Report sent on April 14th to: wlcg-scod@cern.ch

Type of Incident: power cut
Location: IN2P3-CC
Duration: 45 minutes (power), about 5 hours for total recovery
Date: April 8th 2011, from 11h45 to 17h00 CEST
Author: Rolf Rumler

Description

An external power switch was turned off. The remaining backup power supply was just sufficient to keep the machines running but not the cooling. Rising temperature required stopping most of the machines.

Timeline

April 8th (Friday)

- 11:45 An external power switch was turned off. The quickly rising temperature required a manually triggered shutdown / power off sequence for most of the batch system's worker nodes and an automatic emergency shutdown for some servers. Only 15 percent of the worker nodes stayed available.
- 11:55 Announcement of an unscheduled downtime from 10:00 UTC to 14:00 UTC (12:00 to 16:00 local time).
- 12:15 External power was back.
- 12:30 Cooling restarted
- 14:50 Production restarted with a limited number of machines.
- 15:30 VObox for ALICE back into service.
- 17:00 Production back to its normal level.

Analysis

Scheduled work on the electrical network for the new additional machine room required stopping its external power supply. The batteries in place to keep up the remaining external power supply, also for the old machine room, failed prematurely because the required infrastructure in the new building to recharge them was not yet ready. The generator of the old building took over but stopped after a minute. The additionally installed batteries sufficed to keep up workers and servers, not the cooling. The temperature in the machine room raised to critical levels. To avoid damage, especially to storage devices, most of the worker nodes were powered off manually following a proper shutdown sequence, except the remotely hosted ones which stayed up (at the CINES centre of Montpellier).

Some server nodes, all in the same rack, powered off themselves. Grid services concerned were: several VOboxes, some LFCs, some lcg-CEs and a CREAM CE. A back end server of the Operations Portal stopped, too.

External power supply was re-established rapidly and cooling with it. Worker nodes restarted generally without problems after the manually triggered regular shutdown / power off sequence. Servers came back, too. The delay between power return and production restart is due to the time needed to cool down the machine room and the machines and also to verify that every machine restarted correctly. The delay between production restart and return to normal levels is due to repair work on machines with problems.

Impact

All running jobs were lost except those on remote workers, most of the grid services – except for data transfer – went down.

Corrective actions

The immediate solution has just been presented. The interactions between the old and the new electrical network, especially in case of maintenance of one or the other, need further analysis.

Tests of the generator before and after the incident didn't reveal any problems which means that we have to plan for a deliberate power cut one day, to check the correct function of the generator under load.