Proceedings of the Third Conference of the EELA Project
R. Gavela, B. Marechal, R. Barbera et al. (Eds.)
CIEMAT 2007

A Worldwide Production Grid Service Built on EGEE and OSG Infrastructures – Lessons Learnt and Long-term Requirements

Jamie Shiers[1], Maria Dimou[1], Patricia Mendez Lorenzo[1]

[1]*CERN, Geneva, Switzerland*
{Jamie.Shiers,Maria.Dimou,Patricia.Mendez.Lorenzo}@cern.ch

Abstract

Using the Grid Infrastructures provided by EGEE, OSG and others, a worldwide production service has been built that provides the computing and storage needs for the 4 main physics collaborations at CERN's Large Hadron Collider (LHC). The large number of users, their geographical distribution and the very high service availability requirements make this experience of Grid usage worth studying for the sake of a solid and scalable future operation. This service must cater for the needs of thousands of physicists in hundreds of institutes in tens of countries. A 24x7 service with availability of up to 99% is required with major service responsibilities at each of some ten "Tier1" and of the order of one hundred "Tier2" sites. Such a service - which has been operating for some 2 years and will be required for at least an additional decade - has required significant manpower and resource investments from all concerned and is considered a major achievement in the field of Grid computing. We describe the main lessons learned in offering a production service across heterogeneous Grids as well as the requirements for long-term operation and sustainability.

1. Introduction

Grids have proven to be an excellent way of federating resources across computer centres of varying sizes into much larger quasi-homogeneous infrastructures. This matches well with the needs of international science, allowing resources at participating institutes to meet the needs of the entire collaboration. This in turn adds value to the individual sites, leading to a positive feedback situation. This is particularly true in the case of High Energy Physics (HEP), where it is essential that physicists at their home institute, at the host laboratory, as well as "in transit", can all participate equally in contributing to the scientific mission of the experiments in which they are involved. Computing in HEP has traditionally been performed in a distributed manner – CERN having a "one third – two thirds" rule, indicating that the host laboratory would only provide 1/3 of the total resources required. This translates surprisingly well into today's Grid world, where the sum of the resources at the various Tiers (Tier0, Tier1, Tier2) is approximately constant. Nevertheless, Grids have the reputation of poor or problematic reliability and scalability and of presenting a very steep learning curve to new users followed by a constant change of usage methods during operation. All of these issues need to be addressed if long-term, sustainable e-infrastructures are to be achieved. We describe below the main lessons learnt from ramping up services for the Large Hadron

Collider (LHC) at CERN from the stage of early Grid Research and Development to full-scale production services. We address not only the operational issues of deploying a robust Grid infrastructure but also the critical area of application support – essential not only for HEP, where the number of Grid users is expected to explode as first data flows from the LHC, but also to attract new user communities to the Grid. The latter is key to obtaining support and funding for long-term e-infrastructures.

This paper concentrates primarily on user and application oriented issues related to production Grid usage. It is intended to be complementary to another paper submitted to this conference – "Robust & Resilient Services – how to design, build and operate them" [1].

2. Who needs a Grid?

Broadly speaking, applications that need to use a Grid can be categorized as follows:

- Provisioned: the application needs significant quasi-dedicated resources for long periods (years, if not decades). Examples of this category include High Energy Physics experiments, astrophysics / astronomy etc;
- Scheduled: large amounts of resources are required for periods much shorter than the above – perhaps days or weeks. However, given the required scale, dedicated provisioning for such short periods is excluded. Examples in this category range from time-critical – such as disaster response – to non-critical, that can be scheduled well in advance;
- Opportunistic: the least demanding area which – by definition – can effectively soak up any availability resources. Much less time critical that the two previous areas.

