ATLAS 2018 tape recalls analysis

German Cancio german.cancio@cern.ch (mailto:german.cancio@cern.ch)

This workbook contains a comprehensive analysis of recalled files on the ATLAS stager between January and December 2018, across both service classes (“default” and “t0atlas”).

Some basic statistics and histograms

General:
- Total files recalled: 4196626
- Total data volume recalled: 7862.902017 TB
- Median file size: 2513.690742 MB, average: 1873.624673 MB
- Median tape transfer speed (including positioning): 360.5690169 MB, average: 336.6567725 MB
- Median service speed (including positioning): 217.54441 MB, average: 208.6468626 MB

“default” svcclass:
- Total files recalled: 1868860
- Total data volume recalled: 3717.7196584 TB
- Median tape transfer speed: 360.0105982 MB, average: 330.1086319 MB
- Median service speed (including positioning): 185.5007564 MB, average: 184.0814799 MB

“t0atlas” svcclass:
- Total files recalled: 2327766
- Total data volume recalled: 4145.1823586 TB
- Median file size: 2572.6981995 MB, average: 1780.755608 MB
- Median tape transfer speed: 360.7775845 MB, average: 341.9139839 MB
- Median service speed (including positioning): 243.8739069 MB, average: 228.369318 MB

Integrated throughput

Mean integrated throughput:

Overall:
- Mean: 674.3148297 MB/s; median: 278.2708291 MB/s

default:
- Mean: 421.3152712 MB/s; median: 228.4125086 MB/s

t0atlas:
- Mean: 1461.37224 MB/s; median: 972.1831665 MB/s
Drive usage and efficiency

What percentage of the drive session time is spent on reading, positioning, mounting, unmounting?

default:
  - Reading: 39.6114019 %; positioning: 34.3404512 %; mounting: 6.990098 %; unmounting: 19.1881807 %

t0atlas:
  - Reading: 63.7583471 %; positioning: 31.7817281 %; mounting: 1.2005352 %; unmounting: 3.2955313 %

How much total drive utilisation does this represent?

default:
  - Reading: 165.2037296 drive-days; positioning: 143.6923587 drive-days; mounting: 29.2489944 drive-days; unmounting: 80.2900034 drive-days

t0atlas:
  - Reading: 161.375566 drive-days; positioning: 80.4865231 drive-days; mounting: 3.0403289 drive-days; unmounting: 8.3458601 drive-days
Integrated throughput - ATLAS

Integrated throughput - default

Integrated throughput - t0atlas
Histograms of file size, per-drive service and transfer speed

- general
- by service class

**Filesize**

![Filesize chart]

**Per-drive service speed**

![Per-drive service speed chart]
Per-drive service speed - default svcclass

Per-drive transfer speed - default svcclass
Per-drive transfer speed - t0atlas svclass

Per-mount statistics

- Total tape read mounts: 81458
- Distinct tapes mounted: 7285 on 109 distinct drives.
- Average volume read per mount: 96.5270694 GB
- Average files read per mount: 51.5188932

- Number of tape mounts across multiple svcclasses: 325, percentage over total: 0.3989786 %

default service class

- Number of tape mounts on default svcclass: 73788
- Distinct tapes mounted (default): 7123 on 109 distinct drives.
- Average volume read per mount: 50.3705823 GB
- Average files read per mount: 25.2856698

t0atlas service class

- Number of tape mounts on t0atlas svcclass: 7670
- Distinct tapes mounted (t0atlas): 2449 on 108 distinct drives.
- Average volume read per mount: 540.5681214 GB
- Average files read per mount: 303.891395
Plots over time for daily transfer and service speeds

- Transfer speed: actual drive speed
- Service speed: includes positioning time

Average transfer speed, MB/s
Plots for per-svcclass daily mounts, files, data volume
ATLAS 2018 tape recalls analysis

**daily mounts, stacked**

- x-axis: as_datetime(date)
- y-axis: sum_mounts
- svcclass: default, t0atlas

**daily files read, stacked**

- x-axis: as_datetime(date)
- y-axis: sum_files
- svcclass: default, t0atlas
per-svcclass volume per mount and files per mount evolution
11/04/2019
ATLAS 2018 tape recalls analysis

volume per mount (GB), default

files per mount, default

svcclass

default

svcclass

default
Per-svcclass repeated mount rates

Average repeat mount rate (number of times a same tape is mounted during a time interval):

