The Worldwide LHC Computing Grid

WLCG Service Review

Munich Tier 2 Workshop, September 2006

Jamie Shiers, Harry Renshall

CERN, IT/Grid Deployment/Service Coordination Team

Worldwide LHC Computing Grid

Distributed Production Environment for Physics data Processing

Contents:

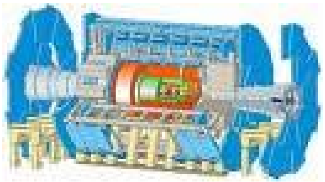
4 main components:

1. High-level experiment-by-experiment review
2. Some Service Issues: problem response / resolution
3. Summary of Service Coordination Observations & Recommendations
4. Next steps/conclusions

SC4 Review

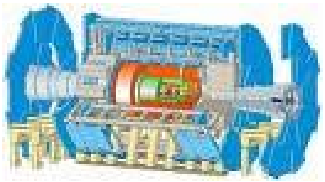
Experiment Review to date

All experiments have ongoing service challenges in 2006



ATLAS Summary (1/3)

- The overall plan for the ATLAS SC4 exercise was to send data out to all ATLAS Tier1 sites at the full nominal rate expected for that site during LHC pp running.
- Whilst these data rates were not achieved for the target of one week, this exercise uncovered a number of problems – many of which have since been resolved – and was clearly an important step towards reaching full nominal rates under realistic conditions.
- Key accomplishments were:
  - Ran a full-scale exercise, from EF, reconstruction farm, T1 export, T2 export with realistic data sizes, complete flow
  - Included all T1s sites in the exercise from first day
  - Included ~ 15 T2s sites on LCG by the end of the second week
  - Maximum export rate (per hour) ~ 700 MB/s (Nominal rate ~ 780 MB/s (with NGDF))
  - ATLAS regional contacts were actively participating in some of the T1/T2 clouds
  - Put in place monitoring system allowing sites to see their rates (disk/tape areas), data assignments, errors in the last hours, per file, dataset, …
  - FTS channels in place between T0 and T1 and now progressing between T1 and T2s
  - Exported a total of 1PB of data by Sunday August 6th
- Problems with VO box load have been identified and resolved, but adequate monitoring of LFC services at Tier1 sites remains an outstanding issue; A range of LFC issues (functionality; usage; deployment; operation) are being addressed
- Major concerns include communication issues with the sites and the serious lack of manpower globally;
- Exercise to be rerun during next 3 weeks – ramp-up from now



ATLAS Summary (2/3)

ASGC	after VO BOX upgrade, went very well. 100 MB/s when ATLAS runs; 40~50 MB/s when CMS runs (should be 60 MB/s); communication problems during start-up of exercise
BNL	not using realistic tape area; suffering from read/write contention when using ‘production’ areas (as opposed to SC4 /dev/null area); very good support for ATLAS
CNAF	unstable Castor-1; now fighting Castor-2 installations. Needs re-evaluation during next phase
LYON	very good service T0->T1 and T1->T2! The only site that was constantly part of the exercise (except for scheduled downtimes).
FZK	after VO BOX upgrade, went better. Still very unstable service (in/out of the exercise all the time)
PIC	stable service; dCache disk area and Castor tape area occasionally suffering some timeouts/overload issues
RAL	not stable; difficult to understand status; could not sustain rate for a few hours. See the LCG Quarterly Report for Q2 2006 for further details of on-going storage issues at RAL.
SARA	very stable service overall
TRIUMF	remains stable; network distance leads to occasional LFC connection glitches

ATLAS
3/3

Tier-1/Tier-2s Data Transfer Functional Test
Preliminary status after 2 days of running

Tier-1	Rqstd/Rcvd Datasets	Success Rate %	Comments
ASGC	3/3	98%	After site problem was fixed (2-3h)
BNL			Site problem : FTS. Fixed, resubscribed
CNAF			Site problem : SE (CASTOR)
GRIDKA	8/8	100%	From the 2nd attempt, initial efficiency 12%
LYON	10/6	20%	FTS errors, after Sep 11 20:00 efficiency ~80%, system was blocked by T2-T1 data transfer, resubscribed
PIC	5/5	25%	FTS errors, resubscribed
RAL	6/6	90%	From the 2nd attempt, initial efficiency 10%, data transfer still in progress (Request : 24h ago)
SARA			After site problem was fixed (2-3h), problems with access to central services and FTS
TRIUMF	5/5	75%	From 1st attempt, FTS errors starting Sep 11 17:00 Data transfer still in progress (Request : 48h ago)

