Service Incident Report

Type of Incident: dCache transfer degradation
Location: IN2P3-CC
Duration: 2 months
Date: September the 23rd 2010 to November the 22nd 2010
Authors: Pierre Girard, Yvan Calas, Ghita Rahal
Report status: Final
Report logs:
- 2011-01-20: Creation by Pierre Girard
- 2011-01-29: Submitted for internal review
- 2011-02-07: Corrections from both Ghita Rahal and Yvan Calas
- 2011-02-08: Submitted to P. Fuhrmann (dCache.org) for validation purposes
- 2011-02-10: Comment/correction from P. Fuhrmann (dCache.org)
- 2011-02-11: Final version sent at **wlcg-scod@cern.ch**.

Description

IN2P3-CC T1 experienced several episodes of low transfer throughputs with its dCache server for both export and import activities of ATLAS. This problem was detected just after a long downtime, ended on September the 23rd 2010, and lasted until the end of ATLAS reprocessing campaign, on November the 22nd 2010. During that period, the problem appeared and disappeared without any clear correlation either with ATLAS activity change, or with the actions taken by IN2P3-CC experts.

Moreover, during each degradation period, the slow transfers were concentrated on some pool servers only, while the transfers from other pool servers were working fine. That resulted in dealing with both normal and slow transfers at the same time.

CMS, which was using the same dCache instance at the same time, did not seem to be impacted as much as ATLAS by slow transfer problems.

Timeline

- From 2010-09-17 to 2010-09-23, cumulative downtimes
  - Security alert (Unscheduled)
  - Maintenance operations (Scheduled)
    - Local network infrastructure upgrade
      - New core switch/router
      - Important changes within local network topology
    - OS upgrade
      - Solaris 10
      - dCache disk servers
    - dCache update
      - dCache 1.9.5-11 -> 1.9.5-22

        - Tape authorization activation
        - Alice VO removal
- From 2010-09-28 to 2010-10-05, low throughputs with some ATLAS export transfers
    - 2010-09-28, TEAM GGUS: #62500
    - 2010-09-30 to 2010-10-01, Network configuration checking. Minor changes done. No progress noticed. Some dCache gridftp doors got a high number of gridftp connections by pool: ~20/pools. The maximum of gridftp connection by pool was consequently decreased to 5/pool. Progress noticed.
    - 2010-10-01 to 2010-10-04, Problem still there.
        - 2010-10-01, TEAM GGUS #62692
        - New configuration tuning tested without significant progress. IPERF tests done. No problem found with network
    - 2010-10-05, the problem disappeared
- From 2010-10-12 to 2010-10-14, some new slow transfers observed on export tranfers to some french sites
    - 2010-10-14, disabling 2 export pools with too many transfer errors.
- From 2010-10-15 to 2010-10-18, Contacts taken with dCache.org people
    - 2010-10-15, P. Mail to P. Fuhrmann to get help about the slow transfers problem.
    - 2010-10-18, Y. Calas (dCache administrator) opened a ticket to dCache.org troubleshooting system (#5906).
        - dCache developers reported that sending data from Solaris 10 servers seems to be limited to less than 100 Mbps/stream.
        - IN2P3-CC expert communicated throughput results from gridftp doors that negated the assumption concerning any Solaris connectivity limitation.
    - 2010-10-19, dCache administrators joined the weekly dCache Tier I meeting to report on IN2P3-CC dCache problem. None of the dCache site community encountered this kind of problem.
- From 2010-10-25 to 2010-10-26, OS downgrade applied to some dCache pools machines. No significant progress observed, and low throughputs were still noticed with the downgraded machines.
- From 2010-10-28 to 2010-12-03, New low throughputs episodes for ATLAS import/export transfers.
    - 2010-10-28, TEAM GGUS ticket #63558
    - 2010-10-29, Main TEAM GGUS ticket #63627, and #63626
    - 2010-11-10, TEAM GGUS ticket #64120
    - 2010-11-12, TEAM GGUS #64202
    - 2010-11-12, Phone conference with Atlas representatives. Decision to stop all ATLAS activities except ongoing reprocessing and export to other Tier1s.
    - 2010-11-12, Mail from P. Fuhrmann providing quick update about dCache.org progress. Its content was wrongly understood as pointing Solaris as being the cause of the hassle.
    - 2010-11-17: ATLAS activities of transfers and computing are gradually restarted.
    - 2010-11-18, Deployment of 9 GridFTP doors on SL5 virtual machines to confirm/negate suspicion on Solaris machines. Low throughput problems were also observed with those SL5 servers.
    - 2010-11-18, Any attempt to put back in production a pool server failed because of a huge increase of CPU Load, P2P connections and GridFTP connections. New monitoring metrics were added to monitor TCP connections status. Many ESTABLISHED and CLOSEWAIT TCP connections were noticed. By the night, Y. Calas finally succeeded in putting back the pool server in production by

