Table of Contents

Week of 090406

WLCG Baseline Versions

WLCG Service Incident Reports

- This section lists WLCG Service Incident Reports from the previous weeks (new or updated only).

GGUS Team / Alarm Tickets during last week

- Tickets: see https://gus.fzk.de/ws/ticket_search.php☑ and select "VO" - this will give all tickets, including team & alarm - or use the following link:
    - ♦ In the GGUS Escalation reports every Monday☑
- Please observe the procedure to become a LHC Experiment VO TEAM member☑ and other most recent GGUS FAQs☑.

Weekly VO Summaries of Site Availability

- ALICE☑
- ATLAS☑
- CMS☑
- LHCb☑

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☑

General Information

- Web archive: https://mmm.cern.ch/public/archive-list/w/wlcg-operations☑
- CERN IT status board: http://it-support-servicestatus.web.cern.ch/it-support-servicestatus/☑
- elogs: https://prod-grid-logger.cern.ch/elog/CCRC'08+Logbook/☑

See the weekly joint operations meeting minutes

Additional Material:

Monday:

Attendance: local(Gavin, Daniele, Jean-Philippe, Alessandro, MariaDZ, Nick);remote(Gareth/RAL, Michael/BNL, Jeff/Nikhef).

Experiments round table:

- ATLAS (Alessandro)- reprocessing working fine during the weekend - waiting for overview on Thursday - ticket submitted to GGUS about DDM/DQ2 because of checksum mismatch (new client will be tested soon) - team ticket to RAL because of problems in transferring files to other Tier1s,

problem known, solved this morning, but should have been an alarm ticket and not a team ticket - many problems at CERN: acron jobs failing on lxplus, VOMS certificates not installed on AFS/UI (installed now). Question from Jeff: a lot of jobs are accessing data not only at SARA/NIKHEF but also at external sites (Roma), overloading the general network, Alessandro will check and talk to Huong.

- CMS reports (Daniele)- Smooth operation - debugging minor issues like slow transfers - also slow reply to CASTOR ticket - pending issues at CNAF to be checked when they are back - 2 GGUS tickets opened: Warsaw (problem is being worked on but ticket not updated), IPHC in France (ticket solved and closed in Savannah) - CMS shifts will restrat after Easter. Comment from MariaDZ: Kreutzer asks for some GGUS development (interfacing GGUS with Savannah); need more information from Kreutzer to avoid duplicate or unnecessary work. Daniele will check with Kreutzer.

- ALICE -

- LHCb -

Sites / Services round table:

- BNL (Michael): power recabling on gridftp and dcap doors; could introduce some instability for a couple of hours but requests are normally automatically retried.

- NIKHEF (Jeff): had to restart the WMS as a workaround for the current bug. Bug should be fixed.

- CERN (Gavin): new version of CASTOR SRM to fix memory leaks; in PPS now; should be tested by Atlas and CMS; will be put in prod later (when tested).

AOB:

- MariaDZ: reminder that a meeting took place on 12th March concerning the migration of OSG sites from GOCDB to OIM: not much progress. Michael says that the modification is intrusive especially for BNL and cannot be done during Atlas reprocessing which is taking place much later than expected; will be done after the reprocessing has been completed.

- Alessandro: what is CNAF status? Daniele reports that CNAF should be up around 17:00 today; they are not extending the downtime.

Tuesday:

Attendance: local(Ricardo, Nick, Gavin, Ewan, Miguel, Jamie, Luca, Jean-Philippe, Harry, Luca, Alessandro, Maria);remote(Jeremy, Michael, Gareth).

Experiments round table:

- ATLAS (Ale) - reprocessing still ongoing. Few problems noticed: 2 zero-sized files moved by FTS - should never happen! Problem with checksums in SARA. Discrepancies between SARA calculation and that stored in catalogs. Checksum calculated directly at DAQ level. Lyon - some jobs failing as some files not staged to disk - Lyon doing reprocessing test after having prestaged files so?? Observed this afternoon. This Thursday status of reprocessing will be given (on request from sites admins). Reprocessing will still be ongoing - interim report.

- CMS reports - apologies - meeting clash.

