Service Incident Report for the Cooling and Power problem at PIC on January 22nd 2012

Outage start: 22-Jan-2012, 14:50 UTC
Outage end: 22-Jan-2012, 20:30 UTC

Description

NOTE: PIC has most of its computing infrastructure in a 125m2/200KVA machine room, referred to as **"the main computing room"** in this document. About 80% of the Worker Nodes are located in a container-like machine room (25m2/80KVA) in the basement, referred to as **"the module"** in this document.

On January 22nd, at 14:50 UTC, the cooling system at PIC main computer room stopped unexpectedly, most probably due to a electric supply incident. At 15:30 UTC PIC started an ordered fast shutdown of the most critical services.

Due to a problem with the supplier electricity quality, a pumping engine was broken in one of the cooling machines which caused a major failure in the whole cooling infrastructure and also activated a leak current detector which cut the electrical supply in the module.

Impact

Most Tier1 services were stopped to avoid computing room overheating. Running jobs and ongoing transfers at the time of the shutdown were lost. Services were restarted about 5 hours later, when cooling could be restablished.

Time line of the incident

All timestamps are in UTC.

22nd Jan 2012:

- 14:50 - Manager On Duty receives a temperature alert on the cell phone and checks that is not a false alarm. Monitoring shows some of the nodes in the module were down. Manager On Duty escalates the problem to the Infrastructure and Services teams responsibles.
- 15:00 - The Infrastructure Responsible (IR) phones the UAB Security department (which also monitors building infrastructure) to check if there is any problem detected with the cooling or electrical infrastructure. They have informed the maintenance team about several alarms on the cooling system in the building where our PIC datacenter is located.
- 15:20 - The IR goes to PIC and confirms that the cooling system is not working in the main room. The temperature is already very high.

- 15:30 - The maintenance technician arrives and informs that 20 minutes are needed to check the installation before starting the cooling system again. At this time decision is taken to shutdown PIC main services, to avoid overheating risk.
- 15:40 - An Unscheduled Downtime is declared in the GOCDB to inform the VOs.
- 16:10 - The Manager on Duty arrives at PIC. The shutdown process is completed.
- 17:00 - The maintenance technician starts the cooling system.
- 17:30 - IR and MoD start bringing up services. The process is completed at around 21:30 when the downtime is declared as finished in the GOCDB.
- 22:15 - Problems with some storage pools are detected. The problem is traced to one of the two controller pairs of the DDN storage system being off. Decision is taken to wait until the next day in the morning to fix this.

Monday Jan 23rd 2012:

- 04:29 - ATLAS opens a GGUS reporting transfer problems at PIC. This is related to the DDN pools that were not switched on (https://ggus.eu/ws/ticket_info.php?ticket=78470)
- 07:45 - Manager On Duty powers on the second DDN cabin which was powered off.
- 13:16 - LHCb opens GGUS ticket reporting that pilot jobs abort at PIC CEs (https://ggus.eu/ws/ticket_info.php?ticket=78487). The cause was that, after the power cut, the batch system did not recovery properly due to the large number of WNs offline and killed jobs. The problem is solved and GGUS closed in one hour.
- 15:15 -The Oracle Streams for LFC LHCb are back in sync and lfclhcb.pic.es is re-enabled in the LHCb production system.

Analysis

Besides the 5 hours of Tier1 service black out due to the power&cooling problems, some of the services remained in “degraded” mode for more time for different reasons:

- *Storage service:* About 45% of the Tier1 disk capacity is provided by a DDN system containing two S2A9900 controller couplets. After the incident, when all PIC services were restarted, one of the couplets was not powered on by mistake. This affected 7 dCache pools which remained unavailable until the next day, Mon Jan 23rd, at 7:45 UTC. The affected pools per VO were:
  - ATLAS: 3 pools, 343 TB
  - CMS: 2 pools, 224 TB
  - LHCb: 2 pools, 144 TB
- *Computing service:* 133 out of 306 Worker Nodes had to be reinstalled after the sudden power cut in the module. This was due to the fact that the network driver was not properly configured to load at boot time. This issue had two consequences:
  - Due to the large number of WNs down and the dead jobs, the Torque/Maui server suffered instabilities until Monday Jan 23rd around 14:00 UTC.
  - In some of these re-installs, the update procedure failed and the nodes booted with an old (vulnerable) kernel. This was detected and fixed on Jan 25th at 13:00 UTC.