Daily:
- default: 1.2532537
- t0atlas: 1.1107892

3 days:
- default: 1.6848402
- t0atlas: 1.167793

1 week:
- default: 2.0166344
- t0atlas: 1.4020468

Positioning times

- Median positioning times: 2.823458s, average: 4.6153875s
- Percentage of positioning times > 20s: 3.3431619 %
- Percentage of contiguously read files (positioning times < 0.1s): 25.7030529 %

default service class:
- Median positioning times: 4.110609s, average: 6.6430978s
- Percentage of positioning times > 20s: 6.0387081 %
- Percentage of contiguously read files (positioning times < 0.1s): 18.0605289 %

**t0atlas service class:**

- Median positioning times: 2.1973055s, average: 2.987429s
- Percentage of positioning times > 20s: 1.1790274 %
- Percentage of contiguously read files (positioning times < 0.1s): 31.8388962 %

**Per-drive, per-library performance**

Check for tape drive and overall library performance figures (positioning time, transfer speed)
positioning time, by drive

positioning time, by library

library
- IBM1L8
- IBM355
- IBM3JD
- IBM455
- IBM4JD
- T10KD5
- T10KD6

libraries
- IBM1L8
- IBM355
- IBM3JD
- IBM455
- IBM4JD
- T10KD5
- T10KD6
Repeated file access

- how many files have been recalled multiple times? What was the time interval distribution between recalls for these?
- how does this differ between “default” and “t0atlas”?

General:

From the 4196626 files read, 489086 files (11.654267 %) have been recalled more than once. Out of these, 83160 files (1.9815919 %) have been recalled from both service classes.

default service class:

From the 1868860 files read, 86665 files (4.637319 %) have been recalled more than once.

t0atlas service class:

From the 2327766 files read, 319261 files (13.7153391 %) have been recalled more than once.

- For the files recalled more than once within 2018, what is the histogram of standard deviation in repeated access time?
- For the files recalled more than once within 24h, how does the histogram look like?
- Are there any differences between t0atlas and default?
File retrieval cohesion

- What is the spread in creation time of files retrieved within a single day? How collocated are recalls with regard to creation time?

Basis is the daily standard deviation of the nsfileID of retrieved files. Average nsfileID creation/day (across all of CASTOR): ~230K files (stable over last 4 years), therefore showing distance divided by 230K files, so roughly corresponding to the number of days between the creation time of these files.

- standard deviation (in ~days) for “default” svcclass (daily plot, histogram)
- standard deviation (in ~days) for “t0atlas” svcclass (daily plot, histogram)
Spread in creation time of recalled files, default svcclass, 500d

Spread in creation time of recalled files, t0atlas svcclass
Per-day recalled nsfileID boxplots

How far away in time are the files being recalled

Average number of distinct creation time days processed within a day

For a given day, in addition to the standard creation time deviation (which represents the overall spread in creation time between all files being processed) it is interesting to know from how many distinct creation days there are files being processed (which gives an idea of how grouped these files are, in addition to the overall creation time spread)

- default: average of 49.5220994 different creation days
- t0atlas: average of 11.7322835 different creation days
Accessed ATLAS datasets

ATLAS dataset names have been identified by extracting the last directory path entry before each file name. Rucio datasets are identified by a naming convention (see https://gitlab.cern.ch/cta/CTA/issues/461 for details). For each of the datasets that have been accessed (at least one file recalled), the following characteristics have been extracted:

- Total number of datasets: 12461, number of Rucio datasets: 12413 (99.6147982 %)
- Total number of files across all datasets: 14.265152 million
- Total volume across all datasets: 3.052964510^4 TB
- Average number of files per dataset: 1144.7838857; median: 55
- The mean and median differ substantially: There are many datasets with just 1-2 files (27.5900811 %) and there is a tail of a few, long datasets (see histograms below)
- Average volume per dataset: 2.4500156 TB; median: 0.0300045 TB.
- Mean and median differ again greatly as a consequence of many datasets being very small.
- Average spread in time (first to last creation time): 2.5432881 days, median: 0.0725652 days.
- Same observation as before. In addition, in the cases where a file has been re-created (e.g. re-import from T1 following a file loss), the spread will become artificially large as the creation time of the recovered file may be much younger than the rest of the files. See also the histograms below.
Dataset access patterns