deactivating the checksum calculation on P2P transfers. The checksum calculation became then very suspected.

- 2010-11-19, Phone conference with dCache.org people. CCIN2P3 dCache/System administrators exposed the details of the dCache infrastructure, the tests made, and symptoms observed. dCache.org people proposed some configuration changes. They confirmed that checksum calculation has to be configured "on transfer" rather than "on write" (P2P configuration parameters). That should then avoid the observed CPU overload and the large amount of CLOSEWAIT TCP connections. Moreover, a new command is available with this new version of dCache that makes possible to better use the 2 network ports on the pool servers, in particular, by avoiding the use of the LHCOPN network port with the internal P2P traffic.
- 2010-11-23, Gerd Behrmann from NDGF met dCache administrator at CCIN2P3 on behalf of dCache.org. Gerd concluded that the load-balancing was not correct when both export and import pools are hosted on the same machine. That explained the phenomenon observed by CCIN2P3 dCache administrators that occurred since September downtime.
- 2010-11-25, Phone conference with dCache.org people. Gerd Behrmann reported about his visit to CCIN2P3. Some configurations recommended, as for example unifying gridftp and P2P queue limits.
- 2010-12-01, Phone conference with dCache.org people. The situation considered as pretty good. Atlas started to increase its activity, and no more slow transfers were observed since.
- 2010-12-02, All transfer activities are back to nominal.
- 2010-12-03, TEAM GGUS #63627 switched to "solved" by IN2P3-CC, not yet verified by Atlas.

Analysis

It was noticed during the problematic period that slow transfers were concentrated on some pools only.

From a dCache point of view, more connections were queued on some pools than on the other ones, even though all pools were similarly configured. Those observations led to think that the problem was due to too many simultaneous writing activities from ATLAS, but a post-mortem analysis of ATLAS activities during a slow transfers episode proved that it was not the case, see Annex A.

In the meantime, many investigations were achieved on the machines hosting the problematic pools. Those investigations led the dCache administrators to take the machines out of production. While trying to put them back in production, it was noticed a sudden overload on the machines which was clearly related with new slow transfers. This is by focusing on this phenomenon that CCIN2P3 dCache administrators started to understand that the checksum calculation was linked to the problem, see Annex B.

From this point, dCache.org people provided help to adapt CCIN2P3 dCache configuration accordingly, to obtain a better load-balancing between pools, and to fully use the network interfaces of the disk servers.

The two following sub-sections summarize the results of the investigations.

What it was not

During the September downtime a lot of changes were applied at the same time, many check tests had consequently to be done to disentangle between the possible suspects.

Below a quick summary of the eliminated suspects:

- Not a hardware problem
- Not a Disk I/O problem
- Not a specific Solaris 10 problem
- Not a Solaris network limitation
- Not a Local Network problem
    - Not a connectivity limitation problem
    - Not a network routing problem
- Not a dCache GridFTP problem
- Not a huge increase of ATLAS Activity

What it really was

The global degradation observed for Atlas in export/import transfers is the result of a complex combination of three problems:

- The checksum calculation “On Write” may cause disk degradation performance. Indeed, once the transferred file is written, its checksum calculation starts by reading it back. This additional read operation can compete with new incoming files and then contribute to a significant increase of the disk I/O.
- The use of the same network interface by an export request and the P2P transfer used to get the requested file from other pool. That can lead to an unexpected sharing of the bandwidth which results in low export/import/P2P transfers and does not make an optimal use of the 2 network interfaces of each machine.
- The load-balancing of the pool manager that can lead to select for a transfer a pool on a machine already loaded by other pool activities

Impact

The degradation problem of CCIN2P3 dCache particularly impacted ATLAS. The apparent random occurrence of the problem made very hard to determine which activities to reduce among various.