- LHCb LFC: There were two issues with this service:
  - After the service stop and restart, the LFC Streams replication did not start properly (in fact it was not running at all). That was because the side effect of running 10g/11g replication environment. CERN had to resync two tables that became inconsistent after abnormal DB stop. After doing this, on Monday Jan 23rd, at 15:15 the Streams were back.
  - On Jan 24th in the morning LHCb reports MC jobs failing to upload the output data. The problem is traced to be in the comunication WN-LFC. In the end, the problem seems to be in the network configuration of the LFC host, which is configured with MTU=9000. When this number is changed to 1500, the problems are fixed. A clear explanation of the issue is not found (MTU=9000 was configured in the LFC host since months) and further investigations are started to try and reproduce the problem in a controlled enviroment.

Follow-up actions

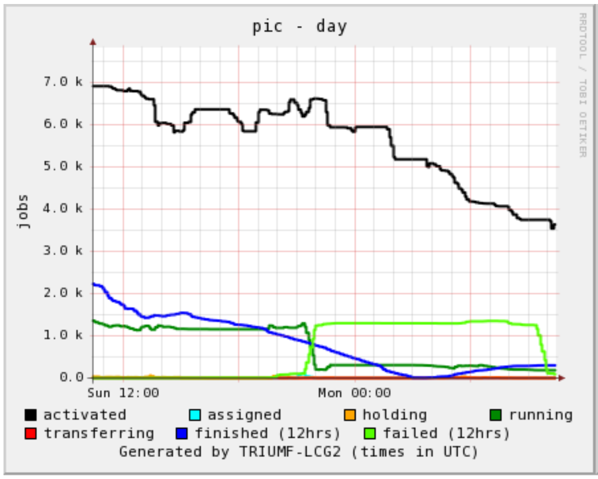
- Reparation of the pumping engine which broke. Building maintenance team.
- Change threshold settings for module RCCB, from 0.5 A to 1 A.
- Review of the automated stop script.
- Review of the ordered startup documentation: Torque/Maui part should include procedure to “purge” eventual killed jobs and unreachable WNs.
- WNs monitoring: improve sensor to report on out-of-date packages (kernel).
- Investigate the MTU=9000 issue in the LFC causing connection failures from WNs.

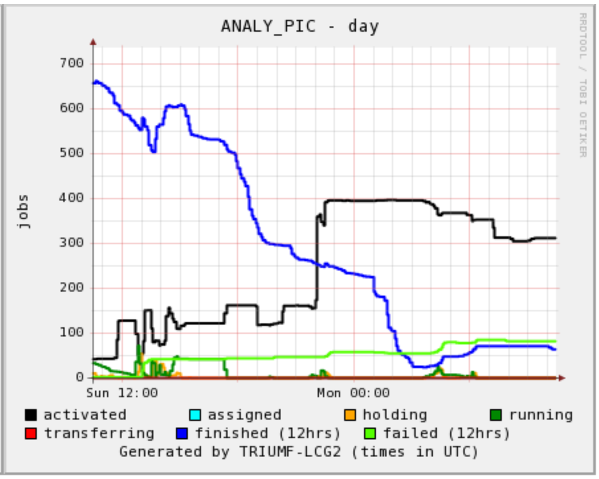
Hardware failures after the incident

- Mainboard and Hard Drive of one workernode (td547.pic.es)
- RAID controller battery failure in two storage pools (no service impact).
- One SFP+ fiber optic transceiver failure in a storage pool group uplink (no service impact).
- Mainboard failure on the installation server (the equipment used for OS installation).
- Mainboard failure on one of the ATLAS Tier3 gridftp doors, dcgftp09.pic.es (no service impact).

