical of new components until they are demonstrated to satisfy not only the application functional requirements, but also stringent requirements on security, stability, robustness and manageability in all their aspects.