CMS Summary (1/2)

- The main activity during this period was preparation work for CMS CSA06. This involved debugging of data rates into and out of CERN (using PhEDEx over FTS), clarification of FTS channel setup, monitoring and operations and testing of the gLite RB;

- Problems resolved include poor transfers both into and out of CERN (related to the use of the loopback interface for SRM transfers and to incorrect handling at the SRM level of duplicate nameserver entries. Once these problems were resolved, and following tuning at the PhEDEx level, CMS were able to drive transfers at the target rate for CSA06 of 150MB/s (1/4 of the nominal rate);

- Following this successful debugging exercise, an attempt to run at 500MB/s out of CERN for at least 3 days was made. Whilst this target was not reached, the ‘threshold’ of 300MB/s was attained, with a daily average of 450MB/s on 8th August, with ATLAS and other transfers proceeding in parallel.

- In the 3 month period ending mid-August CMS transferred over 3.3 PB in wide-area transfers between storage systems. Of this, disk-to-disk SC4 transfers account for just over 3 PB and our recent two high-throughput Tier-0/Tier-1 disk-to-disk tests for most of the rest. This translates to an achieved rate of ~1 PB/month in CMS world-wide.

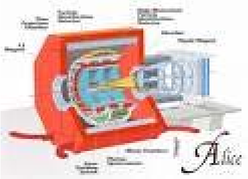
CMS Summary (2/2)

- Specific problems encountered during these tests include various CASTOR2 bugs, such as the fact that CASTOR's reply to the stager_qry command was an arbitrary string that the PhEDEx stager agent had no chance to interpret in a sense that it could determine whether the requested file was on disk or on tape. Therefore it did what it was supposed to do, it submitted a stager_get request for that file. This resulted in a very large number (40K) of stager requests which rapidly overloaded the system. The problem was quickly analyzed and a temporary fix was made available and a permanent fix is expected to be rolled-out by mid September;

- Both CMS and LHCb experienced poor transfer rates into CERN (LHCb from worker nodes used opportunistically, CMS during the centralization of MC data as preparation for CSA06). These problems were eventually traced to a change in a CERN router complex and have now been resolved. However, the intervention on the complex that led to these problems did not follow the agreed procedure for scheduling and announcing such changes and it is imperative that these procedures are rigorously followed in the future;

- Work on patching and tuning the gLite Resource Broker as preparation for CSA06 (in collaboration with ATLAS) has been successful. Thus the CMS requirement to handle 50K jobs / day on less than 10 RBs can be met.

- CSA06 ramp-up has started. Full exercise from 3 October to 15 November.

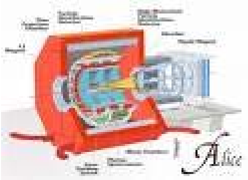


ALICE Summary (1/2)

- PDC’06 includes the integration of the FTS service into the ALICE File Transfer Daemon (FTD) and to test the operation and stability of the combined system.
    - T0-T1: migration of raw data produced at T0 and 1st pass ESDs also produced at T0
    - T1-T2, T2-T1: transfers of ESD and MonteCarlo data and AODs for custodial storage respectively
    - T1-T1: replication of ESDs and AODs

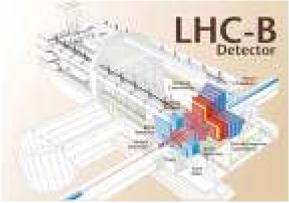
- Multiple successful transfers have been performed to all T1 sites involved in the exercise, however the target rate of 300MB/sec sustained for a week has not been met. The exercise so far has allowed to expose a number of critical areas and as such was a very important step toward achieving full nominal rates under realistic conditions.

