ATLAS Data Carousel

Xin Zhao (BNL)

DOMA general meeting, November 28th, 2018

BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

Outline

- ATLAS data carousel R&D
- Staging test at all ATLAS tape sites
- Discussion points
- Next steps

** Collaborative effort, credit goes to ADC and site experts.*

Data Carousel: Introduction

- To study the feasibility to run various ATLAS workloads from tape
  - Facing the data storage challenge of HL-LHC, ATLAS started this R&D project this June

- *By 'data carousel' we mean an orchestration between workflow management (WFMS), data management (DDM/Rucio) and tape services whereby a bulk production campaign with its inputs resident on tape, is executed by staging and promptly processing a sliding window of X% (5%?, 10%?) of inputs onto buffer disk, such that only ~ X% of inputs are pinned on disk at any one time.*

Data Carousel: Objectives

- Rucio
  - Improve tape usage, e.g.bulk requests to tape, with size tailored to site parameters
- FTS
  - Optimize scheduling of transfers between tape and other storage endpoints, e.g. dedicated FTS instance for tape recall requests
- SE endpoints (dCache, StoRM, Castor, etc)
  - Any bottlenecks and possible improvements on interfacing with respective tape backend ?
- Optimize data placement to tape
  - “do writing right” is the key?
  - Use tape families for files to be read back multiple times
  - Larger file sizes preferred
- Evolving tape scheduler
  - Support high priority, low latency request ?
- PS2
  - Study and optimize prompt processing of data as it appears off of tape --- process immediately when X% of a dataset is staged ?
- WLCG Archival Storage WG
  - Work together, define realistic expectations and evaluate possible evolution

- Touches many aspects of ADC …

Data Carousel: The (*original*) Plan

- First phase
  - Understand tape system performance at all T1 sites
  - Identify workloads (start with derivation), and evaluate performance based on current systems
    - Tape available at ~ 10 sites, while processing happens everywhere
    - Performance with tape vs disk
- Second phase
  - Address issues found in phase 1
  - Deeper integration between workload and data management systems (PanDA/PS2/Rucio)
- Third phase
  - Integrate with production system and run production, at scale, for selected workflows

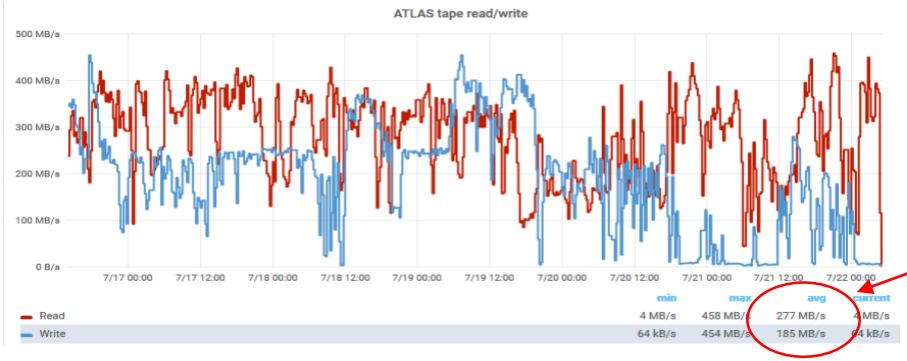
In reality, more of iterative process:
tape test → bottleneck → improvement → tape test → next bottleneck → …

Staging Test at ATLAS Tape Sites

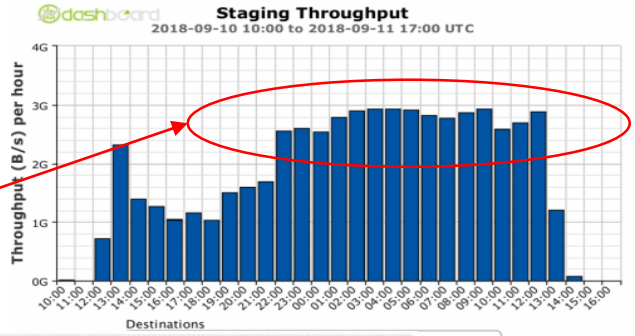
- Goal is to establish ***baseline*** measurement of current tape capacities
- Run the test:
  - Rucio → FTS → Site: staging files from tape to local disk (DATATAPE/MCTAPE to DATADISK)
  - Data sample
    - About 100TB~200TB AOD datasets, average file size 2~3GB
  - Bulk mode
    - Sites can request throttle on incoming staging requests (3 sites)
  - With concurrent activities (production tape writing/reading and other VOs)

Tape Test: Throughput (explained)

- How are various throughputs calculated ?

