PIC
port d'informació
científica

CCRC08 May Post-Mortem
Tier-1 view

Gonzalo Merino, PIC

CERN, 12 June 2008

Outline

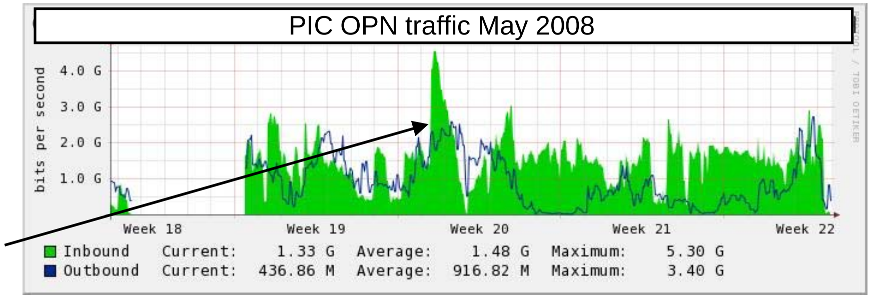
PIC
port d'informació
científica

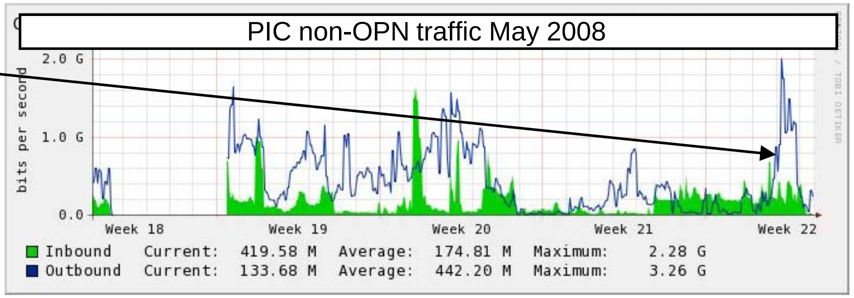
- Overall CCRC08 feedback from PIC
- Review of issues at PIC
- Review of (some) issues from other T1s

Overall T1 service performance at PIC during CCRC08

- In general, the Tier-1 service at PIC ran quite smoothly during May and coped well with the load generated by ATLAS + CMS + LHCb

- Local record data transfer rates reached
  - Above **500MB/s** import during ATLAS T1-T1 test (13-May)
  - Above **300MB/s** export to Tier-2s (mostly CMS, 29-May)
  - LAN traffic WN-disk above **1000MB/s** for CMS skimming (see next slide)


PIC OPN traffic May 2008


PIC non-OPN traffic May 2008

- PIC SAM reliability for May: 96%
  - Good stability of the service, despite the high load levels
  - Avoided making interventions during CCRC08 (priority for stability)
    - No burning-critical update, so all of them were scheduled for post-CCRC08

Computing Element issues

PIC
port d'informació
científica

- We see that sometimes the number of queued jobs reaches very high values (>10x the total slots at the sites)
  - This causes high load on the CEs and might result in problems handling the running jobs
- Question for the VOs: Could we set a limit on the number of jobs you can have **waiting** in the queue?
  - We can publish this number through the Information System:

    ```
    GlueCEPolicyMaxWaitingJobs:
    ```

  - Can the experiment frameworks handle this information?

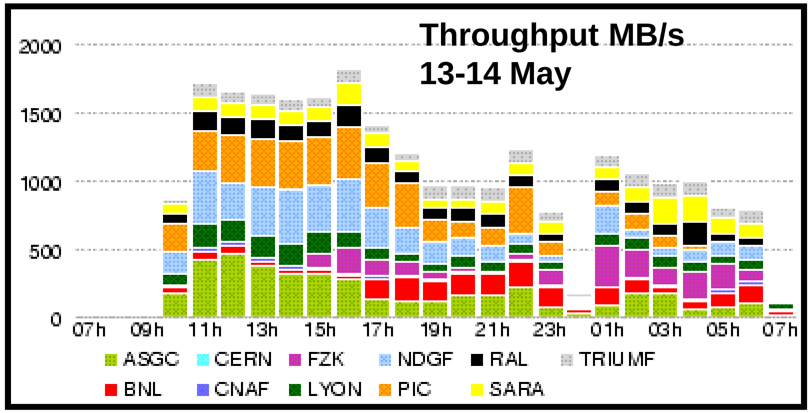
File Families

- This time we tried to group the files into tape File Families (FF) depending on the directory they were landing at
- ATLAS: Besides a “cosmics” FF, only a “test/ccrc08” FF created to ease later bulk deletion.
- CMS: We tried a more fine-grained distribution (5 FFs)
  - Directories created and configured ahead of time to map to these FFs
  - In the end, about half of the data imported to tape during May still went into the catch-all FF .
- LHCb: Simple configuration with two FFs (RAW, DST) also mapping to pre-created directories.
  - Looks like it worked ok. Quite balanced share between both.
- The procedure of manual creation of directories and FF mapping ahead of time it will not scale
  - We should agree in a common way to automate this in the future
  - Use the now available tools to pass the Storage Token to the MSS?