As CCIN2P3 dCache instance is shared by both CCIN2P3 T1, CCIN2P3 T2 and CCIN2P3 T3, it was decided together with ATLAS representatives to keep only activities of the T1. Even with such restrictions, the ATLAS reprocessing campaign, run during Q4, was strongly disturbed and required too much operation from both sides, that is, ATLAS people and CCIN2P3 people.

The unreliability of CCIN2P3 dCache finally led ATLAS to run part of its reprocessing at CERN.

Corrective actions

After confirmation by dCache.org people that the checksum calculation could be the key problem of slow transfers, the change of its mode from “On Write” to “On Transfer” was applied on all the machines.

dCache.org people informed us that new feature was introduced in the version 1.9.5-22 of dCache that makes possible to choose the network interface to be used by the P2P transfers on a pool. That was then applied on CCIN2P3 dCache instance to ensure that all the P2P traffic makes use of the public network interface of pool machines. No more P2P traffic should then interfere on the LHCOPN network interface.

With the help of Gerd Behrmann, from NDGF, the parameters of pool manager cost model were changed to get a better load-balancing between the pools according to CCIN2P3 pool distribution over the disk servers.

Open questions

The CCIN2P3 dCache configuration was unchanged during the downtime of September 2010. That then leads to wonder why CCIN2P3 dCache degradation were not noticed before?

Annex A: Correlation with Atlas activity

We analyzed a specific period (14th of Novembre) where all activities were stopped except the reprocessing and some exports of the reprocessed files to the other Tier1s. During this day, we observed 2 overloads on “put” that can be seen in the graph below:

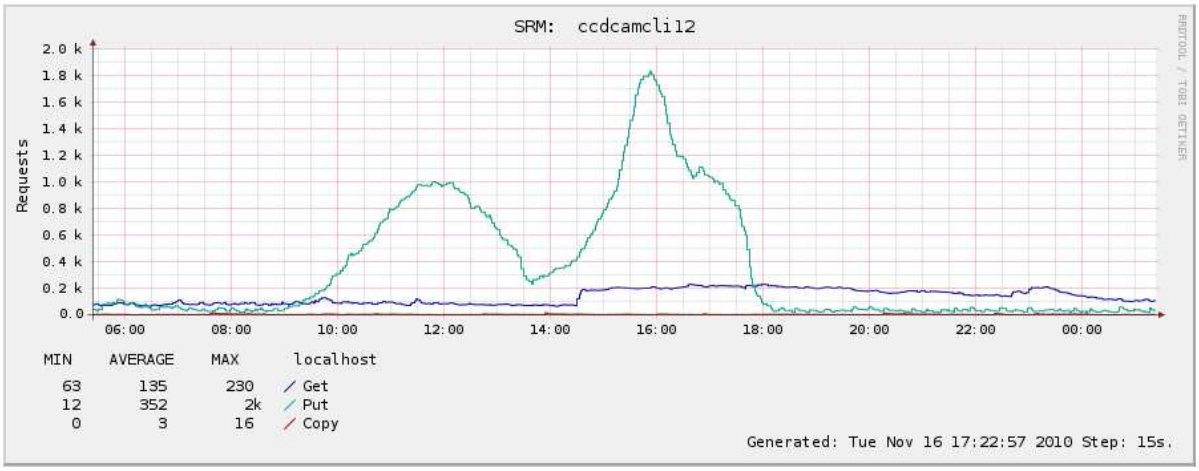

Figure 1: number of open requests on SRM as function of time

We investigated possible correlations between the overload and the number of written files, their size, and their throughput rate.

