Report sent on January 11th 2010 to: wlcg-scod@cern.ch

Type of Incident: Local batch system database server overload
Location: IN2P3-CC
Duration: 6 hours
Date: January 4th 2010 from 14h45 to 20h40
Author: Rolf Rumler

Description

Local batch management system (bqs) failure due to data base server overload.

Timeline

- 14:45 NAGIOS detected time outs in the communication to the LBMS
- 14:45 to 16:00 Research of the origin of the problem; various trials to restart the service.
- 16:00 Downtime declaration
- 16:40 Restart of MySQL demon (which cancelled the long running requests and made the data base functional again)
- 16:50 Re-authorisation of job submission commands
- 17:30 to 20:30 Restart of batch demons on the 580 worker machines
- 20:40 Re-authorisation of job scheduling

Analysis

The local batch system uses a MySQL data base to keep data on job submission and history as well as accounting. A user requested the history for all December 2009 which would have returned more than 2 million records. The answer took too long, the user cancelled the command – which didn't cancel the data base request – and reformulated his request but this time with month and day inverted (January 12th instead of December 1st). The result would have been about 24 million records, the response took even longer, obviously. These requests were done several times and finally crashed the data base service. This in turn crashed a large number of batch system demons on the workers which weren't restarted automatically.

The slowness of the data base response was aggravated by inadequate hardware (slow disks).

Impact

No submission commands and no queries for job status were possible any longer. The running jobs finished but their status could not be shown, neither during the incident nor afterwards as history updates didn't take place. No new jobs were scheduled during the duration of the incident (jobs stayed queued).

An unscheduled downtime had been declared from 4pm to about 10pm for the tier-1 (IN2P3-CC) and the tier-2 (IN2P3-CC-T2).

Corrective actions

The server processes of the data base and the batch system had been restarted.

No action taken in the long term on the local batch system for now, as it will probably be replaced by a new one in the next months and because this type of incident did not arrive in the past.