- This exercise is ongoing and currently there are some concerns about the number of files and the filesizes used for transfers



ALICE Summary (2/2)

CNAF	Unstable overall with different sources of errors, most frequent are related to inaccessible storage (CASTOR2). Max rate achieved: 28.4 MB/s.
RAL	Joined the exercise late (site-wide issues with disk storage). Difficult to debug problems, transfers stay in waiting status without clear reason or fail. Resources for ALICE are not sufficient for the duration of the exercise – 1.8 TB of disk without garbage collector. Still in a setup phase.
CCIN2P3	Generally very stable. Problems with srm_get. Max rate achieved: 121.4 MB/s.
SARA	Problems with the LFC catalogue (backend ORACLE), associated to null comments inserted by ALICE. VO-box instabilities have adverse effect on transfers. Max rate achieved 47.6 MB/s.
GridKA	Unstable overall with a variety of errors: SRM connection refused and transfer timeouts. Max rate achieved 164.3 MB/s.
CERN	Hardware problems with the VO-box, affecting the transfers to all centres. VO-box replaced.

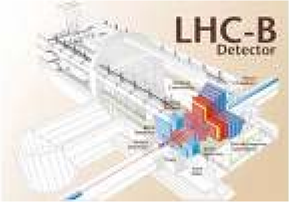


LHCb Summary (1/3)

- The goals of the LHCb DC06 activity are as follows:
  - Distribution of RAW data from CERN to Tier-1's
  - Reconstruction/stripping at Tier-1's including CERN
  - DST distribution to CERN & other Tier-1's

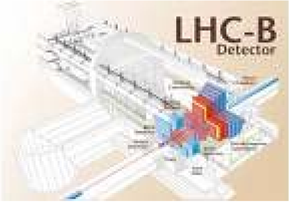
- Simulated data are shipped to the 6 T1s + CERN with a share that depends on the computing power and status of the site. The amount of data processed is correlated to the amount of integrated data transferred out of CERN to various T1. So far the integrated rate is small (but close to a final draft of the computing model : ~3MB/s to each T1).

- Problems at NIKHEF/SARA (dcap callback mechanism incompatible with network setup – resolved in a beta version of dCache) and at Lyon (use of gsidcap not yet supported by a production version of ROOT) impacted production, although temporary workarounds were found in both cases. For the above reason NIKHEF/SARA is not currently participating;



LHCb Summary (2/3)

CERN	ran smoothly its share of jobs during the first month. Some issues with the AFS area serving the Software Installation Area that currently prevents to install jobs through a normal grid job. Problems with the Castor storage in uploading files from simulation jobs running on the small centers (due to the HTAR configuration) and also in the grid mapfile creation that seems to be uncorrelated to VOMS/LDAP mechanism as it happens somewhere else. Flickering behavior of the Information System.
CNAF	potentially CNAF is the largest center and could process the largest share of data. However it suffered a long standing problem with Castor2 stager. Basically CNAF are using a different configuration to at CERN where for each VO there is a dedicated instance of the DB and LSF. There are several reasons behind: 1. The single disk server serving the LHCb requests from LSF was not enough. There was also a limit on the max number of jobs per disk server increased to 300. (Fixed) 2. The DB is overloaded (deadlocks) and all the requests to the stager are stuck (fixed) 3. The pure disk pool (no Garbage Collector) seems to have problem in accessing files in case it becomes full (with consequent pending jobs overloading the LSF queue). Now CNAF should be OK.
IN2P3	ran smoothly DC06. Some minor issues due to the storage. They are using at Lyon the disk only storage instead of the tape endpoint (this last supporting only gsidcap protocol). Length of the largest queue did not fit with the LHCb Simulation jobs. Flickering Information System also experienced there.



LHCb Summary (3/3)

FZK	Poor usage of GridKA for reconstruction jobs of this DC06 (because it prevents to access data directly from the application), it has been rather used for production. The main problem (under investigation) seems related to their gridftp daemons that decide to close their sockets from time to time.
PIC	some issues with the storage; recent issue with pilot jobs that were not picking up any production (either reconstruction or simulation) job. PIC ran its share without any other major problem.
RAL	also ran smoothly DC06 reconstruction jobs without major issues. Experienced a slowness accessing data at some point and problem fixed by adding another disk server.
SARA	NIKHEF/SARA never used for reconstruction: it is currently impossible accessing (through Root) data stored in the WAN connected Storage at SARA from WN via dcache.A patched version of the dCache client has been released for test. This version doesn't require Inbound connectivity on the WN because it wouldn't require calls client back. Site admins at NIKHEF are very collaborative and are pushing for testing/certifying new dcache libraries needed by LHCb. Once there will be proof that new clients are working fine they will install in their nodes without waiting official release of LCG. Until further news, NIKHEF sits out DC06 activity.

