Table of Contents

Week of 091012

WLCG Service Incidents, Interventions and Availability

VO Summaries of Site Availability				SIRs & Broadcasts	
ALICE	ATLAS	CMS	LHCb	WLCG Service Incident Reports	Broadcast archive

GGUS section

- GGUS Escalation reports every Monday (used for WLCG Service Report to MB)
  - Full search: https://gus.fzk.de/ws/ticket_search.php and select "VO" - this will give all tickets, including team & alarm
- **LHCOPN** Tickets in GGUS: see https://gus.fzk.de/pages/all_lhcopn.php and change your selection criteria. Future actions are also listed.
  - If network group participation is necessary, please invite them in time.
- OSG items selected from the GGUS escalation reports.
- Procedure to become a LHC Experiment VO TEAM member
- Other recent GGUS FAQs

Daily WLCG Operations Call details

To join the call, at 15.00 CE(S)T Monday to Friday inclusive (in CERN 513 R-068) do one of the following:

1. Dial +41227676000 (Main) and enter access code 0119168, or
2. To have the system call you, click here
3. The scod rota for the next few weeks is at ScodRota

General Information			
CERN IT status board	M/W PPSCoordinationWorkLog	WLCG Baseline Versions	Weekly joint operations meeting minutes

Additional Material:

STEP09	ATLAS	ATLAS logbook	CMS WLCG Blogs

Monday:

Attendance: local(Jamie, Gang, Eva, Jean-Philippe, Harry, Gareth, Simone, Daniele, Olof, Kors, Dirk, Jason, Alberto, Patricia, Roberto, Miguel, others);remote(MIchael, Gareth, Ron, Jos, Kyle).

Experiments round table:

- ATLAS - On Friday throughput test ramped down. Remaining subscriptions drained over w/e. From today 2 next data distribution activities: ESDs from MC & reprocessing - were only in 1 T1, being replicated. 2nd: placement of data for analysis tests. Will last whole week. Most important point: after 1 year of discussion(!) about to close default CASTOR pool at CERN. Procedure for asking for space for CASTOR T3 pool - pointer to be provided.

- CMS reports - Being of 2nd week of Analysis ex. of Oct. More WMS probs - # aborted jobs seems to be a bit higher - globus error 10. T1s: RAL & ASGC - both transfer errors. RAL: large scale expiration of transfers to all T2s. Recent downtime?? To be checked. ASGC: timeouts in T1->Pisa.

T2s: many issues. See details.. SAM tests - many failures, mostly CE tests.

- ALICE - Stressing SLC5 resources at CERN: WMS & CREAM. LCG CE: ce129 began to show OPS&ALICE tests bad behaviour. This CE was not able to support # jobs submitted through WMS. Around 22:00 Maarten informed submitting > 13K jobs. 50% through ce129. Cleaned up one of WMS. Expected that as soon as ce129 comes back in production some 7K jobs will arrive at this CE.

- LHCb reports - Very few jobs running right now. Validation of previous MC prod ongoing. Over w/e (MC at full steam) filling MC* space tokens at various T1s - see detailed Twiki. Only PIC had 2009 resources available. Others ran out / short of space. Resources not pledged ones... Reduce activity and/or reduce # replicas - still under discussion. Other than disk shortage: hanging connections reported last week seem now to be understood. But.. still problem with hanging connections at Lyon and GridKA. Now TURL resolution at SARA - tickets opened - still under investigation.

Sites / Services round table:

- ASGC: Jason - looking into transfer problems mentioned by CMS. New LTO drives to be installed.

- BNL: over w/e production at T1 impaired by configuration problem in PANDA controller. Production jobs went on hold - not able to put output on SE. Studied and fixed by PANDA team. Production is now continuing. Major intervention coming up tomorrow - announced already. dCache -> 1.9.4. CE -> OSG 1.2. Transparent intervention on Oracle for LFC/FTS to supply O/S and Oracle sec. patches.

- RAL: Situation at RAL - CASTOR service back end Friday pm. OK over w/e. Couple of questions - some CMS files maybe missing from window back in September? Real problem or not? Being checked. 3D DBs not back up - moved to alternative h/w. LHCb: waiting for sync to catchup - in touch with "CERN people" with issues on ATLAS one. Underlying cause of problem: still not known why problems across a # of diskservers. Environmental thing? Power? Earthing? All relevant DBs - incl CASTOR, LFC, 3D - on alt. h/w to allow better investigations on failing h/w.