The result is seen on the 4 graphs below where the 2 periods of overload are marked as 12h and 16h.

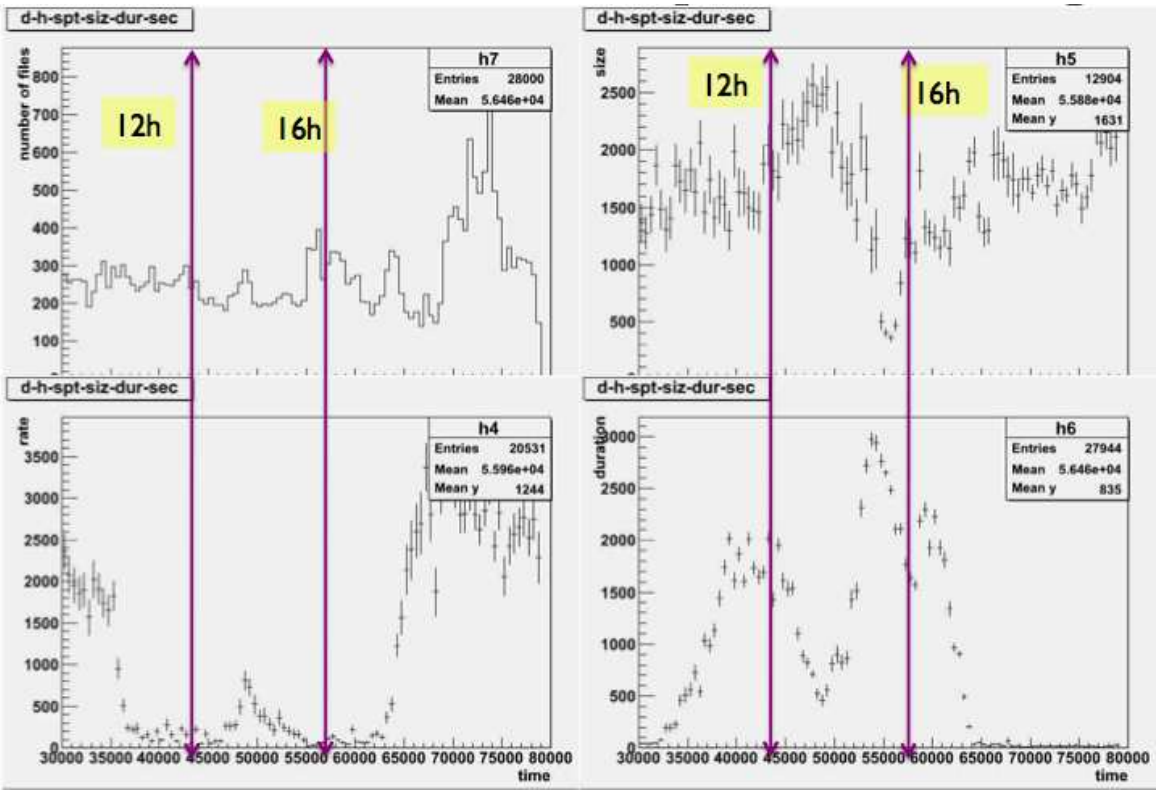

Figure 2: plots of the number of files, size of the files, rate of throughput, duration of the write requests

We observe that the rate of written files is stable during the whole day in the period of stable behavior, as well as during overloads (upper plots, left and right).

The size of the files is also stable around 1.5GB, except around 15:00 where very small files (around 200MB) were transferred.

We can conclude the overload was not due to an increase of activity of writing files.. The behavior seems to be more due to an unexplained build-up bottleneck. This is the sequence of the build-up that is shown comparing lower plots, left and right.

1. The put requests are issued and the rate is correct (around 2.5MB/sec)
2. Then, the rate becomes to drop.
3. New arriving put requests are accumulated with very slow previous open ones.
4. The rates drop more and more until almost 0.
5. Then the timeouts begin to kill the requests after one hour.

As a consequence, the outside activity does not seem to explain the observed degradation.

