Joint ATLAS-CMS tape test

Alexei Klimentov[1], Mario Lassnig[2], Xin Zhao[1]
Benedikt Maier[2], Garyfallia (Lisa) Paspalaki[3]
Fernando Garzon[4], Felipe Gomez-Cortes[2]

[1] BNL
[2] CERN
[3] Purdue University
[4] FNAL

DOMA General, 03/24/2021

Outline

- Introduction
- CMS experience
- ATLAS experience
- Summary

Introduction

- ATLAS and CMS performed joint tape tests on March 2021
  - 3 sites involved (PIC, KIT, IN2P3)
  - ATLAS (staging), CMS (staging *and* writing; T0 → T1) ← still ongoing, no results today
- Goal:
  - see how far we can stress things, compare results from previous tests (CMS: 2017 tape test; ATLAS : 2020 reprocessing campaign)
  - see if there will be any issues if multiple VOs access multi-VO tape sites heavily at the same time, from both VO and site perspective
- For CMS: **first tape test after migration to RUCIO** (2017 test was using PhEDEx)
- For CMS: **transition period**: Sites are migrating to new tape systems (Oracle/HPSS to IBM)

- Performing tests with ~250-300 TB of data
- For KIT and PIC the test was performed simultaneously
- Constant monitoring of the test through FTS (also internal monitoring with sites)
- Constant communication between site admins for the test progress

CMS experience

KIT: *3 datasets*: 126 / 105 / 95 TB;
36k / 34k / 34k files

tape technology: **Oracle SL8500 library**, T10k-D drives, **8 drives** available;

PIC: *2 datasets*: 206 / 124 TB;
45k / 37k files

tape technology: **Oracle library**, with T10k-C tapes, **8 drives** available; and **IBM library**, LTO7M8 tapes, **6 drives** available)

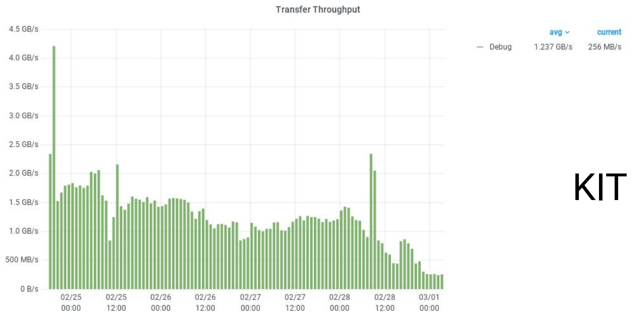
IN2P3: *4 datasets*: 96 / 78 / 73 / 71 TB;
25k / 24k / 15k / 21k files

tape technology: HPSS which provide **45 drives**

Submission of test samples followed the same mechanism as in real life / production

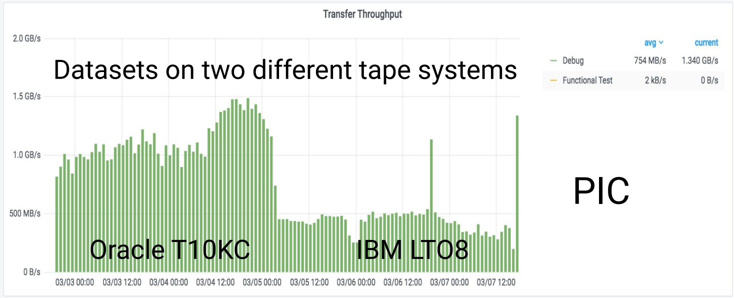
T1_MSS → T1_Disk

- KIT succeeded rate: **1.2 GB/s**
  - (2017 measurement: 200 MB/s)
- PIC succeeded rate: **754 MB/s**
- IN2P3 succeeded rate: **3.56 GB/s**


Transfer Throughput

	avg ⌄	current
— Debug	1.237 GB/s	256 MB/s

KIT


Transfer Throughput

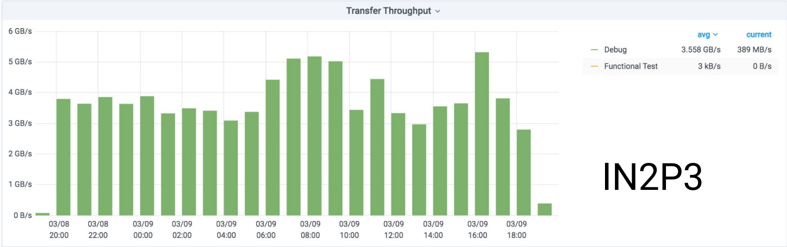
	avg ⌄	current
— Debug	754 MB/s	1.340 GB/s
— Functional Test	2 kB/s	0 B/s

