Power Cut at PIC Tier1

Date start: 21st May 2010 at 13:40 UTC
Date end: 22nd May 2010 at 9:00 UTC

Description

On Friday 21st May 2010 at 13:40 UTC PIC building suffered a power cut that caused the cooling system to stop working. Due to this, the complete site had to be stopped to avoid overheating in the machine room.

The cause of the power cut was an incident in one of the electric panels of the building. In particular, the one that powers the new module for PIC machine room upgrade. A technician was installing a component for protection against lightning in the panel, to follow regulations. When the operators connected back the panel to the main power, this new component exploded.

In some way that the detailed technical reports still have to clarify, the explosion originated a severe short circuit situation that caused the correct activation of the protections of the building, cutting the power input from the main grid and from the diesel engines. The UPS systems kept working, powering the whole IT equipment. However, being the cooling stopped, the temperature in the machine room started to raise. To avoid overheating, PIC started the procedure of ordered shutdown, which stopped the complete site in about 5 minutes.

Few hours later, after confirming that the power panels and the cooling system had not suffered any damage in the explosion, the power and cooling of the building were reestablished. In order to minimize the load on the cooling system, the PIC Director takes the decision to keep the site off until Saturday morning.

On Saturday 22nd at 7:00 UTC the PIC Manager on Duty starts the ordered power on of the PIC site. Around 2 hours later, all of the services are back up and running.

Impact

The whole PIC site is off. All of the PIC Tier1 as well as IFAE ATLAS Tier2 services affected by an outage.

Time line of the incident

Friday 21/05/2010:
- 13:40 UTC: Explosion in the module power panel. The lights of the building go off and cooling stops in the machine room. All the machines still up and running using UPS.
- 13:45 UTC: Fast shutdown of the PIC site started (running jobs and ongoing transfers killed) to avoid machine room overheating.
- 13:50 UTC: PIC site completely off.
- 18:00 UTC: After confirming that power and cooling had been reestablished, the University starts to power on some of their basic services.
  - Decision taken to keep PIC down over night until Saturday morning, to minimize the load on the cooling system.
  - The router and DNS at PIC are powered on and the alias to the LHCB-DIRAC lhcbweb.pic.es can be redirected to the off-site backup server at UB, following the emergency procedure.

Saturday 22/05/2010:
- 6:45 UTC: Manager on Duty (MoD) at PIC. Finds the access door to the machine room closed and the electronic access control not working.
- 7:05 UTC: MoD finds an alternative method to open the machine room door (screwdriver) and begins the start up procedure.
- 9:00 UTC: All of the Tier1 services up and running. Minor issues remain which are discussed in the following "analysis" section.
  - Activity from ATLAS and CMS is reestablished. Issues are found with the SGM software installation mechanism that prevent the LHCb jobs activity to start.
- 16:00 UTC: Software installation service in the Computing service is restored at PIC after fixing the SGM WN.
- 17:30 UTC: PIC back in the production mask for LHCb jobs after the software installation jobs finished successfully.

Sunday 24/05/2010:
- 11:00 UTC: PIC Storage Element unbanned for LHCb. The operation person on duty forgot to do it after the service was restored on Saturday morning.

Analysis

- LHCB-DIRAC service
  - The fast shutdown of the PIC site after the explosion and power cut emergency is executed too fast so the procedure to redirect the DNS alias of lhcbweb.pic.es to the off-site backup at UB can not be executed.
  - On Friday 21st May at 18:00 UTC, once the power and cooling are reestablished, the router and DNS of PIC are powered on and the procedure is followed to point the alias of lhcbweb.pic.es to the off-site backup. Therefore, from 13:50 until 18:00 UTC the attempts to connect to lhcbweb.pic.es service were failing.

- Computing service restart:
  - Network problems in some WNs (about 15%). The problem only affects to those bladecenters with 10GE uplinks. The origin of the problem is found to be that the proper network driver had not been compiled after the last kernel upgrade.
  - Three WNs of another bladecenter (hpblade182, with 16 WNs) do not start. Origin of the problem found to be in a faulty power supply. One of the WNs affected is the software installation (SGM) WN. After spotting this, the SGM is powered on after powering off another one of the WNs in the box. The software installation service is reestablished this way at around 16:00 UTC.

- Storage service restart:
  - PNFS needed to be rebooted up to 3 times to correctly mount internal SSD disk. May be due to too fast power-off. This represented a minor delay in starting up the dCache Storage service.
  - After start up, the Storage service was degraded for a short period ("puts" were working ok, but most of "gets" were failing). This was quickly found to be a sort of "feature" of pool cost calculation in dCache. Configured with a percentage cut-off, when starting from zero activity, the pool cost calculation started lots of pool-to-pool (p2p) copies which overloaded the system and prevented real "gets" to be served.

When this was discovered, p2p copies were disabled and the service was reestablished.
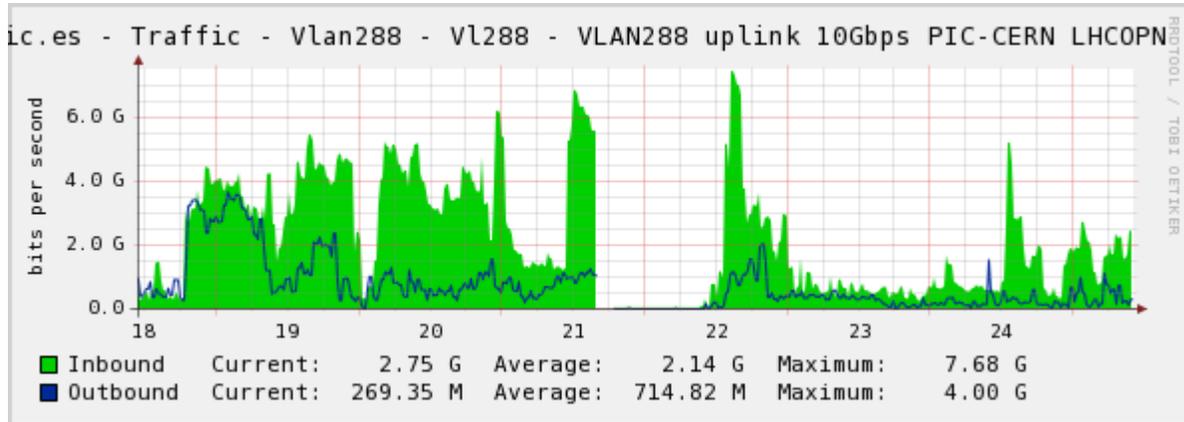
Figures