Annex B: Discoveries while putting back a pool on production

On November the 18th, Y. Calas tried to put back in production a pool server that was taken out of production for investigations because of too many slow transfers. The first attempts immediately triggered new slow transfers. The study of this case led CCIN2P3 dCache administrators to pin down the key problem of the slow transfers.

For a better understanding, CCIN2P3 dCache configuration and some useful dCache concepts are first introduced. The three found problems are analyzed afterwards.

dCache configuration for ATLAS at CCIN2P3

IN2P3-CC dCache is currently configured for Atlas with one export buffer (pool-atlas-xferout), implemented by a dozen of pool servers, and one import buffer (pool-atlas-dq2), implemented by a dozen of pools servers. See Figure 3.

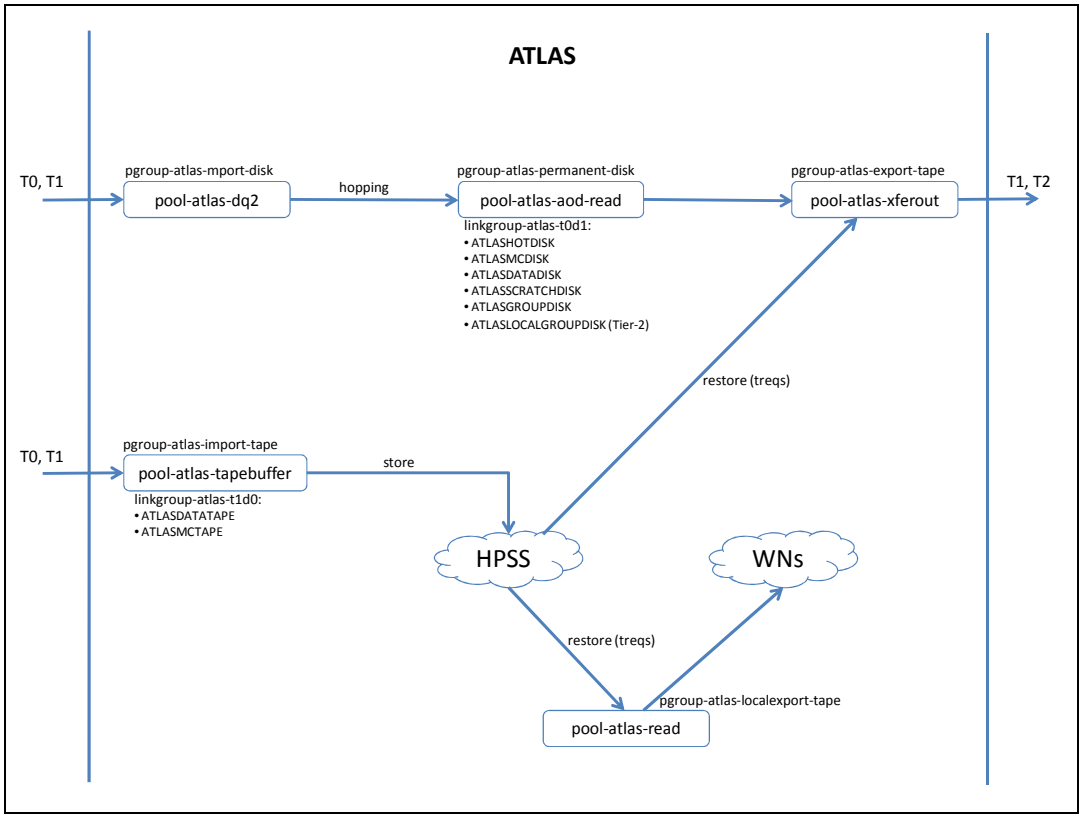

Figure 3: Atlas configuration within IN2P3-CC dCache server

The export buffer is populated by a Pool-to-Pool mechanism which transfers files on demand from data pool servers to export pool servers. The import buffer receives files from gridftp server and uses the Pool-to-Pool mechanism to move the files to the appropriate data pool servers.

The network connectivity of an export/import pool machine is split into 2 network interfaces:

- From/To LHCOPN network, a 2x1Gbps link aggregation. This port is for T0/T1s transfers.
- From/To public network, a 2x1Gbps link aggregation. This port is for all the other types of access.

