LCG Services Report
April – June 2010

2 July 2010

Jamie Shiers

This report concerns the second three months of 2010, which coincide rather precisely with the restart of the LHC and the first extended data-taking period. It also covers the transition from the EGEE III project to the startup of EGI (end April / beginning May). In this respect it can be considered one of the most critical quarters for the WLCG service to date.

During this quarter there were several major service incidents, particularly affecting the Tier0, which resulted in experiments being unable to record raw data for several hours and – in one particular case – in loss of data (much of which was subsequently recovered). These key incidents are described in more detail below. For Tier1 and Tier2 sites and as reviewed at the WLCG Collaboration workshop held from 7 – 9 July at Imperial College in London, the situation was somewhat more positive – whilst there are specific issues that need to be addressed, the service "basically worked."

Further streamlining of the regular reports to the WLCG Management Board (MB) was achieved based on a single view covering the 3 main Key Performance Indicators (GGUS ticket summary, site usability from the experiments' viewpoint and summary of Service Incident Reports and Change / Risk assessments).

The Site Usability plots – which typically show relatively few and short-lived incidents – correlate well with reports from the daily operations meetings and, together with the drill-down on any significant incident, provide a good overview of the service usability during the period concerned.

The GGUS summaries continue to show significantly different usage strategies between experiments: CMS continuing to prefer to perform an initial debugging of any problem before opening a GGUS ticket. Sites and service providers nevertheless stress the importance of opening tickets in a common system and GGUS continues to be that system. Alarm tickets have been opened relatively regularly: an analysis of all alarm tickets since their introduction for CCRC'08 shows that virtually every single alarm to date has been well justified (only one in 2008 was questionable), whereas in recent weeks there have been a number of occasions when an alarm ticket was warranted (based on problem severity) but a team ticket used (the problem being already under investigation). Of more concern, the end-to-end alarm chain continues to be fragile for a variety of reasons. Despite being tested after each GGUS release, failures continue to be uncovered not only by such tests but also by real usage. Both require further investigation and resolution. There are also concerns that the level of expertise performing the triage of tickets (the ticket processing managers – TPMs) is not sufficient to handle real-life problems and this is being addressed with the EGI project. Finally, there is a growing number of tickets that are not resolved sufficiently rapidly – at least a number of which being complex or multi-site and requiring the expertise of several support teams to address. A regular review of key open tickets (as defined by the experiments) has already been instituted at the WLCG Tier1 Service Coordination meeting and metrics will be established to measure improvement.

A similar situation exists with Service Incident Reports (SIRs) which are sometimes never completed or reviewed: at the WLCG Collaboration workshop in July we reviewed the key SIRs from previous quarter and drilled down on a specific multi-site issue in an attempt to make further improvements in this area. Regular reviews of key SIRs will continue.

Summary of Main Service Incidents

Previous quarterly reports have included a table listing by date, site and service the main incidents for which a Service Incident Report was produced. These are typically characterized by a serious degradation or total loss of service of at least several hours and / or when an alarm ticket was generated.

During this period there were over 20 Service Incident Reports for the CASTOR service at CERN alone – more than one per week and more than all other service reports for all services at all sites. The next most frequent category was that of databases with 6 entries, some of which are composite and should probably be counted as separate incidents.

We therefore separate out these two categories from the remainder (where the number of infrastructure incidents, such as power and cooling, is much lower than in the past), as they clearly require special consideration.

Whilst with hindsight these numbers may not appear particularly alarming, they are higher than both historical averages and our targets for this period.

In addition, it is important to stress that the impact of an incident may be more significant to the experiments that are affected than the strict duration as measured by the service provider – this is measured by the impact of an incident on the on-going experiment activities that is not always easy to quantify.

On the other hand, the targets for response to alarms, start of expert intervention and problem resolution continue to be met. This suggests that separate metrics are required on the number and severity of SIRs to those on problem response and resolution.

The information in the tables below is taken from the SIRs directly – these and indeed some of the details in the reports demonstrate clearly that some interpretation and further analysis is required, particularly if these reports are to have a long-term benefit and/or be of use to sites other than the originator.