Fig 1: Weekly network traffic in the PIC OPN link. The power off time interval (in CET) is clearly seen. The high network traffic taking place when the power off occured was an ATLAS massive data replication that was taking place. After service start up on Saturday 22nd it can be seen that the data replication resumed at very high rate and backlog could be recovered.
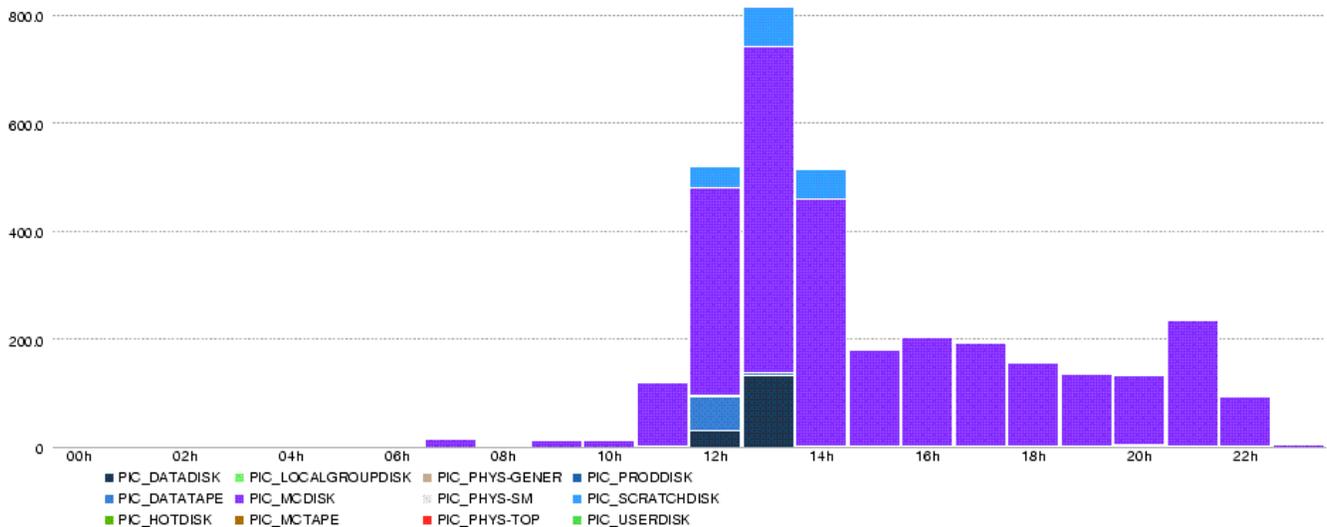

Fig 2: Detail of the hourly import data rate to PIC (in MB/s) on Saturday 22nd from the ATLAS DDN dashboard.
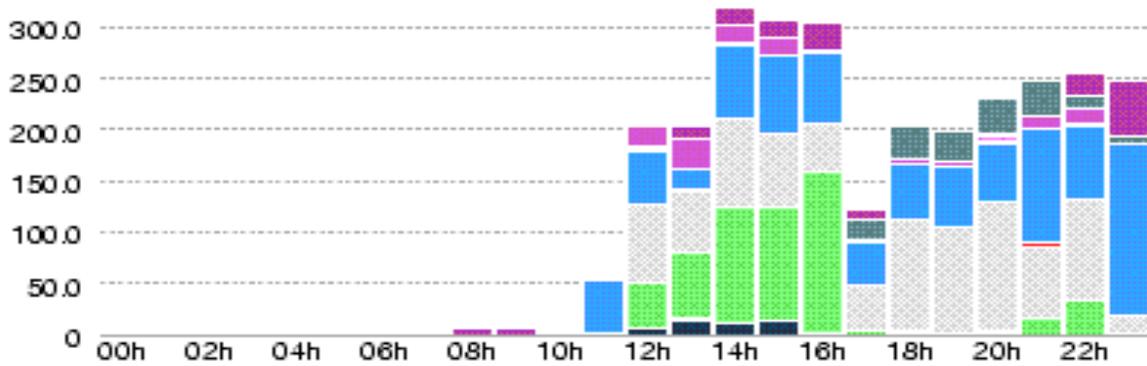
Fig 3: Detail of the hourly export rate in MB/s from PIC to the ATLAS ES Tier2 cloud showing that the data distribution activity was resumed shortly after the service recovery with high performance.

Follow up actions

- Receive detailed report from the electricity installation company doing the works clarifying the technical details of the incident and the event chain causing the power cut.
- Analyze and establish a procedure to open the door of the machine room even if the default electronic access control is not working.
- Improve automation of PIC fast shutdown procedure to ensure all steps are followed (lhcbweb DNS alias point to off-site backup) and all servers are gently shutdown (PNFS).
- Analyze and improve the automatic procedure for kernel upgrades in the WNs, to ensure network drivers and the like are recompiled when needed.