There are 14 import/export machines, which are then able to deliver together a max of 2x28Gbps.

Quick introduction to some useful dCache concepts

We introduce some dCache concepts which are useful to understand problem analysis. For further information on dCache, please refer to dCache documentation at http://www.dCache.org/manuals/.

dCache cache management

The oldest files are removed from a buffer pool when new files must be written and free space is required on disk. This is typically the profile of the pools used for implementing export and import buffers.

Checksum calculation

A file checksum calculation can be activated. In such a case, it is computed for each new file according to two possible modes:

- “on-transfer”: the checksum is computed while transferring the file.
- “on-write”: the checksum is computed once the file is completely transferred.

The checksum calculation is done for each copy operation on the destination pool. At IN2P3-CC, the checksum is activated during file import operation and during a P2P migration as well. IN2P3-CC calculation mode was set to “on-write”.

Pool manager load-balancing

Within dCache, the pool manager implements a load-balancing between pools. It is supposed to select the best pool candidate for storing/reading a file (read/write) according to a cost model which is described at http://www.dCache.org/manuals/Book-1.9.5/Book.shtml#cf-pm-cm.

Note that the load-balancing is applied between pools of a same pool group. A machine can host various pools of different pool groups.

High CPU load due to the checksum calculation

In the Figure 4, the first attempts of putting back the pool into production were achieved without any configuration change. The last one was achieved after disabling the checksum calculation.

The first graph provides the number of TCP connections, while the second graph shows the CPU load in red, and the number of processes in green.