- FZK: two small things: 1) LHCb running out of space - will check tomorrow to see if still some physical space left for the MCM space token. Also understood from Alexei Zhelezov that this is a general T1 issue. 2) Friday PBS stopped accepting jobs. Known issue. SImple reset no good - needs expert. Not as many jobs running as could have been...

- NL-T1: ATLAS had issue with transfers to NIKHEF DPM - checksum errors - still under investigation, also with FTS team. Upcoming downtimes next week: Oct 20 - upgrade to SL5; same week also migration Oracle RAC 3D to new h/w and LFC of ATLAS + LHCb will move as well - to be announced in GOCDB.

- DB: replication CERN to RAL working again. Both DBs are sync-ed. For ATLAS reconfigured streams setup - now going through 6 day backlog. Will take some time. Maybe by tomorrow afternoon? Tomorrow will migrate CMS offline DB to new h/w. LHCb online DB still running on 3 out of 4 nodes - vendor investigating. Gareth - LHCb 3D OK? A: Y.Simone - doubt most recent conditions will be used in next 24H so probably can end downtime. ASGC 3D issues will discuss offline.

- CERN - SRM upgrade for CMS to 2.8 on Oct 14; migrating h/w for DBs for all CASTOR instances. CMS & ALICE for 20-21 Oct. Schedule 1h downtime also for ATLAS & ALICE. Daniele - for CMS Oct 14 is fine.

- OSG - nothing special.

AOB:

Tuesday:

Attendance: local(Jamie, Daniele, Gang, Eva, Olof, Miguel, Patricia, Roberto, Flavia, Simone);remote(Onno, Gareth, Jos).

Experiments round table:

- ATLAS - The ATLAS CASTOR default pool is closed. Please redirect users to atlas-comp-cern-castor-support@cernNOSPAMPLEASE.ch and https://twiki.cern.ch/twiki/bin/view/Atlas/CastorPools. ATLAS jamboree on-going in CERN IT amphitheatre: http://indico.cern.ch/conferenceDisplay.py?confId=66012. Quick update: no major problem; replication of cosmics from detector going on. RAL now in the game - transfers started after yesterday's meeting. 2nd activity that was ongoing but now finished is distr. of data for analysis tests to T1s. From tomorrow distribution from T1s to T2s. Must be finished end week. Reminder: func. tests turned off for moment - will restart Thu lunch.

- CMS reports - the CMS Analysis "October Exercise" is running (5-18 October 2009). Still some issues with gLIte WMS: global error 10 still not fully solved. T1s: ASGC-Pisa now solved; CNAF - installation job hanging; RAL: looks as though files needed are in buffer and not on tape - see full report. RAL - do you know what happened? T2s: huge rampup in # tickets in Savannah: 4-5 tickets/day for T2s up to 12-15; some grouping into categories; full details in report...

- ALICE - # of jobs in production being recovered; checking SLC5 resources at CERN SAM tests for ce128/9 still failing with ALICE. OPS ok for ce129 but not for ce128. Some auth problem?? ALICE production only with SLC4 resources...

- LHCb reports - 2.5K jobs running in the system right now between production/merging and distributed user analysis activity. At "task force" this morning decided to relaunch some stripping production activities to demonstrate and to debug (or to verify the solution adopted about) the problem of hanging connections at the 3 dCache faulty sites - see detailed report. Would like to invite site admins of affected sites to follow SEs closely. Mail from Ron: hanging dcap movers and connections understood by some tuning - default timeout means connections killed after 100 hours - set timeout to ~2hours. After this dcap movers cleaned up. Maybe not bug but just config.Discussed the possibility to reshuffle the allocation of space tokens at sites. (again see detailed report). Confirm gateway at CERN to SLC5 also problematic for LHCb. GGUS ticket opened just before meeting - # in report. Problem with directories with StoRM at CNAF.

Sites / Services round table:

- ASGC: nothing to report.

- NL-T1: a few small issues - about SRM of SARA (LHCb problems) found out that colleague was running stage tests at the time. Running dCache with a bug that SRM cannot handle stage tests well. Probably cause of problems yesterday evening. Last Friday reported broken pool node; last w/e copied file(s) to another pool ndoe and yesterday morning could incorporate new location of files in dCache - should be available now. From NIKHEF DPM: yesterday Ronald reported checksum issue - according to Simone Campana caused by FTS/CASTOR@CERN problem. Simone - still believe that this was the case! Checksum stored in CASTOR - known to be correct - is different to one SRM exposes - leading zero is stripped! Open thread with developers.

