WLCG Service Report

Jamie.Shiers@cern.ch

~ ~ ~

WLCG Management Board, 20th January 2009

Overview

- Summary of GGUS tickets last week
- Summary of outstanding or new “significant” service incidents
  - Several “DB-related” problems **– at least one critical!** (IMHO)
  - dCache problems at FZK
    - Similar problems have been seen at other sites: PIC & now Lyon(?)
  - Outstanding problem exchanging (ATLAS) data FZK<->NDGF
  - ATLAS stager DB at CERN blocked on Saturday: Savannah ticket (degradation of ~1 hour)
  - FTS transfers blocked for 16h also on Saturday: “SIR”
  - Oracle problems with CASTOR/SRM DB: “big ID” syndrome seen now at CERN (PPS) and ASGC(?)

- I got the following error on Atlas and CMS dashboard.

  DESTINATION error during TRANSFER_PREPARATION phase: [STORAGE_INTERNAL_ERROR] No type found for id : 34701226

On Stager DB:

```
SQL> select max(id) from id2type;
  MAX(ID)
----------
9.7454E+24
```

On SRM DB:

```
SQL> select max(id) from id2type;
  MAX(ID)
----------
1.7781E+20
```

GGUS Summary – Has Changed!

VO (submitted by)	USER	TEAM	ALARM	TOTAL
ALICE	2	0	0	2
ATLAS	25	20	0	45
CMS	6	0	0	6
LHCb	13	0	0	13

- After agreeing that the format of the GGUS summary was ok it was changed without prior warning.
  - Split into submitted by, assigned to and affecting.
  - The format has also slightly changed!
- The above table shows tickets submitted by the VOs

Alarm “Tickets”

- There was no GGUS alarm ticket during this period, but ATLAS attempted to use the alarm flow for the FTS problem on Saturday:
  - Sat 06:25 Kors' delegated proxy gets corrupted, and transfers start failing
  - Sat 11:16 GGUS team ticket opened by Stephane Jezequel
  - **Sat 21:28** Stephane sends an e-mail to **atlas-operator-alarm**, and the CERN CC operator calls the FIO Data Services piquet (Jan) at **21:55**.
  - **Sat 22:30** Gavin (called by Jan) fixes the problem, and informs Atlas
  - **Once again, we are hit by 'Major' bug 33641, ("corrupted proxies after delegation"), opened Feb 19, 2008.**
  - **More info here**
- **Follow-up:**
  - [DM group] Increase priority of bug **opened Feb 19, 2008.**
  - [Gavin] Reinstate temporary (huh!) workaround
  - [Jan] verify alarm flow: no SMS received by SMoD's
    - understood: mail2sms gateway cannot handle signed e-mails. Miguel will follow this up [new gateway later this month(?) will at least send subject which is ok]
  - no e-mail received by SMoDs
    - understood: missing posting permissions on it-dep-fio-smod-alarm@cern.ch, unclear how/when they disappeared.

DB Services

- **Online conditions data replication using Streams between the ATLAS offline database and the Tier1 site databases has been restored on Thursday 15.01**
  - This action was pending after the problem affecting the ATLAS Streams setup before the CERN Annual Closure (just over 1 month total)
- **There are still problems with CMS streaming from online to offline: The replication usually works fine but the capture process fails on a regular basis (once a day average).**
  - It seems that the problem is related to some incompatibilities between Oracle Streams and Oracle Change Notification features. We are in contact with Oracle Support on that but in parallel we are looking for ways to abandon using Change Notification functionality.
- **The Tier1 ASGC (Taiwan) has been removed from the ATLAS (conditions) Streams setup. This database is down since Sunday 04.01 due to a block corruption in one tablespace and cannot be restored because they are missing a good backup.**
  - A new h/w setup is being established which will use ASM (as is standard across "3D" installations) and will hopefully be in production relatively soon.
  - 💣 **The missing DBA expertise is still an area of concern and was re-discussed with Simon last Thursday**

DM Services – GridKA (Doris Ressmann)

- GridKa: we had massive problems with SRM last weekend, which we solved together with the dCache developers by tuning SRM database parameters, and the system worked fine again from Sunday till Wednesday early morning. Then we run into the same problem again. There are several causes for these problems, the most important are massive SRMls requests, which hammer on the SRM. If the performance of SRM is ok, then the load is again too high for the PNFS database. So tuning SRM only brings problems on PNFS and vice versa. This resulted in: solving one problem we run immediately into the next problem.

- Furthermore we had some trouble because one experiment tried to write into full pools. This was a kind of "deny of service" attack, since many worker nodes kept repeating this request and the SRM SpaceManager had to refuse it. After detecting this problem we worked together with the experiment representative to free up some space and relaxed this situation.

  However the current situation is that we reduced the amount of SRM threads at a time, to allow the once accessing the system to have a chance to finish. It seems that this has relaxed the situation, we still have transfers failing because we run into the limit, and reject these transfers. But it seems this is the only way to protect us from failing completely. In my opinion we should learn to interpret the SAM tests, in our case a failing SAM test currently does not mean we are down, but fully loaded. Next week we will try to optimise some of our parameters to see if we still can cope with more requests, but currently it seems better to let the system relax and to finish the piled jobs.

  We (experiments, developer and administrator) should all work together to find a solution to not stress the storage system unnecessary, tuning by itself does not solve this problem in the long run.

Summary

- **"Service incidents" happen regularly – in most cases the service is returned "to normal" relatively quickly (hours to days) and (presumably) sufficiently rapidly(?) for 2009 pp data taking (but painful for support teams…)**

➢ **"Chronic incidents" – ones that are not solved (for whatever reason) for > or >> 1 week still continue**

☝ **What can we do to (preferably) avoid the latter or decrease the time it takes to get the service back?**

➢ **N.B. there are real costs associated with the former!**

Proposal

- After 1 week (2? More? MB to decide...) or serious degradation or outage the site responsible for a service ranked as “very critical” or “critical” by a VO using that site should provide a written report to the MB on their action plan and timeline for restoring this service asap
  - e.g. the proposal to perform a clean CASTOR & Oracle installation at ASGC was made back in October and has still not been performed...