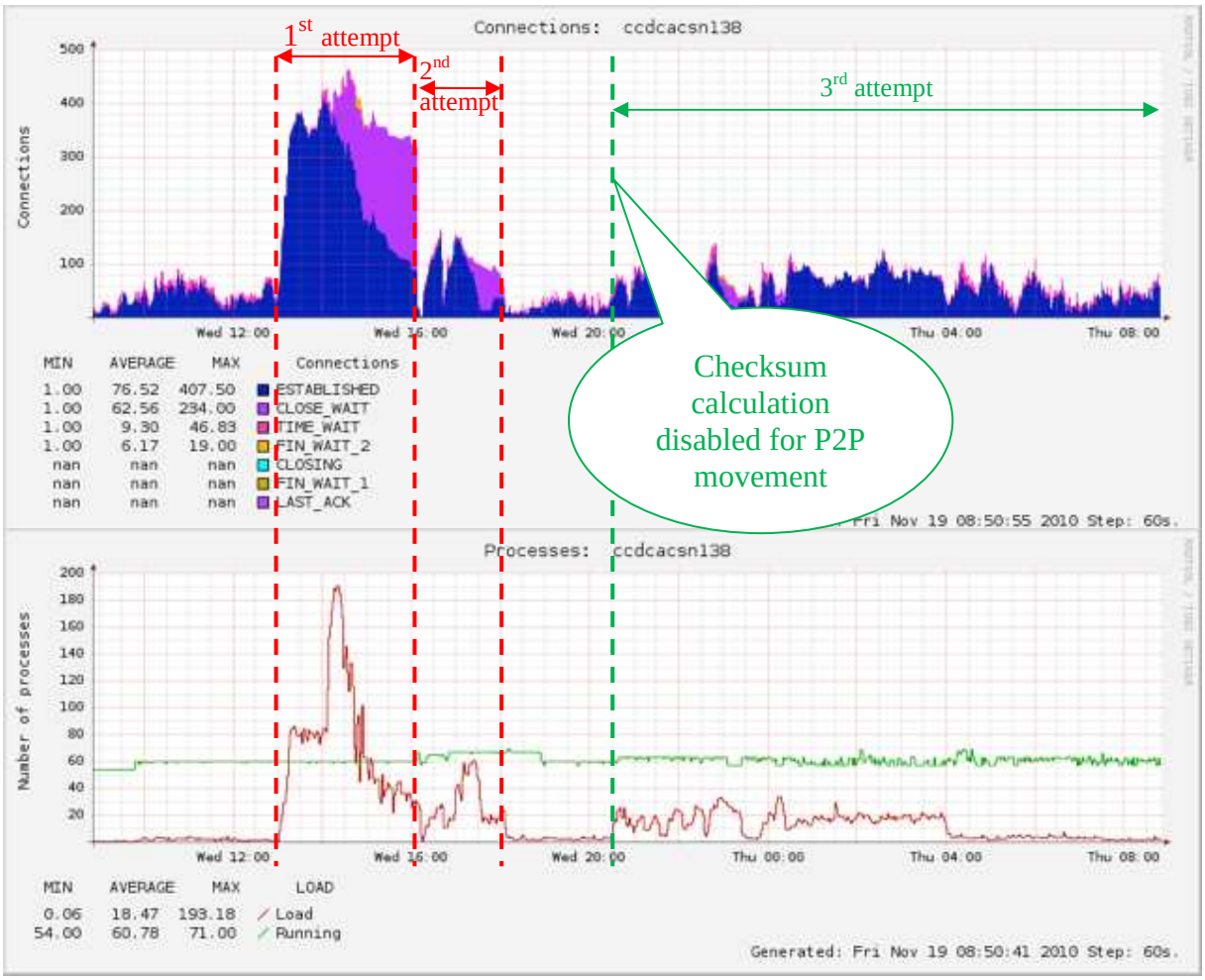

Figure 4: Monitoring of both TCP connections and CPU load while putting back in production a pool server, with and without enabling the checksum calculation

The sudden increase of the TCP connection number is clearly correlated with a peak of CPU load. Such a CPU load increase obviously slowed down the entire system. The machine was then overloaded and became not very responsive.

Yvan started to suspect that the checksum calculation was the reason of this behavior. He therefore launched a new attempt by disabling the checksum. This last attempt succeeded, the machine could be put again in production with both a moderate number of TCP connections and a moderate CPU load.

Yvan' observations were confirmed later by dCache.org people who recommended changing the checksum mode from "on write" to "on transfer". By applying this operation on all the pools, that drastically reduced the load of the disk servers, and all the pools could be put back again in production.

Network interface saturation

By taking a look at the network interface activity, see Figure 5, it was noticed, in green, that only one network interface was used for writing access, for few reading activity, in both light-red and red.

Those activities can be interpreted as being the export activity which triggered a lot of writing activities on the machine to obtain the requested files from other pools through the P2P

mechanism. This P2P activity makes use of the LHCOPN network interface because the original file TURLs, as passed to dCache, explicitly make use of the hostname assigned to this network interface.

By the way, the same network interface is used by export transfers and P2P transfers. It can be noticed that most of the available bandwidth (2x1Gbps) of this interface was used, that is the theoretical value of 250 MB/s, and really a bit more than 200MB/s as shown on the graph.

Given the peak of TCP connections accepted by the pool machine, that explains why the throughput by connection was lower before disabling the checksum calculation than after.

