Update on ATLAS Data Carousel R&D

Xin Zhao (BNL)

WLCG Archival Storage WG

Sep 27th, 2018

BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

Outline

- Ongoing tape test at all ATLAS T1s
- Preliminary results
- Discussion points
- Next steps

** For overview of the R&D, please check out the previous talk*
** Team effort, credit goes to ADC and site experts.*

Ongoing Tape Test at T1s

- Goal is to establish baseline measurement of current tape capacities
- Run the test:
  - Rucio → FTS → Site: staging files from tape to local disk (DATATAPE/MCTAPE to DATADISK)
  - Data sample
    - About 100TB~200TB AOD datasets, average file size 2~3GB
  - Bulk mode
    - Sites can request throttle on incoming staging requests (3 sites)
  - With concurrent activities from tape writing and other VOs
- Status
  - Done at BNL, FZK, PIC, INFN, TRIUMF, CCIN2P3, NL-T1 and RAL
  - Ongoing: NRC
  - Upcoming: NDGF (scheduled for next week)

Preliminary Results

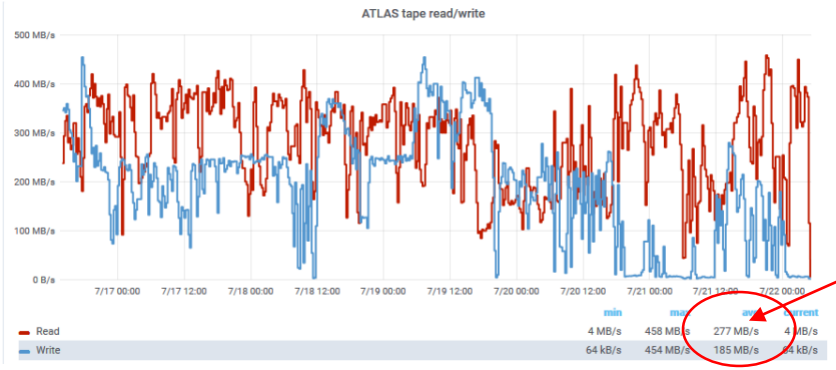
- Throughput

Site	Tape Drives used	Average Tape throughput	Stable Rucio throughput	Test Average throughput
[1]BNL	31 LTO6/7	1~2.5GB/s	866MB/s	545MB/s (47TB/day)
FZK	8 T10KC/D drives	~400MB/s	300MB/s	286MB/s (25TB/day)
INFN	2 T10KD drives	277MB/s	300MB/s	255MB/s (22TB/day)
PIC	5~6 T10KD drives	500MB/s	[2]380MB/s	400MB/s (35TB/day)
[1]TRIUMF	11 LTO7 drives	1.1GB/s	1GB/s	700MB/s (60TB/day)
CCIN2P3	[3]36 T10KD drives	2.2GB/s	3GB/s	2.1GB/s (180TB/day)
SARA-NIKHEF	10 T10KD drives	500~700MB/s	640MB/s	630MB/s (54TB/day)
RAL			2GB/s	1.6GB/s (138TB/day)

[1] dedicated to ATLAS
[2] with 5 drives, later increased to 6 drives
[3] 36 is the max number of drives, shared with other VOs who were not using them during the test

Preliminary Results (continued)

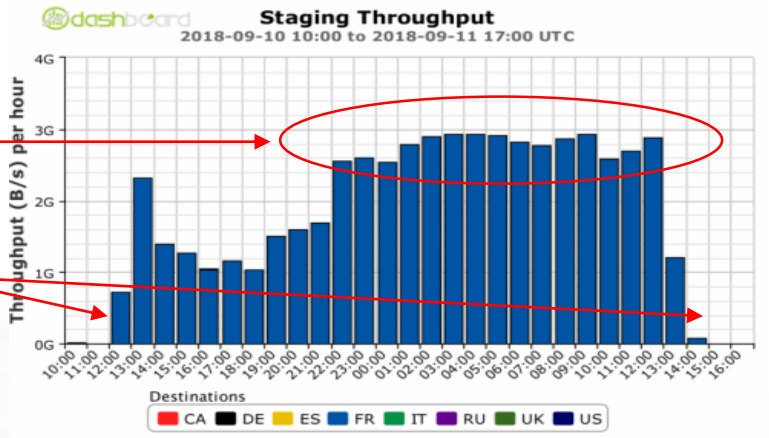
- How are various throughputs calculated ?



(Average) tape throughput is from site tape monitoring directly

Stable Rucio throughput is from Rucio dashboard, over a “stable” run time

Test average throughput = total volume/total walltime, of the test



Discussion Points (1/3)

- Tape frontend --- current bottleneck!
  - Limiting number of incoming staging requests
  - Limiting number of staging requests to pass to backend tape
  - Limiting number of files to retrieve from tape disk buffer
  - Limiting number of files to transfer to the final destination
- Improvements on hardware
  - Bigger disk buffer on the frontend
  - More tape pool servers
- Improvements on software
  - Some dCache questions here
  - Other HSM interface: ENDIT

Discussion Points (2/3)

- Writing is important
  - Good throughput seen from sites who organize writing to tape
    - grouped by datasets on tape
    - Full tape reading, near 0 remounts
    - Fine-grained file families ?
      - Balance between file grouping and tape space usage
  - ATLAS working on
    - Increase file size to tape, target at 10GB
  - Discussion between dCache/Rucio: Rucio provide dataset info in the transfer request ?
    - Such info can help tape system group files before writing to tape

Discussion Points (3/3)

- Bulk request limit
  - 3 sites request a cap on the incoming staging requests from upstream (Rucio/FTS)
    - Consideration factors --- limit from tape system itself, size of disk buffer, load the SRM/pool servers can handle, etc
  - Rucio can set limit per (activity&destination endpoint) pair
    - Adding another knob on limiting the total staging requests, from all activities
  - FTS can set limit on max requests
    - Controlled by each instance? Per link ?
- Find it easier to control from the Rucio side (for ATLAS)
- Can site admin have permission to change the limit themselves ?

Next Steps

- Continue tape test to cover all T1s (very close)
- Rerun test
  - Upon site request
    - For example, after hardware and configuration improvements
  - Repeat the test on one or two sites, changing destination from local to remote DATADISK
- Staging test driven by WFMS
  - All T1s will involve
  - ATLAS adding whole dataset pre-staging to the WFMS
  - Details being worked out

Backup Slides

- ***Estimate*** on data volume required by current ATLAS derivation campaign
  - 260TB/day input AOD data, if run on 100k cores
    - MC data volume not included
    - The same input file will be used multiple times, for different physics studies.