Tape challenge RAL Mar ‘22

Katy Ellis, James Walder, Mark Slater

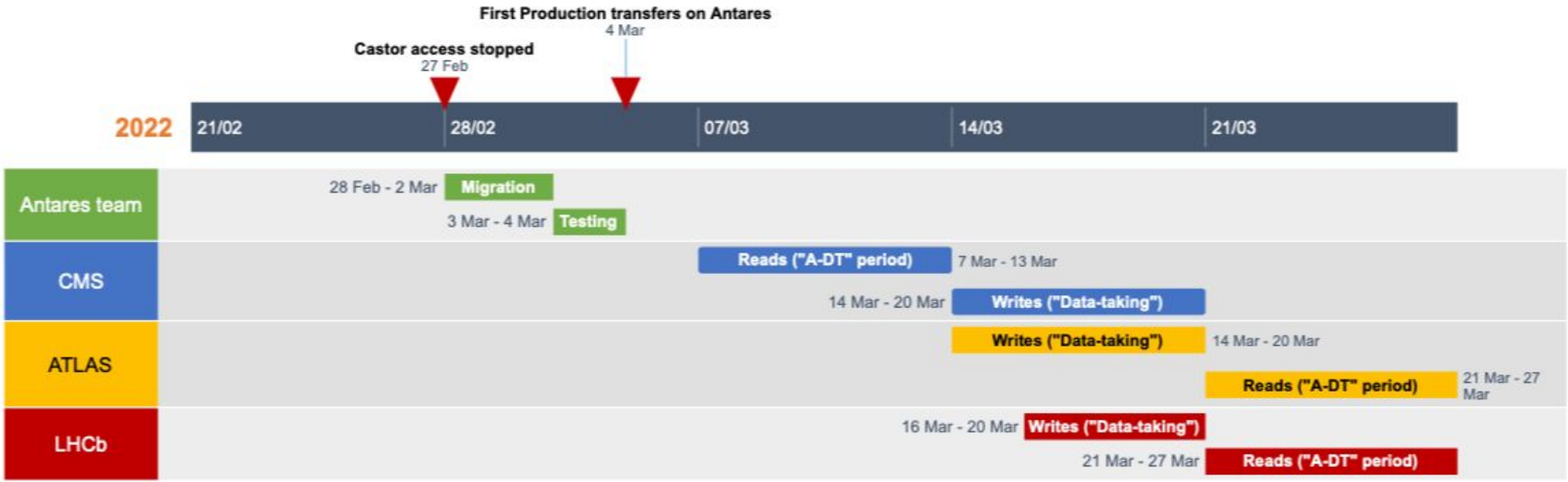
23/03/22, GridPP47

Intentions

- Primary goal: Demonstrate that each tape system can achieve the rates specified by the VOs.
  - These values are set at 10% of the estimated rates required during LHC Run 4.
  - This percentage will be gradually increased in future challenges.
  - Prove readiness for Run 3.
- Secondary goals:
  - Investigate the current limits of the tape systems.
  - Tune configuration, including FTS (File Transfer Service).
  - Observe problems - e.g. failing transfers.

https://docs.google.com/document/d/1jb9PrenXEjo2b7eOq_6wLVyOOCezW48yyJlf1cRJovw/edit#heading=h.k7c6yk8d274t

Challenge schedule

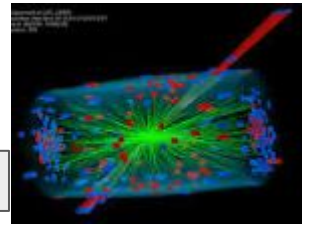


CMS

Expected rates in Run3:
WRITES during data taking - 0.9GB/s
READS outside of data taking - 1.5GB/s

CMS Rucio Antares setup

CMS Rucio name:
T0_CH_CERN_Disk



CERN EOS
“Tier 0 processing”

root

Physical tape

SSD buffer

root

?