SC4 Review

Service Levels, Problem Response and Resolution

LCG

Worldwide LHC Computing Grid

Distributed Production Environment for Physics data Processing

WLCG Service Availability Targets - CERN

- Based on experience of Service Phases of SC3 & SC4, where do we stand with respect to the Service Availability targets in the MoU?

- Take 2 concrete examples:

1. Event reconstruction;
2. Distribution of data to Tier1s during acceleerator run.

WLCG Services

- These two services are characterised by strong dependence on both VO and LCG provided services
- Data export introduces a further coupling to storage services at Tier1 sites
- Typical interruptions seen:
  - 02:00 weekdays until 10:00
  - 14:00 Saturday until Monday 10:00
- Cannot meet targets without on-call services!

WLCG TIER0 MoU Targets

Service	*Maximum delay in responding to operational problems*			*Average availability[1] measured on an annual basis*	
	Service interruption	Degradation of the capacity of the service by more than 50%	Degradation of the capacity of the service by more than 20%	During accelerator operation	At all other times
Raw data recording	4 hours	6 hours	6 hours	99%	n/a
Event reconstruction or distribution of data to Tier-1 Centres during accelerator operation	6 hours	6 hours	12 hours	99%	n/a
Networking service to Tier-1 Centres during accelerator operation	6 hours	6 hours	12 hours	99%	n/a
All other Tier-0 services	12 hours	24 hours	48 hours	98%	98%
All other services[2] – prime service hours[3]	1 hour	1 hour	4 hours	98%	98%
All other services[2] – outside prime service hours[3]	12 hours	24 hours	48 hours	97%	97%

[1] (time running)/(scheduled up-time)

[2] Services essential to the running of the Centre and to those who are using it.

[3] Prime service hours for the Host Laboratory: 08:00-18:00 in the time zone of the Host Laboratory, Monday-Friday, except public holidays and scheduled laboratory closures.

Event Reconstruction- Recent CERN experience

- Event reconstruction was performed using the local batch system, i.e. LSF
- Other services involved include the conditions database service used by the experiment in question (an Oracle-based application for all except ALICE), the experiment-specific book-keeping system(s) (typically based on Oracle and/or MySQL), the LFC (either as a file catalog or as the basis of the CMS DLS), and CASTOR2;
  - In the recent ATLAS Tier0 exercise, DDM/LFC operations were decoupled leaving dependencies only on CASTOR, LSF and AFS;
  - In this exercise, AFS was the primary bottleneck and cause of job failures. This is being followed up (e.g. by the use of volume replication);
  - Overall LSF performed worse than in the previous test – leading to the suggestion that a dedicated instance for first pass processing might be needed;
  - CASTOR exceeded the goal of 1 week of stable operation but with a pool 2-times over-dimensioned and Atlas wasted time trying to understand its performance;
- Steps are being taken to ensure reliable services, although coupling to CASTOR, LSF and AFS (and presumably experiment-specific services) remains. All of these services are complex and problems typically require ‘the expert’ to be solved;

Distribution of Data (1/2)

- This activity is loosely coupled to event reconstruction in that it requires the output of the reconstruction phase. It is, by definition, tightly coupled to the storage management services of the host laboratory (CASTOR + SRM, hence also Oracle and LSF), as well as the FTS (which also depends on Oracle), the experiment-specific framework that drives the FTS, as well as the corresponding storage management services at all of the Tier1 sites supporting a given VO;
- Except in the case of failure or severe degradation of host laboratory services, problems with a single site can, in principle, be tolerated provided that the site in question has the proven ability to catch up with a backlog, however caused (e.g. source/sink error, or both);
- On the assumption that recovery from backlogs is demonstrated, expert coverage can probably be limited to ~12-16 hours per day. Although inter-site problems typically require dialog between experts on both sides, more than 2/3 of the data is sent to European sites, where the maximum time difference is 1 hour;
- Sites must still respond to site-local problems as per MoU