WAN transfer issue: export/import asymmetry

PIC
port d'informació
científica

- During the ATLAS T1-T1 replication test, we observed that the data import efficiency T1→PIC was in general very good (>90%) ☺
  - ☹ However, the export efficiencies PIC→T1 to some T1s (CNAF, FZK) were significantly lower (~50%)
- May be the different FTS parameters at each T1 cause this?
- The FTS at each T1 steers the data import: T1→PIC
  - T1-T1 channels parameters at PIC:
    - N_files = 10 for TRIUMF-PIC, 30 for other T1s-PIC
    - N_streams = 10 for ASGC-PIC, 5 for other T1s-PIC



Throughput MB/s
13-14 May

Data access from jobs issues (1/3) ATLAS CondDB file

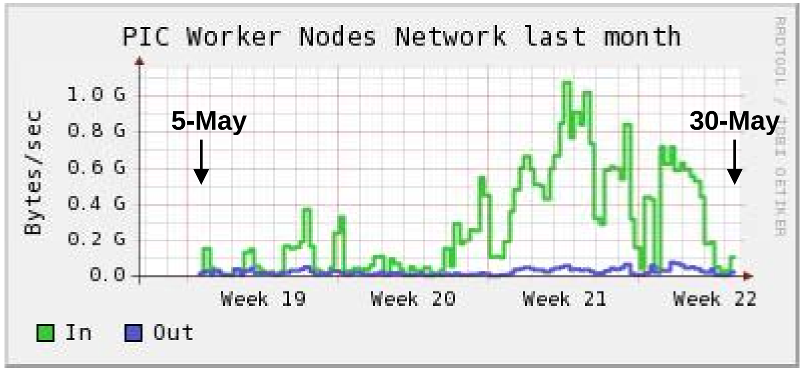
- On 28-May, ATLAS sent ~400 reprocessing jobs to PIC in one shot. All of them tried to access the same single file at the beginning:
  - /pnfs/pic.es/data/atlas/…/CDRelease.28940-28997.v0000.tar.gz
- This is a Conditions DB file
  - ~4.6GB filesize x 400 jobs ➔100MB/s hitting one disk pool for 5 hours
- That pool had still a 1Gbps uplink, which saturated
  - This caused the dCache pool-server control socket to lose packets, and finally the pool hanged
- This week we have finished the upgrade of all the disk-pool uplinks to 4Gbps. The problem will be less severe next time.
  - BUT the issue of hot files is still there (CondDB, Minimum Bias files)
  - Hot files should be in well localized directories so that sites can try and manage the situation (multiple replicas, etc)

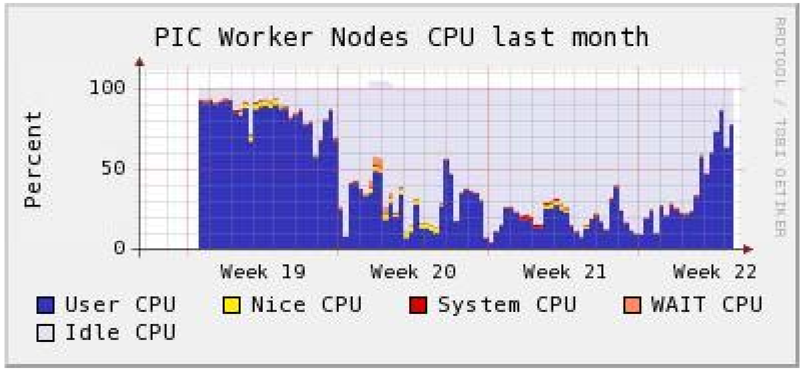
Data access from jobs issues (2/3) CMS reading from T1D0

- CMS ran Skimming and Reprocessing concurrently during May
- We observed that
  - about 2/3 of the load read from tape (T1D0)
  - about 1/3 of the load read from disk (T0D1)
- At PIC we were lucky that we still have big (~100TB) disk buffer in front of T1D0 (heritage from some CMS T1D1 tests in the past)
  - ☺ Most of the files were on the disk cache – had no problems with tape access latency
  - ☹ Most probably in the real use case those jobs will never find the input data in the cache: we missed the chance to test reprocessing recalling from tape (one of the main Tier-1 workflows)
- Other Tier-1 sites experienced important job inefficiencies due to the latency accessing the data on tape (RAL feedback)

