WLCG Operations and the First Prolonged LHC Run

M Girone, J Shiers, CERN

E-mail: Maria.Girone@cern.ch, Jamie.Shiers@cern.ch

Abstract. By the time of CHEP 2010 we had accumulated just over 6 months' experience with proton-proton data taking, production and analysis at the LHC. This paper addresses the issues seen from the point of view of the WLCG Service. In particular, it answers the following questions: Did the WLCG service delivered quantitatively and qualitatively? Were the "key performance indicators" a reliable and accurate measure of the service quality? Were the inevitable service issues been resolved in a sufficiently rapid fashion? What are the key areas of improvement required not only for long-term sustainable operations, but also to embrace new technologies. It concludes with a summary of our readiness for data taking in the light of real experience.

1. Introduction

The first prolonged data-taking period of the Large Hadron Collider (LHC) at CERN started in March 2010 and by the time of its completion the machine had delivered above $40pb^{-1}$ of integrated luminosity. The four main LHC experiments, ALICE, ATLAS, CMS and LHCb, have taken data from proton-proton collisions at 7 TeV centre of mass energy (and – with the exception of LHCb, from Pb Pb collisions at an energy of 2.76 TeV per nucleon pair) with overall data taking efficiency above 90% that has been processed and analysed using the Worldwide LHC Computing Grid (WLCG). This paper focuses on the WLCG operations and service that are required to support the experiments in their activities. Using both quantitative and qualitative metrics, it shows how well the service delivered and which are the areas where improvements are required.

2. Quantitative Metrics

The WLCG service can be measured in terms of what it delivered: the amount of CPU, the number of jobs, the data transfer rates supported, the number of unique analysis users (figures 1 & 2) and so forth. These are notable both on the absolute scale as well as with respect to the foreseen needs – which were met and in some cases exceeded – with 1 million jobs per day / 100k CPU-days per day being delivered (see figures 3 & 4), several hundred different analysis users per month and multi-GB data export rates. ATLAS, for example, regularly exceeded the nominal 2GB/s rate for data export with peaks of 4GB/s out of CERN and 10GB/s grid-wide. It has been widely acknowledged that the WLCG service allowed the experiments to achieve impressive and timely results – further details of which are given in the experiments' experience papers, presented in the same track as this paper [1 – 4].