CMS Rucio names:
T1, T2, T3 sites

CERN EOS

DE

US

webdav

IT

FR

ES

RAL Echo

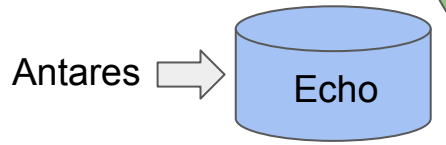
Antares

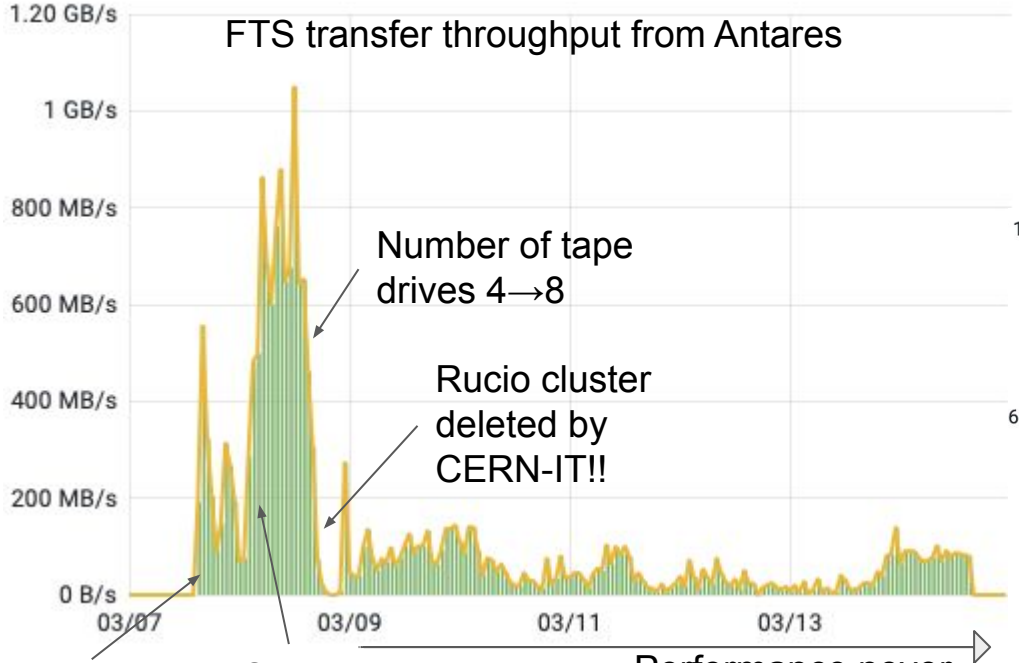
CMS Rucio name:
T1_UK_RAL_Tape

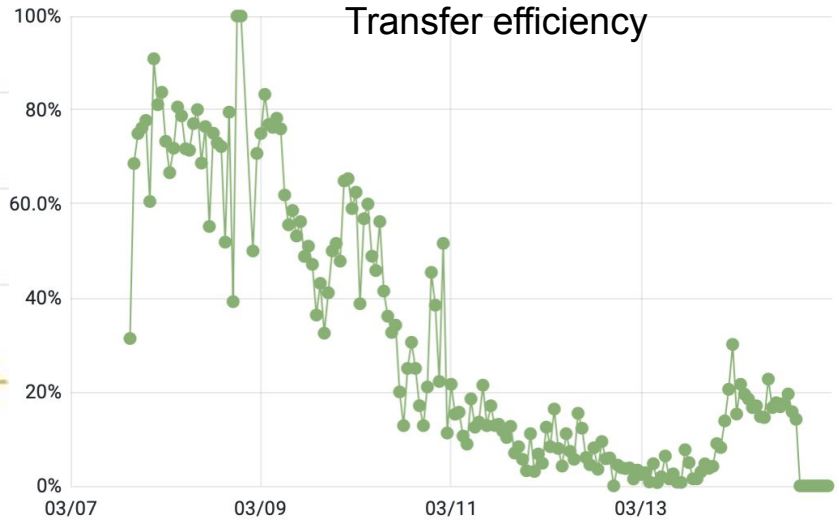
CMS Rucio name:
T1_UK_RAL_Disk

CMS reads from Antares - staging

Plots show challenge data only





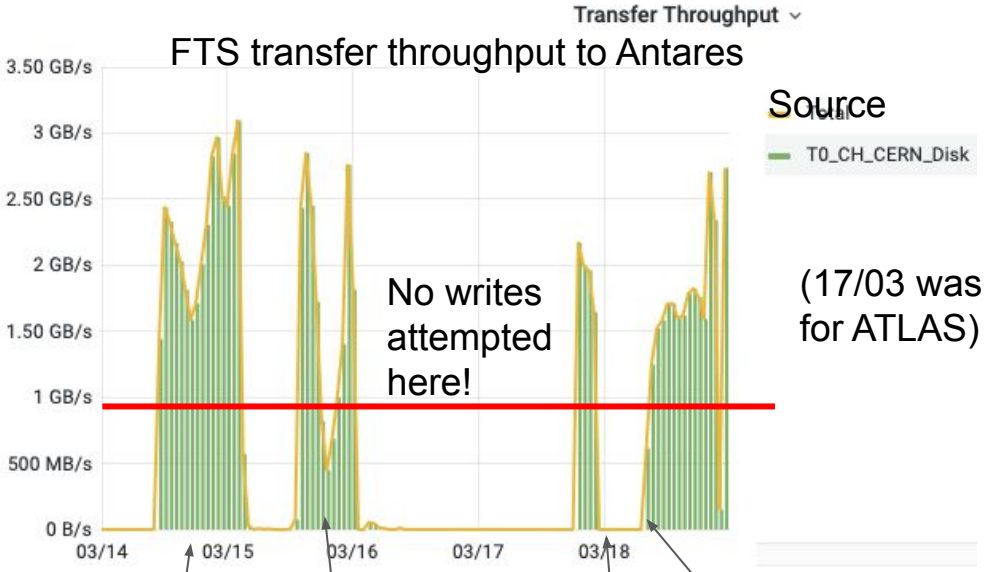


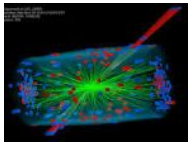
Reads - what went wrong after Rucio was brought back?

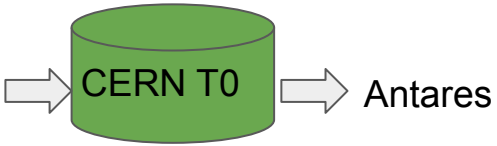
- Some previously-staged files tried to write from buffer to Echo but the files had already been deleted.
- Antares team changed some config
  - Each time this happened, CMS used the *cancel-requests* command in Rucio
  - This is meant to remove existing FTS requests and re-submit
  - However, it never removed the requests - only re-submitted (Rucio bug?)
- Also - I could see **some** successful transfers in FTS monitoring…but not acknowledged in Rucio…another bug??
- New requests (e.g. Production/non-challenge) were transferred efficiently.

Unfortunately CMS never hit the target read rate (1.5GB/s) so I expect this test to be repeated. I also want to do performance multihop tests (Antares→Echo→SiteX)

CMS writes to Antares - archiving

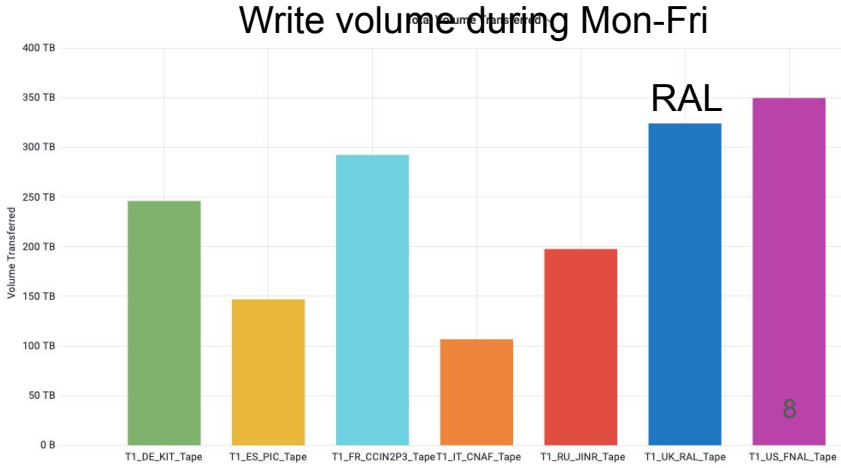






P5

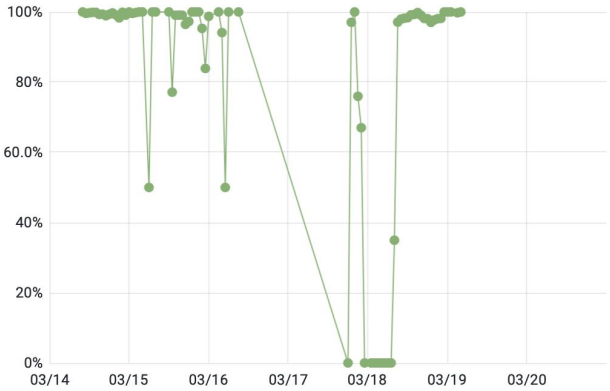
(17/03 was “RAL’s big day” for ATLAS)



Write success/failure

EOS MGM fails

Transfer efficiency



ATLAS

ATLAS Plan: DT

- Two weeks of activity: First Data-taking (DT), i.e. T0 export,
  - Followed by After DT (A-DT); ie. Nomal (non DT) activities: Data consolidation and Data Carousel
- For DT:
  - 2 days of using the SFO (at Point-1) and exercising the full data chain from P1 to T1
  - Two T1 receives a whole ‘fills’ worth of Data on ‘its turn’:
    - Rate of 3.5GB/s would keep up with data rate from SFOs
    - Other T1s receive other streams’ data at rate according to it’s MoU: e.g. 4.6GB/s x MoU
  - Remaining days; start from Cern EOS
    - Read activity from T1 as ‘usual’
  - **Thurs 17th; RAL’s turn**