- RAL: All services up and running on alternative h/w hosting Oracle DBs. Possible problem reported yesterday with some CMS files missing - seems to be a problem, DB restores or something else? Not sure - following up actively.

- FZK: hiccough with ATLAS streams this morning - unknown what caused it - hopefully clear by tomorrow.

- CERN: CMS DB offline intervention has been postponed as CMS running October exercise. This intervention will take place next week. Tomorrow LHCb offline DB->new h/w; ATLAS streams replication CERN->RAL completely synched and running ok.

- CERN: CE - checked problems with service manager; only 1 available (course) - LCG CE -> SLC5 not working correctly & CREAM CE to be upgraded. Look urgently into ce128/9 issues to get submission to SLC5 working corrrectly asap. Some new machines delivered end last week -> prod asap. Simone - ATLAS observed globus error 10 abck in days of 32K files in same directory. Condor could not clean for 'mysterious reason'. Maarten is expert. Roberto - patch for this should be in place at CERN.

AOB:

Wednesday

Attendance: local(Jamie, Eva, Miguel, Julia, MariaDZ);remote(Joel, MIchael, Gareth, Ronald).

On-going GDB with sessions on experiment operations(!):
http://indico.cern.ch/conferenceDisplay.py?confId=45480

Experiments round table:

- ATLAS - Nothing to report.

- CMS reports - CMS affected by Castor@CERN issues (see SLS plot in detailed report). GGUS Alarm opened: #52537. Question: why the GGUS Alarm by LHCb is closed already and the CMS one is still open? Aren't they the same problem? We would need a prompt ping/confirmation of when the situation has recovered, so we can restart safe Ops. Response from CERN FIO (GGUS ticket updated): Problem was solved yesterday within 1h. Please have a look at http://it-support-servicestatus.web.cern.ch/it-support-servicestatus/ for more details. There were a lot of tickets to update yesterday, this one slipped through...

- ALICE - no report

- LHCb reports - Re-running stripping process. After increasing the number of movers to 1500 per pool at SARA and GridKA and IN2p3 things seem a bit better and watch dog killing stuck jobs problem is not reported. CASTOR was not available at all yesterday evening. .ALARM ticket open.Problem reported on the IT Service status board. Scheduled intervention on LHCBR today affecting our LFC and BK. T1 issues: Requested NL-T1 to clean up some old data being apparently not possible from remote; Failure to list directories on GridKA dCache SRM. T2: shared area issues. Joel: stress that at GridKA have a lot of problems to access data - no progress. Forwarded recipe from Ron but no effect.

Sites / Services round table:

- BNL: following dCache upgrade yesterday - completed during scheduled window - observed fairly high load on SRM server incl. gPlazma cell. This cell is taking ~5x more CPU than previously. Causing timeouts and a significant reduction in throughput. Processing of SRM requests taking much longer - timeouts. Deciding what to do... Move gPlazma cell to dedicated machine? One of temporary solutions could be to go back to flat file authentification and investigate gPlazma code.. Otherwise dCache upgrade completed; other upgrades also ok incl. CE & intervention on Oracle DB for FTS & LFC. Shifters reported overnight site services running at BNL reported by SLS to be down.

Investigated - nothing wrong with site services - Panda server is not receiving call backs at times and hence DQ2 is restarted. After 4 times reported to be not working ok in SLS. Have to find out why callbacks not received.

- RAL: We are investigating a loss of data from Castor at the RAL Tier1. This loss follows the restore of the Castor databases from a hardware problem that resulted in no Castor service from the 4th to 12th October. The loss appears to be for ALL data (i.e. all VOs) written into Castor from 00:15:56 on Thursday 24th September local time (23:15:56 on Wednesday 23rd September UTC) until the system crashed on 4th October. We are working to understand this problem and verify the details above. Gareth - CMS reports on loss of data showed big problem. Restore of DBs at end last week had problems. Upshot: lost data during window above. Looking as though any files added in this window are lost. Some concern - having wound DB back - that we are reusing CASTOR file IDs. Short outage of CASTOR tomorrow morning - discussed at CASTOR F2F - to wind fileID on. At the moment all Oracle DBs are running on alternative h/w while we try to understand failures.