Date	Duration	Summary
29 June	4 hours	CASTOR outage due to AFS unavailability
28 June	4 hours	High volume of SRM logs and second DLF overload
22 June	3 hours	LDAP Overloaded
16 June	1 hour	CMS jobmanager daemons stuck on AFS
14 June	2 mins	LSF reconfiguration after node move affected CASTORLHCB
13 June	n/a	two ATLAS "temp" class files lost due to diskserver crash
7 June	3.5 hours	default pool overloaded with disk to disk copies
1 June	6 hours	Jobmanager was not submitting new work
31 May	1.5 hours	LSF reconfiguration after node move affected CASTORPUBLIC
25 May	3 hours	stuck rsyslog affected T0Merge
20 May	6 hours	CMS T0Express caused LSF overload

14 May	2 months	Castor data incorrectly recycled
13 May	3 hours	CMS stress test load
1 May	7 hours	Castor Affected Piquet Call
29 Apr	5 hours	GridFTP checksum errors for large files
22 Apr	4 hours	CMS alarm ticket for long write wait times
21 Apr	2 hours	lxfsrc5706 filesystem error
16 Apr	3 hours	High error rate on SRM
7 Apr	5 hours	Timeout recalling files from t0merge
6 Apr	3 hours	Local SRM BDII stopped working, caused SAM test failures
1 Apr	7 mins	LHCB lhcbhistos service class overload
30 Mar	2 x 1 h?	two short periods of SRM unavailability - all frontend threads stuck
30 Mar	1 hour	T0ATLAS disk servers rebooted

Table 1 – CERN CASTOR-related Service Incident Reports

Comments: failure or degradation of the CERN CASTOR and related services can have a corresponding impact on raw data recording, first pass processing and/or data export *inter alia*.

Site	Date	Duration	Summary
ASGC	29 June	~15 hours	Streams LCRs not applied from central 3D DB for 15 hours
CERN	26 June	1 hour	ATLAS offline DB (ATLR) – 9 Oracle services did not failover properly after a node eviction
CERN, PIC + T1s	24, 25 June	10 hours	LHCb streaming to PIC not working for 10 hours, to other sites not working for 40 minutes
CERN	2 June	?	ATLAS and LHCb online and offline databases – access and QoS compromised
CERN	31 May, 1 June	?	CMS online, LCGR and ATLAS offline databases – services unavailable during patching
CERN	26 May	?	CMS offline database – h/w failure affecting one node

Table 2 – Database Related Service Incident Reports

Comments: online – offline replication is particularly critical, followed by Tier0 services, Tier1 services and inter-site replication (where latency of some hours is normally tolerable.)

Site	Service	Date	Duration	Summary
RAL	SE	30 June	N/A	1083 CMS files lost
CERN	AFS	29 June	5 hours	Complete FC disk array – affected CASTOR and also LHC!

KIT	CMS dCache	22 June	3 hours	Service down
CERN	CREAM CE	7 June	3 hours	Job submission failure
PIC	Power	21 May	19 hours	Whole site out
CERN to ASGC	OPN	12 – 15 May	Days	Reduced bandwidth
CNAF	StoRM	28 & 29 April	9/12 hours	SRM blockage (h/w) followed by MCDISK full and StoRM bug
IN2P3	AFS	26 April	17.5 hours	AFS crashed after server overload. Batch also affected
IN2P3	Batch	24 April	17 hours	Service location service stopped responding blocking most batch system commands
IN2P3	Grid downtime notification	20 April	9 hours & 5 days	Grid downtime notifications impossible after two separate incidents

Table 3 – Other Service Incident Reports

Comments: the above incidents – all arguably “infrastructure” – can be broken down into site, grid and inter-site infrastructure issues. A significant number of GGUS tickets can also be similarly categorized.

Outlook for the remainder of 2010

In past years we have observed significant absences due to vacation at WLCG sites during the summer period, including in the years when LHC operation had been expected. This year – the first summer ever of LHC data taking – will clearly be critical: problem resolution has typically been much longer during the summer months (particularly for complex problems requiring the expertise of multiple people) and this risks to have an impact on production data processing and analysis if site planning has not foreseen this issue adequately.

Summary and Conclusions

Whilst the WLCG service has largely stood up to the challenges of real data taking, processing and analysis, there have been a number of serious service incidents during the past quarter, including both loss of data and of raw data recording capability. There are also a growing number of issues that cannot be readily handled by the daily operations meetings – typically due to their complexity. Concerns have been raised regarding ticket flow / assignment and problem resolution time – these issues are being addressed but are non trivial and will require both time and resources to resolve. Neither are in abundance.