Datasets on two different tape systems

Oracle T10KC

IBM LTO8

PIC

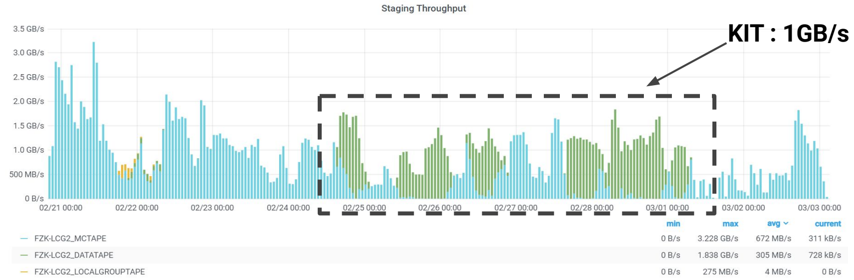

Transfer Throughput ⌄

	avg ⌄	current
— Debug	3.558 GB/s	389 MB/s
— Functional Test	3 kB/s	0 B/s

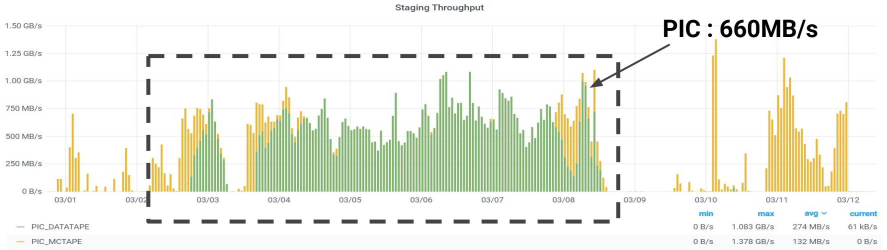
IN2P3

ATLAS experience : test setup and status

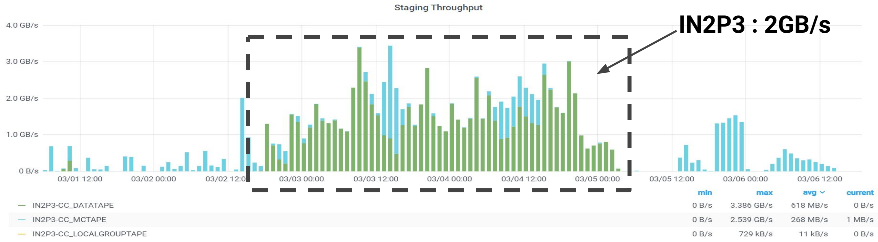
- T1 tape → T1 disk staging test, not writing (production writing not stopped though)
- already done at KIT, PIC and IN2P3
  - Test sample
    - KIT: 22 datasets (AOD), 280TB, 100k files
    - PIC: 29 datasets (AOD/DAOD), 283TB, 104k files
    - IN2P3: 26 datasets (AOD), 307TB, 125k files
  - Mixed with concurrent production staging and migration requests
- Submission of test sample followed the same mechanism as the production one
  - Follow site-staging-profiles
  - Test sample merged with production requests, together throttled by ProdSys2



KIT : 1GB/s



PIC : 660MB/s



IN2P3 : 2GB/s

ATLAS experience : results and observations

- Commonality in staging process between ATLAS and CMS
  - Both use Rucio/FTS
  - Two sites dynamically allocate tape drives between VOs, while PIC has dedicated drives for each VO
- Site staging profile was broken sometimes
  - Not all ATLAS tape access go through Data Carousel at this moment, for example input for user jobs
- Tape throughput & recall efficiency
  - CMS > ATLAS
    - ATLAS has a mix of test sample + production requests, in the staging
    - CMS files are bigger
    - Different way of submitting bulk staging requests between ATLAS and CMS, on the same site
    - Different tape technologies and drives
- Monitoring
  - A lot of information on FTS/DDM/Rucio dashboards, if one knows what/where to look.
  - All sites have tape monitoring, but very few are publicly accessible
  - Lack precise monitoring on tape recall efficiency, at site level