- NL-T1: SARA has done some work on DNS setup. One of slave DNS servers did not take up changes - caused some problems affecting CEs *.sara.nl. Problems found - waiting for caches to be cleared. LHCb issue - cannot delete CCRC08 data from SRM - still under investigation. Manual deletion ok.

- ASGC:

- CERN: will send post-mortem of details, but: yesterday ~16:30 nameserver issues. A few minutes after monitoring alerts; then user complaints; then OPS calls for alarms through GGUS. Around 17:00 problem found - message on site status board. No access to CASTOR. 10' later issue found - in root of catalog /castor fileclass null instead of 0. Fixed - SSB updated. Tried to notify users. GGUS ticket hadn't made it to Remedy.. Not closed correctly in GGUS - only this morning after passing via ROC. Post-mortem with more details will be posted soon. This morning LFC service manager (who had been on course...) was not aware of LHCb DB intervention. LFC was not put in downtime. Understood what problem was. Eva - responsible for DB was checking list of coordinators and service manager for LFC overlooked.

- DB: LHCb offlline has been successfully migrated this morning to new h/w. Tomorrow had announced intervention on LCGR production DB - will be postponed by 1 week. (CMS have requested it due to on-going activities). ALICE dashboard application will be down tomorrow pm for 1h to migrate from ALICE offline DB to LCGR.

Release report: deployment status wiki page

AOB:

Thursday

Attendance: local(Daniele, Gavin, Simone, Jason, Jamie, MariaDZ, Diana, Andrew, Roberto, Gang, Harry, Jean-Philippe, Ricardo, Edoardo, Alessandro, Eva);remote(Xavier Mol (FZK), Jeremy (GridPP), Alexei, Gareth Smith (RAL), Ronald Starink (NL-T1), Michael).

Experiments round table:

- ATLAS - RAL: ATLAS stopped data delivery T0-RAL and from RAL to other sites; MC production in UK cloud still running; problems in getting input data from RAL from time window. Suggested to stop MC production too. Situation in RAL/UK cloud will be re-assessed at ATLAS ops meeting later today: decide then how to proceed. ASGC: problem in copying data from/into CASTOR over night; Now fixed - site back to ~!00% efficiency. Problem staging some files from CASTOR@CNAF - site informed. (Cert problem apparently).

- CMS reports - CMS Analysis "October Exercise" still running; thanks for postponing LCGR DB intervention! October exercise still long way from ramping down... gLite WMS problems still open.. RAL traffic basically suspended. Keeping ticket open & use to track list of files to be invalidated. CNAF: installation job of CMS sw release hanging. Still debugging in CMS layer.. T2s: many issues; file access problems, SAM test failures; more in detailed report.

- ALICE - no report.

- LHCb reports - Not much activity to report. Few MC production (jobs) running. Total of 500 jobs in the system. The stripping rerun yesterday went fine at NL-T1 and IN2P3 but it has reported failure at GridKA still with jobs hanging in connection. LHCb Shifters pursuing that. Issue with server in LHCb data pool stuck yesterday - alarm ticket opened - maybe "a bit tough". CNAF: disk server problem due to cert. RAL: because of the faulty database restore we isolated as definitely lost ~200 files that have to be corrected on the various LHCb catalogs. Many more were lost during the period between the 124th of September and the 4th of October but those were just intermediate files that do not need to be corrected. GridKA still valid the problem (GGUS originally open for that must be used to track this problem down) of hanging connection. GridKA: opened yesterday ticket about failure in listing directories (opened yesterday) is still valid.

Sites / Services round table:

- RAL: Unfortunately we have to confirm the loss of all data for all VOs written into Castor for the period from 00:15:56 on Thursday 24th September local time (23:15:56 on Wednesday 23rd September UTC) until the system crashed on 4th October. This loss was caused by a problem during the restore of the Castor database before services were restarted on Friday 9th October. Specifically, the database restore was not restored to the most recent version. We are working obtaining a list of the files lost during this time.
  A further problem has been brought to our attention and is being investigated. The faulty database restore has meant that the system has been re-using file IDs within Castor. These IDs should be unique but there may be cases where this is not the case. It is theoretically possible that files with re-used IDs may be deleted in the future. We had a short outage at 11am local time (10:00 UTC) this morning to increase this ID beyond the range previously used and we are investigating the level of any risk to data written between 9th October when Castor was restarted and this morning. Chance of files being deleted due to above considered somewhat unlikely. Need to understand cause of disk failures as well as backup / restore issues.