- ALICE -

> • LHCb -

Sites / Services round table:

> • ASGC (Jason): all T1 services should have been restored - more details below:

```
Hi Jamie, and Harry,


sorry for the long delay not updating the progress after the incident.
after relocating the facilities from ASGC DC to IDC, we took another week to resume the power on
also, the complex local management policy have delay the whole progress for another week.

now, all T1 services should have been restored, this including (some time stamp and services refe
slide):


- ASGC Network      : two days after the fire incident (Feb 27)
(relocate into main bld. of AS campus where we have main fiber connectivity from all other servic
service relocated into computer room of IoP are: VOMS, CA, GSTAT, DNS, mail, and list.


- ASGC BDII & core service: first week after the incident (Mar 7) services consider in first relo
 LFC, FTS (DB, and web frontend), VOMRS, UI, and T1 DPM.


- ASGC core services relocate from IoP/4F to IDC at Mar 18th. we took
around two weeks to resume power on trial outside data center area due
to the concerning dusting in the facilities that might trigger VESDA
system in the data center. majority of the system have been relocate
into rack space at Mar 25.


- T2 core services      : Mar 29
this including CE/SEs while the expired crl and out dated ca release
have cause instability of the SAM probes the first few days. we later
integrate the T1/T2 pool that all submission will turn to same batch
scheduler to help utilizing the resources of new quad core, x86-64
computing nodes.


- core services relocated from temporary computer room to IDC at Mar 31.
both 3D, Grid service DB (LFC and FTS) are consider in the same
relocation. migration of more than 20 servers have been perform in the
same removal. the service able to be restored within 4hr, while recovery
of the DB services have been delay for another four hours due to the
fabric problem. replacing the faulty devices able to restored the oracle
db engine shortly.


- T1 core & CEs services   : since last Wed., Mar 25
instability observed due to the screw of system clock and also out dated
crl & CA version. problem have been fixed shortly. other core services
are proxy server, DB (SRM/CASTOR, and 3D), VOBOX of CMS & Atlas, WMS,
RB, main schedulers, file transfer monitoring, as well as all the other
file servers (except T1 MSS).


- T2 DPM pool           : since Apr 2nd,
all 10 disk servers have been restored and resume full functioning.
except for two disk servers showing critical net phys error, other disk
servers all able to be up and functional normally. ether link oftwo
problem disk severs have been redirect to other interface available on
the servers.
```

Tuesday: 3

```
- T1 MSS (CASTOR)   : Apr 5
except for three disk servers in standby pool showing critical h/w
errors, all other pools serving production space tokens have been
onlined 2nd week relocate into IDC.


- T1 Tape system (CASTOR): Apr 8
tape library installation plan to be carried out this Wed, we hope to
deliver full functionality of CASTOR services after attaching tape
library to central resource manager with two temporary LTO3 tape drives.
the other new 6 LTO4 tape drives will be online soon in three weeks
(urgent procurement approve by the authority, as local IBM STG confirm
the tape drives should have been broken already. 12 in total, 8 with
LTO3 and 4 LTO4 install in the system before the incident).


other dual core blade servers are remain offline and expect to be online
soon after ASGC data center area fully restored.

feel free to let me know if having any inquiries.

BR,
J

--
Jason Shih
ASGC/OPS
Tel: +886-2-2789-8311
Fax: +886-2-2783-7653
```

- DB Services - LHCB online DB downtime from 15:00 till 24:00 yesterday: Problem orginally caused by power cut. Systems not brought up correctly - f/s for DB was corrupted. Additional point: streams replication stopped due to wrong apply rule - now fixed (NDGF, IN2P3, CNAF).

```
Hi Jamie,

I forward a short 'post mortem' we wrote on the downtime we had yesterday on LHCB online DB from

Cheers,
L.


-----Original Message-----
From: Dawid Wojcik
Sent: Tuesday, April 07, 2009 11:53 AM
To: lhcb-online-admins (Administrators for LHCb Online systems); Pdb Service; Maria Girone; Niko
Subject: LHCb online DB - failure summary

Dear All,

Here is a summary of yesterday's (6th of April 2009) failure of LHCb online Database:

15:04 - for an unknown reason (to be followed up with LHCb online Admins, power cut in SX8?) we l
All nodes except for lbrac01 reboot for cluster integrity (Oracle's mechanism for IO fencing)

17:55 - We get notification from LHCb Admins that there had been 2 power/network cuts and that ca

18:06 - PDB support reboots all the nodes to see if they come up cleanly.

18:28 - All is up and running (cluster operation restored), but only one FC path is up, we notify

18:32-34 - Some additional storage fails (was rebooting) and DB lost some of the disks while runn

*This accident caused the loss of the ASM diskgroup as 10g ASM cannot handle rolling reboot of st

18:55 - All the storages back in a normal state, DB is still down.
```