Plan for DT week (March 14th~18th)

- Monday, Tuesday
  - Full chain tape write test : P1/SFO -> T0 disk/tape <-> T1 tape/disk
  - Two beam time(“fill”), each lasts for 12h (8am~8pm CET)
  - Activities during the fill
    - SFO will be generating all streams (fake) RAW data at 8GB/s.
    - T0 batch farm to produce (fake) derived data at 2GB/s
    - Data streams to CERN EOS and CTA: 10GB/s, total 864TB, file size 5GB
    - RAW data export from T0 to T1 tape : 8GB/s
      - Main stream (peak rate of 3.5GB/s) data will be sent to one T1 per fill (FZK and INFN)
      - The rest of the (delayed) streams(~4.6GB/s) will be distributed among other T1s based on MoU share
    - Derived data export from T0 to T1 disk (at 2GB/s)
    - Production tape read (from T1s) continues at lower rate
- Wednesday through Friday
  - Dedicated test of peak export rate against individual T1s : T0 EOS → T1 tape
  - Fake data from SFO will be re-used
  - FTS throughput setting will be used to control the inbound rate to T1 tape (set to 3.5GB/s)
  - Production tape read (from T1s) continues at normal rate
- Fake data to be deleted after the challenge: scope name “data_test”

date	Start time (CET)	duration	Raw data streams	Target site
March 14th	8am	12h	All streams (SFO fill1)	main stream to FZK
March 15th	8am	12h	All streams (SFO fill2)	main stream to INFN
March 16th	8am	12h	Main stream (fill1)	NDGF
	2pm	12h	Main stream (fill2)	BNL
	5pm	12h	Main stream (fill1)	TRIUMF
March 17th	8am	12h	Main stream (fill1)	PIC
	2pm	12h	Main stream (fill2)	RAL
March 18th	8am	12h	Main stream (fill1)	IN2P3
	12pm	12h	Main stream (fill2)	SARA
	2pm	12h	Main stream (fill1)	BNL(2)

Schedule of peak (export) rate test for T1s

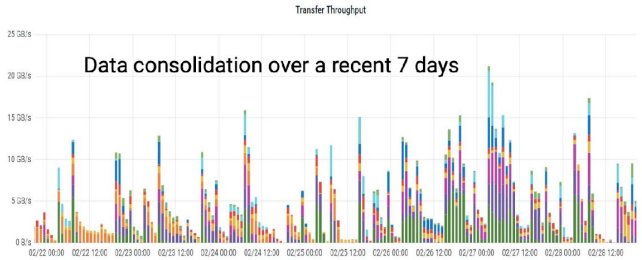
https://indico.cern.ch/event/1134753/contributions/4761039/attachments/2401429/4106944/ATLAS%20Plan.pdf

ATLAS Plan: A-DT

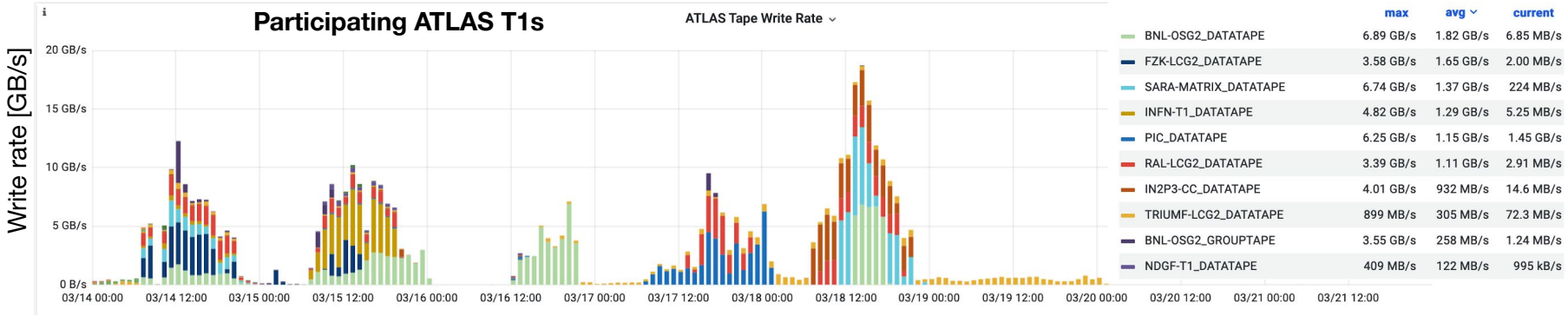
- Plan for After Data-taking activities to be running in ‘normal operations’ mode:
  - FTS tunings to be applied to control / shape throughput for each T1 according to its MoU.
- Currently in A-DT component of TAPE challenge (for ATLAS / LHCb).

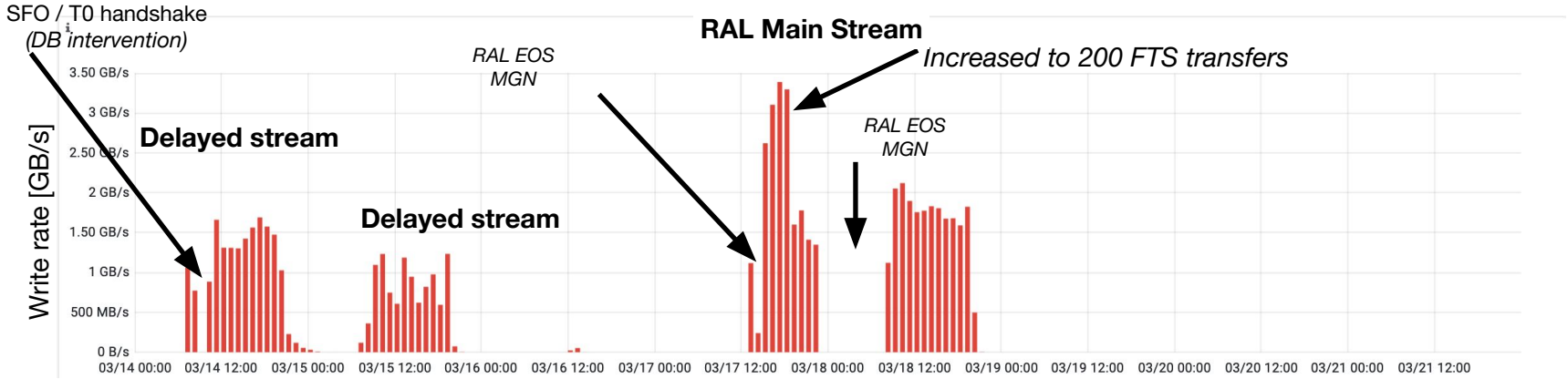
Plan for A-DT week (March 21st ~ 25th)