- FZK: LHCb problem with hanging transfers - installed solution provided over srm deployment list - hopefully problems gone or minimized. Solution written in tickets - reopen if still problems. Roberto - LHCb will run another stripping activity - this could test whether ok.

- ASGC: problem mentioned by ATLAS due to overload of CASTOR DB - both ATLAS & CMS affected. Solved by restarting services. Action from CASTOR F2F to resolve such problems. Have seen overload before. ATLAS DB now cleaned up and ready to prepare streams / transportable tablespace resync. Back in production mid-end next week? TBC.

- NL-T1: change in downtime for LFC for ATLAS ; no longer 22 Oct - new downtime on 27 Oct. (ATLAS acceptance tests): NIKHEF: upgrade WNs to EL5 on Monday 26th Oct.

- BNL: Following dCache upgrade Tuesday spent majority of yesterday cleaning up. Behaviour of authentification method taking too much time - transfer timeouts. Workaround found but not considered satisfactory. Workaround working nevertheless - in contact with developers for better solution.

- CERN: two new LCG CEs pointing to SLC5. With latest version - basic tests ok. Please report any

problems. 2 more CEs to be installed then upgrade ce128/129. (new CEs 130/131). Simone: is this identical SLC5 installation. Problem in sourcing grid env. Import LFC fails. Simone has workaround - should be cured before exposing to public. Any python command using LFC lib will find same problem. Roberto - same problem . Ricardo: known - comes from gLite WN s/w. Problem with installation or general issue? Will follow up.

- DB: after LHCb migration yesterday propagation online-offline down due to delay in adding to list of trusted machines in LHCb online. Connection now restored. Migration of LCGR to new h/w postponed to next Wednesday.

- Network: deploying 2nd 10gig link to BNL. Don't have date for production - will announce later... Ale - will monitoring of this link be visible to expts? A: not yet.

AOB:

- MariaDZ: Got confirmation of announcement of interventions etc. Documented in https://twiki.cern.ch/twiki/bin/view/EGEE/SA1_USAG#Site_scheduled_interventions . This twiki section will be also linked from GGUS home for a number of week as a "Recently created faq".

Friday

Attendance: local(Jamie, Gang, Miguel, Daniele, Dirk, Harry, Jean-Philippe, Roberto, Lola, Simone, Nick, Eva, Andrew, Andrea, Olof);remote(Michael, Onno, Gareth, Alexei, Angela, Gonzalo).

Experiments round table:

- ATLAS - Received request from ATLAS data preparation about distribution of data collected in last 3-4 days. Export started this morning hence peak in activity from 10:30. No problem to report. ASGC had hiccough this morning - srm failed to serve/get data for ~1 hour. Cured. Not sure how.. DB problem? ASGC would like to split SRM endpoint for different VOs. When? Can be done in ~2 weeks. LFC entries would have to be done... To be discussed.

- CMS reports - T1s: RAL closed ticket on list of files to be invalidated - done. CNAF: CMS sw job still problematic. Would like to know why... T2s: things being followed up on all open issues. Few new tickets, falling into existing categories.

- ALICE - Will startup production again - currently just a few jobs running. Setup of 2 new LCG CEs introduced into ALICE production. No issue. Miguel: CEs being used by ALICE and LHCb - will probably upgrade others (2 more new then 2 old).

- LHCb reports - Increasing number of MC production (and merging at T1's) jobs running in the system (ramp up). Now in the system ~7500 production jobs and ~4000 user jobs concurrently running. T0: ce130 and ce131 test jobs run happily and now the two new CEs are part of the production mask in LHCb: T1: problems fixed! (see detail). 2 new GGUS: NL-T1: fixed; some SAM jobs failing at RAL - understood and fixed; CNAF hiccough against StoRM - again fixed but what was action?

Sites / Services round table:

- IN2P3: The parameter specifying the maximum job duration on a worker node has been unfortunately modified to a very low-level value on Wednesday evening. The consequence is that once the long jobs finished only very short jobs were allowed to run. Long time CPU jobs still queued during the incident. A SIR has been uploaded to https://twiki.cern.ch/twiki/bin/view/LCG/WLCGServiceIncidents.