In the section above we have identified the ATLAS datasets that were accessed during 2018. How was the actual access to them, as a function of the service class (default vs t0atlas)?

default:

- total datasets accessed by default: 9482
- total files in these datasets: 12752024; fraction accessed via default: 14.655399 %
- total volume in these datasets: 2.797084610^4 TB; fraction accessed via default: 13.2914095 %
- Average number of existing files per accessed dataset: 1344.8664839; median: 90

t0atlas:

- total datasets accessed by t0atlas: 3214
- total files in these datasets: 2218264; fraction accessed via t0atlas: 104.9363827 %
- total volume in these datasets: 4049.016156 TB; fraction accessed via t0atlas: 102.3750511 %
- Average number of existing files per accessed dataset: 690.1879278; median: 2. Note the low median - 50% of the t0atlas accessed datasets have less or equal than 2 files.
distinct datasets accessed per day

date

Distinct days on which a given dataset is accessed (default)
Mount excess per dataset

How many more mounts did we issue to what it would be neccessary to read out a dataset?

Note: Both tape pools (atlas_raw and atlas_prod) are configured with up to 10 concurrent write drives each. So if there is sufficient data to be written, up to 10 drives will be mounted in parallel for writing. Thus datasets can be spread across 10 tapes, which means reading them back implies 10 mounts (in parallel or sequentially) - for those datasets that have sufficient files of course.

In terms of mount counts, it is difficult to go down below the count of accessed datasets, unless the datasets themselves are grouped within a single tape which is not usually expected. So what is the baseline “excess” factor?

- Number of mounts; number of datasets accessed; “excess” factor (mounts per dataset accessed):
  - default: 73788 ;9482 ; 7.7819026
  - t0atlas: 7670 ;3214 ; 2.3864343
Concurrent number of drives used

How many drives were concurrently used for read by ATLAS?

- default: mean: 3.0763252 concurrent mounts; median: 2 concurrent mounts
- t0atlas: mean: 7.2113379 concurrent mounts; median: 5 concurrent mounts

See also the histograms below showing the frequency against the number of concurrent mounts. The frequency is understood as the amount of minutes where there were N concurrent mounts.

How much total time were the drives active?

t0atlas:

- At least one drive: 8.9944825 %
- At least two drives: 7.4604262 %
- At least five drives: 5.2511416 %
- At least ten drives: 2.4743151 %

default:

- At least one drive: 27.9809741 %
- At least two drives: 16.69121 %
- At least five drives: 5.5064688 %
- At least ten drives: 1.3850837 %
concurrent mounts over time, t0atlas

concurrent mounts over time, default
Relation of number of files read per mount and service speed

The service speed (calculated as volume divided by the sum of transfer and positioning times) is expected to grow with the number of files recalled per tape. This is visualized in the plots below. For getting to high per-drive performances, the number of files read per tape (and therefore the number of files per dataset) would need to grow considerably. Even for t0atlas (where there is systematic organised workflow processing), the average number of files read per dataset is orders of magnitude below the threshold of several thousand files required to achieve increased per-drive performance. In addition, higher collocation on less tapes means that the parallel processing speed will reduce (as less parallel tape mounts can be exploited).
Looking at tape read mounts since 2017:

c2repack (10TB tapes, LHC VO's):
c2atlas:
Queue Analysis

In addition to the number of tape mounts, the average queueing waiting time is a significant metric for understanding what the overall latency for data access is.

Today, in CASTOR, a two-level queueing is used: Tape read requests queue up first within the stager and then on VDQM. The stager holds back submitting tape mounts to VDQM to not exceed a given number of parallel mounts. However, the stager doesn’t distinguish whether submitted jobs are running or just queued (e.g. due to busy library), which may cause a tarpit effect (jobs accumulate on busy/slow/unavailable libraries). As a workaround, jobs are occasionally (manually) released to VDQM, sometimes creating an “avalanche” effect. VDQM tape queues are processed (per library) in FIFO order across all stagers following a FIFO ordering with all VO's having the same priority.

Median and mean latency:

- default: median: 0.4098723 days; mean: 1.7392657 days
- t0atlas: median: 0.2878841 days; mean: 0.6358457 days

Histograms, CDF plots and boxplot graphs:
ECDF for recall latency, t0atlas

recall latency, days

date

fraction

0.00

0.25

0.50

0.75

1.00

days

0

1

2

3

4

svcclass

default
t0atlas