Tape Staging Test

First test on
“How much we can really get out from TAPEs?”

Most of the material already presented last week at the ATLAS Sites Jamboree (plus some more observation)

Why

- In the next years disk space will be scarce
  - Foreseen more CPU growth than disk one
  - Till now, Tape was mostly used to store RAW data (and reprocessing) and backup selected older datasets
- The C-RSG *suggested* the experiments to exploit more tapes
  - We need to investigate how much they can be used, for which workflows
- Many possible scenarios, some ideas:
  - Mixed disk/tape input source - one dataset replica on disk, one on tape - use both for production simultaneously to achieve desired production throughput and resilience
  - Disk for newer, tape for older data - several disk replicas for new data, tape only copies of data older than eg 6 months
  - Using disk as cache only (mostly) storage

“Tape” and “BNL Site Report” @ Jamboree

Tomas Javurek

https://indico.cern.ch/event/579473/contributions/2429473/attachments/1398521/2133259/TAPE_resources_at_ATLAS.pdf

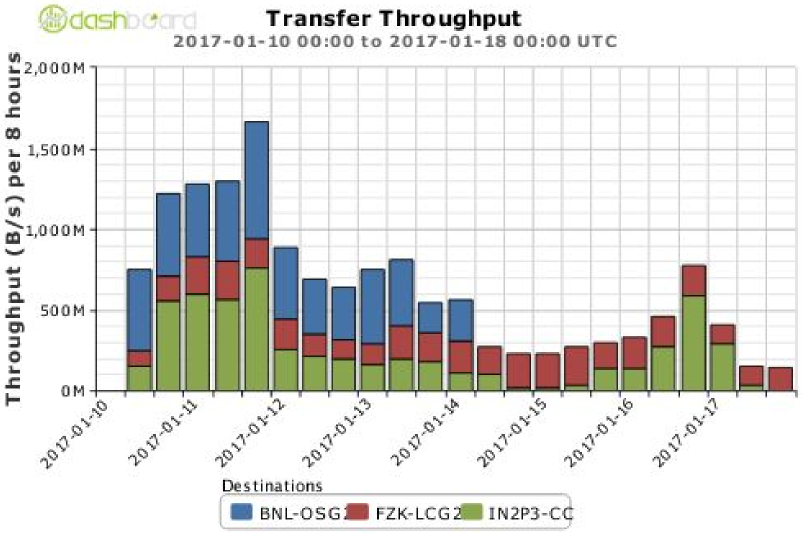
Xin Zhao

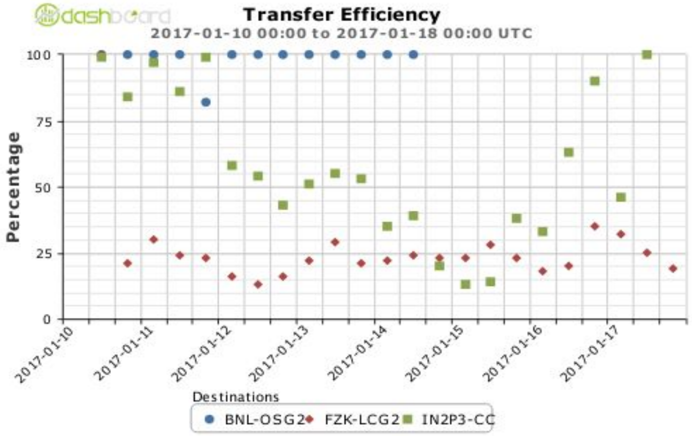
https://indico.cern.ch/event/579473/contributions/2429450/attachments/1397346/2131772/site_report.pdf

From Tomas

Staging tests BNL, FZK, IN2P3-CC

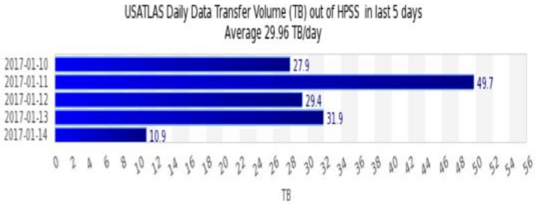
- TAPE → DISK at given site
- 150 TBs of AODs for each of the sites
- 310k files ~ 1.5 GB/file
- Results:
    - BNL: 4 days ⇒ 430 MB/sec.
    - IN2P3-CC: 7 days ⇒ 250 MB/sec.
    - FZK: 100MB/7 days (ongoing) ⇒ 165 MB/sec.
- FZK+BNL+IN2P3-CC ~ 50% MoU ⇒
    - (430+250+165)✖2~1.7 GB/sec from all tapes ⇒
        - 7 days to stage 1 PB (optimistic estimate)