- TRIUMF: Please note that on Tuesday October 20 from 09:00-18:00 (pacific time), we are going to upgrade our tape library infrastructure (expansion frame + drives + tapes) for 2009/2010 capacity. READING from ATLASDATATAPE and ATLASMCTAPE will NOT BE POSSIBLE. Any prestage request will be pending unless data is already on cache buffers. WRITING to tape will STILL WORK since our buffer area is large enough and it is on RAID 6. We think this will have a minimal impact on ATLAS operations at this point in time. Let us know asap if you think otherwise.

- ASGC: problem mentioned by Simone due to CASTOR problem - no more info - but fixed in 1 hour.

- BNL: 1 minor thing - one of T2s: mid-west T2s transfer failing earlier. Understood and fixed. Discrepancy in free space. Service restored.

- NL-T1: LHCb problem - made some changes to LDAP config and hence some auth errors. Fixed this morning now ok.

- KIT: only old problems again - availability decreased but services ok (net KIT-ASGC).

- PIC: reminder: next Tuesday sched. intervention: dCache to golden release. 8:00 - 19:00 storage and computing services down (reminder).

- RAL: managed to restore CASTOR DB to enable us to extract list of lost files - can now supply this list. Working with ATLAS to verify list (LFC/CASTOR). CASTOR fileid was wound on yesterday and service up and running normally. Still ongoing work to understand h/w problems.

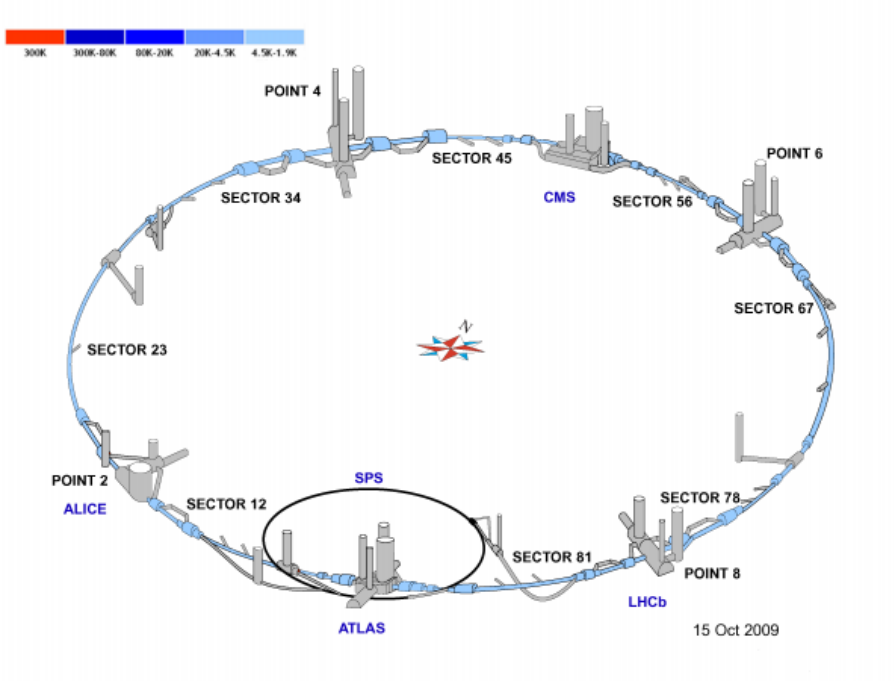
- CERN: starting upgrade campaign on disk servers SLC4 - SLC5. 650 machines to upgrade. Small number at a time. Rolling fashion. Will take > 1 month to complete. May cause some instabilities in data access. Q: are there pools which can be down for an afternoon or a day? These can then be upgraded in one go. This will make process faster. SImone - discuss with ATLAS. Doing a pool seems much better than rolling intervention. Otherwise strange file access errors. CMS pool CMSCAF with 300 servers. Others should be ok for an afternoon. 1150 disk servers; 650 to be upgraded. Others news boxes or already upgraded. T0ATLAS already SLC5. Can see on LEMON.

- DB: next Monday will migrate downstream DBs to new hw/ LHCb am ATLAS pm. Replication from CERN to T1s will be stopped during interventions. Jason confirmed yesterday that ASGC DB is ready but still cannot connect so resync blocked.

AOB:

- SCOD for next week: Jean-Philippe

- LHC status:



-- JamieShiers - 08-Oct-2009

This topic: LCG > WLCGDailyMeetingsWeek091012
Topic revision: r12 - 2009-10-16 - JamieShiers