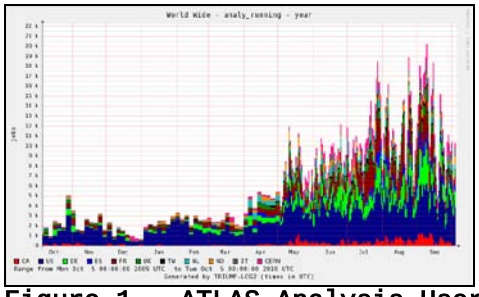

Figure 1 - ATLAS Analysis Users

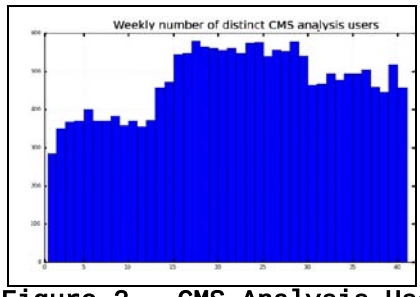

Figure 2 - CMS Analysis Users

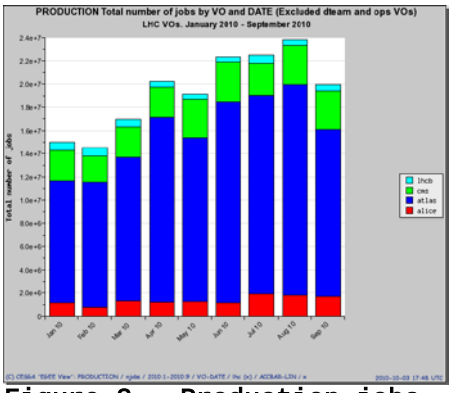

Figure 3 - Production jobs

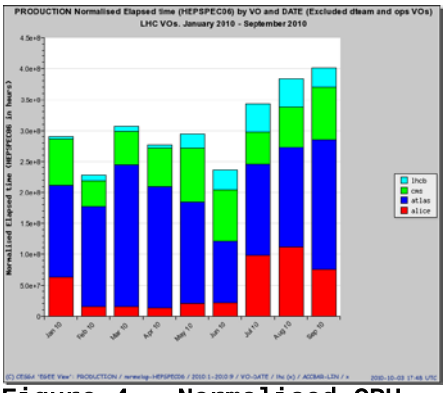

Figure 4 - Normalised CPU

3. Qualitative Metrics

We now present an analysis of the Key Performance Indicators (KPIs) that are used at the regular reports to the WLCG Management Board (MB)[1] and in the WLCG Quarterly Reports to the Overview Board (OB), based on the full data-taking period of 2010.

The KPIs are as follows:

- The Site Availability as measured using Experiment-specific tests;
- A summary of the GGUS tickets, including drill-down on any alarms;
- A summary of any Service Incident Reports (SIRs), triggered whenever one of the service targets defined in the WLCG Memorandum of Understanding (MoU) [5] is not met.

On a short-term basis, these KPIs give a reasonable good overview of the service. However, particularly in the case of the SIRs, the number of incidents in any given week is too small to compare against the targets shown in tables 1-3 below. In this paper we analyse all SIRs produced since the beginning of 2009 – a total of around 100 incidents.

3.1. Site Availability

The following plots show the site availability for the Tier0 and all ATLAS Tier1 sites using ATLAS-specific tests (figure 5) and CMS (figure 6). In figure 5 we can clearly see the prolonged downtime of NL-T1 (NIKHEF and SARA) due to problems with their database services in August / September, (but not in figure 6 as NL-T1 is not a CMS site) plus a number of shorter downtimes shortly after due to sites responding to a security alert. Although a number of "false negatives" still occur – typically a test failure or timeout rather than a real service issue – these plots give a relatively good high-level overview of the service. The analysis of these plots that is performed on a weekly basis is helping to reduce the number of such false negatives, or at least clearly identify in the MB reports which are

[1] Access to the agendas and minutes for these meetings, together with the daily operations call and fortnightly planning meeting, can be found via the WLCG operations Twiki at https://twiki.cern.ch/twiki/bin/view/LCG/WLCGOperationsWeb.

problems due to sites and which are those due to the tests themselves. Globally speaking, it is clear that the site availability is in general above 90%. Individual incidents are explained on a case-by-case basis in the regular reports to the MB that confirms that the service is well under control.

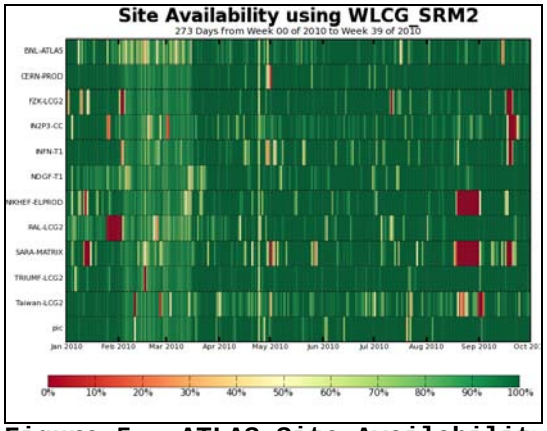

Figure 5 - ATLAS Site Availability

Figure 6 - CMS Site Availability

3.2. GGUS Tickets

The number of GGUS tickets has been relatively stable over 2010, with a small rise in the case of ATLAS (see figure 7). There is a clear difference in usage between ATLAS and CMS, with CMS preferring to perform a first analysis and only issue a GGUS ticket once they have ruled out a problem on their side.