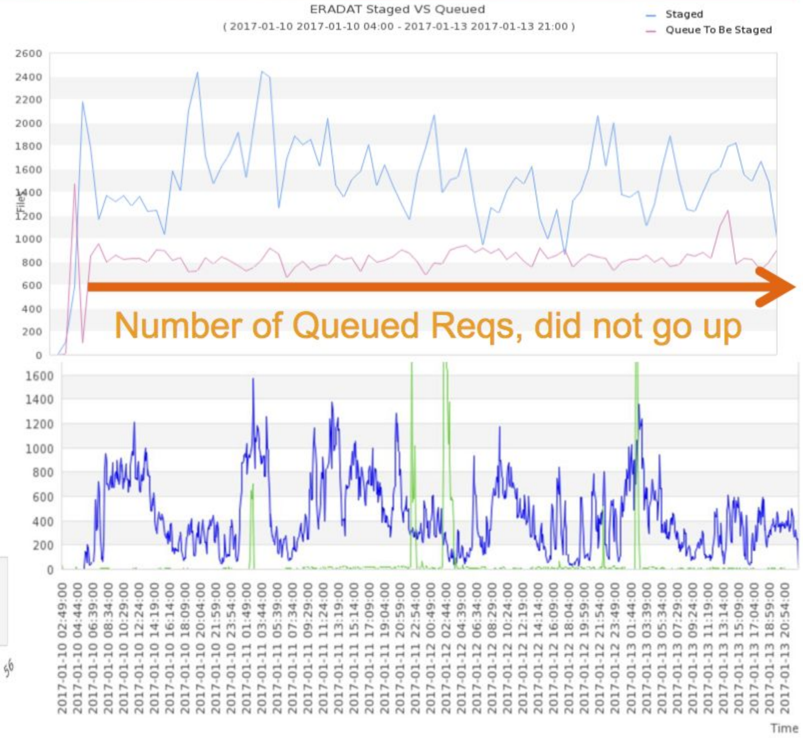




Staging Test from Tapes

- Staging Test (Jan 10-14):
  - replicate 150TB AODs from DATATAPE to DATADISK
  - ~1500 new reqs added, per hour
  - Transfer rate : not constant, average at 385MB/s