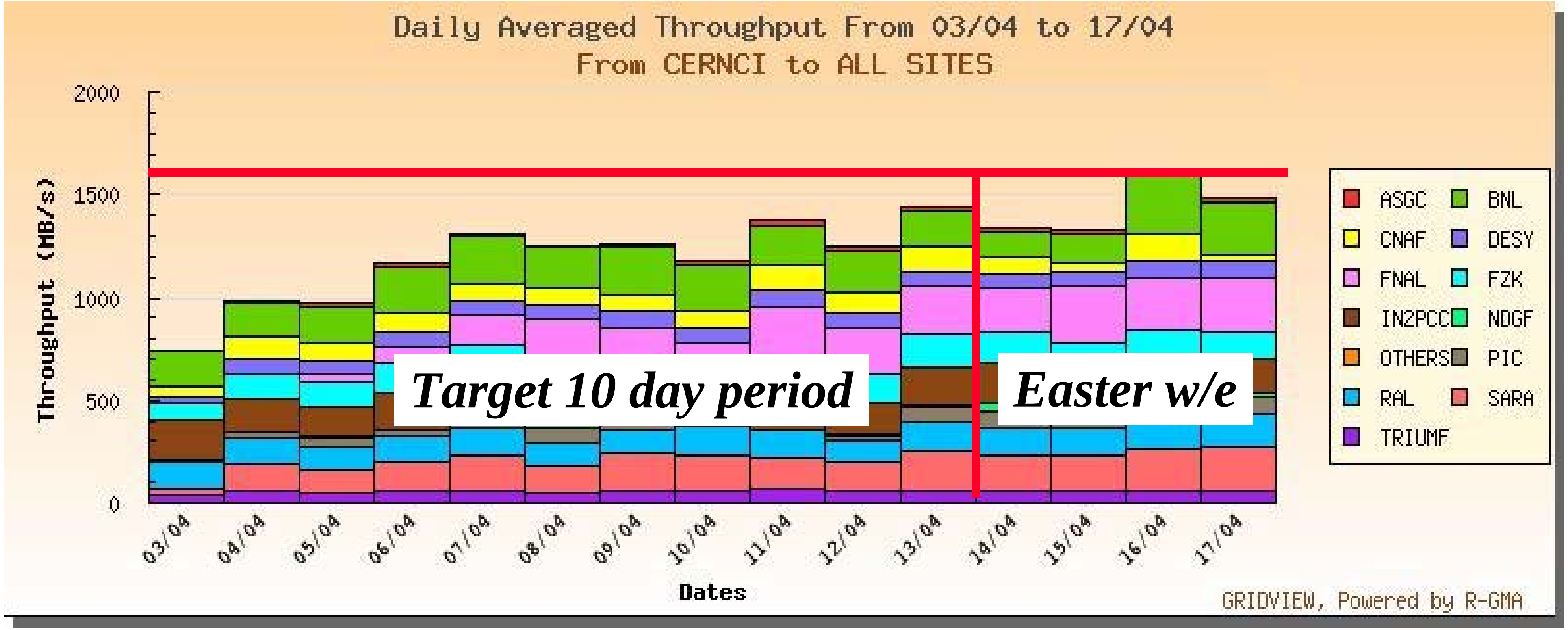
Distribution of Data (2/2)

- In the case of data export to the Tier1 sites, corresponding on-call services are required at the Tier1s as well, together with inter-site contacts and escalation procedures;
- Global Grid User Support and COD (rotating Core Infrastructure Centre On Duty) currently provide a service during office hours only, but should provide the primary problem reporting route during data distribution periods. This requires that realistic VO-specific transfer tests are provided in the Service Availability Monitoring (or equivalent) framework, together with the appropriate documentation and procedures;
- The list of contacts and the procedures for handling out-of-hours problems will be elaborated by the WLCG Service Coordination team and presented to the Management Board for approval. These procedures will be constructed to facilitate their eventual adoption by standard site operations teams as extended cover becomes provided. Such a service must address both problem determination and problem resolution.

