Report sent on March 28th to: wlcg-scod@cern.ch

Type of Incident: server overload
Location: IN2P3-CC
Duration: About 3.5 hours (cumulative)
Date: On March 19th 2011, 6h20 to 8h00 and from about 20h40 to 22h30.
Author: Rolf Rumler

Description

SRM is heavily loaded all the night from Friday evening to Saturday morning and stalled two times that day. Second repair proved to be efficient.

Timeline

March 19th (Saturday)

- 06h20 SRM stalls.
- 06h30 dCache/SRM expert starts with repair.
- 08h00 Service back after index recreation.
- 18h30 Engineer on duty reduces load on SRM
- 20h40 In spite of that SRM stalls. Also, a LHC Alarm ticket is received.
- 22h00 SRM repaired.
- 22h30 All controls done, service ok.

Analysis

On Friday afternoon one of the dCache experts applied a modification to the service configuration as advised by dCache developers.

After the stall of SRM a dCache expert monitoring the service as a follow up on the high CPU load found the evening before, triggered a cleanup manually. SRM worked again but another stall happened in the evening. A cron job for cleanup and statistic update was put in place.

Downtimes of type “outage” were declared corresponding to the incidents. A trial to shorten the first downtime, initially declared for a much longer period than actually needed, was only partially successful due to difficulties with the GOCDB interface felt as unintuitive. An error happened with the second downtime declaration; its start time was wrongly specified as 8am but should have been 8pm; this could not be corrected. Afterwards a downtime “at risk” was declared until Sunday morning but no other incident occurred.

Impact

The dCache/SRM service was completely unavailable during the periods indicated. FTS requests implying access to SRM stalled, too. All other services showed no problem.

Corrective actions

The dCache experts stopped the SRM server on Saturday morning and did a kind of clean-up (via vacuum) and statistic (via analyze - the query planner uses these statistics to help determine the most efficient execution plans for queries) update of the postgres database. Then we restarted the server. It worked fine for about ten hours, then CPU load rose again. We therefore decided to "help" the database by making regular statistic updates of the DB (via a cron job). Since then, the SRM server behaves well.

dCache service managers are currently in touch with the dCache developers to get some advice from them about the best configuration profile we should have for our setup.

The rule not to modify production systems just before a week end has been reminded. More formal change procedures are under study.