In absolute terms, some tens of tickets, with a peak of around 100, are submitted by ATLAS per week. As can be seen, over two thirds of these tickets are team tickets, which are an essential tool for "shifters" as they allow issues not associated with an individual user – e.g. production-related problems – to be followed up by all members of the team. Recently, the possibility to escalate a team ticket to alarm status – which indicates that rapid or immediate action is required, even at nights and weekends – has been implemented.

The percentage of alarm tickets (see figure 8) is very low and this mechanism continues to be an effective way of addressing critical problems – in all cases the expert response is well within targets, although not all incidents are solved in a sufficiently timely manner. This is discussed in more detail in section 3.3 below through the analysis of the Service Incident Reports, described later, associated with each such incident.

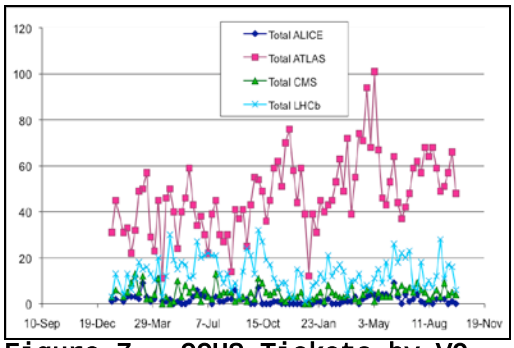

Figure 7 - GGUS Tickets by VO

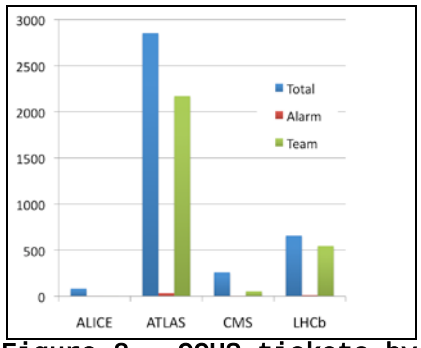

Figure 8 - GGUS tickets by attribute

3.3. Service Incident Reports

Service incident reports were introduced at the time of the Common Computing Readiness Challenge (CCRC'08) [6] to track the cause and follow-up on significant service problems – i.e. where there is a prolonged downtime or degradation with respect to the service targets listed in the WLCG MoU. In

addition to being presented in the regular Operations report to the WLCG MB, follow-up on any new or open reports is performed at the bi-weekly WLCG Service Coordination meetings.

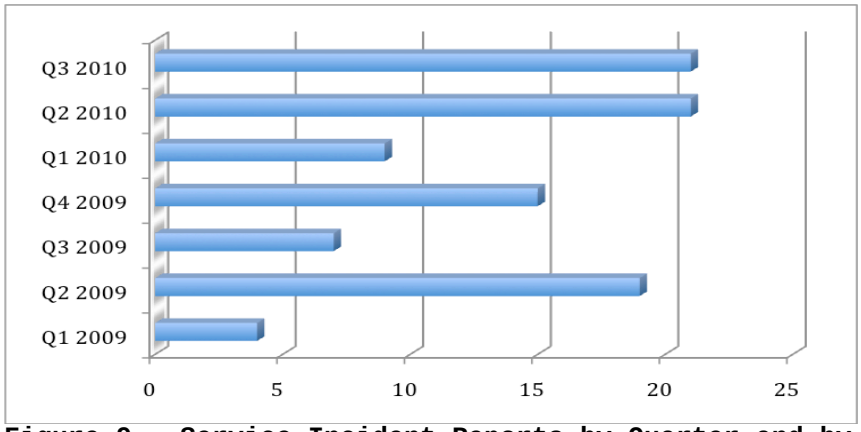

Figure 9 - Service Incident Reports by Quarter and by Area

In figure 9 above a clear correlation between activity and the total number of service incidents can be seen with peaks in 2009 corresponding to STEP'09 (Q2) and the LHC run at injection energy (Q4) and the LHC high energy run as of Q2 2010. As will be described in more detail below, there is a significant variation in severity and duration although all, by definition, exceed the operational targets of the MoU that are described below.
Again in absolute terms, the number of such incidents – peaking just above 20 per quarter and integrated over all Tier1 sites and the Tier0 – does not seem to be too high. At least a fraction of these incidents are probably unavoidable – possibly as much as one third. However, these incidents can be responsible for prolonged site or "cloud" downtimes, resulting in significant disruption to the activities of one or more experiment.