SC4 Review

Summary of Service Coordination Observations & Recommendations

SC4 April High Throughput Results



Target 10 day period

Easter w/e

Observation #1

- We are still not able to demonstrate full nominal Tier0-Tier1 transfer rates (1.6GB/s) over extended periods, let alone recovery rates (targeted at twice nominal);

Observation #2

- However, experiment-driven data transfers (ATLAS and also CMS) achieved rates close to the target of **full nominal rates** (about half of the total rate for all experiments) under much more realistic conditions than for previous DTEAM Service Challenge transfers. For this reason, this is considered a positive result;

Observation #3

- Both ATLAS and CMS have managed to export over **1PB** of data (1 PB of data per month for CMS over a 90-day period, **1.25PB** of data for ATLAS in the two-month period starting 19th June);

- A particular effective model for consistent performance, as demonstrated by Lyon for ATLAS, is to have a contact person for the experiment **both** at the Tier0 and the Tier1;

Observation #4

- By definition, these activities tested site services, such as LFCs, VO boxes, and overall production readiness significantly more than the DTEAM-driven transfers. A number of issues have been found at a variety of sites and solutions have been found or are planned.

- However, they underline the fact that certain sites / regions still have to make **significant progress** to achieve the required service level

Observation #5

- Sites appear to be able to focus their full attention on a specific experiment or challenge for a **few days only**. This is clearly indicative of the high workload at the sites and should be built into the experiments' operational models (i.e. a few days at high priority per month per experiment already completely drains the sites involved);

Observation #6

- Upgrades to CASTOR2 at a number of sites have led to further instabilities. Once all such migrations have been completed, a further test needs to be made to ensure that these sites can now meet both throughput and stability targets;

Observation #7

Several sites have experienced significant power and / or cooling problems, resulting in prolonged service downtime;

Observation #8

- Several – if not many – sites appear to suffer from significant manpower shortages, which impacts both the service level that they are able to provide and the response time to requests (both “setup” and problem resolution);

- This was particularly evident around both the Easter and Summer vacation periods

Observation #9

- Reporting to and attendance at the weekly Joint Operations Meetings[1] has improved since the previous report in May 2006 but still leaves considerable room for further improvement
  - Site reports are often written in a style that is clearly oriented at local consumption,
  - some sites still do not provide reports on a regular basis, even though there is significant activity at that site;

[1] See http://agenda.cern.ch/displayLevel.php?fid=258 to access agendas, reports, action items and minutes.

Observation #10

- Opportunistic use of resources – used or expected to be used by all experiments – may result in the use of CPU resources at sites with insufficient local storage. As an interim solution, unrestricted WAN access to the CERN SE has been provided, but this can result in poor and/or unpredictable network performance and result in problems that are highly complex to debug. It is considered important to clearly separate this opportunistic use of resources from the standard production model, where data is typically written to the local storage element (and eventually archived to the associated Tier1 site in the case of Monte Carlo production at Tier2s.);

Observation #11

- A bug in Oracle 10.2.0.2 led to logical data corruption in the LFC and VOMRS instances at CERN. Once the problem had been sufficiently understood, it was successfully escalated to Oracle as a top priority issue. A work-around was put in place and the experiments and all outside sites were advised accordingly.
  - At the time of writing a patch that passes all test cases has still not been received, although the workaround – effectively to turn off the faulty code path – solves most of the problems and eliminates the risk of further data corruption.
- This can be viewed as an important test case both of our ability to escalate such problems within the Oracle support structure as well as to handle bugs that potentially affect a large number of sites.

Recommendations & Actions

- Streamlining of reporting to the weekly combined operations meeting – now to be held on **Thursdays at 16:00** Geneva time – and the various LCG coordination meetings (LCG **Experiment Coordination** Meeting Mondays at 15:00, LCG Service Coordination Meeting Wednesdays at 10:00) has been proposed to the WLCG Management Board and has been put in place;

Recommendations & Actions

- The use of the EGEE broadcast tool for announcing both scheduled and unscheduled interruptions has greatly improved. Improvements in the tool to clarify broadcast targets are underway. Sites are requested to ensure the nature and scope of the event are clear both from the subject and text of the announcement (and are not, for example, inferred from the e-mail address of the sender);

☺ Tape robot maintenance at CERN 10.30-16.00 Thursday 13 July

☹ Tape access interrupted

Recommendations & Actions

- Site monitoring of local services still needs **considerable further improvement** – many issues that could be spotted locally are still first found by the central Service Coordination Team or – worse still – by the users;
- Sites are encouraged to share their monitoring tools and experience. To this end, a focussed discussion on monitoring was held at the recent Service Challenge Technical Day, September 15th at CERN, and which revealed a very inhomogeneous and confusing set of monitoring tools.

