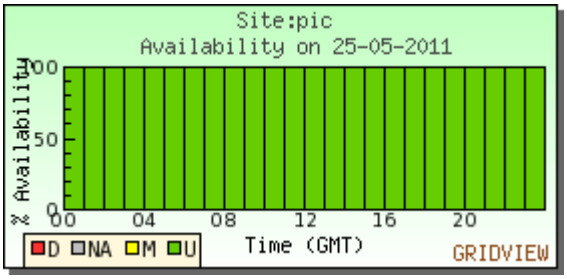
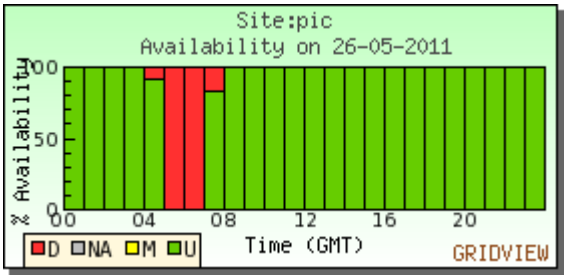
Post Mortem Tier-1 Service Incident Computing Service Incident SSC5 20110525

De WikiPIC

Incident Start: 25th May 2011 18:00 UTC

Incident End: 26th May 2010 06:55 UTC





Contenido

Description

The incident was caused by collateral effects of the security challenge SSC5 (email from security officers received on Wed. 25/May around 15h UTC). As a result of malicious jobs tracking and compromised DNs found, the stipulated protocol recommend to unplug the network of the affected nodes. There were about 55 Worker Nodes (WN) affected, all of them embedded in blade centers. The computing problem was caused when trying to set the network interface down of the affected WNs and caused Computing service instabilities until Thu 26/May 06:55h)

Impact

Computing service was experiencing problems (non-responsive state) during two time periods; 25/05-18:10 to 20:50 UTC and 25/05-23:15 to 26/05-06:55 UTC. About 1000 jobs affected. 55 WNs were directly affected by compromised DNs (running ~400 jobs), but 100 WNs more were also affected by collateral effects. Incident caused CERN SAM/Nagios site unavailability and affected data processing services for ATLAS, CMS and LHCb.

Time line of the incident (UTC)

- 2011/05/25 15:00 Mail reception from NGI security officers alerting about the SSC5 (formal start of the exercise)
- 2011/05/25 15:05 Start of the investigation of the involved DNs
- 2011/05/25 15:15 DN ban process started on all grid services (LFC, FTS, SRM, CE)
- 2011/05/25 16:40 Distributed command issued to isolate affected nodes (55 WNs), start noticing not expected disruptions in not affected WNs.
- 2011/05/25 18:00 Engineers start working on the computing services to recover the nodes. Batch system (Torque) was affected by the instabilities and was in a non-responsive state.
- 2011/05/25 22:00 Situation seemed to reach stable state. Majority of WNs recovered (except the ones affected by SSC5, protocol recommendation is to reinstall)
- 2011/05/26 23:15 The batch system got unstable as a result of the many WNs that were down. Non-responsive state caused failure in SAM probes during the next 7 hours.
- 2011/05/26 06:55 The jobs that caused the batch system instabilities were cleaned manually and computing services were recovered.

Analysis

The collateral effect was caused by the procedure to "unplug" the network within the blade centers, to disable the network interface of the affected WNs. To do so a broadcast command was issued: "ifdown" in the network bond interface. After this, engineers started to observe affectation in different nodes that were not affected and where the command was not even sent.

Follow up actions

The "ifdown" command seems not to be a good idea to disable the network of an specific WN, and that produced disruption at network level on the whole blace center, hence affecting random nodes co-habiting inside the same blade. It was clear at this point that the network stop procedure should have to be reviewed. Luckily, on the next day IFAE was warned about the SSC5 as well and a new procedure was implemented to isolate the nodes: "service network stop" was sent via broadcast instead of "ifdown", nodes affected in this case were 27 and the new procedure worked correctly.