Updates on Data Carousel

Alexei Klimentov(BNL), Mario Lassnig(CERN), Xin Zhao(BNL)

pre-GDB, February 9th, 2021

Team effort: bringing together experts from various domains

- Sites
  - Tier0 and all Tier1s storage and tape experts
- Storage and infrastructure experts
  - dCache, CTA and FTS experts
  - Monitoring team (CERN IT and ATLAS specific)
- ATLAS
  - Workflow management team (WFM)
  - Distributed data management team (DDM/Rucio)
  - Distributed production and analysis team (DPAs)
  - Operations team (Grid/T0)
  - ADC coordinators and experts

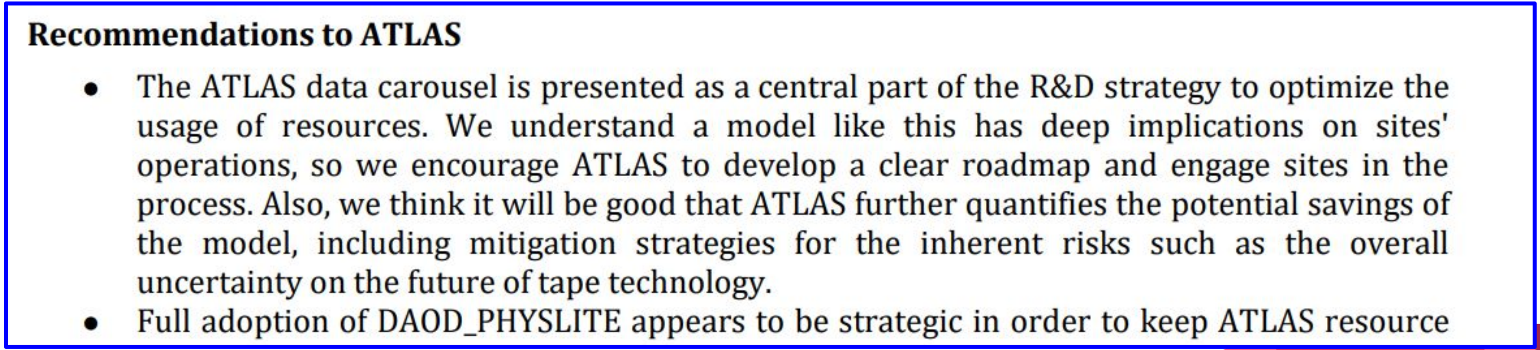
Outline

- What have we achieved
- What do we expect to need
- 2021 plan
- Discussion ……

ATLAS Data Carousel

- Data Carousel R&D
  - Started in June 2018 to tackle the data storage challenge of *HL-LHC*
    - Tape much cheaper than disk (but not free!). Estimates vary a lot - general estimate seems that tape is 3 times cheaper than disk
  - Study the feasibility to get inputs from tape, for various ATLAS workflows
    - So far we have gone through three phases in this R&D (i.e. tape site evaluation, deeper integration of WFM/DDM/facilities and run production at scale)
- All the efforts in this R&D call for close collaboration with sites and any other storage experts!
  - Many findings and improvements not ATLAS specific
  - Generally applicable to experiments, sites and infrastructures

(… from the LHCC review)

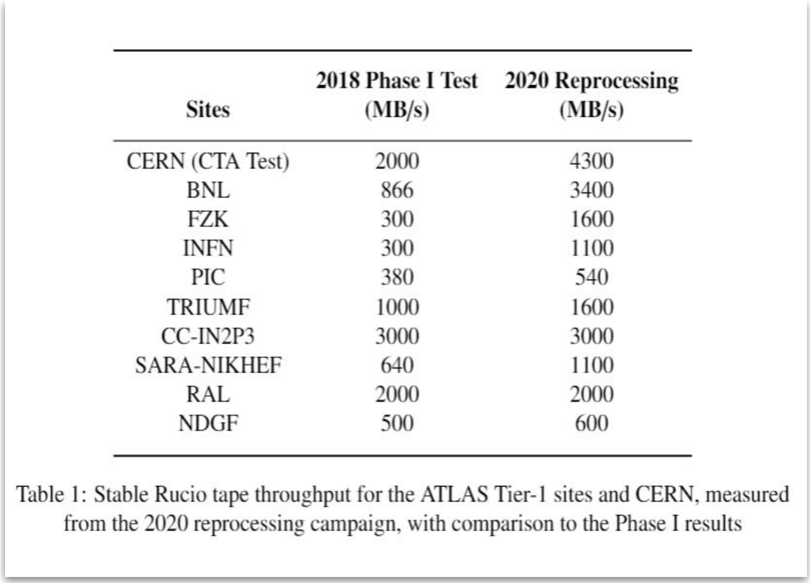


Recommendations to ATLAS

- The ATLAS data carousel is presented as a central part of the R&D strategy to optimize the usage of resources. We understand a model like this has deep implications on sites' operations, so we encourage ATLAS to develop a clear roadmap and engage sites in the process. Also, we think it will be good that ATLAS further quantifies the potential savings of the model, including mitigation strategies for the inherent risks such as the overall uncertainty on the future of tape technology.
- Full adoption of DAOD_PHYSLITE appears to be strategic in order to keep ATLAS resource

What have we achieved ...