Tuesday: 4

Server nodes are rebooted by PDB support.

19:15 - All the nodes are back, DB cannot start, receiving ORA-0600 (internal error). After few m

19:50 - LHCb Admins are asked to notify users of current problems as PDB support cannot post to l

20:00 After some unsuccessful attempts to restart the DB, PDB support starts restoring the databa

21:30 - DB is fully recovered, however there is one issue that is potentially dangerous to open t

23:59 - Cluster is back fully operational again. The database has been recovered to the point in
*some transactions have been lost, but as the activity on the db was low the impact is estimated

00:05 (7th of April) - LHCb Admins asked to notify users of the DB being up again and that some t

Best Regards,
Dawid Wojcik

- RAL (Gareth) - sched. FTS update this morning didn't happen - resched for tomorrow. Includes delegation fix.

- CNAF (Luca) - after 1 week downtime we started to switch on on Friday am. 'Unless major problems up on Monday' - no major problems but during w/e faced issues on network. Doubled all uplinks to core switch using bonding config for all servers - some not only didn't work but caused strange behaviour causing huge delays. Discovered only on Monday morning that some servers were causing this - thought some problem on network itself. Monday morning started to debug - delay in reconfig of shared s/w area. Back only late afternoon yesterday - extended downtime until this morning. Still problems with CASTOR. Issue with CMS mainly. CMS asked us to set ACLs on all CASTOR tree - they can now read without problems but sometimes fail to write. STORM problem with new version - lost gridmap file on all servers after restart. ATLAS - submitted an alarm ticket at 11:00 and both STORM and CASTOR not working. Thought problem not received data from T0. Luca - situation improving. Ale - when can we start reprocessing? Some functional tests first... Have foreseen 48h of functional tests for DDM and also functional tests for prodsys. Luca - yesterday afternoon ATLAS-CNAF people managed to delete files from buffer to start preparing for reprocessing.

- CERN (Gav) - problem with srm public - serves non-LHC VOs (incl OPS & dteam) - appears to be due to socket leak in server.

AOB:

Wednesday

Attendance: local(Nick, Alessandro, Jamie, Greig, Stephane, Gavin, Julia, Jean-Philippe);remote(Michael, Gareth).

Experiments round table:

- ATLAS (Ale) - few issues: reprocessing - status of jobs updated in elog for all T1s. Have asked all T1s about LFC catalog with size 0 files. Due to fact that TRIUMF reported 137 entries in their LFC with size 0. Think useful for all T1s to run same query as TRIUMF - have to share with all T1s query on LFC DB to pick up 0 sized files. RAL - will submit GGUS ticket - datadisk is full. Can RAL deploy more disk or we should delete? Stephane - have to do both. Gareth - don't know how quickly we can deploy more but will follow up.

- CMS reports -

- ALICE -

- LHCb reports - all issues in report. Main thing: open ticket at CNAF for over 6 months! Good if someone from CNAF could at least comment on this! Ongoing problem with dcache at IN2P3 makes it v hard to use it. (Reports wrong locality so doesn't say if file is online correctly - check details).

Sites / Services round table:

- RAL (Gareth) - updated FTS this morning (delegation patch). OPN down from CERN end due to intervention. ATLAS - query on LFC - was request sent as ticket? Stephane - sent to a list managed to atlas of T1 contacts. Try to follow up. Gareth - submit ticket.