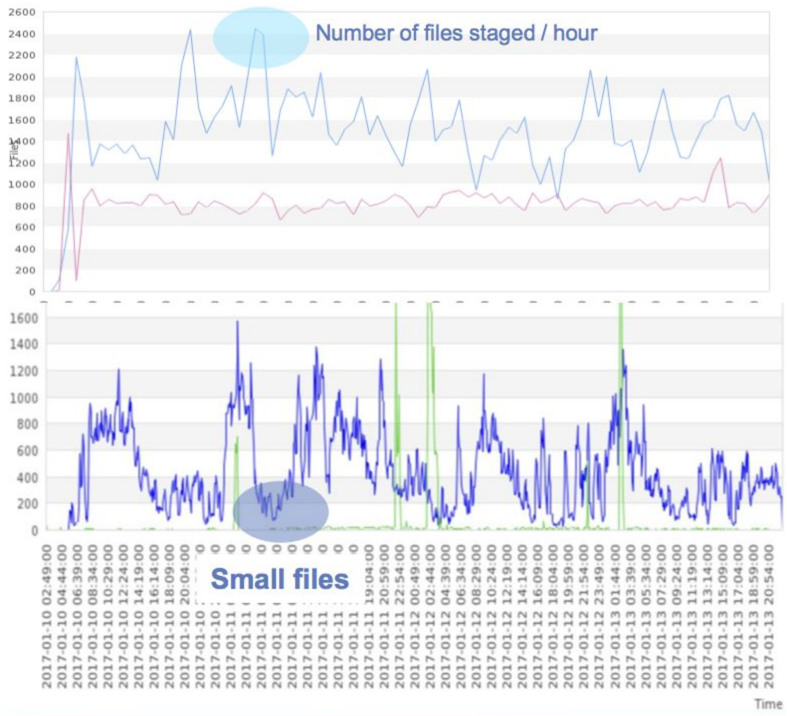


Number of Queued Reqs, did not go up

Staging Test from Tapes

- Improvements ?
  1. Increase File size
  2. Bulk request

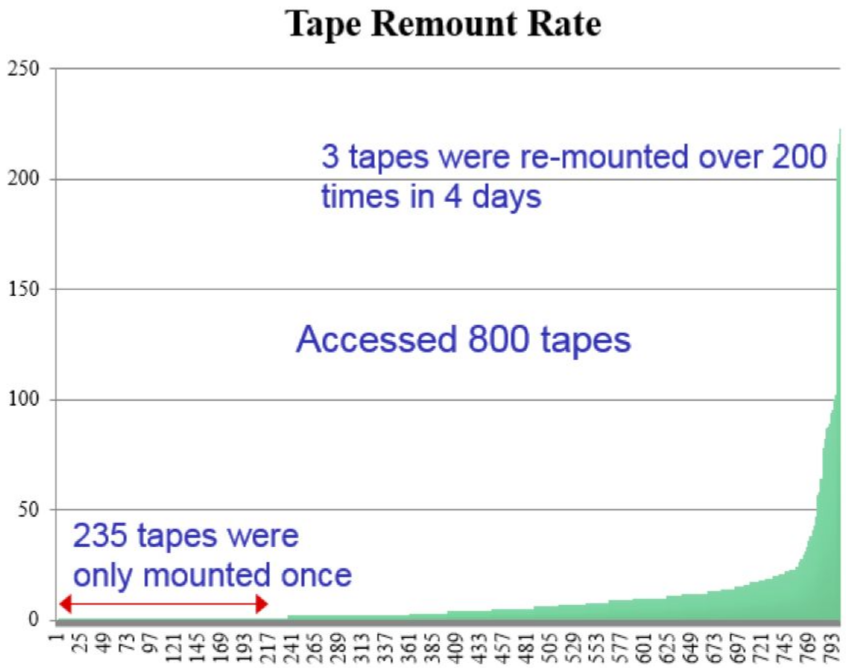
Number of Files Queued VS Staged /hour



Staging Test from Tapes

- Improvements ?

  1. Increase File size
  2. Bulk request: BNL tape system optimizer reduces tape remounts



3 tapes were re-mounted over 200 times in 4 days

Accessed 800 tapes

235 tapes were only mounted once

Observations 1/2

- ATLAS capped to 5k files submitted to FTS
  - This cap can be removed
- Number of files into FTS: 200.
  - It could be increased
- To be discussed (between experiment, FTS and sites) how FTS is “throwing” into the tape systems the requests
  - General statement is that “the more together the better” to better optimize for tape mounts and seek latency
- ATLAS can think about more fancy things
  - If it’s “clear” which are the data that wants to be re-read, tape(file) families is one
  - File size increased, from 1.5GB average to e.g. 10? More? (To be understood transfers if file size is too big)

Observations 2/2

- Monitoring (as discussed in December FTS3 steering meeting) is very approximate
    - As of today the best is the ATLAS DDM dashboard
    - Timeouts are affecting an easy analytics

Next

- We can redo the test.
  - We want to redo the test!
  - With other experiments?
  - Good willing Tier1s who want to check what’s happening to their tape system during the test?
- We plan to remove the 5k file cap
  - Throw everything in FTS
- Proposal is to use 5 Tier1s (also Tier0? Could be…)
- Other suggestions?

Summary

- In general:
  - Tape will be used much more frequently for both production and analysis
- Storage latency should be an integral part of the workflow management system (Panda, Rucio, FTS), eg
  - Hot storage, Warm storage, Cold storage
- Optimizing the production and analysis throughput for different storage technologies is crucial
  - In similar why as the network latency and throughput which are already integrated in ATLAS computing
- The aggregated tape throughput should be "comparable" to the current disk/network usage in ATLAS to make the usage of "cold storage" efficient