Report sent on July 13th to: wlcg-scod@cern.ch
Corrected on August 29th.

Type of Incident: software area unavailable
Location: IN2P3-CC
Duration: 36 hours
Date: June 24th 2012, 4:00AM to June 25th 2012, 11:25AM CEST
Author: Rolf Rumler

Description

Increased activity of ATLAS let overflow the file system for the logs of the SQUID servers for CVMFS at practically the same time. No external failover available.

Timeline

June 24th (Sunday)

- 04:00 Both SQUID servers for CVMFS stopped working because of a file system full condition (/var).
- 10:30 ATLAS created a TEAM ticket (GGUS 83517).
- 10:35 Ticket arrived in local ticketing system.
- 12:27 Ticket assigned to system administrators by engineer-on-duty.

June 25th (Monday)

- 08:36 LHCb submitted a TEAM ticket (GGUS 83528) for IN2P3-CC-T2.
- 08:40 Ticket arrived in local ticketing system.
- 09:10 Local ticket redirected to LHCb support of the site.
- 09:45 GGUS 83517 transformed into an ALARM ticket.
- 10:09 Another ticket from ATLAS about CVMFS problems.
- 10:21 After discussion with system administrators, user support puts ALARM ticket to “in progress”.
- 10:25 Reassignment of 83528 to system administrators by LHCb support.
- 11:15 SQUID servers restarted.
- 11:25 CVMFS up and running.
- 13:16 ALARM ticket put to “solved”.
- 16:42 LHCb ticket 83528 updated with solution and closed.

Analysis

Increasing activity of ATLAS since beginning of June made that the logs on the SQUID servers were growing two times faster than in the past. The frequency of the cleanup mechanism for them was too low. A file system (/var) filled up because of that and the SQUID server crashed. The backup server hit the same condition, the cleanup process succeeded but too late for the SQUID process with the result that the machine was up but SQUID didn't run.

IN2P3-CC assumed erroneously that there was a fail over mechanism foreseen in the CVMFS clients on the worker nodes when the local SQUID servers become unavailable.

The criticality of the incident could not be appreciated correctly by the engineer-on-duty as the mentioned fail over to other sites was thought to be working. Consequently a failure was interpreted as having its origin outside of IN2P3-CC.

Impact

All jobs of ATLAS and LHCb referencing the software area in CVMFS crashed or did not start correctly. The other VOs, not only the remaining LHC ones, occupied the such freed resources which led to an overload of dCache by CMS.

Corrective actions

The log areas were freed manually and the servers restarted. The “logrotate” frequency was increased from four weeks to five days to cope with the higher activity of the experiments.

Note on text revision after initial publication

The initially published version of this report contained an erroneous assumption about the function of CVMFS which consequently led to wrong conclusions.