Service Incident Report for KIT/GridKa cooling system failure on Saturday, July 10 and subsequent problems and failures.

Description

A complete failure of the computing centre's cooling system caused a failure of several local and regional services which could be restored within different periods of time.

Impact

Local and regional Grid services and VO boxes down for several hours.
CMS dCache down for several hours.
Some dCache pools down for several days.
All compute nodes down for 2 days, part of compute nodes down for 4 days.

Timeline (all times in UTC, local time is UTC+2)

10.7. 14:30 Cooling system going down.

10.7. 15:30 First services and compute nodes begin to fail. Nagios was one of the first servers that failed. No alarm has been sent by Nagios.

10.7. 16:45 Network engineers receive alarm from a router.

10.7. 17:30 Network engineer at compute center, calling other on-call engineers and site manager.

10.7. 18:00 Site manager and additional engineers on site.

10.7. 18:30 Technicians of the infrastruture department give OK to power on important systems. 2 out of 4 chiller systems working. System experts for Grid services, storage, databases etc. called.

10.7. 19:00 All required experts on site. Starting to restore services.
CMS dCache headnode down with disk failure.
One storage controller dead (800 TB of CMS dCache pools).

10.7. 20:35 FTS and LFC up. LHCb 3D DB is up. Atlas 3D DB still down with missing SAN block device.

10.7. 22:10 Atlas 3D DB up. Streams latency is ~3h, sync within 10 mins.

11.7. 8:30 Found additional storage controller which failed during the night but was still working in the evening after cooling has bee restored. Several disk enclosures found to be down. Atlas, CMS and LHCb pools affected.

11.7. 12:45 Most dcache pools up. Some raid-groups are in rebuild state. Associated pools disabled for the time of the rebuild.

12.7. 6:00 Starting up some compute nodes.

12.7. 6:15 One of three FTS frontend servers found down with both redundant power supplies dead. Associated channel agents started on remaining two hosts. (Subsequent damage caused by overheat on Saturday?)

12.7. 9:00 All PBS queues online, except CMS, Atlas, Alice (due to remaining SE problems).

12.7. 11:15 Atlas queues online.

12.7. 12:30 Alice queues online.

12.7. 13:00 3 out of 4 chillers working. Powering up compute nodes with best compute power per watt ratio, worth 30000 HEPSPEC'06.

12.7. 18:00 Broken (CMS) disk controller successfully replaced. Associated raid groups and dCache pools up.

13.7. 6:00 Writing to LHCb dCache fails.

13.7. 6:50 LHCb dCache up again. PNFS DB had temporarily lost a disk during cooling incident. This disk was then mounted read-only by the system.

13.7. 12:30 CMS queues online.

14.7. 7:55 LHCb reports SAM test failures / access to certain file in dCache not possible (GGUS 60087). Problem cannot be verified by KIT people since other files used for monitoring work. Got some example file from LHCb, then problem verified.

14.7. All cillers working but analysis of chiller problem not finished. Technical defect cannot be excluded. Outside temperature is reaching 37C again. Some compute nodes stay off to minimize load on the cooling system and minimize the risk of another emergency shutdown.

15.7. 5:40 Found problem with LHCb dCache: a postgres DB table of SRM was corrupted and needed to be re-indexed. Re-indexing started immediately.

15.7. 7:25 LHCb dCache online.

15.7. All chillers working. Powering up remaining compute nodes.

Analysis

The cooling system of the compute center hosting the GridKa resources and services went down at an outside temperature of 37.5C. The system is designed to provide full cooling power up to 40C outside temperature. It is supected that noise-protection walls caused a heat accumulation and the temperature climbed slighly over 40C, thus the cooling system reducing the power.
The temperature of the cooling water climbed, putting even more “load“ on the chillers which in the end caused a complete failure.

The alarm workflow between the KIT central alarm unit and GridKa on-call engineers did not work. This caused a delay of at least 1 -2 hours for GridKa people to react.

Water cooled racks with servers, storage systems and disk enclosures opened the doors as expected to “switch“ to air cooling. The cooling failure also affected air cooling, i.e. the room temperture got also critical after a while. Many disk controllers, hard disk enclosures and disks died due to overtemperature. Automatic shut-down failed in some cases, leading to the long-lasting problems and necessary raid-rebuilds as described in the timeline.

Follow up

We will aim to get direct access to the alarm messages from the cooling system, to feed them directly into our Nagios, bypassing the KIT central alarm unit.

We monitor the water temperature on the input side. If the temperature is climbing more than two degrees, GridKa on-call engineers get an alarm to their mobile phones. This has already been implemented on July 14.

An additional Nagios server will be deployed. The two Nagios servers will monitor each other. (The additional machine is actually ordered since some weeks.)

We will follow up the hardware defects of our storage systems with the manufacturer. These systems are required to shutdown automatically before any damage due to excessive heat occurs.