TIMELINE

All times in CET

```
22.04 07:44 Atlas registered Out of Space event (Savannah bug 66270). FTS failing, job submitting continues
23.04 13:21 First failures on Ethernet switch (connecting half of CMS and ATLAS servers) observed (spontaneous reset)
23.04 17:?? Switch 101-01-01 failed
23.04 19:00 Switch 101-01-01 replaced (reduced bandwidth for bonded interfaces because of change of configuration
parameters, discovered later on)
26.04 09:00-19:00 Scheduled intervention for SAN re-cabling
27.04 16:55 Heavy load on Atlas FS: 700MB/s (ethernet on 50% of NDS servers saturated) outband and 400MB/s in-band (close
to FC up-links limits)
28.04 01:15 Blocking error on StormBE ATLAS: “java: Cannot open logfile /dev/null: Too many open files”
Storm stops responding, Nagios event handler trying to start StormBE fails because previous instance was still present
(action needed: correct event handler procedure to cleanup before restart)
28.04 08:00 Manual restart of StormBE. Normal operations of Atlas restored.
28.04 11.14 observed saturation of FC up-link for ATLAS
28.04 12:?? Atlas filled up MCDISK again. FTS failing, job submitting continues
28.04 12:30 Scheduled down for TSM/GPFS upgrade (affected TSM server, HSM clients, Storm BE and FE)
28.04 ??:?? ATLAS LFC database moved to a new cluster
28.04 16:49 Atlas Requested temporary increase quota for MCDISK
28.04 17:20 Quota for Atlas MCDISK increased by 10TB
29.04 06:?? MCDISK filled up again. FTS failing, job submitting continues
29.04 11:00 in CMS cluster observed huge delays in any kind of operations
29.04 12:00 in Atlas DATATAPE observed similar problems
29.04 19:?? Found that new switch is not balancing Bonded interfaces
29.04 23:00 Switch reconfigured, bandwidth restored

02.05 20.xx CMS FS starts slowing down
02.05 21.?? CMS FS back in production
02.05 21.xx ATLAS StoRM end-point becomes very slow
02.05 23.xx ATLAS end-point operations restored
03.05 08.?? CMS FS stops again
03.05 22.xx CMS operatons restored
```

NOTE: All problems were observed under heavy load conditions (~1GB/s for each experiment), close to the hw limit of present storage system (2010 storage is being deployed).

Configuration of the system

At CNAF, the MSS relies on GEMSS, i.e. StoRM for the srm layer, GPFS for the disk-system and TSM for the tape system.
These are the installed releases for each component:

- GEMSS: 1.1.37-1 (the latest available release)
- StoRM: 1.5.1-3 (the latest available release)
- GPFS: 3.2.1-19 (the latest available release of 3.2 family). On the BEs until April 28 (see below) the installed GPFS version was 3.2.1-14.
- TSM server: 6.1.3-3 upgraded, after IBM advice, on April 28 from 6.1.0-x version.
- TSM clients: 6.1.3-0. The installed version until April 28 was the 6.1.0-2.

All servers, excepting the StoRM ones, have 64 bits architecture with 64 bits OS.

All WLCG experiments have dedicated GPFS clusters.

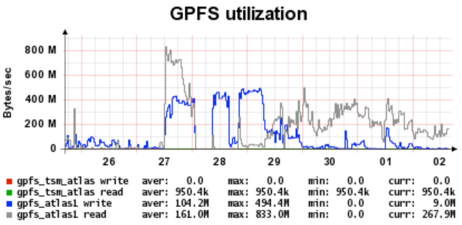
Ante-fact

On Friday April 23 at 13:21 the first failures on an Ethernet switch was observed (spontaneous reset).
After some time (at 17.xx) the switch failed.
This switch connected half of the CMS and ATLAS servers (both gridftp and NSD) to the Tier1 LAN. There was no interruption since the second half of the servers took care of all the load.
Anyway a few hours later (19:00) a new switch was put in production (with the same configuration) and the normal situation restored, the only slight difference being the model of the switch (a more recent one).
The only side effect of this was that the load balancing algorithm for bonded interfaces stopped working properly until Thursday 29 when this was discovered and a new configuration of the switch was put in place (see below).