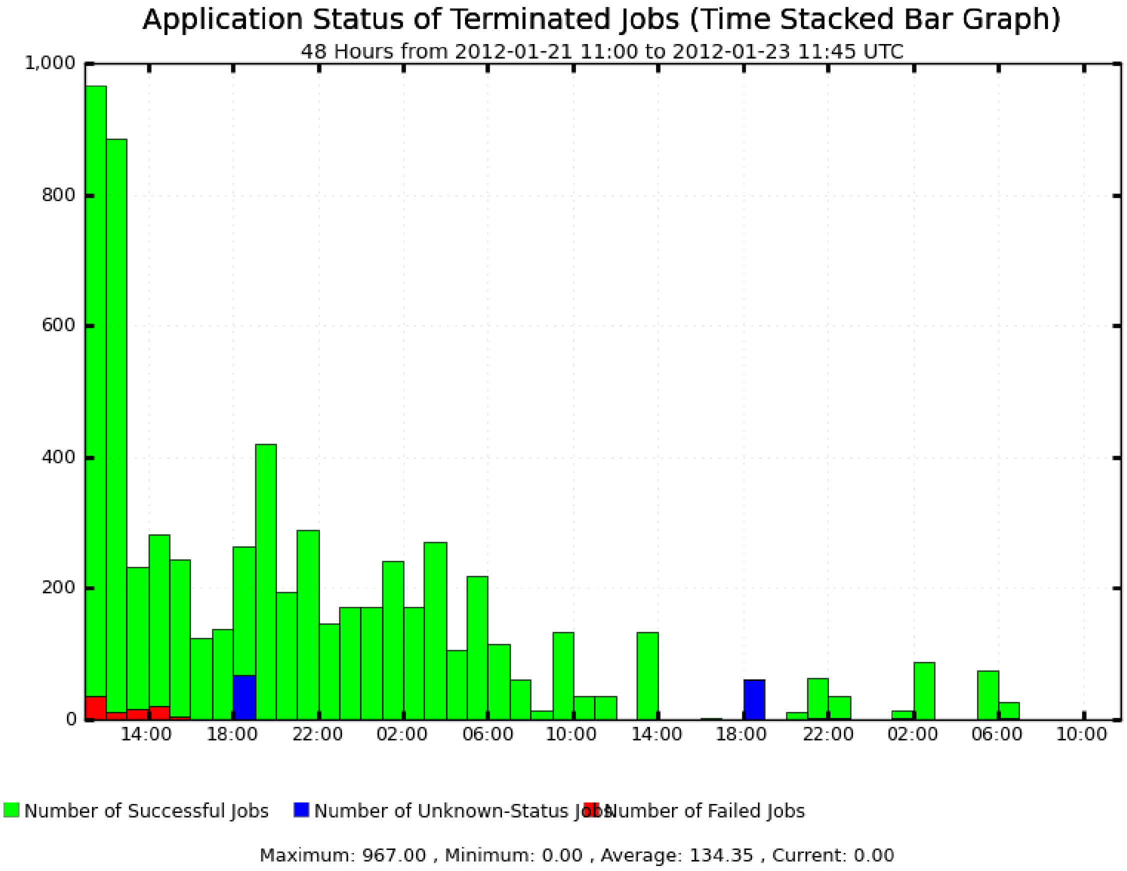
Plots

Next plot shows the ATLAS panda monitoring for PIC. Left: production jobs, Right: analysis jobs. The effect of the incident can be seen as about 1200 production jobs failing at the power off time.

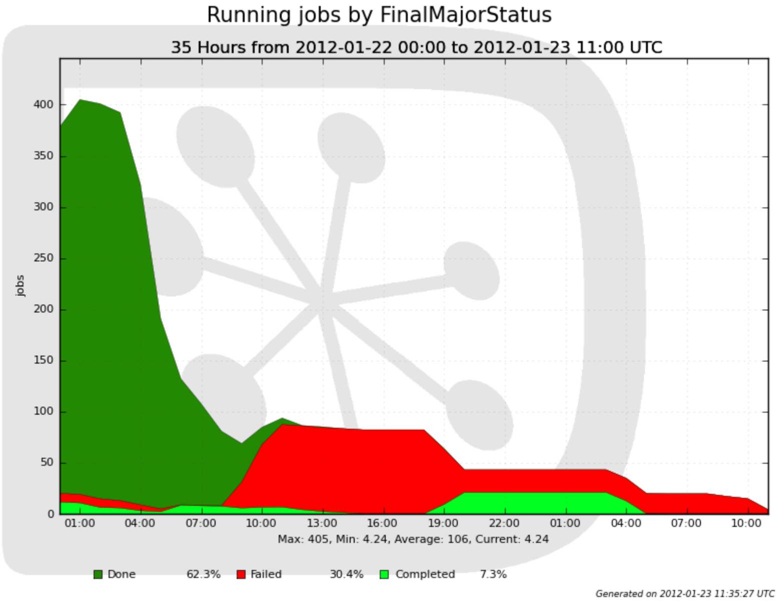




Next plot displays the number of CMS finished jobs at PIC in the 24hrs including the incident. We see there was almost no job activity for CMS.



General LHCb plot:



On the start of the incident LHCb at PIC was on 25% of job running capacity, with MC simulations workload. The jobs were stalled. PIC SE was banned by the LHCb operations team to avoid new production submissions. About 18H the incident was finished at PIC, some jobs were finding the SE was not available so changing to “complete” others still continue in stalled. Later the SE was unbanned and some jobs were “done” success, most jobs were not resubmitted since the MC production was finished.