- Business as usual : normal tape write and tape read activities from production
  - Tape write : data consolidation
    - Run3 target : 5.1GB/s overall
  - Tape read : Data Carousel in full speed
    - Run3 target : 8.4GB/s overall
- T1s to control max tape write rate
  - Set FTS throughput limit for inbound traffic to its tape endpoint : (5.1GB/s * MoU) or above



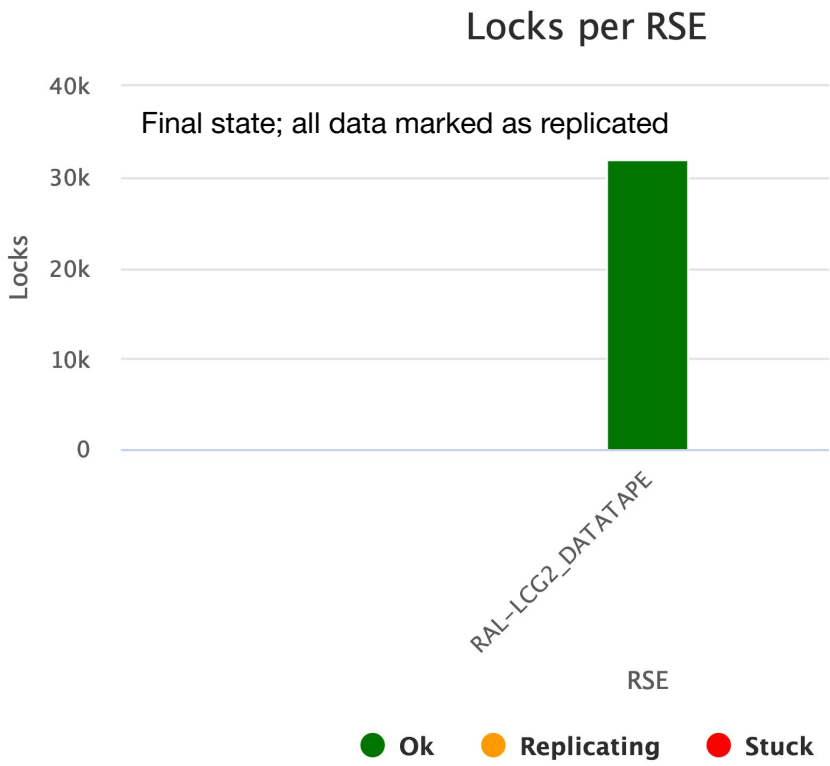
DT Week: ATLAS T0 export





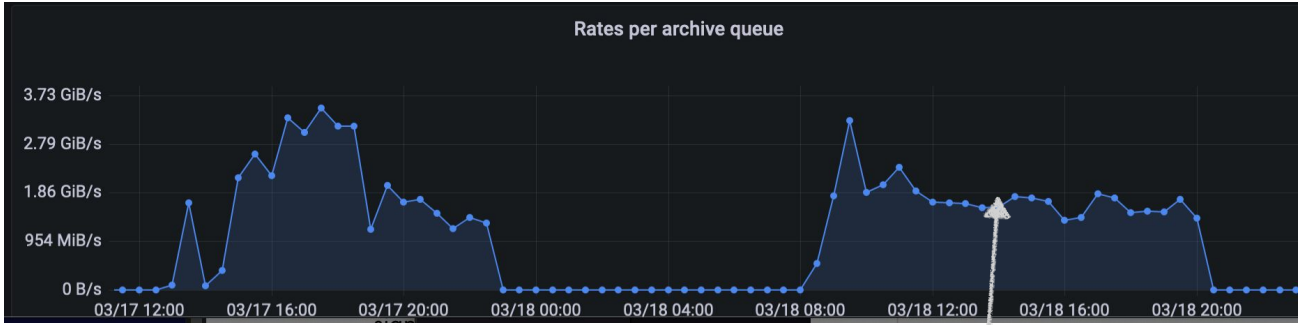
RAL Main stream details

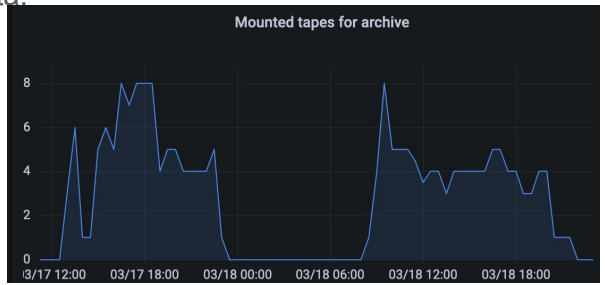
- https://rucio-ui.cern.ch/rule?rule_id=9837f9786f5f46c1a149310397eebdf9#locks
- Data out of Point 1:
  - “Main stream”; the primary physics data stream,
  - delayed stream and other reduced volume streams (e.g calibration, TLA) also produced.
  - Each T1 will get its own Main steam dataset, and rotate through in turn.
- Rule submitted: Thu, 17 Mar 2022 13:13:35 UTC
- data_test.00414125.physics_bulk.daq.RAW
- 31891 files
- ~150TB data
- ~ 5GB/file
- Data to go from CERN (big) EOS to Antares direct via xroot protocol.

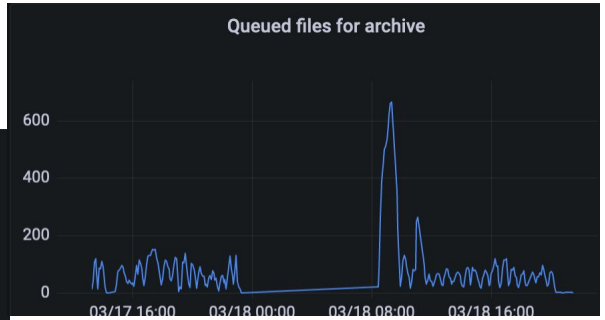


Plots from Antares

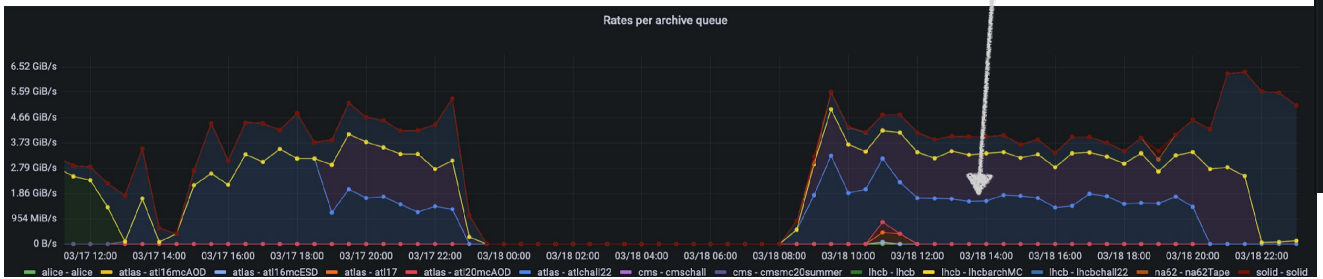
- For the Main stream for RAL; plots showing similar behaviours to ATLAS / Wlcg monitoring.
  - Peaked rate for ATLAS around 3.8 GB/s using up to 8 mounted tapes (typically 4 mounted).
- Apart from at the restart, only O(100) files every queuing for archive for the ATLAS challenge data.