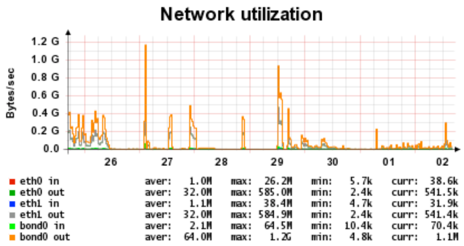
On Monday 26 a scheduled "at risk" intervention took place (09:00-19:00) on the Storage Area Network (it interconnects the storage systems, the disk-servers and the tape drives) without any apparent problem.

During the night April 27-28 at nearly 1:15 AM CET there was a blocking problem on ATLAS StoRM BE. From the log file: "java: Cannot open logfile /dev/null: Too many open files".
As a consequence StoRM stopped processing srm requests. Atlas opened a TEAM ticket (#57733).
To note that a watchdog is active on the BE to restart the process in case of failure but in this case did not work because the previous instance was still present.
The functionality of the overall system (and the normal operations of ATLAS) was restored after a manual restart of the BE nearly at 09.00 AM CET.

Action to take: correct event handler procedure in order to detect zombie process and to clean up before trying to restart.



gridftp servers on D1T0 space-tokens



NSD servers on D1T0 space tokens

We also observed the near saturation conditions of the up-links connecting edge FC switches (Atlas disks) to Tier-1 Core switch where TSM infrastructure (including tape drives) are connected.

Action to be taken: double the up-links (this will be done during May).

In the meanwhile (12:xx PM CET), ATLAS filled up MCDISK space token: hence all FTS transfers started failing (but the job submission continued).

After ATLAS request (05:00 PM CET) quota for MCDISK was temporarily increased by 10 TB (05:20 PM CET), but the new space was completely filled up about a dozen of hours later (about 06.00 AM on April 29) and the FTS started to fail again. This helps to understand the problem occurred on April 29 (see below).

On Wednesday 28 another a scheduled one hour down took place: the TSM server was upgraded together with the TSM/HSM clients (the servers interconnecting the GPFS file-system to the tape drivers). During this down, also the GPFS on the BE component of the StoRM end-points was aligned with the one installed on the other servers (a restart of GPFS and hence of the StoRM process was needed).

Also in this case the intervention showed no evident problems.

For the sake of completeness, we also mention the scheduled intervention, also on April 28, on the ATLAS LFC database which was moved to a new Oracle cluster.

After each of these interventions, the overall system was carefully tested.

From the 4:00 PM CET of April 28 until the morning of April 29, the system showed no problems.

What happened on April 29 - May 3 to the ATLAS end-point

ATLAS stopped working properly on April 29 morning. At first we thought to have to solve the same problem present on the CMS cluster.

The SAM tests were green again in the afternoon but ATLAS notified us that there were 3 jobs still failing due to the storage.

A new check was done and the only odd thing was the high response time of StoRM of the PtP and PtG operations (seconds instead of tens of milliseconds) which we related to the high load of the internal mysql database.

To be noticed that, as stated previously, the space-token MCDISK had been completely filled up and all transfers to this were failing: the failures had quite an high rate (thousands/hour).

This was triggered (probably) by a bug in StoRM: all these failed requests had been inserted in the StoRM mysql database and due to some problem in the query mechanism, the database operations had become very slow and causing many of the other requests to fail as well.

Hence ATLAS was asked (April 30 01:27 AM CET) to stop trying transfers to MCDISK space-token while, on our side, we started digging into the StoRM code. After this analysis we put in place a work-around script to clean the mysql table.

As soon as the table was cleaned, on April 30 afternoon (~ 6:30 PM CET), the ATLAS end-point was back in action.

Action to be taken: bug fix in StoRM

On May 2 09.00 PM MET evening ATLAS FS slowed down again.

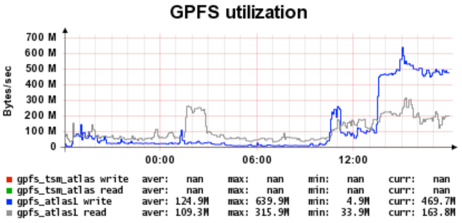
Since we had discovered that TSM client was probably bugged, we decided, as already done for CMS, to roll-back to the previous version.

After this operation the system was back to production, but the disk was still full.

To be noticed that the 3 failing jobs (see above) failed again due to an excessive use of the swap space (problem notified to Atlas).

On May 3 13 TB of data were deleted from MC_DISK by ATLAS to allow more transfers but the disk space was filled up again in the evening and the transfers started failing again.

As of today (May 7) other data have been deleted by Atlas from the MCDISK space token and transfers are going on with quite an high throughput (~ 460 MB/s from the dashboard).



gridftp throughput