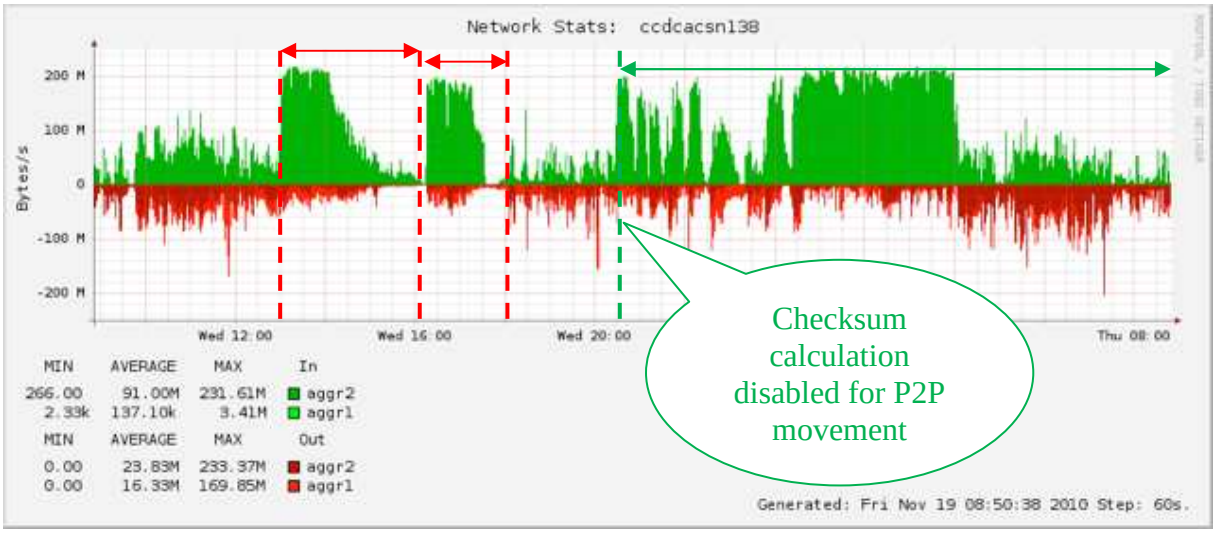

Figure 5: Monitoring of network interface throughputs while putting back in production a pool server, with and without enabling the checksum calculation

The cumulative effect of the checksum calculation on the TCP connections results in a decrease of the throughput.

Pools load-balancing problem

A same machine can host different pools from different pool groups. The poolmanager load-balancing mechanism takes only into account the pools of a same pool group while selecting a pool candidate for a transfer.

The table below, see Table 1, provides an example of a pools distribution over each pool server according to their pool group membership.

dCache pool manager cost model is not well-adapted to such a heterogeneous distribution of pools, because it does not take into account the load involved by the pools of the other pool groups on each machine. Therefore, the pool manager can distribute equitably the transfers over the pools even when some machines are already very busy because of other activities.

Hostname	pgroup-atlas-import-tape	pgroup-atlas-export-tape	pgroup-cms-hpssdata	pgroup-cms-xferout	pgroup-disk-sc3	pgroup-lhcb-tapebuffer	pgroup-atlas-import-disk	pgroup-hpss	pgroup-lhcb-mdst	pgroup-lhcb-xferout
ccdcatsn109.in2p3.fr	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
ccdcatsn110.in2p3.fr	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗
ccdcatsn111.in2p3.fr	✓	✓	✓	✗	✗	✗	✓	✓	✓	✗
ccdcatsn136.in2p3.fr	✓	✓	✓	✓	✗	✓	✗	✗	✗	✓
ccdcatsn137.in2p3.fr	✓	✓	✓	✓	✗	✓	✗	✗	✗	✓
ccdcatsn138.in2p3.fr	✓	✓	✓	✓	✗	✓	✗	✗	✗	✓
ccdcatsn150.in2p3.fr	✓	✓	✓	✗	✗	✓	✓	✗	✗	✓
ccdcatsn151.in2p3.fr	✓	✓	✓	✓	✗	✗	✓	✗	✓	✓
ccdcatsn152.in2p3.fr	✗	✓	✗	✓	✗	✗	✓	✗	✓	✓
ccdcatsn260.in2p3.fr	✓	✓	✓	✓	✗	✓	✗	✗	✓	✓
ccdcatsn278.in2p3.fr	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
ccdcatsn279.in2p3.fr	✗	✓	✗	✗	✓	✗	✗	✓	✗	✗
ccdcatli013.in2p3.fr	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗
ccdcatli014.in2p3.fr	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗
ccdcatli015.in2p3.fr	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗
ccdcacli016.in2p3.fr	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗

Table 1: Distribution of pools by pool group on machines hosting an export pool of ATLAS

For example, if a machine hosts both ATLAS export pool and ATLAS import pool, it can be chosen for an import transfer even when it is already overloaded by P2P transfers due to various export transfer requests. That therefore explains why some slow transfers can occur even with a moderate import activity. Conversely, a moderate export activity can lead to low transfers.

This is such a mix between import activity and export activity that contributed to the overload of some machines, as observed by dCache administrators.