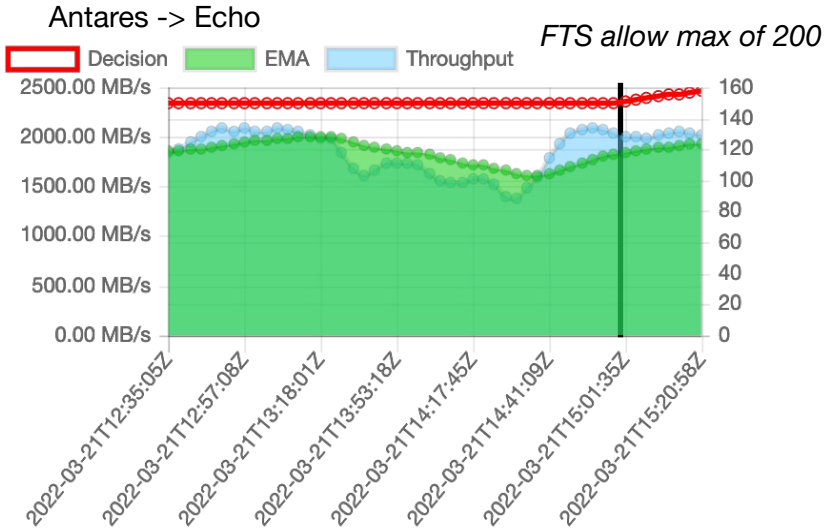
- Total average archive rate during this period (i.e including other VO's ~ 5 – 6GB/s)



After Data-taking Challenge

- Data moved via Multihop transfers through ECHO.
  - Antares -> (xroot:///) -> Echo -> (davs:///) -> Elsewhere …
- And, writing into Echo:
  - Elsewhere -> (davs:///) Echo -> (xroot:///) -> Antares
- Optimisation and tuning of Internal Xroot transfers with FTS ongoing:
  - (Noting that FTS instances are independent actors)
  - Also need to optimise the staging profile (e.g smaller buffer compared to CASTOR might require more ‘streamlined’ approach):
    - Submit more requests with fewer files / request, or,
    - Large O(100k) requests as for CASTOR.

FTS ‘Optimiser’

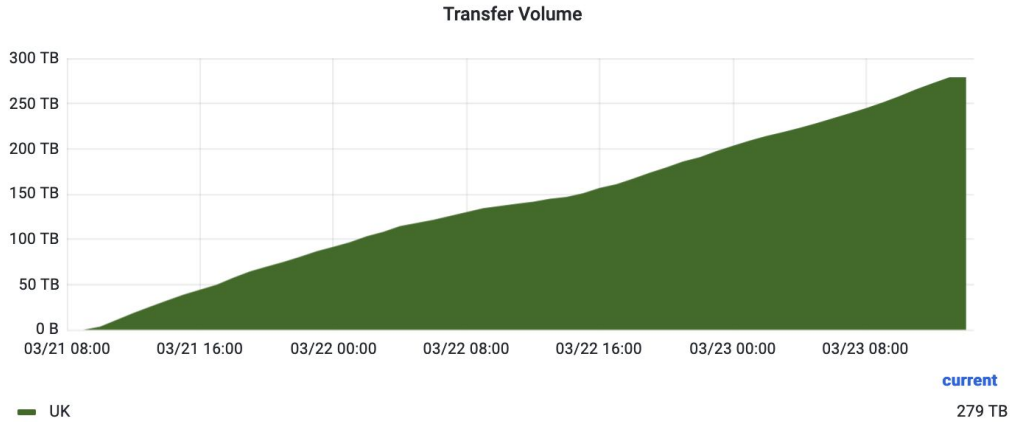


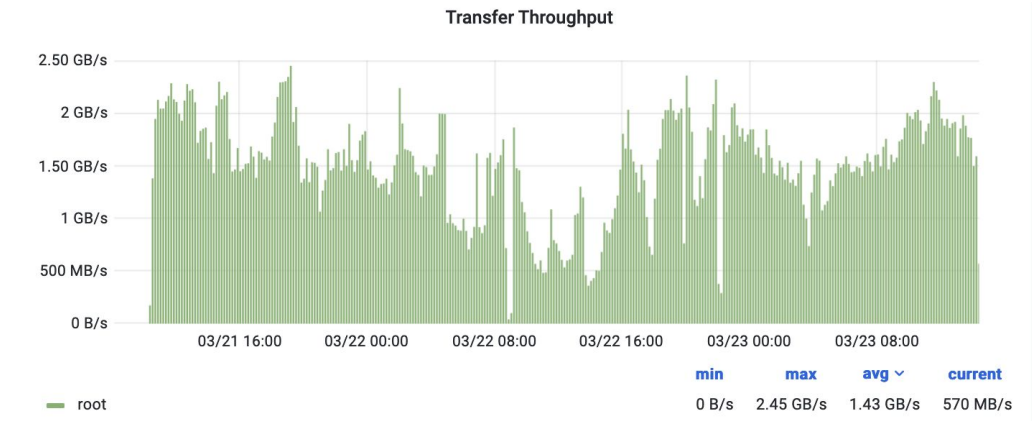
Staging from Antares

- ~100k files for Staging and transfer to Echo: 3-4 GB/s tape read rate.
  - Reached 15 tapes mounted; 230 – 330 MB/s read rate / tape observed

Transfers from Antares to ECHO

- Data transferred from Antares to ECHO via xrootd (on xrootd.echo.stfc.ac.uk):
- (150 max concurrent allowed transfers by FTS) ~ 2GB/s
  - Interesting to find the ‘real’ limit / optimal transfer number
- Some failures due to files already having been evicted from CTA

