Post Mortem Tier-1 Service Incident dCache PNFS overload (10-June-2011)

De WikiPIC

Incident Start: 11th June 2011 16:30 CET

Incident End: 11th June 2011 around 21:20 CET

Contenido

Description

Problem affecting dCache namespace (PNFS unresponsive) lead to errors in data transfers and computing services starting around 10-Jun 16h30. Several actions were performed when trying to identify the problem: jobs were paused, FTS channels were set inactive and finally VO ATLAS was banned in data transfers. When banning ATLAS the problems disappeared but there was nothing wrong for this VO. We found later around 10 Jun 21h a hanged cron used to extract tape accounting, hammering ATLAS PNFS DataBase, hence causing a pile up in requests from all types: FTS, WNs, etc.

Impact

Imports and exports file transfers were failing for all sites trying to access PIC SRM. This issue was also causing very low job efficiency which were failing due to I/O accessing problems from the WNs to the storage system. SAM tests were failing for SRM and CEs during the incident.

Time line of the incident

Incident started on 11th June 2011:

- **10-Jun 16:30 -** Job low efficiency is detected and is opened an investigation to locate the reason of failing jobs: **dccp** commands hanging on the WNs. Is suspected a WN update on some nodes which are causing transfer failures, but few minutes later will be seen that the **dccp** hang is caused due to a deadly storage system.
- **10-Jun 16:35 -** Detected transfer failures in the FTS and in the ATLAS Dashboard. This

clearly points a dCache issue.

- **10-Jun 16:43 -** *GGUS-71447* - PIC Failure to connect with remote SRM (https://ggus.eu /ws/ticket_info.php?ticket=71447) is assigned to PIC.
- **10-Jun ~17:00 -** Multiple problems are seen, but the origin of the problem can not be determined. The following issues are being detected:

  ATLAS aggressive deletion requests: aggressive deletion going on at PIC for ATLAS ~2800/600s and some of them failing as the file is not found
  SRM Bring Online Requests from LHCb are issued with just 30 seconds of difference between the request's place time and the stated expiration time. This probably leads to the SRM to be forced to kill the transfer requests as they start as the true start time is sometimes later than the expiration time. Probably LHCb will be banned.
  FTS Transfers are increasing the load on the SRM and PNFS.
  Thousands of running jobs trying to access to the storage system are heavily increasing the load on the SRM and PNFS.

- **10-Jun 17:27 -** Multiple actions are taken with the purpose to reduce the load on the dCache system:

  ACTION: **User "/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=rsantana /CN=691977/CN=Renato Santana" becomes banned on the dCache system** to avoid SRM Bring Online failures.
  ACTION: **Stop ATLAS deletion** - ATLAS shifter is contacted via telephone and is asked to stop ATLAS deletion for PIC site.
  ACTION: ***atprod* & *atpilot* jobs are suspended -** This action is taken to avoid failing **dccp**s during dCache downtime and to decrease dCache load on startup.

- **10-Jun 18:00 -** Storage System is still unstable, some extra actions are taken:

  ACTION: **glong_sl5 is stopped.**
  ACTION: **FTS channels are paused** by setting all them as *Inactive*.

- **10-Jun ~18:30 -** storage system is stable, but nothing is running (only active transfers from other FTS)
- **10-Jun between 18:30 & 20:00 -** multiple actions are taken but storage system becomes unstable again and again.

  ACTION: **FTS channels are opened**. With this dCache system becomes unstable. Few minutes later, channels are stopped again.
  ACTION: **dCache queue** is decreased from 100 to 50.
  RESULT: It does not solve storage problems.

- **10-Jun between ~20:00 (+/- 30 min) - ATLAS is banned in the dCache system and system becomes stable. Few minutes after, the following actions are taken:**

  ACTION: ATLAS Production Role is opened.
  ACTION: ATLAS transfers are opened.
  RESULT: System becomes unstable after few minutes.

- **10-Jun ~20:45 -** Cron jobs in ui01 are killed. Apparently, those crons are unoffensive.
- **10-Jun ~20:45-21:00 -** Hanged hourly cron job **/root/scripts/Plots_Active_queue.sh** in

enssrv02 is detected. This cron has multiple hanged instances (one per hour since ~6:00am).

> **ACTION:** Hourly cron /root/scripts/Plots_Active_queue.sh is deleted and hanged processes are killed
> **ACTION:** dCache servers are restarted
> **RESULT:** ***system becomes stable***

- **10-Jun ~21:00-21:15 -** computing and file transfers are opened:

> **ACTION:**: FTS channels are set to **Active**
> **ACTION:** computing '**glong_sl5** queue is opened
> **ACTION:** *atpilot* jobs are resumed (unpaused).
> **ACTION:** *atprd* jobs are resumed (unpaused).

- **10-Jun 21:20 -** Incident is solved, but under investigation.
- **10-Jun 21:40 -** ATLAS deletion is enabled from ATLAS side.

Details

- **RT-1812**: Incidència Storage PIC 10 Juny 2011 (http://rt.pic.es/Ticket/Display.html?id=1812)
- **GGUS-71453**: LHCb Bring On line Requests blocking SRM at PIC (https://ggus.eu/ws/ticket_info.php?ticket=71453)
- **GGUS-71447**: PIC Failure to connect with remote SRM (https://ggus.eu/ws/ticket_info.php?ticket=71447)

Analysis

The problem with the cron started before 16h30 but had no impact since it is used only for accounting, but this cron starts a new instance every hour. This caused pile-up of cron jobs listing ATLAS PNFS, and as the pile-up increased this lead to a DoS when trying to access ATLAS PNFS (lcg-cr, srmcp, dccp, etc). This was even worst after noon, as ATLAS was deleting files heavily and the load on the DB increased.

After fixing the cron everything returned to normality, transfers were resumed and jobs un-paused. No more issues during the weekend.

We spotted also that LHCb was sending Bring On Line requests with very small time difference between the placement and the expiration time (also spotted for ATLAS a months ago and corrected):

- **GGUS-71453**: LHCb Bring On line Requests blocking SRM at PIC (https://ggus.eu/ws/ticket_info.php?ticket=71453)

Follow-up

- Some actions are being taken to avoid a similar problem in the future:
  - Ensure that **timeout** is set for all scripts working on the PNFS to avoid new process hangs and PNFS blocking.
  - Need to measure the expiration time for all crons running on the PNFS.

For all those running crons which their execution is higher than 15min, we will set a 2 hours timestamp before next execution. With this, we will ensure that process will finish its execution before a new one.

- Create a new PNFS Manager alarm, raising when more than 100 queued jobs are detected on the PNFS.

  dcmon.pic.es contains this information, so we can use this server to raise the Queued Jobs alarm.