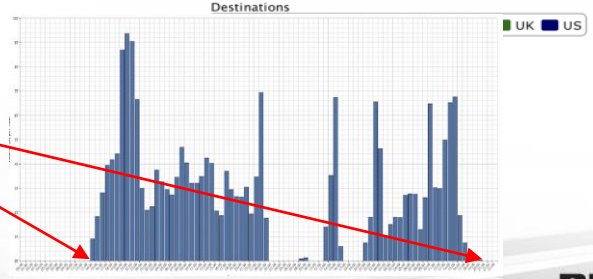


(Average) tape throughput is from site tape monitoring directly



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

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



Tape Test : Throughput

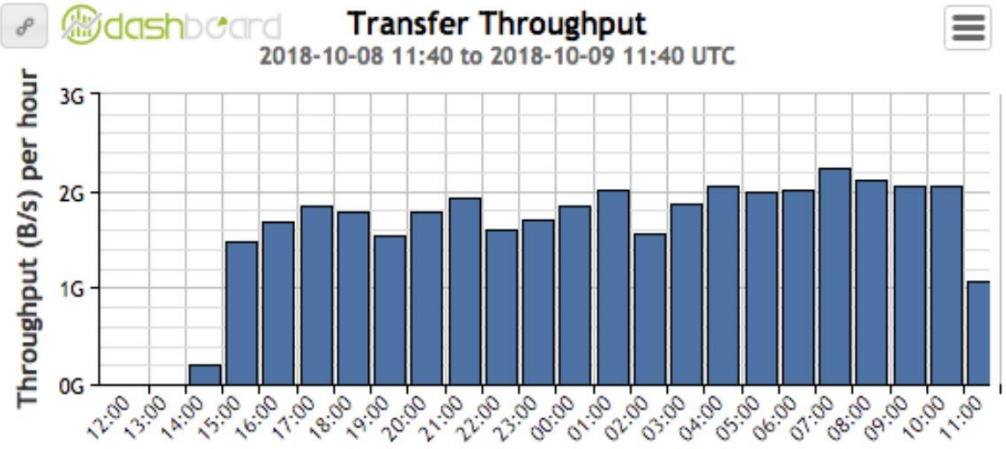
Site	Tape Drives used	Average Tape (re)mounts	Average Tape throughput	Stable Rucio throughput	Test Average throughput
[1]BNL	31 LTO6/7 drives	2.6 times	1~2.5GB/s	866MB/s	545MB/s (47TB/day)
FZK	8 T10KC/D drives	>20 times	~400MB/s	300MB/s	286MB/s (25TB/day)
INFN	2 T10KD drives	Majority tapes mounted once	277MB/s	300MB/s	255MB/s (22TB/day)
PIC	5~6 T10KD drives	Some outliers (>40 times)	500MB/s	[2]380MB/s	400MB/s (35TB/day)
[1]TRIUMF	11 LTO7 drives	Very low (near 0) remounts	1.1GB/s	1GB/s	700MB/s (60TB/day)
CCIN2P3	[3]36 T10KD drives	~5.33 times	2.2GB/s	3GB/s	2.1GB/s (180TB/day)
SARA-NIKHEF	10 T10KD drives	2.6~4.8 times	500~700MB/s	640MB/s	630MB/s (54TB/day)
[4]RAL	10 T10KD drives	n/a	1.6GB/s	2GB/s	1.6GB/s (138TB/day)
[5]NDGF	10 IBM Jaguar/LTO-5/6 drives, from 4 sites	~3 times	200~800MB/s	500MB/s	300MB/s (26TB/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
[4] 8 drives dedicated to this test. Will have 22 shared with other VOs in production.
[5] federated T1, 4 physical sites have tapes

Tape Test : Throughput (continued)

- T0 CTA test
  - Not a full T0 test. Only the CTA part, a validation/commission test, using a limited set of T10KD drives

ATLAS stage out test
- eosctaatlaspps to eosatlas
- 200TB ~90k files
- 3 large FSTs
- 6-10 tape drives
- FROM TAPES



Tape Test : Throughput (continued)

- Results is better than expected
  - ~600TB/day total throughput from all T1s, under “as is” condition
  - Can we repeat it in real production environment ?
- Sites found this test useful
  - System tuning, misconfiguration fixes …, for better performance
  - Bottlenecks spotted, for future improvements
  - Test on prototype system, for production deployment

Discussion Point :Tape frontend (1/3)

- One bottleneck for many (but not every) sites !
  - 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

Discussion Point :Tape frontend (2/3)

