WLCG Service Report

Jamie.Shiers@cern.ch

~ ~ ~

WLCG Management Board, 28th October 2008

Introduction

- This presentation covers 2 weeks – last week's MB was dedicated to preparation of yesterday's OB meeting
- As reported yesterday, Service Incidents that have triggered a "Post-Mortem" report (did we agree to call these "**Service Incident Reports**" from now on?) have averaged about 1/week since June
- The last 2 weeks – **and in particular the last few days** – have been well over this average
- In addition, a number of these incidents (not just those of this weekend!) have still not been fully understood or resolved. Essential to have input from all parties...

Recent Major Incidents

- This morning we were informed of a complete power outage at NIKHEF. Updates in next slide
- On Friday the CASTOR services at ASGC started degrading and were essentially unusable for ATLAS and CMS (100% failure) from Saturday on. Numerous mails on castor-operation-external (CASTOR operation issues at institutes outside CERN) about “ORA-600 Errors in Castor rhserver”
  - Still questions about required patch levels for Oracle services for CASTOR outside CERN
- On Monday, SARA announced an unscheduled downtime of the storage services, following earlier errors (also seen over the weekend?)
  - Intervention log in notes

NIKHEF Power Outage Update

- I don't have specifics on exactly what went wrong with the power (JT)
- We know why many of the services did not come back up : we had recently virtualized many of the services, but the VM configurations were not set up in a way that made it possible for them to auto-start. more concrete information in the post mortem.
- Restoration of service is continuing. we've gotten far enough that our nagios harness is working, so we can get a reasonable overview of what is not working :-) David and Tristan are chained to a console in the machine room :-)

Recent Major Incidents cont.

- On Saturday, Andrew Sansum circulated a preliminary “SIR” on an incident at RAL affecting the CASTOR service for ATLAS with a reported 55h duration
  - http://www.gridpp.ac.uk/wiki/RAL_Tier1_Incident_20081018

- FTS & VOMS services at CERN also suffered on Friday night with a 3-4 hour downtime on some channels and complete failure on others (FTS), only a 5’ interrupt for VOMS.
  - “Triggered” by security scan – follow-up to further understand and avoid; follow-up on alarm handling “got stuck”

- And finally some good news – it seems that we now have a “good patch” for a bug impacting the Oracle Streams service (several iterations – 2 years in total!)

Oracle Streams Service

- The service is designed to decouple as much as possible the different components of the service
  - e.g. “downstream capture” boxes insulate Tier0 services from streaming to Tier1 sites
  - (Logical) change records (LCRs) are kept in a queue – failure of one or more sites to “digest” this queue (dependent obviously on the rate at which it is filled…) can impact the overall service → expert “manual” procedure to split bad site(s) & “re-merge” when solved
    - The above simplified but hopefully not to the point of incorrectness…
  - It is in this area that the “bug fix” should help, allowing much simpler and “non” (i.e. less) expert intervention
- Some better understanding of the overall “real life” needs of this service as well as production (/ test?) usage at Tier1s seems called for – some load related problems in this area (ATLAS conditions: production usage? Tests?)
- [Maybe we should document the service: requirements, setup and operations in a short paper?]

Service Incident Reports

- Now have a web page (wiki) where we will keep pointers to these
- Progressively add “post-mortems” submitted since 1st June 2008 (when we started to be more systematic about it…) plus pointers to earlier reports

Site	Date	Duration	Service	Impact	Report
CERN	24/10/2008	3-4 hours	FTS	(channels)	(link)
etc.					

- These work well for incidents that occur over a relatively short period of time (hours – days) and can be promptly diagnosed
- They work less well for on-going problems – e.g. those “around” the ATLAS conditions service(s) & e.g. storage-related services at RAL, with multiple events over an extended period of time which may or may not (all?) be linked
- As was suggested at a previous MB, maybe simply keep these as “open items” with a regular update
  - Probably need to be rather generic in the description – as is attempted above – to avoid too many discussions about details
  - Take a “service viewpoint” – if it’s the same service that is affected, then its in the same “dossier”

Experiment Activities & Weekly Reports

- As noted in yesterday's report to the OB, the experiments have – judged by what is reported at the daily meetings – ramped up their activities since "09/19"
- It is probably not realistic to report on even the high-level messages at every MB, whereas just mentioning the "SIRs" is probably too high-level…
- What follows is a chronological walk-through of some of the main points affecting the services
- There are also some important experiment-related issues, such as file registration (in LFC) performance problems seen by ATLAS (being debugged, but is not seen when using LFC directly & hence is currently believed to be at DDM level)

Chronological report (1/3)

- 13/10: LHCb requested an investigation into LSF@CERN problems – post-mortem produced
- 13/10: LHCb "data integrity checks" and CASTOR@CERN service – PM produced, operations procedures to pro-actively drain servers out of warranty
- 13/10: various comments about announcements for scheduled & unscheduled maintenance
- 14/10: problems with SARA tape b/e for several hours
- 14/10: gLite 3.1 update 33 withdrawn – bdii problems
- 15/10: more downtime discussions...
- 15/10: extended downtime at RAL; CNAF squid server down

Chronological Report (2/3)

- 16/10: RAL intervention – memory upgrade problematic due to faulty module; longer than expected to import/export DBs; test of failover of FC connecting disk server to Oracle RACs gave problems
- 17/10: ATLAS & LHCb CASTOR services down ~1 hour – DBs bounced. New occurrence of “bad identifiers” in CMS CASTOR DB as seen in August.
- 20/10: ATLAS report >90% failures to RAL – Oracle error. RAL have 2 F/E m/cs for ATLAS; slow DB queries – attempt to improve load balancing went “horribly wrong” – should now be ok
- 20/10: CMS report v high load on ASGC SRM v2 server
- 21/10: discussions on ATLAS conditions, streams et al

Chronological report (3/3)

- 22/10: “transparent” intervention on Lyon DB cluster went wrong ~22:00 – impacted LFC, FTS & CIC portal. Not announced.
- 22/10: Errors seen with Oracle b/e to ASGC storage services. Experts at HEPiX “at hand”.
- 22/10: high load on dCache at SARA. Changed behaviour in gplazma? Users in multiple groups?
- 23/10: performance problems with ATLAS online-offline streams due to delete operation on all rows in a table with neither indices nor primary key.

Outstanding Issues

- Service Incident Reports expected for SARA MSS problems, NIKHEF power outage & ASGC storage problems, all seen in the past few days

- Open “dossiers”
  - The ATLAS conditions papers
  - The RAL CASTOR-Oracle condundrum

- Up-coming workshops:
  - Distributed Database Operations, 11-12 November
    - Will cover many of the points mentioned above plus also requirements gathering for 2009 (experiments’ input)
    - Pre-GDB on storage + GDB also on these days
  - Data Taking Readiness Planning, 13-14 November
    - Volunteers for non-CERN speakers & session chairs welcome!