Report sent on September 12th to: wlcg-scod@cern.ch

Type of Incident: cooling system failure
Location: IN2P3-CC
Duration: 1.5 hours (cooling), about 7.5 hours for total recovery
Date: August 26th 2011, from 3:26AM to 11:00AM CEST
Author: Rolf Rumler

Description

The restart of the cooling system after a long lasting external power cut overloaded a security breaker in the newly constructed machine room. The resulting high temperature lead to an emergency shutdown of most worker nodes.

Timeline

August 26th (Friday)

- 00:20 The external power supply failed. The Centre's electrical generator for the old machine room and batteries for the new one took over, the container with chilled water replaced the cooling generators for the new machine room. Cooling for the old one continued via the power supplied by the generator.
- 01:50 External power is back but technical alarms about the UPS system are still seen.
- 02:30 Sensors for temperature reach “warning” level.
- 03:20 Cooling failure is identified.
- 03:26 High temperature is detected in the new machine room and starts automatic shutdown procedure. At this moment the incident becomes visible for the users.
- 04:15 Cooling generators restart.
- 05:00 Temperature becomes acceptable.
- 05:17 Declaration of unscheduled downtime
- 05:30 Power supplies of several racks are found to be damaged.
- 06:00 Handover to usual operations team.
- 09:30 Batch system restarts.
- 11:00 Batch fully operational, except the damaged racks.
- 11:15 End of downtime.

August 27th (Saturday)

- 10:00 After recharging the batteries, the power supply for the new machine room is again redundant.

Analysis

The external power cut is detected automatically and the fail over to electrical generator and batteries works as designed. The chilled water reserve is used to replace the cooling generators which are not on the UPS, as designed; this helps to preserve battery capacity for CPUs and storage devices.

When power comes back the UPS system signals a defect; this is handled by the engineer on site. He finds out that power is missing for the cooling system and especially, that this is due to a temporary overload of a breaker when power came back and the whole cooling machinery tried to restart at the same moment. It takes some time to restart the cooling system without overloading this breaker. The result is that the temperature alert is given too late for a proper shutdown, also because the monitoring infrastructure in the new machine room is not yet complete.

However, the emergency power off works correctly so that no machines and disks are damaged. The defective power supplies mentioned seem to have failed before the emergency stop, this is still under investigation.

Impact

All jobs running on workers in the new machine room were lost. This means, all jobs running under GridEngine and 80 % of the jobs running under BQS, in numbers 3500 and 4000, respectively. They were either re-run or cancelled, depending on their profile.

Corrective actions

The immediate solution has just been described. The breaker, under-dimensioned for the high temporary load during a power up of the whole cooling system, has been reconfigured correspondingly. Other breakers and power supplies damaged by the power cut have been changed accordingly, too.

The incident handling procedures will be more formalized with respect to the priority of temperature alerts versus other ones.

The already projected monitoring infrastructure of the new machine room will be put in place.

The power outage lasted longer than foreseen in the contract with the external provider who agreed to a meeting to discuss the reasons for this and to take necessary actions.