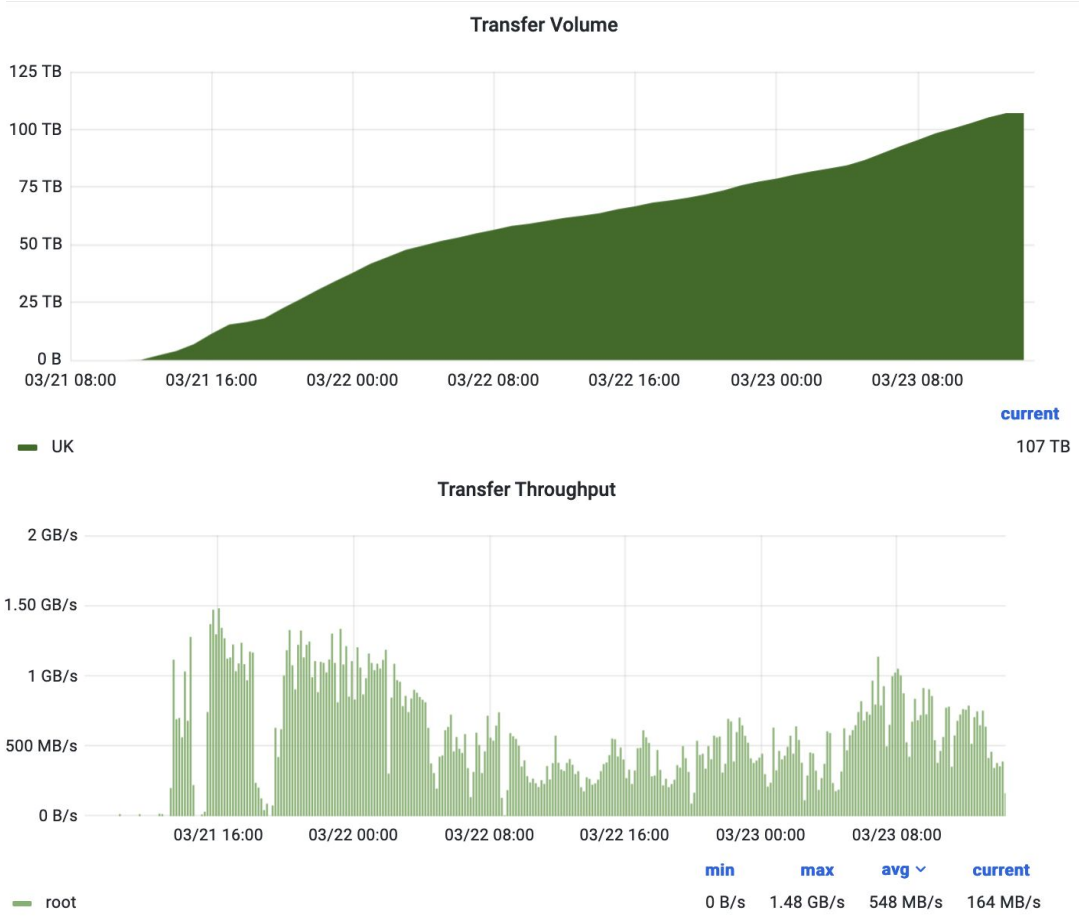




Echo to Antares

- Non “T0 export” writes to Antares go via Multihop (davs://->Echo->root://->Antares).

  - Very high efficiency until 7pm yesterday
  - Number of Operation expired: (destination)
    - Errors

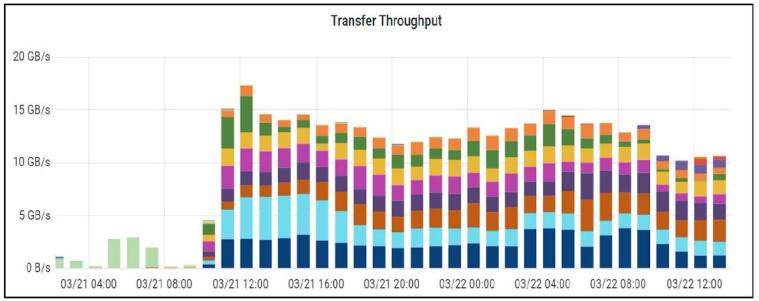


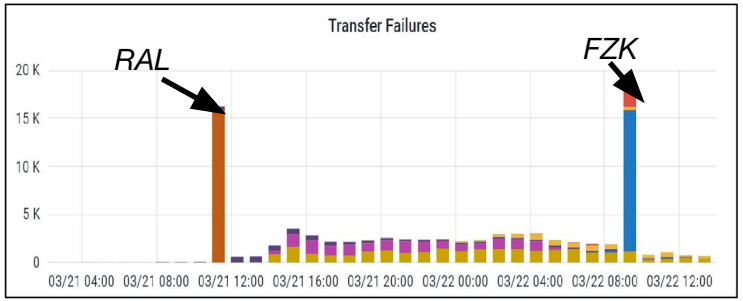
General update from A-DT

- General Summary from Xin, for the first 2-days of A-DT.

Day 2 of A-DT week

- LHCb started staging today as well
- Overall staging rate is ~13GB/s, pretty good
- Two spikes of failure rate so far, one from FZK another from RAL
  - RAL : again EOS MGM issue





https://indico.cern.ch/event/1141724/contributions/4792307/attachments/2412289/4128118/quick%20update%20of%20tape%20challenge%20%2722.pdf

ATLAS: Summary (So far)

- Data-taking part of challenge completed
  - Some items to follow up (e.g. EOS MGM for RAL), otherwise generally positive
    - ATLAS looking to run another test T0 export ~ next week.
  - After data-taking ongoing:
    - Optimisation of data rate from CTA buffer to Echo ongoing.
    - Some failures due to files on buffer being evicted before transfer
      - Some Transfer failures to also investigate
- General ATLAS overview of the TAPE challenge also positive
  - Main reports to be presented to GDB and DOMA meetings

LHCb

Based on slides kindly provided by Chris Haen from LHCb. Any errors will be due to Mark Slater’s alterations!

Challenge Setup

General ‘Data Taking’/Write mode (started 16th of March) was:

- LHCb provide ~ 2 days worth of data
- Transfer from EOS → CTA (T0)
- Transfer EOS → T1-disk → T1-tape
- Remove from T1-disk

This mimics expected data taking in Run 3 closely.

Read/Staging mode (started 22nd of March, not reported on yet)

- T1-tape → T1-disk

RAL Specifics

The specifics for RAL are:

- EOS → ECHO (via Webdav)
- ECHO → Antares (via Webdav)
- Remove from ECHO

During Run 3 this is the mode of operation for processing, i.e. copies are required on both disk and tape