Recommendations & Actions

- Problem resolution – and reporting – needs to be improved, particularly in the case of complex problems which require a range of expertise and / or sites to resolve ;

- Regular reviews of open tickets and identification of complex / unresolved problems are held with escalation (depending on the exact problem) as required.

- This has proved successful in the resolution of chronic LHCb problems as well as the CMS CSA06 preparation issues.

Recommendations & Actions

- Phone and / or physical participation of the experiments in the CERN daily operations meeting[1] (starting at 09:15) is encouraged to highlight new problems and ensure that there is adequate information flow. This has proved very useful, for example, in the current setup phase for CMS CSA06.These meetings are also open to external sites wishing to participate;
- (The meeting starts at 09:00 with a review of CERN internal tickets)

[1] These meetings are typically held in the "openspace" in B513, except when this room is needed for a VIP visit. Dial-in access is via +41 22 767 6000 access code 0175012.

Recommendations & Actions

- A WLCG “Service Dashboard”, allowing both supporters and production managers to clearly see the status of critical components (CASTOR@CERN, FTS, network transfers etc.) should be implemented as soon as possible to replace the laborious manual expert intervention – typically scanning log files – that is currently required;

Recommendations & Actions

- A “Service Coordinator (On Duty – SCOD)” – a rotating, full-time activity for the length of an LHC run (but almost certainly required also outside data taking) should be established. The person assuming this activity would, for their period on duty:
  - Attend the daily and weekly operations meetings, relevant experiment planning and operations meetings, CASTOR deployment meetings;
  - Liaise with site and experiment contacts;
  - Maintain a daily log of on-going events, problems and their resolution;
  - Act as a single point of contact for all immediate WLCG service issues;
  - Escalate problems as appropriate to sites, experiments and / or management;
  - Write a detailed ‘run report’ at the end of the period on duty.
- It is proposed that this rota be staffed by the Tier0 and Tier1 sites, each site manning ~2 2-week periods per year (or 4 1-week periods);

Recommendations & Actions

- A regular (quarterly?) WLCG Service Coordination meeting, where the Tier0 and all Tier1+Tier2 federations as well as the experiments are represented, should be established. This should review the services delivered by that federation, main issues encountered and plans to resolve them, possibly following the model used by GridPP for their collaboration meetings (see, for example Deployment Metrics and Planning, presented at GridPP16). It should also cover the experiments' plans for the coming quarter in more detail than can be achieved at the weekly joint operations meetings (which nevertheless could cover any updates). This meeting should not require physical presence, but would require the reports / presentations to be submitted in advance;

SC4 Review

Outlook & Conclusions

Outlook

- Service Challenge 3 to Service Challenge 4 involved only ‘minor’ changes
- From Service Challenge 4 to LHC startup, we need to understand:
  - Migration to grid middleware level gLite 3.x (some major rewrites);
  - Implications of SL(C)4 both 32 and 64-bit;
  - Deployment of SRM 2.2-compliant data management solutions;
  - Production 3D-services (replicated data bases) as part of WLCG;
  - Critical VO-box and other new services
- We also need coordinated exercises to prepare for Tier1<->Tier1 and Tier1<->Tier2 transfers
- Continue to improve Service Level & Response times!

Summary & Conclusions

- For all its problems, SC3 and more completely SC4 have resulted in improved production services across many sites
- Much has been accomplished; much still outstanding
- Service Coordination two top issues?
  - Collaboration & communication at such a scale requires significant and constant effort (regular remote meetings, workshops such as this one)
  - “Design for failure” – i.e. assume that things don’t work, rather than hope that they always do!