- CERN (Gavin) - postmortem for srm-public outage 7 April (still not understood)

AOB:

- No calls Fri-Mon inclusive (Easter) but Tue-Fri next week as usual.

Thursday

Attendance: local(Harry, Jean-Philippe, Jamie, Olof, Nick, Daniele, Simone, Julia, Andrea);remote(Michael, Luca, Gareth, Greig, JT).

Experiments round table:

- ATLAS ADCoS report◰ (Simone) - ASGC: ATLAS decided to start from a clean situation in ASGC. Asked them to remove or consider 'not existing' SE files and to clean catalog. Done. Before restarting production activities want to verify config of e.g. CASTOR pools. Like to avoid to inherit problems from past. Phone call with Jason to go through various details - not yet ready to start working with them. THis meeting: following GDB discussion ATLAS will start providing daily reports from ATLAS distr. ops shifts - 3 time zones, all fill reports. They will be appended to agenda of this meeting and person attending will report hot issues. dCache & Chimera: chimera installed in ndgf and +ve feedback from them. ATLAS would like at least one of Lyon,FZK, SARA migrating to chimera before step09, i.e. < 1st June. Extremely important to understand if new system works as expected. Tight but would like to push for this hard. Unique opportunity! Ron from SARA gave v +ve feedback from Aachen workshop. Will do more tests especially on migration process. Denial of service for CASTOR reported today - v aggressive user triggering millions of gets from castor pool. Privs removed for this user - ATLAS agrees! Maybe go further - impossible to contact - acct compromised? Remove access to lxplus etc. -> invoke procedure Olof - FIO on interactive clusters, plus also security team perhaps? Simone - no evidence that acct is compromised - wrong email forwarding + other things.. Status of reprocessing (Ale) - INFN T1 now in operation. Started first jobs around noon.

- CMS reports - Top 3 site-related comments of the day: 1. ASGC coming back online, and CNAF fully operational for CMS now (but took long): back to 7 Tier-1 sites now 2. staging issues at FZK still persist. Admins working on the problem. 3. a broken tape at FZK: 4 CMS files lost. Julia - see that backfill is facing at FZK. >50% failure rate. Daniele - not sure if this is related to FZK problems above - have to check with submitters (from US).

- ALICE -

- LHCb reports - Experiment activities
  - ♦ Ran 5000 jobs at T2s over night.
    Top 3 site-related comments of the day: 1. Wrong status set on replicated data at IN2P3 dCache (possible solution to be tested) 2. CNAF-StoRM ticket that has been open for >8

months! 3. Stalled jobs at UKI-NORTHGRID-MAN-HEP . Over hoilday w/e low level activity at T1 and T0 - sites should expect to see some jobs coming through. CNAF ticket opened > 8 months still not actioned! https://gus.fzk.de/ws/ticket_info.php?ticket=38730&from=search

Sites / Services round table:

- NL-T1 (JT) - problem: can't get alot of jobs running as all requests land on 2 disk servers - stager throughput reduced by factor of 5. pre-stage activity. Not known why - being worked on.

- CNAF (Luca) - downtime over, but as consequence have lost some of oldest disk servers. Some tens of TB... Have to reinitialize them to recover. No expts should be affected unless some files not yet migrated to tape have been lost. (CASTOR disk servers). gpfs-storm part not affected. In some ways known problem as some apparatus - switched off properly - always some probability they will not recover. Most affected expt CMS - but only 2-3 files lost on disk buffer space. Q: for team tickets, which is response time requested? Distinction between alarm and team tickets not done at the moment - sms alarms (for us) also triggered for team.

AOB:

- Daniele - currently have several T1s back. Being in STEP or not in STEP: if we see for CMS that a given T1 cannot sustain ops we try to find another T1 that could support load. All T1s for CMS except for FNAL support other VOs. If ATLAS does the same!! Simone - in ATLAS there is no T1 per activity - spread basd on MoU shares. If one T1 offline distribute according to remaining shares. Simone - propose to discuss here (or on wlcg-operations)

Friday

No meeting today!

-- JamieShiers - 02 Apr 2009

This topic: LCG > WLCGDailyMeetingsWeek090406
Topic revision: r10 - 2009-04-09 - AleDiGGi