- Tape resources integrated with ATLAS workflow
  - Reprocessing and derivation productions run in Data Carousel mode now
    - Fair-share and priority among different activities are being added to the production system.
    - Data Carousel is in day to day production now, not only for big campaigns. Site feedbacks (operational issues, new ideas, …) are welcome!
  - Bulk staging requests are enforced for efficient tape access
    - site staging profile, configured by sites in CRIC
- Current tape throughput
  - From 2020 reprocessing campaigns (18PB RAW data), overall throughput
    - from all T1s : 15GB/s
T1s + T0 : ~20GB/s
  - Great progress since 2018
  - Improvements made in all layers
    - ProdSys2, Rucio, FTS and ***sites (thanks!)***



Sites	**2018 Phase I Test (MB/s)**	**2020 Reprocessing (MB/s)**
CERN (CTA Test)	2000	4300
BNL	866	3400
FZK	300	1600
INFN	300	1100
PIC	380	540
TRIUMF	1000	1600
CC-IN2P3	3000	3000
SARA-NIKHEF	640	1100
RAL	2000	2000
NDGF	500	600

Table 1: Stable Rucio tape throughput for the ATLAS Tier-1 sites and CERN, measured from the 2020 reprocessing campaign, with comparison to the Phase I results

What do we expect to need ...

- Improve tape efficiency
  - Right now our overall tape recall efficiency (T1s) is 30~40%, for RAW data
  - Most probably even lower for other data types
- Grow tape capacity towards the needs of the HL-LHC
  - What’s the expected tape throughput for HL-LHC ? Can sites provide it ?

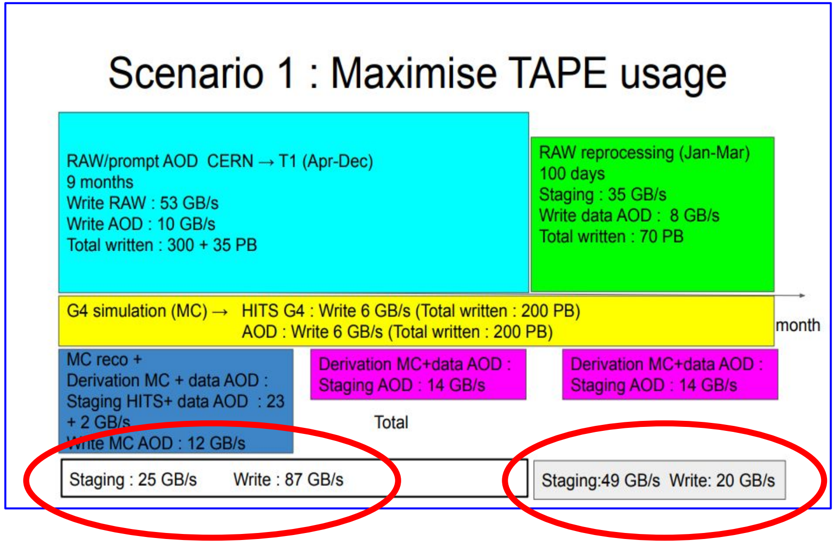
To improve tape efficiency

- Mid-term solution
  - ATLAS will create larger files: Targeting 5~10GB RAW and AOD files for Run3
  - This expect to improve the overall tape recall efficiency to 50%+ in the time scale of Run3
- Long term solution
  - Smart writing
    - Multiple datasets on one tape, but co-locate the same dataset files together
    - Pass metadata from Rucio through FTS to storage
    - Promising results from the tape test at TRIUMF
    - Targeting tape recall efficiency of 70%+ for the long run
  - Some sites already doing this. Some sites have their own plan
  - Also ideas from service providers like dCache team …
- Measure our progress
  - Monitoring improvements, help from sites needed as well

Target tape throughput for HL-LHC ?

- First round of estimation from ATLAS
- Take BNL as an example (23% share)

Throughput	Jan-Mar	Apr-Dec
Read rate (delivered)	12 GB/s	6 GB/s
Read rate (nominal @ 50% eff.)	24 GB/s	12 GB/s
Write rate (delivered)	5 GB/s	20 GB/s
Write rate (nominal @ 80% eff.)	7 GB/s	25 GB/s



Scenario 1 : Maximise TAPE usage

RAW/prompt AOD CERN → T1 (Apr-Dec)
9 months
Write RAW : 53 GB/s
Write AOD : 10 GB/s
Total written : 300 + 35 PB

RAW reprocessing (Jan-Mar)
100 days
Staging : 35 GB/s
Write data AOD : 8 GB/s
Total written : 70 PB

G4 simulation (MC) → HITS G4 : Write 6 GB/s (Total written : 200 PB)
AOD : Write 6 GB/s (Total written : 200 PB)

MC reco +
Derivation MC + data AOD :
Staging HITS+ data AOD : 23 + 2 GB/s
Write MC AOD : 12 GB/s

Derivation MC+data AOD :
Staging AOD : 14 GB/s

Derivation MC+data AOD :
Staging AOD : 14 GB/s

Total

Staging : 25 GB/s Write : 87 GB/s

Staging:49 GB/s Write: 20 GB/s

2021 Plan

- Progressively roll out Data Carousel into production with more workflows
  - Test at scale with more workloads and continue to gain experience and make improvements on various layers (WFM, Rucio, FTS, sites, etc) and collect statistics
    - e.g. more tuning on site staging profile
- Ready to increase file size to tape (RAW, AOD)
  - Automate and test the procedure with Grid jobs and T0 workflow
- Continue with various activities
  - Smart writing
  - Joint exercise with CMS
    - Potentially even beyond tape (disk + network challenges?)
  - Rucio improvements (source throttling, co-location metadata)
  - Continue to refine the estimate of target tape throughput for HL-LHC
  - Replace SRM with new protocol (follow the discussions on DOMA TPC, such as dCache API)
  - New tape sites (NET2) coming up

Questions & comments ...