Given the ostensibly differing needs of these application areas, can a single shared Grid infrastructure meet their requirements? What are the pros and cons of such an approach – in particular, what is the motivation for funding bodies to invest in such a shared infrastructure and why would application communities be willing to use it in this non-exclusive way? The answer to this question is that no individual site or application would ever be able to assemble, on its own, the resources and expertise required to satisfy the needs of any of the above categories. And the funding bodies would never be justified to invest in favour of only one community. The example of disaster response scenario easily demonstrates this rationale. By definition, one cannot know when such an event will take place, even if there might be times of increased risks for certain types of disaster. However, when such an event occurs, time is of the essence and any delay will have a corresponding effect in the response to any such problem. Governments and other organizations involved in disaster response cannot realistically be expected to have the necessary computing infrastructure on hot-standby in case of these hopefully rare events (in contrast to regularly used response teams, such as fire brigades and ambulances). Furthermore, unless the needed infrastructure is regularly exercised, it is likely that it will not be in a usable state on the hopefully rare occasion that it is needed. Further arguments in terms of economy can be found both in the areas of scheduled (non-time-critical) and opportunistic use. Finally, "provisioned" applications offer an excellent long-term load that can be guaranteed to ensure that the Grid infrastructure is permanently in full production status and evolves continuously to remain state-of-the-art and hence competitive.

3. First steps in the Grid

In order to begin to use the Grid, both a user and the Virtual Organisation (VO) to which the user belongs need to be established. Today, this is somewhat akin to obtaining a visa: first, the host country has to be recognized – typically a lengthy and complex process. Once this has been achieved, individual citizens of the newly recognized country may apply for visas for other countries – somewhat quicker, but still non-trivial (in our extended analogy, Australia's Electronic Travel Authority (ETA) System is exemplary in this respect). Although not quite as lengthy in the Grid world, the delays in setting up new VOs and in registering users are a significant impediment to certain types of Grid usage. The current process would be unacceptable – primarily by nature of the time required – to many types of disaster response, which needs – by definition – to be rapid and may well be triggered out of hours or during holidays / weekends. Thus a well-defined and preferably (largely) automated mechanism of VO and user registration is required (provided that the necessary authentication control that is required for security purposes is satisfied.)

Once this has been achieved, the process of porting the application(s) to the Grid can begin. Today, this requires not only in-depth Grid knowledge, but also a good understanding of the computing model of the applications involved. Only once both are sufficiently mastered – typically by involving Grid experts in a close dialogue with those of the application – can porting to the Grid commence. Again, this is an area that needs stream-lining: one cannot simply scale the support team with the number of Grid applications. As it is not reasonable to expect that any 'central' authority can ever master the design and coding internals of all applications that will join the Grid, the only area where we can focus on is simplifying the Grid infrastructure expansion and the Grid middleware evolution and deployment. Similarly, adding new applications is valuable to the Grid community, helping to motivate funding for a longer term, sustainable e-infrastructure. To address these challenges, the Grid 'standardisation bodies', which define middleware design requirements should provide straight-forward hooks to individual applications and make them known to new applications contemplating to join the Grid. This will give a mutual 'black-box interconnection' or 'lego building' nature to the community and will remove the scalability problem.

At the time of writing, Grids are in production use by numerous HEP experiments around the world. However, as we get closer to the start-up of the LHC, one of the main challenges ahead is to extend this usage from a relatively small number of "production users", but to large numbers – possibly hundreds or more – of less specialist physicists, whose primary job is to perform analysis (and write papers). To paraphrase a senior ATLAS physicist – the Grid should not "get in the way", but rather facilitate.

This should not, however, be seen as an impossible task. By analogy, in the early 1990s, the number of people at CERN, or indeed the world, who were able to make content available on the Web was rather small. Of course, this changed very rapidly – so much so that the Web is perhaps more of our daily life than a television or telephone. Had Mosaic and then Netscape and much more not appeared, together with web authoring tools that masked the HTML complexity and had the Apache web not come up automatically with the Unix OS installation, the web take-off wouldn't have some so early and to such numbers. The Grid is missing such tools still today. Similar changes have also taken place regarding network and cluster computing. Again, in the early days of the LEP collider at CERN, these were considered almost "dark arts", but have long since become part of the expected skill set of an

even novice IT professional when they became off-the-shelf machine attributes and/or plug-and-play network components

