Report sent on February 25th 2010 to: wlcg-scod@cern.ch

Type of Incident: Lost network connections of worker nodes
Location: IN2P3-CC
Duration: 4.25 hours
Date: February 15th 2010 from 14:12 to 18:30
Author: Rolf Rumler

Description

All worker nodes lost network connection at nearly the same time.

Timeline

- 14:30 LBMS main server found to have a huge amount of pending connections
- 14:37 The automatic take-out mechanism for stalled workers starts to eliminate the first workers from the list of available machines.
- 14:45 System administrators signal a large amount of worker nodes having lost sshd
- 15:45 Downtime declaration
- 15:50 Mass reboot of worker nodes starts
- 18:00 After various verifications, about 90 percent of the worker nodes are back to production
- 18:30 End of downtime

Analysis

The logs of every impacted worker node showed that between 14:12 and 14:23 a signal USR1 has been sent to a bunch of processes, namely acpid, rsshd, sshd, atd, crond, ypbind, bqs, ipmievd and others.

This looks like either a direct human error or an indirect one via an ill parametrised automatic procedure.

Impact

All jobs running at the moment of the incident can considered to be lost. No new jobs scheduled before about 16:00.

An unscheduled downtime had been declared from 3:45pm to 6:30pm for the tier-1 (IN2P3-CC) and the tier-2 (IN2P3-CC-T2).

Corrective actions

Reboot of all worker nodes.

This is an isolated incident but with an important impact. Actions in the long term may include finer logging, especially of cluster wide tools like rssh, and deployment of a monitoring and restart system for basic system processes.