3.4. WLCG Services and Operational Targets

WLCG services can be broken down into the following categories:

1. **Middleware services** – generic services at Grid middleware layer, typically operated by WLCG
2. **Infrastructure services** – fabric-oriented services operated by the sites
3. **Storage services** – at all sites and critical at Tier0 / Tier1s
4. **Database services** – mainly at Tier0 & Tier1s
5. **Network** – connecting the sites (OPN and GPN)

This breakdown will be used in the following analysis to understand if there are issues or trends within these categories.
Also essential for the experiments' operations are:

6. **Experiment services** – developed, maintained and operated by the collaborations themselves (typically run in "VO boxes").

After discussions at the WLCG Overview Board and other bodies such as the Grid Deployment and Management Boards, the following targets for response to and resolution of critical service problems have been set. In particular for the Tier0 and Tier1s, these targets have been used in all Management Board reports since the beginning of 2008. In general, was have seen that the response time to problems is good – typically better than the targets listed.

Time Interval	Critical Tier0 Services (see MoU)	Target
30'	Operator response to alarm / call to x5011	99%
1 hour	Operator response to alarm / call to x5011	100%
4 hours	Expert intervention in response to above	95%

8 hours	Problem resolved	90%
24 hours	Problem resolved	99%

Table 1 - Targets for Incident Response and Resolution for Critical Tier0 Services

Time Interval	Tier1 Services	Target
1 working day	All services – problem resolved	95%

Table 2 - Targets for Tier1 Services

Time Interval	Tier2 Services	Target
1 working day	All services – problem resolved	90%

Table 3 - Targets for Tier2 Services

On the other hand, as shown in figure 10 below, a significant fraction of such problems take longer than 24 hours (as many as one third) – or even 96 hours (up to 20%) – to be resolved. To understand these issues further, we first break them down into the above service categories (figure 11).

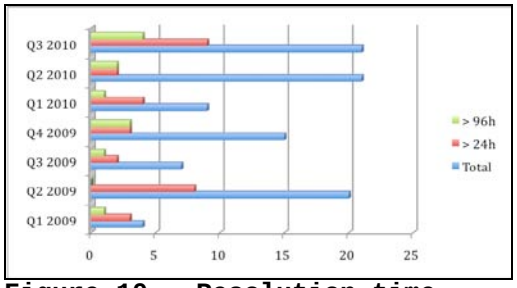

Figure 10 - Resolution time

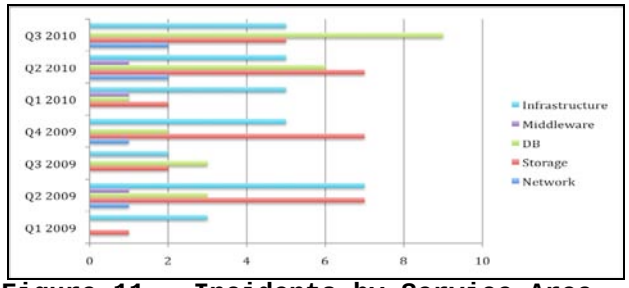

Figure 11 - Incidents by Service Area

Here we see the following:

- **Middleware services**: there are very few incidents that result in a SIR, at least partly due to the way that the experiments use these services. Those problems that do exist are typically relatively short-lived instabilities;
- **Infrastructure services**: the number of such problems is rather constant quarter to quarter and at least some (e.g. power and cooling) are probably unavoidable. As has been seen on several occasions a power cut of very short duration – even 1 second or less – can cause many hours or even days of downtime;
- **Network services**: typically degradations but which can take a long time to be completely resolved. A new WLCG procedure, described below, has recently been agreed for handling such problems;
- **Storage and Database services**: typically complex problems that sometimes cannot be resolved within one or more days and hence the area where improvements can be expected to yield the largest benefits.