4. Bringing New Communities to the Grid

The Grid Support team at CERN has developed an effective procedure that has successfully allowed the introduction of a significant number of applications into the Grid environment. This procedure has been based on an in-depth analysis of the requirements of each individual community together with a continuous and detailed follow up of their Grid production activities. The "gridification" procedure is based on the following 3 blocks:

1. The Grid team ensures the availability of a significant amount of stable resources and services. This Grid capacity is targeted at production work, rather than a basic training infrastructure. Based on the experience gained with several applications, the support team at CERN has provided the Grid infrastructure for a generic VO that is managed by Grid members for communities to whom support has been agreed;
2. The support team then provides continuous follow-up of the application – no knowledge of the Grid environment or technologies is assumed at the outset – this should not be a barrier to using the Grid;
3. Finally, a set of user-friendly tools is provided in order to rapidly adapt the appplication to the Grid. These tools must fulfill the requirements of the application in terms of reliability, tracking of job status and fast access to job outputs. All communities have different goals and requirements and the main challenge of the support team is the creation of a standard and general software infrastructure to allow rapid adaption to the Grid. This general infrastructure effectively "shields" the applications from the details of the Grid (the emphasis here is to run applications developed independently from the Grid middleware). On the other hand, it has to be stable enough to require an aceptable level of monitoring and support from the Grid team and also of the members of the user communities. Finally, it must be flexible and general enough to match the requirements of the different productions without requiring major changes to the design of the tool. As general submission, tracking and monitoring tool the support team have chosen the Ganga/DIANE infrastructure as the official tool for all new "gridifications". This infrastructure is adapted to the requirements of each production with a minimum impact in the general tool. It also includes a layer to MonALISA to monitor the status of the jobs at each site and keep information on processing history.

This infrastructure has been able to attract many applications to the Grid environment and these three fundamental blocks will have likely be relevant for future "gridifications" for current or future Grid projects. How to generalise this procedure (mostly in terms of support) to ensure a continuous flow of communities, which will be able to use the Grid infrastructure with a minimum and standard support infrastructure is an issue which will have to be considered by any Grid project. Special emphasis in a stable infrastructure together with training and dissemination sessions will be fundamental to ensure that mentioned continuous flow of new applications and communities.

5. Operations and Support Issues

A tentative list of functions that a future pan-Grid operations 'body' – centralised, federated or otherwise – needs to provide is listed below. This applies whether or not the direction towards national Grid remains, with a possible overall coordination at the European level. It is essential that the various Grid services, operational and support infrastructures are set-up with a view to reliability, robustness and automation. Any issues that require human intervention have a significant and measurable cost associated. Consider, for example, a problem ticket that requires 1 hour to resolve (many require much more – typically of an expert). The number of such problems that can be solved per week – and indeed their cost – can be simply calculated. The number of problems per 'user-week' has to be correspondingly low: if each of 1000 users has a single problem per year that requires one hour to solve, this is already close to saturating a single support person (based on 50 40-hour weeks per year at 50% efficiency for direct problem solving tasks – almost certainly an over-estimate).

- Define and maintain standards for implementing services with a view to robustness (e.g. resilience to short-term glitches, transparency of common upgrades etc.). This includes all aspects from middleware development through service deployment and operation, including hardware set-up, monitoring and alarms, procedures & documentation etc.;
- Define and maintain operational tools for testing the agreed services under realistic conditions and for reporting errors to the sites in question and / or any international / national Grid infrastructures;
- Define and maintain accounting mechanisms with the agreed level of granularity (by VO, by obfuscated user etc.);
- Define and maintain procedures for registering VOs and users within VOs with an acceptable response time (i.e. days or less);
- Define and maintain security and related policies (e.g. 'acceptable use policy', policies regarding operating system, database and middleware security patches);
- Define and maintain service level agreements with the NGIs and other sites providing resources to the overall Grid, if necessary on a per-VO basis;
- Staff and run a 'central' (possibly relaying based on time-zones, not just one) high-level 24x7 operations centre that is complementary to those provided at national level, with eventually some 'catch-all' functionality;
- Implement and maintain the necessary tools and fault-tolerant services / database structures required for the above.