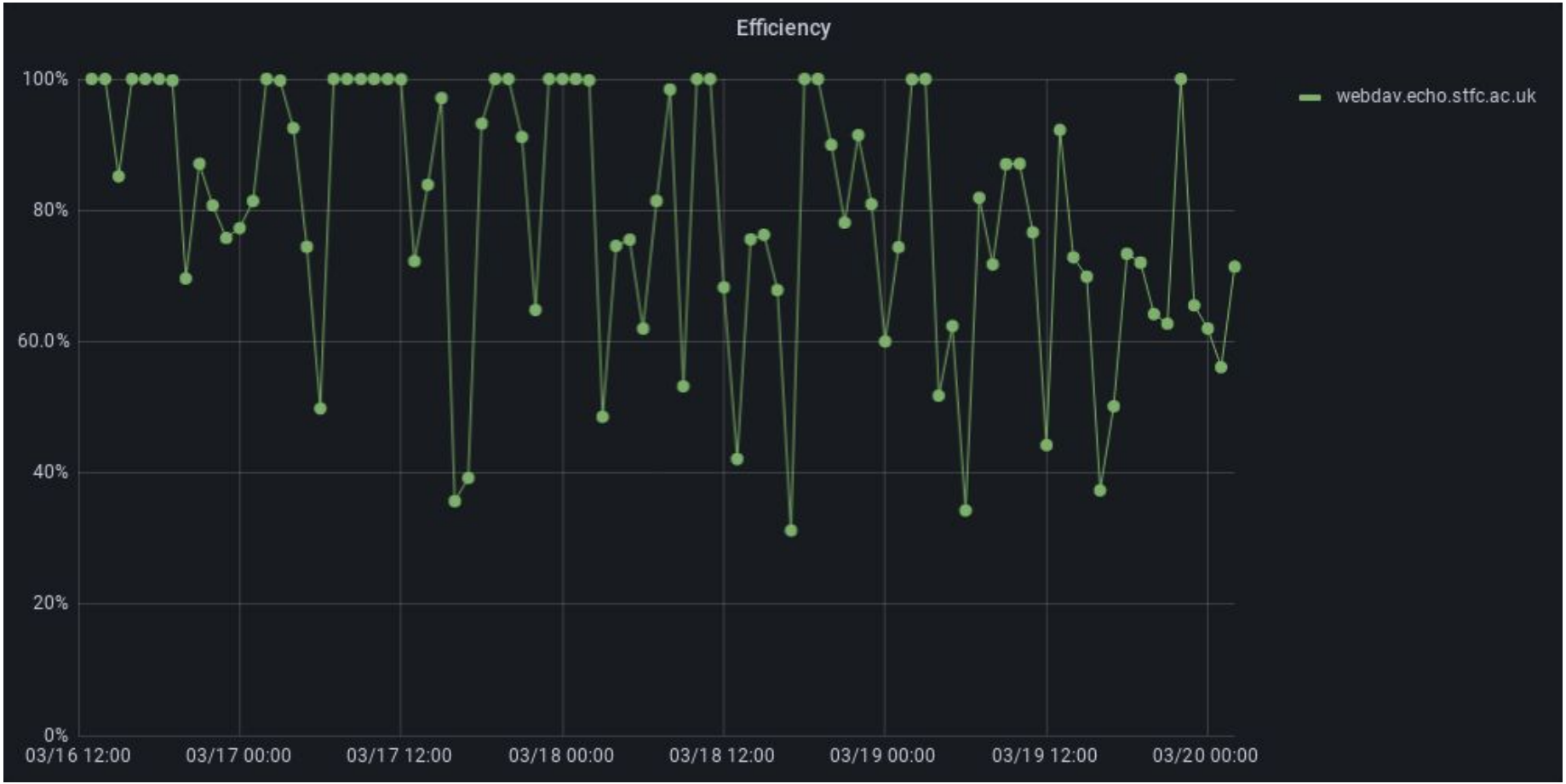
Note that this doesn't include **direct** TPC EOS → Antares (via Webdav) which will also happen for preservation purposes

Site	shares	Data Written (TB)	Export Speed (GB/s)	Staging Speed (GB/s)	Staging Duration (hours)
CERN		2117	11.00	1.90	
CNAF	15.7%	331.91	1.72	1.35	68.43
GRIDKA	20.3%	429.88	2.23	1.36	87.98
IN2P3	11.3%	240.26	1.25	0.98	68.43
NCBJ	12.0%	253.49	1.32	0.91	76.98
PIC	1.8%	38.24	0.20	0.17	61.59
RAL	26.9%	570.33	2.96	1.93	82.11
RRCKI	2.2%	47.51	0.25	0.21	61.59
SARA	9.7%	205.37	1.07	0.74	76.98
Total Tier1s	100.0%	2117.00	11.00	7.65	

