Report sent on Mai 5th 2010 to: wlcg-scod@cern.ch

Type of Incident: Distributed file system outage (AFS)
Location: IN2P3-CC
Duration: 17.5 hours
Date: April 26th 17:00 to April 27th 10:30
Author: Rolf Rumler

Description

The site's distributed file system, AFS, crashed after the overload of one of its servers.

Timeline

- April 26th, 17:00 Many processes on virtually all worker nodes and servers stalled. Operations stopped job initiation immediately (i. e. the batch is stopped).
- 17:30 AFS was identified as being at the origin of the problems. AFS administrators got involved. A downtime was declared until next day, 10:30.
- A very high request rate had been issued against one of the AFS read/write servers on behalf of one specific user group. Administrators stopped all processes with pending connections and started to move the files of the mentioned user group to a less critical server. The ATLAS software installation area, hosted by the crashed server, was also moved elsewhere.
- 22:00 File moves finished, servers back: AFS up and running.
- Various other services got verified and restarted if necessary.
- April 27th, 0:30 Batch processing is re-opened but site still in downtime and running with reduced capacity. Only jobs already queued on site and jobs of local users can run. Jobs of the user group at the origin of the high load on AFS are locked out.
- 10:30 Site back to normal, downtime ended.

Analysis

AFS knows two types of file spaces, read only ones and read/write. The latter one is obviously more complicated to handle in a distributed environment. At the origin of the incident was a very high request rate from one particular user group against an AFS server which had to deal with write requests, too, especially also for the ATLAS software installation. For this reason overloading this server not only interrupted the service for the originating user group but also for various other ones, including ATLAS.

Impact

No logins were possible and no job submissions as AFS is used during login and the batch commands are located in AFS. Various other services got stuck for similar reasons. Load balanced servers which used AFS were no longer available. Not finding any available server, the load balancing mechanism which is integrated into the DNS no longer recognized several aliases for basic services. This resulted in a complete batch system outage and the unavailability of grid services like BDII, LFC, and FTS.

Corrective actions

The files of the user group at the origin of the load were placed on a different server. The ATLAS software installation was also placed on a separate server.