6. A definition of Service

Ian Foster's famous paper "What is the Grid? A Three Point Checklist**"** [2][1] lists 3 criteria that are proposed for determining whether a given system is "a Grid" or not.

1. Computing resources are not administered centrally;
2. Open standards are used;
3. Non-trivial quality of service is achieved.

The Worldwide LHC Computing Grid (WLCG) [3] is a system that uses resources provided by two major production Grids – namely the Enabling Grid for E-SciencE (EGEE) [4] in Europe and elsewhere, and the Open Science Grid (OSG) [5] primarily in the US. By

definition, WLCG satisfies the first two criteria – these two major Grids are clearly separate management domains and at least a workable degree of *de-facto* standards is needed for successful production services to be offered. This paper addresses the third point in this checklist – quoted in full below – describing in detail the lessons learnt from offering world-wide production services across many sites for a number of years.

"… to deliver nontrivial qualities of service. (A Grid allows its constituent resources to be used in a coordinated fashion to deliver various qualities of service, relating for example to response time, throughput, availability, and security, and/or co-allocation of multiple resource types to meet complex user demands, so that the utility of the combined system is significantly greater than that of the sum of its parts.)"

7. The WLCG Experience

The purpose of the WLCG service is to satisfy the data processing and analysis needs of the LHC experiments at CERN. The WLCG service implementation is based on a hierarchical model as a natural evolution of the HEP move to distributed processing.

As a reminder, the main responsibilities of the different tiers of the WLCG computing model are as follows:

• Tier0 (CERN): safe keeping of RAW data (first copy); first pass reconstruction, **distribution of RAW data and reconstruction output (Event Summary Data or ESD) to Tier1**; reprocessing of data during LHC down-times;
• Tier1: safe keeping of a proportional share of RAW and reconstructed data; large scale **reprocessing** and safe keeping of corresponding output; **distribution of data products to Tier2s** and safe keeping of a share of simulated data produced at these Tier2s;
• Tier2: Handling **analysis** requirements and proportional share of **simulated event** production and reconstruction.

Sites that are members of the WLCG collaboration sign a Memorandum of Understanding (MoU) [6] that lists the specific responsibilities for that site, the resources that they will offer each supported virtual organisation (VO), the maximum time for intervening in the case of service degradation or loss, as well as the annual availability that should be provided. However, even in the simplest case, these "services" in fact involve numerous components – that sometimes involve other (WLCG) sites and/or third parties, such as network operations. An example is given below for Tier1 sites:

"Tier1 services must be provided with excellent reliability, a high level of availability and rapid responsiveness to problems, since the LHC Experiments depend on them in these respects.
The following services shall be provided by each of the Tier1 Centres in respect of the LHC Experiments that they serve, according to policies agreed with these Experiments. …:

i. acceptance of an agreed share of raw data from the Tier0 Centre, keeping up with data acquisition;
ii. acceptance of an agreed share of first-pass reconstructed data from the Tier0 Centre;
iii. acceptance of processed and simulated data from other centres of the WLCG;"

8. WLCG Service Challenges

In 2004 the WLCG Service Challenge programme [7][7] was launched, aimed at *"achieving the goal of a production quality world-wide Grid that meets the requirements of the LHC experiments in terms of functionality and scale."* Whilst most widely known for their contribution in ramping up data movement and data management services, an often over-looked but extremely important aspect of this programme was that of delivering full production services. Indeed, whilst the first two challenges focussed on basic infrastructure setup and network tuning, the bar was raised considerably for Service Challenges 3 and 4. Both challenges included not only tests performed using the *dteam* virtual organisation, but more importantly included extensive production use by all four of the major LHC experiments. As such – and for the first time – an attempt was made to identify and deploy all of the needed services at the participating sites. The target date for the deployment of these services at the Tier0 was May 2005. This was a highly ambitious goal – not only was this date well in advance of the delivery of the final "Baseline Services" working group report [8][8], but also a number of the middleware components behind the corresponding services had never previously been deployed in production conditions, nor had they been tested by the experiments, nor integrated into their data processing environments.

