Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules

Data access @ CC-IN2P3:
experience and challenges with LHC

Outline

ATLAS Data Carousel R&D
Tests & results
Discussion points
Next steps

CCIN2P3

by **Aresh Vedaee** (12.12.19)

CCIN2P3

ATLAS perspective on the data storage challenge of HL-LHC:

➢ 'Opportunistic storage' basically doesn't exist

➢ Format size reduction and data compression are both long-term goals, require significant efforts from the software and distributed computing teams

➢ Tape storage is 3~5 times cheaper than disk storage, increasing tape usage is a natural way to cut into the gap of storage shortage for HL-LHC
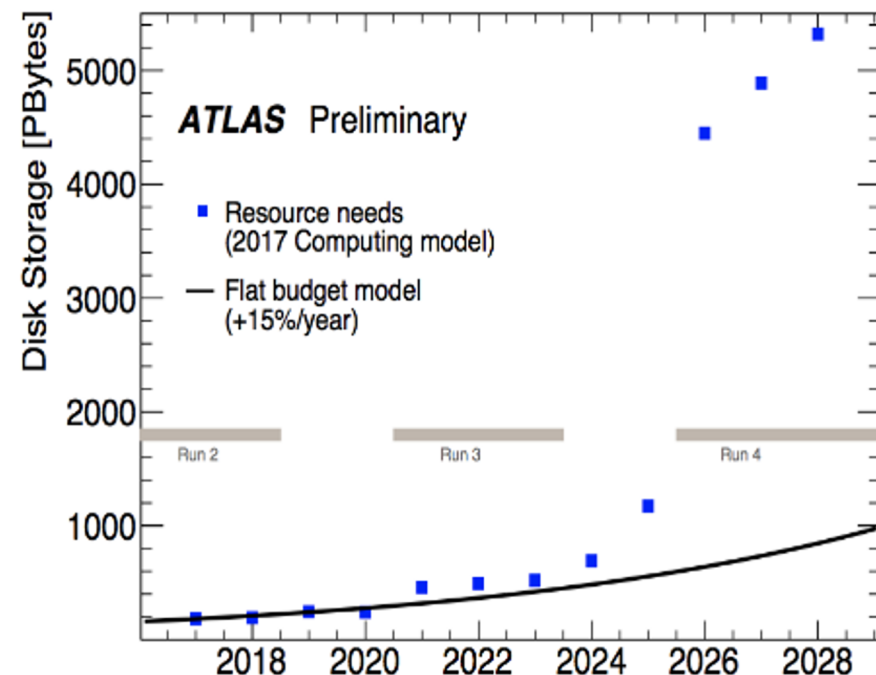
Exploit more tape usage

- Facing the data storage challenge of HL-LHC, ATLAS started this R&D project in June 2018 to study the feasibility to run various ATLAS workloads from tape, based on the Data Carousel Model
    - *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.*
    - Coordinate integration between several agents: PS2, Rucio, FTS, Disk and Tape endpoints

- **First phase** (2018)
  - Understand tape system performance at all T1 sites
  - Identify workflows (starting with derivation) and performance parameters for implementing the 'Data Carousel' mechanism
  - Identify bottlenecks and tune storage systems to boost performance

- **Second phase** (2019)
  - Address issues found in phase 1
  - Deeper integration between workload and data management systems (PanDA/PS2/Rucio)

- **Third phase** (for Run3)
  - Integrate with production system and run production, at scale, for selected workflows

- Run the test (P1R2):
  - 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)
- Status : all done at 10 T1s
  - BNL, FZK, PIC, INFN, TRIUMF, CCIN2P3, NL-T1 and RAL, NRC and NDGF

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

Meilleurs résultats des T1 !

Mais :
- 36 drives utilisés
- Taux remontage : 5,33 x / bande

- Results is better than expected
  - ~600TB/day total throughput from all T1s, under "as is" condition
  - Can we repeat it in real production environment ?

- The test (P2R2) performed during last 2018 RAW reprocessing campaign:
  - Timeline : started on the 8th August 2019
  - Data volume varies by site
  - No warning
  - Job released by ATLAS PS2 after 90% of staged input data
  - Monitoring tool
    - https://bigpanda.cern.ch/datacardash/

Source	Progress plot	DDM Dash	Datasets Active	Datasets Done (+ 90% readiness)	Files Remaining	Files Done
BNL-OSG2_DATATAPE	〰	-->	3	0 (+0)	171	3
CERN-PROD_RAW	〰	-->	3	0 (+3)	0	22886
FZK-LCG2_DATATAPE	〰	-->	1	0 (+0)	1	0

Show 10 ▼ entries   Search:

Campaign	Request ID	TaskID	Status	Total Files	Staged Files	Progress (%)	Source RSE	Time Elapsed	Started At	Rucio Rule
Archive	27866	19899431 〰	staging	22299	22287	100		6 days, 18:11:08.719864	2019-11-28T13:56:03.295620	c5cf5e27841b4e3090bf42ff4e6a05eb
data15_13TeV	23576	18352838 〰	done	58	1	2	BNL-OSG2_DATATAPE	1 day, 21:26:47.140179	2019-12-03T10:40:24.875153	19ff7762384443f081fa98ebb9a92ccc

sites	tape	SE	Tape drives (max. reading)	P1 avg throughput	P1 stable throughput	P2R2 avg throughput (100% staged)	P2R2 avg throughput (90% staged)	Capacity pledge (2019)
CERN	CTA	EOS	(expect 60GB/s for Run3)	2GB/s	2GB/s	763MB/s	1.2GB/s	100%
BNL	HPSS	dCache	22 LTO7 17 LTO6	545MB/s	866MB/s	900MB/s	1.4GB/s	23%
FZK	TSM → HPSS	dCache	30 T10KD	286MB/s	300MB/s	316MB/s	324MB/s	13%
RAL	CTA	Echo	21 T10KD	1.6GB/s	2GB/s	850MB/s	1.3GB/s	13%
CCIN2P3	HPSS	dCache	56 T10KD	2.1GB/s	3GB/s	401MB/s	524MB/s	12%
TRIUMF	Tapeguy	dCache	20 LTO8 12 LTO7	700MB/s	1GB/s	366MB/s	330MB/s	10%
INFN	TSM	StoRM	16 T10KD 19 TS1600	255MB/s	300MB/s	N/A	N/A	8%
NL-T1	DMF	dCache	8 T10KC, 2 T10KD, 10 LTO8	630MB/s	640MB/s	626MB/s	630MB/s	8%
NDGF	TSM	dCache	N/A	300MB/s	500MB/s	214MB/s	371MB/s	6%
PIC	Enstore	dCache	4~6 T10KD	400MB/s	380MB/s	179MB/s	170MB/s	4%

311K files, 0.6PB data in 21 days → AVG THROUGHPUT << 1GB/s << P1R2 (2.1GB/s)

ATLAS ACTIVITY ON HPSS

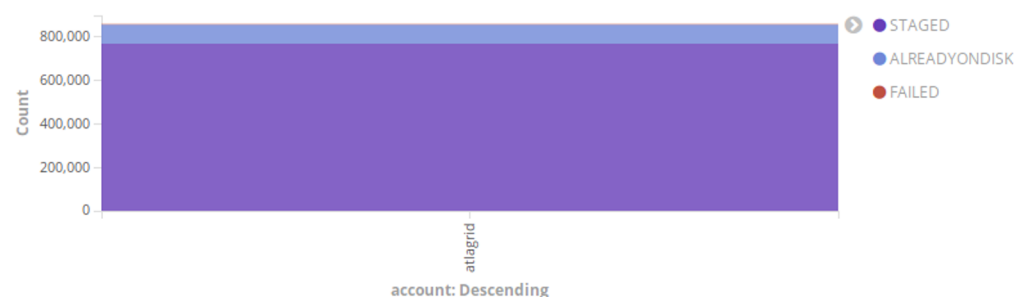**860K files, 1.7PB data in 21 days (~1GB/s) : quite different stats wtr ATLAS report !**

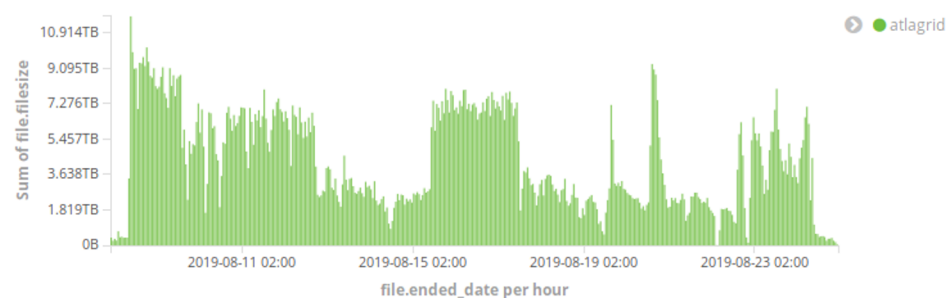## 860,523 1.724PB
Requests   Total Size

Status	Count	File size	Avg File size
STAGED	769,821	1.642PB	2.237GB
ALREADYONDISK	87,764	79.816TB	953.618MB
FAILED	2,938	4.027TB	1.403GB
	860,523	1.724PB	4.572GB

