Report sent on April 30th 2010 to: wlcg-scod@cern.ch

Type of Incident: Downtime notifications impossible
Location: IN2P3-CC
Duration: Between 6 to 12 hours, then 5 days
Date: April 20th 2010 in the early morning hours to April 26th, about 11am
Author: Rolf Rumler

Description

Downtime notifications sent out late and finally, not sent at all. During investigation it was found that there were **two** consecutive incidents.

Timeline

First incident

- Somewhere between 0:00 and 6:00 in the morning, April 20th, an Oracle DB table space holding the downtime notifications got full. The corresponding log files are no longer available, so the time cannot be given more precisely.
- The only member of the portal team on site found an error message (ORA-01653) in the portal logs and informed the DBAs at 8:56am.
- The database administrators allocated more table space at around 11:30am.
- The notification system restarted working. No messages were sent to sites, neither at the start nor at the end of the incident.
- GGUS ticket 57515 was filed and reported during the daily WLCG operations meeting at 3pm, April 21st.
- A preliminary investigation started which led to an explanatory mail to scod at 16:44, April 21st.

Second incident

- At 13:06, April 21st, the DBAs announced the application of a security patch during the afternoon. While patching one of the servers it crashed and rebooted. After that the server was no longer usable. Reason for the crash yet unidentified.
- During the night the remaining three servers (for this kind of database which is not directly used by LCG) got overloaded: all possible connections were taken; response times became very slow for an unidentified reason.
- At about 20:00 a broadcast was sent out announcing the instabilities of the downtime notification mechanism. It was also said that the operations dashboard was not affected by the problem.
- On Thursday morning at about 9:30, the three servers got restarted by the DBAs. No downtime was announced for the corresponding portal functions. The batch system was restricted to those jobs which didn't need the databases affected.
- The restarted servers were back at noon but the accumulated load degraded the performance immediately.
- The fourth server got reinstalled and put into production during the afternoon.
- However, the batch wasn't reopened for the users of the database in question, the portal and underlying software were stopped because of the remaining severe performance problems. Broadcasts were impossible but the various ROD teams were contacted by other means. The operational dashboard is still up and running.
- Friday, April 23rd, in the morning all four servers got restarted (up at about 11:30) to allow migration of various database schemas to other servers, but which are on a cluster with the next generation of Oracle. One of the schemas migrated was the one for the portal.
- Still on Friday, 15:00, the migrations are finished. The operations portal restarted. It

was discovered that the database requests of the portal were incompatible with the new version of Oracle. It was decided to leave the portal offline.
- Monday, April 26th in the morning, the patches of the Oracle database are rolled back and an incident is opened at Oracle.
- The portal is back to normal operations which is announced at about 12:50.

Analysis

The first incident impacted only the downtime notification mechanism. The detection of the table space overflow wasn't automatic, neither on the level of the database administration nor on the one of the portal.

The second incident's detection suffered from the arrival of the first one. Also, the fail over mechanism foreseen is not conceived for unscheduled outages; it needs some preparation. The distribution of control – on one side the database administration, on the other the portal team – introduces delays and misunderstandings in the communication. This is underlined by the fact that the operational dashboard was available all the time, as it is based on a database local to the dashboard itself (based on MySQL), a consequence of the regionalisation of this function during EGEE-III.

In addition, key people were out of office, either still on travel back from Uppsala (EGEE User Forum) with the difficulties induced by the outage of air traffic in Europe, or on vacations (planned and unplanned), or participating in other grid events in France. Especially the database administrator with the most experience with the portal's use of databases wasn't available. The reprogramming of the interface of the portal to the database, necessary because of the version change of the DB, was severely hindered by this lack of manpower.

The cause of the second incident was the application of a defective patch from Oracle.

Impact

In what concerns LCG (and EGEE), the downtime notification mechanism was delayed for the first incident and completely out of function for the second one. Downtimes could still be declared though, as the GOCDB is fully independent for this.

The second incident pulled down all other services of the portal – except the operations dashboard as already noted and the site of the ENOC for people who know the URL – which means:

- No broadcasts
- No access to VO ID cards
- No access to ROC reports
- No YAIM configurator for VO integration into sites
- No alarm notifications and subscriptions (except those via the dashboard)
- No resource distribution overview by VO and/or sites
- No user tracking (i. e. contacting a user via the DN only)

and other less significant services.

Corrective actions

The review of the current fail over mechanism for the portal (or better, the operations toolbox) will be accelerated, especially under the light of the experience gained with the operations dashboard.

Interventions on the Centre's databases of the type which led to the incident were considered to be lightweight in the past, so they weren't integrated into the normal planning for outages. This will be changed.

The technical problems encountered weren't on the side of the portal, so no direct action seems to be needed on that side. However, a self contained solution for the portal will be studied, especially in the light of ease of installation and configuration on other sites.

The patch at the origin of the second problem is accepted as being in error by Oracle whose staff is analysing it now.

The bottom line is that some operations procedures of the site have to be changed.