Realising that there were two distinct goals to be achieved and understanding the unlikelihood of deploying the new services perfectly the first time, two separate instances of the main new services were deployed in the production environment. These were a so-called *pilot service* – the goal of which was to expose the new service to the experiments in order to allow them to gain experience with it, integrate into their software and to provide early feedback and the standard *production service* – to be used both for *dteam* and – after and urgent fixes or enhancements from the experience with the pilot system – for the experiments' production processing. The requirements on these two instances in terms of stability versus rapid updates were clearly different and this model continues to prove valid for making available new features in the production system today.

Other issues that compounded the task for initial service deployment were the lack of clear understanding of how these services would be used by the experiments – making resource estimation an impossible task – as well as the lack of available hardware resources. As is true for many laboratories, hardware resources at CERN are acquired through competitive tender and are typically over-subscribed. Thus, we had little choice but to deploy the services on the only boxes that were then available – typically batch worker nodes, lacking even dual power supplies and mirrored system disks – deferring the choice of suitable systems with which to target the needed availability and reliability to a later date.

9. WLCG Services – the "a priori" analysis

Starting in August 2005, and based on the service levels implied in the WLCG MoU, an *a priori* analysis of the Tier0 WLCG services was performed. This targeted not only the hardware needs, but also the middleware requirements, operational procedures and all other service aspects involved in setting up robust and reliable services. In addition, the feedback and experience from the early months of Service Challenge 3 called for a significant number of service updates. In order to perform these, a "long shutdown" of several days was scheduled during October 2005. It was well understood that such intervention could not normally be performed on a production service, but this was felt to be the least intrusive method available at that time to perform the numerous pending upgrades – including not only deployment of new middleware releases, but also network reconfiguration, hardware moves and reallocation. Unfortunately, sufficient hardware was still unavailable to redeploy the

services in an optimal manner, and their redeployment continued over a period of many months. This was first done using a regular "intervention slot" – simplifying not only scheduling of such interventions with the experiments but also their production planning. However, it was soon realized that the coupling between the various services – not to mention their impact that in many cases extended way beyond the host site and was often Grid-wide – called for a less intrusive manner of performing such changes.

10. Expecting the (un-)expected

It is a truism to state that anything that can go wrong will do so – this is often referred to as "Murphy's law". Whilst this is even part of popular culture, it is still often ignored – who has not lost one or more files due to human error, hardware failure or even a combination, only to find out (or often to realise, in the case of a personal computer) that no adequate backup exists? However, do we systematically prepare for common failures or problems – let alone less likely scenarios? Experience from previous generations of HEP experiments – such as those at the LEP collider at CERN - remind us that there can be many causes of data loss or corruption, including software failures. Whilst naively some such scenarios would appear to be so unlikely that they can be readily dismissed, experience over more than two decades of running production services suggest that preparation for all eventualities is a much safer strategy. In the early days of attempting to deploy the European DataGrid (EDG) Replica Location Service (RLS) as a file catalogue, it was even claimed that if the release procedure were correctly followed, it would be "impossible" for a bug to appear in the production system. More valuable lessons that were (re-)learnt by a new generation of service providers were the length of time that it takes to deploy a full production service and the amount of detail that is required in the associated planning process. Some concrete examples of events that have taken place that are more or less expected, depending on one's viewpoint, are listed below:

- The Expected:
  - When services / servers don't respond or return an invalid status / message;
  - When users use a new client against an old server;
  - When the air-conditioning / power fails (again & again & again);
  - When 1000 batch jobs start up simultaneously and clobber the system;
  - **A disruptive and urgent security incident… (again, we've forgotten…)**
- The Un-expected:
  - When disks fail and you have to recover from backup – and the tapes have been overwritten;
  - When a 'transparent' intervention results in long-term service instability and (significantly) degraded performance;
  - When a service engineer puts a Coke into a machine to 'warm it up'…
- The Truly Un-expected:
  - When a fishing trawler cuts a trans-Atlantic network cable;
  - When a Tsunami does the equivalent in Asia Pacific;
  - When Oracle returns you someone else's data…
  - When mozzarella is declared a weapon of mass destruction…

11. Generic middleware versus an application oriented approach

Experience in offering production services to a variety of applications shows that there is a clear distinction between "generic middleware" and those middleware components which require detailed knowledge of the computing model of the application area. Even within the WLCG community, the four LHC experiments have significant differences in their computing models which strongly impact on the suitability and usage of existing middleware components. As an example, only two (ATLAS and LHCb) explicitly use a generic file catalog component, but in very different ways (ATLAS as a 'local' file catalog, covering a Tier1-Tier2 cloud (in their nomenclature); LHCb as a global catalog with a R/O replica (eventually) at all of their Tier1 sites (R/W master at the Tier0 – CERN. Another example is the gLite File Transfer Service, where again the computing models, whilst globally similar at the high level, differ sufficiently in detail that specific developments and deployment models had to be foreseen. Thus, we propose two categories of middleware: *generic middleware*, which is neutral to the computing models of at least several applications, and *application-oriented middleware*, which takes the latter into account. This means that the corresponding support must be organized on the same lines and that for successful operation of an application in the Grid environment first detailed analysis of the computing model(s) involved, followed by on-going application-oriented support, must be foreseen. This is a non-negligible cost that must be understood for a generic infrastructure, where Grid experts must work hand-in-hand with application experts. Any simplification or streamlining that can be performed in this area can be expected to have major benefits and should be considered a requirement if the Grid is reach the same level of ubiquity as the Web – even in its state as it was ten plus years ago.

12. Summary and conclusion

From the time when the Grid was just an assembly of articles published as a book, we have come a long way. Specifications, policy statements and wishful thinking have actually become installable and configurable products. They are still not smoothly deployable and expandable to the required extent for the Grid to become a success story.

Our WLCG service experience showed that the way to:

- Achieving consensus on how policy is paved;
- Translating policy statements into products is done;
- Monitoring and testing sites made great progress and proved valuable for knowing what portion of the Grid is actually available for use at any given point;
- Secure registration of VOs and users and tracing back actions to the original author are done or being worked on.

Areas still needing more work are the expansion of VOs, sites, users, applications and the passage to new middleware releases. These are therefore candidate areas for any future centralised (or coordinated) effort to focus on.

All policies, procedures, Service Level Agreements and methods should aim at obtaining concrete and measurable results. These are the only indicators of a healthy service status. They are also useful to anticipate and plan changes of direction in the required strategy.

References

[1] Robust & Resilient Services – how to design, build and operate them – presented to this conference.
[2] I. Foster, Argonne National Laboratory and University of Chicago, What is the Grid? A Three Point Checklist, 2002.
[3] The Worldwide LHC Computing Grid (WLCG), http://lcg.web.cern.ch/LCG/.
[4] The Enabling Grids for E-sciencE (EGEE) project, http://public.euegee.org/.
[5] The Open Science Grid, http://www.opensciencegrid.org/.
[6] Memorandum of Understanding for Collaboration in the Deployment and Exploitation of the Worldwide LHC Computing Grid, available at http://lcg.web.cern.ch/LCG/C-RRB/MoU/WLCGMoU.pdf.
[7] The LCG Service Challenges – focus on SC3 rerun, Jamie Shiers, in the proceedings of the International Conference on Computing in High Energy Physics, Mumbai, India, February 2005.
[8] LCG Baseline Services Group Report – available at http://lcg.web.cern.ch/LCG/peb/bs/BSReport-v1.0.pdf.
[9] The European Grid Initiative Design Study – website at http://www.eu-egi.org/.