- Most of the issues/failures happened at this layer

code	Sample	Total /13564
201	TRANSFER [110] TRANSFER Transfer canceled because the gsiftp performance marker timeout of 360 seconds has been exceeded, or all performance markers during that period indicated zero bytes transferred	13090
63	TRANSFER [5] TRANSFER HTTP 500 : Unexpected server error: 500	201
127	STAGING [70] error on the bring online request: [SE][StatusOfBringOnlineRequest][SRM_INTERNAL_ERROR] Failed to abort transfers	100
132	TRANSFER [70] SOURCE SRM_GET_TURL error on the turl request : [SE][StatusOfGetRequest][SRM_INTERNAL_ERROR] Pin operation timed out	43
80	TRANSFER [13] TRANSFER Authentication error, reached maximum number of attempts	25
118	SOURCE [70] Error reported from srm_ifce : 70 [SE][Ls][SRM_INTERNAL_ERROR] Request to [>SpaceManager@local] timed out.	24
100	SOURCE [70] Error reported from srm_ifce : 70 [SE][Ls][SRM_INTERNAL_ERROR] Failed to abort transfers	21
174	TRANSFER [110] SOURCE SRM_GET_TURL srm-ifce err: Connection timed out, err: [SE][StatusOfGetRequest][ETIMEDOUT] httpg://srmatlas.pic.es:8443/srm/managerv2: User timeout over	16
451	STAGING [5] error on the bring online request: [SE][StatusOfBringOnlineRequest][SRM_FAILURE] Failed to pin file [rc=10011,msg=org.springframework.dao.CannotSerializeTransactionException: PreparedStatementCallback; SQL [UPDATE pins SET state = ?,request_id = ? WHERE id = ?]; ERROR: could not serialize access due to concurrent update; nested exception is org.postgresql.util.PSQLException: ERROR: could not serialize access due to concurrent update].	15
218	TRANSFER [70] DESTINATION SRM_PUTDONE call to srm_ifce error: [SE][PutDone][] httpg://srmatlas.pic.es:8443/srm/managerv2: CGSI-gSOAP running on fts800.cern.ch reports Error reading token data header: Connection closed	13
225	TRANSFER [70] DESTINATION SRM_PUTDONE call to srm_ifce error: [SE][PutDone][] httpg://srmatlas.pic.es:8443/srm/managerv2: CGSI-gSOAP running on fts800.cern.ch reports Error reading token data header: Connection reset by peer	10
240	TRANSFER [70] DESTINATION SRM_PUT_TURL srm-ifce err: Communication error on send, err: [SE][GetSpaceTokens][SRM_INTERNAL_ERROR] httpg://srmatlas.pic.es:8443/srm/managerv2: Authentication failed (server log contains additional information).	4
44	TRANSFER [110] TRANSFER Operation timed out	1
47	TRANSFER [112] TRANSFER (Neon): Unknown error.	1

13K staging failures due to GFTP Performance Markers issues
Hundreds staging failures due to SRM issues

- Retries will get all the requests done eventually.

Discussion Point :Tape frontend (3/3)

- Improvements on hardware
  - Bigger disk buffer on the frontend
  - More tape pool servers
- Improvements on software
  - Feedbacks to dCache team
  - Other HSM interface: ENDIT ?

Discussion Point: writing (1/2)

- Writing is important
  - Better throughput seen from sites who manage writing to tape in more organized way
  - Usually the reason for performance difference between sites with similar system settings

Discussion Point: writing (2/2)

- Write in the way you want to read later
  - File family is good feature provided by tape system, most sites use it
  - There are more … group by datasets!
    - Full tape reading, near 0 remounts observed with sites doing that
    - Discussion between dCache/Rucio: Rucio provide dataset info in the transfer request ?
- File size
  - ADC working on increasing size of files written to tape, target at 10GB
  - Could be a big improvement to tape throughput

Discussion Point: bulk request limit (1/2)

- Need knob to control bulk request limit
  - 3 sites requested 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
  - Save on operational cost
    - Autopilot mode, smooth operation
    - Sacrifice some tape capacities

Discussion Point: bulk request limit (2/2)

- Three places to control the limit
  - 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
    - Each instance sets its own limit, need to orchestrate multiple instances
  - dCache sites can control incoming requests by setting limits on:
    - Total staging requests, in progress requests and default staging lifetime
- Find it easier to control from the Rucio side, while leaving FTS wide open

Next Steps (1/2)

- Follow up on issues from the first round test
  - What dCache team can offer ?
  - What tape experts can offer ?
    - tape BoF session at the last HEPiX
- Rerun the test upon site requests
  - after site hardware/configuration improvements
  - different test conditions: destination being remote DATADISK

Next Steps (2/2)

- Staging test in real production environment
  - Can we get the throughput observed from individual site test, in real production environment?
  - Planning
    - ADC discussion on additional pre-staging step in WFMS/DDM, for tasks/jobs with inputs from tape
    - More monitoring needed
    - (Derivation) jobs will run on the grid, not only T1s
    - All T1s will involve
    - Timing will be random
    - ……

Questions ?