Export: Raw ⬇ Formatted ⬇

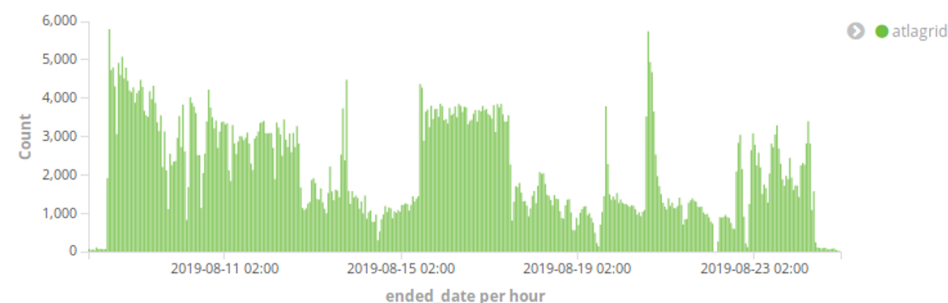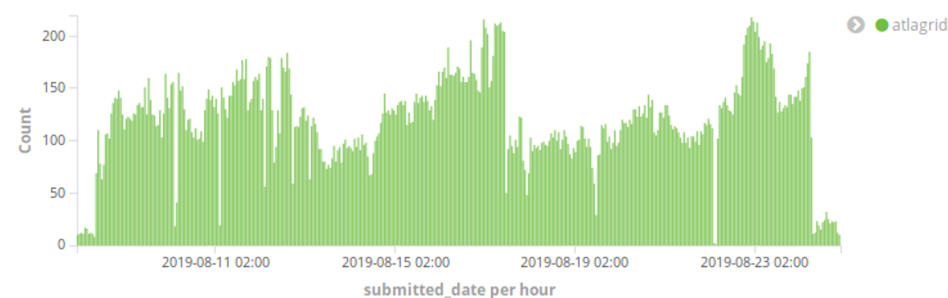TREQS2 : Requetes par utilisateurs

TREQS2: File requests by hour

TREQS2: Stage rate by users

TREQS2: Tape count by users

- **FTS issues @CERN:**
  - FTS Scheduler degraded
    - FTS was not able to schedule transfers between the tape buffer to the disk on time → files got garbage collected → transfer requests failed → FTS optimizer throttled to the minimum parallel transfer requests
  - SRM prepareToGet overkill
    - Redundant tape recalls → nearline failures → FTS optimizer throttling
  - FTS Daemons' crashes
    - Failing to recover "already started" requests → untracked staged files

    The main consequence of this problem was the CC-IN2P3 poor staging stats !

- More details :
  - https://indico.cern.ch/event/843988/contributions/3543611/attachments/1904532/3146356/FTS_Data_Carousel_PostMortem.pdf

- During P2R2 the performance of CC-IN2P3 storage system was lower than P1R2.

- **But it wasn't a site issue nor a performance issue :**
  - As later understood, it was mainly due to several FTS issues @CERN
  - Data carousel tests were dataset specific but site performance varies wrt the whole of concurrent activities within and outside a given VO, sharing same storage resources
  - Besides, the number of requests per recall pool during P2R2 @CC-IN2P3 was even increased wrt P1R2
  - Still not clear what is the required avg throughput

- Future improvements :

  - Move to IBM Entreprise class tape cartridges&drives (Jaguar)

- New test with reprocessing campaign (after FTS upgrade ?) postponed to next week:

  - Site Profile ([google doc](#)) :

    - Bulk size max is 10,000

    - Bulk size min is 5,000 (Default)

    - Time delay between bulk request : 50 % (Might be adjusted during the test according the performances)

    - Results in the next ATLAS SW&C week (10 Feb 2020)?

      - https://indico.cern.ch/event/823140/

- HL-LHC challenge obliges LHC VOs to dramatically revise their data management computing model

  - Among other solutions, greater usage of tape systems

- ATLAS Data Carousel working group is addressing this issue since June 2018:

  - Data Carousel Model prototyped and implemented for efficient tape data access

  - During Phase1 and Phase2 tests, CC-IN2P3 tape system was among the best performers across all T1 sites:

    - There are margin of improvements but it would still be useful to know required/recommended avg throughput from tape
  - CC-IN2P3 is looking forward for the next tests

- https://indico.cern.ch/event/756338/attachments/1723845/2784624/update-atlas-data-carousel-wlcg-wg.pdf
- https://indico.cern.ch/event/651359/contributions/3208536/attachments/1752789/2840658/atlas-data-carousel-GDB-nov2018.pdf
- https://indico.cern.ch/event/865577/contributions/3646827/attachments/1951569/3240027/CC_IN2P3_DATA_CAROUSEL_2019.pdf
- https://indico.cern.ch/event/823341/