Report sent on March 1st to: wlcg-scod@cern.ch

Type of Incident:
Location: IN2P3-CC
Duration: 3 hours
Date: February 13th 2011 from 01h04 to 03h50
Author: Rolf Rumler

Description

Core network switch outage due to CPU card failure.

Timeline

- 01:04 Local network monitoring system detects outage, NAGIOS starts to signal problems with various services (dCache, HPSS, NIS, batch/BQS with intermittent errors, ...). SMS sent to engineer on duty.
- 01:50 Robotics down.
- 02:00 Automount down on about 650 workers.
- 03:50 CPU card changed, switch back to operation, batch system again fully operational.
- 05:00 Robotics back to operation.
- 10:00 HPSS has to be taken down completely.
- 13:30 NIS back to operation (major part already before 11am), automount back on all 650 workers.
- 14:00 HPSS restarts, fully functional.

Analysis

A CPU card of a local core network switch failed permanently.

Impact

The failing switch was responsible for connecting about 50 percent of the site's storage servers. Correspondingly, all storage services were more or less impacted; dCache was hit only indirectly by accumulating a lot of connections on one of its servers which normally would have been more fairly distributed.

As the mount server for the robotics was disconnected (ACSLS), HPSS had problems with tape migration and staging, in addition to missing disk servers. Here manual intervention was necessary to bring the service back.

The network failure revealed a problem in NIS. This in turn made impossible job submission and job start, but had only an impact on job execution for jobs needing other (storage) services. Though the exact number of jobs having suffered from this is unknown, the about 11 000 jobs running at the time of the incident did not decrease significantly more than which would have been expected by normal job ends.

Corrective actions

The major action was to replace the failing CPU card. This was possible rapidly because a proper card was on site, the engineer on duty was by chance a network administrator, and he knew how to replace the defective component.

To cope with a potential re-occurrence of the NIS problem, an automatic restart procedure for this service has been installed on the NIS slaves.

HPSS and automount problems were handled manually and no special action has been taken beyond this.

Additional redundancy for the switch is not planned at the current stage, as the incident is considered as extremely rare.