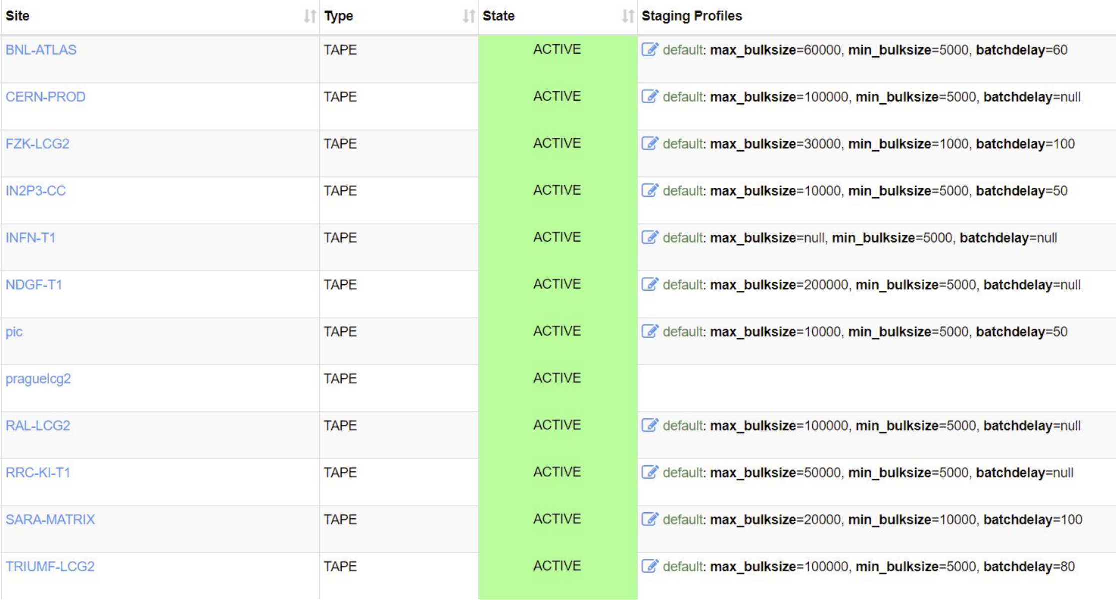
Summary

- Common approach used in tape staging by both ATLAS and CMS (Rucio/FTS)
  - A lot of common ground to work together with, and many improvements are applicable to both sides
- Monitoring is a crucial part
  - Central monitoring for tape activities across VO
    - Rucio team has started to look into building a central place, integrating FTS/Rucio/sites
    - More exposure from site monitoring (only needed for the crucial metrics)
- Site staging profile
  - Already configurable in CRIC
  - Applicable to all VO ? One profile per VO ?
- CMS write test results will be summarized and communicated to ATLAS

Backup

Throttle limit on staging requests at each T1 (ATLAS)

- Defined in CRIC by the site staging profile



Site	Type	State	Staging Profiles
BNL-ATLAS	TAPE	ACTIVE	default: **max_bulksize**=60000, **min_bulksize**=5000, **batchdelay**=60
CERN-PROD	TAPE	ACTIVE	default: **max_bulksize**=100000, **min_bulksize**=5000, **batchdelay**=null
FZK-LCG2	TAPE	ACTIVE	default: **max_bulksize**=30000, **min_bulksize**=1000, **batchdelay**=100
IN2P3-CC	TAPE	ACTIVE	default: **max_bulksize**=10000, **min_bulksize**=5000, **batchdelay**=50
INFN-T1	TAPE	ACTIVE	default: **max_bulksize**=null, **min_bulksize**=5000, **batchdelay**=null
NDGF-T1	TAPE	ACTIVE	default: **max_bulksize**=200000, **min_bulksize**=5000, **batchdelay**=null
pic	TAPE	ACTIVE	default: **max_bulksize**=10000, **min_bulksize**=5000, **batchdelay**=50
praguelcg2	TAPE	ACTIVE	
RAL-LCG2	TAPE	ACTIVE	default: **max_bulksize**=100000, **min_bulksize**=5000, **batchdelay**=null
RRC-KI-T1	TAPE	ACTIVE	default: **max_bulksize**=50000, **min_bulksize**=5000, **batchdelay**=null
SARA-MATRIX	TAPE	ACTIVE	default: **max_bulksize**=20000, **min_bulksize**=10000, **batchdelay**=100
TRIUMF-LCG2	TAPE	ACTIVE	default: **max_bulksize**=100000, **min_bulksize**=5000, **batchdelay**=80