Data access from jobs issues (3/3) CMS skimming

- 3rd week of May, CMS filled the PIC farm (~600 slots) with skimming jobs with a huge read throughput rate from storage
  - 600 to 1000 MB/s WN reading from SE sustained for several days
- ☹ Resulting problem: Some of the switches saturated (WN-SE network designed for 1-2 MB/s/job)
  - Timeouts WAN-exporting data sitting in pools connected to those switches
  - The WNs CPU was idle (bottleneck in the LAN)
- ☺ Positive part: The service stayed up and running all the time
  - Disk servers coped with the load
  - Central services not affected





Data access from jobs issues (3/3) CMS skimming

PIC
port d'informació científica

- Cause of the problem understood by CMS the last week of CCRC08:
  - DCACHE_READ_AHEAD configuration
- Default was **1 MB** buffer
  - Each job was reading more than 10x the same data (~30MB/s/job)
  - Read ahead is a problem for erratic read patterns
- Reducing it to **128 kB** solved the problem
  - Skimming jobs back to ~2 MB/s/job
- Question: will CMS keep the “remote open” paradigm for the jobs, or will try out the “copy to local disk first” possibility?

Summary of issues at PIC

PIC
port d'informació
científica

- The overall behaviour of the T1 service has been rather positive. Good stability. We believe much of this was thanks to our decision of making no intervention since last week April and throughout May.
- Still, some issues have arisen. From less to more relevant:
  - CE load peaks due to lots jobs waiting in the queues
  - Tape File Families: granularity level and automatic creation
  - ATLAS "hot files" (CondDB and MinBias)
  - ATLAS T1-T1 replication tests, asymmetry import/export, FTS parameters
  - Feel that we are still missing a realistic test of tape recalling from ordered reprocessing workload
  - CMS skimming jobs collapsing the LAN of the centre

Gsidcap problems (IN2P3)

- Quoting IN2P3:
- LHCb has been having problems accessing dCache-managed data in our site, through gsidcap.
- Their jobs get stuck accessing files without a reproducible pattern: the same job ran several times get stuck reading different files each time. We know that the problems happen with files residing in different servers, and the same job can succeed reading a file and get errors reading another one, both on the same file server. Experts of LHCb software are working with the experts of the site trying to understand the causes of this.
- To my knowledge, this behaviour is only observed by LHCb in our site; conversely, no other experiment using our site has reported such a problem.
- Sadly, this is a very long standing issue and without doubt our Top 1 issue.

LFC daemon crashing (TRIUMF, IN2P3)

PIC
port d'informació
científica

- Already reported this morning (SARA dCache report)
- Also reported as a relevant issue by other T1s (IN2P3, TRIUMF …)
  - **IN2P3**: “We observed problems of instability of the regional LFC service for Atlas. The symptoms were that the LFC daemon suddenly stopped working without lefting anything in the logs helping us identifying a potential cause. Initially we thought it was related to the hardware so we used a second machine (different hardware configuration) to deploy the same LFC version. The same symptom was observed with this instance. After asking some help from the Atlas experts, we could not correlate it to any particular activity of Atlas at the times the daemons stopped working. To mitigate the problem, we implemented a LFC service composed of a 2 DNS-load balanced machines. The developers were informed but the cause is still unknown for us. Basically, the problem is still there”
  - **TRIUMF**: “we created a GGUS ticket and there is now a new version which we didn't deploy yet since it is not official yet.”

MSS-dCache issues (IN2P3)

- Quoting IN2P3:
- We know we need to improve (i.e. make more intelligent) the mechanism used by dCache for writing/reading files to/from the backend MSS in order to improve the usage of the tape drives.
- We know we need to improve the (few) existing monitoring tools in order to be able to follow the activity of the MSS for each experiment, for instance input/output rates, number of tape drive used, etc.

Quality Control for the SRM releases (NIKHEF)

- Already reported this morning (SARA dCache report)
- The advertised required dCache version for CCRC08 was 1.8.0-15
- NIKHEF installed the last patch level available (15p3) which did not work
  - Later, were given p4 which fixed that bug but apparently introduced another one
- We should have a certified version at least a week before any large-scale operation like this, and the testing & quality control of any patches needs to be MUCH better

- NB: Several T1s did run CCRC08 with an older version of dCache (v12) with apparently no missing key functionality problems

END