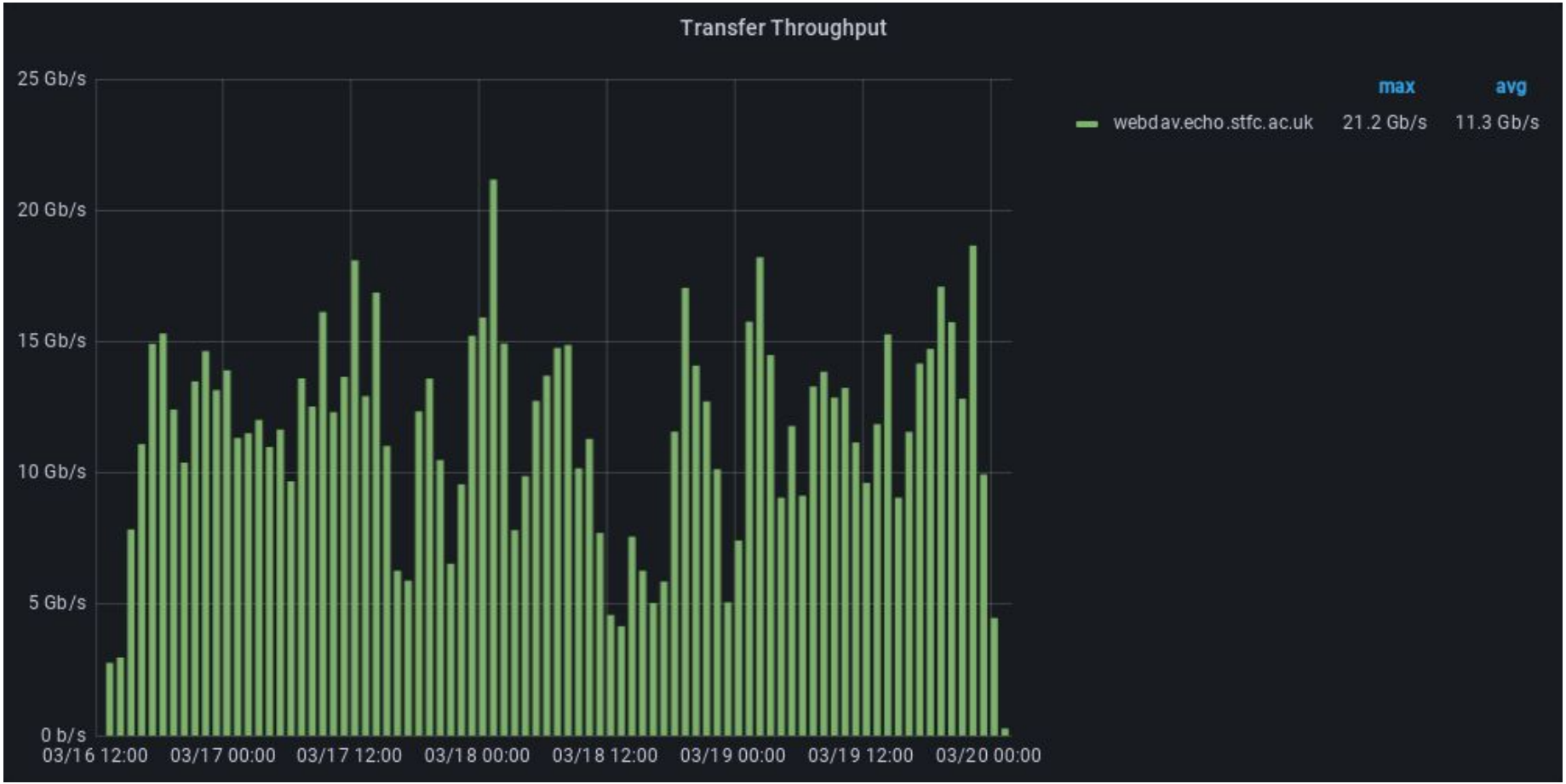
Planned Distribution of Data During TC

Site	expected Speed (GB/s)	Max Speed (GB/s) disk/tape	Avg Speed (GB/s) disk/tape	Duration (hours) disk/tape	Success
CERN	11.00	NaN/13.13	NaN/11.31	NaN/38	Yes
CNAF	1.72	2.24/3.43	1.5/1.21	45/56	~Yes
GRIDKA	2.23	4.46/3.46	2.24/2.03	39/43	Yes
IN2P3	1.25	2.99/3.03	1.31/1.13	37/43	Yes
NCBJ	1.32	0.84/NaN	0.64/NaN	82/NaN	No
PIC	0.20	0.77/0.98	0.21/0.22	37/36	Yes
RAL	2.96	2.65/4.11	1.41/1.28	86/85	~No
RRCKI	0.25	0.57/0.30	0.29/0.13	22/49	Yes
SARA	1.07	3.16/2.23	1.12/1.04	37/40	Yes
Total Tier1s	11.00				

Write Performance Overview

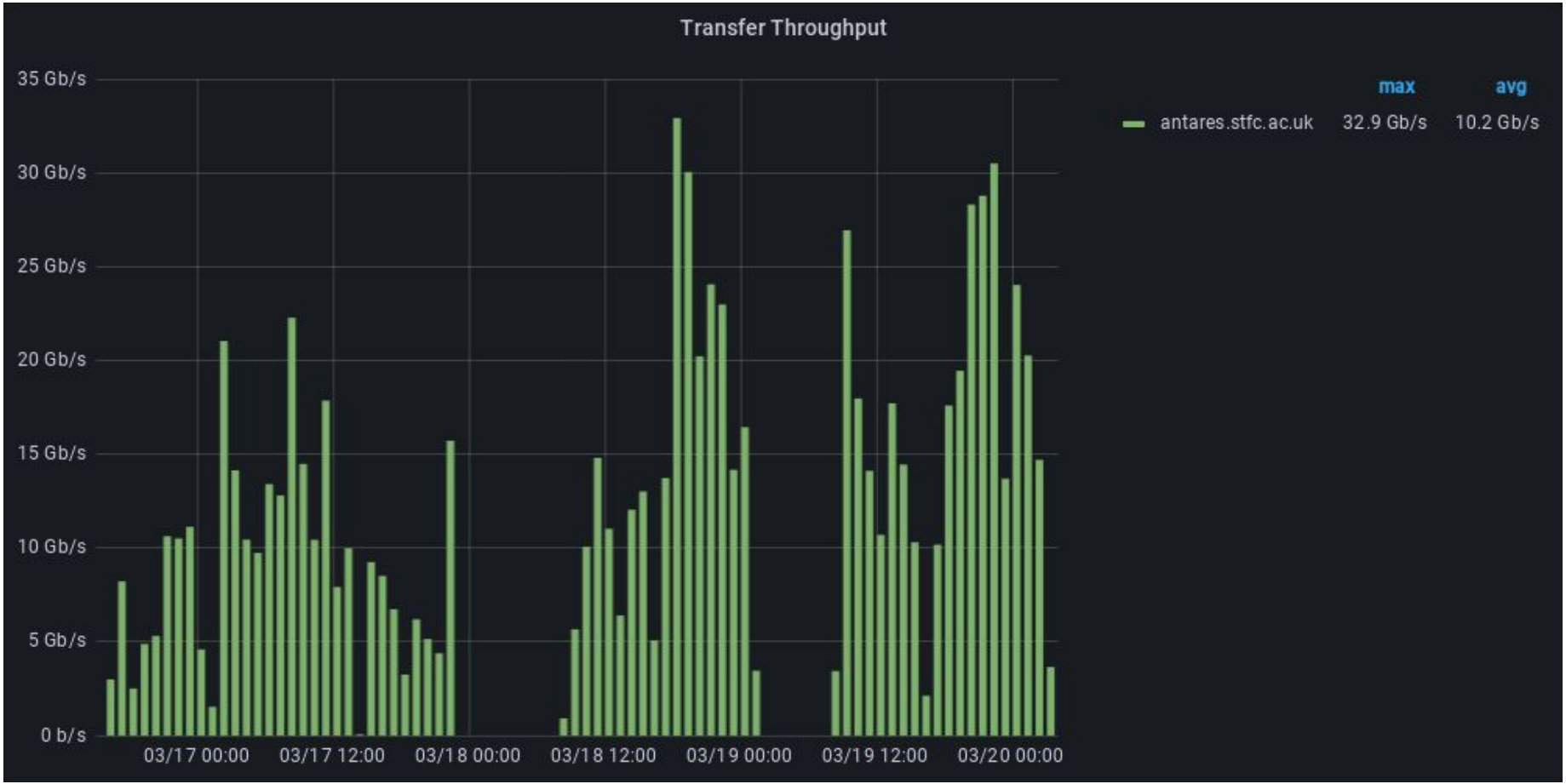


EOS → ECHO Efficiency



EOS → ECHO Throughput (Gb/s rather than GB/s)

ECHO → Antares Efficiency



ECHO → Antares Throughput (Gb/s rather than GB/s)

More Details on RAL Performance

- Antares suffered 2 outages during the challenge which lowers the overall **TAPE** efficiency (not disk)
- Discounting those, the efficiency was very good for a new service
- ECHO efficiency was affected by some issues with webdav. More gateways are planned to alleviate these problems (GGUS:156277)
- Overall throughput was a factor 2 below target but reasons for this are understood

From the LHCb POV

- It was a good stress test
- Made good use of lessons from DC1
- Some issues on the LHCb side with data ‘running out’
  - Partially due to a bug (now fixed)
  - Partially due to the (fake) distribution strategy
- SRM + HTTPS validated
- Happy to redo tests with sites when they wish