4. Areas of Improvement

4.1. Network

In order to improve the resolution of network problems, a new procedure has been established following CHEP. It is applicable to all types of problem, from degradation to total outage (e.g. fibre cuts) and is independent of the network type (general purpose network or LHCOPN).

To illustrate this new procedure, we consider the case when one or more VO observes high failure rates and / or low performance in transfers between two sites: A & B. After some basic debugging, for example to rule out problem related to the source or target storage element or the transfer software

itself (middleware or experiment-specific), the issue is declared to be a network issue and raised at the WLCG operations meeting. (The procedure is equally valid for fibre cuts or other problems, but these are typically much easier to debug and resolve).
It clarifies that the site representatives are responsible for interacting with their local network experts, who in turn are responsible for testing the connection between the two sites and ensuring any problem is resolved. This includes all interactions with the network service providers concerned. The site representatives should update the corresponding ticket on all changes of state, as well as at regular intervals (daily for urgent problems).
It is too early yet to see whether this procedure has had a measurable impact on such problems but will be included in future analyses of WLCG Service Incidents.

4.2. Storage and Databases

The service areas that continue to have a relatively high number of problems that can take days or more to resolve are those of storage and/or databases (storage services often involve databases for one or more component). In some cases, such as that of the ATLAS "cloud" file catalogs, a disruption to the corresponding service can have an impact on many sites and on a significant fraction of the experiments' resources. Whist this is an area that is still being discussed, both ATLAS and LHCb are reviewing the way that they handle detector conditions – currently accessed by running jobs from a local Oracle instance at the Tier0 and Tier1s – as well as ATLAS' file catalog strategy. It is not unlikely that more reliance will be placed on caching techniques, consistent with the direction being pursued also for event data. Again, it will be important to measure quantitatively whether such changes result in fewer and/or shorter service incidents.

5. Conclusions

In summary, the WLCG service was ready for the LHC restart and delivered both quantitatively – up to and beyond the foreseen scale – as well as qualitatively. In the latter respect, expert response has always been within targets and many problems have been solved within 8 hours – some problems unavoidably take longer. Whilst steps will be taken to improve the fraction resolved within the target time interval, pragmatic approaches to prolonged service and/or site downtime will be required in the foreseeable future. Strategies for this have been proposed by CMS and WLCG operations can offer assistance in coordinating between the different VOs affected, the sites and WLCG management. In absolute terms, halving the number of incidents that result in a Service Incident Report is probably an over ambitious target. Given the increased load expected in the 2011 / 2012 run, maintaining the service at the 2010-level of stability is probably a more realistic goal, whilst at the same time decreasing the time for problem resolution, particularly in the case of problems lasting several or even many days.

Acknowledgments

The qualitative plots in this paper were produced by Andrea Sciaba.
The GGUS information was provided by Maria Dimou-Zacharova together with the GGUS team.

References

[1] Ikuo Ueda, ATLAS Operations: Experience and Evolution in the Data Taking Era, in the proceedings of this conference;
[2] Daniele Bonacorsi, Experience with the CMS Computing Model from Commissioning to Collisions, in the proceedings of this conference;
[3] Costin Grigoras, Grid Interoperability and Operations Challenges Under Real Load for the ALICE Experiment, in the proceedings of this conference;
[4] Philippe Charpentier, The LHCb Computing Model in Front of Real Data, in the proceedings of this conference;
[5] WLCG Memorandum of Understanding, available at http://lcg.web.cern.ch/LCG/mou.htm
[6] Common Computing Readiness Challenge (CCRC'08) – post-mortem workshop.